Year: 2024

Where rare meets common: Leveraging population cohorts to study rare copy-number variants

Auwerx Chiara Maria Paula

Auwerx Chiara Maria Paula, 2024, Where rare meets common: Leveraging population cohorts to study rare copy-number variants

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_1BFE79774F1F5

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Centre Intégratif de Génomique &
Département de Biologie Computationnelle

Where rare meets common: Leveraging population cohorts to study rare copy-number variants

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Chiara Maria Paula AUWERX

Master de science en biologie
Eidgenössische Technische Hochschule Zürich (ETHZ).

Jury

Prof. Philipp Engel, Président
Prof. Alexandre Reymond, Directeur de thèse
Prof. Zoltán Kutalik, Co-directeur de thèse
Prof. Tim Frayling, Expert
Prof. Michael Talkowski, Expert

Lausanne
(2024)

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Ecole Doctorale
Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Monsieur	Prof.	Philipp	**Engel**
Directeur·trice de thèse	Monsieur	Prof.	Alexandre	**Reymond**
Co-directeur·trice	Monsieur	Prof.	Zoltán	**Kutalik**
Expert·e·s	Monsieur	Prof.	Tim	**Frayling**
	Monsieur	Prof.	Michael	**Talkowski**

le Conseil de Faculté autorise l'impression de la thèse de

Chiara Maria P. Auwerx

Master of Science ETH in Biology, ETHZ, Zürich

intitulée

Where rare meets common: Leveraging population cohorts to study rare copy-number variants

Lausanne, le 28 juin 2024

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Philipp Engel

doctoral thesis

Where rare meets common: Leveraging population cohorts to study rare copy-number variants

CHIARA AUWERX

Spring 2024

University of Lausanne

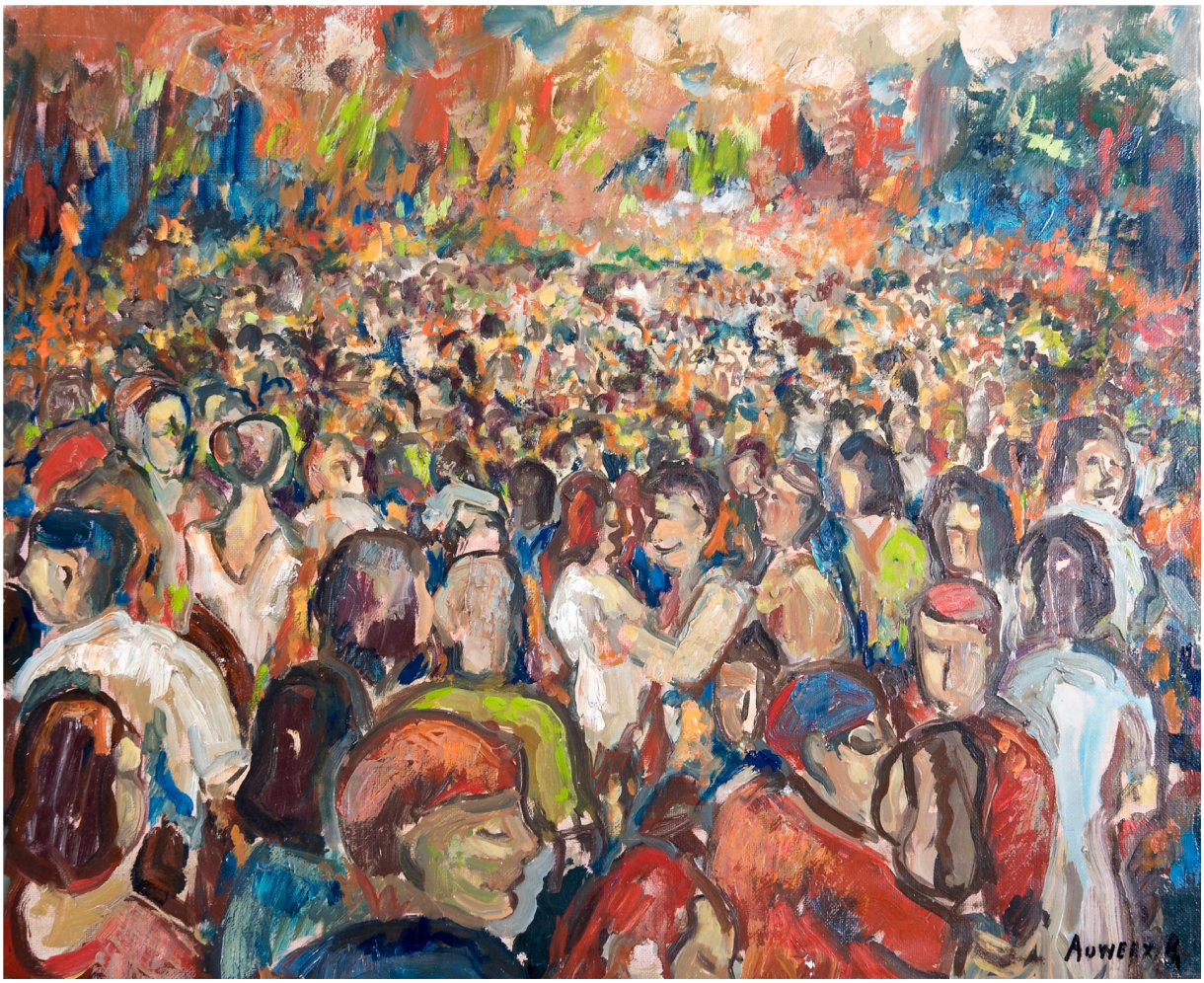
WHERE RARE MEETS COMMON:
LEVERAGING POPULATION COHORTS TO STUDY RARE COPY-NUMBER VARIANTS.
2024
Chiara Maria Paula Auwerx

Publisher
First printed in 2024 by University of Lausanne

Preface

We must understand life as it is and understand that diversity is its most essential feature.

– Mary Parker Follett



Ever since childhood, I have been fascinated by the incredibly many forms that life can take. By acting as a substrate for evolution, genetic diversity has been intertwined with life since its origin, defining taxonomic branches down to its smallest unit, the species. In humans, diversity manifests itself in terms of appearance, predisposition to diseases, culture, interests, and behavior. While this multiplicity is at the root of many conflicts, it also represents a source of growth and strength, and only by embracing it can we truly grasp the complexity of our world. As such, understanding how variation encoded in our genomes is translated into the tremendous diversity observed in our species represents one of the most enticing but also challenging questions in biology. And what better way is there to do so, than a PhD in genetics?

Painting by Rene Auwerx. Reproduced with permission for https://www.auwerx.org/.

Acknowledgments

Art is I, Science is We.

– Claude Bernard

During our first feedback meeting, Zoltán told me: "I had some reservations at first, but now my doubts are gone". These words resonated with me, as for me too, the transition from wet lab to computational biology was daunting. Yet, this journey has led me onto a path I love and want to pursue. For that, I am grateful.

First and foremost, I am grateful to **Zoltán** and **Alex** for taking a bet on me and introducing me to the world of human genetics. Like yin and yang (I let you decide who is who :D), you both have unique and complementary strengths that combined, make for the best supervision one could ask for. **Zoltán**, your quick thinking and statistical knowledge will never cease to impress me, but what I admire most is your ability to transmit this knowledge without ever losing patience. As a mentor, you are highly dependable and help people develop their strengths and reach their full potential. Even after four years, I continue to learn from you and I will certainly miss our weekly meetings. **Alex**, your curiosity is boundless and reflected in the many anecdotes you have to share on anything related to genetics and biology. As a supervisor, it allows you to bring a fresh eye to scientific questions and more importantly, leave your students the freedom to explore and forge their own paths. Yet, you still provide guidance and I am truly appreciative of the many conferences that I was able to attend, where you introduced me to leaders in the field.

I also want to thank my unofficial mentors, **Eleonora**, **Jolanda**, and **Katrin**, for always having an open ear, being supportive, and providing a female perspective on career advice in academia and beyond. A special thanks to **Eleonora**. I still remember the day I met you in office 3032. You recommended a co-supervision by Zoltán and Alex, which turned out to be one of the best career advice I have ever received!

The PhD also offered me the opportunity to supervise others. **Charlie**, **Nicolò**, **Caterina**, and **Samuel**, I hope that you enjoyed as much as I did our time working together. In particular, I would like to thank **Sam**. Supervising you was one of the most enriching parts of my PhD. Looking back, I am glad that you decided to join the group to pursue your own PhD, keeping me company in the upstairs office. Not only are you a dedicated scientist, but also an extraordinarily generous, reliable, and nice person.

Through these four years, I became convinced that not only great science but also great people make for a successful PhD. I found those in the **Statistical Genetics Group (SGG)** and beyond. I could not have wished for a better PhD companion than **Marie**. You are a brilliant scientist and an amazing person and I feel incredibly lucky to have shared my PhD with you. I have learned a lot – both about science and life – from you and I can say without a doubt that our collaborations, coffee breaks, retreats, and conferences have made my PhD journey more enriching, fun, and successful. I would also like to thank **Liza**, for making every newcomer feel welcome. You are one of the most helpful people I have ever met and your many social initiatives were key to building the group's cohesion. In some way, **Leona**, you have taken over the task of representing and keeping the SGG alive. While I might not be as athletic as you, I recognize myself in your dedication and love for order and always enjoyed discussing politics and philosophy with you over lunch. **Sven** and **Marion**, I am so glad that our group "adopted" you. **Sven**, your kindness and smarts are unparalleled, and **Marion** I admire your generosity and resilience (and also your enthusiasm for board games!). But above all, I am grateful for our friendships that lasted beyond your time at Unil. **Adriaan**, thank you for boycotting home office and being the most reliable lunch partner. I appreciate your honesty, good mood, humbleness, and knowledge about random facts, which make for the most entertaining conversations. These might only be

topped by **Tabea's** crazy life stories, which could be turned into a movie. You joke about wanting to be an eternal postdoc, but you are so resourceful that if one person can manage this, it is you. **Kaido**, your short stay in the group was sufficient to reveal the calm, reflective, and genuinely kind person you are, always willing to help. **Robin**, your creativity makes it inspiring to bounce scientific ideas back and forth and I am glad you decided to join our group. Finally, **Diana**, I think we can consider you as an unofficial member of the SGG, given the many retreats and conferences we spent together. You are a bit crazy but I love that about you!

From the third floor, I would like to thank **Jacqueline**, for showing me around whenever I needed to find something and **Clara** for motivating me to pick up yoga. **Sissy**, I am glad I got the chance to know the sincere and caring person you are by being your ASHG conference roommate. **Shanaz**, I am grateful for your company in the office. Your good mood and open ear for gossip and random discussions have often brightened my day. Finally, a special thanks to **Corinne**, **Mariona**, and **Michelle**, who were always helpful and smoothly helped me navigate the administrative hurdles of being affiliated with multiple departments.

Outside of the University of Lausanne, I also got the chance to collaborate with terrific people. **Maarja**, thank you for always promptly helping me replicate my findings in the Estonian Biobank and involving me in some of your own projects. My only regret is to not have had the chance to meet you in person. **Malú**, I loved to have another person working on CNVs in the group and enjoyed our collaboration. You are a super nice person and I am glad to have built a lasting connection with you. **Ruth**, collaborating with you felt energizing and I am still impressed by how efficiently we coordinated our efforts. I learned a lot from your clinical knowledge and enjoyed the translational angle of the 16p11.2 BP2-3 project. **Fabrice** and **Guillaume**, thank you for helping us take our findings to the next level and validating them *in vivo*.

Lastly, I want to thank my whole family. **Mama**, **Papa**, how lucky was I to grow up in a family of scientists? Be hardworking, stay curious, follow your passion. I certainly owe many of my successes to the values you taught me since I was a kid and I am glad that I could celebrate some milestones, such as my first talk at a conference or published paper, with you. But even more importantly, you showed me the meaning of unconditional love and support. You are my biggest supporters and I know that I can always turn to you for advice. For that, I am truly grateful. **Hannah**, while you have a special gift to drive me mad, I am also constantly impressed by your energy and determination and I know that you'll achieve whatever you set your mind to. But you also have a big heart and I know that I can count on you during hard times. **Andi**, you were there every day, for the highs and the lows, which you know all too well can be more extreme than on a roller coaster. While your rational view of the world can sometimes be beyond my comprehension, it often helped me to see the forest for the trees. You help me find balance and bring the best out of me and I know that with you by my side, I can accomplish more. Among your many qualities, you have always understood the large place my career takes in my life and I am thankful that you support my dreams and ambitions. I look forward to our next adventures together!

Thank you! Merci! Dank u! Danke!

Abstract

Completion of the Human Genome Project has democratized access to genomic data and led to the creation of large biobanks that couple genotype information with phenotypic and medical data for hundreds of thousands of individuals. The latter fostered the boom of genome-wide association studies (GWASs) that aim at identifying associations between common single nucleotide polymorphisms (SNPs) and complex traits. In parallel, studies in clinical cohorts ascertained for neurodevelopmental disorders have revealed that large, recurrent copy-number variants (CNVs) represent the etiology of various genomic disorders. CNVs are a class of structural variants defined by the deletion or duplication of large (> 50 bp) DNA fragments. Despite their evident relevance to human health, technical challenges linked to their detection have prevented assessment of their presence and phenotypic consequences in the general population.

The research conducted during my PhD aimed at filling this gap. Firstly, we called CNVs based on microarray data for 500,000 individuals from the UK Biobank. About 40% of individuals carried at least one high confidence CNV. We then developed a framework to conduct **CNV-GWAS** and applied it to test the association between the copy number of CNV-proxy probes and 117 medically relevant quantitative traits and complex disease diagnoses. We identified over 200 independent CNV-trait associations, as well as a negative impact of a high CNV load on an individual's disease burden, socio-economic status, and proxied lifespan, suggesting profound repercussions of this mutational class on global health. Follow-up studies revealed general patterns related to the CNV architecture of complex traits, as well as insights into the epidemiology and biology of specific examples. First, our signals colocalized with both common SNP-GWASs and rare Mendelian disorder genes, suggesting **phenotypic convergence** of different genetic lesions at the same locus. Second, CNVs exhibited **variable expressivity**, illustrated by well-established pathogenic CNVs leading to subclinical phenotypic alterations and heterozygous CNVs causing phenotypic changes reminiscent of the recessive associated disease. Overall, the same physiological systems are implicated by both clinical and population studies, suggesting that the same CNV can generate a spectrum of phenotypic consequences with variable degrees of severity. Third, many multi-genic CNVs exhibited high levels of **pleiotropy**. Dissection of these signals revealed insights into molecular mechanisms, highlighting putative drivers for specific phenotypes. Fourth, CNVs were found to act through **different dosage mechanisms**, even at the same locus. Fifth, CNVs cause **early disease onset**, in line with the earlier onset of diseases with a genetic etiology and supporting the view that common diseases represent aggregates of multiple rarer conditions.

The second part of my thesis focused on three genomic disorders with variable expressivity caused by CNVs mapping to chromosomes 22q11.2 (DiGeorge syndrome), 16p11.2 BP2-3, and 16p11.2 BP4-5. Specifically, we used phenome-wide association scans (PheWASs) to reveal the full pleiotropic spectrum of these CNVs. We leveraged data from known Mendelian disorders, rare protein-coding burden tests, SNP-GWASs, and gene expression to gain insights into driver genes and the **molecular mechanisms** of uncovered associations. In a second time, we used causal inference approaches, such as Mendelian randomization, and matched-control analyses, to disentangle **direct pleiotropic effects from secondary consequences** of the CNV's impact on intermediate mediator traits, such as adiposity levels and socio-economic factors.

Overall, the body of work presented in my thesis provides numerous methodological aspects that help to address challenges linked to performing association studies with rare CNVs in the general population. More importantly, it sheds light on the previously underappreciated mechanisms through which rare CNVs contribute to shaping human traits and disease risk, with important implications in terms of personalized medicine.

Résumé

L'achèvement du *Human Genome Project* a démocratisé l'accès aux données génomiques, entraînant la création de biobanques qui associent informations génotypiques à des données phénotypiques et médicales pour des centaines de milliers d'individus. Ces dernières ont favorisé l'essor des études d'association pangénomique (en anglais *genome-wide association study*, GWAS) qui visent à établir des liens entre polymorphismes nucléotidiques (en anglais *single nucleotide polymorphism*, SNP) et traits complexes. En parallèle, des études menées dans des cohortes cliniques de troubles neurodéveloppementaux ont révélé que des variants récurrents du nombre de copies (en anglais *copy number variation*, CNV) représentaient l'étiologie de divers troubles génomiques. Les CNVs sont une classe de variants définis par la délétion ou la duplication de long fragments d'ADN. Malgré leur importance évidente pour la santé humaine, les défis techniques liés à leur détection ont empêché d'évaluer leur présence et leurs conséquences phénotypiques dans la population générale.

Les recherches menées dans le cadre de ma thèse ont visé à combler cette lacune. Nous avons détecté des CNVs sur la base de données de puces à ADN pour 500'000 individus de la *UK Biobank*. Environ 40% des individus étaient porteurs d'au moins un CNV. Nous avons ensuite développé une méthodologie pour effectuer des GWAS entre CNV et 117 traits quantitatifs et diagnostics de maladies complexes. Nous avons identifié plus de 200 associations, ainsi qu'un impact négatif des CNVs sur la comorbidité, le statut socio-économique, et l'espérance de vie, suggèrant de profondes répercussions sur la santé de ce type de variants. Des analyses complémentaires ont révélé des tendances générales décrivant l'architecture CNV des traits complexes, ainsi que de nouvelles connaissances sur l'épidémiologie d'exemples spécifiques. Premièrement, nous observons une **convergence phénotypique** de différentes lésions génétiques communes et rares sur le même locus. Deuxièmement, les CNVs présentent une **expressivité variable**, illustrée par des CNVs pathogènes entraînant parfois seulement des altérations subcliniques, ainsi que des CNVs hétérozygotes entraînant des changements évoquant des maladies récessives. Globalement, les mêmes systèmes physiologiques sont impliqués par les études cliniques et les études de population, suggèrant qu'un même CNV peut générer un spectre de conséquences à sévérité variable. Troisièmement, la plupart des CNVs multigéniques récurrents sont **pléiotropiques**. Quatrièmement, les CNVs agissent par le biais de **mécanismes de dosage distincts**. Cinquièmement, les CNVs entraînent une **apparition précoce de maladies**, conformément à leur étiologie génétique et à la thèse selon laquelle les maladies communes représentent des agrégats de plusieurs affections plus rares.

La deuxième partie de ma thèse se concentre sur trois troubles génomiques à expressivité variable causés par des CNVs des chromosomes 22q11.2 (syndrome de DiGeorge), 16p11.2 BP2-3, et 16p11.2 BP4-5. Nous avons utilisé des études d'association phénomique pour révéler le spectre pléiotropique de ces CNVs. Se basant sur des données provenant de troubles mendéliens connus, de tests d'association pangénomique et d'expression génique, nous avons mis en évidence des gènes présumés responsables. Dans un deuxième temps, nous avons utilisé des approches d'inférence causale, comme la randomisation mendélienne, et des analyses de contrôle appariés, pour **distinguer les effets pléiotropiques directs des conséquences secondaires** des CNVs sur des traits médiateurs, tels que le taux d'adiposité et les facteurs socio-économiques.

En conclusion, ma thèse présente de nombreux éléments méthodologiques permettant de surmonter les difficultés inhérentes à la réalisation d'études d'associations avec des CNV rares dans la population générale. Cette recherche met en lumière les mécanismes jusqu'ici sous-estimés par lesquels les CNV rares façonnent les traits humains et le risque de maladie, avec d'importantes implications en termes de médecine personnalisée.

Contents

Preface iii

Acknowledgments v

Abstract vii

Résumé ix

Contents xi

Introduction 1

1 Introduction 3
1.1 The human genome . . . 4
1.1.1 DNA as a vehicle to store genetic information . . . 4
1.1.2 Inheritance of genetic information . . . 6
1.1.3 The landscape of human genetic variation . . . 7
1.2 Biobanks . . . 10
1.2.1 Clinical cohorts . . . 11
1.2.2 Birth cohorts . . . 11
1.2.3 Healthcare cohorts . . . 12
1.2.4 Population cohorts . . . 12
1.3 Link genotype to phenotype . . . 15
1.3.1 Basic statistical concepts behind GWASs . . . 15
1.3.2 Fixed effect models . . . 17
1.3.3 Reporting & interpreting GWASs . . . 18
1.3.4 GWAS model extensions . . . 20
1.3.5 Leveraging molecular phenotypes . . . 26
1.4 Copy-number variants . . . 31
1.4.1 CNV mechanisms . . . 31
1.4.2 CNV detection tools . . . 33
1.4.3 Functional consequences of CNVs . . . 37

Developing a framework for CNV-GWAS 45

2 Quantitative traits 47
2.1 Aims . . . 47
2.2 Key Findings . . . 48
2.3 Author Contributions . . . 48
2.4 The individual and global impact of copy-number variants on complex human traits . . . 49

3 Common diseases 83
3.1 Aims . . . 83

3.2 Key Findings . . . 84
3.3 Author Contributions . . . 84
3.4 Rare copy-number variants as modulators of common disease susceptibility . . . 85

Approaches to dissect the pleiotropy of recurrent genomic rearrangements 125

4 22q11.2 127
4.1 Aims . . . 127
4.2 Key Findings . . . 127
4.3 Author Contributions . . . 128
4.4 The impact of 22q11.2 copy-number variants on human traits in the general population . . . 129

5 16p11.2 BP2-3 147
5.1 Aims . . . 147
5.2 Key Findings . . . 147
5.3 Author Contributions . . . 148
5.4 Chromosomal deletions on 16p11.2 encompassing *SH2B1* are associated with accelerated metabolic disease . . . 149

6 16p11.2 BP4-5 173
6.1 Aims . . . 173
6.2 Key Findings . . . 173
6.3 Author Contributions . . . 174
6.4 Disentangling mechanisms behind the pleiotropic effects of proximal 16p11.2 CNVs . . . 175
6.5 The pleiotropic spectrum of proximal 16p11.2 CNVs . . . 194

Discussion 227

7 Discussion 229
7.1 Lessons learned from CNV-GWAS . . . 229
7.1.1 Methodological advances . . . 229
7.1.2 Beyond CNV-GWAS . . . 230
7.1.3 The future of CNV-GWAS . . . 231
7.2 From global patterns to translational knowledge . . . 231
7.2.1 Pleiotropy . . . 232
7.2.2 Molecular mechanisms . . . 232
7.2.3 Variable expressivity . . . 233
7.3 Perspectives . . . 233
7.3.1 Mechanisms of CNV action . . . 233
7.3.2 Modulators of CNV impact . . . 236
7.3.3 Clinical translation . . . 240
7.4 Conclusions . . . 241

Bibliography 243

List of abbreviations

I here provide a list of abbreviations used recurrently throughout my dissertation. Some abbreviations, which are only used once but whose acronym is more common than the full spelling, may not be listed here.

A	**ADHD**	Attention-deficit hyperactivity disorder
	AGRE	Autism Genetic Resource Exchange
	AKI	Acute kidney injury
	ALP	Alkaline phosphatase
	ALT	Alanine aminotransferase
	ApoB	Apolipoprotein B
	ASD	Autism spectrum disorder
	AST	Aspartate aminotransferase
	ATAC-seq	Assay for transposase-accessible chromatin sequencing
B	**BAF**	B allele frequency
	BDNF	Brain-derived neurotrophic factor
	BioVU	Vanderbilt University Medical Center's biobank
	BMI	Body mass index
	bp	Base pair
	BP	Breakpoint
C	**CAKUT**	Congenital anomalies of kidney and urinary tract
	CGH	Comparative genomic hybridization
	ChIP-seq	Chromatin immunoprecipitation sequencing
	CHUV	Lausanne University Hospital (Centre Hospitalier Universitaire Vaudois)
	CI	Confidence interval
	CKD	Chronic kidney disease
	cM	Centimorgans
	CMA	Chromosomal microarray analysis
	CN	Copy number
	CNV	Copy-number variant
	CNVR	CNV region
	COPD	Chronic obstructive pulmonary disease
	CoxPH	Cox proportional hazards
	CRISPR	Clustered regularly interspaced short palindromic repeats
	CRP	C-reactive protein
D	**DD/ID**	Developmental delay and intellectual disability
	DECIPHER	Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources
	DEE	Developmental and epileptic encephalopathies
	DNA	Deoxyribonucleic acid
E	**EA**	Educational attainment
	eGFR	Estimated glomerular filtration rate
	EHR	Electronic health record
	ENIGMA-CNV	Enhancing NeuroImaging Genetics through Meta-Analysis CNV
	eQTL	expression QTL
	EstBB	Estonian Biobank
F	**FISH**	fluorescence *in situ* hybridization
	FVC	Forced vital capacity

G	**GD**	Genomic disorder
	GGT	Gamma-glutamyltransferase
	GH	Growth hormone
	GnomAD	Genome aggregation database
	GoF	Gain-of-function
	GTEx	Genotype-Tissue Expression Project
	GW	Genome-wide
	GWAS	Genome-wide association study
H	**HbA1c**	Glycated hemoglobin
	HBOC	Hereditary breast and ovarian cancer
	HDL cholesterol	High-density lipoprotein cholesterol
	HMM	Hidden Markov model
	HPO	Human Phenotype Ontology
	HR	Hazard ratio
	HTN	Essential hypertension
I	**ICCA**	Infantile convulsion with choreoathetosis
	ICD-10	International Classification of Diseases, 10th revision
	IGF-1	Insulin-like growth factor 1
	IHD	Ischemic heart disease
	IMPC	International Mouse Phenotyping Consortium
	iPSYCH	Danish Lundbeck Foundation Initiative for Integrative Psychiatric Research
	IQ	intelligence quotient
	IV	Instrumental variable
	IVW	Inverse-variance weighted
K	**kb**	Kilobase pair
L	**LCR**	Low copy repeat
	LDL cholesterol	Low-density lipoprotein cholesterol
	LDLR	LDL receptor
	LD	Linkage disequilibrium
	LDSC	Linkage disequilibrium score regression
	LEPR	Leptin receptor
	LOEUF	LoF observed over expected upper bound fraction
	LoF	Loss-of-function
	LOH	Loss of heterozygosity
	LRR	Log R ratio
	LRS	Long-read sequencing
M	**MAF**	Minor allele frequency
	Mb	Megabase pair
	metQTL	Metabolite QTL
	MoBA	Norwegian Mother and Child Cohort Study
	MPV	Mean platelet volume
	mQTL	DNA methylation QTL
	MR	Mendelian randomization
	mRNA	Messenger RNA
	MVMR	Multivariable MR
N	**NAHR**	Non-allelic homologous recombination
	NDD	Neurodevelopmental disorder
	NGS	Next-generation sequencing
	NHEJ	Non-homologous end joining
	NMR	Nuclear magnetic resonance
O	**OA**	Arthrosis
	[O]MIM	[Online] Mendelian Inheritance in Man
	OR	Odds ratio

P	**PC**	Principal component
	PCA	Principal component analysis
	PFB	Population frequency of the B allele
	PGS	Polygenic score
	pHaplo	Probability of haploinsufficiency
	PheWAS	Phenome-wide association study
	PKD	Paroxysmal kinesigenic dyskinesia
	pLI	Probability of LoF intolerance
	pLoF	Predicted LoF
	POMC	Pro-opiomelanocortin
	pQTL	Protein QTL
	pTriplo	Probability of triplosensitivity
	PTV	Protein-truncating variant
Q	**QC**	Quality control
	QQ plot	Quantile-quantile plot
	QS	Quality score
	QTL	Quantitative trait locus
R	**RCAD**	Renal cyst and diabetes
	Rh	Rhesus
	RNA	Ribonucleic acid
	RNA-seq	RNA-sequencing
S	**SCr**	Serum creatinine
	SCZ	Schizophrenia
	SD	Standard deviation
	SE	Standard error
	SeLIE	Self-limited familial and non-familial infantile epilepsy
	SFARI	Simons Foundation Autism Research Initiative
	SH2B1	Sarcoma homology 2 B adaptor protein 1
	SHBG	Sex hormone binding globulin
	Simons VIP	Simons Variation in Individuals Project
	SKAT	Sequence kernel association test
	SNP	Single-nucleotide polymorphism
	SNV	Single-nucleotide variant
	SPA	Saddlepoint approximation
	SPARK	Simons Foundation Powering Autism Research for Knowledge
	SSC	Simons Simplex Collection
	SV	Structural variant
T	**T1D**	Type 1 diabetes
	T2D	Type 1 diabetes
	T2T	Telomere-to-Telomere
	TAD	Topologically associating domain
	TDI	Townsend deprivation index
	TWMR	Transcriptome-wide MR
U	**UCSC**	University of California Santa Cruz
	UKBB	UK Biobank
U	**WES**	Whole-exome sequencing
	WGS	Whole-genome sequencing
	WHR	Wasit-to-hip ratio
	WHRadjBMI	WHR adjusted for BMI

Introduction

1 Introduction

Genetics has always turned out to be much more complicated than it seemed reasonable to imagine. Biology is not like physics. The more we know, the less it seems that there is one final explanation waiting to be discovered.

– Steve Jones

In this dissertation entitled "*Where rare meets common: Leveraging population cohorts to study rare copy-number variants*" I present the work that I conducted over the last four years under the supervision of Alexandre Reymond and Zoltán Kutalik at the University of Lausanne, Switzerland. The main aim of my thesis was to develop a framework to study the phenotypic consequences of copy-number variants (CNVs) within the general population. Focusing on rare, mostly recurrent, large CNVs typically associated with severe phenotypic outcomes, this work shed light on the pleiotropy and variable expressivity of these CNVs, providing insights into the CNV architecture of complex traits.

I start my dissertation with an introduction that provides background as to the state of the field and defines concepts that are key to the understanding of the ensuing research. Human genetics and genomics are rapidly evolving disciplines, and what was state-of-the-art four years ago, when I started my PhD in 2020, might not reflect the current stand of the field. To name only a few major advances, the last years have seen the completion of the first Telomere-to-Telomere (T2T) reference genome (*1*), the release of hundreds of thousands of whole exome and genome sequences coupled to telomere, protein, and metabolite measurements, as well as electronic health record data (*2*–*9*), the generation of the first long-read sequencing datasets (*10*–*14*), and an overall push towards increasing diversity in genetic datasets. This chapter aims to provide a clear view of the motivations that fueled the research I conducted during my PhD. To provide background to the discussion, I included references to some advances that emerged over the last four years.

Next, I dedicate five chapters to studies to which I provided a major contribution. In each of these chapters, I introduce the study by briefly summarizing the aims and major findings, before presenting an extended version of the work, that in the case of published manuscripts integrates supplemental content. Finally, I finish by discussing how these works relate to each other and contribute to advancing our understanding of the role of CNVs as modulators of complex traits, while providing a perspective on future challenges and open questions.

1.1 The human genome 4
1.1.1 DNA as a vehicle to store genetic information 4
1.1.2 Inheritance of genetic information 6
1.1.3 The landscape of human genetic variation 7
1.2 Biobanks 10
1.2.1 Clinical cohorts 11
1.2.2 Birth cohorts 11
1.2.3 Healthcare cohorts 12
1.2.4 Population cohorts 12
1.3 Link genotype to phenotype . 15
1.3.1 Basic statistical concepts behind GWASs 15
1.3.2 Fixed effect models 17
1.3.3 Reporting & interpreting GWASs 18
1.3.4 GWAS model extensions . . 20
1.3.5 Leveraging molecular phenotypes 26
1.4 Copy-number variants 31
1.4.1 CNV mechanisms 31
1.4.2 CNV detection tools 33
1.4.3 Functional consequences of CNVs 37

Some of the figures were created with BioRender.

1.1 The human genome

1.1.1 DNA as a vehicle to store genetic information

A core characteristic of all living organisms is the ability to transmit the information required to build the next generation. The molecule encapsulating this information is called **DNA (deoxyribonucleic acid)**. DNA is a macromolecule composed of four distinct building blocks, namely nucleotides harboring either of the bases adenine (A), cytosine (C), guanine (G), or thymine (T), which are assembled into polynucleotide chains, also called nucleic acids. In humans, DNA is most commonly found in a double-stranded form, wherein two complementary polynucleotide chains face each other – through a base pairing rule wherein A always opposes T and C opposes G – and coil into a double helix (*15*). According to the latest estimates from the T2T Consortium, each copy of the human genome is composed of 3.055 billion base pairs (bp) (*1*), which are split into 23 separate entities called **chromosomes** (Figure 1.1).

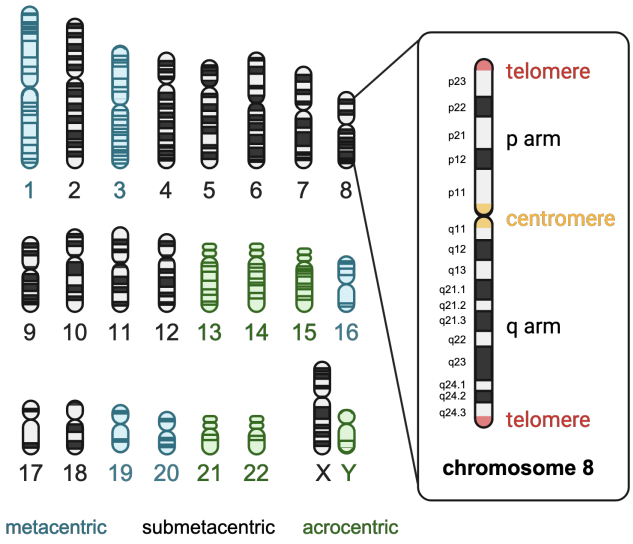


Figure 1.1: Human chromosomes. Haploid set of 22 autosomes and sex chromosomes (X or Y), colored according to the position of the centromere, with females being XX and males XY. Except for gametes, each human cell contains two sets of chromosomes. Zoom on chromosome 8 to highlight the different regions of the chromosome and illustrate the cytogenic bands used to define genomic positions.

Humans are **diploid** organisms, meaning that each cell harbors two copies of the human genome, one being inherited from the mother and the other from the father. These 46 chromosomes can be further categorized into 22 homologous autosomal pairs, labeled from 1 to 22 by decreasing size, and one sex chromosome pair that will define an individual's biological sex: females have two copies of chromosome X, while males carry a single X chromosome and a much smaller Y chromosome. At their ends, the two sex chromosomes share two regions of homology, called pseudoautosomal regions, which allow them to pair during male gamete formation. Zooming in, each chromosome can further be divided into five regions (Figure 1.1). At the tips are two **telomere regions** composed of repetitive DNA sequences that act as buffers, protecting genomic content from degradation. Each chromosome also has a **centromeric region**, composed of repetitive DNA sequences. The centromere plays a key role in cell division by acting as the connecting point for sister chromatids and the attachment point for the mitotic spindle. It also delimits the short (**p arm**, from the French "petit") from long (**q arm**)

chromosomal arm, which are often used as reference points in **cytogenetic nomenclature**. The latter describes genomic positions as **cytobands**[1], which reflect approximate chromosomal locations defined based on chromosomal staining techniques. The length ratio between the two arms will determined if the chromosome is metacentric (equal length), submetacentric (slightly longer q arm), or acrocentric (much longer q arm).

Zooming in, chromosomes harbor **genes**, which represent the most basic genetic units and account for ~1-2% of the human genome sequence. Humans harbor ~19,969 protein-coding genes (*1*), i.e., genes whose sequence will be transcribed into **RNA (ribonucleic acid)**[2] and eventually translated into proteins, following the **central dogma** of biology (Figure 1.2). More specifically, **transcription** of protein-coding genes results into RNA molecules from which introns are cut out in a process called **splicing**, resulting in messenger RNA (mRNA) molecules that only contain **exons**. Importantly, what is defined as intron and exon is transcript-specific, so that sequences retained in one transcript might be eliminated into another one, resulting in different **isoforms** through alternative splicing. After being processed, mRNA molecules leave the nucleus, where the DNA is stored, for the cytoplasm. There, ribosomes read mRNA in a process called **translation**. Specifically, each group of three mRNA nucleotides forms a codon corresponding to one of the 21 amino acids used by humans. The mRNA molecule is thus translated, codon by codon, into a chain of amino acids that will fold into a 3D functional **protein**, depending on the size and physio-chemical properties of the amino acids it is composed of. In turn, proteins will carry out the vast majority of cellular processes.

1: They are termed by chromosome number, arm, region, band, and sub-band, the three last ones being numbers of increasing value from centromere to telomere (e.g., 22p11.2).

2: **RNA** molecules have a similar structure to DNA. These mobile copies of the genetic content distinguish themselves from DNA by i) usually being single-stranded, ii) harboring nucleotides containing ribose instead of deoxyribose, and iii) using uracil instead of thymine.

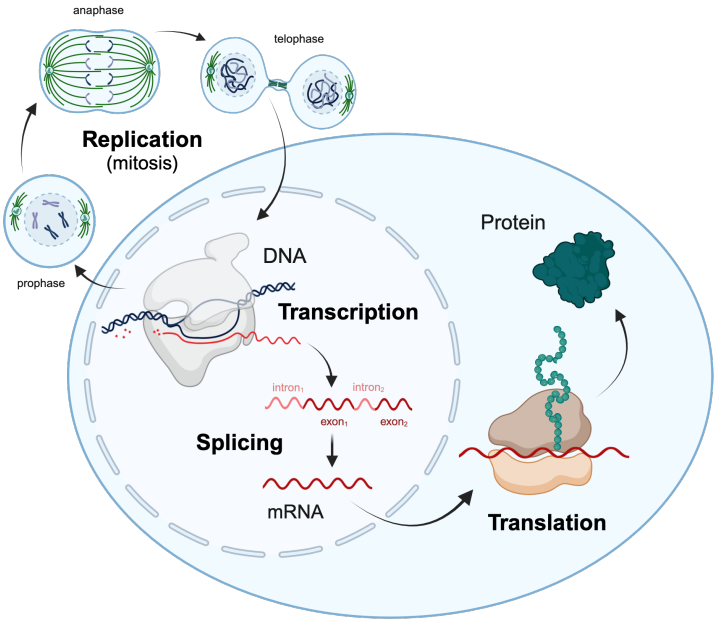


Figure 1.2: The central dogma. Schematic representation of the central dogma of biology. Genetic information is encoded into DNA, which is transmitted through replication. Once DNA polymerase has duplicated the genetic content, the cell can undergo cell division, illustrated here by some key steps of mitosis, generating two cells harboring identical genetic content. Within a cell, information flows from DNA to RNA through transcription, which is mediated by RNA polymerase (gray) in the nucleus. The RNA molecule is processed and spliced to retain only protein-coding exons in the mRNA molecule. In the cytoplasm, the mRNA molecule is translated into a polypeptide chain of amino acids by the ribosome. The latter in turn folds to adopt a final, stable conformation, forming a functional protein.

1.1.2 Inheritance of genetic information

The double helix conformation of DNA implies that both strands contain the same information, allowing DNA replication and transmission of genetic material to the next generation (Figure 1.2). While the replication process operates with high fidelity and includes proofreading mechanisms, it is not perfect and can lead to ***de novo* mutations** – in opposition to mutations inherited from either parent – at a rate of 1 mutation per $10^{8\text{-}10}$ nucleotides (*16*, *17*). In the human germline, this corresponds to about 50 to 90 new mutations per generation (*17*). **Germline mutations** are present in the gametes and can thus be transmitted to offspring. They oppose **somatic variants** that occur post-fertilization in non-germline cells. Because their detection and interpretation require special considerations, this introduction, as well as the research conducted in the following chapters focuses on germline mutations.

When a mutation appears, it faces two possible outcomes. Either it will remain in the population and become an **allele**[3] that can be further transmitted, or it will be wiped out as the individuals carrying it fail to transmit it. Which of these outcomes will materialize depends on genetic drift and natural selection. While genetic drift is a random process, the impact of natural selection depends on the fitness cost of the mutation: if it provides an advantage in a given environment, its frequency in the population will rise, sometimes up to **fixation**[4]; otherwise, it will be eliminated. Throughout millions of years, these processes – in combination with geographical and environmental factors – have led to speciation events wherein two groups of individuals (populations) accumulate such a large number of genetic variation that they become incompatible, i.e., they cannot mate and produce fertile offspring.

3: One of two or more versions of a DNA sequence at a given genomic position. The two alleles present in a diploid individual form the **genotype**.

4: Process through which an allele becomes the only one present in a population, eliminating variation at that locus.

Even within a single species, a large amount of genetic variation can be found. Recent efforts sequencing hundreds of thousands of individuals revealed the existence of billions of mutations, with the number of rare mutations being discovered steadily increasing with the addition of more samples (*3*, *9*). While this threshold is arbitrary, mutations present in less than 1% of a population, i.e., with a **minor allele frequency (MAF)** < 1%, are referred to as rare – or even as private when documented only in a single individual or family. This opposes common polymorphisms, present in > 1% of a population. Of note, as humans are diploids, within a single individual a given allele or mutation, can be present in a **heterozygous** (single copy) or **homozygous** (two copies) state (Figure 1.3). Importantly, allele frequencies are population-specific, so that alleles that are common in one ancestry group might be rare in another one. This observation is particularly relevant as it means that the **minor allele**, defined as the allele with the lowest frequency, is population-specific. This is not the case for the **reference** and **alternate alleles**, which have been predefined historically from the first human reference genome build produced by the Human Genome Project completed over 20 years ago (*18*). Because the reference genome was derived from individuals of European ancestry, the alternate allele often matches the minor allele in the latter ancestral group, but not necessarily in populations of other ancestries. To counter this, recent efforts by the Human Pangenome Reference Consortium have pushed for the transition to a pangenomic reference that better captures the full spectrum of human genetic diversity (*19*, *20*).

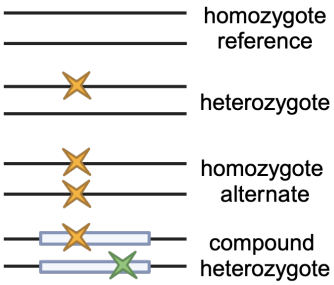


Figure 1.3: Zygosity.
At each genomic position, an individual can either carry zero (homozygous reference), one (heterozygous), or two (homozygous alternate) copies of a given variant (yellow star). Compound heterozygosity is a special case where two distinct mutations affect the two different alleles of a single entity, represented by the blue box (e.g., a gene). Compound heterozygotes can functionally mimic homozygotes alternates, leading to recessive phenotypes (see Table 1.5) despite each variant being heterozygous.

1.1.3 The landscape of human genetic variation

Genetic variation can take on different forms. One way to distinguish them is by the number of base pairs they affect: **short variants** affect 1 to 50 bp of DNA sequence while **structural variants (SVs)** affect ≥ 50 bp (Figure 1.4).

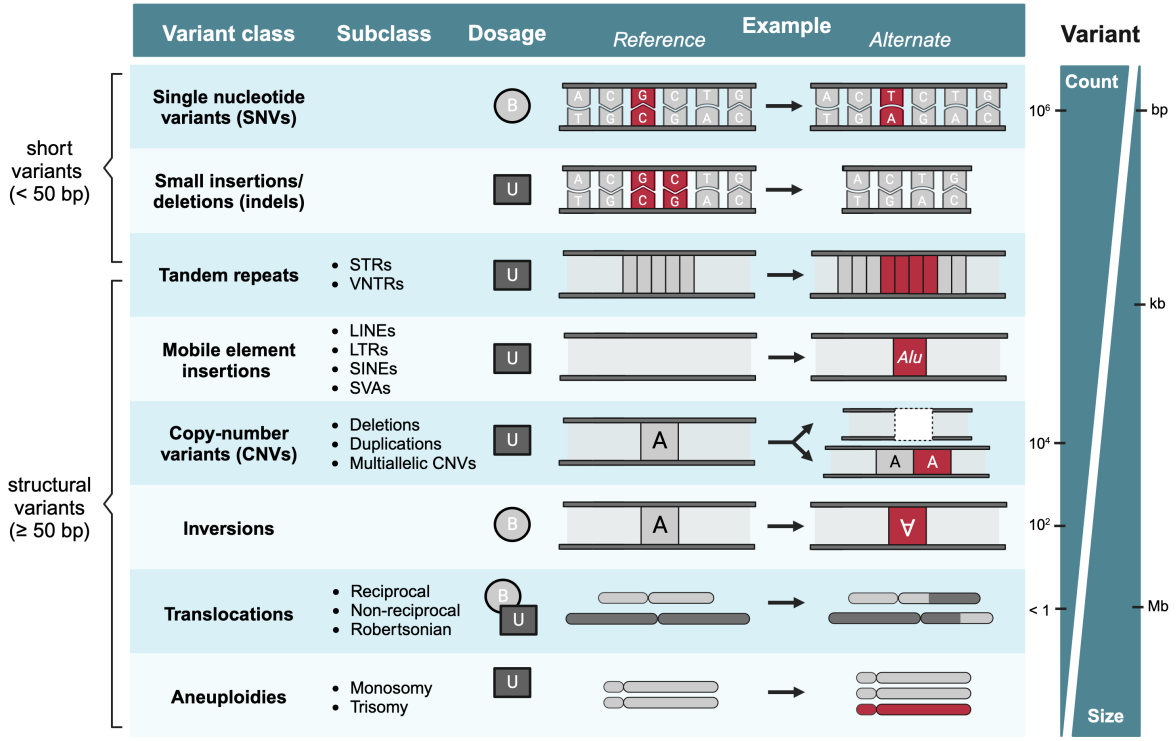


Figure 1.4: Landscape of genetic variation.
Summary table of the main classes of genetic variants observed in the human genome, adapted from (*21*). Common subclasses are listed, along with whether the mutations are balanced (B; light gray circle) or not (U; dark gray square). One example is given for each mutational class, with the changed sequence marked in red. Variant classes are ordered by increasing size, which roughly negatively correlates with the number of events per genome. LINEs = long interspersed elements (e.g., ~6k L1 elements); LTRs = long terminal repeats; SINEs = short interspersed elements (e.g., ~300 bp *Alu* elements); STRs = short tandem repeats (1-6 bp); SVA = SINE-VNTR-*Alu*; VNTR = variable number tandem repeats (few kb).

Short variants can be further subdivided into **single nucleotide variants (SNVs)** – also referred to as **single nucleotide polymorphisms (SNPs)** when their frequency exceeds 1% – where a single base is swapped for another one, and short insertions and deletions, collectively referred to as **indels**. Together, they represent by far the most common type of mutation, with an estimate of ~4-5 million SNVs and ~0.6-1.2 million indels in each genome, compared to the reference (*3*, *11*, *22*). When arising in coding regions, mutations are typically categorized depending on their impact on the protein sequence. **Synonymous** mutations do not affect the protein sequence and often result from an SNV in the last nucleotide of the codon, which due to redundancy of the genetic code, will still produce the same amino acid. Although synonymous changes can impact splicing, translation speed, and mRNA stability, their contribution to human disease is perceived as marginal (*23*). **Missense** mutations change the codon so

that it encodes for another amino acid. The consequence of such a change depends on the exact amino acid substitution and how it disrupts protein folding, activity, or stability. Tools such as SIFT (*24*) or PolyPhen-2 (*25*) leverage sequence conservation to predict deleteriousness, while more recent tools, such as AlphaMissense (*26*), additionally leverage variant population frequencies, sequence context, and protein folding prediction tools to improve pathogenicity predictions. **Nonsense** mutations, also refereed to as **protein-truncating variants (PTVs)**, lead to the premature insertion of a stop codon. Transcripts harboring such mutations are usually degraded by nonsense-mediated decay (i.e., if the mutation is before the last exon junction complex) to prevent the accumulation of defective proteins. Finally, **frameshift** mutations, resulting from indels that are not multiples of three, shift the transcript reading frame, altering the downstream amino acid sequence.

Mutations can also be classified according to their impact on cellular function, often by considering how the mutation affects encoded gene products. One such classification scheme is provided by Muller's morphs (Table 1.1). While any protein-coding mutation can result in protein **loss-of-function (LoF)**, i.e., reduction or absence of functional protein, nonsense and frameshift mutations, as well as splicing variants[5], are the most likely to do so (*27*). LoF variants are particularly pathogenic when they occur in a homozygous state or affect a **haploinsufficient** gene, for which a single functional copy is not sufficient to sustain the wildtype phenotype. This opposes **triplosensitivity**, which denotes intolerance to excess of gene product caused e.g., by **gain-of-function (GoF)** mutations. How well LoF mutations are tolerated is described by evolutionary constraint scores, such as the **probability of LoF intolerance (pLI)** or the **LoF observed over expected upper bound fraction (LOEUF)** scores (*29*, *30*), which have been derived for most protein-coding genes by the genome aggregation database (GnomAD) (*29*). These scores interpret absence of LoF mutations in a large population as a sign of purifying natural selection on that gene. This concept was extended to build genome-wide constraint maps for non-coding variants by e.g., the CADD score (*31*) or the more recent Gnocchi score (*32*), that take a step towards improving the interpretation of non-coding SNVs/indels, which remains comparatively more complex (*33*).

5: Located at or near exon-intron boundaries, splicing variants are not necessarily coding but they can disrupt splicing, leading to erroneous exon skipping or intron retention (*27*, *28*).

Table 1.1: Muller's morphs.
Mutation classification scheme according to its consequence on the gene product (i.e., protein) and the interaction of that product with the one resulting from the wild type allele. Type indicates if the morph is generally considered as loss-of-function (LoF) or gain-of-function (GoF).

Morph	Type	Consequence
Wild type	reference	Reference gene product
Amorph	LoF	No active gene product (**null allele**)
Hypomorph	LoF	Incompletely functioning gene product (**leaky allele**)
Hypermorph	GoF	More of the same, active gene product
Neomorph	GoF	Active gene product with a new, different function
Antimorph	other	Antagonizing or interfering gene product
Isomorph	other	Identical gene product

Small-scale mutations are abundant, relatively easy to detect, and represent the best-studied type of genetic variants. Calling of SVs is much more challenging (see section 1.4.2) and our knowledge about this highly diverse mutational class lags behind. Recent studies estimate the number of SVs per genome to 3,000-12,000 using short-read (*3*, *34*–*36*) and 23,000-28,0000 using long-read (*10*–*14*) sequencing. While true numbers are more likely to be close to the latter estimate, long-read sequencing approaches have not been applied at very large scale yet, with the largest

studies assessing ~1000 genomes. Despite the number of SVs being much lower than short variants (~2 orders of magnitude), they affect a much larger number of base pairs (*22*), making them a major source of human genetic variation.

SVs can be classified according to various characteristics, including their size and copy number. A first consideration is whether the mutation is **balanced** or not, i.e., whether there is a net gain or loss of genetic material compared to the two expected copies in humans. Balanced SVs tend to be rare and difficult to detect. They include **inversions**, where a DNA fragment is positioned in the wrong direction, and **translocations**, where large DNA segments break off and are swapped between chromosomes. For unbalanced SVs, the number of copies should be considered. **Deletions** and **duplications**, collectively referred to as **copy-number variants (CNVs)**, reflect the loss or gain of at least one copy, respectively, and are the best-studied type of SV. CNVs of median length (~1 kb) are common (*21*); more rarely, they can affect several Mb or even a full chromosome, leading to **aneuploidies**[6]. Usually, a deviation of a single copy is observed but some multiallelic CNVs can exhibit higher copy number changes. Multiallelic CNV regions have often been linked to evolution, as through gene duplication, the additional copy is free to accumulate changes that might result in the acquisition of new properties, while the original copy retains its primary function (*38*). Higher copy number can also be adaptive through increased expression of encompassed genes (*39*). This has been suggested for the iron-metabolism gene *BOLA2* (*40*), present in 3-8 copies in modern humans but only in a single copy in archaic humans (*41*). Another example is the starch-digesting gene *AMY1*, whose copy number correlates with salivary amylase levels and is present in a higher copy number in populations with a starch-rich diet (*42*). SVs present in extremely high copy numbers are referred to as repeats, which are further divided depending on their relative position to each other. Interspersed repeats, also called **mobile or transposable elements**, propagate through **retrotransposition**[7] so that 42.4% of our genome is estimated to be composed of these elements (*43*). They contrast with **tandem repeats**, which are composed of consecutive repetitive DNA units highly prone to expansion. Over 1 million tandem repeat sites exist genome-wide, making them important sites of human genetic variation (*44*). SVs that do not fit in any of these categories are termed **complex SVs**. They typically involve multiple DNA segments and might be generated by rare events, such as chromothripsis, where a single catastrophic event shatters one or multiple chromosomes which are then reassembled erroneously (*45*). Complex SVs can also happen at a smaller scale, and a median of a few dozen events per genome has been estimated (*34*).

One global trend across the spectrum of human genetic variation is that the size of a mutational class negatively correlates with the number of events per genome and the frequency of individual events (Figure 1.4) (*21*, *34*, *35*). This aligns with larger variants having a higher disruptive potential, making them more likely to exert phenotypic consequences and to be pruned out by natural selection. In the remainder of the introduction, I describe how genetic variants can be detected and linked to phenotypic variation and how this fits into a global understanding of the genetic architecture of complex traits, with a special focus on CNVs.

6: **Aneuploidy** refers to an abnormal chromosome number. In humans, only trisomy 13, 18, and 21 – also known as Patau, Edwards, and Down syndromes, respectively – are viable, even though only carriers of the latter are expected to survive infancy (*37*). Sex chromosome aneuploidies are better tolerated, the most common one being Turner (45,X), Klinefelter (47,XXY), triple X (47,XXX), and 47,XYY syndromes.

7: Briefly, the transposable element is transcribed, resulting in an RNA molecule that is used as a template to generate a DNA fragment through reverse transcription that is then inserted in the genome.

1.2 Biobanks

To fully capture the spectrum of human genetic diversity, one has to scrutinize many genomes. This is where **biobanks**, which collect large amounts of genetic, biological, medical, and environmental data for research purposes come into play. Through their large sample size, biobanks provide the statistical power to detect associations between variation genome-wide and **phenotypes**. The latter are defined as any measurable or observable traits resulting from the interaction between an individual's genome and its environment. Phenotypes can be classified according to their characteristics (Table 1.2). In genetic research, biobanks typically link genetic information with phenotypes such as questionnaire data regarding health, lifestyle, and occupation, external information related to demographics, socio-economic status, and environmental exposures, anthropometric and physical (e.g., cognitive, cardiac, pulmonary, auditory, or ophthalmic function) measurements, blood and urine biomarkers, stool samples for microbiome analysis, medical imaging, electronic health record data, drug usage and purchase information, and large-scale measurements of transcript, protein, and metabolite levels, collectively referred to as **omics** data. Depending on the biobank's protocol, longitudinal data might be available, along with genetic and phenotypic data for related individuals. Some biobanks can also return useful medical information to the participants or recall individuals to validate hypotheses generated from the initial data. This is particularly useful as phenotypic data might be noisy, especially if relying on self-reported data (46). Hence, careful thought should be put into selecting and defining phenotypes used to answer targeted research questions.

Table 1.2: Types of phenotypes. Classification of different types of phenotypes with examples.

Category	Definition	Example
Quantitative	Trait that can be measured numerically in each individual, taking a **continuous** or **discrete** value.	Height
Ordinal	Trait with multiple categories that have an implied order.	Relative size (smaller vs taller)
Nominal	Trait with multiple values that cannot be ordered.	Ethnicity
Binary	Trait that can take two values.	Disease status (case vs control)

Besides sample size and phenotyping depth, sample diversity in terms of ancestry, demographics, and **ascertainment** are important, as discussed in more detail in Chapter 6. Indeed, different genetic variants will be identified depending on how participants were recruited. Similarly to the common saying "*All models are wrong but some are useful*", most biobanks will present with some form of selection bias and are thus not truly representative of their target population (Figure 1.5). Yet each is useful in its own way, as long as one is aware of the recruitment protocol and takes the biases ascertainment induces into account when interpreting results.

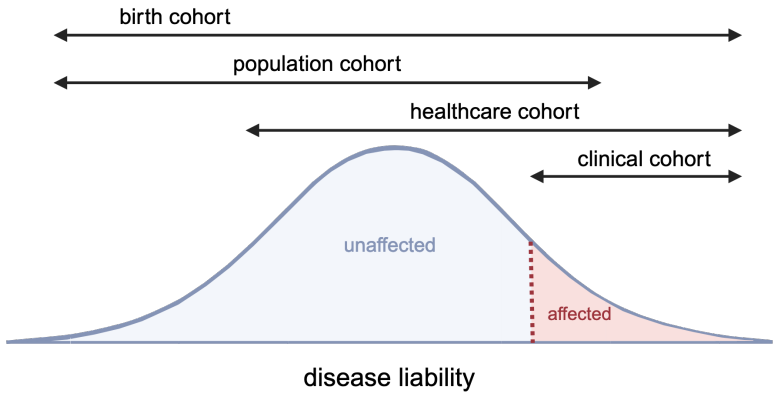


Figure 1.5: Ascertainment bias.
The disease liability distribution (see 1.3.4 – *Age of disease onset*) of the target population forms a Gaussian distribution, with individuals surpassing a threshold (dashed red line) being affected (red area), while others are not (blue area). Disease liability can refer to a specific disease or general health. The range in which individuals from different types of cohorts are sampled is indicated. Except for birth cohorts, none sample the full distribution, biasing variant frequency, effects size, and penetrance estimates derived from these cohorts.

1.2.1 Clinical cohorts

Clinical cohorts where patients are recruited based on the presence of a given phenotype are enriched for genetic variants associated with that phenotype. This approach is useful to recruit individuals with a homogeneous phenotypic presentation to study the etiology of that condition. For some rare disorders, it is the only approach to gather a sufficiently large number of patients. For instance, the Simons Simplex Collection (SSC) (*47*) and the more recent Simons Foundation Powering Autism Research for Knowledge (SPARK) (*48*) coordinated by the Simons Foundation Autism Research Initiative (SFARI), have collected genetic and phenotypic data on over 50,000 **simplex**[8] autism spectrum disease (ASD) families. Given the high genetic heterogeneity of ASD, SFARI launched a subprogram, the Simons Variation in Individuals Project (Simons VIP, now part of Simons Searchlight) that studies a subset of over 200 individuals with a common etiology of ASD, namely 16p11.2 CNVs. Despite the tremendous positive impact of such initiatives, clinical cohorts are not well suited to estimate the prevalence of phenotype-associated variants and are likely to overestimate their impact. Indeed, individuals who carry these variants but present with a different, milder, or no phenotype will not be present in the cohort. This concept is well illustrated by initiatives such as the Danish Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH). The study has two branches: 57,000 individuals affected with at least one severe psychiatric condition and 30,000 randomly sampled individuals from the same birth cohort that are not ascertained based on any phenotype (*49*). As detailed in Chapter 6, 16p11.2 BP4-5 duplications – a major factor for psychiatric diseases – are twice as prevalent in the cases vs control branch of the study.

8: **Simplex** families only have one affected individual, the proband. This opposed **multiplex** families, in which there are multiple affected individuals.

1.2.2 Birth cohorts

To obtain unbiased estimates of disease prevalences, allele frequencies, and effect sizes, one would ideally study **birth cohorts**, which aim at providing an unselected population sample. These studies are designed to enroll all or a representative sample (e.g., control branch of iPSYCH) of individuals born during a given time period at a given place. Participants, and sometimes close relatives, are then followed up on over several years,

allowing collection of genetic and environmental data and monitoring of their phenotypic and clinical development. Some examples include the Avon Longitudinal Study of Parents and Children (ALSPAC; N = 13,000) (*50*), the Northern Finland Birth Cohorts (NFBC; N = 22,000) (*51*), the Danish National Birth Cohort (DNBC; N = 100,000) (*52*), or the Norwegian Mother and Child Cohort Study (MoBa; N = 114,000) (*53*), as well as a series of UK national birth cohorts covering four decades hosted by the Centre for Longitudinal Studies (CLS; N > 17,000 per cohort). In practice, birth cohorts represent a considerable time and resource investment and are thus often limited in size, which represents a major limitation when studying rare genetic variants, as will be detailed in section 1.3.4. An additional consideration is that participant drop-out can further reduce sample size and generate selection bias.

1.2.3 Healthcare cohorts

Healthcare cohorts make use of existing infrastructure. Specifically, individuals within a given healthcare system are offered the possibility to donate a biological sample for research purposes. Genetic information can be extracted from these samples and linked to pre-existing medical records. Over the last years, many hospitals and health systems have established such cohorts, including the Vanderbilt University Medical Center's biobank (BioVU; N = 300,000; whole genome sequencing planned for 250,000 samples) (*54*), the Geisinger MyCode Community Health Initiative (N = 245,000; including 185,000 with genetic information through the DiscovEHR Project) (*55*, *56*), the Icahn School of medicine Bio*Me* program (N = 55,000, with genetic information), or the Mass General Brigham Biobank (N = 145,000, including 65,000 with genetic information) (*57*). Importantly, these programs are constantly growing, which represents one of the main strengths of healthcare cohorts. Yet, because participants are ascertained based on their interaction with the healthcare system, there is usually an overrepresentation of diseased individuals, which is reflected in the higher prevalence of pathogenic mutations (*58*).

1.2.4 Population cohorts

Population biobanks are initiatives that invite a large number of voluntary participants. Despite, aiming at sampling a representative subset of the general population, population biobanks have been shown to exhibit **healthy cohort bias** (*59*), i.e., participants are more educated, have a healthier lifestyle, and experience a lower disease burden compared to the general population. Population biobanks also tend to have a higher percentage of female participants (Table 1.3). The extent of this bias depends on the recruitment scheme and will be strongest in cohorts like the **UK Biobank (UKBB)**, which is composed of ~500,000 individuals that agreed to participate in the study, out of the nine million that were invited (*60*, *61*). Conversely, it might be milder for national biobanks that sample a larger fraction of their population, such as the **Estonian Biobank (EstBB)** that with ~200,000 participants samples about a fifth of the country's adult population (*62*). UKBB and EstBB have pioneered the

Sample sizes for healthcare cohorts are reported as of March 2024.

field of large-scale biobanks, and they are the biobanks I have worked with in the context of my dissertation. Over the last few years, many new biobanks have been created (Table 1.3), while efforts to facilitate collaboration and meta-analyses across biobanks, such as the Global Biobank Meta-analysis initiative (GBMI) (*63*), have been undertaken. Thanks to their large sample size, population biobanks offer the opportunity to study rare variants in a new light, i.e., in a non-clinical setting. Given the healthy cohort selection bias, population biobank studies are likely to underestimate the true allele frequency and effect of these variants. While strategies to account for this bias have been proposed (*64*), these studies can also be seen as complementary to the ones conducted in clinical and healthcare cohorts. If population cohorts typically have a defined sample size, it is not uncommon for them to accrue the number of phenotypes available for the fixed set of participants, making them a great resources for studying the global health impact of genetic variants. In the following section, I will review the basic principles of phenotype-genotype association studies and elaborate on some more sophisticated approaches that allowed to tackle challenges encountered in my research.

Table 1.3: List of 10 major population biobanks.
Major population biobanks ordered by decreasing number of individuals with available genetic information. The country of recruitment is indicated, along with the dominant ancestry group of the cohort. Target sample size is given and current status is indicated for cohorts that are still recruiting participants. Note that EstBB and deCODE do not have a target sample size. Demographics, including age range in years and proportion of female individuals (♀) are given. Available genetic and selected phenotypes are listed with the number of individuals (N) for which the data are available in parenthesis. If no sample size is indicated, data are available for the entire current sample. The baseline assessment typically involves basic questionnaire and physical examination. RNA-seq = RNA-sequencing; WGS = whole genome sequencing; WES = whole exome sequencing; EHR = linkage to electronic health record.

Name	Country (ancestry)	Size (current)	Enroll	Age	♀	Genotype (N)	Phenotypes (N)
UK Biobank (UKBB) (*60*, *61*)	UK (European)	500,000	2006-2010	40-69	54%	Genotyping WES WGS	Baseline assessment EHRs Biomarkers Telomere length Metabolomics Proteomics (50,000) Imaging (50,000)
All of Us (*9*)	US (European; African)	1,000,000 (400,000)	2018-	≥ 18	61%	Genotyping (300,000) WGS (250,000) Long-reads (1,000)	Baseline assessment EHRs (250,000) Biomarkers (250,000)
FinnGen (*8*)	Finland (European)	500,000 (400,000)	2017-	≥ 18	57%	Genotyping (250,000) WGS (3,700)	EHRs
Estonian Biobank (EstBB) (*62*)	Estonia (European)	200,000	2002-	≥ 18	65%	Genotyping WES (2,500) WGS (3,000)	Baseline assessment EHRs Biomarkers Metabolomics RNA-seq (600) DNA methylation (800) Microbiome (2,500)
BioBank Japan (BBJ) (*65*)	Japan (East Asian)	200,000	2003-2008	20-80	47%	Genotyping	Baseline assessment Medical records review
deCODE (*10*, *66*)	Iceland (European)	160,000	1996-	≥ 18	56%	Genotyping; WGS (60,000) Long-reads (3,500)	EHRs Proteomics (35,500)
Taiwan Biobank (TWB) (*67*)	Taiwan (Han Chinese)	200,000 (150,000)	2012-	20-70	64 %	Genotyping WGS (2,000)	Baseline assessment EHRs Biomarkers Imaging (38,000) DNA methylation (2,500) Metabolomics (1,100)
China Kadoorie Biobank (CKD) (*68*)	China (Han Chinese)	500,000	2004-2008	30-79	59%	Genotyping (100,000)	Baseline assessment EHRs Biomarkers (35,000) Imaging (35,000) Proteomics (3,000)
Trøndelag Health Study (HUNT) (*69*, *70*)	Norway (European)	230,000	1984-2019	18-90	52%	Genotyping (88,000) WGS (2,200)	Baseline assessment EHRs Biomarkers (123,000) Imaging (1100) Microbiome (13,000)
Uganda Genome Resource (UGR) (*71*)	Uganda (African)	7,800	2011	13-60	56%	Genotyping (5,000) WGS (2,000)	Baseline assessment Self-reported diseases Vaccination Biomarkers

Sample sizes for population biobanks are reported as of March 2024.

1.3 Link genotype to phenotype

One of the key goals of quantitative genetics is to establish relations between genetic variants and phenotypes. Early human genetics studies used **family studies** to demonstrate the genetic basis of a particular phenotype (*72*). These typically rely on **segregation analysis** in pedigrees to confirm the genetic basis of the phenotype and determine its inheritance mode, followed by **linkage mapping** to locate the allelic region responsible for the phenotype, and targeted investigation of candidate genes. Completion of the Human Genome Project in the early 2000s (*18*, *73*) has marked a shift in human genetics by facilitating access to genotype information and bypassing the need for linkage and candidate genes studies. This opened the doors for **genome-wide association study (GWAS)**. Conceptually, GWASs consist of running a large number of association tests to probe the phenotypic impact of genetic variants scattered over the entire human genome. The first large-scale GWAS by the Wellcome Trust Case Control Consortium in 2007 showcased the effectiveness of GWAS in identifying regions contributing to the genetic susceptibility of seven diseases (*74*). The study also emphasized the importance of large sample sizes and the need to adequately control for **population structure**[9]. Since then, GWASs have become a staple of the statistical genetics toolbox and have been applied to a broad range of human traits. In the following section, I describe multiple strategies to perform genotype association tests, with an increasing degree of sophistication, and a focus on methodology relevant for the ensuing chapters.

9: Also known as **population stratification**, this phenomenon described the cryptic presence of multiple subpopulations with different allele frequencies, which can lead to spurious associations.

1.3.1 Basic statistical concepts behind GWASs

The ultimate goal of a GWAS is to estimate the **effect** of a variant on the studied phenotype. For simplicity, let's start by assuming a haploid genome, where at a given position, an individual carries either the **effect allele** G_1 or the other allele G_0. At its most basic form, the effect β of G_1 on a **quantitative trait** can be viewed as the difference in mean phenotype values between individuals carrying the effect allele (μ_{G_1}) and those that do not (μ_{G_0})

$$\beta = \mu_{G_1} - \mu_{G_0} \tag{1.1}$$

Estimated effect size, $\hat{\beta}$, can take values between $-\infty$ and $+\infty$. Both the effect **magnitude** and **significance** are considered to determine if G_1 has a meaningful impact on the phenotype. The magnitude of the effect indicates how much G_1 increases (positive sign) or decreases (negative sign) the phenotype, with values around zero corresponding to no effect. To establish whether an effect is significant, uncertainty around parameters is taken into account. The most frequently used measure of uncertainty is the **standard deviation (SD)** or σ, which describes the variability across observations. The standard error (SE), which describes how accurately a parameter can be estimated, can be derived from the SD by dividing it by the square root of the sample size. The SE is used to calculate the confidence intervals (CI)[10], which reflects the range in which the parameter is expected to be contained 95% of the time. The SE of μ_{G_1} (SE_{G_1}) and μ_{G_0} (SE_{G_0}) can also be used to derive a test statistic

10: Typically, one estimates the 95% CI:

95% CI = $\hat{\beta} \pm 1.96 \cdot SE$, assuming that the parameter estimate comes from a Gaussian distribution

$$t = \frac{\mu_{G_1} - \mu_{G_0}}{\sqrt{SE_{G_1}^2 + SE_{G_0}^2}} \tag{1.2}$$

The estimator t can be used to perform a **t-test** or a **z-test**[11], from which a **p-value** can be derived. The two-sided p-value corresponds to the probability of obtaining a value of $|t|$ at least as extreme as the one observed under the assumption that the true effect size is null. The latter are typically considered statistically significant when the probability is lower than 5%.

11: The difference is whether one assumes an error (**t-test**) or not (**z-test**) in the SE estimators. At small sample sizes, the t-test is more accurate but the resulting p-values become virtually indistinguishable at large sample sizes.

Table 1.4: Contingency table.
a and b represent the number of cases, and c and d represent the number of controls, carrying G_0 and G_1.

	G_0	G_1
case	a	b
control	c	d

While **ordinal traits** can be treated as quantitative traits, associations between genotype and **binary traits** are typically assessed through **Fisher tests** (or chi-squared tests) which return the effect of the genotype on the trait as an **odds ratio (OR)**. Odds indicate how many times more likely it is for an event to occur (e.g., case) than it is not to occur (e.g., control), so that if an event has a probability p, its odds are $p/(1-p)$. Hence, the OR_{G_1} reflects the relative change in odds for individuals carrying G_1, compared to those carrying G_0. Based on the contingency Table 1.4[12] it is defined as

12: The **Cochran–Armitage test** for trend offers an alternative for 2×3 contingency tables, allowing to test the association with an ordinal variable (e.g., three genotype groups under an additive model).

$$OR_{G_1} = \frac{b/d}{a/c} = \frac{bc}{da} \tag{1.3}$$

Estimated $\widehat{OR_{G_1}}$ can take values between 0 and $+\infty$, indicating whether G_1 increases (OR > 1) or decreases (OR < 1) the risk to be a case. OR are often transformed to log(OR), which analogously to β, range from $-\infty$ and $+\infty$. This transformation notably allows estimation of the uncertainty parameter

$$SE(log(\widehat{OR_{G_1}})) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \tag{1.4}$$

which can then be used to calculate the 95% confidence of the $\widehat{OR_{G_1}}$

$$95\%\ CI = e^{log(\widehat{OR_{G_1}}) \pm 1.96 \cdot SE(log(\widehat{OR_{G_1}}))} \tag{1.5}$$

While the above-described approaches allow to obtain an estimate for genetic effect sizes, along with measures for the parameter uncertainty and significance, they suffer a major caveat, i.e., they fail to account for **covariates**, which are defined as variables that show an association with the phenotype of interest. Typically, these include demographic factors such as age, age^2, and sex, as well as the principal components (PCs) of the SNP genotype matrix, which capture population stratification patterns (*75*). Adjusting for covariates is paramount to obtain unbiased genetic effects and avoid detection of spurious associations (e.g., adjusting for population structure) and can increase statistical power by accounting for phenotypic variance that is not caused by genetic factors (e.g., adjusting for age or sex)[13]. Hence GWAS are practically always implemented using **multivariate regression models** that allow to account for possible confounding factors and offer more flexibility in terms of genotype encoding.

13: As reported by others (*76*) and explored in further detail in Chapter 6, adjustment for covariates should not be applied blindly as it can bias effect size estimates.

1.3.2 Fixed effect models

Early GWASs were conducted in cohorts of unrelated individuals of homogenous ancestry, with the choice of the regression model depending on the nature of the assessed trait. For quantitative (and ordinal) traits **fixed effect linear regressions** are typically used. For each variant, one fits a model:

$$Y = \beta_0 + X\beta_G + W\beta_C + \epsilon \quad \text{, where } \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{1.6}$$

In this equation, the input data represent:

- Y: a vector of **quantitative phenotype** values, that can be normalized to obtain effect sizes in SD instead of phenotype units.
- X: a vector of **genotype** values for the investigated variant. GWAS typically consider SNV genotypes under an **additive model**, which assumes that the phenotype is proportional to the dosage of the effect allele (Table 1.5; Figure 1.8).
- W: a matrix of **covariates** tailored to the research question but typically including age, sex, and genotype PCs.

Both X and W are referred to as **predictors** of the **outcome** Y. For each predictor, we estimate a parameter[14] that reflects the weight of this predictor in determining the outcome. These correspond to:

- β_0: the **intercept**, which is typically not further considered and corresponds to the phenotype value if all other predictors are null.
- β_G: **genetic effect** of the effect allele on the the phenotype (Figure 1.8).
- β_C: **covariates effects** on the phenotype, which are typically not further considered.

14: There are multiple methods to estimate parameter values, all of which also provide a measure for the parameter's uncertainty and significance. A popular method is the **least-squares approach** that aims at minimizing the sum of the squared residuals (i.e., the difference between observed and fitted values) to identify the best fitting equation model.

For **binary traits**, the method of choice is **fixed effect logistic regression**, a form of generalized mixed model that explains the logarithm of the odds for an event with a probability p to occur, as a linear function of the genotype X and the covariates W

$$log(\frac{p}{1-p}) = \beta_0 + X\beta_G + W\beta_C \tag{1.7}$$

As such, the same parameters as for the linear regression presented in equation (1.6) are estimated, usually through **maximum likelihood estimation**[15]. Following the reverse procedure than the one described in section 1.3.1, $\widehat{OR_G}$ can be derived as the exponential of $\widehat{\beta_G}$. Alternatively, one can also derive the probability p as

$$p = \frac{1}{1 + e^{-(\beta_0 + X\beta_G + W\beta_C)}} \tag{1.8}$$

15: This framework aims at estimating parameter values by maximizing the probability of observing the data at hand, through an iterative process. It further assumes that the parameter follows a particular distribution, namely a Binomial distribution for the data sample, where each example is one outcome of a Bernoulli trial.

Unlike linear regression, logistic regression might fail to **converge**, i.e., the maximum likelihood estimation could not find an appropriate solution. This can occur due to the inclusion of a too large number of predictors (i.e., covariates) compared to the number of cases, high correlation between predictors, sparseness in data, or data **separation**[16].

16: Scenario wherein the predictor(s) perfectly predict(s) the outcome, generating infinite coefficients with large SEs.

1.3.3 Reporting & interpreting GWASs

Summary statistics & data reporting

Effect size estimates for all assessed variants, including non-significant ones, are reported in **GWAS summary statistics**. Because GWAS summary statistics do not contain any individual-level data, they can be publicly shared without data privacy concerns. This is often done through the NHGRI-EBI GWAS Catalog (*77*, *78*), a free and user-friendly database hosting over 60,000 summary statistics. In an effort to promote data portability and reproducible, the GWAS-SSF format has been proposed (*79*), which requires reporting of specific information in a standardized way (Figure 1.6A).

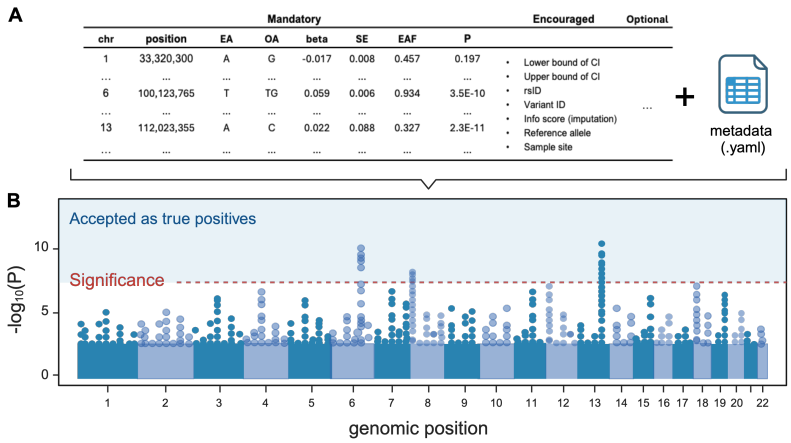


Figure 1.6: GWAS summary statistics. (**A**) Example of summary statistics in GWAS-SSF format. Mandatory columns include variant chromosome and position, effect (EA) and other (OA) allele, effect size (beta) with standard error (SE), EA frequency (EAF), and p-value. Data encouraged to be reported are listed, with the option to add further columns. A metadata file should be provided. (**B**) Representation of the GWAS summary statistics as a Manhattan plot showing the negative logarithm of the p-value (y-axis) against the genomic position (x-axis; labels indicate chromosomes) for each variant (dot). Variants surpassing $-\log_{10}(5 \times 10^{-8})$ (red dashed line) are considered as significant.

GWAS quality control and interpretation

Once associations have been computed for all variants, results are typically visualized as **Manhattan plots** (Figure 1.6B). In addition, deviation of the observed -$\log_{10}$ transformed p-values against the expectation under a null model of no significant associations is assessed through **QQ plots (quantile-quantile plot)**, where one expects a late and abrupt deviation from the expectation. This is often complemented by calculating the **genomic inflation factor** λ[17], whose value is expected to be close to 1. Genomic inflation ($\lambda > 1$) indicates either poor control for relatedness and/or population structure or a **polygenic** genetic architecture (*80*). Polygenicity refers to traits influenced by a large number of independent loci and represents the dominant genetic architecture of complex phenotypes (*81*). More rarely, λ will show deflation ($\lambda < 1$), which can be caused by pre-correction of the phenotype for population stratification (assuming a strong correlation between the latter and the tested variants), rare variants, or strong correlation across variants (*82*, *83*).

17: λ corresponds to the median of observed chi-squared test statistics divided by the expected median of the chi-squared distribution

Significant associations are selected based on a **Bonferroni correction** that accounts for **multiple testing**. Indeed, if each test has a 5% chance for **type I error**[18], the number of false positives becomes unacceptable (e.g., 1,000,000 × 0.05 = 50,000). Bonferroni correction adjusts for this by controlling the family-wise error rate, defined as the probability of making at least one type I error. For an association to be significant, $p < p_{Bonferroni} = \frac{\alpha}{m}$, where α is the accepted probability for type I

18: False positives (**type I error**) are associations deemed significant but for which the true effect is null; They oppose false negatives (**type II error**), which are not deemed significant but have a true non-null effect.

Number of summary statistics hosted by the GWAS Catalog as of February 2024.

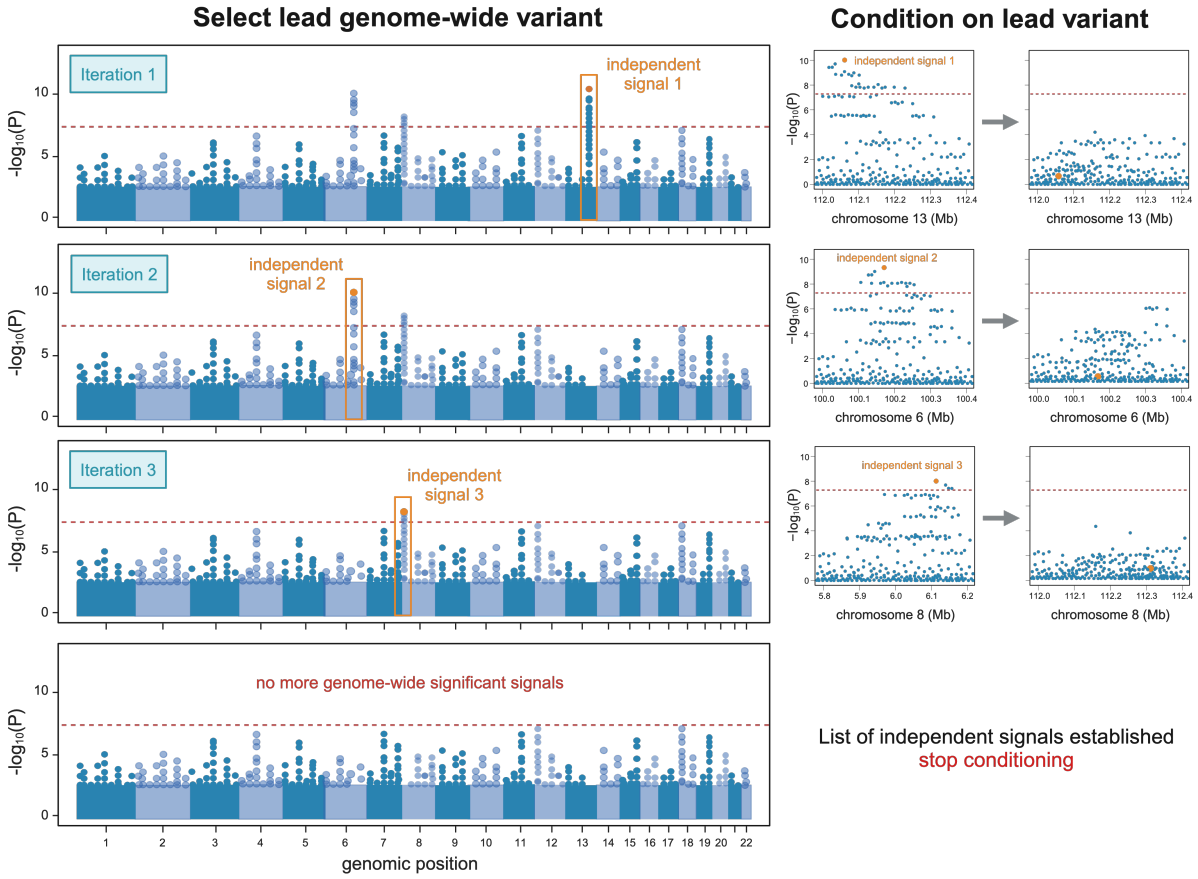


Figure 1.7: GWAS stepwise conditional analysis.
Schematic representation of the stepwise conditional analysis procedure. In the first iteration, the lead genome-wide signal (i.e., most significant) is selected (orange dot). The GWAS is performed anew, conditioning on the genotype of the lead variant, i.e., by including it as a covariate in the regression model. As a result the association p-value of that variant – and all the ones in linkage disequilibrium with it – nears one. The same procedure is repeated on the new summary statistics and repeated until there are no more genome-wide significant signals. Variants selected as lead variants generate a list of independently associated signals. Note that it is possible to have multiple independent signals in the same "skyscraper" of the Manhattan plot.

error (usually $\alpha = 0.05$) and m is the number of independent tests performed. While GWASs making used of **imputed variants**[19] typically perform several million tests, these are not independent due to **linkage disequilibrium (LD)**. It has been estimated that there are about 1,000,000 independent loci or **effective tests**. Hence, the commonly accepted threshold for significance is set at $p \leq 0.05/1{,}000{,}000 = 5 \times 10^{-8}$. This threshold might need adjustment when working with other mutational classes to accommodate the number of effective tests performed. The latter can be determined as the number of eigenvalues required to explain 99.5% of the variance in the genotype matrix (*85*).

Because of LD, the number of significant tests might not correspond to the number of independent signals, and many variants will appear significant because they are correlated with the **causal variant**. This phenomenon can be visualized as the Manhattan plot "skyscrapers" formed by correlated data points. **Stepwise conditional analysis** determines the number of independent signals by iterative conditioning on the lead variant (Figure 1.7). A caveat of this approach is that the most significantly associated variant is not necessarily the causal one, which often is not even assessed. The process of prioritizing and identifying causal variants is called **fine-**

19: **Imputation** refers to the inference of genetic variants that were not directly genotyped using a references panel, i.e., a set of high-quality, deeply sequenced genomes. Imputation is possible as blocks of variants in **linkage disequilibrium**, i.e., **haplotypes**, tend to be co-inherited. Hence, the presence of a variant predicts the genotype at adjacent locations (*84*). Because patterns of linkage disequilibrium vary across populations, the reference panel should match the ancestry of the imputed samples.

mapping and often relies on Bayesian models that incorporate prior assumptions, e.g., on the number of causal effects and their distribution, and will output a credible set of plausible causal variants. Commonly used tools include FINEMAP (*86*) or SuSIE (*87*), which generate lists of independent, prioritized GWAS signals that form the starting point for mechanistic and functional studies.

1.3.4 GWAS model extensions

Linear and logistic regression represent the foundation of GWASs and have been widely implemented in software such as PLINK (*88*). Since, extensions have been implemented that address various shortcomings.

Alternative association models
While there exist multiple **models of inheritance** (Table 1.5), determining which one is the best fitting comes at the cost of power as it requires estimation of multiple parameters. For this reason, GWAS favor **additive models**, which only necessitate estimation of a single parameter and were shown to capture the bulk of phenotypic variability explained by genetic variance, also known as **heritability** (*89*, *90*). Sometimes, genotype-phenotype relations are better captured by models that use a different genotype encoding (Table 1.5; Figure 1.8). **Recessive models** are particularly useful in populations with a high degree of **consanguinity**[20], such as in Pakistani individuals, where a recent study identified 185 recessive effects, 82% of which did not show an additive effect (*91*). Another situation in which recessive models should be applied are isolated populations having undergone a population **bottleneck**, such as in Finland (*92*) or Greenland (*93*). Bottlenecks are characterized by a strong population size reduction that diminishes genetic diversity. While this leads to the loss of many rare alleles, the ones that are retained will be present at higher frequencies, increasing homozygosity and facilitating the study of recessive phenotypes (*94*, *95*). Examples include the "Finish Disease Heritage" (*96*) or the "Jewish Genetic Disorders" (*97*), which represent primarily recessive disorders that are particularly frequent in Finish and Ashkenazi Jewish populations, respectively, due to the high rate of carriers in these populations.

20: **Consanguinity** refers to mating between close relatives. It will increase homozygosity (also termed **loss of heterozygosity (LOH)**) and thereby the expression of recessive phenotypes.

An alternative approach to genotype re-coding is to adjust phenotypes for known additive effects and test for residual contribution to trait variation, which allows to identify loci that contribute to the phenotype in a non-additive fashion. This approach is termed **dominance GWAS**, with dominance being defined as any deviation from a purely additive effect, thus encompassing all non-additive models in Table 1.5. Applying this approach to 1,060 phenotypes in UKBB revealed 183 such loci (*98*). Still, contribution of non-additive effects to phenotypic variance was minimal and sample sizes in the orders of millions will be necessary to capture dominance effects with effect sizes similar to the ones currently detected with additive models (*98*).

Heritability

Phenotypic variance can be decomposed into genetic and environmental variance. The former is composed of **additive**, **dominance**, and

epistatic variance, which capture the phenotypic contribution of individual alleles, within locus interactions, and cross-loci interactions, respectively. Based on this, **broad-sense heritability (H^2)** is defined as the proportion of phenotypic variance that can be explained by the *total* genetic variance in a given population at a given time point. It contrasts with **narrow-sense heritability (h^2)**, that is attributable to *additive* genetic variance (*99*). Historically, twin and family studies were used to estimate heritability. Because these methods rely on assumptions regarding the relatedness of the individuals and the extent of shared environment, they tend to overestimate heritability (*100*). More recently, results from GWASs have been used to estimate SNP-based h^2_{SNP} in unrelated individuals, including contribution of variants that do not pass the genome-wide significance threshold (*101*). Two popular methods to estimate h^2_{SNP} include genomic relatedness restricted maximum-likelihood (GREML) implemented in GCTA (*102*) and linkage disequilibrium score regression (LDSC) (*103*). Many extensions have been developed that go beyond the scope of this introduction. An interesting concept is that these approaches allow partitioning of heritability estimates based on genomic region, annotations, or allele frequency (e.g., (*104*–*107*)), allowing to compare the contribution of various sets of variants to phenotypic variability.

In the late 2000s, h^2_{GWAS} estimated from genome-wide significant GWAS signals were largely discrepant from the ones estimated through twin studies (*108*). For instance, height had a h^2_{GWAS} of ~5%, compared to a ~80% heritability estimated based on twin studies. Explanations for this ***missing heritability*** included unmeasured contribution of:

- Variants with small effect size.
- Rare variants.
- Structural variants.
- Non-additive effect (i.e., dominance and epistasis).
- Inadequate accounting for shared environment.

Fifteen years later, we know more about the contribution of these individual factors. Variants with small effect sizes indeed contribute to a vast part of the missing heritability, with recent h^2_{SNP} estimates for height being at ~50% (*109*). Despite still being far from estimates from twin studies, the study shows that in individuals of European ancestry, h^2_{SNP} from common additive effects reached saturation (*109*). This gap can partly be filled when accounting for additive effects from rare variants (MAF > 0.01%), yielding a heritability estimate of ~70% (*110*). Yet, assessing a broader spectrum of phenotypes, the average heritability explained by rare coding variants was estimated to only ~1% (*106*). This discrepancy suggests that rare *non-coding* variants might substantially contribute to heritability, as supported by the recent identification of rare non-coding variants regulating height (*111*). On the other hand, the global contribution of non-additive effect was repeatedly also shown to be minimal (*89*, *90*, *98*), in spite of striking examples of clinically relevant epistasis (*112*).

The contribution of SVs to heritability has not yet been properly assessed for complex traits, but several studies have shown that SVs contribute to up to 8% of heritability of gene expression (*113*–*116*). In Chapter 3, I crudely explore the contribution of large and rare CNVs to the heritability of the disease burden in the UKBB, suggesting a minimal contribution.

Table 1.5: Inheritance models.
Models describing genotype-phenotype relation used in GWASs. Encoding of the homozygous reference, heterozygous, and homozygous alternate genotypes is indicated in the right column. For the additive model, the encoding corresponds to the dosage of the alternate or effect allele. Intermediate models of e.g., dominance with reduced penetrance (0-0.8-1) or recessiveness with rare expressing heterozygotes (0-0.2-1) are possible. As described in Chapters 2 and 3, GWAS with other types of variants, such as CNVs, opens the door to an even larger number of association models.

Model	Description	Encoding
Additive	Each additional copy of the effect allele contributes equally to the phenotype. All genotype groups have a different phenotype.	0-1-2
Dominant	Carrying a single copy of the effect allele is sufficient to produce the phenotype. Both individuals that are heterozygous and homozygous for the effect allele have the same phenotype.	0-1-1
Recessive	Carrying two copies of the risk allele is required to express the phenotype. Only individuals homozygous for the effect allele exhibit the phenotype.	0-0-1
Over-dominant	Heterozygous individuals exhibit a different phenotype than the two homozygous groups, which exhibit the same phenotype.	0-1-0

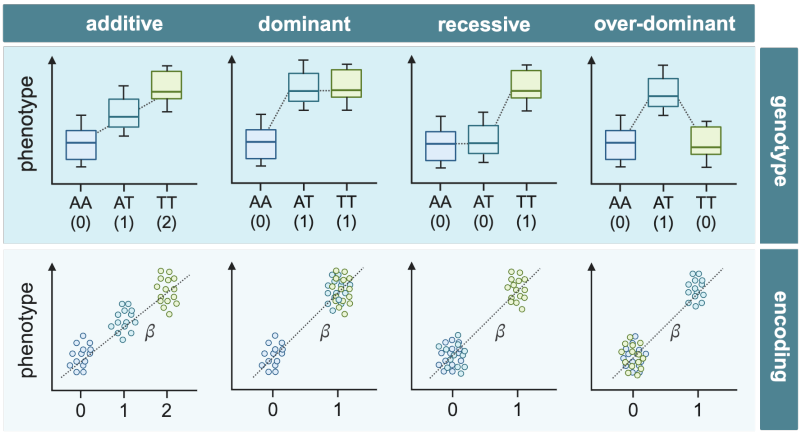


Figure 1.8: GWAS association models.
Schematic representation of the models in Table 1.5. For each variant, phenotypic values (quantitative traits) or disease risk (binary trait) (y-axis) are plotted against genotype groups (x-axis). Top: Qualitative genotype groups (top x-axis) are encoded numerically (bottom x-axis) to fit different association models. Bottom: Plotting of data according to their numerical encoding allows to estimate the effect of the variant, β, as the slope of the fitted regression line (dotted gray line).

Mixed effect regression

The introduction of **mixed effect models** that condition on various covariates without estimating each regression coefficient, but only certain distributional properties of them, represent a major advancement in GWAS methodology (*117*, *118*). One of the first applications has been to include a second error term, u, that captures the heritable component of random variation as its variance is proportional to the kinship matrix. The latter can be derived from genetic data and encompasses fine-scale population structure and cryptic relatedness, which represent important confounding factors in GWASs. By conditioning on u, the false positive rate and the error variance of estimated genetic effects are reduced. Because exclusion of related individuals becomes obsolete, these models can be applied to larger sample sizes, leading to a gain in power. Another mechanism through which mixed models increase power is by conditioning on genetic variants – other than the variant of interest – that are linked to the phenotype by either being causal, tagging causal variants, or through confounding mechanisms. This concept, known as **whole-genome regression**, reduces phenotypic variance that is not caused by the variant of interest, leading to more accurate effect size estimation (*101*, *119*).

Although computationally intensive due to the requirement of large matrix inversions, several software tools have implemented efficient versions of both linear and generalized linear mixed models. These allow for flexible prior distribution of SNV effect sizes[21] and can deal with both quantitative (e.g., GEMMA (*120*), BOLT-LMM (*121*), or fastGWA (*122*)) and binary (e.g., SAIGE (*123*)) traits. Typically, these methods rely on a two-step approach. Firstly, a model is fitted to a restricted set of

21: Spike-and-slab priors, which assume that only a fraction of all variants are causal, are popular. Being more realistic than earlier Gaussian priors for effect size distribution, they explain a larger fraction of phenotypic variance.

genome-wide SNVs. This model is used to generate individual-level phenotype predictors through a leave-one-chromosome-out scheme, wherein one predictor per chromosome is built, using variants located on all other chromosomes. In the second step, a larger set of variants (from the same sample) is tested for association, conditioning on the step 1 predictor that does not include the chromosome on which the assessed variant is located. By fully decoupling steps 1 and 2 and allowing them to be calculated in parallel for multiple quantitative and binary traits[22], REGENIE (*124*) tremendously reduced computation time and memory usage without compromising statistical efficiency.

22: **Multi-trait analysis** or simultaneous analysis of multiple traits reduces computation time by requiring a single pass over the genetic data. It was originally developed by BEGENIE.

Case-control imbalance

A common problem when performing GWASs in population biobanks that tend to be depleted of disease cases (Figure 1.5), is that the number of controls outnumbers the number of cases, sometimes by several orders of magnitude. Imbalance in the case-control ratio leads to inflation of type I error, a phenomenon further exacerbated in the context of rare variant analyses due to invalidation of the asymptotic assumptions[23] for logistic regression (*125*). Two approaches are commonly used to mitigate this issue. **Firth bias-corrected logistic regression** (*126*) uses a penalized likelihood function to correct the parameter estimates, resulting in well-calibrated type 1 error and sensible effect size and SE estimates. A second approach is to use a **saddlepoint approximation (SPA)** (*127*, *128*) which approximates the probability density function in a more flexible manner[24]. Unlike SPA, which better approximates the full shape of the distribution, the normal approximation only considers the distribution's mean and variance and thus performs poorly at the tails, especially if these are skewed as would be the case when the case-control ratio is imbalanced. While both Firth and SPA corrections efficiently control type I errors in genetic studies (*125*, *129*), they are computationally intensive. A fast SPA implementation has been proposed (*129*) and is used in SAIGE (*123*), although it was found to inflate effect size estimates (*124*). REGENIE includes both SPA and Firth correction, including a 60-times faster approximate Firth correction yielding highly concordant results to exact Firth regression (*124*). This approximation is also implemented in the generalized linear model function of PLINK v2 (*88*).

23: When sample size or case count is too low, the log-likelihood function will not take the assumed asymptotically quadratic shape, and resulting p-values estimations will be off.

24: The probability density function is estimated with a flexible with a flexible function of the cumulant generating function, which corresponds to the logarithm of the moment generating function, i.e., $K_X(t) = \log(E[e^{t \cdot X}])$.

Age of disease onset

Diseases with a stronger genetic, as opposed to environmental, component tend to have an earlier age of onset (*130*), suggesting that the age of onset captures information that could be included in GWAS. Modern biobanks are often coupled with electronic health records that include the date on which a diagnosis was received. This information can be incorporated through time-to-event or **survival analysis**. The latter focuses both on whether an event has occurred and when it occurred, leading to a gain in power over logistic regression under certain circumstances (*131*–*133*). In a GWAS context, time-to-event analysis can inform whether a specific mutation associates with an earlier or later age of disease onset.

Models used for time-to-event analysis need to be able to handle **censored data**, a type of missing data where the subject did not experience the event of interest during the follow-up time (Figure 1.9A). For instance, multivariable **Cox proportional hazards** (CoxPH) models (*134*) have been efficiently implemented for genome-wide applications in both fixed and random models and often make use of SPA to deal with heavily

censored data or low-frequency variants (135–137). CoxPH regression is a semi-parametric model of the **hazard function**

$$\lambda(t) = \lambda_0(t) \cdot e^{\beta_0 + X\beta_G + W\beta_C} \tag{1.9}$$

The instant risk for experiencing an event at time t, given by $\lambda(t)$, depends on a non-parametric and parametric component. The non-parametric component, $\lambda_0(t)$, is termed **baseline hazard**. It varies with time (monotonic increase) and corresponds to the hazard if the parametric part of the equation is null. The latter represents the additional contribution to the hazard by a set of predictors – here the the genotype vector X and the covariate matrix W, using the same notation as in the linear regression Equation 1.6 – and is not time-dependent. The independence between the parametric component of the equation and time is key to the proportional hazard assumption underlying the model. It implies that the hazard function is proportional over time, making the **hazard ratio (HR)** time-invariant (Figure 1.9B). The HR of the genetic effect represents the ratio between the hazards of individuals that carry the effect allele (G_1) and those that do not (G_0)

$$HR = \frac{\lambda(t|X = G_1)}{\lambda(t|X = G_0)} = e^{\beta_G} \tag{1.10}$$

The HR thus represents the exponent of the effect size in the survival model, which is estimated by maximizing the Cox partial likelihood function. Estimated for each predictor in the parametric part of the equation, HR takes values between 0 and $+\infty$[25]. HR > 1 indicates increased risk (i.e., earlier occurrence) and HR < 1 indicates decreased risk (i.e., later occurrence). A common way of visualizing results from survival analysis is by leveraging a non-parametric approach to survival analysis based on the **Kaplan-Meier** estimator (Figure 1.9C).

25: Despite many similarities, HR and OR differ in that they reflect the instantaneous risk over the study period, as opposed to the cumulative risk. This makes HR less prone to biases linked to the selection of the assessment endpoint.

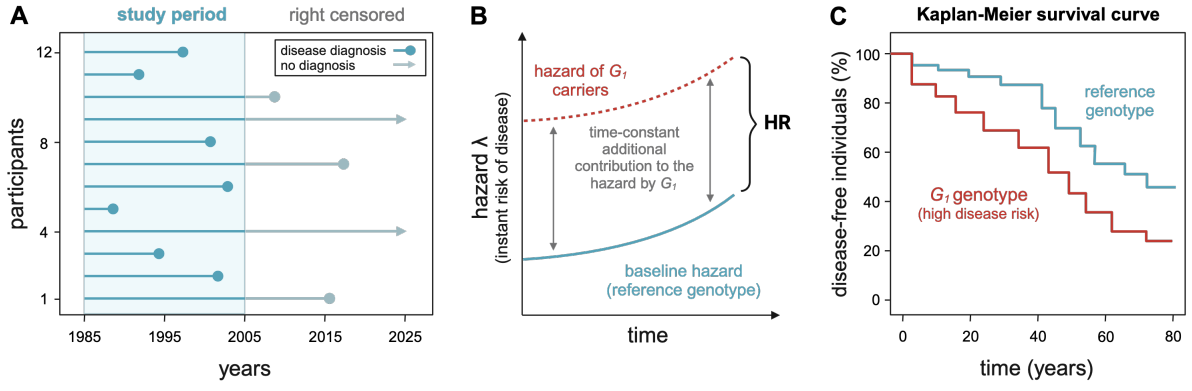


Figure 1.9: Survival analysis.
(**A**) Example of 12 individuals assessed for a disease diagnosis (dot) in a study period from 1985-2005 (blue window). Some individuals developed the disease after the assessment window or have not developed it to date (arrow), corresponding to right censoring. (**B**) Hazard (y-axis) over time (x-axis) for a univariable Cox proportional hazards model whose only predictor is the presence of the risk increasing G_1 allele. The baseline hazard function (blue curve) has a monotonic increase over time and corresponds to the hazard function of individuals that do not carry G_1. Individuals carrying G_1, have a time-constant additional contribution to the baseline hazard, resulting in a higher hazard (red dashed curve). Because hazard functions are proportional, the hazard ratio (HR) is constant over time. Note that baseline hazard does not have to be estimated to infer the HR. (**C**) Representation of time-to-event analysis as a Kaplan-Meier survival curves that depict the percentage of individuals that did *not* receive a diagnosis for the disease of interest (y-axis) against time (x-axis) for individuals belonging to the two same groups as in (B).

Recently, creative alternatives to CoxPH have been developed to integrate age of disease onset information. These are based, for example, on **liability threshold models** (*138*) that stipulate that for a given disease, each individual has a liability l that follows a normal distribution (Figure 1.5). If l exceeds a certain threshold, the individual is a case, otherwise a control. Notably, the probability for l to exceed the threshold corresponds to the disease's lifetime prevalence. The ADuLT framework estimates each individual's genetic liability by incorporating birth year, sex, and age-of-onset information to generate personalized thresholds (*139*). The estimated genetic liability can then be used as a phenotype to perform GWAS through any software handling quantitative traits.

Rare variants

Historically, GWASs have focused on common SNPs, i.e., the ones that are either directly genotyped or that can be imputed with good accuracy. Sequencing has made it possible to assess rare variants. By definition, these variants are only present in a handful of individuals. At constant **power**, there is a hyperbolic increase in the required (squared) effect size to detect an association with a variant as its frequency decreases (Figure 1.10)[26]. As such, only variants with extremely large effects will be identified as genome-wide significant. This loss of power also makes rare variant association testing more prone to **Winner's curse**[27], as in a situation where power is not adequate, signals called significant are more likely to have been over-estimated (*140*).

A common strategy to counter the loss of power linked to rare variant association testing is to perform a joint analysis of multiple rare variants grouped into a single analysis unit, most often a gene or an exon (*141*). Included variants are usually selected by applying various filters or **masks** on the variant's frequency or predicted functional impact (e.g., LoF). The advantage of such an approach is double as not only does it increase signal strength, but it also reduces the number of performed statistical tests, hence the multiple testing burden. Simple **burden tests** summarize information across considered variants either as i) a binary variable reflecting the presence of at least one rare allele, ii) a discrete variable reflecting the count of rare alleles, or iii) a weighted sum that gives more weight to specific variants (e.g., based on frequency). A limitation of this approach is that it assumes that all variants affect the phenotype with the same magnitude and direction of effect, which is not always realistic and can lead to power loss. Variance component tests, such as the sequence kernel association test (SKAT) (*142*), do not rely on these assumptions and allow simultaneous testing of both rare and common variants. The optimal unified SKAT-O test further improves on both approaches by selecting the best linear combination of the burden test and SKAT to maximize test power, even in low sample sizes (*143*). These methods have been implemented in major GWAS software, including REGENIE (*124*) and SAIGE-GENE+ (*144*), and have been applied to perform gene-level burden tests in cohorts such as UKBB on thousands of phenotypes (*6*, *145*). Results are publicly available through Genebass (*145*).

Phenome-wide association studies

Instead of testing all variants genome-wide for association with a specific trait, one can also test the association between a single genetic variant and a large spectrum of phenotypes, an approach termed **phenome-wide association studies (PheWASs)**. Conceptually similar yet complementary

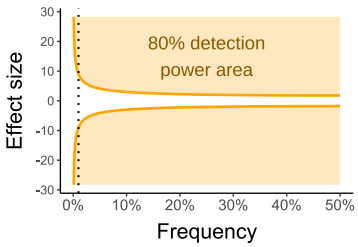


Figure 1.10: GWAS detection power. The orange area depicts the range of variant effects size (y-axis) and frequency (x-axis) combinations in which signal detection power is > 80%. The dotted black line delimits the 1% frequency threshold used to define rare variants.

26: **Statistical power** is defined as the probability to detect a true effect, i.e., with 80% power, on average, 8 out of 10 true effects will be called significant. In a linear regression model, to detect a non-zero effect at a fixed significance threshold (α), approximately the following inequality needs to hold: $-\sqrt{2q(1-q)} \cdot |\beta| \cdot \sqrt{N} < T_\alpha$, where q is the variant's frequency, β its effect size, N the sample size, and $T_\alpha = \Phi^{-1}(\alpha/2)$. When the variant is extremely rare ($(1-q) \approx 1$), the lower bound for the squared effect size is $\beta^2 > \frac{1}{q} \cdot \frac{T_\alpha^2}{2N}$.

27: The term is derived from auctions, where each bidder estimates in an unbiased way the true value of the item, with an error margin. Yet, the bidder that overestimates the true value will provide the highest bid and win the auction.

to GWASs, PheWASs require special consideration regarding phenotype definition, as well as tailored multiple-testing correction approaches (*146*). Notably, PheWASs reveal the **pleiotropy** of genetic variants, allowing the detection of shared genetic mechanisms. These are particularly relevant when considering intermediate molecular phenotypes, as I discuss in the following section.

1.3.5 Leveraging molecular phenotypes

Molecular quantitative trait loci (QTLs)

GWASs can establish associations between genetic variants and phenotypes, yet due to phenomena such as confounding and reverse causality, correlation does not equate to causation. In genetics the concept of **causality** is more often probabilistic than deterministic. Yet, the concept remains key in the optic of developing interventional approaches, which often rely on knowledge about the gene whose disruption causes the altered phenotype. Only 2–3% of GWAS signals are fine-mapped to coding variants and thus have a clear candidate gene (*81*). For the remaining signals, the gene closest to the GWAS signal has been shown to often be the causal gene (*6*, *161*, *162*), despite some notably misleading examples[28]. A strategy that gained traction over the last decade is to leverage **molecular phenotypes** to identify causal genes. GWAS can be performed on these phenotypes to identify molecular **quantitative trait loci (QTLs)**, i.e., genetic variants that affect the levels of DNA methylation (mQTLs), gene expression (eQTLs), proteins (pQTL), or metabolites (metQTLs) (Table 1.6). Given the high combinatorial space (genome-wide variants × hundreds of molecular traits) and the strong holding of the nearest gene hypothesis (*151*, *155*), QTL studies often restrict association testing to the ***cis*** **region**, which corresponds to variants ± 1 Mb away from the

28: A textbook example is a GWAS signal for obesity mapping to *FTO* (*163*, *164*). Later mechanistic studies showed that the variant acts by de-repressing expression of the adipocyte regulator genes *IRX3* and *IRX5*, located hundreds of kilobase pairs away (*165*).

Table 1.6: List of major molecular QTL studies.
List of studies coupling genotype to molecular phenotype measurements, enabling QTL analyses. For each study, the used technology, number of participants (N), number of assessed molecular entities (Phenotypes), and analyzed tissue(s) are indicated. Methylation/expression array and RNA-sequencing (RNA-seq) rely on the same principle as genotyping arrays and short-read sequencing, which are described in the context of CNV detection in section 1.4.2. SOMAscan and Olink are affinity proteomics approaches that rely on binding of proteins by nucleotide-based antibodies/reagents, enabling protein detection and quantification through standard DNA analysis tools. Advantages of these technologies over mass spectrometry-based proteomics include higher throughput and low sample volume requirement, at the cost of being targeted, and having lower molecular specificity (*147*). Metabolomics quantification typically uses mass spectrometry or nuclear magnetic resonance (NMR). While the former offers the best sensitivity, NMR spectroscopy has the advantage of being highly reproducible, requiring low sample preparation, and being non-destructive (*148*).

Omics	Study	Technology	N	Phenotypes	Tissue
Methylome (mQTLs)	GTEx (*149*)	Methylation array	424	~750,000	9 tissues
	GoDMC (*150*)	Methylation array	27,750	~420,000	whole blood
Transcriptome (eQTLs)	GTEx (*151*)	RNA-seq	838	~27,000	49 tissues
	eQTL Catalog (*152*)	RNA-seq & expression array	5,714	~35,000	69 tissues
	MetaBrain (*153*)	RNA-seq	6,518	~19,000	7 brain tissues
	eQTLGen (*154*)	RNA-seq & expression array	31,684	~20,000	whole blood
Proteome (pQTLs)	INTERVAL (*155*)	SOMAscan	3,301	3,622	plasma
	SCALLOP (*156*)	Olink	30,931	90	plasma
	deCODE (*157*)	SOMAscan	35,559	4,719	plasma
	UKBB (*4*)	Olink	54,219	2,923	plasma
Metabolome (metQTLs)	KORA/TwinsUK (*158*)	Mass spectrometry	7,824	486	plasma
	CLSA (*159*)	Mass spectrometry	8,299	1,091	plasma
	Meta-analysis (*160*)	Mass spectrometry	8,569-86,507	174	plasma
	UKBB (*5*)	NMR	118,461	249	plasma

methylation site or gene encoding for the assessed transcript/protein. This opposes *trans* studies that estimate association with all variants genome-wide. Due to differences in the throughput of technologies used to measure molecular phenotypes, eQTLs have benefited from larger sample sizes, which combined with their straightforward interpretation and proximity to the genotype[29], makes them the most widely studied type of QTL. Here, I briefly discuss two methods that leverage QTLs: colocalization and Mendelian randomization (MR).

29: Due to increased exposure to external factor and regulatory processes, QTL effects weaken across omic layers, the further downstream from DNA (e.g., transcript > protein > metabolite) (*166*).

Integrating QTLs & GWAS results

Colocalization can be viewed as an extension of fine-mapping to multiple traits. Providing that two traits show an association at a given genetic locus, colocalization leverages the association signal of a large number of variants in the region to distinguish between two scenarios: the signals are caused by the same causal variant (i.e., colocalization) or by two different variants, that are possibly in LD[30]. Often, one of the traits will be a molecular phenotype, e.g., expression levels of a gene mapping to the region. Colocalization would then provide support that the same variant that affects gene expression also affects the other phenotype, providing support for the involvement of that gene. One of the most popular colocalization methods is *coloc* (*167*). It makes use of a Bayesian framework that outputs a posterior probability for five hypotheses (Figure 1.11):

30: **Marginal** effects are estimated by testing one variant at the time and can be affected by other variants in LD. They oppose **joint effects**, which do account for variants in LD.

- H_0: none of the two traits are significantly associated.
- H_1: trait 1 is associated but not trait 2.
- H_2: trait 2 is associated but not trait 1.
- H_3: both traits are associated but through different causal variants.
- H_4: both traits are associated through the same causal variant.

The latter hypothesis denotes colocalization, which is often accepted when the posterior probability for H_4 is > 0.8. One important caveat is that *coloc* works under the assumption of a single causal variant per trait. This assumption is relaxed in SuSiE-coloc (*168*), as well as by other colocalization software tools, such as eCAVIAR (*169*).

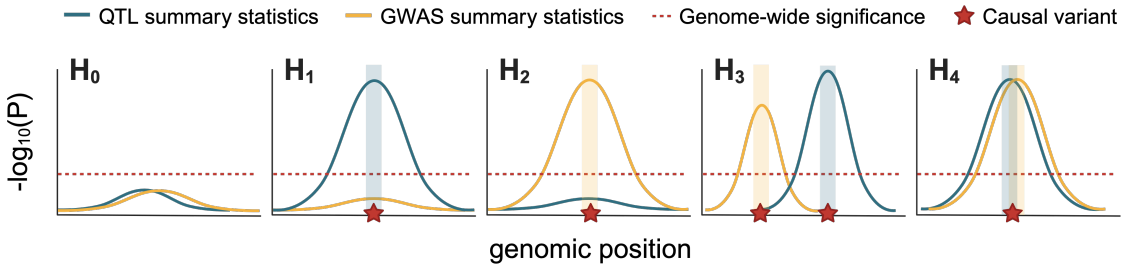


Figure 1.11: Colocalization.
Schematic representation of the hypotheses (H) tested by *coloc*, plotting the negative logarithm of the association signal (y-axis) against the local genomic position (x-axis). Of interest are the posterior probabilities for H_3 and H_4, which determine if the QTL (blue) and GWAS (orange) summary statistics are concordant with the presence of distinct or a common causal variant (red star), respectively. A high probability for H_4 indicates colocalization.

A related, yet distinct, approach for causal gene identification is **Mendelian randomization (MR)**. MR is a causal inference framework that leverages genetic variants, termed **instrumental variables (IVs)**, to estimate whether an **exposure** has a causal impact on an **outcome** (Figure 1.12). Specifically, IVs are selected to be significantly associated with the exposure before assessing their impact on the outcome, from which the causal effect is derived. The most commonly used MR method is called **inverse-variance weighted (IVW)**. It estimates the causal effect (α) as a weighted meta-analysis of each IVs' **Wald ratio**, computed as the effect of the IV on the outcome ($\beta_{outcome}$) divided by its effect on the exposure ($\beta_{exposure}$). Importantly, MR relies on three key assumptions whose violation can bias causal effect estimates (*170*):

1. **Relevance:** IVs are associated with the exposure.
2. **Exchangeability:** no confounder affects both the IVs and the outcome (e.g., due to **assortative mating**[31] or population stratification).
3. **Exclusion restriction:** IVs affect the outcome only through the exposure. This assumption is often violated by **pleiotropy**.

31: Non-random mating of individuals that share more (positive) or less (negative) often than by chance a given phenotype. **Assortative mating** can also occur across two distinct traits.

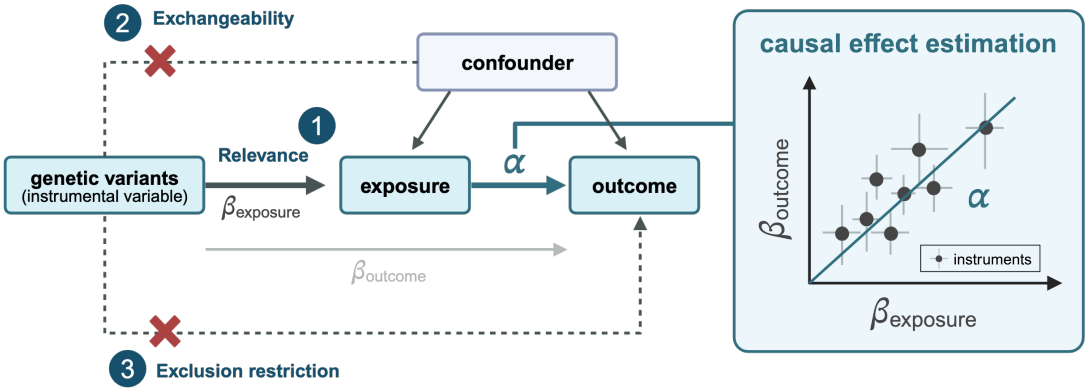


Figure 1.12: Mendelian randomization assumptions and causal effect estimation.
Mendelian randomization (MR) framework and its three core assumptions, which state that the genetic variants used as instrumental variables (IVs) should impact the exposure (1), not be associated with a confounder of the exposure-outcome effect (2), and not impact the outcome through pathways other than the exposure (3). Providing these assumptions hold, the causal MR effect of the exposure on the outcome (α; blue arrow) is estimated from the IV's effects on the outcome ($\beta_{outcome}$; thin light gray floating arrow) divided by its effect on the exposure ($\beta_{exposure}$; thick gray arrow), both of which originating from GWAS/QTL summary statistics. This estimator is termed the Wald ratio. To obtain more robust causal estimates, multiple IVs are typically used. The most popular multi-IV approach is inverse-variance weighted (IVW) MR. IVW-MR estimates can be computed as a meta-analysis of the Wald ratio obtained for each IV, weighted by the inverse variance of $\beta_{outcome}$. This is equivalent to regressing $\beta_{outcome}$ (y-axis) on $\beta_{exposure}$ (x-axis), weighted by the inverse variance of $\beta_{outcome}$, as depicted in the blue box. In this scatter plot, each dot represents an IV whose error bars represent effect size standard errors. The IVW estimate, α, is the slope of the best-fitting line through these data points that also passes through the origin (blue line).

Pleiotropy

Pleiotropy refers to the genetic phenomenon wherein a single genetic variant or locus is associated with multiple phenotypes. After a decade of GWAS and PheWAS powered to detect ever smaller genetic effects on a broad variety of phenotypes, it has become apparent that the vast majority of the genome impacts the human phenome in some way and that most loci (90%) are associated with multiple traits (*171*)

implying the ubiquitousness of pleiotropy. With an increasing number of molecular QTL studies measuring intermediate phenotypes, we can expect these trends to become even more pronounced.

Pleiotropy is particularly relevant for MR, as its presence often leads to violation of the 3rd assumption. Yet, not all types of pleiotropy are the same. Mechanistically, we distinguish between **direct** or **horizontal pleiotropy**, where the genetic variant is independently associated with different phenotypes, and **indirect** or **vertical pleiotropy**, where the genotype first affects one trait, which in turn affects another trait (Figure 1.13). Horizontal pleiotropy can bias MR causal effect estimates, especially in the presence of **correlated pleiotropy**, i.e., when there is correlation between the IV-exposure effect and the pleiotropic effect. The latter is likely to occur when the IVs associate with a confounder of the exposure-outcome association. Conversely, vertical pleiotropy does not bias causal effect size estimate and can even help identify mediators of the exposure-outcome relation (*170*). Of note, unlike MR, association studies cannot disentangle the order of vertical pleiotropy as both traits will appear associated.

As discussed in section 1.4.3, some recurrent CNVs show extreme pleiotropy. Throughout Chapters 4 to 6, I describe studies aiming at assessing and better understanding the mechanisms of pleiotropy of three distinct CNV regions.

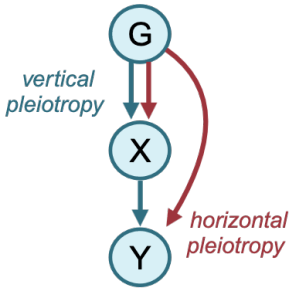


Figure 1.13: Types of pleiotropy. Schematic representation of pleiotropy types in an MR framework, where a genetic variant (G) affects two traits: the exposure, X, and the outcome, Y. The blue arrows depict indirect vertical pleiotropy, where the effect of G on Y goes through X. The red arrows depict direct horizontal pleiotropy, where G directly impacts both X and Y.

While the first MR assumption can formally be tested, this is not the case for the other ones. IVW further relies on the **InSIDE assumption** (INstrument Strength Independent of Direct Effect) that states that there is no imbalanced (i.e., mean effect of IVs through paths other than the exposure is null) or correlated pleiotropy (*172*). Hence, it is common to perform sensitivity analyses with other MR methods that relax some of these assumptions (*170*). Which method is best suited depends on the research question. Historically, MR has most often been used as an epidemiological tool to assess the impact of risk factors (e.g., alcohol consumption) on health outcomes (e.g., cardiovascular disorders). More recently, it was adapted to assess the causal impact of molecular traits. During my PhD, I collaborated with colleagues from the Statistical Genetics Group that developed such frameworks, e.g., to infer the causal impact of the change in expression of one or multiple genes on complex traits by **transcriptome-wide MR (TWMR)** (*173*), as well as the reverse causal effect of diseases on the transcriptome (*174*) (Figure 1.14A). Using multivariable MR (MVMR) approaches, these frameworks can be adapted to decipher causality chains across molecular layers. For instance, to explore the mediatory role of gene expression on DNA methylation-phenotype relations (*175*), or the role of metabolites in mediating transcript-phenotype associations (*176*) (Figure 1.14B).

One important specificity of MR with molecular exposures is that unlike complex exposures, for which there exist genome-wide IVs, the number of IVs fulfilling the first assumption is often limited to one or a few uncorrelated variants clustering at the locus encoding for the assessed molecular trait. Under such circumstances, a parallel can be drawn between MR and colocalization, which by definition focuses on a single genetic region. Strengths of both approaches are exploited by MR-link-2, which, unlike most MR methods that rely on independent IVs, makes

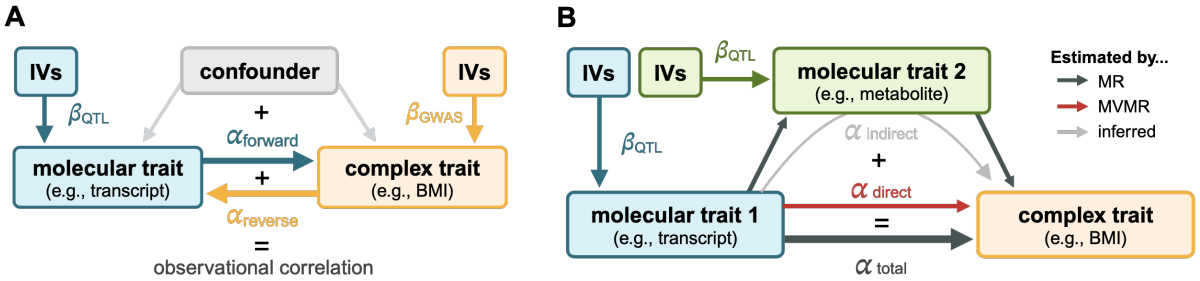


Figure 1.14: MR frameworks with molecular exposures.
(**A**) Schematic representation of bidirectional causal MR effect estimation between a molecular phenotype (blue) and a complex trait (orange). Instrumental variables (IVs) for the forward (blue) and reverse (orange) effect estimation are selected from QTL (β_{QTL}) and GWAS (β_{QTL}) summary statistics, respectively (thick arrows). To calculate MR effects as in Figure 1.12, the effects of these IVs on the outcome are assessed. The sum of forward and reverse MR effects plus the confounder contribution equals the observational correlation. Porcu *et al.*, show that the observational correlation between the transcriptome and BMI, triglycerides, and HDL is mainly due to confounding, followed by reverse effects (10-30%), and only a marginal contribution of forward effects (*174*). (**B**) Schematic representation of a multivariable MR (MVMR) framework for mediation analysis of a molecular trait (blue) on a complex phenotype (orange) by another (downstream) molecular trait (green). Dark gray arrows are estimated using MR, using IVs selected from QTL (β_{QTL}) summary statistics (blue/green thick arrows), following the same framework as depicted in (A). The effect of the main molecular trait (blue) on the complex trait (orange) is defined as the total effect (α_{total}). The total effect is the sum of the direct effect between these traits, estimated by MVMR (red arrow; α_{direct}) accounting for the molecular mediator (green), and the indirect effect (light gray arrow; $\alpha_{indirect}$), mediated by the second molecular phenotype (green). Auwerx *et al.*, shows that while the majority (77%) of transcript-to-trait effects show no mediation through metabolite levels, mediation analysis recovers biological plausible mediated transcript-metabolite-trait triplets that are not captured by simple transcript-to-trait MR (*176*).

use of LD information to estimate pleiotropy-robust causal MR effects (*177*). The method was shown to improve power and reduce type 1 error over other MR methods (*177*). Higher false positive rates are a major distinction between MR and colocalization methods. While MR is prone to false positives, colocalization is more conservative as it requires strong statistical evidence of associations with traits to conclude colocalization, potentially generating false negatives (*178*). Other caveats are common to both methods. For instance, some genes lack QTLs altogether, while others share QTLs with other genes in proximity, making it hard to distinguish the causal gene. In addition, many QTLs are context-dependent and can only be captured in certain cell types, developmental stages, or environmental exposures (*151*–*153*, *179*). Recent years have seen the release of ever larger and deeper QTL studies (Table 1.6), which should at least partially address the mentioned caveats.

The presented extensions to GWASs only detail a fraction of the methodological developments aiming to address challenges in the field of genetic association studies, which mainly focused on SNPs and more recently rare SNVs. One area that remains underexplored is how to adapt these tools to study a broader set of variants, notably SVs, which require particular considerations. In the context of my dissertation, I focused on CNVs. The following section provides an overview of mechanisms of CNV formation, tools to detect them, and special considerations when interpreting their phenotypic consequences.

1.4 Copy-number variants

1.4.1 CNV mechanisms

The specific mechanisms of formation underlying a given CNV can often be inferred from the genomic context surrounding its breakpoints as different mechanisms leave different mutational signatures. Here we will focus on the three main mechanisms at the origin of most CNVs (*180*–*182*), although it should be noted that these can further be divided into more precise subcategories. A first consideration is whether the CNV is **recurrent**, i.e., whether unrelated individuals have the same breakpoints. If this is the case, the CNV is likely to be mediated by **non-allelic homologous recombination (NAHR)** between regions of repeated elements sharing high similarity called **low copy repeats (LCRs)**. Also called **segmental duplications**, these regions are ≥ 1 kb long and typically share ≥ 90% sequence identity (*180*). Frequency of NAHR is affected by the distance between LCRs, their degree of homology, size, orientation, as well as the identity of the sequence itself (e.g., GC content or presence of *PRDM9* recombination hotspot motifs). Recombination between sequences with lower degree of homology[32], have been described to cause intragenic CNVs (*183*–*187*), but current consensus suggests that they arise through mechanisms distinct from NAHR (*180*, *181*). Due to their high degree of similarity, non-allelic LCRs might align, or rather *misalign*, during meiosis or mitosis. Ensuing recombination between LCRs with **direct orientation** (i.e., same direction) typically results in one cell with a deletion and one with a duplication (Figure 1.15A). More precisely, during meiosis, faulty recombination can occur between homologous chromosomes (interchromosomal), sister chromatids (interchromatid), or paralogous LCRs on the same chromatid (intrachromatid). The latter results in a deletion and a ring chromosome, stipulating that deletions should be more frequent than duplication, with the difference in frequencies reflecting the frequency of intrachromatid NAHR. This prediction was validated experimentally in a study identifying twice as many deletions than duplications at NAHR hotspots in the sperm population of five men (*188*). It remains unclear whether this ratio holds for all loci, varies between male and female gametogenesis[33], and to which extent it is affected by differential selection on deletions and duplications. Furthermore, twin studies have suggested that other genetic, environmental, and life history factors might regulate NAHR rates (*194*). Balanced SVs result from NAHR between LCRs with **inverted orientation** (i.e., opposing direction) or on different chromosomes (Figure 1.15B-C).

3R CNV mechanisms

Most CNVs appear through either of these three mechanisms:

- **Recombination**
- **Repair**
- **Replication**

32: Termed **hom*e*ologous** sequences (75-91% homology), these include for instance mobile *Alu* elements.

33: Some CNVs have been found to predominantly arise in male (*189*–*191*) or female (*192*, *193*) gametogenesis.

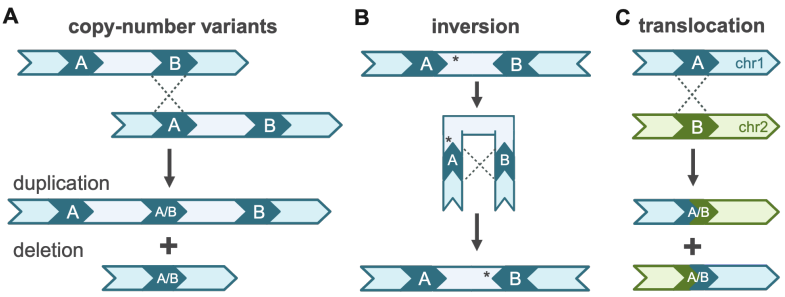


Figure 1.15: Non-allelic homologous recombination outcomes.
(**A**) NAHR between LCRs with direct orientation on the homologous chromosomes or sister chromatids results in a deletion and a duplication. (**B**) NAHR between LCRs with inverted orientation on the same chromosomal arm results in inversions. (**C**) NAHR between LCRs on different chromosomes results in a translocation.

While NAHR is the best-studied mechanism for CNV formation, the majority of CNVs do not have recurrent breakpoints (*195*). Non-recurrent CNVs are usually attributed to either **non-homologous end joining** (NHEJ) or **replication-based mechanisms**. NHEJ is a salvage cellular mechanism for the repair of physiological and pathological double-strand DNA breaks. CNVs caused by this mechanism are characterized by a scaring pattern caused by the editing of DNA ends (i.e., small deletions and insertions) before reattachment (Figure 1.16), and in some cases, short stretches (1-3 bp) of shared nucleotides at breakpoints, i.e. *microhomology* (*196*). Double-strand DNA breaks can also be repaired by another error-prone mechanism, microhomology-mediated end joining (MMEJ) (Figure 1.16). Unlike NHEJ, MMEJ is dependent on microhomologies but leaves a similar scaring pattern than NHEJ, making it difficult to distinguish these mechanisms (*197*).

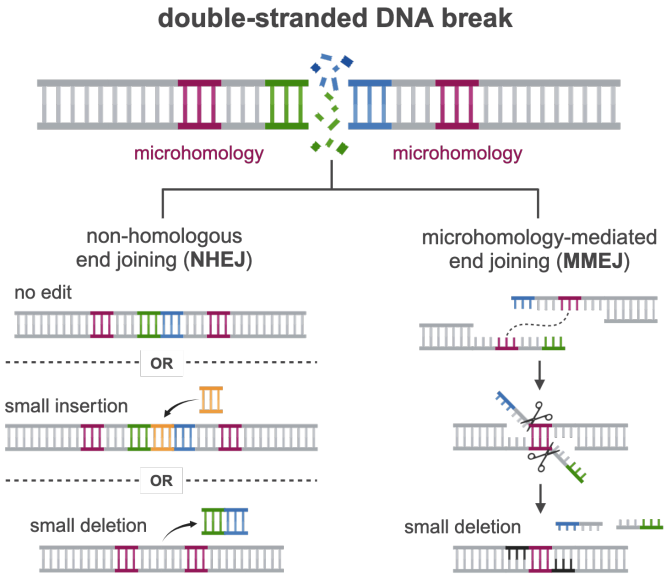


Figure 1.16: CNV formation through repair mechanisms.
Schematic representation of two error-prone double-stranded DNA break repair mechanisms. Non-homologous end joining (NHEJ; left) will rejoin broken ends, leading either to an identical sequence or the insertion (yellow) or deletion (blue/green) of a few nucleotides. Microhomology-mediated end joining (MMEJ; right) occurs through annealing of two single-stranded overhangs with exposed microhomology (pink). After trimming of the 3′ extended strands, the gaps are filled (black) and ligated, resulting in a small deletion.

Non-recurrent CNVs generated by replication-based mechanisms have larger regions of microhomology (2-33 bp) and frequently insert template fragments of up to 100 bp around breakpoint regions. There exist many subcategories of replication-based mechanisms, including fork stalling and template switching (FoSTeS), serial replication slippage, break-induced replication (BIR), and microhomology-mediated break-induced replication (MMBIR) (*180*, *197*) but a detailed review of these mechanisms goes beyond the scope of this introduction. Notably error-prone, replication-based mechanisms have been linked to complex genomic rearrangements with multiple breakpoints, as well as increased local mutation rate, creating clusters of *de novo* SNVs and indels near breakpoints (*195*, *198*).

Importantly, non-recurrent CNVs do not have a random distribution and tend to cluster at specific sites, often co-occurring with repeated sequences. This suggests that while repeats do not mediate NHEJ or replication-based mechanisms, they do promote non-recurrent CNV formation. One explanation is that sequences such as fragile sites or short inverted repeats lead to the formation of secondary DNA structures (*180*, *197*). Secondary structures might promote stalling and eventually

collapse of the replication fork, generating double-strand DNA breaks and leading to genomic instability. Many other features and mechanisms might contribute to sensitizing some regions to CNV formation and advances in CNV technologies will be key in unraveling these.

1.4.2 CNV detection tools

Cytogenetics

The first detected CNVs were gross chromosomal abnormalities (> 5 Mb) that could be detected through **karyotyping**, a technique developed in the 1950s. Through staining of cytobands and alignment of condensed chromosomes of mitotic cells, karyotypes enable detection of CNVs with low **resolution**. In the context of CNV detection, resolution refers to the ability to detect small events, which is linked to the concept of **breakpoint resolution**, which reflects the ability to determine the exact position of a CNV's breakpoints. As a remnant of the early days of cytogenetics, large recurrent CNVs are still named according to the cytogenic band in which they occur. A major improvement came from **fluorescence *in situ* hybridization (FISH)** in the early 1980s. FISH relies on fluorescently labeled **hybridization probes**. The latter are short sequences of single-stranded nucleic acids that bind to the targeted sequence with a high degree of complementarity in the probed genome. This allows visualization of the position of the targeted genomic region (or its absence) under the microscope, allowing detection of events with a resolution of 100-200 kb.

Microarrays

The major limitations of FISH are that its resolution is bound to the resolution of the used microscope and that it only allows detection of CNVs in actively probed regions. **Chromosomal microarray analysis (CMA)**, addresses this by providing a solid support on which hundreds of thousands of probes are immobilized. This allows to probe hybridization at a genome-wide scale in a single experiment. Two major types of arrays are used for CNV detection. **Comparative genomic hybridization (CGH) arrays** were specifically designed for CNV detection in the 1990s (Figure 1.17). An alternative approach is to use the signal produced by **SNP arrays**, later referred to as (micro)arrays. As this is the technology that I used to infer CNV calls in my research, I will dedicate the largest fraction of this section to this technology, even though sequencing-based technologies offer considerable advantages and have become increasingly used over the course of my PhD.

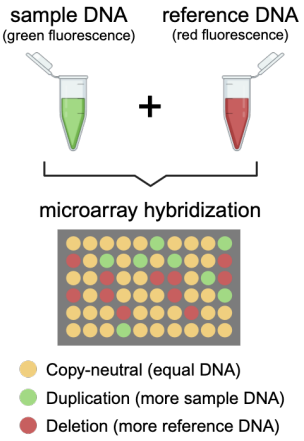


Figure 1.17: Comparative genomic hybridization CNV detection.
In CGH, sample and reference DNA labeled with different fluorescent dies are added simultaneously to the array, where they hybridize competitively. For each probe (dot on the array), the intensity of the two colors is assessed. Probes, where the sample and reference do not have the same signal intensity, correspond to the genomic locations of CNVs.

The clear advantage of SNP arrays is that they allow the simultaneous detection of SNPs and CNVs. Given that small mutations are more widely studied, SNP microarray data are already abundantly available for numerous samples and can be generated at low cost. From a technical perspective, SNP arrays have the advantage that they also detect regions with LOH. LOH regions are informative for clinical diagnosis as they indicate consanguinity or **uniparental disomy (UPD)**[34]. SNP arrays differ from CGH arrays in the fact that no reference DNA sample is used. Instead, two readouts termed **log R ratio (LRR)** and **B allele frequency (BAF)** are used to detect CNVs (Figure 1.18). Hybridization intensity is quantified through the LRR, which is computed as the $\log_2$ of the ratio between the observed and expected signal[35]. In addition, detection

34: Genetic abnormality wherein both chromosomes (or chromosome segments) originate from the same parent.

35: The logarithmic behavior of the LRR makes it easier to detect deletions than duplications, as the increase in LRR does not grow linearly with the copy number. This renders multiallelic CNV detection particularly difficult. Like for CGH, position of additional copies cannot be determined, but sequencing studies suggest that they most often appear in **tandem** (i.e., adjacent) with direct orientation (*199*).

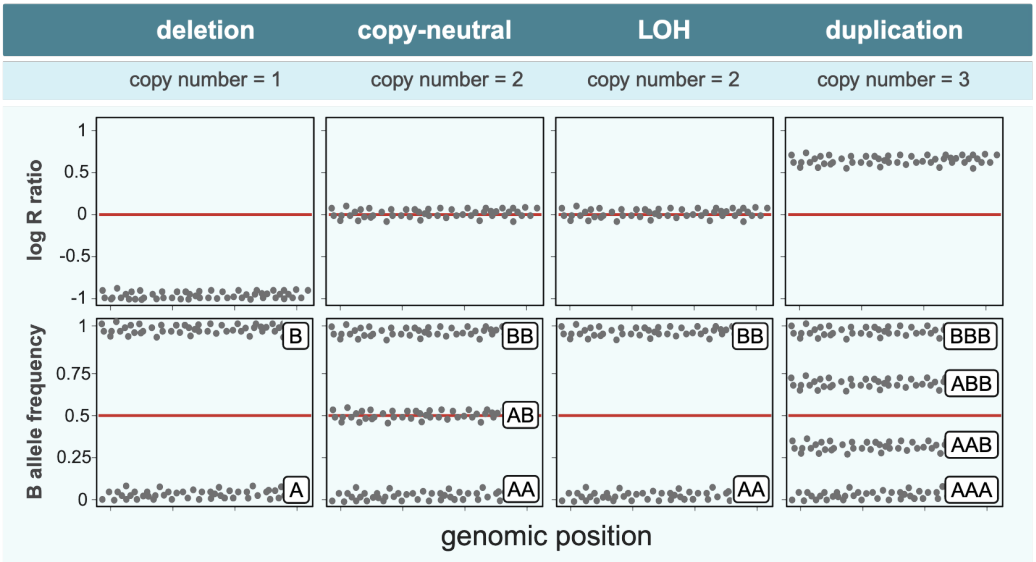


Figure 1.18: SNP array CNV detection.
Signature of deleted, copy-neutral, loss of heterozygosity (LOH), or duplicated regions, with corresponding copy number state. Each dot represents a probe assessed by the SNP microarray, ordered by genomic position (x-axis). The top plots show the log R ratio (LRR; y-axis), which is null (red line) for regions present in two copies ($\log_2(2/2) = 0$). Conversely, it is negative for deletions ($\log_2(1/2) = -1$) and positive for duplications ($\log_2(3/2) = 0.58$). The bottom plots the B allele frequency (BAF; y-axis). In deleted regions, a single copy is present with the only possible genotypes being A (BAF = 0/1) or B (BAF = 1/1). Copy-neutral regions have three possible genotypes: AA (BAF = 0/2), AB (BAF = 1/2), or BB (BAF = 2/2). In LOH, only the two homozygous genotypes are observed. Duplications generate a larger spectrum of BAF values, i.e., AAA (BAF = 0/3), AAB (BAF = 1/3), ABB (BAF = 2/3), and BBB (BBB; BAF = 3/3). Genotype groups are indicated on the plot.

algorithms also make use of a qualitative readout derived from the array's primary purpose as an SNV detection tool, the BAF. The BAF corresponds to the fraction of the intensity signal generated by the minor (B) allele. How these are used to detect CNVs is described in Figure 1.18.

Various software tools have been developed to call CNVs based on genome-wide BAF and LRR values (*200*–*202*). Among the most popular ones is PennCNV (*203*), which relies on a **hidden Markov model (HMM)** with six states. These correspond to copy numbers ranging from zero to four, plus an additional state for LOH. HMMs aim at inferring a sequence of states, based on the assumption that the probability of observing a given state depends *only* on the current state. In other words, if we consider two adjacent probes p_0 and p_1, the copy-number state of p_1 will depend only on the copy-number state of p_0. HMMs have two parameter types, whose values are known: **emission probabilities**, which corresponds to the probability of each state, and **transition probabilities**, which correspond to the probability of transitioning from one state to another. The latter are used to estimate the most likely state-transition path across probes on a chromosome, given the observed data. Besides the BAF and LRR values, PennCNV uses the population frequency of the B allele, the distance between SNPs, and the GC content to improve its performance. PennCNV outputs CNV calls, along with CNV- and sample-level parameters. Because multiple probes are required to reliably call a CNV[36], resolution is determined by the density of probes in a given region and is thus strongly dependent on the array design. Hence, regions of known biological importance tend to have a higher probe density, while others whose function is less well known or for which it is hard

36: Microarray readouts are particularly noisy and prone to biases e.g., due to probe cross-hybridization or variability in CG content.

to design probes (e.g., LCRs) have lower probe coverage. As LCRs are key to CNV formation (see section 1.4.1), some arrays have added special non-polymorphic probes in these regions (*204*). Overall, a resolution in the orders of 10-100 kb can be expected for modern microarrays.

PennCNV still suffers from high false positive rates (*205*). To address this, Aurélien Macé, a previous PhD in the group, introduced an approach to attribute a **quality score (QS)** to each PennCNV call, reflecting the probability that this CNV is a consensus call (> 70% probe overlap) with two other calling algorithms, QuantiSNP (*202*) and CNVpartition, and thus a true positive (*206*). The score is defined as:

$$QS = \frac{1}{1 + e^{(-\beta_0 + \sum_{i=1}^{n} \beta_i V_i)}} \tag{1.11}$$

where β_i corresponds to the weights for the PennCNV output parameters[37] that were found to significantly affect the probability for a CNV call to be consensus, and V_i the actual parameter value. Weights have been pre-computed for deletions and duplications separately and are publicly available. By multiplying the QS by -1 for deletions, a continuous value reflecting the **probabilistic dosage** of the region is obtained, with values close to -1, 0, and, 1 reflecting a deletion, copy-neutral, and duplication state, respectively. The QS was shown to perform better than previously used filtering strategies and simulations found that it could yield up to 20% power increase in the context of association studies (*206*). The QS framework was further developed to use sequencing, RNA expression, and DNA methylation data to gauge the reliability of PennCNV calls (*207*). The rationale is that a true positive CNV call is more likely to be detectable by an alternative detection method (e.g., sequencing), to alter expression levels of the genes it deletes, disrupts, or duplicates, and to affect the intensity of DNA methylation signal in the overlapping region. This omics-informed QS outperforms the consensus-based QS by explaining a larger fraction of the phenotypic variation in real-data association studies conducted in the EstBB and UKBB. While a probabilistic dosage has many benefits, in practice, QSs tend to follow a bimodal distribution with peaks around 0 and 1 (*208*). Hence, high-confidence CNVs can be selected by applying a cutoff at $|QS| > 0.5$ without losing much information. As is described in detail in the ensuing chapters, this is the approach I followed to retain CNVs for association studies.

37: Considered were three CNV-level (CNV confidence score, length [bp], length [probes]) and seven sample-level (mean LRR, LRR SD, mean BAF, BAF SD, BAF drift, waviness factor, and total CNV numbers) parameters. The three BAF parameters have weights of zero for deletion QS computation. Length in number of probes has the highest weight.

Sequencing

By assessing the identity of every single nucleotide, **sequencing strategies** can contend with some of the major limitations of microarray CNV calling, namely the dependence on probe coverage, the low resolution, the inability to accurately determine breakpoints, the calling of multiallelic CNVs, and the inference of the position and orientation of duplications. Sequencing approaches can be split according to the length of generated **reads**, which correspond to the inferred sequences.

Modern **short-read sequencing**, also referred to as **next-generation sequencing (NGS)**, is based on massively parallel sequencing of millions of short reads that are 50-300 nucleotides long. NGS can be applied to perform either **whole exome sequencing (WES)**, which uses capture and enrichment methods to only sequence the protein-coding portion of the genome, or **whole genome sequencing (WGS)**, where the entire

Table 1.7: Approaches to short read sequencing-based CNV detection.
Summary of the four main approaches to CNV detection based on short-read sequencing data, with weaknesses and strengths. The second column indicates whether the approach relies on paired-end sequencing (PR), i.e., both ends of the DNA fragments are sequenced. Y = yes. Approaches can be combined for optimal detection.

Approach	PE	Description	Strengths	Weaknesses
Paired-end mapping	Y	Identify mate pairs with discordant mapping. Specifically, mates will map too far apart in the presence of a deletion and too close or in the wrong order in the presence of a duplication.	- Repeats (providing the reads do not map within them)	- Resolution limited by DNA fragment size
Read-depth		Read coverage is proportional to the region's copy number, i.e., increased for duplications and decreased for deletions.	- Intuitive, popular	- Balanced SVs - Repeats
Split-read	Y	Detect mate pairs where only one mate can be uniquely mapped. The other mate is assumed to overlap the SV breakpoint and is split into multiple fragments that are tentatively mapped at the presumed breakpoint. The latter's position is inferred from the mapped mate and the DNA fragment length.	- Accurate breakpoint	- Read length - Repeats
Sequence assembly		Perform *de novo* assembly of the sample genome and compare it to the reference.	- Detect small events	- High coverage - Repeats

genome is assessed. WES is cheaper and represents a good option to detect small, exon-level CNVs that are typically missed by microarray approaches and can be of clinical significance (*209*–*211*). Furthermore, a recent study found that WES CNV detection performed based on a workflow integrating multiple CNV calling algorithms has similar sensitivity than high-resolution CMA to detect pathogenic CNVs in a patient cohort (*212*). WES CNV calling can be further improved by leveraging off-target reads (i.e., up to 60% of reads generated by WES do not map to regions specifically enriched for) to call CNVs genome-wide (*213*). Still, WGS remains the gold standard to fully address limitations of microarray-based CNV calling, simultaneously allowing the detection of small and large CNVs, as well as other SVs, at most locations in the genome. Many software tools have been developed to call CNVs from short-read sequencing data (*214*). Among the most widely used are cn.MOPS (*215*), LUMPY (*216*), DELLY (*217*), Manta (*218*), and GATK-gCNV (*210*). They can be divided into four approaches (Table 1.7), **read-depth approaches** being the most popular. Despite the abundance of tools, no method so far performs better than the others on all fronts, and each has its own strengths and caveats. As such, multiple tools are typically used to generate a set of high-confidence CNV calls for downstream analyses, paralleling the idea behind the consensus QS previously discussed in the context of microarray CNV calls.

One important limitation of NGS is that the read length is shorter than many SVs, hindering their detection. With reads that can reach over 200 kb in length and new generation sequencers having high accuracy, **long-read sequencing (LRS)** or third generation sequencing, has come forward as a solution to this problem (*219*). The first large LRS studies systematically identify about three times more SVs than studies based on short reads (*10*, *11*, *14*, *34*–*36*). Another advantage of LRS is that it allows for nucleotide modification detection, such as DNA methylation. Tools for CNV detection from LRS data have been developed, including Sniffles2 (*220*), pbsv, cuteSV (*221*), or SVIM (*222*), which are often used in combination. Elected method of the year by Nature Methods in 2022 (*223*), there is no doubt that LRS will revolutionize sequencing. Yet the technology remains expensive and has limited throughput, which – so far – has hindered its application to large biobanks with hundreds of thousands of samples. LRS is usually used in a WGS setting but a solution

to lower its cost is to apply it in a targeted fashion (*224*).

Optical mapping
Another technology in the arena of SV detection is **optical mapping**, which relies on fluorescent labeling of specific motifs in large DNA fragments (> 250 kb) that are subsequently passed through nanochannels for imaging. The signal is converted into a digitized genomic map used to call CNVs by comparing it to a reference map. Optical mapping has a much higher resolution than classical cytogenetic techniques, allows balanced SV detection unlike array-based approaches, and is less biased than sequencing-based methods. It thus represents a suitable detection tool for large and complex SVs, especially those involving repetitive structures such as segmental duplications (*225*), even though it was noted that the technology is not ideal for aneuploidies and **Robertsonian translocations**[38] (*226*).

38: Translocation between acrocentric chromosomes (Figure 1.1), leading to the fusion of their q arms. Fused p arms are often lost, resulting in 45 chromosomes. As the p arms contain rRNA copies present on other acrocentric short arms, no critical genetic material is lost.

1.4.3 Functional consequences of CNVs

Through their size and diversity, SVs can induce a plethora of functional consequences (*21*, *227*, *228*). Often affecting multiple protein-coding genes and/or non-coding regulatory elements simultaneously, predicting their functional impact is typically more complex than for short variants. Because of their large genetic footprint, SVs are more likely to disrupt important genetic entities or configurations, which translates into pathological consequences in the carrier. This idea is supported by SV size inversely correlating with frequency (*34*, *35*), suggesting negative selection against larger – and likely more deleterious – SVs. Indeed, SVs tend to have low frequencies in the general population, and excess of rare SVs cannot be explained by lower *de novo* rates (~0.3 events per generation) (*21*, *34*, *35*). Here, I briefly review the functional impact of SVs at the molecular level before focusing on their pathophysiological consequences.

Molecular consequences
A key distinction is whether the SV affects protein coding sequence or not. Full deletions or duplications of a single gene have the most straightforward interpretation, as the change in copy number will generally lead to lower (i.e., a "human knockdown") or higher gene expression, respectively. It should be noted, however, that the change in expression is usually not entirely proportional to the copy number, i.e., deletions will not halve the expression but reduce it, while duplications will increase it but not by fifty percent (*227*). There are exceptions to this paradigm, and even if such a change is observed at the transcript level, it might not be reflected at the protein level. Alternatively, SV breakpoint might fall within the gene body, which most of the time results in LoF. Notable exceptions include gene fusion events that can lead to GoF (through newly acquired function or dominant negative effect) if the genes have the same orientation and the reading frame remains intact.

Severity of functional consequences is determined by whether affected genes are sensitive to changes in **dosage**. SVs leading to gene LoF have a functional impact if they affect a haploinsufficient gene or if they unmask another pathogenic variant in the remaining copy of a gene associated

with a recessive inheritance (Table 1.5), leading to compound heterozygosity (Figure 1.3). Conversely, gene duplication SVs have a functional impact if they affect triplosensitive genes. While haploinsufficiency can be inferred from abundant data on LoF SNVs associating with dominant phenotypes, knowledge on triplosensitivity is more sparse. Throughout my PhD, continuous metrics assessing the probability for haploinsufficiency (pHaplo) and triplosensitivity (pTriplo) of all autosomal genes have been developed. Overall, 2,987 haploinsufficient and 1,559 triplosensitive genes – including 648 that are uniquely triplosensitive – were identified (*229*), helping with the interpretation of novel SVs.

SVs that do not affect protein-coding regions can exert their impact through **positional effects** leading to altered gene expression (*230*). SVs occurring within the boundaries of a **topologically associating domain (TAD)**, defined as conserved units of self-interacting genetic region, might alter dosage of regulatory elements. By breaking up activating (e.g., enhancer) or repressive (e.g., silencer) elements, SVs decrease or increase the expression of genes under the control of these elements, respectively. Similarly, duplication of an enhancer can lead to gene overexpression. Accordingly, SVs are depleted from regulatory elements (*34*, *35*, *231*). In Chapter 2, I describe a non-coding, height-associated CNV region in the gene desert surrounding *SHOX* (*208*). CNVs disrupting distant *SHOX* regulatory elements (> 250 kb) have been shown to **phenocopy**, or cause the same phenotype, as disruption of the gene itself (*232*), illustrating how SVs can have functional consequences over long distances. Indeed, non-coding SVs form potent eQTLs. Common SVs (MAF $\geq$ 5%) are more likely to act over long ranges and have stronger *cis* effects than SNV-eQTLs, accounting for about 8% of the transcriptome's heritability (*113*). Others found that rare CNVs strongly contributed to extreme patterns in gene expression (*114*–*116*). Integrating gene expression with long read-sequencing data was shown to efficiently prioritize causal SVs in rare disease patients lacking a genetic diagnosis (*233*). Alternatively, SVs can also span across TAD boundaries, leading to more profound changes in the **genome 3D organization**. Typically, deletions cause TAD fusion, duplications create neo-TADs, while inversions and translocations shuffle TADs. TAD disruption can lead to **enhancer adoption**, which describes the ectopic expression of a gene brought under the control of an enhancer normally located in another TAD, or **enhancer disconnection**, leading to loss of expression of the enhancer's target due to the enhancer being shuffled to another TAD. Such positional effects can have severe consequences, especially when they affect TADs encompassing key developmental genes and their regulatory elements, the context in which they have been predominantly studied (*230*).

Phenotypic consequences

SVs disrupting a single gene mediating a key physiological process and whose dosage is tightly regulated can have serious phenotypic consequences, often confined to a single or a few related physiological systems (Figure 1.21A). Table 1.8 provides a few examples of SVs involved in specific phenotypes' **etiologies** or causes. Disruption of these genes by other classes of variants (e.g., SNVs) can also give rise to the same phenotypes, so that SVs only account for a fraction of cases[39]. Among these examples, the link between genotype and phenotype can be explained through different models that reflect different **genetic architectures**.

39: Disorders whose onset is triggered by **repeat expansions** (e.g., Huntington's disease or Friedreich's ataxia) represent an exception, as for them, SVs tends to represent the main genetic etiology.

Genetic architecture & variable expressivity and penetrance

The **genetic architecture** describes the number and characteristics of genetic variants that contribute to a given phenotype (Figure 1.19).

In a classical view on genetic architecture, rare diseases are caused by a single or a few variants with low frequency but a strong effect size. These diseases are often termed **monogenic** (i.e., caused by a single gene) or **Mendelian**, in reference to the fact that disruption of the gene is both necessary and sufficient to develop the disorder. This definition implies that the variant has a near **complete penetrance** (Figure 1.20) so that the presence of the variant in a family pedigree tracks with the presence of the phenotype and allows to infer its inheritance mode. In Table 1.8, one example is the LoF of *LDLR* causing familial hypercholesterolemia. At the other end of the spectrum are common variants with small effect sizes, which tend to be linked to altered susceptibility to common diseases. In line with the liability threshold model, a large number of such risk-increasing variants – as well as environmental factors – is required to develop a disease that has a **polygenic** (i.e., caused by multiple genes) or **complex** architecture. One example is the copy number polymorphism at *LPA*, which represents one of the many other genetic and environmental (e.g., diet) risk factors for developing coronary heart disease. Some common disorders, such as Alzheimer's disease, can be caused either by the cumulative effect of unknown etiologies (i.e., **idiopathic**) or they can be primarily driven by a single, rare, and highly penetrant variant, such as the *APP* duplication. The latter strongly predisposes for early-onset familial Alzheimer's disease through a Mendelian, autosomal dominant inheritance pattern. Deviating from the archetypal rare variant - rare disease vs common variant - common disease paradigm, the latter example illustrates a more complex view of genetic architecture, where rare variants can contribute to common diseases. Such observations have led to the interesting perspective that many common diseases encompass subsets of rarer diseases with distinct genetic etiologies, subtle phenotypic specificities, and different (usually earlier) ages of onset. Dissection of common disorders into smaller entities is at the heart of personalized medicine strategies and is receiving increasingly more attention. An excellent discussion about the distinctions and commonalities between rare and common diseases can be found in "*Rare Diseases and Orphan Drugs: Keys to Understanding and Treating the Common Diseases*" (*130*).

In recent years, it has become clear that the above-described model represents an oversimplification of a more nuanced reality. Indeed, many variants linked to Mendelian disorders have **incomplete penetrance** and **variable expressivity** (Figure 1.20) (*234*–*237*). Furthermore, the cumulative effect of common variants with small effect size, captured through **polygenic risk score (PGS)**, have been shown to modulate disease risk or severity on top of the risk attributed to rare variants (*238*–*240*). This realization can in part be attributed to a shift from studying rare pathogenic variants solely in clinical cohorts, that by definition are ascertained for the studied phenotype, to assessing their prevalence and effect in population cohorts, where these variants were found in individuals that lacked the associated phenotype (*241*).

My dissertation embraces this perspective by leveraging population cohorts to study the phenotypic expression of CNVs which have historically been studied in clinical settings. Throughout the next chapters,

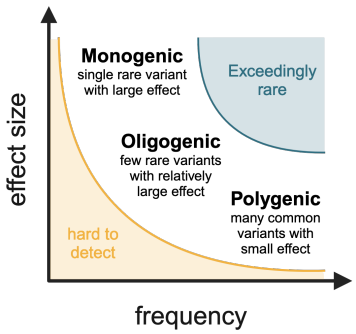


Figure 1.19: Genetic architecture. Genetic architecture is typically defined based on the frequency (x-axis) and effect size (y-axis) of the variants contributing to a trait. It forms a continuum ranging from a monogenic (Mendelian) to a polygenic (complex) architecture. Some complex diseases might have monogenic forms. The genetic contribution of variants in the orange area (rare, small effect size) is hard to detect due to low power (see Figure 1.10). Variants in the blue area (common, large effect) are rare as they are pruned by selection or fixed.

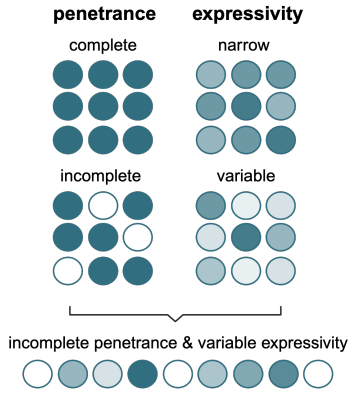


Figure 1.20: Penetrance & expressivity. Penetrance refers to the percentage of individuals with a given genotype that expresses a phenotype. It can be complete (i.e., all individuals express the phenotype) or incomplete (i.e., a fraction expresses it, here 66%). Expressivity refers to the extent to which a phenotype is expressed in individuals with a given genotype. It can be narrow (i.e., similar across individuals) or variable (i.e., strong in some individuals and weak in others). Incomplete penetrance and variable expressivity describe a scenario wherein only a fraction of the individuals carrying the genotype express the phenotype and those that do, express the phenotype more or less strongly.

I demonstrate that rare CNVs have a highly variable expressivity and associate with common disease risk.

Table 1.8: Disorders caused by SVs.
Examples of phenotypes linked to SVs affecting a single or two adjacent genes, alphabetically ordered by phenotype category. Genes are linked to their OMIM page. The type of involved SV is indicated. For repeats, an approximate threshold for pathogenicity is provided. *indicates that the SVs represents a polymorphism, often associated with altered susceptibility. hom. = homozygous.

Category	Phenotype	Gene	Causal SV	Ref
Metabolic	Pseudoxanthoma elasticum	*ABCC6*	Deletions	(*184*, *185*)
	Isolated growth hormone deficiency type 1A	*GH1*	Hom. deletions	(*242*)
	Familial hypercholesterolemia	*LDLR*	Deletions	(*243*)
	Coronary heart disease susceptibility	*LPA*	Repeats (< 22)*	(*244*)
	Rotor syndrome	*SLCO1B1*, *SLCO1B3*	Hom. deletions	(*245*)
Neurvous system	Alzheimer's disease	*APP*	Duplications	(*246*, *247*)
	Duchenne muscular dystrophy	*DMD*	Frameshift CNVs	(*248*)
	Friedreich's ataxia	*FXN*	Hom. repeats (> 200)	(*249*)
	Huntington's disease	*HTT*	Repeats (> 40)	(*250*)
	Spinal muscular atrophy	*SMN1*	Hom. deletions	(*251*)
	Spinal muscular atrophy attenuation	*SMN2*	Duplications*	(*252*)
Hematologic	Iron deficiency anemia susceptibility	*BOLA2*	Low gene copy number*	(*40*)
	Hemophilia A	*F8*	Inversions, deletions	(*253*, *254*)
	Alpha-thalassemia	*HBA1*, *HBA2*	Hom. deletions	(*255*)
	Rhesus negative blood group	*RHD*	Deletions*	(*256*)
Immune	Systemic lupus erythematosus susceptibility	*C4*	Low gene copy number*	(*257*)
	HIV/AIDS susceptibility	*CCL3L1*	Low gene copy number*	(*258*)

Table 1.9: Diagnostic yield of genetic testing for unexplained DD/ID and/or congenital anomalies.
Diagnostic yield for different genetic testing approaches for unexplained DD/ID and/or congenital anomalies (*259*). *Include CMA, candidate single-gene testing, or large gene panel testing.

Diagnostic yield	
Standard testing*	21%
WES	34%
WGS	43%

Although the examples in Table 1.8 illustrate how CNVs are involved in a broad range of disorders, there is one clinical area where SVs have been shown to represent a major risk factor, i.e., neurodevelopmental disorders (NDD) characterized by developmental delay and intellectual disability (DD/ID) and high prevalence of psychiatric conditions, such as ASD. The number of genes associated with NDD is extensive (> 1,500) (*260*) and the contribution of SVs to these disorders is such that in 2010, the American College of Medical Genetics recommended CMA-based CNV screening as the first-tier diagnostic approach for unexplained cases of DD/ID, ASD, and/or multiple congenital anomalies (*261*, *262*). Since 2021, a new consensus was reached, suggesting the usage of genome-wide sequencing-based technologies as a first- or second-tier test, based on the higher diagnostic yield of these technologies (Table 1.9) (*259*). SVs affecting NDD genes can lead to isolated DD/ID or ASD, in which case the carrier will not suffer from any other comorbidities, or lead to syndromic forms of the disease. **Syndromes** are defined as clinical diagnoses requiring the presence of multiple clinical features to co-occur in an affected individual. While for some genetic etiologies the set of features is relatively consistent, other etiologies will present variability in the number and severity of comorbidities present in a given individual.

Since the advent of modern cytogenetics, large CNVs have been associated with syndromic forms of NDD (*263*), which are at the origin of **genomic disorders**. Framed by LCRs, these CNVs recurrently appear *de novo* through NAHR despite negative selection promoting their elimination from the population (*264*). By spanning multiple kilobase pairs and affecting up to several dozen genes, these CNVs have a strong pathogenic potential and represent important susceptibility loci for a

wide range of conditions. Unlike the previously described examples that link a single gene to a single phenotype, these CNVs link multiple genes to multiple phenotypes. To date, close to 100 genomic disorders have been described and an expert-curated set of 66 microdeletion and microduplication syndromes involved in developmental disorders is cataloged in the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER) (*265*). In Table 1.10, I present a selection of well-described CNVs linked to genomic disorders, focusing on those with variable expressivity that are discussed in later chapters.

Table 1.10: Selection of 10 genetic regions linked to genomic disorders.
The genomic coordinates (GRCh37/hg19), size, and number of genes in the locus are provided. In case the phenotype is unambiguously linked to a single gene, the latter is indicated. The name of the recurrent deletion and reciprocal duplication are indicated and the link to the GeneReviews is provided, when available. A selection of hallmark clinical features is provided, although it should be noted that most of these CNVs have highly heterogenous clinical presentation and carriers might not present with all listed features. ASD = autism spectrum disorder; CAKUT = congenital anomalies of kidney and urinary tract; CMTA1 = Charcot-Marie-Tooth disease, type 1A; DD/ID = developmental delay and intellectual disability; DGS/VCFS = DiGeorge/Velocardiofacial syndrome; HNPP = Hereditary neuropathy with liability to pressure plasy; RCAD = renal cyst and diabetes; TAR = thrombocytopenia-absent radius

Locus (GRCh37)	Length	Genes	CNV	Hallmark clinical features
1q21.1 (chr1:145.4-145.7)	360 kb	15 (*RBM8A*)	1q21.1 deletion (TAR syndrome)	Hypomegakaryocytic thrombocytopenia and bilateral radial aplasia.
			1q21.1 duplication	/
15q11-q13 (chr15:23.6-28.4)	4.8 Mb	14 (*UBE3A*)	Maternal 15q11-q13 deletion (Angelman syndrome)	DD/ID, happy demeanor, seizures, sleep problems, microcephaly, gait ataxia, facial dysmorphism.
			Paternal 15q11-q13 deletion (Prader-Willi syndrome)	DD/ID, stereotypical behavior, short stature, scoliosis, obesity, hypotonia, hypogonadism, facial dysmorphism.
15q13.3 (chr16:30.9-32.4)	1.5 Mb	7	15q13.3 BP4-5 deletion	DD/ID, psychiatric disorders, seizures, facial dysmorphism.
			15q13.3 BP4-5 duplication	DD/ID, psychiatric disorders.
16p13.11 (chr16:15.0-16.5)	1.5 Mb	16	16p13.11 deletion	DD/ID, seizures, microcephaly, facial dysmorphism.
			16p13.11 duplication	DD/ID, psychiatric disorders, seizures, facial dysmorphism.
16p12.1 (chr16:21.9-22.4)	520 kb	8	16p12.1 deletion	DD/ID, psychiatric disorders, seizures, short stature, cardiac defects, facial dysmorphism.
			16p12.1 duplication	/
16p11.2 (chr16:28.6-29.2)	220 kb	9	16p11.2 BP2-3 deletion	DD/ID, ASD, macrocephaly, obesity.
			16p11.2 BP2-3 duplication	DD/ID, ASD, microcephaly, underweight.
16p11.2 (chr16:29.6-30.2)	600 kb	27	16p11.2 BP4-5 deletion	DD/ID, ASD, seizures, short stature, vertebral anomalies, macrocephaly, obesity, CAKUT.
			16p11.2 BP4-5 duplication	DD/ID, psychiatric disorders, seizures, microcephaly, underweight.
17p12 (chr17:14.1-15.6)	1.4 Mb	9 (*PMP22*)	17p12 deletion (HNPP)	Hereditary neuropathy with liability to pressure palsy.
			17p12 duplication (CMTA1)	Charcot-Marie-Tooth disease, type 1A.
17q12 (chr17:34.8-36.2)	1.4 Mb	15 (*HNF1B*)	17q12 deletion (RCAD)	DD/ID, CAKUT and tubulointerstitial disease, maturity-onset diabetes of the young, facial dysmorphisms, hyperparathyroidism.
			17q12 duplication	DD/ID, seizures.
22q11.2 (chr22:18.9-21.5)	2.5 Mb	46	22q11.2 LCR A-D deletion (DGS/VCFS)	DD/ID, psychiatric disorders, cardiac defects, cleft palate, immune deficiency, facial dysmorphism.
			22q11.2 LCR A-D duplication	DD/ID, hypotonia.

Importantly, many of these loci are associated with **incomplete penetrance and variable expressivity** (Figure 1.20), further complicating the understanding of underlying molecular mechanisms and the identification of genes that drive phenotypic associations. Golzio and Katsanis propose models to describe the genotype-phenotype relation at these loci (*266*). The **single gene model** stipulates that there is a single primary driver whose altered dosage is necessary and sufficient to produce the phenotype (Figure 1.21B). There are only few examples of recurrent CNVs encompassing multiple genes where this model seems to hold. One of them is located on chromosome 17p12 and represents the primary etiology for Charcot–Marie–Tooth Type 1A (duplication) and hereditary neuropathy with liability to pressure palsies (deletion). The CNV increases disease risk through altered dosage of the dosage-sensitive gene *PMP22* (*267*, *268*), as corroborated by small *PMP22* LoF mutations and animal models (*269*–*271*). A more prevalent model seems to be the one of ***cis*-epistasis**. In its simplest form, a main driver gene is modulated through epistatic (i.e., non-additive) contribution of other genes within the locus (Figure 1.21C). This model can be validated experimentally in animal models with double gene knockdowns. For instance, manipulation of the expression of the 16p11.2 BP4-5 gene *KCTD13* in zebrafish recapitulates the brain size phenotype observed in human CNV carriers (*266*), but was subject to epistatic interactions from other genes in the region (*266*, *272*). Yet, in mouse models, *Kctd13* knockout only leads to neuroanatomical changes upon additional knockout of epistatic interactors (*273*). A recent study showed that over a dozen of 16p11.2 BP4-5 syntenic mouse orthologs contribute to neuroanatomical phenotypes (*274*), underscoring the **oligogenic** architecture (i.e., influenced by multiple genes) of the trait. As such, the *cis*-epistasis model can be further extended to a complex model with multiple driver genes, whose effects are modulated by additive and epistatic contributions of **modifier genes** (Figure 1.21D). Additional complexity results from accounting for the pleiotropic effect of the CNV, as different phenotypes might be regulated by distinct sets of genes (Figure 1.21E). For instance, *TBX6* has been associated with the musculoskeletal and genitourinary phenotypes of 16p11.2 BP4-5 deletion carriers (*275*–*278*), yet at least four other genes in the locus have been involved in modulating these phenotypes (see Chapter 6). Complex interaction models imply that while small-scale mutations affecting individual genes can inform about the individual contribution of genes to specific phenotypes, they cannot fully phenocopy the impact of CNVs. While the above-described models focus on the CNV locus itself, there is now abundant evidence that the phenotypic expression of CNVs linked to genomic disorder is also influenced by interactions with genes outside of the locus (*273*, *279*–*284*). This was initially stipulated by the **"two-hit" model** that explains the phenotypic heterogeneity observed in carriers of recurrent CNVs by the presence of additional rare variants that determine the severity of the expressed phenotype (*285*–*288*) (Figure 1.21F). More recently, evidence has emerged that the entire genetic background on which a CNV occurs contributes to phenotypic heterogeneity (Figure 1.21G). Indeed, an individual's **PGS** can both exacerbate or decrease phenotypic expression (*289*–*291*), even though it should be noted that the power of these studies remains limited.

Challenges linked to CNV-GWASs

As described extensively in the introductions of the following chapters,

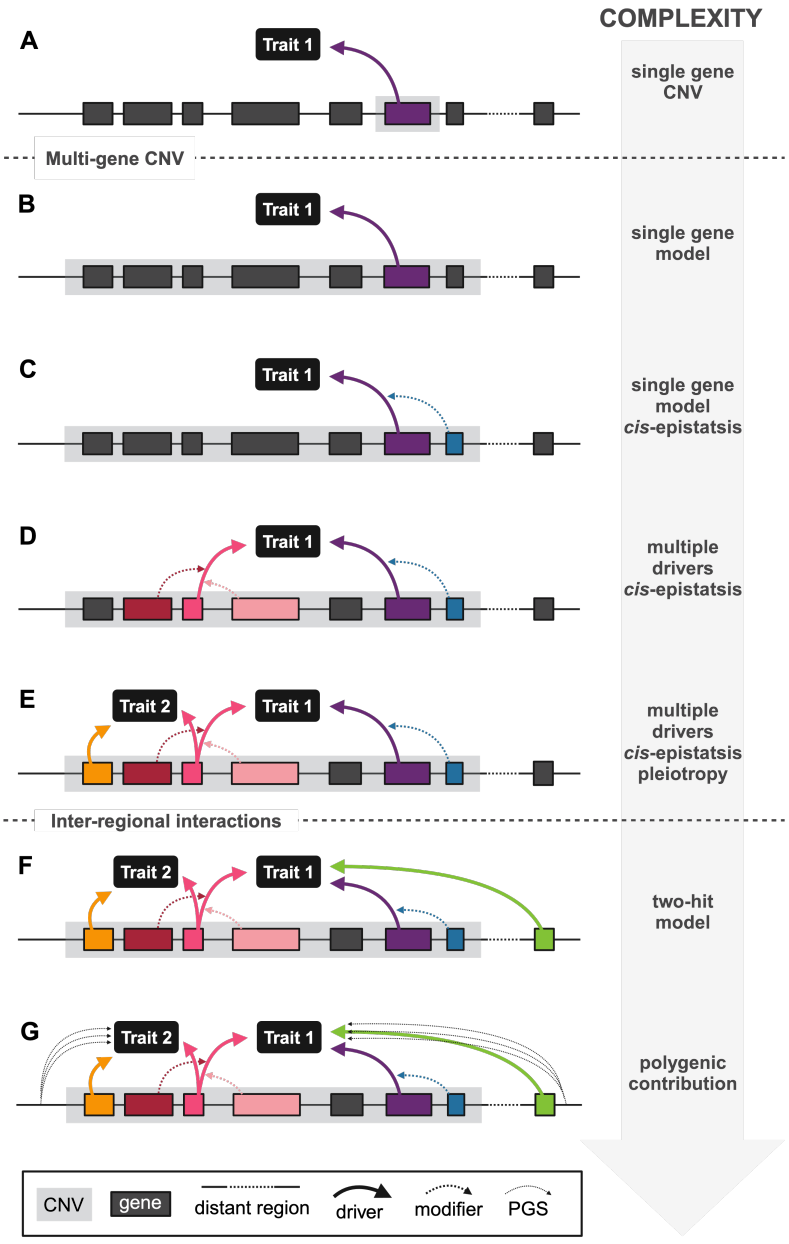


Figure 1.21: Models for CNV-phenotype relationship.
Schematic representation of increasingly complex models that explain the relation between CNVs and associated phenotypes. Contribution of non-genetic factors (e.g., lifestyle, sex, socio-economic status) is not considered here. (**A**) Simple model where a CNV affects a single gene, often leading to a single phenotype (e.g., Table 1.8). The genotype of the present copies of the CNV region contribute to the final phenotypic expression. (**B**-**E**) CNVs that affect multiple genes often associate with multiple traits (i.e., are pleiotropic, e.g., Table 1.10). These phenotypes result from the altered dosage of one or multiple driver genes subject to *cis*-epistatic effect from other modulator genes within the region. (**F**-**G**) Phenotypic expression is influenced by other rare variants through a "two-hit" model (F), as well as by the polygenic contribution of many small-effect size variants genome-wide (G).

at the start of my PhD, only few studies had thoroughly assessed the role of CNVs in complex traits and common diseases (292–295). The genetic basis of complex traits is typically studied through a GWAS framework (1.3) in population cohorts (1.2). The main aim of my thesis was to develop a CNV-GWAS framework to comprehensively assess the role of CNVs in the general population and gain deeper insights into the CNV architecture of complex traits. I implemented this framework in the UKBB, which at the start of my PhD had genotype microarray data available for ~500,000 samples. Yet, compared to SNP-GWAS, additional methodological considerations were needed (205). Indeed, array-based CNV detection suffers from high false positive and negative rates, requiring additional steps to deal with uncertainty in CNV call, including stringent quality control steps. Confidence in CNV calls increases with the number of probes that cover the CNV. A correlate of this is that the CNVs used in our association studies tend to be large and confined to

genomic regions with high probe density. In either case, these CNVs are under strong selective pressure – especially in population cohorts – and are thus rare. This reduces statistical power and emphasizes the need for large sample sizes. Another consideration is that CNVs span multiple base pairs. Grouping CNVs is complicated by true biological variability in breakpoints and the low breakpoint resolution of microarray-based CNV calls, making it important to adequately define the testing unit to avoid further power loss. The choice of the testing unit is also an important consideration to allow replication and/or meta-analysis of CNV calls originating from a cohort genotyped with another array. Finally, is crucial to redefine an appropriate multiple-testing strategy that accounts for the unique characteristics of microarray-based CNV calls. Because CNV-GWASs are not yet standard practice, there is no consensus about how to handle these issues.

Throughout my PhD, I hope to have contributed to addressing some of these challenges and helped define good practices for the implementation of CNV-GWASs. This in turn revealed new biological insights regarding the pleiotropy and variable expressivity of CNVs linked with genomic disorders, with important consequences in terms of personalized medicine. The research I conducted also opened the discussion about new avenues to further deepen our understanding of the mechanisms through which CNVs exert their phenotypic consequences, as I will detail in the last two chapters.

Developing a framework for CNV-GWAS

Quantitative traits 2

Once you come up with a premise, you have to work out how it all happened. It's a bit like coming up with a spectacular roof design first. Before you can get it up there, you need to build a solid foundation and supporting structure.

– Linwood Barclay

2.1 Aims 47
2.2 Key Findings 48
2.3 Author Contributions 48
2.4 The individual and global impact of copy-number variants on complex human traits 49

This chapter describes *"The individual and global impact of copy-number variants on complex human traits"* (*208*), which was published in the *American Journal of Human Genetics* and forms the foundation of my dissertation. Here, I present an extended version of the study that incorporates supplemental material.

The study was the most read paper of the *American Journal of Human Genetics* in 2022 and was selected as a remarkable output by the Swiss Institute of Bioinformatics. I furthermore presented this work at international conferences, including the American Society of Human Genetics, the European Society of Human Genetics, and the Swiss Society of Medical Genetics annual meetings, where it received several distinctions.

Finally, I collaborated with the GWAS Catalog to release the first copy-number variant genome-wide association study (CNV-GWAS) summary statistics available through the platform, paving the way for increased awareness around the role of CNVs in shaping complex traits in the general population.

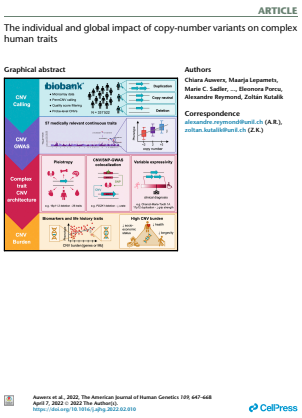


Figure 2.1: Cover of *Auwerx et al., 2022*.

2.1 Aims

This study extends on previous work in the group carried out by Aurélien Macé, who called CNVs from genotype microarray data for ~120,000 UKBB participants and performed association studies between the latter and four anthropometric traits: body mass index, weight, height, and waist-to-hip ratio (*295*). To provide a more comprehensive view of the role and medical relevance of CNVs in the general population, the current study had the following goals:

1. Generate CNV calls for ~500,000 individuals and describe the CNV landscape of the UK Biobank. At the time, this represented one of the largest sets of CNV calls, providing the necessary statistical power to study the impact of rare CNVs in the general population.
2. Develop a statistically robust pipeline to establish associations between genome-wide CNVs and quantitative traits through different dosage mechanisms of action: a mirror model assessing the additive consequence of each additional copy and duplication- and deletion-only models that assess the individual contribution of duplications and deletions to the phenotype, respectively.

Outreach:

- → Unil press release (French)
- → SIB press release
- → Radio interview (French)
- → AJHG journal club

Data & code availability:

- → CNV frequencies
- → CNV-GWAS summary statistics
- → GitHub

3. Apply the developed pipeline to 57 medically relevant quantitative traits. The large number and diversity of assessed phenotypes put forward general patterns describing the CNV genetic architecture of complex traits and allow in-depth follow-up analyses of specific associations.
4. Assess the global impact of CNVs on quality of life and lifespan.

2.2 Key Findings

Describing the landscape of large – and thus often deleterious – CNVs detectable by microarrays in the UK Biobank, we found that 39% of participants carried at least one high confidence CNV and that these affected over 80% of the human genome. These higher-than-anticipated numbers indicate that CNVs represent a major source of genetic variation within the general population and that biobanks can be used to study the role of this mutational class outside of the clinical setting in which they have typically been described.

We developed a CNV-GWAS framework that allowed us to overcome two of the major hurdles related to CNV association studies: i) deal with variability in CNV breakpoints across individuals and ii) distinguish between distinct CNV dosage mechanisms of action. We identified 131 associations. These involve 47 of the 57 tested phenotypes and 28 unique genomic regions, most of which had previously been linked to genomic disorders and accordingly were found to exhibit high levels of pleiotropy.

By dissecting over twenty associations, we provided insights into the biology and epidemiology of these examples and established bridges between rare and common diseases. Indeed, 38% of our CNV-GWAS signals mapped to regions in which common genetic variation had previously been linked to a similar phenotype. Conversely, we identified CNVs that led to subclinical manifestations reminiscent of the phenotypes characterizing the Mendelian disorders linked to these loci. Together, this speaks for the presence of convergent genetic mechanisms, strengthening our confidence in the biomedical relevance of our findings. It also brings forward a complex and nuanced model of variable expressivity and incomplete penetrance that links a given genetic region to a spectrum of phenotypic alterations with variable clinical severity.

Finally, we investigated the global impact of CNVs on human health by aggregating CNV calls into a CNV burden. This revealed the widespread negative consequences of a high CNV load on over half of the assessed phenotypes. These deleterious consequences extended to an individual's socio-economic status and proxied lifespan, emphasizing how CNVs impact the overall quality of life of carriers.

2.3 Author Contributions

This study was conceived by Zoltán Kutalik, Alexandre Reymond, and myself. I performed the CNV calling in the UKBB, with guidance from

Eleonora Porcu and Maarja Jõeloo (previously Lepamets). I carried out the bulk of the analyses, including the CNV-GWAS and burden analyses in the UKBB, as well as the follow-ups on specific examples. Statistical analyses were supervised by Zoltán Kutalik. Zoltán Kutalik and I coordinated the generation of supportive evidence:

- The Estonian Biobank Research Team (Tõnu Esko, Andres Metspalu, Lili Milani, Reedik Mägi, Mari Nelis) coordinated genotyping and sequencing data acquisition in the EstBB and Maarja Jõeloo performed the CNV calling, comparative analyses, and replication study under the supervision of Reedik Mägi.
- Marie Sadler carried out the Mendelian randomization analyses under Zoltán Kutalik and Eleonora Porcu's supervision.
- David Baud and Miloš Stojanov collected and provided access to the CHUV maternity cohort data. Data were analyzed by Marion Patxot.

Results were interpreted by Zoltán Kutalik, Alexandre Reymond, and myself. I designed all the figures and drafted the manuscript, with critical revisions made by Zoltán Kutalik and Alexandre Reymond.

2.4 The individual and global impact of copy-number variants on complex human traits

Chiara Auwerx [1,2,3,4], Maarja Lepamets [5,6], Marie C. Sadler [3,4], Marion Patxot [2], Miloš Stojanov [7], David Baud [7], Reedik Mägi [6], Estonian Biobank Research Team [6], Eleonora Porcu [1,3,4], Alexandre Reymond [1,*], and Zoltán Kutalik [2,3,4*].

Abstract

The impact of copy-number variations (CNVs) on complex human traits remains understudied. We called CNVs in 331,522 UK Biobank participants and performed genome-wide association studies (GWASs) between the copy number of CNV-proxy probes and 57 continuous traits, revealing 131 signals spanning 47 phenotypes. Our analysis recapitulated well-known associations (e.g., 1q21 and height), revealed the pleiotropy of recurrent CNVs (e.g., 26 and 16 traits for 16p11.2 BP4-BP5 and 22q11.21, respectively), and suggested gene functionalities (e.g., *MARF1* in female reproduction). Forty-eight CNV signals (38%) overlapped with single-nucleotide polymorphism (SNP)-GWAS signals for the same trait. For instance, deletion of *PDZK1*, which encodes a urate transporter scaffold protein, decreased serum urate levels, while deletion of *RHD*, which encodes the Rhesus blood group D antigen, associated with hematological traits. Other signals overlapped Mendelian disorder regions, suggesting variable expressivity and broad impact of these loci, as illustrated by signals mapping to Rotor syndrome (*SLCO1B1/3*), renal cysts and diabetes syndrome (*HNF1B*), or Charcot-Marie-Tooth (*PMP22*) loci. Total CNV burden negatively impacted 35 traits, leading to increased adiposity, liver/kidney damage, and decreased intelligence and physical capacity.

[1] Center for Integrative Genomics, University of Lausanne, Lausanne 1015, Switzerland; [2] Department of Computational Biology, University of Lausanne, Lausanne 1015, Switzerland; [3] Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland; [4] University Center for Primary Care and Public Health, Lausanne 1010, Switzerland; [5] Institute of Molecular and Cell Biology, University of Tartu, Tartu 51010, Estonia; [6] Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu 51010, Estonia; [7] Materno-fetal and Obstetrics Research Unit, Department Woman-Mother-Child, CHUV, Lausanne 1011,Switzerland; * **Correspondence**.

Thirty traits remained burden-associated after correcting for CNV-GWAS signals, pointing to a polygenic CNV architecture. The burden negatively correlated with socio-economic indicators, parental lifespan, and age (survivorship proxy), suggesting a contribution to decreased longevity. Together, our results showcase how studying CNVs can expand biological insights, emphasizing the critical role of this mutational class in shaping human traits and arguing in favor of a continuum between Mendelian and complex diseases.

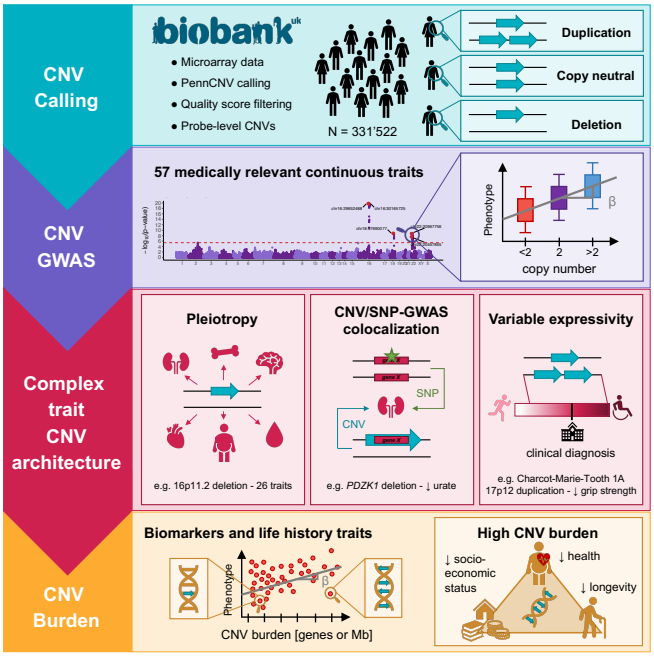


Figure 2.2: Graphical abstract of *Auwerx et al., 2022*.

Introduction

With the advent of genome-wide association studies (GWASs), the polygenic architecture of complex human traits has become apparent (*81*, *171*, *296*). Still, single-nucleotide polymorphisms (SNPs) do not explain the totality of observed phenotypic variability – a phenomenon referred to as *missing heritability* – and one proposed explanation is the contribution of additional types of genetic variants, such as copy-number variants (CNVs) (*108*).

Characterized by the deletion or duplication of DNA fragments ≥ 50 bases (*36*), CNVs represent a highly diverse mutational class that, due to their possibly large size, constitute potent phenotypic modifiers that act through e.g., gene dosage sensitivity, truncation or fusion of genes, unmasking of recessive alleles, or disruption of *cis*-regulatory elements (*297*). Hence, CNVs have been acknowledged to play an important role in human diseases and were identified as the genetic etiology of 65 rare and debilitating genomic syndromes by DECIPHER (*265*). However, early GWASs failed to establish clear links between CNVs and complex traits and diseases (*298*, *299*). Several factors, specific to genome-wide copy-number association studies (CNV-GWASs), contributed to these negative results, such as the low frequency and variable breakpoints of CNVs in the population, as well as uncertainty and low resolution of CNV calls originating from genotyping microarrays (*205*). In recent

years, methodological development, as well as the creation of large biobanks, has allowed bypassing of some of these hurdles. Focusing on a curated set of CNVs, a series of studies characterized the impact of well-established pathogenic CNVs on cognitive performance (*300*), physical measurements (*294*, *301*), common medical conditions (*293*, *302*), and blood biomarkers (*303*). Alternatively, unbiased genome-wide (GW) studies have been conducted (*292*, *295*, *304*–*306*), involving loci not covered by targeted approaches and adding to the growing body of evidence implicating CNVs in complex traits. Notably, a recent study made use of the UK Biobank (UKBB) (*61*) to assess the impact of CNVs on over 3,000 traits, providing the research community with a large population-based CNV-to-phenotype resource (*292*). Using an independent CNV calling and association pipeline and focusing on a set of 57 medically relevant continuous traits, we here confirm previously established associations, uncover biological insight through in-depth analysis of particular CNV-trait pairs, and expose a nuanced role of CNVs along the rare versus common disease spectrum, suggesting that the deleterious impact of CNVs contributes to decreased longevity in the general population.

Materials and methods

Study material

Cohort description

The UK Biobank (UKBB) is a volunteer-based cohort of ~500,000 individuals (54% females) from the general UK population (*61*). Individuals were aged 40-69 years at recruitment and underwent microarray-based genotyping and extensive phenotyping, which is constantly extended and includes physical measurements, blood biomarker analyses, socio-demographic and health-related questionnaires, as well as linkage to medical health records. Participants signed a broad informed consent form and data was accessed through the application number 16389.

Software versions:

- CNV calling: PennCNV v1.0.5 (*203*).
- CNV QC: (*206*).
- PLINK v1.9 and PLINK v2.0.26 (*88*).
- Gene annotation: ANNOVAR (*307*).
- Meta-analysis: GWAMA v2.2.2 (*308*).
- Statistical analyses: R v3.6.1.
- Graphs: R v4.0.3.

The Estonian Biobank (EstBB) is a population-based cohort encompassing ~20% of Estonia's adult population (~200'000 individuals; 66% females) (*62*). Individuals underwent microarray-based genotyping at the Core Genotyping Lab of the Institute of Genomics, University of Tartu, and a subset of ~2'500 samples underwent whole-genome sequencing (WGS). General data, including body measurements, were collected at recruitment. Project-based questionnaires were sent later and filled on a voluntary basis. Health records are updated through linkage with the national Health Insurance Fond and other relevant databases, providing sporadic access to blood biomarker measurements and medical diagnoses. All participants signed a broad informed consent form and analyses were carried out under ethical approval 1.1-12/624 from the Estonian Committee on Bioethics and Human Research and data release N05 from the EstBB.

The Lausanne University Hospital (CHUV) maternity cohort was designed as a serological surveillance study of maternal toxoplasmosis infections and approval from the Ethics Committee of Vaud (CER-VD) was obtained for data reusage under the project ID 2019-00280 to investigate maternal and fetal outcomes. Rhesus (Rh) blood groups were serologically determined for 5,164 women. Reticulocyte count, platelet count, glycated hemoglobin (HbA1c) levels, intrapartum reports, and

International Classification of Diseases, 10th Revision (ICD-10) codes were sporadically collected between 2009-2014.

The CNV landscape of the UK Biobank

Genotype data
Data acquisition and quality control (QC) have been described (*61*). Briefly, UKBB participants were genotyped on two similar arrays (95% probe overlap): 438,427 samples (95 batches) were genotyped with the Applied Biosystems UK Biobank Axiom Array (825,927 probes) and 49,950 samples (11 batches) were genotyped with the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix (807,411 probes). All results in this study are based on the human genome reference build GRCh37/hg19.

Sample selection
Related, gender mismatched, high missingness, non-white British ancestry, and retracted samples were excluded (`used.in.pca.calculation = 0` and `in.white.British.ancestry.subset = 0` in Sample-QC v2 file). To protect the analysis from somatic chromosomal aberrations, we excluded individuals with self-reported (#20001, codes: 1047, 1048, 1050, 1051, 1052, 1053, 1055, 1056, 1056; UKBB update 03/2020) and/or hospital diagnosed (#41270; International Classification of Diseases, 10th Revision [ICD-10] codes mapping to *cancer of lymphatic and hematopoietic tissue's* exclusion range in the Phecode Map 1.2 [beta], accessed 09/12/2020 (*309*); UKBB update 08/2019) blood malignancy. CNV outliers were later removed (see *CNV calling and quality control*). All reported results are for 331,522 unrelated white British UKBB participants (54% females).

CNV calling and quality control
Autosomal and pseudoautosomal CNV calling was performed in parallel for the 106 genotyping batches using PennCNV. Individual intensity files were generated from the B allele frequency (BAF) and Log R Ratio (LRR) files available on the UKBB portal. Missing values (−1) were set to NA. Batch-specific population frequency of the B allele (PFB) files were generated. Probes with missing PFB were removed in a batch-specific way. The hidden Markov Model file for Affymetrix genome-wide 6.0 array was downloaded as part of the PennCNV-Affy package and used without training. The GC model file was generated following instructions of `cal_gc_snp.pl`, using gc5Base downloaded from the UCSC Genome Browser (03/2020). Above-described files were used to call CNVs with confidence score using `detect_cnv.pl` with genomic wave adjustment. Adjacent CNVs (gap ≤ 20% of merged CNV length) were merged with `clean_cnv.pl`. Chromosome X CNVs were called separately with the PennCNV inbuilt arguments `−chrx` and `−sexfile` when running `detect_cnv.pl`. Copy-neutral losses of heterozygosity resulting from male chromosome X hemizygosity were excluded, and adjacent CNVs were merged.

CNVs originating from samples genotyped on plates with a mean CNV count per sample > 100 or from samples with > 200 CNVs or a single CNV > 10 Mb were excluded, as these might be indicative of batch effects, genotyping errors, or extreme chromosomal abnormalities. Hurdles linked to CNV analysis include high false positive rates and variability in breakpoints. To mitigate these issues, a post-PennCNV processing pipeline was used to attribute a quality score (QS) to each CNV and transform calls to the probe level (*206*, *295*). QSs range from -1 to 1 and

reflect the probability of a CNV to be a true positive (−1 = likely deletion; 1 = likely duplication; ∼ 0 = low confidence CNVs). Briefly, PennCNV `filter_cnv.pl` was run separately on autosomal/pseudoautosomal and chromosome X CNVs (with `−chrx`), resulting in two sample-level QC files. These were combined by adding the number of chromosome X CNVs to the autosomal/pseudoautosomal sample-level QC file. A single file containing all called CNVs, as well as associated CNV- and sample-level QC metadata, was generated and used to attribute a QS to each called CNV. Next, linear PennCNV coordinates were transformed into per-chromosome $probe \times sample$ matrices, with entries reflecting the QS attributed to the CNV mapping to these probes. Copy-neutral probes are indicated by 0 and individuals with no CNVs were added as all-0 columns.

Converting CNV calls into PLINK format
QS matrices were converted to PLINK binary file sets. Probes with ≥ 1 high-confidence CNV, stringently defined by $|QS| \geq 0.5$, were retained and encoded into three file sets to accommodate analyses according to a mirror ($PLINK_{CNV}$), duplication-only ($PLINK_{DUP}$), or deletion-only ($PLINK_{DEL}$) association model (`−−make−bed` PLINK v1.9; Table 2.1).

Table 2.1: PLINK encoding of CNVs. Encoding of high-confidence CNVs from quality score (QS) matrices into three PLINK file sets.

	Mirror	**Duplication-only**	**Deletion-only**
PLINK file set	**$PLINK_{CNV}$**	**$PLINK_{DUP}$**	**$PLINK_{DEL}$**
Deletion (QS < -0.5)	AA	00	TT
Copy-neutral (-0.5 ≤ QS ≤ 0.5)	AT	AT	AT
Duplication (QS > 0.5)	TT	TT	00

CNV frequency calculation
Genotype counting was performed for the 740,434 probes stored in $PLINK_{CNV}$ (`−−freqx` PLINK v1.9). 41,670 array-specific probes with genotype count missingness > 5% were excluded and each probe's CNV, duplication, and deletion frequencies were calculated (%)[1], with $Num_{non-CNV}$, Num_{DUP}, and Num_{DEL}, the number of individuals carrying 2, > 2, and < 2 copies of that probe, respectively, and $Num_{CNV} = Num_{DUP} + Num_{DEL}$.

1:
CNV frequency:

$$q_{CNV} = \frac{100 \cdot Num_{CNV}}{Num_{CNV} + Num_{non-CNV}}$$

Duplication frequency:

$$q_{DUP} = \frac{100 \cdot Num_{DUP}}{Num_{CNV} + Num_{non-CNV}}$$

Deletion frequency:

$$q_{DEL} = \frac{100 \cdot Num_{DEL}}{Num_{CNV} + Num_{non-CNV}}$$

CNV association studies in the UK Biobank

CNV probe selection and number of effective tests
Association studies were restricted to probes with a CNV, duplication, or deletion frequency ≥ 0.005% for the mirror, duplication-only, or deletion-only models, respectively. To group probes at the core of CNV regions while retaining variability at breakpoints, we pruned probes at $r^2 > 0.9999$ in $PLINK_{CNV}$, $PLINK_{DUP}$, and $PLINK_{DEL}$ (`−−indep−pairwise 500 250 0.9999` PLINK v2.0). As retained CNV-proxy probes remain highly correlated – much more so than SNPs would be due to classical linkage disequilibrium patterns – the number of effective tests, N_{eff}, was determined (*85*, *295*). Per-chromosome $probe \times sample$ genotype matrices G were generated, with genotypes taking values of −1 (deletion), 0 (copy-neutral), and 1 (duplication). Chromosome-wise N_{eff} were defined as the number of eigenvalues required to explain 99.5% of the variance in G and were summed up, resulting in a GW N_{eff}. N_{eff} was estimated at 11,804, setting the GW threshold for significance at $p \leq 0.05/11{,}804 = 4.2 \times 10^{-6}$. Accounting solely for duplications or deletions resulted in

lower N_{eff} estimates but the same conservative threshold was used for all models.

Phenotype selection
Fifty-seven continuous traits were selected based on data availability and presumed high heritability. Fifty-four were defined as the mean of measured instances. Three were composite traits: Grip strength, as the mean of *hand grip strength left* (#46) and *right* (#47); waist-to-hip ratio (WHR), as the ratio between *waist* (#48) and *hip circumference* (#49); WHR adjusted for body mass index (WHRadjBMI), by correcting WHR for BMI. Two were male-specific (*relative age of first facial hair* (#2375); *hair/balding pattern* (#2395)) and three were female-specific (*age when periods started* (menarche) (#2714); *age at menopause (last menstrual period)* (#3581); *birth weight of first child* (#2744)). Entries "do not know", "only had twins", "prefer not to answer" were set as missing. Traits were inverse normal transformed prior correction for sex (except for sex-specific traits), age (#21003), age^2, genotyping batch, and principal components (PCs) 1-40. Normal phenotypic ranges were retrieved and converted from Symed MediCalc.

Genome-wide copy-number association studies
Associations between the copy number (CN) of selected probes and normalized covariate-corrected traits were performed (`--glm omit-ref no-x-sex hide-covar allow-covars` PLINK v2.0). To avoid interference between the two-letter CNV encoding (Table 2.1) and the assumption of male chromosome X hemizygosity, we (falsely) labeled all individuals as female. For sex-specific traits, samples from the opposite sex were excluded. Three association models were applied: the mirror model ($PLINK_{CNV}$) assessed the additive effect of each additional copy of a probe, the duplication-only model ($PLINK_{DUP}$) assessed the impact of a duplication while disregarding deletions, and the deletion-only model ($PLINK_{DEL}$) assessed the impact of a deletion while disregarding duplications. Given CNV encoding (Table 2.1), effects were homogenized to "T" by multiplying β by -1 when A1 was "A". GW-significant associations ($p \leq 4.2 \times 10^{-6}$) were retained. Stepwise conditional analysis was performed on CNV-GWAS results to determine the number of independent signals per trait. For traits with ≥ 1 GW-significant signal, CNV information was extracted for the lead probe, with genotypes taking values of -1 (deletion), 0 (neutral), and 1 (duplication) for the mirror model, and setting deletions or duplications to missing when considering the duplication-only or deletion-only models, respectively. Lead CNV probe effect was regressed out of the phenotype and association studies were conducted anew. These steps were repeated until no GW-associated probes were identified.

CNV region definition, merging, and annotation
CNV region (CNVR) boundaries were defined by the most distant probe within ± 3 Mb and $r^2 \geq 0.5$ of each independent lead probe (`--show-tags --tag-kb 3000 --tag-r2 0.5` PLINK v1.9). Signals from the different models were merged when involving i) the same trait, ii) overlapping CNVRs, and iii) directional concordance according to a mirror model. CNVR boundaries were defined as the maximal CNVR and characteristics of the most significant model were retained. CNVRs were annotated with `annotate_variation.pl`, with hg19 RefSeq gene names (`--geneanno`; 08/06/2020) and NHGRI-EBI GWAS Catalog (*77*) asso-

ciations (`--regionanno`; 27/10/2021) via ANNOVAR. GWAS Catalog trait synonyms considered are listed in Table S2.1. For each trait, focusing on autosomes, we performed a two-sided binomial test to compare GWAS Catalog SNP-GWAS signal density (accessed 27/10/2021) within CNVRs as compared to the entire genome. The number of SNP-GWAS signals falling within trait-associated CNVRs represents successes, the total length of trait-associated CNVRs [bp] represents trials, and total number of SNP-GWAS signals divided by the autosomal genome length (2,881,033,286 bp) represent the hypothesized signal density.

Replication in the Estonian Biobank

Comparative analysis of CNV quality
About 7,750 EstBB participants were genotyped with Illumina Infinium OmniExpress-24 genotyping array. Samples with genotype call rate < 98%, Hardy-Weinberg equilibrium test p-value < 1×10^{-4}, or mismatched sex based on chromosome X heterozygosity were excluded. Intensity files (LRR and BAF) were created with Illumina GenomeStudio v2.0.4. A PFB file was generated from all samples. Only autosomal probes were carried over to the CNV detection step (709,358 probes), which was performed with PennCNV, analogously to what has been described for the UKBB. Samples with > 200 CNVs or a total length of CNV calls > 10 Mb were excluded. CNVs were attributed a QS, as previously described for the UKBB, and CNVs with $|QS| \geq 0.5$ were retained. To harmonize data with WGS-based CNV calls, duplications < 2 kb and deletions < 1 kb were excluded.

WGS data (30x coverage) was available for ~2,500 EstBB samples. WGS-based autosomal CNVs were called in 5 batches using the Genome STRiP pipeline (*39*). Eleven samples with a number of calls exceeding the median plus three absolute median deviations were removed. The union of the discovered sites was genotyped with the Genome STRiP SVGenotyper module in all batches separately and merged. Duplicate calls were removed using standard Genome STRiP duplicate removal settings (overlap > 50% and duplicate score > 0). Low-quality CNVs and CNVs with call rate < 90% were excluded. Duplications < 2 kb and deletions < 1 kb were excluded. To harmonize data with microarray-based CNV calls, CNVs > 10 Mb were excluded and adjacent CNVs were merged (gap ≤ 20% of merged CNV length).

Quality-controlled Illumina Infinium OmniExpress-24 microarray genotyping and WGS data were available for 966 overlapping and unrelated samples. We constructed a cross-sample PennCNV-CNV profile for each of the 709,358 genotyped probes, taking values of −1 (deletion), 0 (copy-neutral), and 1 (duplication). Similar profiles were constructed based on STRiP-CNV calls and Pearson's coefficient of correlation and number of CNV carriers according to both methods were calculated for each genomic location. For probes with ≥ 1 PennCNV call but no STRiP call, all correlated probes ($r \geq 0.5$, according to PennCNV profiles) within ± 250 kb were retrieved and maximal PennCNV-WGS correlation among these probes was retained. Analyses were repeated on a subset of 5,566 probes overlapping UKBB-trait-associated CNVRs.

Phenotype data
Analyzed traits were queried in the EstBB: height, weight, and BMI

were collected at enrollment; age at menarche and menopause were collected by project-based questionnaires; 41 traits were retrieved from parsed notes in health registries; 11 did not have any corresponding term. Because most phenotypic measurements originate from health registries, they were gathered at different time points and by different practitioners and were only available for a limited subset of participants. In case of repeated measurement, the most recent one was retained. Traits with sample size ≥ 2,000 were selected and inverse normal transformed prior correction for sex (except for sex-specific traits), age, age^2, genotyping batch, and PCs 1-20.

CNV calling and copy-number association studies
Twelve batches containing 202,282 EstBB participants were genotyped with Illumina GSAv1.0, GSAv2.0, GSAv2.0_ESTChip, and GSAv3.0_-ESTChip2. Samples with genotype call rate < 98%, Hardy-Weinberg equilibrium test p-value < 1×10^{-4}, or mismatched sex based on chromosome X heterozygosity were excluded and one of each duplicated sample was retained. Genotypes were re-clustered by manual realignment of cluster location. Intensity files (LRR and BAF) were created with Illumina GenomeStudio v2.0.4. A PFB file was generated from 1,000 randomly selected samples from batch 1. Only autosomal probes overlapping all GSA versions (excluding custom ESTChip probes) were carried over to the CNV detection step (671,035 probes; 242,091 probes overlap with UKBB). Autosomal CNVs were called for 193,844 individuals. Samples originating from two batches with outlier genotyping intensity parameters, as well as genotyping plates with > 3 samples with either > 200 called CNVs or a total length of CNV calls > 10 Mb were excluded. Individual samples meeting these criteria were removed. Among related pairs (KING kinship coefficient > 0.0884), the sample with the most available phenotypes was retained. CNV calls were attributed a QS and encoded into three PLINK binary file sets, following the procedure described for the UKBB. CNV, duplication, and deletion frequencies among the 89,516 unrelated samples remaining after QC were calculated for 671,035 probes and association studies were run as previously described for the UKBB. Using the most significant association model for the 131 merged UKBB signals, we selected the most significantly associated EstBB probe within the boundaries of the UKBB-defined CNVR. EstBB p-values were adjusted to account for directional concordance with UKBB effects[2]. Sufficient genomic variability and phenotypic data were available to assess replication of 61 out of 131 signals, setting the replication threshold for significance at p ≤ $0.05/61 = 8.2 \times 10^{-4}$. Simulations were conducted to estimate the power of our replication study. We defined $\beta_{i,j}$ as the standardized effect of probe i on trait j observed in the UKBB; $q_{i,DUP}$ and $q_{i,DEL}$ the duplication and deletion frequencies of probe i in the EstBB, respectively, and N_j the sample size for trait j in the EstBB. Considering duplication-only, deletion-only, and mirror signals, CNVs were simulated for N_j samples[3]. Normally distributed error terms ϵ were simulated according to $\epsilon \sim \mathcal{N}(0, \sigma^2)$ for N_j samples. For mirror signals, the noise variance σ^2 was defined as $\sigma^2 = \sigma_j^2 - var(CNV) \cdot \beta_{i,j}^2$, with σ_j^2 the observed standardized variance for trait j in the EstBB equaling 1. For duplication-only and deletion-only signals, σ^2 was defined as $\sigma^2 = \sigma_j^2 - q_{i,DUP}(1 - q_{i,DUP}) \cdot \beta_{i,j}^2$ and $\sigma^2 = \sigma_j^2 - q_{i,DEL}(1 - q_{i,DEL}) \cdot \beta_{i,j}^2$, respectively. Phenotypes Y were simulated for N_j samples as $Y = CNV \cdot \beta_{i,j} + \epsilon$. When simulated data contained ≥ 1 CNV carrier, the p-value for the estimated effect size from

2: Direction agreement: $p_{new} = \frac{p_{old}}{2}$; else: $p_{new} = 1 - \frac{p_{old}}{2}$

3:
Duplication-only:
$CNV \sim Bernoulli(q_{i,DUP})$

Deletion-only:
$CNV \sim Bernoulli(q_{i,DEL})$

Mirror:
given $z \sim U(0,1)$

$$CNV = \begin{cases} -1, & z \leq q_{i,DEL} \\ 0, & q_{i,DEL} < z < q_{i,DUP} \\ 1, & z \geq 1 - q_{i,DUP} \end{cases}$$

the linear regression $Y \sim CNV$ was computed and retained. Otherwise, the p-value was set as missing. For each signal, 10,000 simulations were conducted, and power was defined as the fraction of non-missing p-values $\leq 8.2 \times 10^{-4}$. Expected number of replications was estimated as the average power across assessed signals multiplied by the number of assessed signals.

Extended phenotypic assessment

Disease diagnosis

To assess patients' disease status, ICD-10 diagnoses were used (#41270)[4]. Self-reported high alcohol consumption (#1558 as *daily or almost daily*) and γ-glutamyl transferase GGT-increasing drug usage (#20003 as 2038459704 (carbamazepine), 1140865426 (cimetidine), 1140909708 (furosemide), 1140869848 (methotrexate), 1140910706 (phenobarbital), 2038460076 (phenytoin)) (*310*) were further evaluated as potential lifestyle confounders of the 22q11.23-GGT association.

Socio-economic status and life history traits

Six socio-economic and life history traits were additionally considered in the burden analysis[5]. Entries matching "do not know" or "prefer not to answer" were set as missing and if available, average over measured instances were used. Traits were inverse normal transformed prior correction for sex, age (#21003), age^2, genotyping batch, and PCs 1–40, except for *age at recruitment*, which was not corrected for age and age^2.

RHD and hematological traits

Transcriptome-wide Mendelian randomization

Using univariable transcriptome-wide Mendelian randomization (TWMR) (*173*), the causal effect of differential *RHD* and *RSRP1* expression on reticulocyte count, platelet count, and HbA1c was estimated based on independent ($r^2 < 0.01$) genetic variants. Expression quantitative trait loci (eQTLs) were obtained from the eQTLGen consortium and included *cis*-eQTLs (false discovery rate < 0.05, two-cohort filter) for ~16,900 transcripts (*154*). GWAS effect sizes originate from the Neale UKBB summary statistics. Exposure and outcome datasets were harmonized, standardized effect size estimates were obtained by dividing z-scores by the square root of the sample size, and palindromic variants and variants with allele frequency difference > 5% between the two datasets were removed. Robustness of estimates was ascertained by excluding rs55794721, which had an extreme effect on both exposures and outcomes.

Association between Rhesus blood group and hematological traits

Impact of Rh$^-$ blood group on platelet count, reticulocyte count, and HbA1c was assessed in the CHUV maternity cohort through multivariate linear regression that incorporates the covariates: age at measurement, gestational week at measurement, whether the woman was pregnant at measurement (57.5% for reticulocyte count, 35.6% for platelet count, 23.4% for HbA1c), and whether the women had a child prior to the measurement (78.9% for reticulocyte count, 72.7% for platelet count, 96.7% for HbA1c). For women with multiple measurements, one was randomly selected, giving preference to measurements taken outside of pregnancy and excluding measurements taken during a pregnancy that resulted in stillbirth or multiple births. For measurements taken during

4: **Diseases:**

- **GGT-altering diseases:** *heart failure* (I50), *malignant neoplasm of liver and intra-hepatic bile ducts* (C22), *gallbladder* (C23), *other unspecified parts of biliary tract* (C24), and *diseases of the liver* (K70-K77) and the *gallbladder, biliary tract, or pancreas* (K80-K87).
- **Rotor syndrome (MIM: 237450):** classified with Dubin-Johnson syndrome (MIM: 237500) under *other disorders of bilirubin metabolism* (E80.6).
- **Charcot-Marie-Tooth type 1A (MIM: 118220):** classified as *hereditary motor or sensory neuropathy* (G60.0), a diagnosis encompassing all forms of Charcot-Marie-Tooth and related neuropathies.

5: **Socio-economic and lifestyle:**

- **Townsend deprivation index:** #189.
- **Household income:** #738.
- **Educational attainment:** #845.
- **Age at recruitment:** #21022.
- **Parental age at death:** *Mother's* (#3526) and *father's* (#1807) *age at death* meta-analyzed as parental lifespan.
- **Leukocyte telomere length:** #22191.

pregnancy, gestational week at measurement was calculated from date and gestational age at delivery. When gestational age at delivery was missing (52.9%), mean gestational age at delivery of the cohort (39.13 weeks) was used. For measurements outside of pregnancy, gestational week at measurement was coded as 0. When age at measurement was missing (12.1% for reticulocyte count, 19.1% for platelet count, 22.2% for HbA1c), data was imputed with multivariate imputation by chain equations including covariates. Ten complete imputed sets were analyzed and estimates were combined with `pool()`(R package `MICE` v3.13.0 (*311*)). One-sided p-value were calculated as $p_{new} = \frac{p_{old}}{2}$ in case of directional agreement with the effect observed in the UKBB.

CNV burden analysis in the UK Biobank

CNV burden calculation
An individual's CNV burden was defined as the number of Mb or genes affected by high-confidence autosomal CNVs ($|QS| \geq 0.5$). For the latter, we retained CNVs overlapping exons, splice sites, non-coding RNA, 3'UTR, and 5'UTR (CNV region definition and annotation) to assess number of disrupted genes. Duplication and deletion burdens were calculated similarly, and correlation between the six metrics was assessed with Pearson's coefficient of correlation. We used a two-sided unpaired Wilcoxon rank-sum test to assess differences in CNV burden between males and females.

CNV burden analysis
Linear regressions were performed between burden metrics and the same 57 normalized, covariate-corrected traits investigated by GWAS. For sex-specific traits, samples from the opposite sex were excluded. We set the significance threshold at $p \leq 0.05/63 = 7.9 \times 10^{-4}$ to account for six additional life history traits. Linear regressions were computed between non-normalized, covariate-corrected *mother's* and *father's age at death* and the burden to get effects on the years/[Mb or gene] scale. We meta-analyzed results with GWAMA to assess the impact on parental lifespan.

Burden analysis correction for modifier CNVRs
To assess the impact of the CNV burden on a trait, we collected CNVRs associating with that trait under the mirror model into a $sample \times CNVR$ matrix G. G Takes a value of -1 or 1 if the sample carries a CNVR-overlapping (≥ 1 bp) deletion or duplication, respectively, and 0 otherwise. G was regressed out of the trait and burdens were adjusted by subtracting the number of Mb or genes affected by CNVR-overlapping CNVs before performing associations anew. For the duplication and deletion burdens, CNVRs found through the duplication-only and deletion-only models, respectively, were considered and CNVR-overlapping deletions and duplications, respectively, were set to 0 in G.

Fraction of inherited CNVs
Rate of CNV inheritance was estimated by examining the fraction of shared CNVs among siblings pairs defined by kinship coefficient 0.2–0.3 and fraction of SNPs with identity by state at $0 \geq 0.0012$ (*61*). We retained 16,179 pairs with one individual among samples selected for the main CNV-GWASs. Shared CNVs were defined as high-confidence duplications ($QS \geq 0.5$) or deletions ($QS \leq -0.5$) on the same chromosome with

≥ 25 kb overlap. For each pair, we calculated the fraction of CNVs the individual in the main analysis shared with their sibling (number of shared CNVs/total number of CNVs in that individual) and averaged the results over all pairs to obtain the mean fraction of shared CNVs. As a control, the analysis was repeated by pairing the 16,179 individuals from the main analysis with random individuals sampled without replacement from the main pool of individuals.

Results

The CNV landscape of the UK Biobank

We used PennCNV (*203*) to call autosomal, pseudoautosomal, and chromosome X CNVs in 332,935 unrelated white British UKBB participants with no reported blood malignancy. Calls were processed by a pipeline that excluded 1,413 CNV outlier samples and attributed a probabilistic QS to each CNV (*206*). Out of 1,329,290 identified CNVs, 176,870 high-confidence CNVs with $|QS| \geq 0.5$ were retained for follow-up analyses (Figure 2.3A). As the fraction of homozygous CNVs (CN = 0 or 4) was negligible (1.1%; Figure 2.3B), we define deletions and duplications as having a CN smaller or larger than two, respectively, for the remainder of this study. Duplication length varied between 366 bp and the upper boundary, set at 10 Mb (17–3,968 probes), with a median of 297 kb (133 probes), and deletion length between 217 bp and 10 Mb (8–4,017 probes), with a median of 137 kb (60 probes) (Figure 2.3C-D). Overall, 129,263 (39%) participants carried at least one high-confidence CNV and 34,804 (10%) carried more than one (Figure 2.3E). In samples with ≥ 1 CNV, the total length of affected bases ranged between 217 bp and 14.2 Mb, with a median of 292 kb (Figure 2.3F). Analyzing the global CNV burden of the cohort, 70% was caused by duplications, which were both more numerous (54%) and 213 kb longer, on average, than deletions (Figure 2.3B–D). No differences in CNV burden, measured as the number of Mb or genes affected by CNVs, was detected across sexes (two-sided, unpaired Wilcoxon rank-sum test: $p_{Mb} = 0.793$; $p_{genes} = 0.748$). This contrasts with the excess of deleterious CNVs reported in females with neuro-psychiatric/developmental disorders (*312*–*315*), suggesting that this observation is trait dependent.

To bypass issues related to inter-individual variability in recurrent CNV breakpoints, we transformed CNV calls to the probe level for frequency calculation (*295*). A large fraction of the genome was subjected to CNVs as 662,247 probes (82%) were found in a CN-altered state in at least one participant, even if 81% of these had a CNV frequency ≤ 0.005% (n ≤ 16). The fraction of never-deleted probes (43%) was 1.73 higher than the fraction of never-duplicated probes (26%), and with some notable exceptions, deletion frequencies tended to be lower than duplication frequencies (Figure 2.4). For most loci with high CNV frequency, duplication and deletion frequencies did not mirror each other (Figure 2.4)). Overall, these results are in line with the common paradigm that CNVs are individually rare but collectively common (*34*, *35*, *292*).

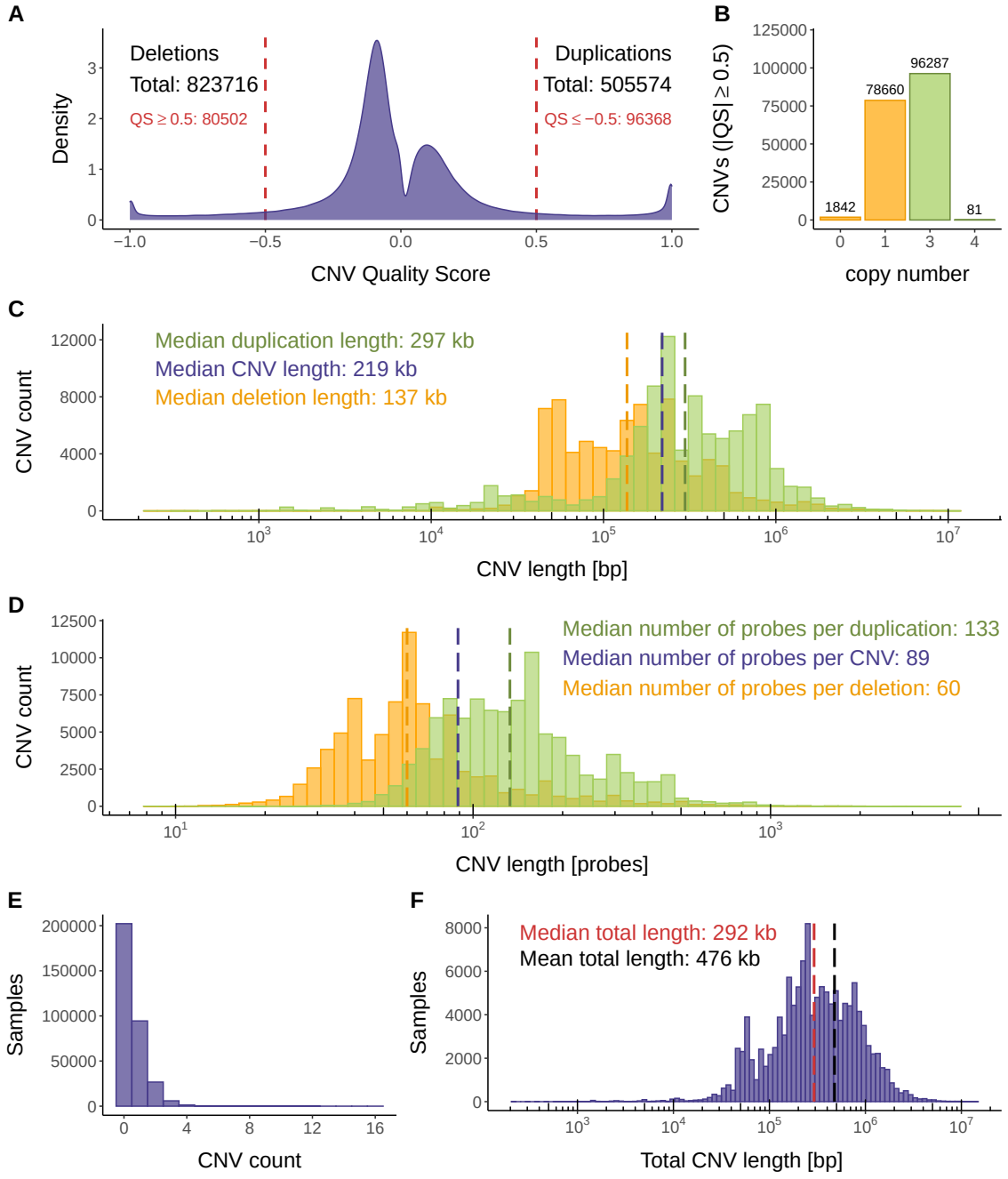


Figure 2.3: Distribution of high confidence CNVs in UKBB.
(**A**) Density plot of QS for the 1,329,290 called CNVs. High confidence duplications (QS ≥ 0.5) and deletions (QS ≤ -0.5), indicated per the red dashed lines, were retained for downstream analyses. (**B**) Distribution of the CN of high confidence CNVs. Deletions (CN = 0 or 1) are in orange, duplications (CN = 3 or 4) are in green. Number of CNVs in each category is indicated on top of the bars. Distribution of high confidence duplications (green) and deletions (orange) length in base pairs (**C**) and number of probes (**D**) on a logarithmic scale. Dashed lines show the median duplication (green), CNV (purple), and deletion (orange) length. (**E**) Distribution of high confidence CNV counts per individual. (**F**) Distribution of the total amount of bases affected by high confidence CNVs per individual on a logarithmic scale; red and black dashed lines show the median and mean number of bases affected by CNVs among individuals with ≥ 1 CNV, respectively.

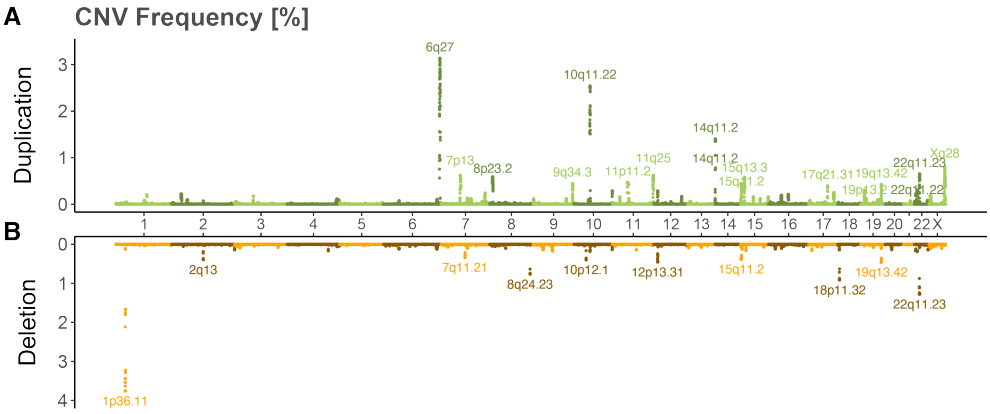


Figure 2.4: CNV frequency landscape in the UK Biobank.
Miami plot of high-confidence probe-level duplication (**A**) and deletion (**B**) frequencies (%) in the UKBB. Consecutive probes with identical duplication and deletion frequencies were clustered so that each dot represents one probe group. Loci with duplication frequency ≥ 0.3% or deletion frequency ≥ 0.2% are labeled according to the affected cytogenic bands.

The pleiotropic impact of recurrent CNVs

To assess the phenotypic impact of the UKBB CNV landscape, we selected 57 medically relevant phenotypes – including anthropometric traits, cardio-pulmonary assessments, hematological measurements, blood biomarkers, neuronal functions, and sex-specific attributes – with presumed high heritability (Table S2.1; Figure 2.5). GWASs were performed between the CN of pruned ($r^2 > 0.9999$) CNV-proxy probes with a CNV, duplication, and deletion frequency ≥ 0.005% and aforementioned traits according to a mirror (28,257 probes; Figure 2.6A), duplication-only (14,070 probes; Figure 2.6B), and deletion-only (9,936 probes; Figure 2.6C) association model, respectively. As the number of statistical tests is much lower than for classical SNP-GWASs and retained probes remain highly correlated due to the recurrent nature and large size of assessed CNVs, we calculated the number of effective (i.e., independent) tests, setting the GW threshold for significance at $p \leq 0.05/11{,}804 = 4.2 \times 10^{-6}$. Stepwise conditional analysis narrowed signals down to 86, 50, and 68 GW-significant associations for the mirror, duplication-only, and deletion-only models, respectively, of which 45, 22, and 32 reached the conventional SNP-GWAS threshold of $p \leq 5 \times 10^{-8}$. These signals were combined into 131 independent associations spanning 47 phenotypes (Figure 2.6D; Table S2.2; 62 signals across 32 phenotypes at $p \leq 5 \times 10^{-8}$). Following previous works (*292*, *295*, *306*), we omitted to account for the number of assessed traits, but even with a stringent experiment-wide threshold for significance ($p \leq 0.05/11{,}804 \times 57 = 7.4 \times 10^{-8}$), 68 out of 131 (52%) CNV-GWAS signals remained significantly associated. All summary statistics are made available on the GWAS Catalog (GCST90027274–GCST90027444).

Among signals identified through the mirror model, 63 (73%) replicated with either type-specific model, often reflecting the most common CNV type (Figure 2.6D, top). Five (6%) signals replicated with both type-specific models, providing examples of *true mirror effects* (i.e., opposite impact of duplications and deletions), such as the association between height and the CN of a Xp22.33 pseudoautosomal CNVR (chrX:285,850–1,720,422; $\beta_{mirror} = 2.33$ cm; $p = 7.2 \times 10^{-36}$; Figure 2.6E) encompassing the short-stature homeobox gene *SHOX* (MIM: 312865).

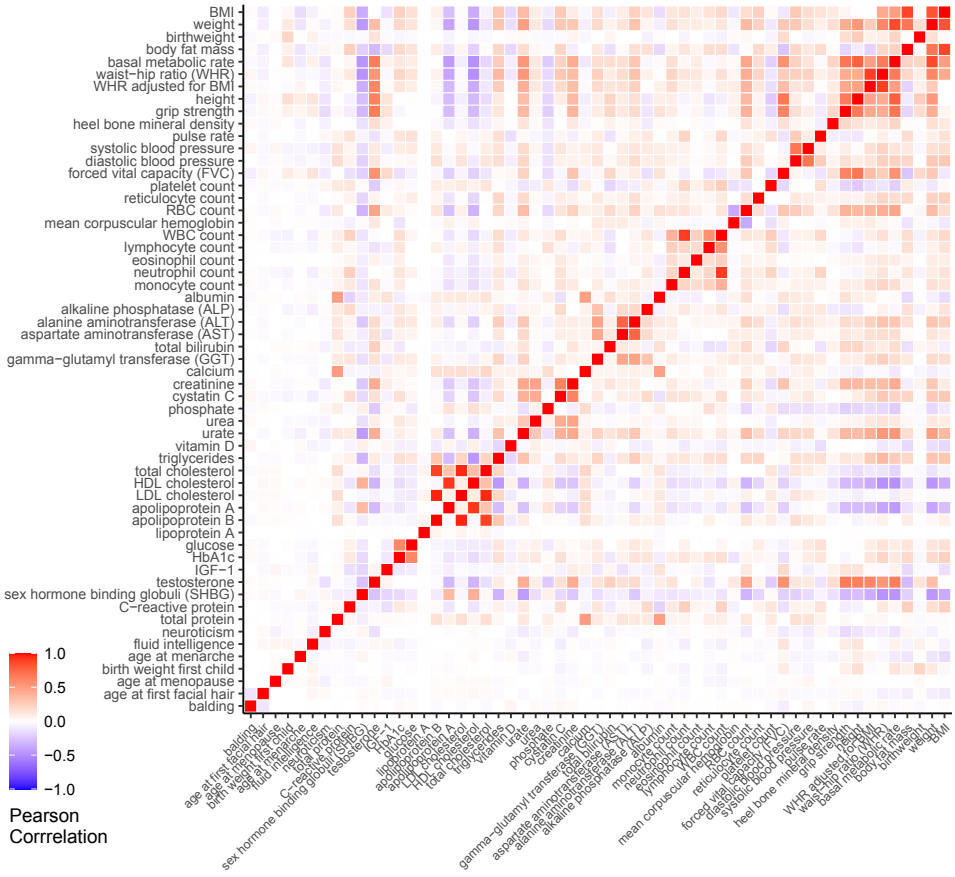


Figure 2.5: UKBB trait correlation.
Pearson correlation across the 57 continuous traits assessed by CNV-GWAS.

This association aligns with the short stature observed in individuals suffering from Turner syndrome (i.e., females with partial or complete loss of one chromosome X) and *SHOX* deficiency disorders (Leri-Weill dyschondrosteosis [MIM: 127300]; Langer mesomelic dysplasia [MIM: 249700]; idiopathic short stature [X-linked] [MIM: 300582]) (*316*, *317*). Less established is the impact of increased CN of *SHOX* and/or its regulatory regions (*318*), which we found to be associated with tall stature. CN and deletion of overlapping CNVRs further associated with WHR adjusted for BMI (chrX:514,930–618,611; β_{mirror} = 0.12 SD; p = 2.3 × 10^{-6}) and hand grip strength (chrX:762,346–2,219,659; β_{DEL} = -4.73 kg; p = 3.7 × 10^{-7}), respectively. While skeletal muscle hypertrophy has been reported in patients with Leri-Weill dyschondrosteosis (*319*), we hypothesize that the reduced grip strength in deletion carriers might result from the Madelung deformity characterizing the disorder, which is known to cause wrist pain and decreased grip strength (*320*), and/or the correlation between grip strength and height (Figure 2.5). Unlike mirror effects, partially overlapping signals between decreased forced vital capacity or grip strength and the 22q11.21 low copy repeat (LCR) A-B (chr22:19,024,651–20,311,646; deletion-only) and 22q11.21 LCR A-D (chr22:19,024,651–21,407,523; mirror and duplication-only) hinted at U-shaped effects (i.e., deletion and duplication shift the phenotype in the same direction) (MIM: 188400 and 192430), demonstrating the existence of different mechanisms of gene dosage (Figure 2.6D).

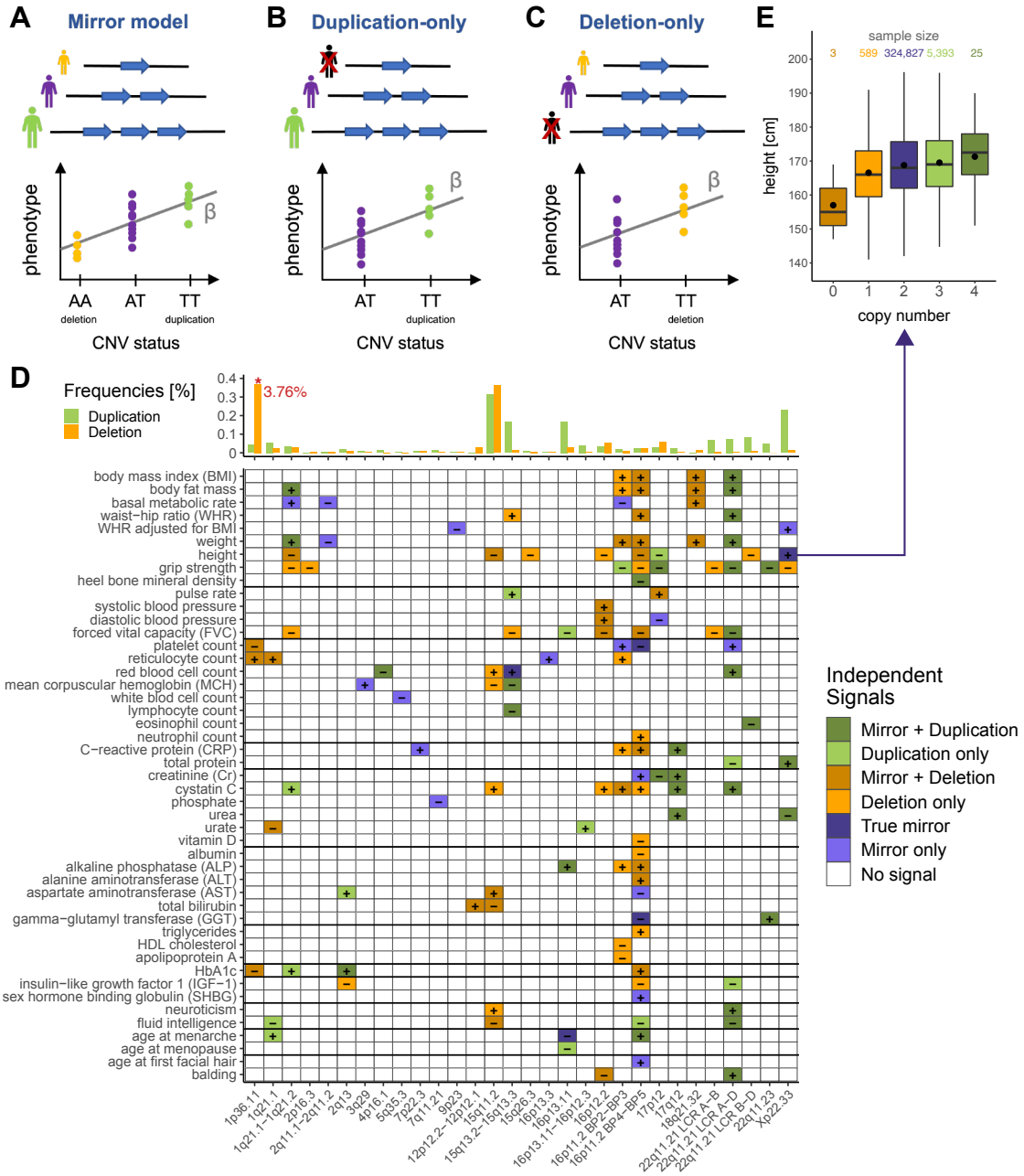


Figure 2.6: CNV-GWAS roadmap of the UK Biobank.
CNV-GWAS association models with PLINK encoding: the mirror model assumes equal-sized but opposite-direction effect of deletion and duplication and estimates the impact of each additional copy (**A**); the duplication-only model disregards deletion carriers and estimates the effect of duplications (**B**); the deletion-only model disregards duplication carriers and estimates the effect of deletions (**C**). (**D**) Independent genome-wide significant associations between CNV regions (x-axis; as cytogenic bands) and traits (y-axis). Color tiles represent the model(s) through which the association was detected (true mirror: mirror, duplication-only, and deletion-only) and signs show directionality, so that the duplication (greens), deletion (oranges), or copy number (purples) of a CNV region associated with a phenotypic increase (+) or decrease (−). 16p11.2 (BP2-BP3 and BP4-BP5) and 22q11.21 (LCR B at chr22:20,400,000) recurrent CNVs are assessed separately. For each CNVR, average duplication and deletion frequencies (%) of the lead probe (according to the most significant model) are depicted at the top. Deletion frequency of 1p36.11 was truncated from 3.76%. (**E**) Boxplot of height in individuals with CNVs overlapping the Xp22.33 pseudoautosomal region (chrX:285,850–1,720,422). Sample size is reported for each copy-number category at the top; dots show the mean; outliers are not shown.

Most signals involved large recurrent CNVRs (mean = 901 kb; median = 612 kb) and we confirm multiple well-established associations, such as the negative impact of the 1q21.1–1q21.2 deletion (MIM: 612474) on height (*321*–*323*) (chr1:146,478,785–147,832,715; β_{DEL} = -6.67 cm; p = 2.5

$\times 10^{-21}$), the negative correlation between BMI and the CN of 16p11.2 BP4-BP5 (MIM: 611913 and 614671) (chr16:29,596,230–30,208,637; β_{DEL} = 6.11 kg/m^2; p = 3.6×10^{-29}) (*324*–*326*) and 16p11.2 BP2-BP3 (MIM: 613444) (chr16:28,818,541–29,043,450; β_{DEL} = 4.25 kg/m^2; p = 5.3×10^{-8}) (*279*, *325*, *327*), or the more recently reported positive association between 16p11.2 BP4-BP5's CN and age at menarche (chr16:29,596,230–30,208,637; β_{mirror} = 1.16 years; p = 1.2×10^{-10}) (*328*). In addition, our results revealed the broad pleiotropic impact of these loci: 26, 16, and 12 traits associated with the 16p11.2 BP4-BP5, 22q11.21, or 16p11.2 BP2-BP3 regions, respectively. Some of these previously poorly described associations might help shed light on the molecular mechanisms linking involved loci to phenotypes, as exemplified by the association between the 16p11.2 BP4-BP5 deletion (chr16:29,596,230–30,208,637) and reduced levels of insulin-like growth factor 1 (IGF-1; β_{DEL} = -3.26 nmol/L; p = 2.9×10^{-7}). In children, diseases characterized by low levels of IGF-1 (e.g., IGF-1 deficiency [MIM: 608747], Laron syndrome [MIM: 262500], or growth hormone [GH] deficiencies [MIM: 262400, 612781, 173100, 307200, 618157, and 615925]) typically result in short stature (proxied by height), while symptoms of adult GH deficiency include increased adipose mass (proxied by BMI, body fat mass, weight, and WHR), decreased muscle mass and strength (proxied by grip strength), altered lipid profile (proxied by triglycerides), and insulin resistance (proxied by HbA1c) (*329*), all of which are affected in a directionally concordant fashion by the 16p11.2 BP4-BP5 deletion. Conversely, some regions only associated with a single trait, e.g., the CN of a 3q29 region (chr3:195,725,157–196,035,229) associated with increased mean corpuscular hemoglobin (β_{mirror} = 1.92 pg; p = 1.1×10^{-9}), whose levels indirectly reflect iron load in erythrocytes (*330*). The CNVR harbors the transferrin receptor gene, *TFRC* (MIM: 190010), which is involved in cellular iron uptake and was shown to associate with mean corpuscular hemoglobin through SNP-GWAS (*331*). Together, these results emphasize the potent role of CNVs as phenotypic modifiers.

Replication in the Estonian Biobank

We next assessed our ability to detect CNVs and sought to replicate identified signals in an independent cohort, the EstBB (*62*). Taking advantage of 966 unrelated samples with both microarray-based (PennCNV) and WGS-based (STRiP) CNV calls, we calculated the correlation between the CNV profiles obtained with these two methods for 709,358 quality-controlled, autosomal probes (Figure 2.7A). Due to small sample size, most probes (630,819 probes; 89%) were monomorphic. Among the 20,963 probes detected in a CNV state in at least one sample by both methods, 71% (14,976 probes; 2.1% of all probes) showed high ($r \geq 0.75$) agreement in calling profiles. We detected 39,847 (5.6%) apparent false positives (i.e., probes only detected in a CNV state by PennCNV). Forty percent of these were in linkage disequilibrium (± 250 kb and $r \geq 0.5$) with probes showing high microarray-WGS concordance (Figure 2.7B), suggesting that they are true positives mislabeled as false positives due to fragmentation of STRiP CNV calls. We also observed 17,717 (2.5%) false negatives (i.e., probes only detected in a CNV state by STRiP). Size distribution – both in number of base pairs (Figure 2.7C) and probes (Figure 2.7D) – of consecutive stretches of false negative probes was smaller than for the other assessed categories, confirming the poor ability

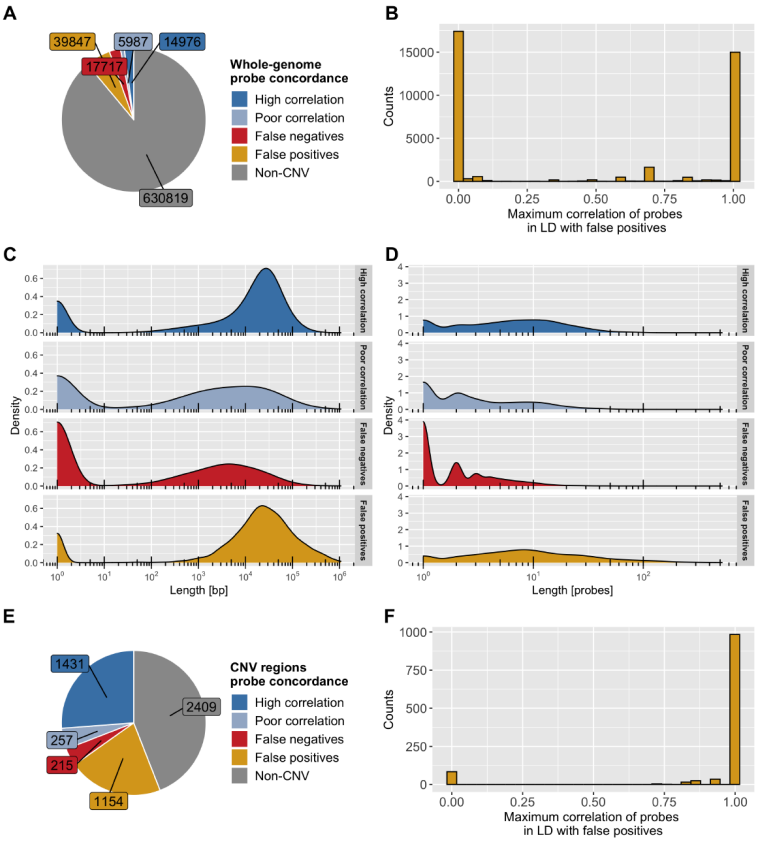
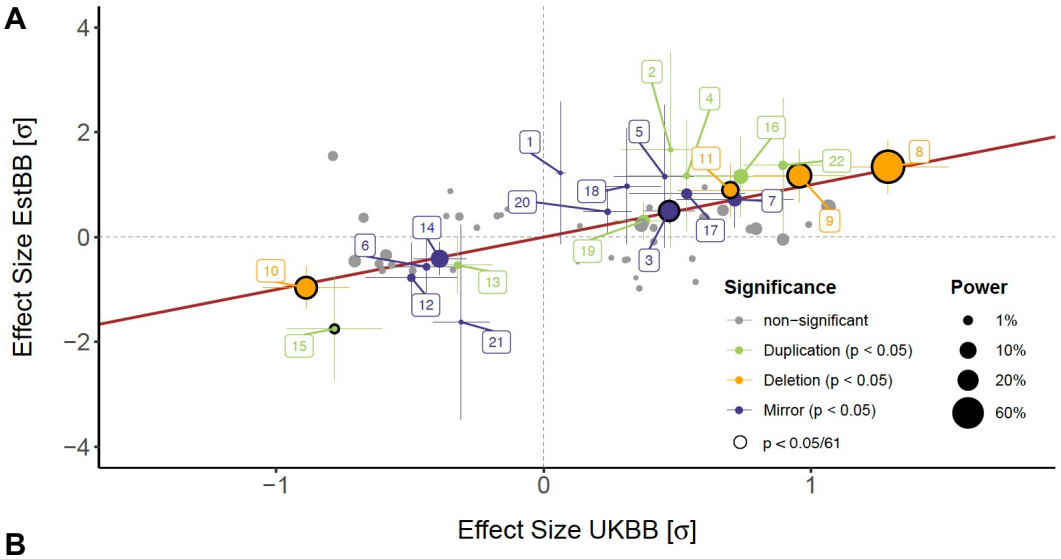


Figure 2.7: Comparative analysis of microarray vs WGS CNV call quality. (**A**) Concordance between PennCNV (microarray) and STRiP (whole-genome sequencing) CNV calls for 709,358 quality-controlled, autosomal probes from the OmniExpress-24 genotyping array in EstBB. Probes with high ($\rho \geq 0.75$) and low ($\rho < 0.75$) correlation between PennCNV and STRiP CNV profiles are in dark and light blue, respectively; false negatives (i.e. probes only detected in a CNV state by STRiP) are in red; false positives (i.e. probes only detected in a CNV state by PennCNV) are in gold; monomorphic probes are in gray. (**B**) Distribution of the maximal PennCNV-STRiP correlation among probes in linkage disequilibrium (± 250 kb and r ≥ 0.5) with false positive probes from (A). Size distribution in base pairs (**C**) and probes (**D**) of consecutive stretches of probes mapping to non-monomorphic categories in (A). (**E**) EstBB concordance between PennCNV and STRiP CNV calls at 5,566 probes overlapping UK Biobank trait-associated CNV regions; identical color scheme to (A). (**F**) Distribution of the maximal PennCNV-STRiP correlation among probes in linkage disequilibrium (± 250 kb and r ≥ 0.5) with false positive probes from (E).

to detect small CNVs with microarray data (*205*). If false negatives hinder discovery, they do not affect the validity of detected associations. We next repeated the analysis on 5,566 probes overlapping UKBB trait-associated CNVRs (Figure 2.7E) and observed i) an increased fraction of highly correlated probes (1,431 probes; 71% vs 85%), ii) an increased fraction of apparently mislabeled false positives in linkage disequilibrium with highly correlated probes (1,061 probes; 40% vs 92%; Figure 2.7F), and iii) a decreased proportion of false negatives among non-monomorphic probes (215 probes; 23% vs 7%), indicating good sensitivity and specificity to detect CNVs at trait-associated genomic loci.

To replicate association signals, microarray-based CNV data were available for 89,516 unrelated individuals. Phenotypic measurements originating from national health registries were only available for a limited subset of participants, ranging from ~60,000 for anthropometric measurements, to < 1,000 for specialized biomarkers (Table S2.1). Restricting ourselves to autosomal signals with sample size ≥ 2,000 and ≥ 1 CNV carrier, data were available for 61 (47%) CNVR-trait pairs (Table S2.2; Figure 2.8A). Six signals replicated with Bonferroni correction for multiple testing ($p \leq 0.05/61 = 8.2 \times 10^{-4}$; Figure 2.8B) and we observed 7.2-times more nominally significant signals than expected by chance (22 signals; two-sided binomial test: $p = 7.8 \times 10^{-14}$; Figure 2.9A). Effect size estimates followed closely the ones detected in the UKBB (Figure 2.8). Given the low sample sizes, we conducted simulations to assess the power of the replication study. Assuming effect sizes matching those observed in the UKBB, the average replication power was estimated at 5.5% ($\alpha = 0.05/61$; Figure 2.9B). This corresponds to an expected number of replicated signals of

3.4, slightly below the six observed, and argues in favor of the robustness of the original UKBB CNV-GWAS findings.



B

	Cytogenic Band (CNV region)	Trait	Effect	UK Biobank β	UK Biobank *p*	UK Biobank Del – Neutral –Dup	Estonian Biobank β	Estonian Biobank *p*	Estonian Biobank Del – Neutral –Dup
3	**1q21.1-1q21.2** (1:146'478'785-147'832'715)	height	cm/CN	4.3	7.8×10^{-22}	93 – 330'619 – 125	4.6	3.3×10^{-5}	11 – 67'844 – 78
8	**16p11.2 BP4-BP5** (16:29'596'230-30'208'637)	BMI	kg/m²/del	6.1	3.6×10^{-29}	75 – 330'321 – 101	7.0	3.4×10^{-7}	12 – 67'877 – 21
9		WHR	unit/del	0.1	1.5×10^{-27}	75 – 330'750 – 101	0.1	7.6×10^{-6}	9 – 59'381 – 16
10		height	cm/del	-8.2	3.3×10^{-27}	76 – 330'660 – 101	-9.0	2.5×10^{-6}	12 – 67'900 – 21
11		weight	kg/del	11.1	5.0×10^{-12}	75 – 330'431 – 101	15.2	5.2×10^{-4}	12 – 67'891 – 21
15	**17p12** (17:14'098'277-15'468'444)	Cr	µmol/L/dup	-13.8	6.5×10^{-18}	196 – 316'488 – 102	-94.1	5.2×10^{-4}	4 – 16'247 – 5

Figure 2.8: Replication of CNV-GWAS signals in EstBB.
(**A**) Estonian (EstBB; y-axis) vs UK (UKBB; x-axis) Biobank standardized effect sizes. The identity line is in red; size reflects power at $\alpha = 0.05/61$; non-significant signals are in gray; nominally significant signals with 95% confidence intervals are colored according to replication models; multiple-testing correction surviving signals ($p \leq 8.2 \times 10^{-4}$) are circled in black and listed in (B) with the first column's color corresponding to the association model and numbers matching labels in (A). (**B**) Effect sizes (β; unit in the effect column) and p-values (p) for the UKBB and EstBB GWAS, along with the number of individuals with available phenotypic data carrying a deletion, no CNV, or a duplication overlapping the CNVR. Labels: (1) Platelet count - 1p36.11; (2) HbA1c—1q21.1 - 1q21.2; (3) Height—1q21.1 - 1q21.2; (4) Age at menarche - 1q21.1; (5) Platelet count - 16p11.2 BP2-BP3; (6) Weight - 16p11.2 BP2-BP3; (7) Age at menarche - 16p11.2 BP4-BP5; (8) BMI - 16p11.2 BP4-BP5; (9) WHR - 16p11.2 BP4-BP5; (10) Height - 16p11.2 BP4-BP5; (11) Weight - 16p11.2 BP4-BP5; (12) ALT - 16p11.2 BP4-BP5; (13) Age at menopause - 16p13.11; (14) Age at menarche - 16p13.11; (15) Creatinine - 17p12; (16) Creatinine - 17q12; (17) CRP - 17q12; (18) Platelet count - 22q11.21 LCR A-D; (19) BMI - 22q11.21 LCR A-D; (20) Weight - 22q11.21 LCR A-D; (21) Eosinophil count - 22q11.21 LCR B-D; (22) GGT - 22q11.23.

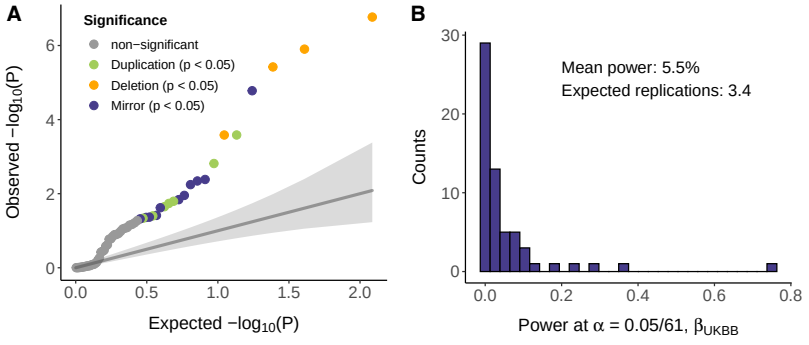


Figure 2.9: Replication power in EstBB.
(**A**) Expected vs observed negative logarithm of p-values for the 61 EstBB replicated CNV-trait pairs colored by significance and replication model. Shade represents the 95% confidence interval. (**B**) Distribution of replication power – assuming similar effect sizes to the ones observed in the UKBB – for signals in (B). Mean power is 5.5%, so that the number of multiple testing correction surviving signals is $5.5\% \times 61 = 3.4$.

CNVs as modifiers of complex traits

To assess whether CNV-GWAS signals mapped to regions previously identified by SNP-GWASs for the same trait, we annotated CNVRs with associations reported by the GWAS Catalog (*77*). From the 126 autosomal CNV associations considered, 48 (38%) harbored a SNP signal for the same trait (Table S2.2). A similar fraction (31%) of CNV-GWAS signals with $4.2 \times 10^{-6} \geq p \geq 5 \times 10^{-8}$ is supported by SNP-GWAS signal, backing the reliability of intermediate-significant associations. We further tested whether SNP-GWAS signal distribution was denser within trait-associated CNVRs, as compared to the rest of the genome. While this was the case for nine traits (two-sided binomial test: $p = 0.05/56 = 8.9 \times 10^{-4}$; Table S2.3), enrichment did not seem to depend on the type of trait, total number of SNP-GWAS signals (Figure 2.10), or length of trait-associated CNVRs (Figure 2.10, insert). Nevertheless, colocalization of SNP and CNV signals reinforces confidence that involved loci play a role in shaping associated traits, as illustrated with four examples. The first example relates to a 1.7 Mb 2q13 CNV (chr2:111,398,266–113,115,598). Deletion of the region associated with decreased IGF-1 (β_{DEL} = -5.67 nmol/L; $p = 6.3 \times 10^{-10}$), an important regulator of glucose and insulin metabolism (*332*), and duplication associated with increased HbA1c (β_{DUP} = 3.47 mmol/mol; $p = 1.4 \times 10^{-7}$). The interval encompassed an IGF-1-associated intronic *ACOXL* SNP (*305*) upstream of *BCL2L11* (MIM: 603827) and two HbA1c-associated SNPs (*305*, *333*) downstream of *BCL2L11*. These SNP signals were reported in 2021, indicating that with increased statistical power, signal colocalization will increase. Both traits have not been thoroughly assessed in carriers of the recurrent reciprocal 2q13 CNV, who present with neurodevelopmental/psychiatric disorders, dysmorphisms, congenital heart disorder, hypotonia, seizures, micro-/macrocephaly, and microphallus at variable penetrance and expressivity (*334*–*338*); the two latter features are reminiscent of growth defects potentially mediated by dysregulation of the GH/IGF-1/insulin axis. Multiple genes overlapping the CNVR play a role in cell cycle (*BUB1* [MIM: 602452], *ANAPC1* [MIM: 608473]), cell survival (*MERTK* [MIM: 604705]), and apoptosis regulation (*BCL2L11*), which is negatively regulated by IGF-1 (*339*)). Our data support the variable penetrance and expressivity of this CNV – not listed as a DECIPHER CNV syndrome – and prompts follow-up studies to confirm and refine understanding of the genetic mechanisms linking the locus to the associated phenotypes.

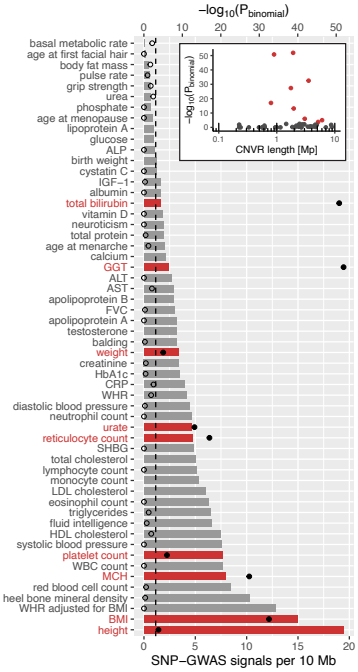


Figure 2.10: Genome-wide vs CNVR SNP-GWAS signal distribution. Number of SNP-GWAS signals per 10 Mb autosomal genome (bottom x-axis) for 56 assessed traits (y-axis). Top x-axis indicates the negative logarithm of the p-value (as a dot) for the binomial test assessing if SNP-GWAS signal distribution within CNVRs is higher than in the rest of the genome. Non-significant traits are in gray (empty dot); significant traits ($p \leq 0.05/56$; above the black dashed line) are in red (full dots). Inset shows the negative logarithm of binomial test p-values against the total length of trait-associated CNVRs in Mb on a logarithmic scale for each trait. Significant traits are in red and non-significant ones are in gray.

The second example links the 382 kb 1q21.1 deletion (MIM: 274000) to decreased serum urate levels (chr1:145,383,239–145,765,206; β_{DEL} = -48.32 mmol/L; $p = 5.8 \times 10^{-13}$; Figure 2.11). The rearranged interval encompasses 15 genes (Figure 2.13), including *PDZK1* (MIM: 603831), which encodes a urate transporter scaffold protein (*340*) and was associated with serum urate levels by SNP-GWAS (*341*–*344*). Recently, *in vitro* experiments elucidated the mechanism through which the urate-increasing "T" allele of rs1967017 leads to increased *PDZK1* expression (*345*), while the *PDZK1* protein-truncating variant rs191362962 was found to associate with decreased serum urate (*305*), both suggesting that decreased *PDZK1* expression – an expected outcome of *PDZK1* deletion – decreases serum urate levels. Dividing deletion carriers into groups harboring a full (start < 145.6 Mb) or a partial (start ≥ 145.6 Mb) deletion revealed that the small deletion, encompassing *PDZK1* and three other genes (Figure 2.13), was

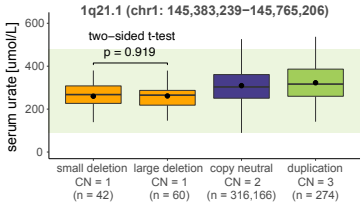


Figure 2.11: Urate & 1q21.1 CNVs. Boxplots of serum urate levels in individuals with a 1q21.1 overlapping small (start ≥ 145.6 Mb) or large (start < 145.6 Mb) deletion, copy-neutrality, or duplication. Copy number (CN) and sample size (n) are reported; dots show the mean; outliers are not shown; light green backgrounds show normal clinical range: 89–476 mmol/L. Two-sided t-test p-value compares urate levels of small and large 1q21.1 deletion carriers.

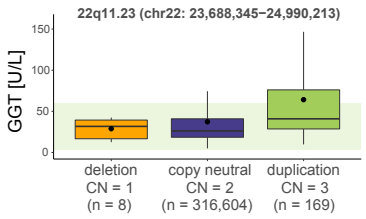


Figure 2.12: GGT & 22q11.23 CNVs. Boxplots of γ-glutamyltransferase (GGT) levels in individuals with a 22q11.23 overlapping deletion, copy-neutrality, or duplication. Copy number (CN) and sample size (n) are reported; dots show the mean; outliers are not shown; light green backgrounds show normal clinical range: 4–6 U/L.

sufficient to alter urate levels (two-sided t-test: p = 0.92; Figure 2.11).

The third example involves a 1.3 Mb long 22q11.23 duplication and increased GGT (chr22:23,688,345–24,990,213; β_{DUP} = 37.2 U/L; p = 9.3 × 10^{-32}; Figure 2.12) (MIM: 612365). The region harbors several independent GGT SNP-GWAS signals (*305*, *346*–*351*) and five genes involved in glutathione metabolism (KEGG pathway hsa00480), including *GGT1* (MIM: 612346) and *GGT5* (MIM: 137168) (Figure 2.14A), suggesting that an additional copy of these genes associates with increased levels of the encoded protein. As multiple factors can elevate GGT levels (*310*), we used binomial tests to verify that the 180 duplication carriers were not enriched for GGT-altering drug usage (p = 0.55), high alcohol consumption (p = 0.85), heart failure (p = 0.23), or cancer (p = 1) and other diseases (p = 0.64) of the liver, gallbladder, and bile ducts, as compared to control individuals. Visualization of GGT levels in individuals with two or three copies of the CNVR showed that the 22q11.23 duplication increased serum GGT independently of and additively to other GGT-increasing factors (Figure 2.14B-F).

Finally, we focused on the most frequent CNV in our cohort (frequency = 3.76%; Figure 2.4), the 50 kb 1p36.11 deletion (chr1:25,599,041–25,648,747), which encompasses *RHD* (Rhesus [Rh] blood group D antigen [MIM: 111680]) and *RSRP1* and associated with increased reticulocyte count (β_{DEL} = 2.7 × 10^9 cells/L; p = 7.8 × 10^{-14}), decreased platelet count (β_{DEL} = -3.7 × 10^9 cells/L; p = 1.4 × 10^{-12}), and decreased HbA1c (β_{DEL} = -0.3 mmol/mol; p = 9.3 × 10^{-8}; Figure 2.15A). Overlap with SNP-GWAS signals for various hematological traits (*352*, *353*) combined with subsequent replication of the reticulocyte count association based on WES-based CNV calls (*209*) prompted the investigation of the expression of these genes in whole blood. Tissue-specific transcriptomic data from the GTEx project v8 (*151*) revealed that *RHD*, a protein whose presence/absence on erythrocyte cell membranes is critical in determining an individual's Rh blood group (*256*), was almost exclusively expressed in whole blood (Figure 2.15B), whereas *RSRP1* was ubiquitously expressed, with lower expression in whole blood (Figure 2.15C). Selecting *RHD*'s (ENST00000328664) and *RSRP1*'s (ENST00000243189; Figure 2.15D) most highly expressed isoforms in whole blood, we mapped exons to the association plot, showing that *RSRP1*'s isoform does not overlap the CNVR, in contrast to RHD's, which is fully encompassed by it (Figure 2.15A). We next used transcriptome-wide Mendelian randomization (*173*) (TWMR; Table S2.4) to establish a directionally concordant causal link between *RHD* expression and reticulocyte count (α_{TWMR} = -0.013, p = 1.6 × 10^{-4}; Figure 2.16A), platelet count (α_{TWMR} = 0.031, p = 2.3 × 10^{-9}; Figure 2.16B), and HbA1c levels (α_{TWMR} = 0.017, p = 3.5 × 10^{-7}; Figure 2.16C). *RSRP1* TWMR resulted in directionally concordant and significant effects, but the gene had suboptimal number of instruments (three) for robust causal inference (Figure 2.16D-F). Furthermore, both genes' signals were driven by a strong upstream expression quantitative locus (rs55794721; Figure 2.16A-F). Strengthening the causal role of *RHD*'s CN, lack or strongly reduced expression of all Rh antigens, a rare condition named Rh deficiency or Rh_{null} syndrome [MIM: 617970 and 268150], is associated with increased erythrocyte osmotic fragility, resulting in hemolytic anemia (*354*). Hemolytic anemia is characterized by increased reticulocyte count (*355*) and can falsely lower HbA1c levels because of

decreased erythrocyte lifespan (*356*), putting forward the hypothesis that heterozygous deletion of *RHD* leads to subclinical phenotypes akin to hemolytic anemia. To gauge the generalizability of these results, we looked for similar trends in individuals with Rh$^-$ blood type, which can be caused by various polymorphisms (*256*). Because Rhesus groups were unavailable for the UKBB, we turned to a maternity cohort from the Lausanne University Hospital. Despite low samples sizes, concordant trends of increased reticulocyte count (β_{Rh^-} = 1.07 °/oo; $p_{one\text{-}sided}$ = 0.134; n = 741) and decreased platelet count (β_{Rh^-} = -2.8 × 10^{-9} cells/L; $p_{one\text{-}sided}$ = 0.126; n = 5,034) and HbA1c levels (β_{Rh^-} = -0.22%; $p_{one\text{-}sided}$ = 0.050; n = 418) were observed in Rh$^-$ women (Table S2.5). Of note, reticulocyte and platelet counts have been reported to increase and decrease, respectively, along pregnancy (*357*), and despite correcting for pregnancy status and gestational weeks, interaction between Rh$^-$ blood group and pregnancy cannot be excluded. Impact of Rh blood type on hematological traits awaits validation but these examples illustrate how studying CNVs at SNP-GWAS loci can pinpoint causal genes and shared genetic mechanisms.

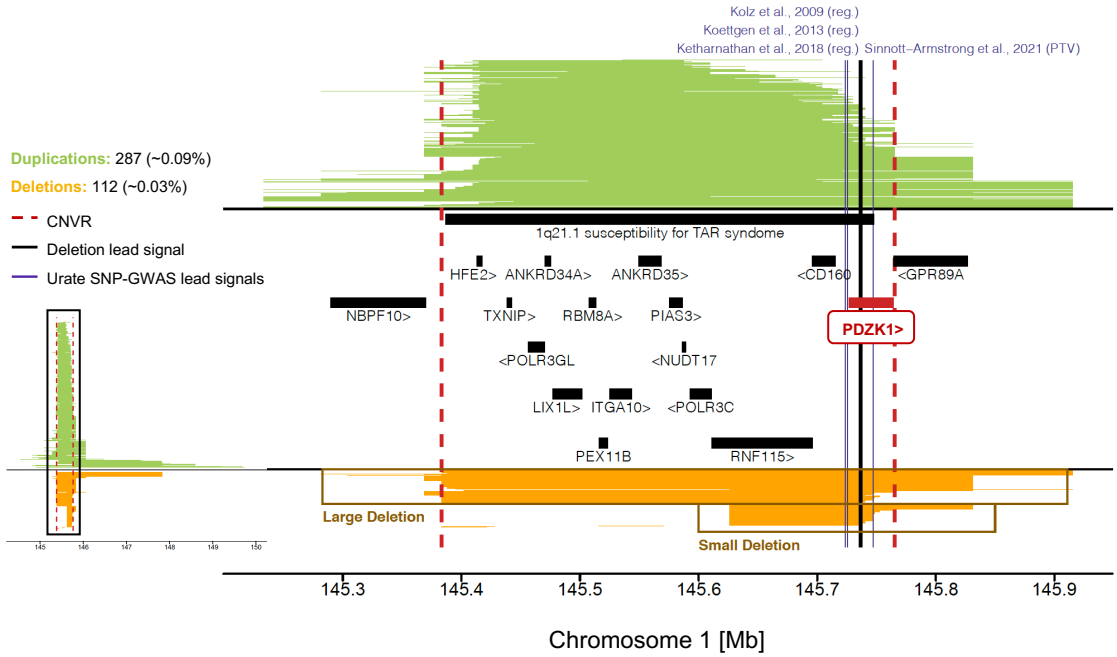


Figure 2.13: 1q21.1 deletion and decreased serum urate levels.
Mapping of CNVs overlapping the 1q21.1 (chr1:145,383,239-145,765,206) deletion region associated with decreased serum urate levels. Number and frequency of duplications and deletions are at the top left; left plot shows all overlapping CNVs; right plot focuses on the central CNV region represented by red dashed lines. Duplications are in green, deletions in orange; black line indicates the lead signal for serum urate (deletion-only); purple lines indicate serum urate-associated SNPs (*305*, *341*, *342*, *345*) (reg. = regulatory variant; PTV = protein-truncating variant). DECIPHER recurrent CNV and overlapping protein-coding genes are in black, except for *PDZK1* in red. Brown boxes separate large (start < 145.6 Mb) vs small (start ≥ 145.6 Mb) deletion carriers.

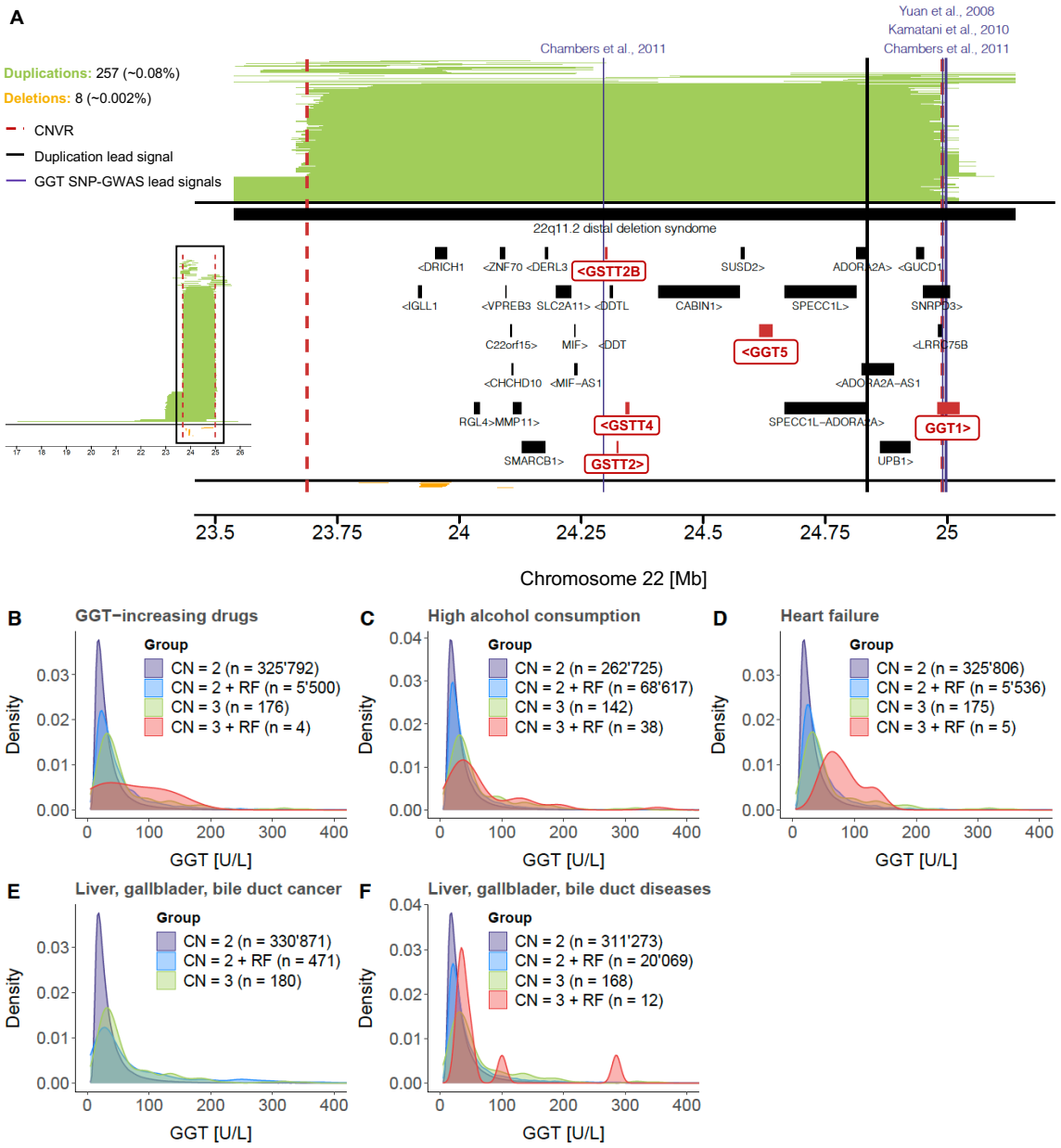


Figure 2.14: 22q11.23 duplication and increased serum γ-glutamyl transferase levels.
(**A**) Mapping of CNVs overlapping the 22q11.23 (chr22:23,688,345-24,990,213) duplication associated with increased γ-glutamyl transferase (GGT) levels. Number and frequency of duplications and deletions are at the top left; left plot shows all overlapping CNVs; right plot focuses on the central CNV region represented by red dashed lines. Duplications are in green, deletions in orange; black line indicates the lead signal for GGT (duplication-only); purple lines indicate GGT-associated SNPs (*346*–*348*). DECIPHER recurrent CNV (truncated) and overlapping protein-coding genes are in black, except for genes involved in glutathione metabolism in red. Density plots showing the distribution of GGT levels in copy-neutral (CN = 2) and 22q11.23 overlapping duplication carriers (CN = 3) with or without various risk factors (RF) for increased GGT: (**B**) GGT-increasing drugs, (**C**) high alcohol consumption, (**D**) heart failure, and (**E**) cancer or (**F**) other diseases of the liver, gallbladder, or bile ducts. Sample size (n) is indicated. Plots were truncated to 400 U/L (max.: 1167 U/L).

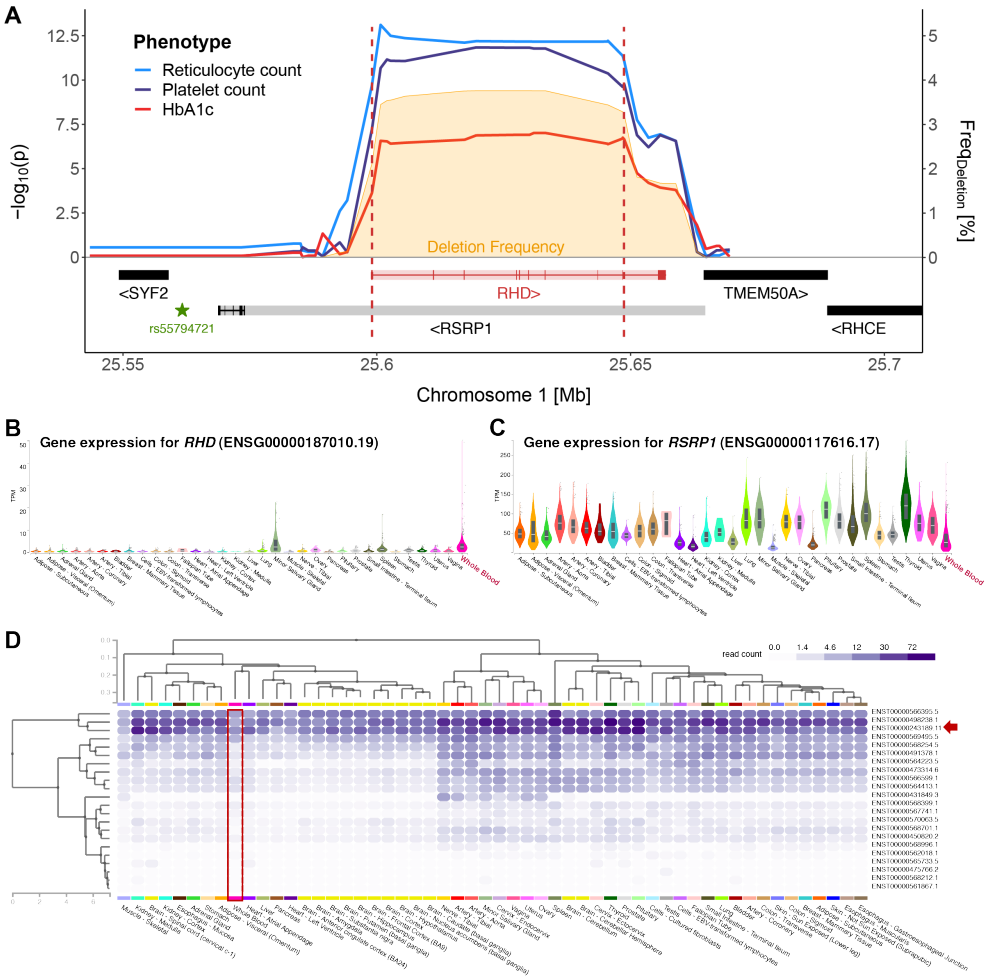


Figure 2.15: *RHD* deletions modulate hematological traits.
(**A**) Association plot for the 1p36.11 deletion (chr1:25,599,041–25,648,747). Red dashed lines delimit the deletion-only CNVR; left y-axis shows the negative logarithm of association p-value for reticulocyte count (blue), platelet count (purple), and glycated hemoglobin (HbA1c; red); right y-axis shows deletion frequency (%) (orange); encompassed genes are schematically represented at the bottom; retained exons for the most strongly expressed isoform in whole blood are shown for *RHD* (ENST00000328664) and *RSRP1* (ENST00000243189), and shaded color represents the full gene sequence; star indicates the *RHD* and *RSRP1* expression quantitative locus rs55794721. GTEx gene expression in 33 tissues for *RHD* (**B**) and *RSRP1* (**C**). Brain, cervix, esophagus, and skin are not shown for visibility. Whole blood is shown with a red label. (**D**) Isoform expression for *RSRP1* in 54 tissues, with whole blood circled in red. Red arrow points at the isoform with the highest expression in whole blood, ENST00000243189.

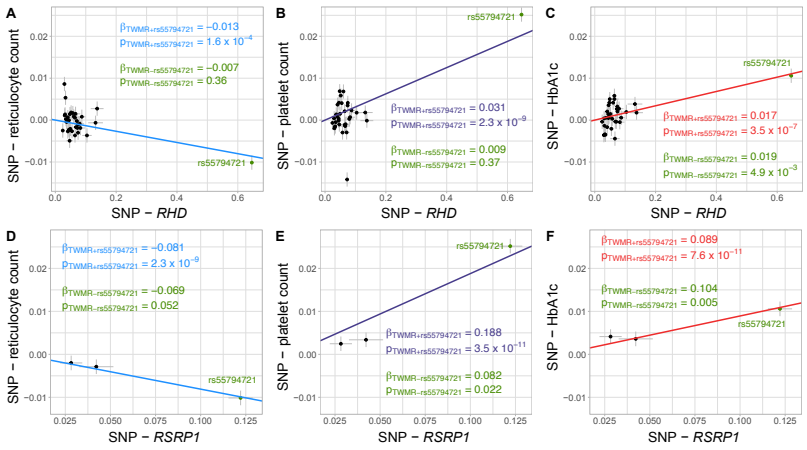


Figure 2.16: 1p36.11 TWMR.
SNP-exposure (x-axis) vs SNP-outcome (y-axis) plots for transcriptome-wide Mendelian randomization (TWMR) analyses estimating the causal effect of *RHD* on (**A**) reticulocyte count, (**B**) platelet count, and (**C**) glycated hemoglobin (HbA1c), and of *RSRP1* on (**D**) reticulocyte count, (**E**) platelet count, and (**F**) HbA1c. Gray horizontal and vertical lines represent standard errors of the SNP-exposure and SNP-outcome estimates, respectively. Causal effect estimates (β) and associated p-values are reported. The outlier SNP rs55794721 is in green. Causal effects estimated without rs55794721 are in green.

CNVs at Mendelian disorder loci

Despite the lower-than-average disease burden of UKBB participants (*59*), several associations comprised loci involved in Mendelian disorders. The heterozygous 395 kb 12p12.2-p12.1 deletion, which associated with a non-pathological increase in total bilirubin (chr12: 21,008,080–21,403,457; $\beta_{DEL} = 3.1\,\text{mmol/L}$, $p = 2.2 \times 10^{-13}$; Figure 2.17A) and harbors SNP-GWAS signals for bilirubin levels (*305*, *358*–*362*), overlaps the Rotor syndrome locus (MIM: 237450), an extremely rare disorder whose main clinical manifestation is hyperbilirubinemia. Rotor syndrome (*245*) is caused by the homozygous disruption of *SLCO1B1* (MIM: 604843) and *SLCO1B3* (MIM: 605495) (Figure 2.17B), which encode for the hepatic transporters OATP1B1 and OATP1B3, respectively, involved in the uptake of various drugs and metabolic compounds, including bilirubin (*363*). Concordantly, UKBB participants diagnosed with Rotor syndrome or the related and more common Dubin-Johnson syndrome (MIM: 237500) presented above-normal levels of total bilirubin (Figure 2.17A). Interestingly, individuals carrying a partial deletion that only affects *SLCO1B1* (start ≥ 21.1 Mb; Figure 2.17B) exhibited significantly milder increase in total bilirubin (two-sided t-test: $p = 3.1 \times 10^{-4}$; Figure 2.17A), illustrating how mutations pathogenic in a digenic recessive framework can contribute to subtle changes in disease-associated phenotypes when present in an isolated heterozygous state.

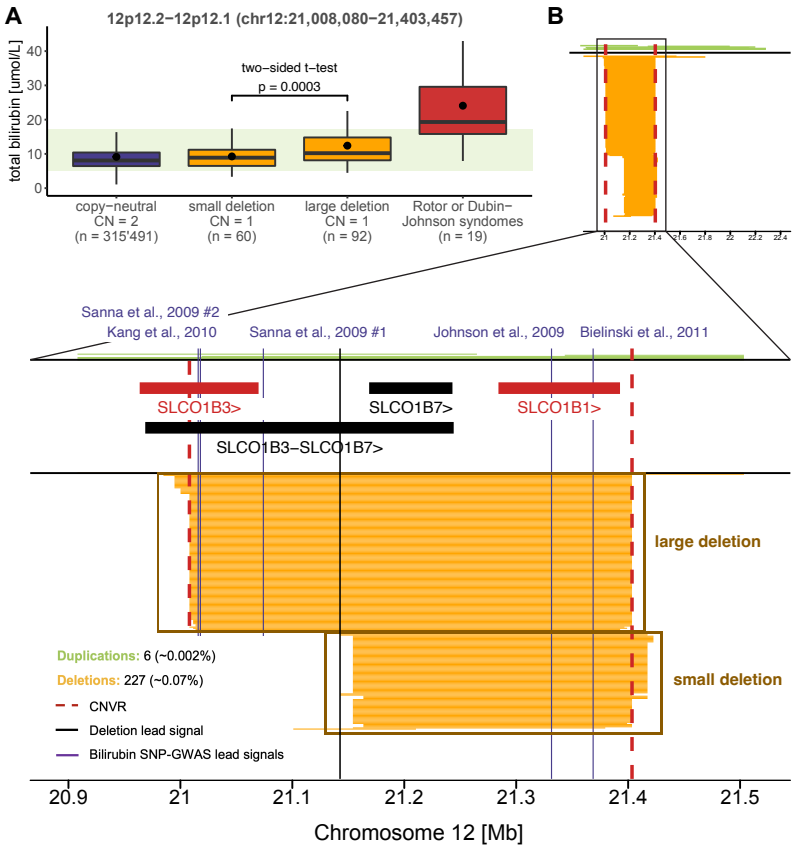


Figure 2.17: 12p12.2-p12.1 deletion and increased total bilirubin levels.
(**A**) Boxplots of total bilirubin levels in copy-neutral individuals, and small (start ≥ 21.1 Mb) or large (start < 21.1 Mb) 12p12.2-p12.1 deletion carriers, and Rotor or Dubin-Johnson syndrome-affected individuals (ICD-10 E80.6); dots show the mean; outliers are not shown; light green backgrounds show normal clinical range: 5–17 mmol/L. p-value of a two-sided t-test comparing total bilirubin levels of small and large deletion carriers is shown. (**B**) Mapping of CNVs overlapping the 12p12.2-12p12.1 CNVR (top) with zoom on the black box (bottom). Duplications are in green, and deletions are in orange. Red dashed lines represent the trait-associated CNVR; black line indicates the lead signal for total bilirubin (deletion-only); purple lines indicate serum bilirubin-associated SNPs (*358*–*361*); overlapping protein-coding genes are in black, except for Rotor syndrome-associated genes – *SLCO1B1* and *SLCO1B3* – in red. Brown boxes separate large from small deletion carriers.

A second example links the 1.5 Mb long 17q12 duplication (MIM: 614526) (chr17:34,797,651–36,249,489) and increased levels of kidney damage biomarkers, including cystatin C (β_{DUP} = 0.15 mg/L, p = 4.2×10^{-17}; Figure 2.18A), serum creatinine (SCr; β_{DUP} = 13.0 mmol/L, p = 2.7×10^{-16}; Figure 2.18B), and serum urea (β_{DUP} = 0.93 mmol/L, p = 9.1×10^{-10}; Figure 2.18C), as well as the inflammation biomarker C-reactive protein (CRP; β_{mirror} = 2.3 mg/L, p = 1.1×10^{-6}; Figure 2.18D). Deletion of this interval (Figure 2.18E), as well as point mutations in overlapping *HNF1B* (MIM: 189907), cause the highly pathogenic and penetrant autosomal dominant renal cysts and diabetes syndrome (RCAD [MIM: 137920 and 614527]). RCAD is characterized by heterogeneous structural and/or functional renal defects, neurodevelopmental/psychiatric disorders, and maturity-onset diabetes of the young (*364*). Because of the small number of deletion carriers (n = 6, regardless of phenotypic data availability), the deletion's effect was not assessed by CNV-GWASs, but elevated levels of cystatin C (Figure 2.18A), SCr (Figure 2.18B), and urea (Figure 2.18C) in these individuals align with RCAD's clinical description. Conversely, penetrance of the reciprocal duplication remains debated and only 20% of diagnosed patients report renal abnormalities (*365*). In line with a lower pathogenicity, we detected 16 times more duplication than deletion carriers. Still, these individuals showed strong alterations in kidney biomarkers, suggesting tight gene dosage control on *HNF1B*.

Third, we zoomed in on the 1.4 Mb long 17p12 duplication (Figure 2.19A) known as the main etiology of Charcot-Marie-Tooth type 1A (MIM: 118220), a peripheral demyelinating neuropathy characterized by progressive muscle wasting (*366*). Correspondingly, duplication carriers showed decreased hand grip strength (chr17:14,098,277–15,457,056; β_{DUP} = -9.8 kg, p = 4.1×10^{-39}; Figure 2.19B) and lower SCr (chr17:14,098,277–15,468,444; β_{DUP} = -13.8 mmol/L, p = 6.5×10^{-18}; Figure 2.19C; EstBB: β_{DUP} = -94.1 mmol/L, p = 5.2×10^{-4}; Figure 2.8), indicating decreased muscle mass (*367*). We next assessed the proportion of duplication carriers diagnosed with Charcot-Marie-Tooth or related hereditary motor and sensory neuropathies and detected 48 and 38 diagnoses among the 331,206 copy-neutral individuals and 107 duplication carriers, respectively. While there is a clear enrichment for neuropathy diagnoses among duplication carriers (Fisher's exact test: odds ratio = 3,668, p < 2.2×10^{-16}), only 36% of duplication carriers were clinically identified. To test whether these individuals presented with more extreme clinical manifestations, we compared grip strength and SCr levels in duplication carriers with or without a neuropathy diagnosis. The former group exhibited lower grip strength (one-sided t-test: p = 0.005; Figure 2.19B) but no difference was detected in SCr levels (one-sided t-test: p = 0.384; Figure 2.19C). Importantly, there was no age difference between diagnosed (mean = 55.5 years) and undiagnosed (mean = 56.2 years) duplication carriers (two-sided t-test: p = 0.650), indicating that results do not reflect biases regarding age of disease onset. These examples show that well-established pathogenic CNVs can modulate disease-associated phenotypes in the general population without necessarily causing clinically diagnosable disorders, supporting a model of variable expressivity (*234*–*236*, *368*).

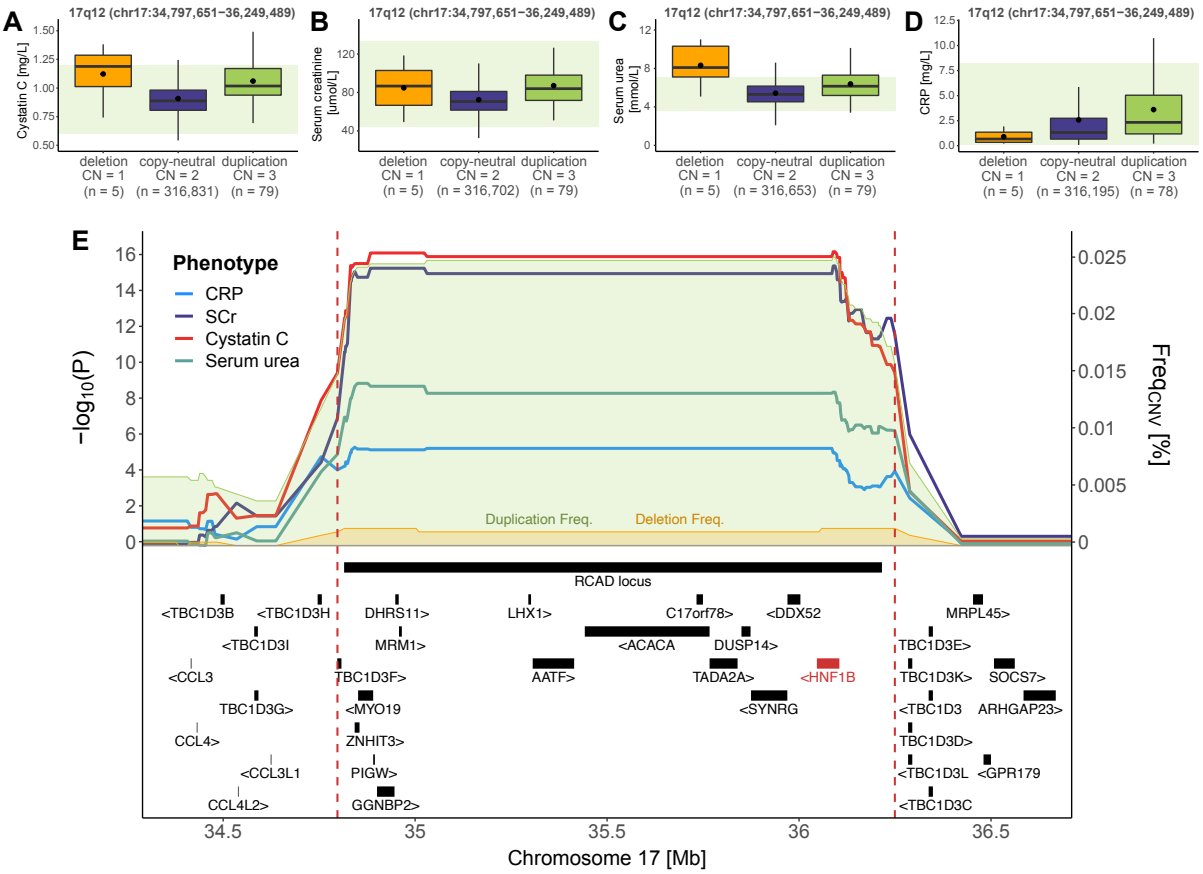


Figure 2.18: 17q12 CNVs and renal phenotypes.
Boxplots representing levels of (**A**) cystatin C, (**B**) serum creatinine (SCr), (**C**) serum urea, and (**D**) C-reactive protein (CRP) in individuals with a 17q12 overlapping deletion, copy-neutrality, or duplication. Copy number (CN) and sample size (n) are reported; dots show the mean; outliers are not shown; green bands show normal clinical range for (A) cystatin C: 0.6–1.2 mg/L; (B) SCr: 44.2-132.6 µmol/L; (C) serum urea: 3.6-7.1 mg/L, and (D) CRP: 0.07-8.2 mg/L. (**E**) Association plot for the 17q12 CNVR. Red dashed lines delimit the duplication-only associated CNVR; left y-axis shows the negative logarithm of association p-value for CRP (blue), SCr (purple), cystatin C (red), and serum urea (turquoise); right y-axis shows CNV frequency (%), with duplication frequency in green and deletion frequency in orange; overlapping DECIPHER recurrent CNV and genes are in black, except for the putative causal gene, *HNF1B*, in red.

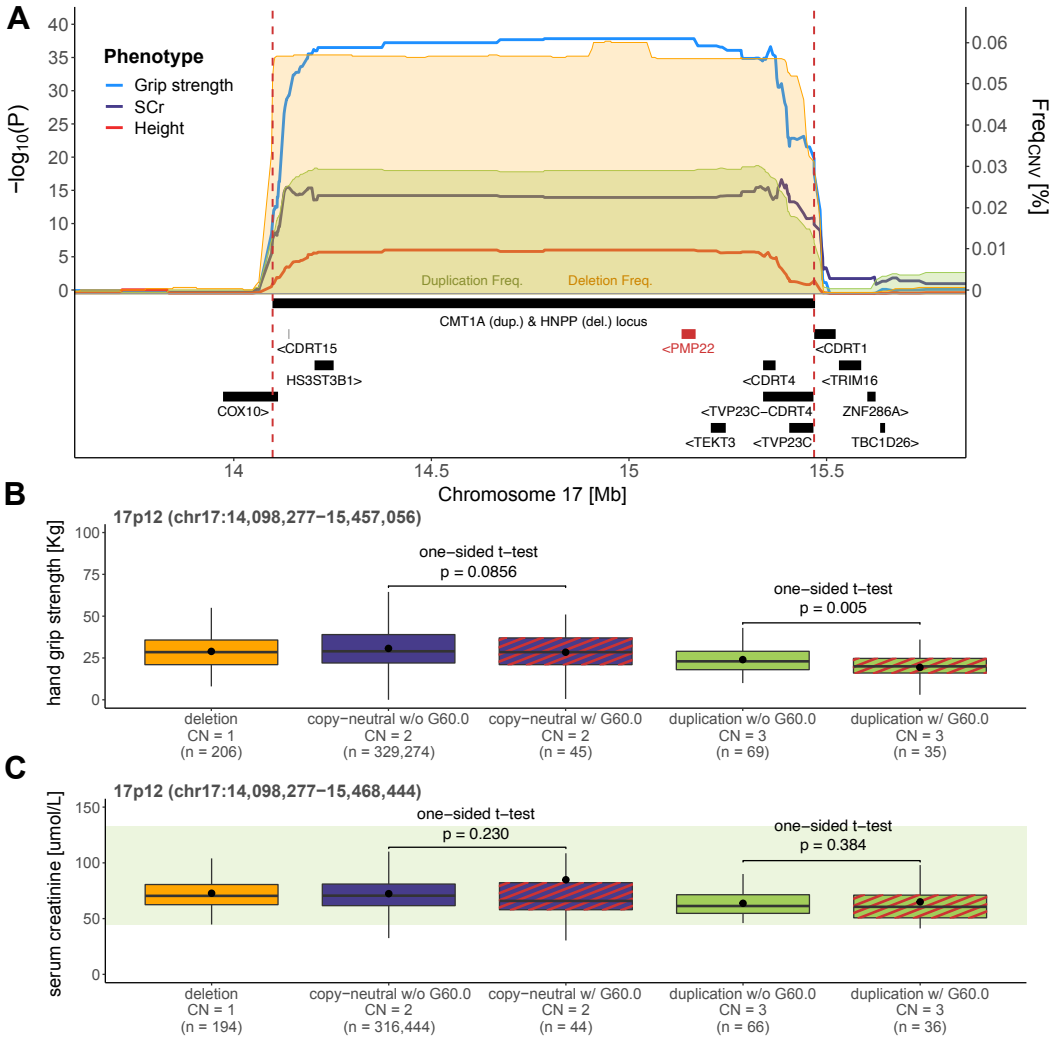


Figure 2.19: 17p12 duplication and muscle phenotypes.
(**A**) Association plot for the 17p12 overlapping CNVR. Red dashed lines delimit the duplication-only associated CNVR; left y-axis represents the negative logarithm of association p-values for hand grip strength (blue), serum creatinine (SCr; purple), and height (red); right y-axis represents CNV frequency (%), with duplication frequency in green and deletion frequency in orange; overlapping DECIPHER recurrent CNV and genes are in black, except for the putative causal gene, *PMP22*, in red. Boxplots representing (**B**) grip strength and (**C**) SCr levels of individuals with a 17p12 overlapping deletion, copy-neutrality, or duplication, the two latter being split according to the presence (w/) or absence (w/o) of a hereditary motor or sensory neuropathy diagnosis (ICD-10 G60.0; red stripes). Copy number (CN) and sample size (n) are reported; dots show the mean; outliers are not shown; green bands show normal SCr clinical range: 44.2-132.6 µmol/L.

CNV-GWAS signals suggest gene functionalities

CNV-GWAS signals can corroborate or generate hypotheses regarding the function of encompassed genes, as shown by the association between the CN of a 1.2 Mb 16p13.11 interval and female reproductive traits. Specifically, duplication of the region correlated with decreased age at menarche (chr16:15,120,501–16,308,285; β_{DUP} = -0.6 years, p = 2.0×10^{-10}) and menopause (chr16:15,151,451–16,308,285; β_{DUP} = -1.8 years, p = 1.7×10^{-6}), whereas its deletion correlated with increased age at menarche (chr16:15,120,501–16,308,285; β_{DEL} = 1.1 years, p = 3.6×10^{-7}), suggesting a shift in reproductive timing associated with the region's CN (Figure 2.20A-B) that aligns with a low, albeit positive, genetic correlation between the two traits (from Neale UKBB genetic correlations). Duplication effect on age at menarche (β_{DUP} = -0.6 years, p = 1.8×10^{-2}) and menopause (β_{DUP} = -2.6 years, p = 4.5×10^{-2}) were confirmed with nominal significance

in the EstBB (Figure 2.8A) and a SNP-GWAS signal for age at menarche (rs153793) colocalized with the CNVR (*369*) (Figure 2.20C).

Literature supports the role of *MARF1* (MIM: 614593) in this association. First, *MARF1* (observed/expected ratio [o/e] = 0.05 [0.03–0.12]; probability of loss-of-function intolerance [pLI] = 1) and *MYH11* (o/e = 0.22 [0.16–0.30]; pLI = 0.77; [MIM: 160745]) are the only encompassed genes under evolutionary constraint (upper bound of o/e <0.35) according to gnomAD v2.1.1 (*29*) (Figure 2.20C; Table S2.6). Second, *MARF1* was shown to play an essential role in murine oogenesis by fostering successful completion of meiosis and cytoplasmic maturation and protecting germline genomic integrity (*372*). The gene's function is supported by studies in fly (*373*) and goat (*374*), as well as two human case reports of females with *MARF1* mutations and reproduction phenotypes (*375*, *376*).

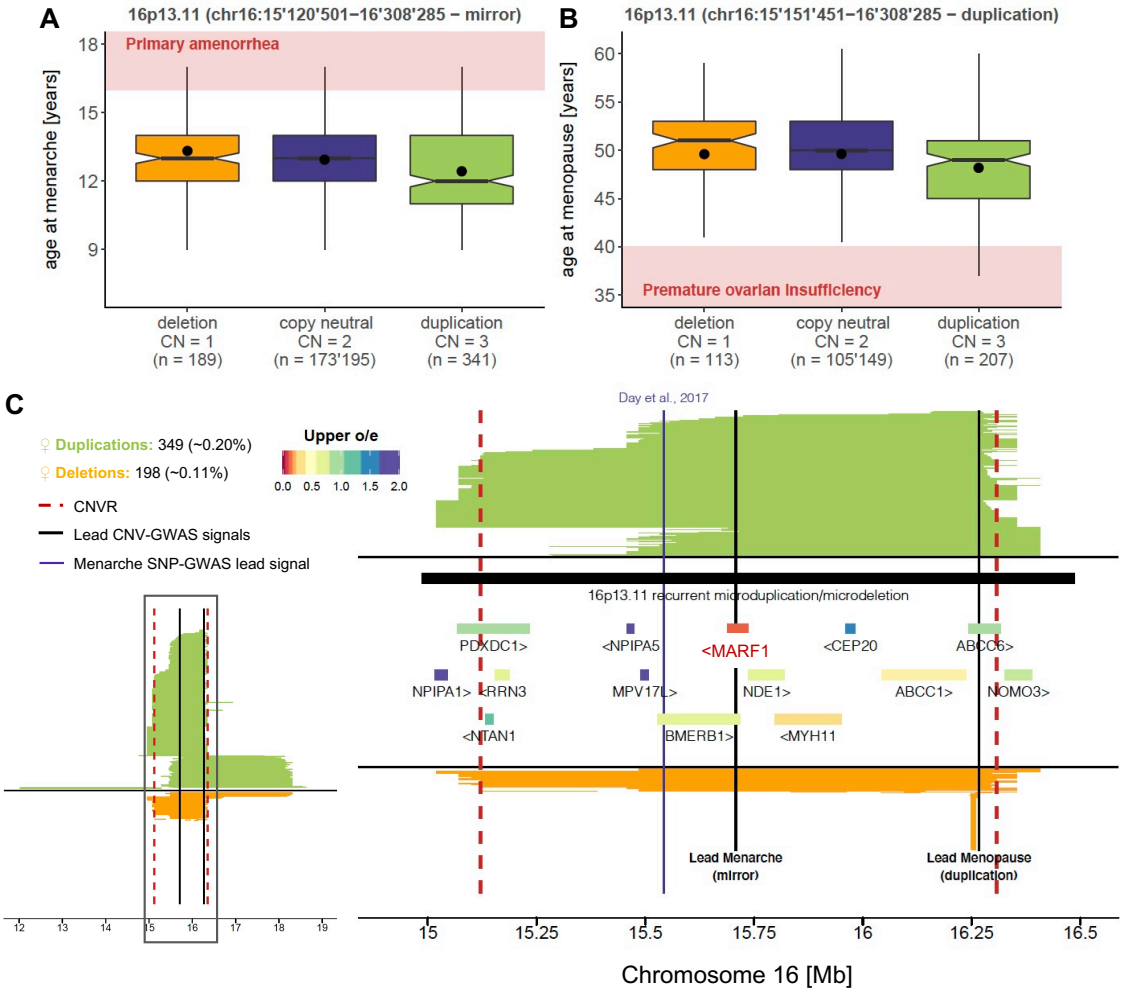


Figure 2.20: *MARF1* as a putative gene involved in human female reproduction.
Boxplots of age at (**A**) menarche and (**B**) menopause in individuals with a 16p13.11 overlapping deletion, copy-neutrality, or duplication. Copy number (CN) and sample size (n) are reported; dots show the mean; notches represent median ± 1.58 · IQR $\sqrt{n}$; outliers are not shown; light red backgrounds indicate pathogenic values corresponding to (A) primary amenorrhea (age at menarche > 16 years) (*370*) and (B) premature ovarian insufficiency (age at menopause < 40 years) (*371*), respectively. (**C**) Mapping of CNVs overlapping the 16p13.11 CNVR (left) with zoom on the black box (right). Number and frequency of duplications (green) and deletions (orange) in females are at the top left. Red dashed lines represent the trait-associated CNVR; black lines indicate the lead signal for age at menarche (mirror) and menopause (duplication-only); purple line indicates age at menarche-associated SNP (*369*); overlapping recurrent DECIPHER CNV is shown in black and protein-coding genes are colored according to the upper bound of the confidence interval for the observed/expected (o/e) mutation ratio (LOEUF score) from gnomAD.

The female-specific role of *MARF1* (372–377) aligns with the absence of association with our proxies for male sexual maturation (i.e., age at first facial hair and balding). Although further investigations are warranted to characterize the function of *MARF1* in human female reproduction and assess the contribution of nearby genes and regulatory elements, it illustrates how CNV-GWASs can be leveraged to generate plausible hypotheses regarding gene functionalities.

The deleterious impact of a high CNV burden

Moving beyond single CNVs, the impact of an individual's total CNV burden on complex traits was estimated. Each participant's autosomal CNV, duplication, and deletion burden was calculated in number of affected Mb or genes. Both Mb and gene burden metrics correlated well (ρ: 0.71–0.74) and while we observed high correlations (ρ: 0.40–0.92) between the CNV and duplication/deletion burdens, the two latter were uncorrelated (Figure 2.21A). From the 57 traits analyzed by CNV-GWASs, 35 (61%) significantly associated with at least one burden metric ($p \leq 0.05/63 = 7.9 \times 10^{-4}$), showcasing negative health consequences such as increased levels of adiposity, liver/kidney damage biomarkers, leukocytes, glycemic values, or anxiety and decreased global physical capacity or intelligence (Figure 2.21B; Table S2.7). Harmful phenotypic consequences were often best captured by the number of deleted genes, in line with a higher sensitivity to decreased (i.e., haploinsufficiency) rather than increased (i.e., triplosensitivity) gene dosage (229).

We then corrected each individual's phenotype and burden for the presence of trait-associated CNVs and performed the burden analysis anew to ensure that signals were not solely driven by significantly trait-associated CNVs (Figure 2.21C; Table S2.7). Whereas the association was lost for albumin, balding, body fat mass, GGT, triglycerides, and weight, indicating a mono- or oligogenic CNV architecture, 30 traits remained associated. Among these, birth weight, total cholesterol, low-density lipoprotein (LDL) cholesterol, and apolipoprotein B (ApoB) were significantly associated with the burden (Figure 2.21D) but lacked CNVR associations (Figure 2.6D). This indicates that, as established for SNPs (81, 171, 296), the CNV architecture underlying most complex traits is polygenic, suggesting the presence of additional associations that we currently lack the power to detect.

The CNV burden extended its impact to global aspects of an individual's life, as illustrated by the negative correlation with several socio-economic factors, including decreased educational attainment (EA; β_{burden} = -0.07 years/Mb, $p = 4.4 \times 10^{-11}$) and income (β_{burden} = -1,593 £/year/Mb, $p = 2.9 \times 10^{-60}$), and increased Townsend deprivation index (β_{burden} = 0.04 SD/Mb, $p = 3.6 \times 10^{-7}$) (Figure 2.21E; Table S2.8). While we did not observe any effect of the CNV burden on age- and sex-corrected telomere length, the trait specifically associated with the *BRCA1* cancer locus (378) (MIM: 113705) (chr17:41,197,733–41,258,551, β_{DUP} = 0.45 SD, $p = 1.9 \times 10^{-8}$), paralleling findings that long telomere-associated SNPs also associate with increased cancer risk (379). Because of the low number of deceased UKBB participants, we used proxies to assess the impact of the CNV burden on lifespan; we observed a negative association between an individual's CNV burden and both parental lifespan (β_{burden} = -0.21 years/Mb, $p = 1.4 \times 10^{-5}$) and age (survivorship proxy; β_{burden}

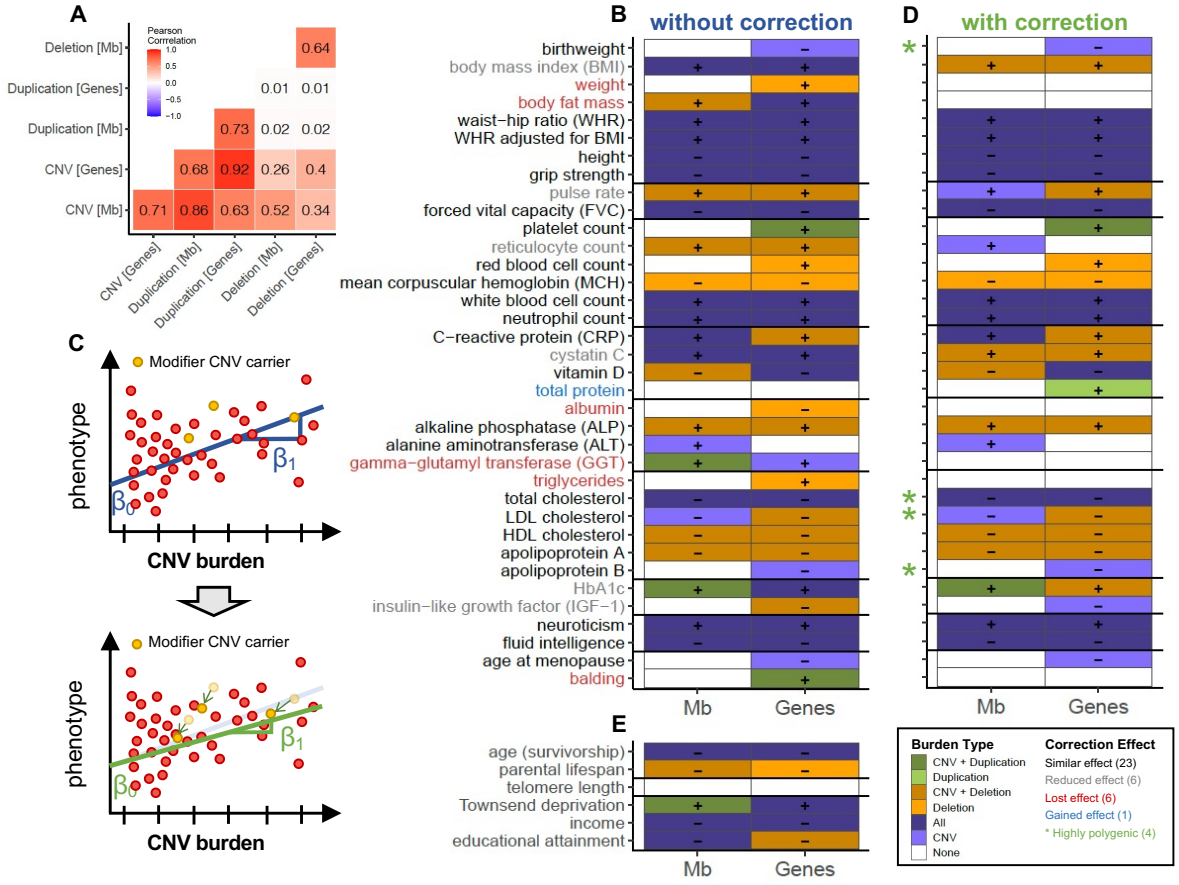


Figure 2.21: The negative impact of the CNV burden on complex traits.
(**A**) Pearson correlation across six burden metrics. (**B**) Significant associations ($p \leq 0.05/63 = 7.9 \times 10^{-4}$) between the CNV burden, expressed as the number of Mb or genes affected by CNVs (x-axis), and traits assessed through CNV-GWASs (y-axis). Color represents the type of burden found to increase (+) or decrease (−) the considered phenotype. (**C**) Schematic representation of the correction for modifier CNVs. Top: individuals carrying a CNV overlapping a CNV-GWAS region were identified (i.e., modifier CNV carrier; yellow). Bottom: Phenotype and burden were corrected (green arrows) and a new linear regression was fitted. (**D**) Significant associations after correction for modifier CNVs. Phenotype label color indicates whether the number of associated metrics between the CNV burden and the trait was fully lost (red), decreased (gray), identical (black), or increased (blue) after the correction. Green stars mark highly polygenic traits associating with the CNV burden without having any significant CNV-GWAS signals. (**E**) Significant associations between the CNV burden and life history traits (y-axis). (D and E) follow the legend in (B).

= -0.18 years/Mb, $p = 1.1 \times 10^{-7}$), suggesting that the deleterious impact of CNVs contributes to decreased longevity (Figure 2.21E; Table S2.8). Given this, we questioned whether the CNV burden was transmitted at a Mendelian rate. Taking advantage of the presence of a UKBB sibling for 16,179 individuals assessed in our previous analyses, we calculated that the average fraction of shared CNVs among siblings was 27%. Whereas substantially higher than for random pairs (0.7%), it only represents 54% of the expected fraction of shared additive genetic variance among siblings (50%) (*380*). Together, these results describe the broadly deleterious impact of CNVs on a wide range of complex traits in the general population and suggest that most traits are influenced by a polygenic CNV architecture.

Discussion

By coupling CNV calls to the phenotypic data available in the UKBB, we

generated a roadmap of clinically relevant CNV-trait associations that allowed us to gain deeper insights into specific biological pathways and put forward general patterns describing the role of CNVs in shaping complex human traits in the general population.

Our UKBB CNV landscape matched previous reports (*292*), and while some of the 131 CNV-GWAS signals overlapped known associations (*292*, *294*, *295*, *305*) our analyses shed light on others that have not been studied extensively. The combined use of three association models revealed general patterns through which CNVs modulate phenotypes, and while geared toward the discovery of mirror effects, we also witnessed U-shape effects, illustrating different mechanisms through which altered dosage influences phenotypes. We further provide evidence for a broad and nuanced role of CNVs in shaping complex traits, as both common (frequency ≥ 1%) and rare (frequency < 1%) CNVs mapping to regions involved by SNP-GWAS contribute to phenotypic variability in the general population, and rare CNVs have larger effects sizes than common ones. Other signals mapped to regions involved in Mendelian disorders. Studying pathogenic CNVs in the general population, as opposed to clinical cohorts selected based on phenotypic criteria or family history, makes it possible to re-assess their frequency, penetrance, expressivity, pleiotropy, and inheritance pattern. Matching the increasing awareness around variable penetrance and expressivity (*234*–*236*, *289*, *368*, *381*), we show that pathogenic dominant CNVs can impact disease-associated traits without causing clinically diagnosable disorders, whereas recessive CNVs can impact disease-related biomarkers at the heterozygous state. Together, these results provide a more complex and nuanced – but also broader – understanding of the phenotypic impact of CNVs at odds with the classical dichotomy between common complex diseases and rare Mendelian disorders.

Confirming the deleterious influence of a high CNV load on anthropometric traits (*295*, *382*, *383*) and EA (*300*, *384*, *385*) in a nonclinical cohort, we extended this observation to over 30 global health biomarkers. We show how the CNV burden – limited to large and rare CNVs detectable by microarrays – shapes intermediate molecular phenotypes that predate or are consequences of disease processes in a population-based cohort, consistent with its known contribution to a wide range of disorders (*386*–*390*). Our data further show that the CNV load negatively impacts socio-economic factors and longevity proxies. The lower CNV burden observed in individuals with advanced age matches the depletion of life-shortening alleles in older UKBB participants (*391*), suggesting improved health/decreased mortality in individuals with a low CNV load. Parental lifespan negatively correlated with the CNV burden. While lower than expected, a substantial fraction of CNVs (27%) was shared among siblings and thus inherited from either parent. As inclusion of haplotype sharing information in CNV calling mainly increases the detection of small (< 10 kb) but not that of large CNVs (*306*), we hypothesize that large events recurrently appear *de novo* on multiple backgrounds and are rapidly eliminated from the population through transmission bias or from the cohort through ascertainment bias (i.e., increased participation of healthier siblings) because of their deleteriousness. Our analysis of CNV call quality in the EstBB suggests marginal contribution of false CNV calls but confounders – such as CNV length, which affects both detection

capacity and pathogenicity – prevent the assessment of these factors separately. Nevertheless, the lower-than-expected CNV inheritance allows speculating that an even stronger association with lifespan would be obtained providing access to parental CNV genotypes. If further studies are required to confirm the life-shortening effect of a high CNV load, our data clearly illustrate the deleterious impact of CNVs on an individual's global health.

Both CNV-GWASs and burden analyses results improve our understanding of the CNV architecture underlying studied traits. Many CNV-GWAS loci involve rare but recurrent CNVRs. Due to the difficulty of gathering large cohorts of carriers, complete phenotypic characterization of these loci is still missing and limited to easily assessed anthropometric traits or severely debilitating neurodevelopmental and psychiatric disorders. Our results provide a map of the pleiotropic consequences of these CNVRs on over 50 medically relevant traits. Some traits are not typically assessed or reported in patient cohorts and targeted study of their distribution among cases might refine diagnostic criteria and help clinicians identify and follow-up on patients with mild and/or atypical presentation. Mechanistically, most assessed CNVRs are large, potentially harboring several causal genes. One of the next challenges will be to narrow down causal region(s) in pleiotropic multi-genic CNVRs to untangle primary from secondary associated traits, as some, such as obesity, are known to causally alter multiple biomarkers (*392*–*394*). The substantial overlap between CNV- and SNP-GWAS signals speaks for the presence of shared genetic mechanisms, so that both mutational classes can be exploited synergistically to pinpoint causal genes and elucidate their biological function. In parallel, we observed a high degree of CNV-polygenicity, as 30 out of 35 traits remained associated with the CNV burden after correction for modifier CNVRs. For six traits, CNV-GWAS signals captured the bulk of phenotypic variability caused by CNVs, while ApoB, birth weight, LDL cholesterol, and total cholesterol were solely associated with the CNV burden. This indicates a polygenic CNV architecture that might arise from rare high-impact CNVs that were not assessed by CNV-GWASs (frequency < 0.005%) and/or more frequent CNVs with mild effects; indeed, most high-frequency CNVRs do not overlap CNV-GWAS signals (Figure 2.4 and Figure 2.6D). Among these, decreased birth weight, which is associated with a high CNV load, has been linked to increased risk for metabolic syndrome, obesity, and various other diseases in adulthood (*395*, *396*), opening the question as to whether some of the deleterious effects of the CNV burden are rooted in early development. Strikingly, the three other traits are plasma lipids with few CNV-GWAS signals. Speaking for their high polygenicity, a GWAS on 35 blood biomarkers in the UKBB found an average of 87 vs 478 associations per trait for non-lipid compared to lipid traits (*305*). Collectively, these results illustrate a more complex than expected contribution of CNVs in shaping the genetic architecture of complex human traits.

It is important to keep in mind limitations of the current study. First, CNVs were called based on microarray data with PennCNV. In addition to the high false positive rates associated with array-based CNV calls, this renders the study blind to variants in regions not covered by the array, limits resolution—both in length and exact breakpoint location—and hinders the detection of high copy-number states (CN ≥ 4) and

deviations thereof. To mitigate these issues, we stringently filtered CNVs and transformed calls to the probe level (*206*, *295*), which at risk of missing true associations guarantees the identification of trustworthy CNV-trait pairs. Few cohorts have sufficiently large genetic and phenotypic coverage to replicate UKBB findings at adequate power, so that we relied on literature evidence to gauge the validity of our results, highlighting the need for large-sized biobanks for studying (rare) CNVs. Future release of large sequencing datasets combined to progress in CNV detection tools could resolve these issues and lead to novel discoveries (*209*, *306*, *397*, *398*). Second, despite substantial evidence of CNV- and SNP-GWAS signal colocalization, we did not perform robust enrichment analyses, as the non-random genomic distribution and complex nature of CNVs renders simulating the null scenario beyond the scope of this paper. Signal colocalization is likely to be underestimated, as manual literature searches revealed overlaps missed by our annotation pipeline (e.g., 16p13.11 age at menarche signal (*369*)) and we obtained a 7% increase in signal colocalization by using GWAS Catalog annotation 6 months apart (31% 04/2021 vs 38% 10/2021). Third, our study is limited to individuals of white British ancestry. As CNV frequencies vary across populations (*36*, *399*–*401*), assessing diverse ancestral groups is likely to unravel new associations, even though smaller sample sizes represent a limiting factor. Finally, UKBB suffers from healthy cohort bias (*59*). Focusing on the impact of CNVs in healthy populations, we used this bias to our advantage through the inclusion of CNV carriers with subclinical phenotypes, providing lower bounds for effect size estimates (*235*, *236*, *402*). However, this means that the cohort is depleted for severely affected cases and extremely rare (frequency < 0.005%) but pathogenic CNVs were not tested for associations. Extending the analysis to low frequency/high impact CNVRs would allow for better distinguishing of mechanisms of action – with the remaining caveat that effects will be underestimated due to selection bias – and will be the focus of future work.

In conclusion, our study provides a map of high-confidence CNV-trait associations. While we explored some of the reported signals, collective efforts will be required to validate and interpret these discoveries and we hope that this resource will be useful for researchers and clinicians aiming at improving the characterization of recurrent CNVs. Our study revealed the nuanced role of CNVs along the rare versus common disease spectrum, their shared mechanisms with SNPs, as well as a widespread polygenic CNV architecture, consolidating the growing body of evidence implicating CNVs in the shaping of complex human traits.

Acknowledgments

We thank all biobank participants for sharing their data. UKBB and EstBB computations were carried out on the JURA server (University of Lausanne) and the High-Performance Computing Center (University of Tartu), respectively. This work was supported by funding from the Department of Computational Biology (ZK) and the Center for Integrative Genomics (AR) from the University of Lausanne, as well as grants from the Swiss National Science Foundation (31003A_182632, AR), Horizon2020 Twinning projects (ePerMed 692145, AR), and the Estonian Research Council (PRG687, ML and RM). Critical reading of the draft by

Johan Auwerx and Matthew Robinson was appreciated.

Declaration of interests

The authors declare no competing interests.

Supplemental tables

Supplemental tables are available for download as a single Excel file.

- **Table S2.1** Description of analyzed complex traits.
- **Table S2.2** Significant CNV-GWAS signals in the UKBB and replication in the EstBB.
- **Table S2.3** GWAS Catalog SNP-GWAS signals for assessed traits.
- **Table S2.4** TWMR causal effects of differential *RHD* and *RSRP1* expression on hematological traits.
- **Table S2.5** Effect of Rhesus blood type on hematological traits in the CHUV maternity cohort.
- **Table S2.6** GnomAD constraint metrices for *MARF1* and *MYH11*.
- **Table S2.7** Association between complex traits and the CNV burden.
- **Table S2.8** Association between life history traits and the CNV burden.

Common diseases 3

All interest in disease and death is only another expression of interest in life.

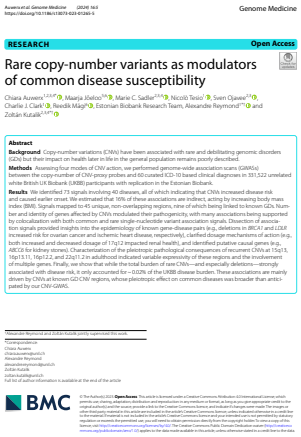
– Thomas Mann

3.1 Aims 83
3.2 Key Findings 84
3.3 Author Contributions 84
3.4 Rare copy-number variants as modulators of common disease susceptibility 85

This chapter describes *"Rare copy-number variants as modulators of common disease susceptibility"* (*82*), which was published in *Genome Medicine* and represents a major extension to the previous chapter both in terms of methodological development and description of new biological insights. The version presented in the dissertation integrates supplemental content.

I presented this study at multiple international conferences, including the Swiss Society of Medical Genetics, the European Society of Human Genetics, and the American Society of Human Genetics, where the abstract was selected as a semifinalist for the 2023 predoctoral award for excellence in human genetics research.

As for the previous study, all CNV-GWAS summary statistics were made available through the GWAS Catalog.


Auwerx et al. Genome Medicine (2024) 16:3
https://doi.org/10.1186/s13073-023-01265-5

Genome Medicine

RESEARCH

Open Access

Rare copy-number variants as modulators of common disease susceptibility

Abstract

Figure 3.1: Cover of *Auwerx et al., 2024*.

Outreach:

- → Unil press release (French)
- → SIB press release

Data & code availability:

- → CNV-GWAS summary statistics
- → GitHub

3.1 Aims

After having demonstrated the crucial role of CNVs in modulating complex *quantitative* traits within the general population (*208*), we wanted to probe the clinical relevance of these findings by determining whether CNVs also modulate susceptibility to common diseases. To achieve this, the study's goal included:

1. Extend the CNV-GWAS pipeline to accommodate binary traits by switching from linear to logistic regression. To address statistical challenges linked to the low prevalence of both CNVs and disease diagnoses in UKBB, we developed a series of pre-processing steps and validation strategies aiming at improving computational time and increasing the robustness of our results.
2. Incorporate a fourth, U-shape, association model that assesses the impact of any deviation in copy-number against the copy-neutral state, based on the intuition that some genetic regions are dosage-sensitive.
3. Develop a time-to-event framework that determines whether CNVs are associated with age of disease onset.
4. Apply the newly developed framework to 60 common disease diagnoses defined through manual curation. Diseases cover a wide range of physiological systems, including disorders for which the contribution of CNVs remains unknown.
5. Estimate the contribution of CNVs to the total disease burden in UKBB and determine the genomic location and copy-number alteration that most strongly contribute to it.

3.2 Key Findings

Our study revealed 73 associations involving 41 traits and 45 unique genomic regions. Thanks to our multiple validation strategies, we could classify these associations in confidence tiers that reflect the amount of supporting data gathered. All of our associations were replicated through time-to-event analysis, showing that CNVs lead to an earlier age of diagnosis. Some regions, such as 17q12, followed a U-shape association model with both duplications and deletions having deleterious consequences, suggesting tight dosage control.

Unlike the previous study, only nine of the uncovered regions had previously been linked to genomic disorders. Some of the associations involve single-gene CNVs, e.g., *BRCA1* or *LDLR*, which are associated with highly penetrant forms of common diseases, i.e., ovarian cancer and ischemic heart disease, respectively. While these associations are well-established, they nevertheless represent the first description of CNVs in these genes within UKBB. We use these examples to showcase how the rich phenotypic data available for UKBB participants can be leveraged to gain new insights into the epidemiology and comorbidities of these gene-disease pairs.

Regions linked to genomic disorder exhibited complex pleiotropy, whose extent is likely to be currently underestimated. We describe multiple novel associations and confirm the variable expressivity of these loci by linking them to the same physiological systems previously implicated by studies in clinical cohorts. We also illustrate how pleiotropic association signals can be dissected to gain insights into the pathological mechanisms of recurrent CNVs, allowing the identification of putative causal genes or suggesting an oligogenic contribution to the phenotype. These results are of key clinical relevance as a better understanding of the spectrum of adult comorbidities linked to each genomic disorder can translate into preventive measures.

Finally, our CNV burden analysis revealed that CNVs increased the risk of 20 individual diseases, despite only contributing to ~0.02% of the UKBB disease burden – with slightly higher contribution to psychiatric disorders. Disease risk was primarily driven by deletions affecting regions linked to known genomic disorders. Overall, this emphasizes the role of rare CNVs as modulators of common disease susceptibility and aligns with the paradigm that the latter are of marginal importance at the population level but are highly relevant in terms of personalized medicine, as they strongly impact the disease risk of carriers.

3.3 Author Contributions

This study was conceived by Zoltán Kutalik, Alexandre Reymond, and myself. I carried out the bulk of the analyses, under the supervision of Zoltán Kutalik. Co-authors have contributed in the following ways:

- The Estonian Biobank Research Team (Tõnu Esko, Andres Metspalu, Lili Milani, Reedik Mägi, Mari Nelis) coordinated genotyping and

sequencing data acquisition in the EstBB and Maarja Jõeloo did the replication study under the supervision of Reedik Mägi.
- Marie Sadler provided SNP-GWAS data for locus zoom plots.
- Nicolò Tesio validated and refined definitions of cases and controls.
- Sven Ojavee provided guidance for time-to-event analysis.
- Charlie Clark proofed the binary association pipeline.

Results were interpreted by Zoltán Kutalik, Alexandre Reymond, and myself. I designed all the figures and drafted the manuscript, with critical revisions made by Zoltán Kutalik and Alexandre Reymond.

3.4 Rare copy-number variants as modulators of common disease susceptibility

Chiara Auwerx [1,2,3,4,*], Maarja Jõeloo [5,6], Marie C. Sadler [3,4], Nicolò Tesio [1], Sven Ojavee [2,3], Charlie J. Clark [1], Reedik Mägi [6], Estonian Biobank Research Team [6], Alexandre Reymond [1,*], and Zoltán Kutalik [2,3,4*].

Abstract

Copy-number variations (CNVs) have been associated with rare and debilitating genomic disorders (GDs) but their impact on health later in life in the general population remains poorly described. Assessing four modes of CNV action, we performed genome-wide association scans (GWASs) between the copy-number of CNV-proxy probes and 60 curated ICD-10-based clinical diagnoses in 331,522 unrelated white British UK Biobank (UKBB) participants with replication in the Estonian Biobank. We identified 73 signals involving 40 diseases, all of which indicated that CNVs increased disease risk and caused earlier onset. We estimated that 16% of these associations are indirect, acting by increasing body mass index (BMI). Signals mapped to 45 unique, non-overlapping regions, nine of which being linked to known GDs. Number and identity of genes affected by CNVs modulated their pathogenicity, with many associations being supported by colocalization with both common and rare single-nucleotide variant association signals. Dissection of association signals provided insights into the epidemiology of known gene-disease pairs (e.g., deletions in *BRCA1* and *LDLR* increased risk for ovarian cancer and ischemic heart disease, respectively), clarified dosage mechanisms of action (e.g., both increased and decreased dosage of 17q12 impacted renal health), and identified putative causal genes (e.g., *ABCC6* for kidney stones). Characterization of the pleiotropic pathological consequences of recurrent CNVs at 15q13, 16p13.11, 16p12.2, and 22q11.2 in adulthood indicated variable expressivity of these regions and the involvement of multiple genes. Finally, we show that while the total burden of rare CNVs – and especially deletions – strongly associated with disease risk, it only accounted for ~0.02% of the UKBB disease burden. These associations are mainly driven by CNVs at known GD CNV regions, whose pleiotropic effect on common diseases was broader than anticipated by our CNV-GWAS. Our results shed light on the prominent role of rare CNVs in determining common disease susceptibility within the general population and provide actionable insights for anticipating later-onset

[1] Center for Integrative Genomics, University of Lausanne, Lausanne 1015, Switzerland; [2] Department of Computational Biology, University of Lausanne, Lausanne 1015, Switzerland; [3] Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland; [4] University Center for Primary Care and Public Health, Lausanne 1010, Switzerland; [5] Institute of Molecular and Cell Biology, University of Tartu, Tartu 51010, Estonia; [6] Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu 51010, Estonia; * **Correspondence**.

comorbidities in carriers of recurrent CNVs.

Introduction

Copy-number variants (CNVs) refer to duplicated or deleted DNA fragments (≥ 50 bp) and represent an important source of inter-individual genetic variation (*36*, *324*). As a highly diverse mutational class, CNVs can alter the copy-number of dosage-sensitive genes, induce gain- or loss-of-function (LoF) through gene fusion or truncation, unmask recessive alleles, or disrupt regulatory sequences, thereby representing potent phenotypic modifiers (*227*, *403*). As such, their role in human disease has mainly been studied in clinically ascertained cohorts, often presenting with congenital anomalies and/or severe neurological (e.g., developmental delay and intellectual disability or epilepsy) or psychiatric (e.g., autism or schizophrenia) symptoms (*335*, *386*, *387*, *389*). Today, close to 100 genomic disorders (GDs), i.e., diseases caused by genomic rearrangements, have been described (*229*, *265*). Despite their deleteriousness, some of these CNVs are flanked by repeats and recurrently appear, remaining at a low but stable frequency in the population (*180*).

The emergence of large biobanks coupling genotype to phenotype data has fostered the study of CNVs in the general population. Whole genome sequencing represents the best approach to characterize the full human CNV landscape (*34*–*36*) but current long- and short-read sequencing association studies have a limited sample size (*10*, *397*, *398*). Alternatively, larger sample sizes are available for exome sequencing data, offering the possibility to assess the phenotypic consequence of small CNVs (*209*–*211*), while microarray-based CNV calls are better-suited for the study of large CNVs and have been successfully used in association studies (*208*, *229*, *292*–*295*, *300*, *302*, *304*–*306*, *404*). Performing a CNV genome-wide association study (GWAS) on 57 medically relevant continuous traits in the UK Biobank (UKBB) (*61*), we previously identified 131 independent associations, including allelic series wherein carriers of CNVs at loci previously associated with rare Mendelian disorders exhibited subtle changes in disease-associated phenotypes but lacked the corresponding clinical diagnosis (*208*). Paralleling findings for point mutations (*234*–*237*), this supports a model of variable expressivity, where CNVs can cause a wide spectrum of phenotypic alterations ranging from severe, early-onset diseases to mild subclinical symptoms, opening the question as to whether these loci are also associated with common diseases.

While continuous traits can be objectively measured in any individual, population cohorts, such as UKBB, have lower numbers of diseased individuals compared to the population as a whole (*59*), leading to a case-control imbalance that reduces power compared to a balanced cohort of the same size. Moreover, defining cases relies on the dichotomization of complex underlying pathophysiological processes (*138*). Beyond the inherent loss of power associated with the usage of binary variables (*405*), cases might be missed because an individual did not consult a physician, was misdiagnosed due to atypical clinical presentation, or is in a prodromal disease phase. Studies investigating CNV-disease associations in the general population have either focused on only a few diseases (*277*, *304*, *406*–*410*) or well-established recurrent CNVs (*40*, *293*, *411*–*413*). Alternatively, high-throughput studies have assessed a broad range of continuous and binary traits simultaneously (*209*, *292*, *306*) without any

precautions to accommodate the aforementioned challenges. To date, the largest disease CNV-GWAS meta-analyzed ~1,000,000 individuals (*229*). While boosting power through increased sample size, it comes at the cost of extensive data harmonization, resulting in the exclusion of smaller CNVs (≤ 100 kb) and usage of broader disease categories (e.g., *immune abnormality*). Moreover, as this study includes several clinical cohorts, phenotypes are biased towards neuropsychiatric disorders (24 out of 54 phenotypes) for which the role of CNVs is well-established (*335*, *386*, *387*, *389*).

Using tailored CNV-GWAS models mimicking four mechanisms of CNV action and time-to-event analysis, we investigate the relationship between CNVs and 60 carefully defined common diseases affecting a broad range of physiological systems in 331,522 unrelated white British UKBB participants. Extensively validating our results, we report associations according to confidence tiers and take advantage of rich individual-level phenotypic data to demonstrate the contribution of CNVs to the common disease burden in the general population.

Materials and methods

Study material

Discovery cohort: UK Biobank

The UK Biobank (UKBB) is composed of ~500,000 volunteers (54% females) from the general UK population for which microarray-based genotyping and extensive phenotyping data – including hospital-based International Classification of Diseases, 10th Revision (ICD-10) codes (up to September 2021) and self-reported conditions – are available (*61*). UKBB data were accessed through application 16389. UKBB has approval from the North West Multi-centre Research Ethics Committee as a Research Tissue Bank and all participants signed a broad informed consent form.

Replication cohort: Estonian Biobank

The Estonian Biobank (EstBB) is a population-based cohort of ~208,000 Estonian individuals (65% females; data freeze 2022v01 [12/04/2022]) for which microarray-based genotyping data and ICD-10 codes from cross-linking with national and hospital databases (up to end 2021) are available (*62*). The activities of the EstBB are regulated by the Human Genes Research Act, which was adopted in 2000 specifically for the operations of the EstBB. Individual-level data analysis in the EstBB was carried out under ethical approval 1.1–12/624 from the Estonian Committee on Bioethics and Human Research (Estonian Ministry of Social Affairs), using data according to release application 3–10/GI/34668 [20/12/2022] from the EstBB. All participants signed a broad informed consent form.

Software versions:

- CNV calling: PennCNV v1.0.5 (*203*).
- CNV QC: (*206*).
- PLINK v1.9 and PLINK v2.0.26 (*88*).
- Gene mapping: ANNOVAR (*307*).
- UCSC Browser LiftOver (*414*)
- Statistical analyses: R v3.6.1.
- Graphs: R v4.1.3.

CNV association studies in the UK Biobank

Microarray-based CNV calling

All results in this study are based on the human genome reference build GRCh37/hg19. UKBB genotype microarray data were acquired from two arrays with 95% probe overlap (Applied Biosystems UK Biobank Axiom Array: 438,427 samples; Applied Biosystems UK BiLEVE Axiom Array by Affymetrix: 49,950 samples) (*61*) and used to call CNVs as previously described (*208*). Briefly, CNVs were called using standard PennCNV

settings. Chromosome X CNVs were called using dedicated PennCNV modalities. To avoid interference between the two-letter CNV encoding (Table 3.1) and the male chromosome X hemizygosity assumption of PLINK, all individuals were (falsely) labeled as female when performing genetic analyses in PLINK. CNV outlier samples based on genotyping plate or extreme CNV profile were excluded (Table 3.2). Remaining CNVs were attributed a probabilistic quality score (QS) that reflects the likelihood that the CNV call is a true positive. The QS ranges from −1 (likely deletion) to 1 (likely duplication), with intermediate values around 0 reflecting less confident CNV calls (*206*). High-confidence CNVs, stringently defined by $|QS| > 0.5$, were retained and encoded in chromosome-wide $probe \times sample$ matrices with entries of 1, −1, or 0 indicating probes overlapping a high-confidence duplication, deletion, or no/low-quality CNV, respectively. CNV matrices were encoded into three PLINK binary file sets (`--make-bed` PLINK v1.9; Table 3.1). To reduce file size and facilitate parallelized computation, files are split at the chromosome level (i.e., for each PLINK encoding there are 24 files: 22 autosomes, pseudoautosomal regions, chromosome X). PLINK file sets were used to fit four association models mimicking different modes of CNV action: mirror, U-shape, duplication-only, or deletion-only (Table 3.1).

Table 3.1: Numerical and PLINK encoding of CNVs.
Encoding of high-confidence CNVs into numerical $probe \times sample$ quality score (QS) matrices (Num.) and three PLINK file sets (PLINK), mimicking encoding of single-nucleotide variants. *The U-shape model is assessed with the `hetonly` modifier of PLINK's `glm` function, allowing to compare the effect of deletion and duplications against copy-neutral individuals.

	Mirror		**U-shape**		**Duplication-only**		**Deletion-only**	
PLINK file set	$PLINK_{CNV}$		$PLINK_{CNV}$*		$PLINK_{DUP}$		$PLINK_{DEL}$	
Encoding	Num.	PLINK	Num.	PLINK	Num.	PLINK	Num.	PLINK
Deletion (QS < -0.5)	-1	AA	1	AA	NA	00	1	TT
Copy-neutral (-0.5 ≤ QS ≤ 0.5)	0	AT	0	AT	0	AT	0	AT
Duplication (QS > 0.5)	1	TT	1	TT	1	TT	NA	00

Case–control definition and age of disease onset calculation
A pool of 331,522 unrelated white British UKBB participants (54% females) was considered after excluding related individuals (≤ 3rd degree), individuals with high genotype missingness (≥ 0.02), individuals that are not of white British ancestry (self-reported + genetically confirmed), CNV outlier samples based on genotyping plate or extreme CNV profile, and individuals reporting blood malignancies (Table 3.2).

Cases and controls were defined for 60 ICD-10-based clinical diagnoses using *diagnosis − ICD10* (#41270), *cancer code, self-reported* (#20001), and *non-cancer illness code, self-reported* (#20002) to build exclusion and inclusion lists. For each disease, starting with the selected subset of 331,522 individuals previously described, we identified cases as individuals having received a specific, restricted set of ICD-10 codes matching our disease definition (i.e., inclusion list). We then defined our controls as individuals lacking both ICD-10 codes matching the case definition and self-reported or ICD-10 diagnoses of a broad set of conditions related to the assessed disorder (i.e., exclusion list). For instance, breast cancer controls should not have other cancers or radio-/chemotherapy, while schizophrenia controls should not have mood or personality disorders. For second-level ICD-10 codes, all subcodes are considered, otherwise

Table 3.2: Summary of sample filtering procedure.
Ordered list of filters applied to select individuals for the CNV-GWAS analysis. Description of criteria and rationale are provided for each step, along with the number of excluded individuals ($N_{exc.}$) and the number of individuals remaining after applying each filter ($_{rem.}$).

STEP	Filter	Description	$N_{exc.}$	$N_{rem.}$
START				**488,377**
1	Relatedness	Samples were excluded if "0" in `used.in.pca.calculation` (i.e., not used for principal component analysis (PCA) calculation) in the sample QC file (v2) described in UKBB resource 531. PCA was calculated based on unrelated individuals (KING software (*415*) `--related --degree 3`), with missing rate on autosomes ≤ 0.02, and no mismatch between inferred and self-reported sex (*61*). Focusing on unrelated individuals prevents p-value deflation due to correlated residual noise.	81,158	407,219
2	Ancestry	Only sample with `in.white.British.ancestry.subset` as "1" (i.e., self-identify as "white British" and cluster with that group based on single nucleotide polymorphism (SNP) PCA analysis (*61*)) in the sample QC file (v2) described in UKBB resource 531 were retained. This allows to obtain a sample with homogenous genetic ancestry.	69,674	337,545
3	Retracted	Samples that were redacted or retracted their participation at the time the project was initiated (August 2020) were excluded.	80	337,465
4	Genotype plate outlier	Samples that were genotyped on a genotyping plate with a mean CNV count per sample > 100 were excluded as this might indicate systematic error during the genotyping and lead to the inclusion of artifactual CNV calls.	569	336,869
5	Extreme CNV profiles	Individuals with an extreme CNV profile, i.e., > 200 CNVs/sample or a single CNV > 10 Mb were excluded. The former could either indicate poor quality genotyping, the presence of a large CNV that was called as many small CNVs, or extreme events such as chromothripsis. Extremely large CNV can reflect aneuploidies or other extreme chromosomal aberrations. As we expect these events to be rare, with potentially massive phenotypic consequences, we decided to exclude these individuals.	924	335,972
6	Blood cancer	Individuals with a known blood malignancy (i.e., UKBB field #20001: 10047, 1048, 1050, 1051, 1052, 1053, 1055, 1056, 1056; #41270: ICD-10 codes mapping to Phecode exclusion range of *cancer of lymphatic and hematopoietic tissue* (*309*)) were excluded as these individuals might harbor somatic CNVs, which are not within the scope of this study.	4,450	331,522
END				**331,522**

only the specified ones. Finally, the disease burden was calculated as the number of diagnoses (out of the 60 assessed ones) an individual has received. For male- (prostate cancer) and female- (menstruation disorders, endometriosis, breast cancer, ovarian cancer) specific diseases, downstream analyses were conducted excluding individuals from the opposite sex.

Based on the *date at first in-patient diagnosis – ICD10* (#41280) and the individual's *month* (#52) and *year* (#34) of birth (day assumed on average to be the 15^{th}), age at diagnosis was calculated by subtracting the earliest diagnosis date for codes on the inclusion list from the birth date and converting it to years by dividing by 365.25 to account for leap years.

Covariate and probe selection
Relevant covariates and probes were selected to fit tailored main CNV-GWASs and reduce computation time.

For each disease, a logistic regression was fitted to explain disease probability as a function of age (#21003), sex, genotyping array, and the 40 first principal components (PCs) from the SNP genotyping data. Nominally significantly associated covariates ($p \leq 0.05$) were retained for the main GWAS. Number of retained covariates ranged between two (sarcoidosis and multiple sclerosis) and 24 (hypertension, arthrosis,

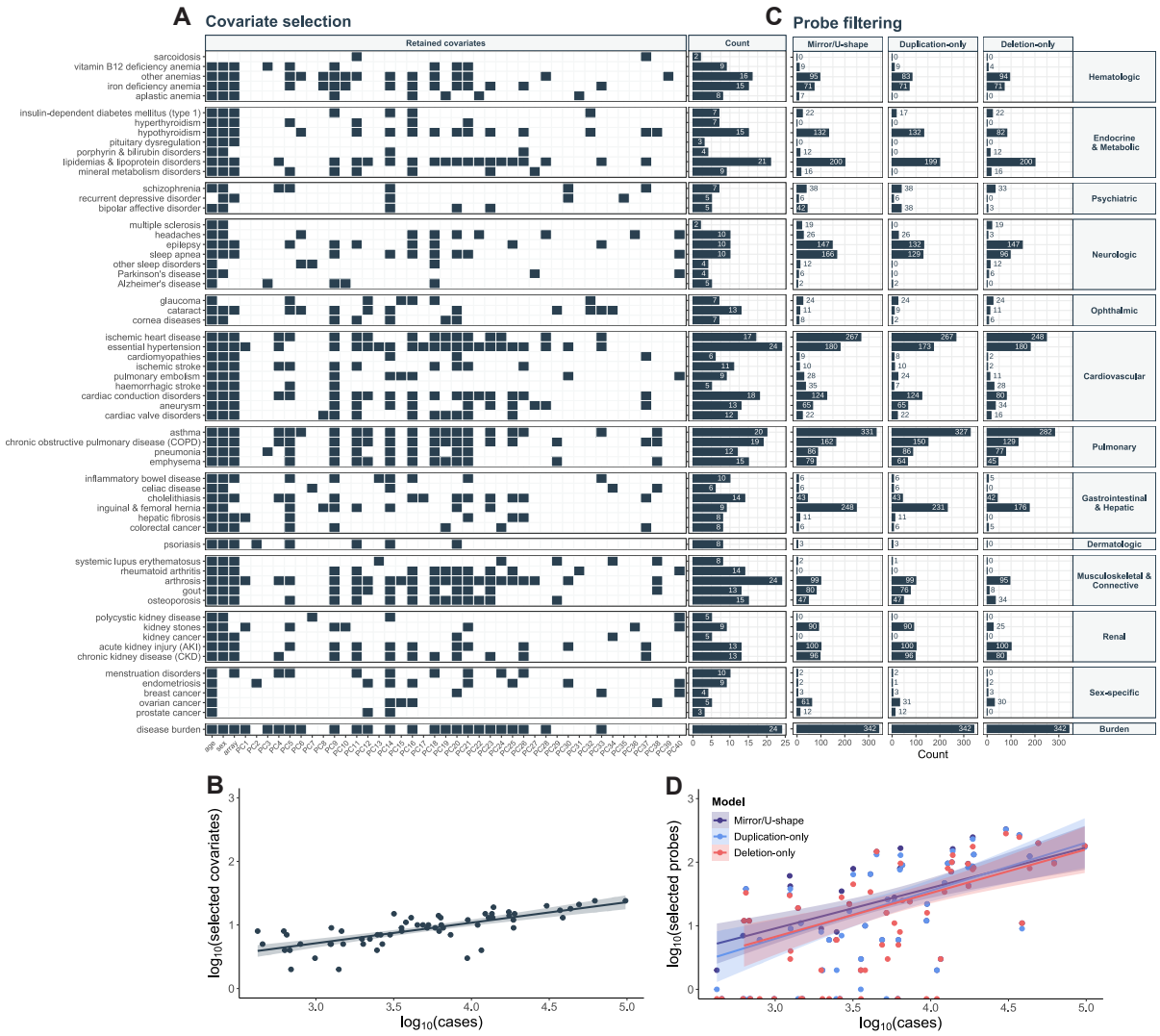


Figure 3.2: Covariate selection and probe filtering in the UK Biobank.
(**A**) Left: Dark gray tiles indicate covariates (x-axis) retained for the corresponding disease and/or disease burden (y-axis) ($p \leq 0.05$). PC = principal component. Right: Number of retained covariates per disease. (**B**) The logarithm of number of selected covariates (y-axis) against the logarithm of the number of cases (x-axis) for each of the 60 assessed diseases. Linear regression equation with 95% confidence intervals is shown. (**C**) Number of probes retained (x-axis) for the mirror and U-shape (left), duplication-only (middle), and deletion-only (right) models for each of the 60 investigated diseases and the disease burden (y-axis). (**D**) The logarithm of number of selected probes (y-axis) against the logarithm of the number of cases (x-axis) for each of the 60 assessed diseases, split by association model. Linear regression equations with 95% confidence intervals are shown.

and disease burden) (Figure 3.2A) and correlated with case number of the disease, aligned with the expected gain in power for more frequent diseases (Figure 3.2B). Covariates used for the main CNV-GWAS are listed in Table S3.2.

Probe-level CNV frequency was calculated as previously described (*208*). Briefly, for the 740,434 probes stored in $\text{PLINK}_{\text{CNV}}$, we counted the number of times a genotyped probe was found in a deleted (N_{DEL}), copy-neutral ($N_{non-CNV}$), and duplicated (N_{DUP}) state among a subset of 331,522 selected individuals (`--freqx` PLINK v1.9). We excluded 41,670 array-specific probes with genotype count missingness > 5%. For the remaining probes, we calculated the probe-level CNV, duplication, and deletion frequencies[1] , with $N_{CNV} = N_{DUP} + N_{DEL}$. Probes with a CNV frequency < 0.01% were excluded. The 70,631 remaining probes

1:
CNV frequency:

$$q_{CNV} = \frac{Num_{CNV}}{Num_{CNV}+Num_{non-CNV}}$$

Duplication frequency:

$$q_{DUP} = \frac{Num_{DUP}}{Num_{CNV}+Num_{non-CNV}}$$

Deletion frequency:

$$q_{DEL} = \frac{Num_{DEL}}{Num_{CNV}+Num_{non-CNV}}$$

were pruned at $r^2 > 0.9999$ in $PLINK_{CNV}$ (`--indep-pairwise 500 250 0.9999` PLINK v2.0), based on their CNV genotype, resulting in 18,725 probes. Pruning at such a high threshold will retain only a single probe at the core of a CNVR, where due to the recurrent nature of CNVs the correlation is extremely high. However, it will retain multiple probes around the CNV breakpoints (BPs), where we expect variability due to true biological variation or uncertainty of the CNV calling algorithm.

For each disease, 2-by-3 genotypic Fisher tests assessed dependence between disease status and probe copy-number (rows: control vs case; columns: deletion vs copy-neutral vs duplication; `--model fisher` PLINK v1.9; TEST column `GENO`). For each phenotype, quantile-quantile (QQ) plots of the Fisher's test p-value were generated. For the disease burden, p-values from linear regression were used instead. The genomic inflation factor, λ, was calculated as the median of the chi-squared test statistics derived from the Fisher test p-values divided by the expected median of the chi-squared distribution. Overall, there was no sign of strong p-value inflation (Figure 3.3A). λ values above 1.1 indicate genomic inflation, which can be caused by population structure, linkage disequilibrium, or polygenicity (*80*) and was observed only for six highly polygenic traits, with a maximum value of 1.33 for the disease burden. On the other hand, 42 traits exhibited λ values below 0.9. Deflated p-values can be caused by extremely rare variants. To verify this hypothesis, λ values were calculated anew, excluding probes with the 5-80% lowest CNV frequency (in incremental steps of 5%), to determine the impact of CNV frequency on genomic deflation (Figure 3.3B). We observed a trend of increasing λ values when excluding low-frequency probes, indicating that the deflation is caused by probes with low CNV frequency. Importantly, low λ values do not increase false positive rates. λ values are available in Table S3.3, along with minimal CNV frequency after probe exclusion.

Finally, probes with $p_{Fisher} \leq 0.001$ and a minimum of two disease cases among CNV, duplication, or deletion carriers were retained for assessment through the mirror/U-shaped, duplication-only, or deletion-only model, respectively. The number of probes retained across all models ranged from 0 (sarcoidosis, hyperthyroidism, pituitary dysregulation, rheumatoid arthritis, polycystic kidney disease, and kidney cancer) to 342 (disease burden) (Figure 3.2C) and correlated with case number, aligned with the expected gain in power for more frequent diseases (Figure 3.2D). Number of probes retained according to different models for the main CNV-GWAS is listed in Table S3.2. The rationale behind this pre-selection is to reduce computation time, as it is highly unlikely that a genotypic test with $p > 0.001$ would yield a genome-wide significant ($p \leq 7.5 \times 10^{-6}$) logistic regression p-value.

Figure 3.3: Genomic inflation factor of probe genotypic Fisher tests.
(**A**) QQ plots depicting the expected (y-axis) against observed (x-axis) negative logarithm of the genotypic Fisher test's p-values assessing the association strength between the copy-number status of 18,725 probes that passed the CNV frequency filter of ≥ 0.01% and pruning at $r^2 > 0.9999$ and the 60 diseases and disease burden. Data points are expected to follow the dark gray line (95% confidence interval as gray shaded area). Phenotypes are ordered by decreasing genomic inflation factor (λ), whose value is indicated in the top left corner. (**B**) Boxplots of λ values across all 61 phenotypes (y-axis) obtained when excluding an increasing percentage (0-80%) of probes with the lowest CNV frequency (x-axis). The red line indicates $\lambda = 1$ (i.e., no inflation).

Genome-wide significance threshold
Due to the recurrent nature of CNVs, the copy-number statuses of the 18,725 probes retained after frequency filter and pruning remain highly correlated and are thus not independent. Accounting for these 18,725 probes would result in an overly strict multiple-testing correction. Using an established protocol (*85*, *208*, *295*), we estimated the chromosome-level number of effective tests and summed them up, resulting in an estimate of N_{eff} = 6,633, setting the genome-wide (GW) threshold for significance at $p \leq 0.05/6{,}633 = 7.5 \times 10^{-6}$. This threshold is of the same order of magnitude as what others have estimated for disease CNV-GWASs (*229*). We also assessed the number of associations surviving an experiment-wide threshold for significance that further accounts for the 60 assessed diseases (plus the disease burden), defined as $p \leq 0.05/(6{,}633 \times 61) = 1.2 \times 10^{-7}$. Enrichment for tier 1 and 2 associations (see *Statistical confidence tiers*) among experiment-wide, as opposed to genome-wide, significant signals were assessed with a two-sided Fisher test.

Main CNV-GWAS model
Association between disease risk and copy-number of CNV-proxy probes was assessed through logistic regression with Firth fallback (`--glm firth-fallback omit-ref no-x-sex hide-covar --ci 0.95 --covar-variance-standardize` PLINK v2.0), using disease- and model-specific probes and covariates. Four association models were assessed: the mirror model assessed the additive effect of each additional copy ($PLINK_{CNV}$ file set); the U-shape model assessed a consistent effect of any deviation from the copy-neutral state ($PLINK_{CNV}$ file set, using the `hetonly` option in `--glm` PLINK v2.0); the duplication-only model ($PLINK_{DUP}$ file set) assessed the impact of a duplication while disregarding deletions; the deletion-only model ($PLINK_{DEL}$ file set) assessed the impact of a deletion while disregarding duplications.

Given the encoding of CNVs in PLINK (Table 3.1), we want to obtain the effect of carrying an additional "T" for the mirror (i.e., effect of each additional copy), duplication-only (i.e., effect of the duplication), and the deletion-only (i.e., effect of the deletion) models. PLINK selects the effect allele ("A1") as the minor allele, so that depending on the deletion and duplication frequencies, it will report the effect of "A" or "T". In the former case, odds ratios (OR) and their 95% confidence interval (CI) were harmonized to "T"[2]. Because we use the `hetonly` modifier for the U-shape model, PLINK systematically reports the effect of being "AT", i.e., copy-neutral. To instead obtain the effect of having a CNV, the same transformation as described above was applied to all probes. For the disease burden, which was assessed through linear regression, $\beta_T = -\beta_A$ was applied when PLINK reported the effect of "A". Similarly, the CI was multiplied by −1 and inverted, i.e., the lower bound becomes the upper bound and vice-versa.

2: $OR_T = \frac{1}{OR_A}$, with $95\%CI_T = e^{log(OR_T) \pm 1.96 \cdot SE_{log(OR_A)}}$

Because of the high correlation between the copy-number state of tested probes, it is important to determine the number of independent CNV-disease associations identified. Genome-wide significant associations ($p \leq 7.5 \times 10^{-6}$) were pruned at $r^2 > 0.8$ (`--indep-pairwise 3000 500 0.8` PLINK v2.0). As PLINK preferentially keeps probes with higher non-major allele frequencies, we inputted a scaled negative logarithm of association p-value as frequency (`--read-freq` PLINK v2.0) to instead prioritize probes with the strongest association p-value. For the U-

shape model, pruning was performed using custom code by extracting probes from $PLINK_{CNV}$ and re-coding them to match U-shape numerical encoding. Number of independent signals per disease was determined by stepwise conditional analysis. Briefly, for each disease and association model, the CNV genotype of the lead probe (i.e., probe with the most significant association p-value at each iteration; encoding numerically as Table 3.1) was included along selected covariates in the logistic regression model and association studies were conducted anew. This process was repeated iteratively, always including the next lead probe as an additional covariate, until no probes passed the genome-wide significant threshold. Characteristics of the most significant model (i.e., *best* model) are reported. The *main* model indicates which CNV type mainly drives the association, i.e., when associations were found through multiple models, priority was given to either the duplication-only or deletion-only models, otherwise to the model yielding the lowest p-value.

Due to the quantitative nature of the disease burden, the CNV-GWAS for that phenotype was based on linear regressions correcting for selected covariates (`--covar-variance-standardize --glm firth-fallback omit-ref no-x-sex hide-covar --ci 0.95` PLINK v2.0). Post-GWAS processing was performed as previously described (*208*).

CNV region definition and annotation
CNV region (CNVR) boundaries were defined by the most distant probes within ± 3 Mb and $r^2 \geq 0.5$ of each independent lead probe (`--show-tags --tag-kb 3000 --tag-r2 0.5` PLINK v1.9; U-shape model: custom code). When multiple disease-CNV associations mapped to overlapping (≥ 1 bp) genomic coordinates, the CNVRs were merged, resulting in 45 unique, non-overlapping, disease-associated CNVRs, whose boundaries are defined as the maximal CNVR. Manual inspection ensured substantial overlap between merged CNVRs. CNVRs were annotated with hg19 HGNC and ENSEMBL gene names using `annotate_variation.pl` from ANNOVAR (`--geneanno`). Number of genes mapping to a CNVR was calculated and set to zero for CNVRs with REGION not equaling `exonic`.

Statistical confidence tiers
Following primary assessment through logistic regression, three statistical approaches were implemented to gauge the robustness of the lead probe's association signal. First, we assessed *post hoc* the p-value of the 2-by-3 genotypic Fisher tests. Second, we transformed the binary disease status into a continuous variable by computing the response residuals of the logistic regression of disease status on disease-relevant covariates. This allowed to use linear regressions to estimate the effect of the CNV status, encoded according to all significantly associated models in the primary analysis, on disease risk. The model generating the lowest p-value for the CNV encoding is reported. Third, time-to-event analysis was used to assess whether CNVs influence the age of disease onset using Cox proportional hazards (CoxPH) models, the latter requiring an estimate for the age at last healthy measurement. For cases, the latter was defined as the age at disease diagnosis. For controls, age at last healthy measurement was defined by subtracting birth date from cutoff date (30/09/2021) and the resulting period was converted into years. CoxPH models were fitted including disease-relevant covariates and numerically encoded CNV genotype for either of the four association models as

predictors, using the `coxph()` function from the R `survival` package (*416*). The model with the lowest CNV genotype p-value is reported. CNV-disease associations were classified in confidence tiers depending on whether they were confirmed by three (tier 1), two (tier 2), or one (tier 3) of the above-described approaches at the arbitrary validation significance threshold of $p \leq 1 \times 10^{-4}$. Above-described validation strategies are not suitable for disease burden associations. As quantitative variables do not suffer from the same caveats as binary traits, we classified all disease burden associations as tier 1.

Literature-based supporting evidence
Using three literature-based approaches, we examined whether disease-associated CNVRs had previously been linked to relevant phenotypes. First, we investigated the colocalization of autosomal CNVRs with SNP-GWAS signals. GRCh38/hg38 lifted CNVR coordinates were inputted in the GWAS Catalog and associations ($p \leq 1 \times 10^{-7}$) relevant to the investigated disease (i.e., disease itself, synonyms, continuous proxies, or major risk factors) were identified through manual curation. Second, we overlapped OMIM morbid genes (i.e., linked to an OMIM disorder; `morbidmap.txt`) with disease-associated CNVRs. Through manual curation, we flagged OMIM genes associated with Mendelian disorders sharing clinical features with the common disease associated through CNV-GWAS. Third, we examined if implicated CNVRs overlapped regions at which CNVs were found to modulate continuous traits (*208*) or disease risk (*293*, *306*).

Replication in the Estonian Biobank

Disease cases and disease burden were defined using the same inclusion and exclusion criteria as for the UKBB, with the notable exceptions of excluding "Z12" (routine preventive screens for cancer) and "D22-23" (benign skin lesions) subcodes from the exclusion list of cancer traits as due to differences in recording practices, these were much more frequent in the EstBB than in the UKBB, strongly reducing the number of controls. Furthermore, as there are no self-reported diagnoses available in the EstBB, the latter could not be used as an exclusion criterion for disease definition in the EstBB.

Autosomal CNVs were called from Illumina Global Screening Array genotype data for 193,844 individuals that survived general quality control and had i) matching genotype-phenotype identifiers, ii) matching inferred vs reported sex, iii) SNP-call rate ≥ 98%, iv) were of European ancestry (i.e., Europe (East), Europe (South West), Europe (North West), Finland, and Italy assignments from the `bigsnpr` R package function `snp_ancestry_summary()` (*417*)), and v) were included in the EstBB SNP imputation pipeline. CNV outlier samples based on genotyping plate or extreme CNV profile, as well as individuals reporting blood malignancies, were excluded, using the same criteria as for the UKBB. High confidence CNV calls (i.e., $|QS| > 0.5$) of the 156,254 remaining individuals were encoded into three PLINK binary file sets, following the procedure described for the UKBB (Table 3.1).

Related individuals were pruned (`--make-king-table --geno 0.05 --king-table-filter 0.0884 --maf 0.01` PLINK v2.0; kinship coefficient > 0.0884 corresponding to < 3rd degree relatedness), prioritizing

individuals whose disease status was least often missing, leaving 90,211 unrelated samples for the replication study. Disease-relevant covariates were selected among sex, year of birth, genotyping batch (1-11), and PC1-20. For each of the unique 68 autosomal CNVR-disease association signals identified in the UKBB, we identified EstBB probes that were overlapping the CNVR's genomic coordinates. Probes with an EstBB CNV, duplication, or deletion frequency $\geq 0.01\%$, were retained, depending on whether the mirror/U-shape, duplication-only, or deletion-only was the best UKBB model, respectively, and 11 signals were excluded due to null/low CNV frequency for all probes in the CNVR. Association studies were performed on the remaining probes using disease-specific covariates and the best UKBB model, following the previously described procedure. Probes for which the regression failed to converge were discarded, leading to the exclusion of 8 signals for which all regressions failed. Summary statistics of the EstBB probe with the closest genomic location to the lead UKBB probe were retrieved for the remaining 49 signals, setting the replication threshold for significance at $p \leq 0.05/49 = 1.0 \times 10^{-3}$. P-values were adjusted to account for directional concordance with UKBB effects by rewarding and penalizing the 35 directionally concordant and 14 non-concordant signals, respectively[3]. One-sided binomial tests (`binom.test()`) were used to assess enrichment of observed vs expected significant replications at various thresholds ($\alpha = 0.1$ to 0.005 by steps of 0.005), with the R function arguments: `x` the number of observed signals at α, `n` the number of testable signals (i.e., 49), and `p` the expected probability of signals meeting α (i.e., α).

3: Direction agreement: $p_{new} = \frac{p_{old}}{2}$; else: $p_{new} = 1 - \frac{p_{old}}{2}$

BMI confounding analysis

We sought to assess whether some of our associations might be driven by the CNVR's effect on body mass index (BMI). Average BMI (#21001) over available instances was used. For an association to be tested for possible confounding, we required that i) BMI significantly associated with disease risk ($p \leq 0.05/61 = 8.2 \times 10^{-4}$) in a model including all disease-specific covariates and ii) the CNV genotype of the lead probe encoded numerically according to the best model to significantly associate with BMI ($p \leq 0.05/73 = 6.8 \times 10^{-4}$) previously inverse normal transformed and corrected for age, age^2, sex, genotyping batch, and PC1-40. Twenty-five association signals matched these criteria and for them, we fitted a logistic regression (or linear regression for the disease burden) with disease status as outcome and lead probe encoded numerically according to the best model, disease-specific covariates, and BMI as predictors. Significant differences in CNV effect sizes upon BMI adjustment were assessed by a two-sided t-test and deemed significant at $p \leq 0.05/25 = 0.002$. Associations likely driven by BMI were defined as those for which the CNV effect p-value dropped below the GW significance threshold upon adjustment for BMI.

CNV region constraint analysis

Evolutionary constraint of genes overlapping disease-associated CNVRs, i.e., "disease genes," was assessed by comparing their probability of LoF intolerance (pLI), loss of function observed/expected upper bound fraction (LOEUF), probability of haploinsufficiency (pHaplo), and probability of triplosensitivity (pTriplo) scores to the ones of "background

genes" with a two-sided Wilcoxon rank-sum test. Background genes were identified by annotating ranges of one or multiple consecutive probes with CNV frequency ≥ 0.01% with ANNOVAR (hg19 HGNC gene names) and excluding disease genes. For pLI and LOEUF, all disease genes were considered together. For pHaplo and pTriplo, two disease gene groups were considered: genes overlapping CNVRs with at least one association through the duplication-only model and genes overlapping CNVRs with at least one association through the deletion-only model. As many CNVRs associated through both models, the analysis was repeated considering genes overlapping CNVRs with at least one association through the duplication-only and none through the deletion-only model and vice versa.

Extended phenotypic assessment

17q12 deletion

For time-to-event analysis, the same chronic kidney disease (CKD) definition as in the main analysis was used. Low-quality CNVs ($|QS| \leq 0.5$) were excluded from analyses. Time-to-event analysis was performed as previously described, modeling both 17q12 deletions and duplications in the same CoxPH model adjusted for sex, age, age^2, array, and PC1-40. Estimated glomerular filtration rate (eGFR) was calculated based on the CKD-EPI equation using #30700 (creatinine [μmol/L]), accounting for age, sex, and ancestry (*418*).

BRCA1 deletion

Medical history of female *BRCA1* deletion carriers is based on #41270 (*diagnosis – ICD10*) and age at diagnosis was calculated as previously described. Relevant and prevalent diagnoses were manually selected for display. For the hereditary breast and ovarian cancer (HBOC) prevalence and time-to-event analysis, we included the following ICD-10 diagnoses on the inclusion list[4], and used the same exclusion list as for ovarian cancer. Duplications and low-quality CNVs ($|QS| \leq 0.5$), as well as male individuals, were excluded from the analyses. Difference in prevalence was assessed with a two-sided Fisher test. Time-to-event analysis was performed as previously described to estimate the effect of the *BRCA1* deletion, using age, age^2, array, and PC1-40 as covariates.

4: **Inclusion list:**

- **C50:** malignant neoplasm of breast.
- **C53:** malignant neoplasm of cervix uteri.
- **C54:** malignant neoplasm of corpus uteri.
- **C55:** malignant neoplasm of uterus, part unspecified.
- **C56:** malignant neoplasm of ovary.
- **C57:** malignant neoplasm of other unspecified female genital organs.

LDLR deletion

Medical history of low-density lipoprotein (LDL) receptor (*LDLR*) deletion carriers is based on #41270 (*diagnosis – ICD10*) and age at diagnosis was calculated as previously described. Drug usage data originates from #20003 (*treatment/medication code*) (Table 3.3) and a minimum of three individuals was required for a code/drug to be displayed in figures.

For prevalence and time-to-event analysis, only E78.0 (pure hypercholesterolemia) was considered on the inclusion list; the same exclusion list as for lipidemia was used. Duplications and low-quality CNVs ($|QS| \leq 0.5$) were excluded from analyses. Difference in prevalence was assessed with a two-sided Fisher test. Time-to-event analysis was performed as previously described to estimate the effect of the *LDLR* deletion, using sex, age, age^2, array, and PC1-40 as covariates.

LDL cholesterol measurements were available for seven *LDLR* deletion carriers in #42040 (*GP clinical event records*). LDL levels of earliest measurement on record (primary care) were compared to LDL levels from

Table 3.3: Hypolipidemic agents and antihypertensive/antianginal drugs. List of considered hypolipidemic agents and antihypertensive/antianginal drugs from #20003, generated based on drug.com (accessed: 29/09/2022). UK Biobank encoding is provided in the last column. * = sachet powder.

Category	Description	UKBB code
statins	atorvastatin	1141146234
	lipitor (10mg tablet)	1141146138
	fluvastatin	1140888594
	lescol (20mg capsule)	1140864592
	pravastatin	1140888648
	rosuvastatin	1141192410
	crestor (10mg tablet)	1141192414
	simvastatin	1140861958
	zocor (10mg tablet)	1140881748
	zocor heart-pro (10mg tablet)	1141200040
	eptastatin	1140910632
	velastatin	1140910654
cholesterol absorption inhibitors	ezetimibe	1141192736
	ezetrol (10mg tablet)	1141192740
fibrates	fenofibrate	1140861954
	gemfibrozil	1140861856
	gemfibrozil product	1141157262
	lopid 300 (capsule)	1140861858
	clofibrate	1140861944
	bezafibrate product	1141157260
	bezafibrate	1140861924
	bezalip (200mg tablet)	1140861926
	bezalip-mono (400mg m/r tablet)	1140861928
bile acid sequestrants	cholestyramine+aspartame (4g*)	1140861942
	cholestyramine	1140865576
	cholestyramine product	1141157416
	questran (4g*)	1140861936
	colestipol	1140888590
	colestid (5g/sachet granules)	1140861848
cardioselective beta-blocker	atenolol	1140866738
	bisoprolol	1140879760
	cardicor (1.25mg tablet)	1141171152
ACE inhibitor	ramipril	1140860806
	perindopril	1140888560
	lisinopril	1140860696

standardized blood biochemistry measurement (#30780) taken at assessment (#53) using a one-sided paired t-test. P12 was excluded as blood biochemistry LDL levels preluded the first primary care measurement. Based on #42039 (*GP prescription records*), P5 and P13 were identified as being prescribed statins by their general practitioner despite no record of statin usage in #20003.

Subgrouping of CNV carriers

When analyzing complex CNVRs, CNV carriers were split into subgroups based on visual inspection of breakpoints and segmental duplications overlapping the region. Following criteria were used to define groups (Table 3.4). CNVs not matching any of the groups are referred to as "atypical" CNVs.

Comparisons between groups of CNV carriers and copy-neutral individuals always exclude low-quality CNV carriers ($|QS| \leq 0.5$) altogether. Differences in prevalence q[5] were assessed with a two-sided Fisher test. For continuous traits, comparisons were based on two-sided t-tests.

5: SE depicted in graph is calculated as:

$q = \frac{c}{n}$, with $SE(q) = \sqrt{\frac{q(1-q)}{n}}$

where c and n represent the number of cases and total number of individuals in a group.

Table 3.4: CNV carrier subgrouping.
Selection criteria for different CNV carrier subgroups considered for analyzed CNV regions. Minimum and maximum start and end positions reflect the range in which the CNV breakpoints are to be for a given CNV to be assigned to a subgroup. "-" indicates an open end. *For 16p13.11 cat5 CNVs, coordinates correspond to the coordinates of *ABCC6*. All positions are in GRCh37/hg19.

CNVR	group	chr	min. start [bp]	max. start [bp]	min. end [bp]	max. end [bp]
16p13.11	cat1	16	-	15,000,000	16,250,000	16,750,000
	cat 2	16	15,000,000	15,200,000	16,250,000	16,750,000
	cat 3	16	15,200,000	15,800,000	16,250,000	16,750,000
	cat 4	16	-	15,800,000	17,500,000	-
	cat 5*	16	16,242,785	-	-	16,317,379
15q13	BP4-5	15	30,250,000	31,250,000	32,300,000	33,100,000
	D-CHRNA7-BP5	15	31,700,000	32,300,000	32,300,000	33,100,000
22q11.2	LCR A-D	22	18,500,000	19,200,000	21,250,000	21,900,000
	LCR A-B	22	18,500,000	19,200,000	20,150,000	20,600,000
	LCR B-D	22	20,000,000	20,850,000	21,250,000	21,900,000
	LCR C-D	22	21,000,000	21,150,000	21,250,000	21,900,000

CNV burden analyses

CNV burden association studies
In the UKBB, individual-level CNV, duplication, and deletion burden were calculated as the number of Mb or genes affected by high-confidence ($|QS| > 0.5$) autosomal CNVs, duplications, and deletions, respectively, yielding six CNV burden metrics, as previously described (*208*). Variance explained by these six CNV burden metrics was estimated by fitting logistic or linear regressions predicting disease outcome or disease burden as a function of the CNV burden metric (without any covariates) and assessing the McFadden pseudo-R^2 or the adjusted R^2 of the regression, respectively. Association between the six CNV burden metrics and the 60 diseases (logistic regression) or the disease burden (linear regression) were assessed including disease-relevant covariates in the model. Accounting for the 61 evaluated traits, significance was defined at $p \leq 0.05/61 = 8.2 \times 10^{-4}$.

CNV burden association studies corrected for CNV-GWAS signals
For each disease, CNVs, duplications, and deletions overlapping (≥ 1 bp) a CNVR significantly associated with the disease of interest through CNV-GWAS were omitted from the CNV, duplication, and deletion burden calculations if the CNVR had been found to associate with the disease through the mirror/U-shape, duplication-only, or deletion-only model, respectively. Association studies were repeated as previously described using corrected burden values.

Partitioned CNV burden association studies
To determine which part of the genome was driving the associations between disease risk and the CNV burden, we defined five genomic partitions:

1. **CNVR partition:** 40 autosomal disease-associated CNVRs identified in this study. All CNVRs were considered for the CNV, duplication, and deletion burden, except for CNVRs yielding associations uniquely through the duplication-only or deletion-only models, which were considered only for the duplication and deletion burdens.

2. **GD partition:** 92 GDs curated by Crawford et al. (*293*). Duplication syndromes were considered for the duplication burden, deletion syndromes for the deletion burden, and all genomic disorders were considered for the CNV burden.
3. **R1 partition:** Intersect between the CNVR and GD partitions, encompassing nine disease-associated CNVRs and 20 GDs caused by 10 reciprocal CNVs.
4. **R2 partition:** 72 GDs not included in the R1 partition.
5. **R3 partition:** 31 autosomal CNVRs not included in the R1 partition.

For every individual, we identified and summed up the subset of CNVs, duplications, and deletions (measured in number of genes or number of Mb) that overlaps these partitions (i.e., *subset* burden). Overlaps were defined either as i) any overlap ($\geq$ 1 bp) with the regions defined by the partition, or more stringently, ii) by reciprocal 50% bp overlap (i.e., the CNV covers > 50% of the partition's region and the partition's region covers > 50% of the CNV). The subset burden was subtracted from the total burden (i.e., *corrected* burden). Association studies were repeated as previously described using subset and corrected burden metrics.

Results

The spectrum of common diseases in the UK Biobank

Sixty disorders spanning 12 ICD-10 chapters were selected to cover a wide range of physiological systems, favoring conditions with sufficiently large sample size and a likely genetic basis (Figure 3.4; Figure 3.6; Table S3.1). We used a three-step approach to designate cases and controls in the UKBB (Figure 3.4A; top): Starting from 331,522 unrelated white British individuals, we defined cases based on a narrow list of hospital-based diagnoses (i.e., ICD-10 codes) and excluded self-reported cases, as well as self-reported and hospital diagnoses of related conditions from controls. Except for systemic lupus erythematosus (N = 422) and polycystic kidney disease (N = 454), all diseases had over 500 cases. Nineteen diseases had a case count > 10,000, with arthrosis (N = 62,175) and essential hypertension (N = 97,860) being the most frequent. Seven diseases had a median age of onset $\leq$ 60 years, predominantly female reproductive disorders, autoimmune conditions, and psychiatric diseases. Conversely, the nine diseases with a median age of onset $\geq$ 70 years were mainly degenerative disorders of the brain, eye, and kidney, overall aligning with epidemiological knowledge of the respective diseases (Figure 3.4B).

Copy-number variant genome-wide association study

To assess whether disease susceptibility is modulated by CNVs, we performed CNV-GWASs, i.e., test if the copy-number of CNV-proxy probes influences the probability of developing a disease or an individual's disease burden (Figure 3.4A; middle). Briefly, microarray-called CNVs for 331,522 unrelated white British individuals were transformed to the probe level after quality control (*208*). To further reduce the number and complexity of implemented logistic regressions, pre-processing steps selected relevant covariates and probes for each disease and model combination, thereby lowering computation time (Tables S3.2-3). As CNVs

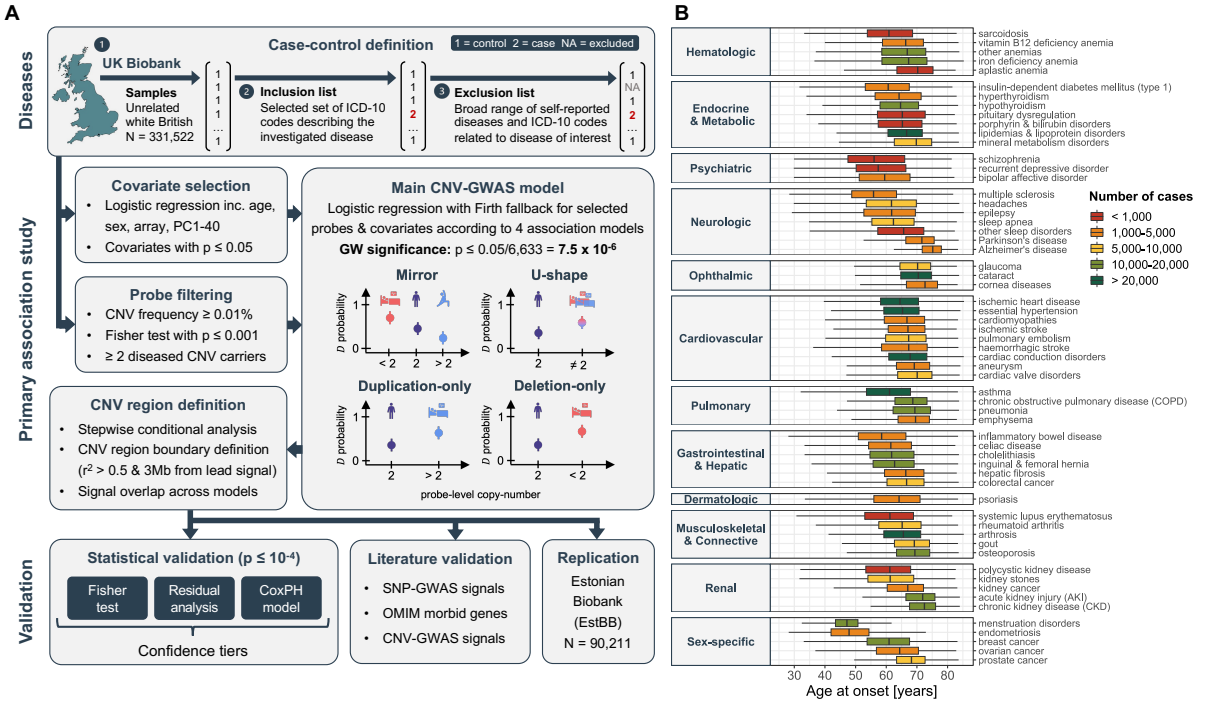


Figure 3.4: Study overview.
(**A**) Schematic representation of the analysis workflow. Diseases: For each of the 60 investigated diseases, 331,552 unrelated white British individuals were divided into three subsets: controls (encoded as 1; step 1), cases with the disease (encoded as 2; step 2), and a subset of individuals who were excluded because they had conditions similar but not identical to the disease (encoded as NA; step 3). Primary association study: Disease-specific relevant covariates were selected. Probes were prefiltered based on CNV frequency, Fisher test association p-value, and presence of ≥ 2 diseased carriers. Disease- and model-specific covariates and probes were used to generate tailored genome-wide CNV association scans (CNV-GWASs) based on Firth fallback logistic regression according to a mirror, U-shape, duplication-only (i.e., considering only duplications), and deletion-only (i.e., considering only deletions) models. Independent lead signals were identified through stepwise conditional analysis and CNV regions were defined based on probe correlation and merged across models. Validation: Statistical validation methods (i.e., Fisher test, residuals regression, and Cox proportional hazards model (CoxPH)) were used to rank associations in confidence tiers. Literature validation approaches leverage data from independent studies to corroborate that genetic perturbation (single-nucleotide polymorphisms (SNP), rare variants from the OMIM database, or CNVs) in the region are linked to the disease. Independent replication in the Estonian Biobank. (**B**) Age of onset for the 60 assessed diseases, grouped based on ICD-10 chapters and colored according to case count. Data are represented as boxplots; outliers are not shown.

can act through different gene dosage mechanisms, four association models were assessed: mirror and U-shape models consider deletions and duplications simultaneously, assuming that they impact disease risk in opposite or identical directions, respectively, while the CNV type-specific duplication- and deletion-only models assess independently the effect of duplications and deletions, respectively. All summary statistics are publicly available (GCST90297568-GCST90297771).

Stepwise conditional analysis narrowed GW significant associations ($p \leq 7.5 \times 10^{-6}$) to 40, 41, 21, and 38 independent signals for the mirror, U-shape, duplication-only, and deletion-only models, respectively. These were combined into 70 risk-increasing (i.e., no disease-protecting CNV) disease associations and three disease burden associations that map to 45 unique, non-overlapping, disease-associated CNVRs (Figure 3.5; Table 3.5; Table S3.4), among which nine (20%) could be unambiguously linked to a known GD. Forty-five associations (45 out of 73 = 62%) were supported at GW significance by multiple models, the lowest p-value (i.e., *best* model) being obtained through the mirror, deletion-only, U-shape, and duplication-only models for 24, 23, 21, and 5 of the signals, respectively. No association was detected at GW significance by both

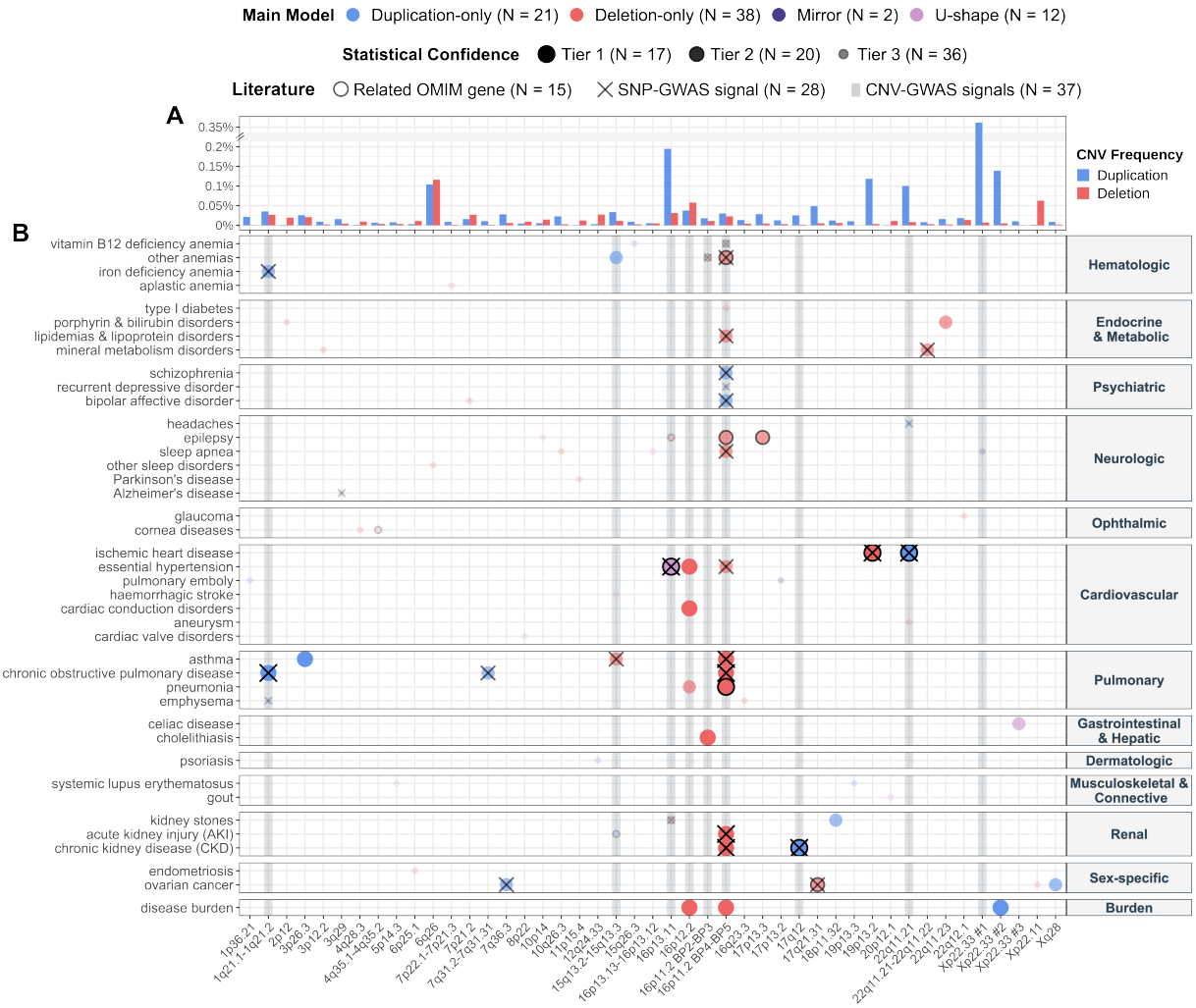


Figure 3.5: CNV-disease association map.
(**A**) Duplication and deletion frequencies (y-axis; break: //) of the lead probe for each unique and non-overlapping disease-associated CNVR, labeled with the corresponding cytogenic band (x-axis; 16p11.2 is split to distinguish BP2-3 and BP4-5 CNVRs; non-overlapping CNVRs on the same cytogenic band are numbered). If signals mapping to the same CNVR have different lead probes, the maximal frequency was plotted. **B** Associations between CNVRs (x-axis) and diseases (y-axis) identified through CNV-GWAS. Color indicates the main association model. Size and transparency reflect the statistical confidence tier. Black contours indicate overlap with an OMIM gene causing a disease with shared phenotypic features. Black crosses indicate overlap with SNP-GWAS signal for a related trait. Gray, shaded vertical lines indicate CNVRs with continuous trait associations (*208*). N provides count for various features.

the duplication-only and deletion-only models, so that each signal was attributed a *main* model that indicates whether the association is primarily driven by duplications or deletions (Figure 3.5). The main model should be interpreted with caution as both deletions and duplications might influence disease risk but only one CNV type-specific model might reach GW significance, due for instance to higher frequency of one CNV type. This is particularly relevant as 73% of the 45 disease-associated CNVRs have a higher duplication than deletion frequency (Figure 3.5A). Hence, 20 out of 21 (95%) signals mainly driven by duplications were also identified by the mirror/U-shape model(s), and contribution of deletions cannot be excluded.

Validation of identified CNV-GWAS signals

Across the 45 disease-associated CNVRs, CNV frequencies were low, ranging between 0.01% (our frequency cutoff) and 0.36%, with 87%

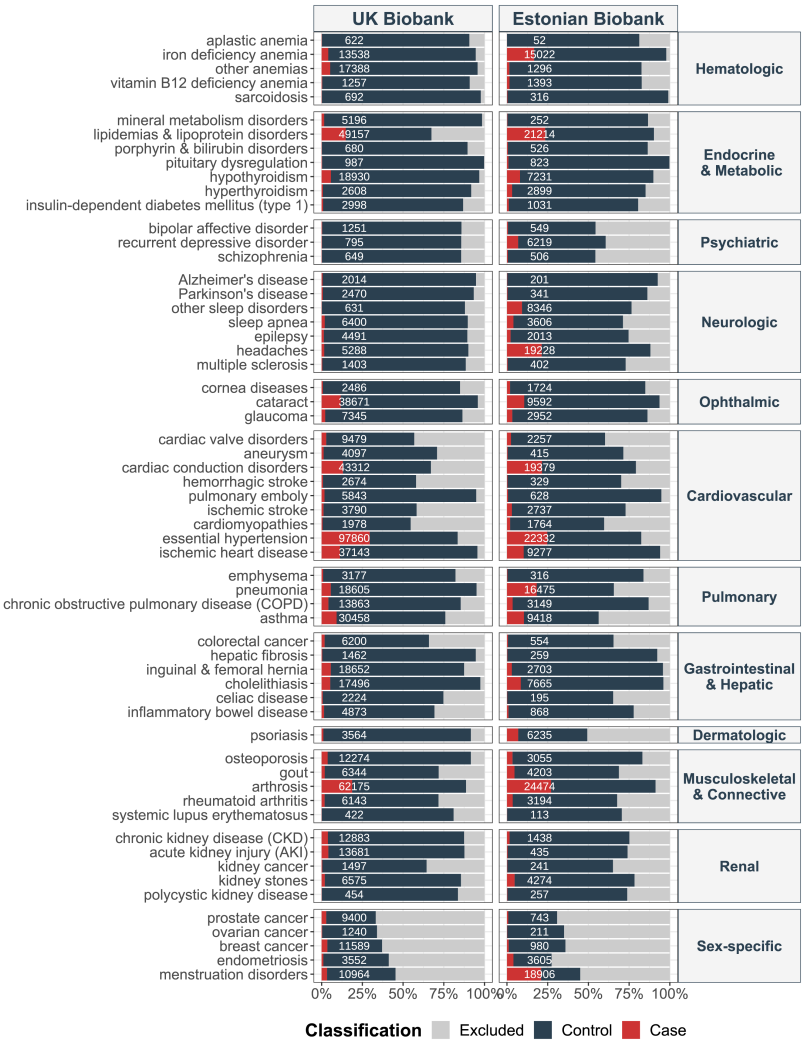


Figure 3.6: Case-control distribution in the UK and Estonian Biobanks. UKBB (left) and EstBB (right) percent stacked bar chart (x-axis) of cases (red), controls (dark gray), and excluded (gray) individuals, for each of the 60 assessed diseases (y-axis; left) categorized according to their ICD-10 chapter (y-axis; right). Case count is indicated in white.

(39 out of 45) of CNVRs having a frequency ≤ 0.1% (Figure 3.5A). Consequently, associations rely on a low number of diseased CNV carriers and require validation (Figure 3.4A; bottom; Figure 3.5B; Table S3.4). We used three statistical approaches to assess the robustness of CNV-disease associations: i) Fisher test, ii) residual regression, and iii) time-to-event analysis through CoxPH modeling. We replicated 28 (40%), 23 (33%), and 70 out of 70 (100%) of the associations with the respective methods at the arbitrary validation threshold of $p \leq 1 \times 10^{-4}$. This allowed to stratify associations in confidence tiers, with 17 signals replicating with all methods (tier 1), 20 with two (tier 2), and 36 only through time-to-event analysis (tier 3). Importantly, time-to-event analysis showed that CNVs always contributed to an earlier age of disease onset (Table S3.4), in line with the paradigm that diseases with a strong genetic etiology have earlier onset (*130*). Finally, when accounting for the number of assessed traits by using a stringent experiment-wide threshold for significance ($p \leq 1.2 \times 10^{-7}$), 32 out of 73 (44%) CNV-GWAS signals remained significantly associated. These signals were enriched for tier 1 and 2 associations ($p_{Fisher} = 0.05$).

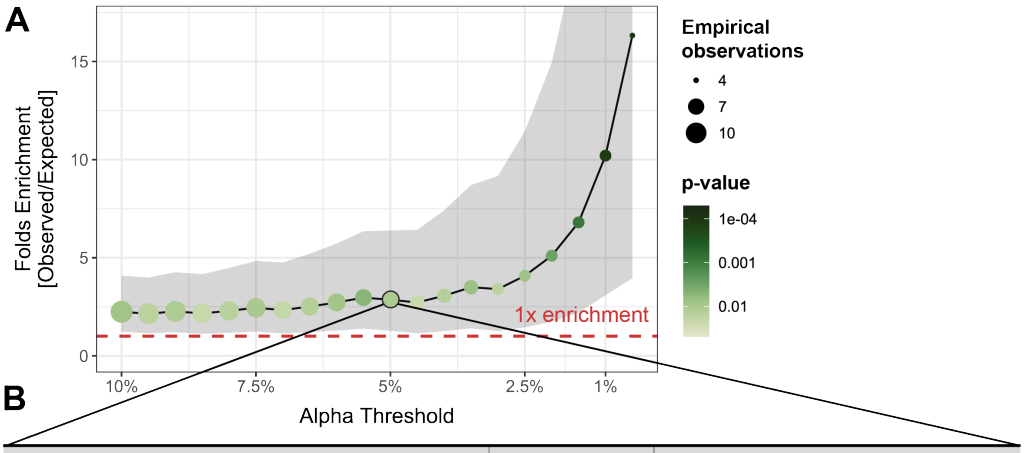
In parallel, we gathered literature evidence linking genetic variation at CNVRs with relevant phenotypes (Table S3.4). Forty-eight signals (48 out of 73 = 64%) mapped to a CNVR harboring a least one OMIM

Table 3.5: Forty-five disease-associated CNV regions.
Cytogenic band and genomic coordinates (GRCh37/hg19) of the 45 unique, non-overlapping, disease-associated CNVRs depicted on the x-axis of Figure 3.5. For each CNVR, length in kb is given ("Size"). GD indicates whether the CNVR matches any of the 92 genomic disorders (GD) compiled by Crawford et al. (*293*): "Y" = yes; *Partial overlap with the 22q11.2 distal CNVR (chr22:21,920,000–23,650,000). Disease associations mapping to that CNVR are listed, with bold font indicating that the association is likely mediated by increased body mass index (BMI). All the models through which the association was detected at genome-wide significance are indicated in superscript: "U" = U-shape; "+" = duplication-only; "-" = deletion-only ; "M" = mirror models.

Cytogenic band	Chr	CNVR start	CNVR end	Size [kb]	GD	Disease associations
1p36.21	1	12,854,105	13,038,285	184		pulmonary emboly $^{U,+,M}$
1q21.1-1q21.2	1	146,478,785	147,832,715	1,354	Y	chronic obstructive pulmonary disease $^{U,+}$, emphysema $^{+,M}$, iron deficiency anemia $^{U,+}$
2p12	2	78,376,475	78,680,202	304		Disorders of bilirubin metabolism $^{U,-,M}$
3p26.3	3	2,141,411	2,465,091	324		asthma $^{+,M}$
3p12.2	3	80,344,634	83,400,564	3,056		disorders of mineral metabolism $^{-,M}$
3q29	3	196,953,177	197,331,898	379	Y	Alzheimer's disease ^{U}
4q28.3	4	136,510,759	136,952,267	442		cornea diseases $^{U,-,M}$
4q35.1-4q35.2	4	186,687,554	187,182,384	495		cornea diseases ^{U}
5p14.3	5	20,254,182	20,924,403	670		systemic lupus erythematosus ^{U}
6p25.1	6	4,235,784	4,658,277	422		endometriosis $^{-,M}$
6q26	6	162,705,164	162,873,489	168		sleep disorders $^{-,M}$
7p22.1-7p21.3	7	7,260,027	7,504,011	244		aplastic anemia ^{U}
7p21.2	7	15,074,783	15,249,515	175		bipolar disorder $^{-,M}$
7q31.2-7q31.31	7	117,399,981	119,333,169	1,933		chronic obstructive pulmonary disease $^{U,+,M}$
7q36.3	7	158,530,132	158,953,160	423		ovarian cancer $^{U,+,M}$
8p22	8	17,599,136	17,719,930	121		cardiac valve disorders ^{U}
10p14	10	6,677,540	6,833,390	156		epilepsy ^{U}
10q26.3	10	135,217,002	135,237,176	20		sleep apnea $^{-}$
11p15.4	11	5,322,902	5,417,034	94		Parkinson's disease $^{U,-,M}$
12q24.33	12	131,611,538	131,825,359	214		psoriasis $^{+,M}$
15q13.2-15q13.3	15	30,912,719	32,516,949	1,604	Y	AKI $^{+}$, anemia $^{U,+}$, asthma $^{-,M}$, **hemorrhagic stroke** ^{U}
15q26.3	15	101,319,208	101,613,151	294		vitamin B12 anemia $^{U,+,M}$
16p13.13-16p13.12	16	12,516,765	12,659,427	143		sleep apnea ^{U}
16p13.11	16	15,120,501	16,353,166	1,233	Y	epilepsy $^{-}$, hypertension ^{U}, kidney stones $^{-}$
16p12.2	16	21,946,523	22,440,319	494	Y	AKI $^{+}$, anemia $^{U,+}$, asthma $^{-,M}$, hemorrhagic stroke ^{U}
16p11.2 BP2-BP3	16	28,775,159	29,043,450	268	Y	**anemia** $^{-}$, **cholelithiasis** $^{-,M}$
16p11.2 BP4-BP5	16	29,596,230	30,208,637	612	Y	AKI $^{U,-,M}$, **anemia** $^{U,-}$, **asthma** $^{-}$, bipolar disorder $^{U,+,M}$, chronic obstructive pulmonary disease $^{U,-,M}$, CKD $^{U,-}$, disease burden $^{U,-,M}$, **epilepsy** $^{-}$, **hypertension** $^{-}$, **lipidemias and lipoprotein disorders** $^{-}$, pneumonia $^{U,-,M}$, recurrent depressive disorder $^{U,+,M}$, schizophrenia $^{U,+,M}$, **sleep apnea** $^{-,M}$, **type I diabetes** $^{-}$, vitamin B12 anemia ^{U}
16q23.3	16	82,954,230	83,133,760	180		emphysema $^{-}$
17p13.3	17	631,380	738,187	107		epilepsy $^{-,M}$
17p13.2	17	4,378,105	4,498,641	121		pulmonary emboly $^{-}$
17q12	17	34,755,219	36,249,489	1,494	Y	CKD $^{U,+,M}$
17q21.31	17	41,197,733	41,276,111	78		ovarian cancer $^{-,M}$
18p11.32	18	685,968	1,266,259	580		kidney stones $^{U,+}$
19p13.3	19	6,873,527	6,881,286	8		systemic lupus erythematosus $^{U,+,M}$
19p13.2	19	11,210,904	11,218,188	7		ischemic heart disease $^{-}$
20p12.1	20	14,523,969	14,652,973	129		gout ^{U}
22q11.21	22	19,024,651	21,463,545	2,439	Y	aneurysm $^{-}$, **headaches** $^{+,M}$, **ischemic heart disease** $^{U,+,M}$
22q11.21-22q11.22	22	21,797,101	22,661,627	865	*	disorders of mineral metabolism $^{-,M}$
22q11.23	22	23,627,256	23,658,006	31	*	disorders of bilirubin metabolism $^{-,M}$
22q12.1	22	25,929,538	25,994,013	64		glaucoma $^{U,-}$
Xp22.33	X	1,746,850	2,046,202	299		sleep apnea ^{M}
Xp22.33	X	2,128,228	2,361,712	233		disease burden $^{U,+,M}$
Xp22.33	X	2,814,160	2,945,477	131		celiac disease ^{U}
Xp22.11	X	22,946,631	23,087,940	141		ovarian cancer $^{U,-,M}$
Xq28	X	152,703,776	152,887,811	184		ovarian cancer $^{U,+,M}$

morbid gene and in 15 cases, the gene was linked to a Mendelian disorder sharing phenotypic features with the associated common disease. For instance, the association between 4q35 CNVs and corneal conditions (chr4:186,687,554–187,182,384; $OR_{U\text{-}shape}$ = 18.2; 95%-CI [5.2; 63.1]; p = 5.0×10^{-6}) mapped to *CYP4V2* (MIM: 608614), a gene associated with autosomal recessive Bietti crystalline corneoretinal dystrophy (MIM: 210370), a disorder that impairs vision and progresses to blindness by age 50–60 years (*419*). We next assessed whether SNPs overlapping disease-associated CNVRs were reported to associate with the implicated disease or a biomarker thereof in the GWAS Catalog. This was the case for 28 (28 out of 66 = 42%) autosomal signals, a similar proportion (38%) than for continuous trait CNV-GWASs (*208*). For instance, distal 22q11.2 CNVs increased risk for disorders of mineral metabolism (chr22:21,797,101–22,661,627; OR_{mirror} = 0.02; 95%-CI [0.006; 0.083]; p = 9.9×10^{-9}) and overlapped heel bone mineral density SNP-GWASs signals, while 3q29 CNVs increased Alzheimer's disease risk (chr3:196,953,177–197,331,898; $OR_{U\text{-}shape}$ = 11.8; 95%-CI [4.0; 34.7]; p = 6.6×10^{-6}) and overlapped with SNP-GWAS signal for PHF-tau levels, and suggestive signals ($p < 5 \times 10^{-6}$) for frontotemporal dementia and cognitive decline in Alzheimer's disease. Finally, 37 signals (37 out of 73 = 51%) mapped to nine CNVRs previously found to be associated with complex traits (*208*), among which eight correspond to known GDs.

We also set out to replicate association signals in 90,211 unrelated EstBB individuals (*62*), using similar case definitions as in the UKBB analyses (Figure 3.6). A total of 49 out of 73 associations could be evaluated, among which three were strictly replicated ($p \leq 0.05$ out of 49 = 1.0×10^{-3}) and four additional ones reached nominal significance ($p \leq 0.05$) (Table S3.4). Compared to what would be expected by chance, this corresponds to a 2.9-fold ($p_{binomial}$ = 0.011) and 16.3-fold ($p_{binomial} = 1.1 \times 10^{-4}$) enrichment for replication at $p \leq 0.05$ and $p \leq 5 \times 10^{-3}$, respectively (Figure 3.7A). We have previously shown that the smaller sample size of the EstBB strongly limits replication power (*208*). Hence, despite only 7 out of 49 (14%) associations being nominally replicated, the strong enrichment for significant results supports the validity of the primary UKBB association signals. Replicated associations harbor SNP-GWAS signals for related phenotypes (5 out of 7), relevant morbid OMIM genes (2 out of 7), or map to CNVRs previously associated with similar diseases (5 out of 6) or biomarkers (4 out of 7) (Figure 3.7B). Among them is the association between 15q13 duplications and increased risk for acute kidney injury (AKI; chr15:30,946,160–31,881,106 | UKBB: OR_{DUP} = 4.6; 95%-CI [2.5; 8.4]; p = 7.1×10^{-7} | EstBB: p = 2.7×10^{-4}). Homozygous LoF mutations in *FAN1* (MIM: 613534), one of the five genes mapping to this CNVR, have been linked to karyomegalic interstitial nephritis (MIM: 614817) (*420*), opening the possibility that both increased and decreased dosage of this region have negative consequences on renal health. Importantly, integrating evidence provided by statistical, literature-based, or independent replication helps prioritize the most promising associations for follow-up studies and pinpoint plausible candidate genes.



		UKBB discovery					EstBB replication			Literature support			
	Disease	CNV Region	Model	OR	P	Tier	OR	$P_{one\text{-}side}$	Case/ CNV	SNP-GWAS	OMIM genes	CNV association	
												Crawford 2019	*Auwerx 2022*
Bonferroni	AKI	**16p11.2 BP4-5** (16:29,596,230-30,208,637)	DEL	16	5.6e-20	1	42	2.1e-4	1/10	✓		renal failure (DEL; p = 6e-5)	creatinine; cystatin C
	AKI	**15q13.2-15q13.3** (15:30,946,160-31,881,106)	DUP	4.6	7.1e-7	3	12	2.7e-4	2/40	✗	*FAN1* (KIN)	✗	*
	CKD	**17q12** (17:34,755,219-36,249,489)	U	6.5	5.9e-9	1	18	8.6e-4	3/13	✓	*HNF1B* (RCAD)	renal failure (DUP; p = 3e-4)	creatinine; urea; cystatin C
Nominal significance	HTN	**16p11.2 BP4-5** (16:29,596,230-30,208,637)	DEL	4.6	3.4e-7	2	5.0	4.0e-3	5/10	✓		hypertension (DEL; p = 9e-6)	*
	asthma	**15q13.2-15q13.3** (15:30,912,719-32,516,949)	M	0.17	1.2e-6	2	0.29	6.2e-3	7/40	✓		asthma (DEL; p = 2e-4)	forced vital capacity
	CKD	**16p11.2 BP4-5** (16:29,596,230-30,208,637)	U	4.4	2.7e-8	1	8.4	0.033	1/13	✓		renal failure (DEL; p = 6e-5)	creatinine; cystatin C
	PD	**11p15.4** (11:5,322,902-5,417,034)	M	0.05	1.9e-7	3	0.05	0.047	0/11	✗		not tested	✗

Figure 3.7: Replication of CNV-disease associations in the Estonian Biobank.
(**A**) Enrichment for signal replication (y-axis; 95% confidence interval as gray ribbon) at different levels of significance (x-axis) in EstBB. Color and size indicate the p-value of the enrichment (one-sided binomial test) and the number of observed associations, respectively. Dashed red line indicates one-fold enrichment, i.e., the number of observed associations matches the number of expected ones. (**B**) Associations replicated at nominal significance in the EstBB, color-stratified according to whether they meet the Bonferroni replication ($p \leq 1.0 \times 10^{-3}$; green) or nominal ($p \leq 0.05$; light green) significance threshold. Disease (CKD = chronic kidney disease; AKI = acute kidney injury; HTN = hypertension; PD = Parkinson's disease), cytogenic band and coordinates, best model (M = mirror; U = U-shape; DUP = duplication-only; DEL = deletion-only), odds ratio (OR), p-value (P), and statistical confidence tier are given for the UK Biobank (UKBB) discovery analysis. OR, one-sided p-values, and number of cases among CNV carriers are provided for the EstBB replication. Overlap with SNP-GWAS signals for a related trait (✓ = yes; ✗ = no) or a relevant OMIM gene (RCAD = renal cyst and diabetes; KIN = karyomegalic interstitial nephritis) is indicated. Previous association with diseases (*293*) (duplication (DUP) or deletion (DEL) was associated with indicated disease; no association (✗); some CNVRs were not tested) and continuous traits (*208*) (disease-relevant biomarkers are specified; other traits (*); no association(✗)) are listed.

CNV-disease associations driven by BMI

Large recurrent CNVs have been linked to altered body weight (*208*, *292*, *294*, *295*), which itself represents a risk factor for a broad range of common diseases. We identified 25 CNV-disease associations for which both disease risk and CNV status associated with BMI, indicating that the latter might confound these associations. While including BMI as an additional covariate did not result in significantly different CNV effects, 12 out of 25 associations did not meet the strict GW significance threshold anymore (Table 3.5; Figure 3.8; Table S3.5), so that 16% of the 73 associations uncovered by our CNV-GWAS are likely driven by the CNV's propensity for increasing adiposity in its carriers. In line with expectations, associations showing the strongest confounding include cardiometabolic diseases such as lipidemia, or sleep apnea, while pulmonary, renal, and psychiatric diseases, along with the disease burden were less affected. Importantly, only one CNVR lost all its associations upon BMI adjustment, i.e., the *SH2B1*-overlapping distal 16p11.2 BP2-3 deletion, which is known to cause severe, early-onset obesity (*325*, *412*).

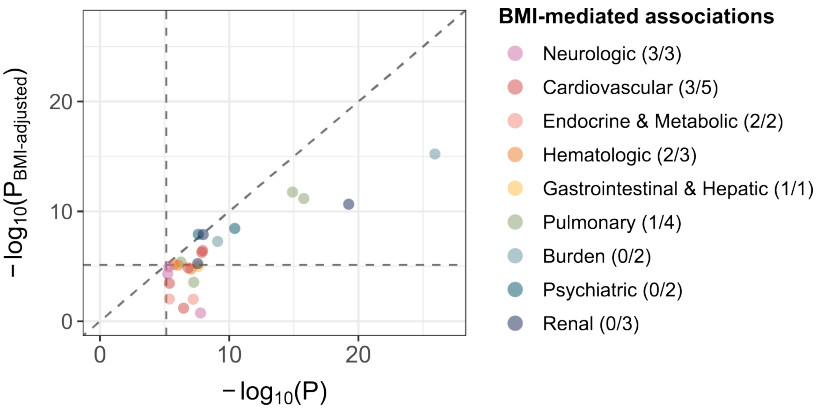


Figure 3.8: BMI adjustment for possibly confounded CNV-disease associations. Negative logarithm of CNV-disease association p-value with (y-axis) and without (x-axis) adjustment for body mass index (BMI) for the 25 CNV-GWAS signals potentially confounded by the latter. The horizontal and vertical dashed lines represent the genome-wide significance threshold at $p \leq 7.5 \times 10^{-6}$; the diagonal dashed line represents the identity line. Associations are colored by ICD-10 chapter, with the number of associations that fail to reach genome-wide significance upon adjustment for BMI indicated in parenthesis.

Global characterization of disease-associated CNV regions

We sought to identify global characteristics that distinguish disease-associated CNVRs (Table S3.6). Number of protein-coding genes embedded in the 45 disease-associated CNVRs ranged from 0 to over 30 and generally correlated with the number of encompassed probes (ρ = 0.50; p = 4.2×10^{-4}; Figure 3.9A). Exceptions include single-gene CNVRs overlapping well-known pathogenic genes captured thanks to high probe coverage, such as *BRCA1* or *LDLR*. Seven CNVRs (16%) associated with multiple diseases, all of which mapped to known GD regions. One CNVR that stood out is the 600 kb 16p11.2 BP4-5 region (Figure 3.5B; Table 3.5). Originally identified as a major risk factor for autism, schizophrenia, developmental delay and intellectual disability, macro-/microcephaly, epilepsy, and obesity/underweight (*326*, *421*–*426*), we previously found the region to associate with 26 continuous complex traits (*208*). Here, we show that 16p11.2 BP4-5 deletions increase the risk of 12 diseases across multiple organ systems as well as the disease burden (+ 3 diseases/deletion; p = 1.2×10^{-26}), five of which, alongside the disease burden, remain significant upon adjustment for BMI (Table 3.5; Table S3.5). On the other hand, the region's duplication drove increased risk for psychiatric conditions (i.e., bipolar disorder, schizophrenia, and depression), in line with previous findings (*425*).

Next, we assessed whether disease genes were under stronger evolutionary constraint than genes affected by CNVs at the same frequency but not associated with any disease (i.e., "background genes"). Compared to background genes, the 231 disease genes had more constrained pLI ($p_{Wilcoxon} = 1.3 \times 10^{-4}$; Figure 3.9B) and LOEUF ($p_{Wilcoxon} = 1.9 \times 10^{-7}$; Figure 3.9C) scores, suggesting stronger intolerance to LoF mutations. Splitting CNVRs depending on whether they have at least one association through either the duplication-only or deletion-only model, we evaluated whether embedded disease genes were more susceptible to haploinsufficiency (Figure 3.9D) or triplosensitivity (Figure 3.9E). No significant difference in pHaplo scores were observed but genes overlapping regions whose duplication ($p_{Wilcoxon} = 9.0 \times 10^{-19}$) and deletion ($p_{Wilcoxon} = 1.0 \times 10^{-23}$) have been linked to diseases were more likely to be triplosensitive than background genes. Similar trends were observed considering genes overlapping CNVRs involved uniquely through the duplication-only and deletion-only models and not the other CNV type-specific model (Figure 3.9F-G). Overall, our results indicate that a CNVR's pathogenicity

is influenced both by the number and characteristics of affected genes, even though our study did not explore whether part of the observed phenotypic consequences is driven by disruption of regulatory regions (*227*).

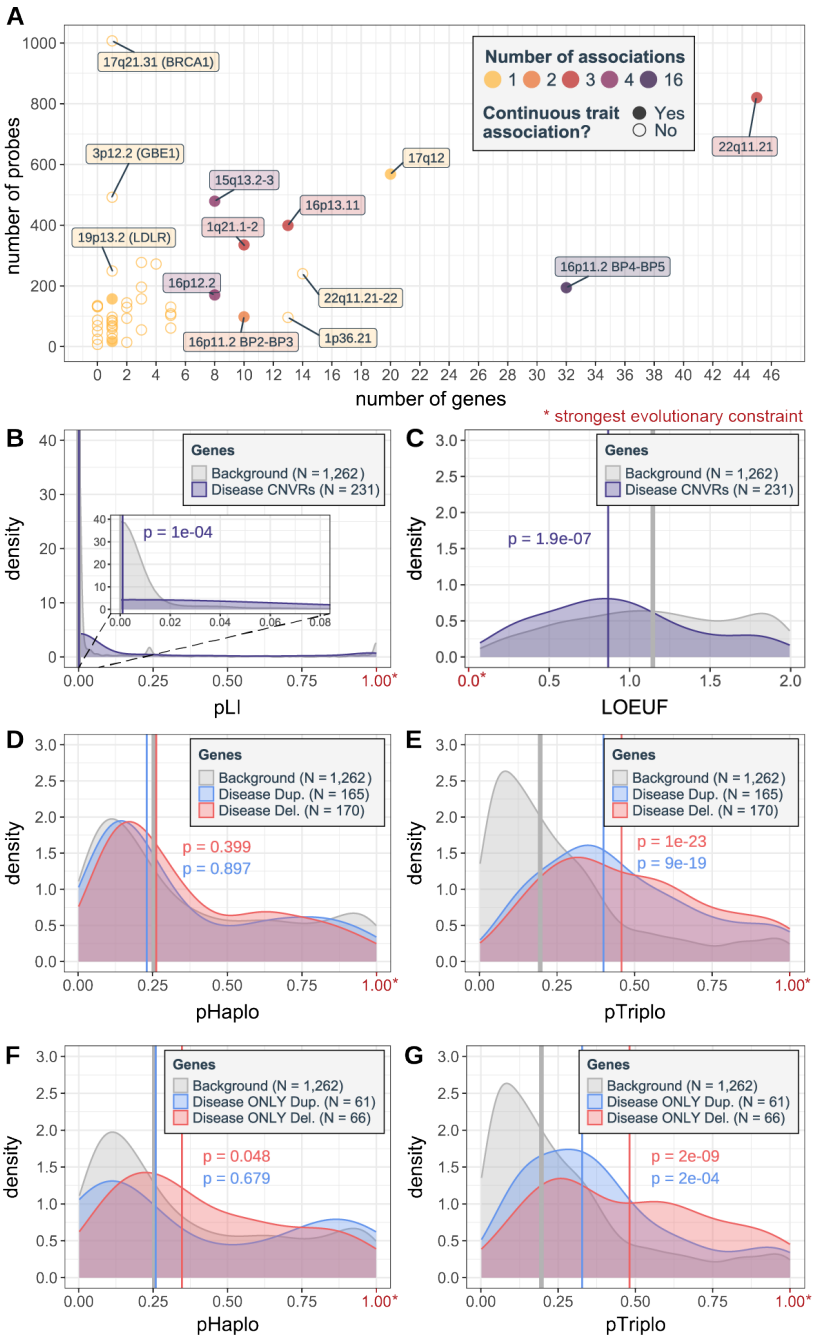


Figure 3.9: Constraint analysis of disease-associated CNV regions. (**A**) Number of probes (y-axis) vs number of affected genes (x-axis) for disease-associated CNVRs. Color reflects the number of associations, with full circles indicating previous association with continuous traits (*208*). CNVRs affecting ≥ 6 genes or single-gene CNVRs affecting > 200 probes are labeled with cytogenic bands. Evolutionary constraint of CNVR-encompassed genes (i.e., "disease genes"): Distribution of (**B**) pLI and (**C**) LOEUF scores for disease vs background genes (i.e., genes overlapping regions with a CNV frequency ≥ 0.01% but no disease association). Distribution of (**D**) pHaplo and (**E**) pTriplo scores for genes overlapping CNVRs significantly associated with a disease through the duplication-only or deletion-only models vs background genes. Distribution of probability of (**F**) pHaplo and (**G**) pTriplo scores for genes overlapping CNVRs uniquely associated to a disease through the duplication-only or deletion-only model vs background genes. Number of genes (N) and the median score (vertical line) are indicated for each group. P-values compare groups vs background gene medians (two-sided Wilcoxon test). Direction of the strongest evolutionary constraint is indicated in red with a star.

New insights in known disease genes

Two out of 12 female *BRCA1* deletion carriers were diagnosed with ovarian cancer (chr17:41,197,733-41,276,111; OR_{DEL} = 284.3; 95%-CI [24.6; 3290.8]; $p = 6.1 \times 10^{-6}$; Figure 3.10A). *BRCA1* (MIM: 113705) is a tumor suppressor gene whose LoF represents a major genetic risk factor for the development of HBOC (MIM: 604370) (*378*). Exploring the clinical records of the 12 deletion carriers, we found five diagnoses of breast cancer (a

trait assessed by CNV-GWAS but that did not yield a GW-significant association), one of endometrial cancer, and one of Fallopian tube cancer, so that eight carriers (67%) had received a HBOC diagnosis (Figure 3.10B). Not only was prevalence of HBOC higher among *BRCA1* deletion carriers (OR_{Fisher} = 31.0; p = 1.1×10^{-6}), but disease onset was earlier (HR = 17.0; p = 1.3×10^{-15}; Figure 3.10C). Among the four carriers with no HBOC, two had received cancer prophylactic surgery, *de facto* reducing the penetrance of the deletion. Surgeries were likely carried out based on family history of HBOC, which was reported for 6 carriers (50%), suggesting that these deletions are inherited. We did not observe higher prevalence of other cancer types (Figure 3.10B).

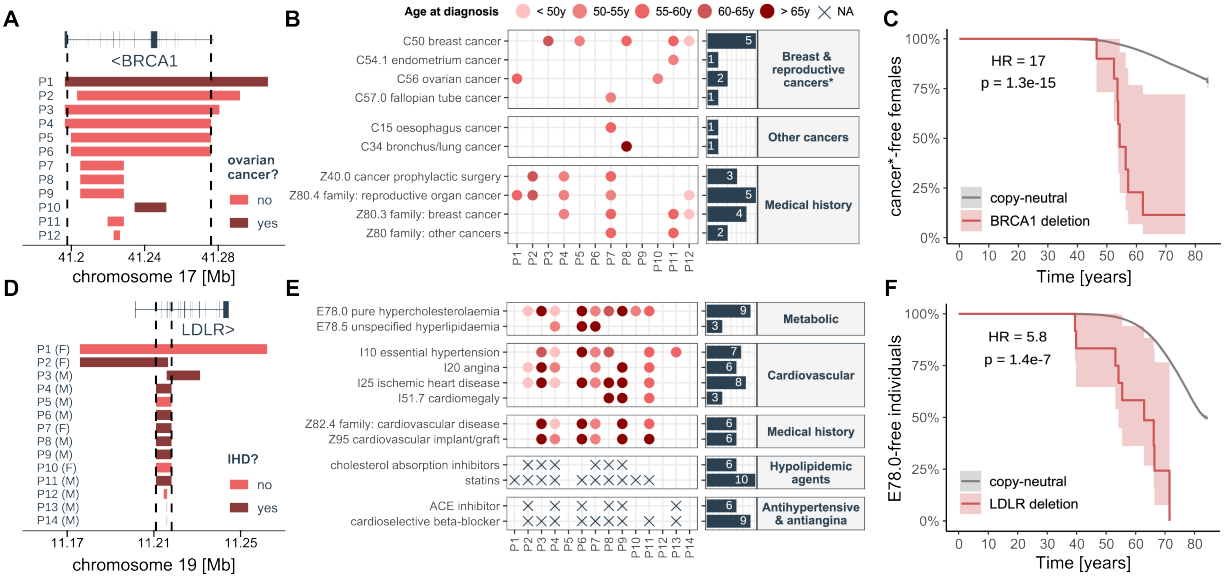


Figure 3.10: New insights into *BRCA1* and LDLR deletions.
(**A**) Genomic coordinates of the 12 females (P1-12; y-axis) carrying a *BRCA1* deletion (CNVR delimited by vertical dashed lines), colored according to ovarian cancer diagnosis. (**B**) Left: Cancer and related family/personal diagnoses received by individuals in (A). Color indicates age of diagnosis. Right: Counts per ICD-10 code. (**C**) Kaplan-Meier curve depicting the percentage, with 95% confidence interval, of females free of female-specific cancers over time among copy-neutral and *BRCA1* deletion carriers. Hazard ratio (HR) and p-value for the *BRCA1* deletion are given (CoxPH model). (**D**) Genomic coordinates of the 14 individuals (P1-14; y-axis) carrying an *LDLR* deletion (CNVR delimited by vertical dashed lines), colored according to ischemic heart disease (IHD) diagnosis. Sex of the individuals is indicated (M = male; F = female). (**E**) Left: Medical conditions and family/personal diagnoses and medication received by ≥ 3 *LDLR* deletion carriers in (D). Color indicates age of diagnosis. Right: Counts per ICD-10 code. (**F**) Kaplan-Meier curve depicting the percentage, with 95% confidence interval, of individuals free of pure hypercholesterolemia (E78.0) among copy-neutral and *LDLR* deletion carriers. HR and p-value for the *LDLR* deletion are given (CoxPH model).

High abundance of *Alu* repeats make the *LDLR* gene (MIM: 606945) susceptible to CNVs (244). We found that deletion of exon 2-6 increased risk for ischemic heart disease (chr19:11,210,904-11,218,188; OR_{DEL} = 31.2; 95%-CI [7.1; 137.8]; p = 5.6×10^{-6}) in a BMI-independent fashion. The condition was present in 8 of 14 deletion carriers (Figure 3.10D). Heterozygous - and less frequently homozygous – mutations in *LDLR* represent the main genetic etiology for familial hypercholesterolemia (243), which is characterized by elevated LDL cholesterol and predisposition for adverse cardiovascular outcomes (427). Previously identified in clinical studies of familial hypercholesterolemia (428), the CNVR implicated by our analysis specifically encompasses the ligand-binding domain of *LDLR* (243). Confirming widespread prevalence and family history (43%)

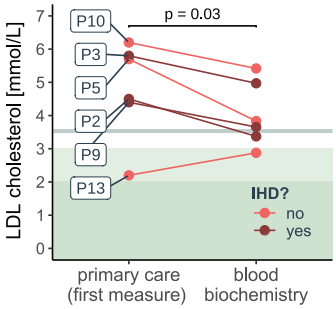


Figure 3.11: Statins mask genetic effect of *LDLR* deletion.
LDL-cholesterol levels (y-axis) from primary care data (first available measurement) and blood biochemistry (average over instances) for six deletion carriers (Figure 3.10D) with antecedent primary care LDL-cholesterol measurement, colored according to IHD diagnosis. Paired one-sided t-test p-value. Gray horizontal line represents the median blood biochemistry LDL value in non-carriers. Light and darker green backgrounds represent recommended target values for low (≤ 3 mmol/L) and high (≤ 1.8 mmol/L) risk individuals, respectively.

of cardiovascular diseases (Figure 3.10E), medical records of deletion carriers further revealed higher prevalence (OR_{Fisher} = 11.6; p = 7.9 × 10^{-5}) and earlier onset (HR = 5.8; p = 1.4 × 10^{-7}; Figure 3.10F) of pure hypercholesterolemia (E78.0), a code included in our lipidemia definition but that did not yield a signal picked-up by the CNV-GWAS. As we previously did not find the CNVR to associate with standardized blood biochemistry LDL levels (*208*), we hypothesized that the latter were lowered by hypolipidemic agents (Table 3.3). Ten (71%) deletion carriers were on statins and six (43%) were additionally using cholesterol absorption inhibitors, while the remaining four did not receive a dyslipidemia or ischemic heart disease diagnosis and harbored smaller deletions (i.e., P12-14; Figure 3.10D-E). We concluded that drugs likely masked genetically determined LDL levels, as shown by higher LDL levels in the first primary care measurement on record, measured prior to the standardized LDL measurement ($p_{t\text{-}test}$ = 0.03; Figure 3.11). Despite this, the recommended target of ≤ 1.8 mmol/L for high-risk individuals (*429*) was never met. By recovering known gene-disease pairs typically studied in clinical cohorts, we showcase how the rich phenotypic data from biobanks can generate insights into the mechanisms, epidemiology, and comorbidities of these diseases, implicating CNVs as important genetic risk factors.

CNV-biomarker associations tag pathophysiological processes

Integration of biomarker and disease CNV-GWAS signals can identify high-confidence, clinically relevant associations. Heterozygous LoF of *HNF1B* (MIM: 189907) and 17q12 deletions cause renal cyst and diabetes (RCAD) (MIM: 137920), a severe disorder characterized by renal abnormalities and maturity-onset diabetes of the young (*430*, *431*). While we previously showed that renal biomarkers were increased in duplication carriers (*208*), here, we demonstrate that both 17q12 deletions and duplications increase CKD risk (chr17:34,755,219–36,249,489; $OR_{U\text{-}shape}$ = 6.5; 95%-CI [3.4; 12.1]; p = 5.9 × 10^{-9}; Figure 3.12A), with a prevalence of 33.3% ($p_{t\text{-}test}$ = 0.026) and 16.9% ($p_{t\text{-}test}$ = 6.8 × 10^{-5}) among deletion and duplication carriers, respectively, versus 4.4% in copy-neutral individuals (Figure 3.12B). Results replicated in the EstBB (p = 8.6 × 10^{-4}; Figure 3.7B) and are supported by 20% of CNV carriers showing signs of kidney disease based on eGFR (< 60 ml/min/1.73m^2), compared to 2.2% in copy-neutral individuals (Figure 3.12C). Importantly, both 17q12 deletion and duplication lower age of CKD onset (HR ≥ 4.6; p ≤ 1.3 × 10^{-7}; Figure 3.12D), providing strong evidence of the deleterious consequences on kidney health of altered dosage of 17q12. These results align with two recent clinical studies that found that 17q12 deletions were observed in ~2% of individuals with congenital kidney anomalies (*277*) and that the 17q12 CNV was the most common GD etiology within a cohort of 6,679 CKD cases, in which nine deletion and seven duplication carriers were identified (*409*).

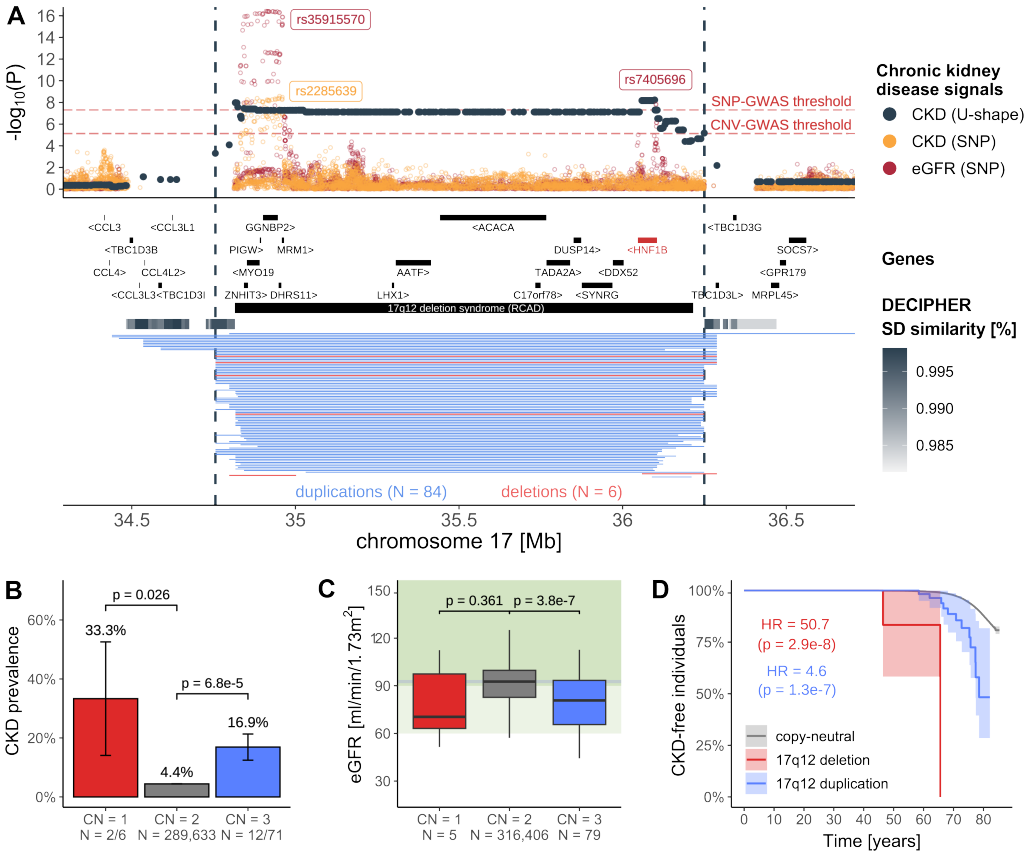


Figure 3.12: Increased and decreased dosage of 17q12 impairs kidney function.
(**A**) 17q12 association landscape. Top: Negative logarithm of association p-values of CNVs (dark gray; CNVR delimited by vertical dashed lines) and single-nucleotide polymorphisms (SNPs) with chronic kidney disease (CKD; orange) (*432*) and SNPs with estimated glomerular filtration rate (eGFR; red) (*433*). Lead SNPs are labeled. Red horizontal dashed lines represent the genome-wide threshold for significance for CNV-GWAS ($p \leq 7.5 \times 10^{-6}$) and SNP-GWAS ($p \leq 5 \times 10^{-8}$). Middle: Genomic coordinates of genes and DECIPHER GD, with *HNF1B*, the putative causal gene in red. Segmental duplications are represented as a gray gradient proportional to the degree of similarity. Bottom: Genomic coordinates of duplications (blue) and deletions (red) of UKBB participants overlapping the region. (**B**) CKD prevalence (± standard error) according to 17q12 copy-number (CN). P-values compare deletion (CN = 1) and duplication (CN = 3) carriers to copy-neutral (CN = 2) individuals (two-sided Fisher test). Number of cases and samples sizes are indicated (N = cases/sample size). (**C**) Boxplots of eGFR levels according to 17q12 CN; outliers are not shown. P-value comparisons as in (B) (two-sided t-test). Gray horizontal line represents median eGFR in non-carriers. Light and darker green backgrounds represent mildly decreased (60–90 ml/min/1.73m^2) and normal (≥ 90 ml/min/1.73 m^2) kidney function, respectively. (**D**) Kaplan–Meier curve depicting the percentage, with 95% confidence interval, of individuals free of CKD over time among copy-neutral and 17q12 deletion and duplication carriers. Hazard ratio (HR) and p-value for deletion and duplication are given (CoxPH model).

In another similar example, the blood pressure-increasing 16p12.2 deletion (chr16:21,946,523-22,440,319) (*208*, *294*) increased risk for hypertension (OR_{DEL} = 2.7; 95%-CI [1.9; 3.8]; $p = 1.3 \times 10^{-8}$) and cardiac conduction disorders (OR_{DEL} = 3.3; 95%-CI [2.2; 4.9]; $p = 1.1 \times 10^{-8}$), suggesting a role in cardiovascular health (Figure 3.13A-D). Primarily associated with developmental delay and intellectual disability (*434*, *435*) – proxied by decreased fluid intelligence ($p_{t\text{-test}} = 8.7 \times 10^{-5}$) and income ($p_{t\text{-test}} = 1.4 \times 10^{-12}$) in the UKBB (Figure 3.13E-F) – cardiac malformations are reported in ~38% of clinically ascertained cases (*436*). Among 193 UKBB deletion carriers, two (1%) had congenital insufficiency of the aortic valve (Q23.1), corresponding to a higher but not significantly different prevalence of cardiovascular malformations (Q20-28) than in copy-neutral individuals (OR_{Fisher} = 2.1; p = 0.251). The deletion also associated with increased risk for pneumonia (OR_{DEL} = 3.0; 95%-CI [1.9; 4.6]; $p = 5.4 \times 10^{-7}$) and decreased forced vital capacity (*208*) (Figure 3.13G-H) and peak expiratory flow (*294*).

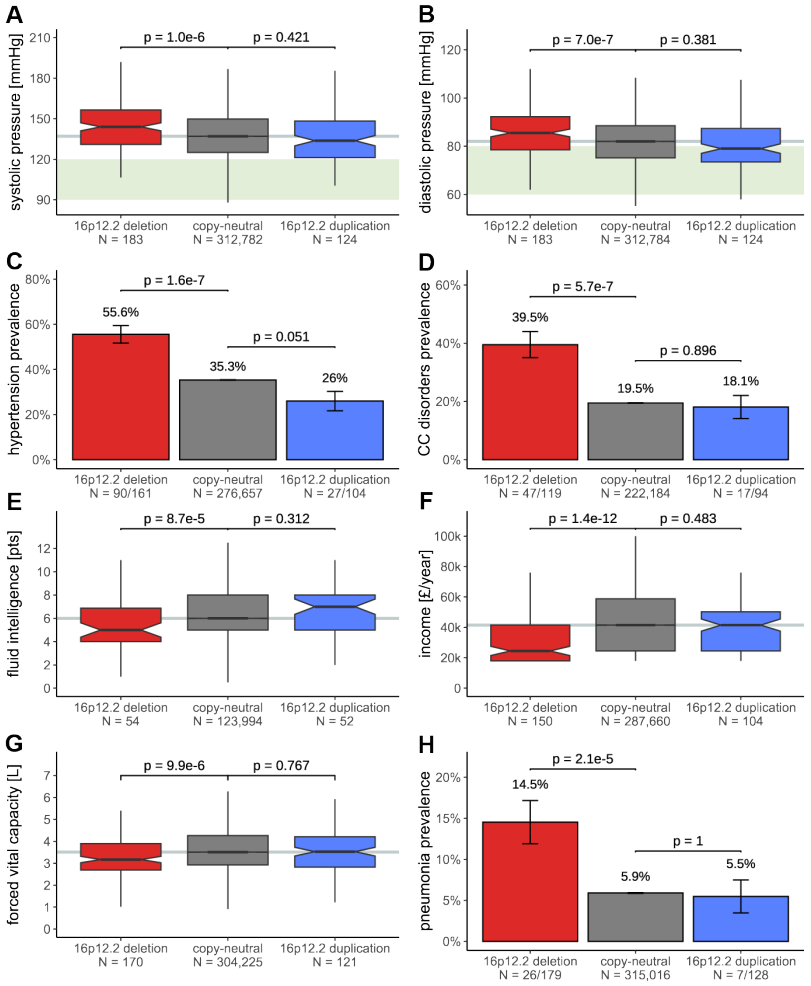


Figure 3.13: Cardiopulmonary phenotypes in 16p12.2 deletion carriers. Boxplots of (**A**) systolic (UKBB field #4080) and (**B**) diastolic (#4079) blood pressure according to 16p12.2 copy-number (CN). Green background represents optimal blood pressure (systolic: 90-120 mmHg; diastolic: 60-80 mmHg). Bar plots of (**C**) essential hypertension and (**D**) cardiac conduction (CC) disorders prevalence according to 16p12.2 CN. Boxplots of (**E**) fluid intelligence score (#20016; maximum = 13 points), (**F**) average yearly total household income before taxes (#738: ≤ £18k to £18k; £18k-30.9 to £24.5k; £31k-51.9 to £41.5k; £52k-100k to £76k; ≥ £100k to £100k), and (**G**) forced vital capacity (#3062) according to 16p12.2 CN, shown as boxplots. (**H**) Pneumonia prevalence according to 16p12.2 CN. For boxplots, outliers are not shown; p-values compare deletion and duplication carriers to copy-neutral individuals (two-sided t-test); gray horizontal line represents median among copy-neutral individuals; N indicates sample sizes. For bar plots, error bars represent ± the standard error; p-values compare prevalence among deletion and duplication carriers to the one in copy-neutral individuals (two-sided Fisher test); N indicates case count over sample size.

Dissecting complex pleiotropic CNV regions

While some CNV signals converge onto the same underlying physiological processes, others tie apparently unrelated traits to the same genetic region, suggesting genuine pleiotropy. 16p13.11 harbors multiple, partially overlapping recurrent groups of CNVs that allow fine-mapping of signals to different subregions of the CNVR (Figure 3.14). Through different association models, the CNVR was linked to uncorrelated traits including epilepsy, kidney stones, hypertension, alkaline phosphatase (ALP), forced vital capacity, and age at menopause and menarche. We previously proposed *MARF1* as a candidate gene for the female reproductive phenotypes (*208*) and will focus here on the remaining traits.

The 654 duplications and 355 deletions overlapping the maximal CNVR (chr16:15,070,916–16,353,166) were grouped into 5 categories (cat1-5) based on their breakpoints (Figure 3.14A). Matching previous findings (*410*), risk for epilepsy was increased in deletion carriers (chr16:15,122,801–16,353,166; OR_{DEL} = 6.2; 95%-CI [2.8; 13.4]; p = 4.4×10^{-6}; Figure 3.14B), with a prevalence of 8.2% among cat1-4 deletion carriers compared to less than 1.5% among copy-neutral and duplication carriers (Figure 3.14C). Previously associated with epilepsy in clinical cohorts (*389*, *440*, *441*), the region harbors *NDE1* (MIM: 609449), a gene associated with autosomal recessive lissencephaly (MIM: 614019) and microhydranencephaly (MIM: 605013)

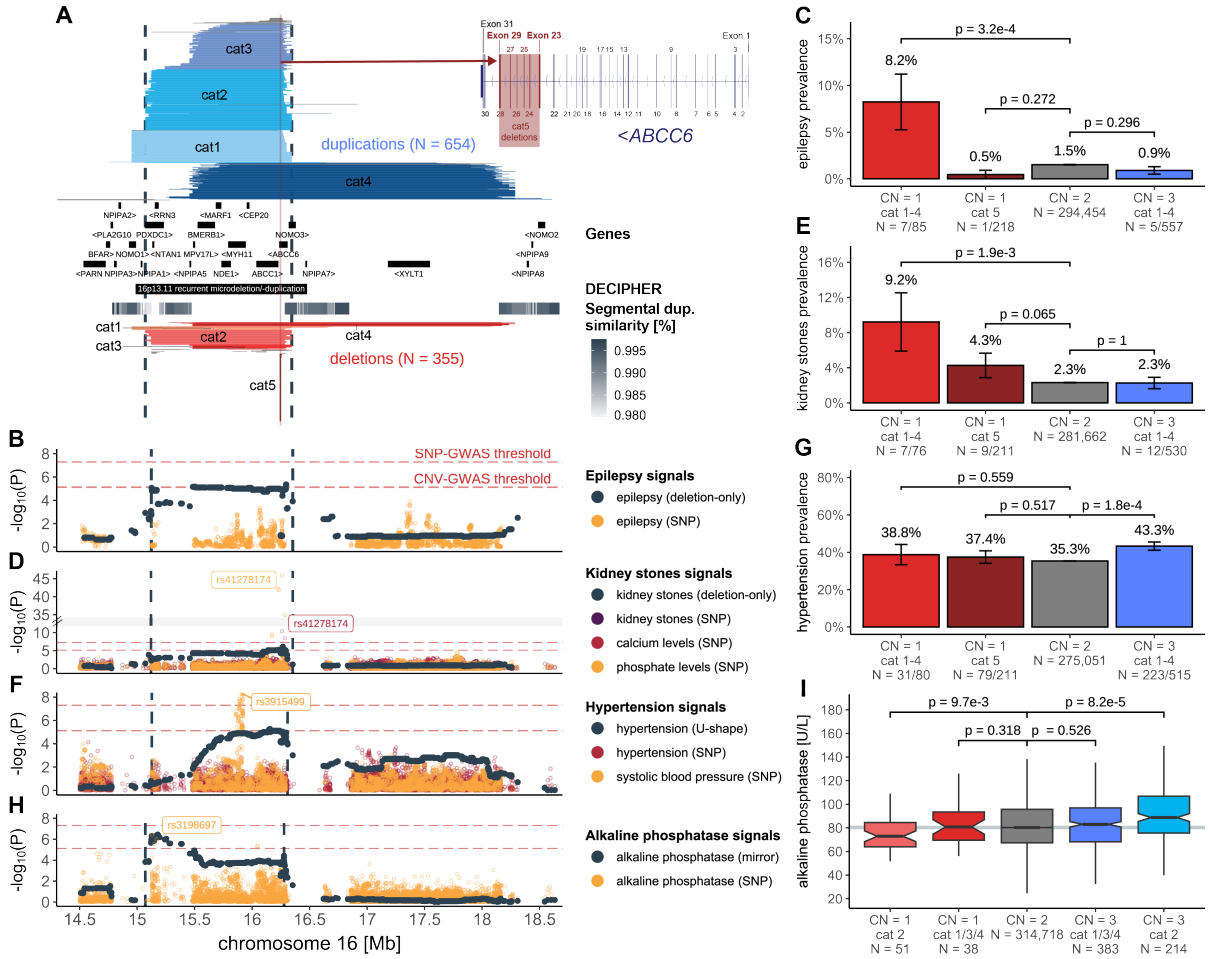


Figure 3.14: Dissection of complex pleiotropic patterns of recurrent CNVs at 16p13.11.
(**A**) 16p13.11 genetic landscape. Coordinates of UKBB duplications (shades of blue; top) and deletions (shades of red; bottom) overlapping the maximal CNVR (delimited by vertical dashed lines) associated with epilepsy, kidney stones, hypertension, and alkaline phosphatase (ALP). CNVs are divided and colored according to five categories (cat1-5) to reflect recurrent breakpoints, with atypical CNVs in gray (Table 3.4). Breakpoints reflect segmental duplications, represented with a gray gradient proportional to the degree of similarity. Middle: genomic coordinates of genes and DECIPHER GD. Inset: Overlap between *ABCC6*'s exonic structure and cat5 deletions. Negative logarithm of association p-values of CNVs (dark gray; model in parenthesis; CNVR delimited by vertical dashed lines) with (**B**) epilepsy, (**D**) kidney stones, (**F**) hypertension, and (**H**) ALP and SNPs with (B) epilepsy (*437*), (D) kidney stones (*438*), calcium and phosphate levels (y-axis; break: //); (F) hypertension and systolic blood pressure (*439*), and (H) ALP. Lead SNPs are labeled. Red horizontal dashed lines represent genome-wide thresholds for significance for CNV-GWAS ($p \leq 7.5 \times 10^{-6}$) and SNP-GWAS ($p \leq 5 \times 10^{-8}$). Prevalence (± standard error) of (**C**) epilepsy, (**E**) kidney stones, and (**G**) hypertension according to 16p13.11 copy-number (CN) and CNV categories from (A). P-values compare carriers of specific deletions (CN = 1) and duplications (CN = 3) to copy-neutral (CN = 2) individuals (two-sided Fisher test). Number of cases and sample sizes are indicated (N = cases/sample size). (**I**) ALP levels according to 16p13.11 CN and CNV category, are shown as boxplots; outliers are not shown. P-values compare carriers of specific deletions (CN = 1) and duplications (CN = 3) to copy-neutral (CN = 2) individuals (two-sided t-test). Gray horizontal line represents median ALP value in non-carriers.

and whose mutation has been linked to epilepsy (*442*, *443*). Deletions also increased risk for kidney stones (chr16:15,120,501–16,353,166; OR_{DEL} = 5.9; 95%-CI [2.9; 11.9]; $p = 7.3 \times 10^{-7}$), with the CNV-GWAS signal peaking close to a missense variant (rs41278174 G > A; $frequency_A$: 2.6%) in exon 23 of *ABCC6* (MIM: 603234) associating with calcium and phosphate levels through SNP-GWASs (Figure 3.14D). These signals coincide with the recurrent cat5 deletion that covers 29 probes spanning exons 23–29 of *ABCC6* (Figure 3.14A; inset). Kidney stones prevalence reaches 4.3% among cat5 deletion carriers, in-between estimates for larger cat1-4 deletion carriers (9.2%) and copy-neutral individuals (2.3%) (Figure 3.14E). A wide range of variants affecting *ABCC6* have been identified and linked to

the calcification disorder pseudoxanthoma elasticum through recessive (MIM: 264800) – and more rarely dominant (MIM: 177850) – inheritance (444–447), with the *Alu*-mediated cat5 deletion representing one of the most frequent variants (184, 185). *ABCC6* is expressed in the kidney and recent estimates from clinical cohorts suggested that kidney stones are an unrecognized (i.e., not used to establish clinical diagnosis) but prevalent (11–40%) feature of pseudoxanthoma elasticum (448–450). Our data support kidney stones as a clinical outcome of *ABCC6* disruption with partial gene deletions having lower penetrance than larger 16p13.11 deletions. Unlike epilepsy and kidney stones, both deletion (38.8%) and duplication (43.3%) carriers are at increased risk for hypertension (chr16:15,127,986–16,308,285; $OR_{U\text{-}shape} = 1.5$; 95%-CI [1.3; 1.8]; $p = 5.5 \times 10^{-6}$; Figure 3.14F), compared to copy-neutral individuals (35.3%) (Figure 3.14G). The CNVR overlaps a SNP-GWAS signal for systolic blood pressure mapping to *MYH11* (MIM: 160745) (Figure 3.14F). Expressed in arteries, *MYH11* encodes for smooth muscle myosin heavy chains and has been linked to dominant familial thoracic aortic aneurysm (MIM: 132900), for which hypertension represents a leading risk factor. Intermediate prevalence (37.4%) of hypertension among cat5 deletion carriers implicates *ABCC6*, suggesting that multiple genes might contribute to hypertension risk at 16p13.11. Consistent with this model, *ABCC6* plays a role in vascular calcification as the causal gene for generalized arterial calcification of infancy (MIM: 614473) (451, 452), typically diagnosed by hypertension in newborns. Interestingly, the previously described mirror association with ALP (chr16:15,070,916–16,276,964; $\beta_{mirror} = 6.6$ U/L; $p = 3.5 \times 10^{-7}$; UKBB field #30610) peaks at the distal end of the CNVR (208), nearby a suggestive SNP-GWAS signal for ALP levels (Figure 3.14H). Splitting ALP levels by CNV category revealed that this mirroring effect is driven by individuals with cat2 deletions (mean = 76.4 U/L; $p_{t\text{-}test} = 9.7 \times 10^{-3}$) and duplications (mean = 92.9 U/L; $p_{t\text{-}test} = 8.2 \times 10^{-5}$), as other CNV carriers had ALP levels indistinguishable from those of copy-neutral individuals (mean = 83.6 U/L) (Figure 3.14I). Hence, we propose the distal region of the CNVR to harbor the critical region regulating ALP levels, even though no obvious candidate gene could be identified in the literature.

The proximal 22q11.2 region, previously linked to DiGeorge (MIM: 188400) and velocardiofacial (MIM: 192430) syndromes, harbors four low-copy repeats (LCR A-D) (453). Building on evidence of complex association patterns within this CNVR (411), we report novel associations between CNVs spanning LCR A-D and ischemic heart disease (IHD; chr22:19,024,651-21,463,545; $OR_{U\text{-}shape} = 2.1$; 95%-CI [1.6; 2.8]; $p = 1.5 \times 10^{-7}$), LCR B-D and aneurysm (chr22:20,708,685-21,460,008; $OR_{DEL} = 41.8$; 95%-CI [10.0; 175.1]; $p = 3.2 \times 10^{-7}$), and LCR A-C and headaches (chr22:19,024,651-21,110,240; $OR_{mirror} = 3.7$; 95%-CI [2.1; 6.5]; $p = 4.8 \times 10^{-6}$) (Figure 3.15A). Based on three LCR B-D deletion carriers with aneurysm, this corresponds to a 22-times higher prevalence than in copy-neutral individuals (Figure 3.15B). Association with IHD is better powered, with a prevalence of 12%, 21%, 16%, and 20% among copy-neutral individuals and carriers of LCR C-D, B-D, and A-D CNVs, respectively (Figure 3.15C). Unlike the association with aneurysm, association with IHD was lost upon adjustment for BMI. This suggests that IHD risk is driven by increased adiposity which scales with the amount of affected genetic content, supporting the presence of multiple driver genes. Collectively,

our data indicate that altered 22q11.2 dosage can result in a spectrum of cardiovascular afflictions, ranging from well-described congenital malformation (*453*, *454*) to adult-onset aneurysm or IHD.

Another region exhibiting complex pleiotropic patterns is 15q13. Deletions spanning BP4-5 (MIM: 612001) – and to a lesser extent duplications – have been associated with neuropsychiatric and developmental conditions (*455*, *456*), with the nicotinic acetylcholine receptor ion channel *CHRNA7* being proposed as the driver gene based on the presence of similar phenotypes in individuals with a smaller deletion (D-CHRNA7-BP5) only affecting *CHRNA7* (*457*) (Figure 3.16A). BP4-5 duplication carriers – but not the ~10-times more numerous D-CHRNA7-BP5 duplication carriers – showed higher prevalence of AKI (EstBB-replicated: Figure 3.7B and Figure 3.16B), hemorrhagic stroke (chr15:30,912,719–31,982,408;

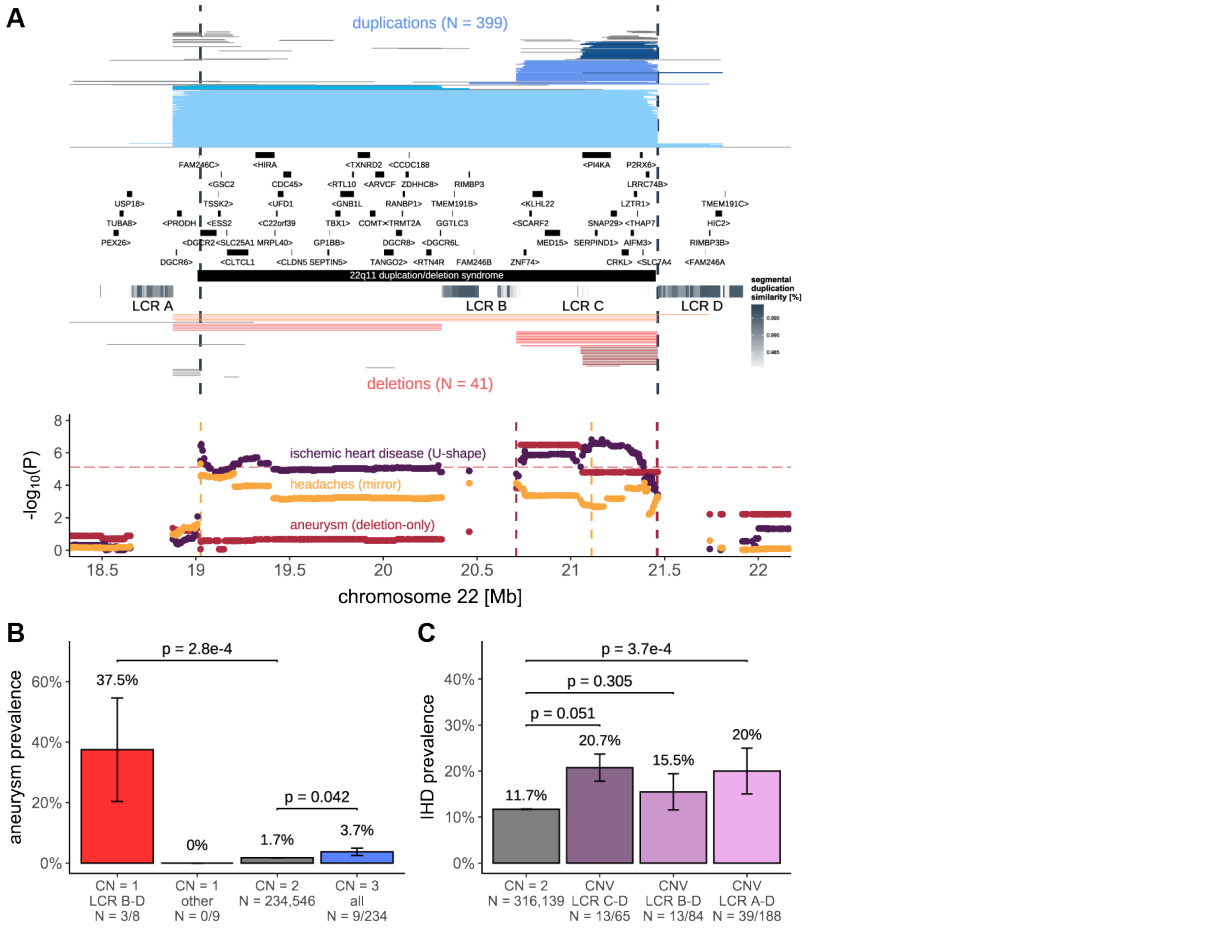


Figure 3.15: Dissection of complex pleiotropic patterns of recurrent CNVs at 22q11.2.
(**A**) 22q11.2 genetic landscape. Top: Coordinates of duplications (shades of blue; top) and deletions (shades of red; bottom) overlapping the maximal CNVR (delimited by vertical dashed lines) associated with ischemic heart disease (IHD), headaches, and aneurysm. CNVs are divided and colored according to four groups to reflect breakpoints at low-copy repeats (LCRs) spanning the region: A-D, A-B, B-D, C-D, with atypical CNVs in gray (Table 3.4). LCRs are composed of segmental duplications, represented as a gray gradient proportional to the degree of similarity. Genomic coordinates of genes and DECIPHER GD are displayed. Bottom: Negative logarithm of association p-values of CNVs (best model in parenthesis) with IHD, headaches, and aneurysm. Disease-specific CNVRs are shown with colored vertical dashed lines. Red horizontal dashed line represents the genome-wide threshold for significance for CNV-GWAS ($p \leq 7.5 \times 10^{-6}$). (**B**) Prevalence of aneurysm according to 22q11.2 copy-number (CN) and CNV group (A). P-values compare deletion (CN = 1) and duplication (CN = 3) carriers from various groups (other = A-D, A-B, C-D; all = A-D, A-B, B-D, C-D) to copy-neutral (CN = 2) individuals (two-sided Fisher test). (**C**) Prevalence of IHD according to CNV groups (A). P-values compare IHD prevalence among individuals carrying a CNV (duplication or deletion) spanning LCR C-D, B-D, or A-D to copy-neutral (CN = 2) individuals (two-sided Fisher test). Error bars represent ± standard error; number of cases and sample sizes are indicated (N = cases/sample size).

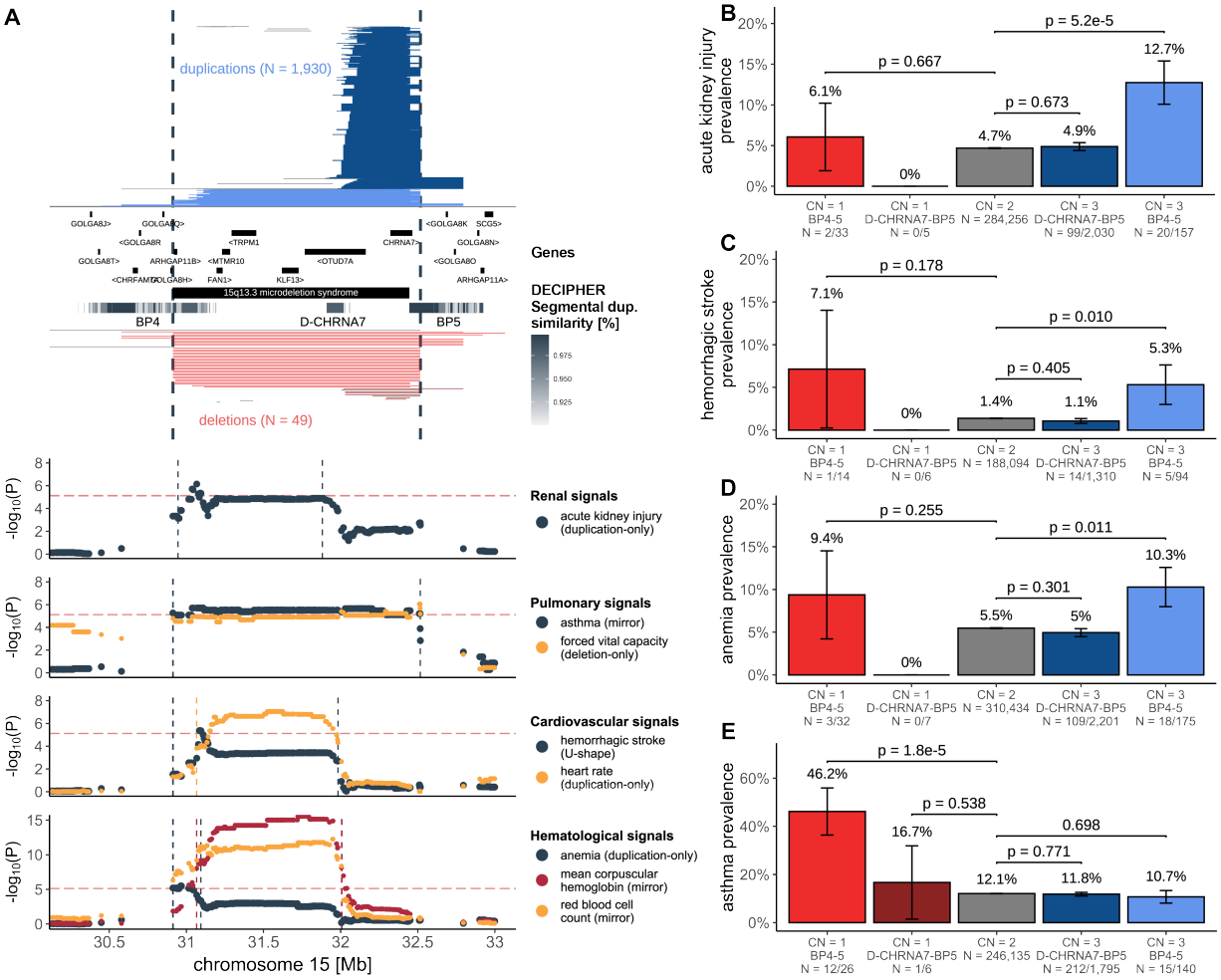


Figure 3.16: Dissection of complex pleiotropic patterns of recurrent CNVs at 15q13.
(**A**) 15q13 genetic landscape. Top: Coordinates of duplications (shades of blue; top) and deletions (shades of red; bottom) overlapping the maximal CNVR (delimited by vertical dashed lines) associated with acute kidney injury (AKI), asthma, forced vital capacity, hemorrhagic strokes, heart rate, anemia, mean corpuscular hemoglobin, and red blood cell count. CNVs are divided and colored according to whether they span breakpoint (BP) 4 to 5 or D-CHRNA7 to BP5, with atypical CNVs in gray (Table 3.4). Breakpoints reflect segmental duplications, represented as a gray gradient proportional to the degree of similarity. Genomic coordinates of genes and DECIPHER GD are displayed. Bottom: Negative logarithm of association p-values of CNVs (best model in parenthesis) with renal, pulmonary, cardiovascular, and hematological traits. Traits-specific CNVRs are shown with vertical dashed lines. Red horizontal dashed line represents the genome-wide threshold for significance for CNV-GWAS ($p \leq 7.5 \times 10^{-6}$). Prevalence (± standard error) of (**B**) AKI, (**C**) hemorrhagic stroke, (**D**) anemia, and (**E**) asthma according to 15q13 copy-number (CN) and groups from (A). P-values compare BP4-5 and D-CHRNA7-BP5 deletion (CN = 1) and duplication (CN = 3) carriers to copy-neutral (CN = 2) individuals (two-sided Fisher test). Number of cases and sample sizes are indicated (N = cases/sample size).

$OR_{U\text{-}shape}$ = 7.5; 95%-CI [3.2; 17.9]; $p = 4.3 \times 10^{-6}$; Figure 3.16C; note that this association is possibly confounded by BMI; Table 3.5; Table S3.5), and anemia (chr15:30,912,719–31,094,479; OR_{DUP} = 4.9; 95%-CI [2.5; 9.7]; $p = 3.2 \times 10^{-6}$; Figure 3.16D), reminiscent of associations with pulse rate, mean corpuscular hemoglobin, and red blood cell count (*208*, *294*). Replicating an association with asthma (*293*) (chr15:30,912,719–32,516,949; OR_{DEL} = 0.17; 95%-CI [0.08; 0.35]; $p = 1.2 \times 10^{-6}$) which parallels the previously reported decreased forced vital capacity (*208*) and peak expiratory flow (*294*), this was the only association at the locus driven by deletions, with prevalence being increased in only BP4-5 (46.2%; $p_{t\text{-}test} = 1.8 \times 10^{-5}$) but not D-CHRNA7-BP5 deletion carriers (16.7%; $p_{t\text{-}test} = 0.538$), compared to copy-neutral individuals (12.1%) (Figure 3.16E). Hence, non-neurological disorders appear to specifically involve dosage of the genes within BP4-D-CHRNA7 and not *CHRNA7*.

CNV burden at known genomic disorder CNVRs increases overall disease risk

By aggregating CNVs into a burden, we capture the effect of ultra-rare CNVs (frequency ≤ 0.01%), as well as those whose effect is not strong enough to reach GW significance under current settings, increasing our power to detect the global pathogenic impact of CNVs on human health. Individual-level autosomal CNV (duplication + deletion), duplication, and deletion burdens were calculated as the number of Mb or genes affected by the considered type of CNV. The predictive value of these six CNV burden metrics on the same 60 diseases (and the disease burden) previously assessed through CNV-GWAS was estimated (Figure 3.17A; middle). Disease burden strongly associated with a high CNV load (β_{DEL} = + 0.03 disease per deleted gene; $p = 3.7 \times 10^{-27}$) and risk for 20 individual disorders was increased by at least one type of CNV burden ($p \leq 0.05/61 = 8.2 \times 10^{-4}$; Figure 3.17B; *total* burden; Table S3.7). Overall, the deletion burden tended to yield more significant associations than the duplication burden and strongest effect sizes were observed for psychiatric disorders, such as bipolar disorder ($OR_{Mb\text{-}DEL} = 1.4$; $p = 6.9 \times 10^{-4}$), schizophrenia ($OR_{Mb\text{-}DEL} = 1.4$; $p = 4.1 \times 10^{-5}$), or epilepsy ($OR_{Mb\text{-}CNV} = 1.1$; $p = 8.3 \times 10^{-5}$), in agreement with CNVs representing important risk factors for these complex and polygenic disorders. Still, we note that the CNV burden only accounts for ~0.02% of the variability in disease burden, with up to 0.1% of schizophrenia and bipolar disorder cases being explained by the CNV burden (Table S3.8).

To ensure that we do not merely capture the effect of individual CNV-disease associations previously isolated by CNV-GWAS, we corrected CNV burdens for CNV-GWAS signals. Specifically, we excluded from the burden calculation CNVs overlapping disease-associated CNVRs in a disease- and burden-type-specific fashion. We then estimated the predictive value of corrected burdens on disease risk (Figure 3.17A; left). Overall association strength dropped but signal was lost only for type 1 diabetes and chronic obstructive pulmonary disease (Figure 3.17B; *GWAS-corrected*; Table S3.7). However, if we exclude CNVs overlapping the 40 autosomal unique disease-associated CNVRs systematically, i.e. not in a disease- and burden-type-specific fashion, the bulk of association signals disappears (Figure 3.17B; *CNVR-corrected*; Table S3.7), indicating that the genomic partition uncovered by our CNV-GWAS increases disease risk beyond the 73 CNV-disease pairs reaching genome-wide significance.

To further explore this hypothesis, we calculated subset CNV burdens (Figure 3.17A; right) overlapping three different genomic partitions (Figure 3.17C) composed of i) nine disease-associated CNVRs that map to known GDs (R1), ii) regions of known GDs that did not yield any association in our CNV-GWAS (R2), iii) and disease-associated CNVRs uncovered by our CNV-GWAS that were not linked to a known GD (R3). Risk for 25 diseases, as well as the disease burden, were significantly increased by the R1 CNV burden subset and included associations with eight diseases that were not picked up by the total burden association (Figure 3.17B; *R1 burden*; Table S3.7). We observed a substantial contribution of the R2 burden subset to the risk of diseases such as epilepsy, hypertension, cardiac conduction disorders, AKI, CKD, and hypothyroidism, even though the pleiotropy of this partition was more moderate

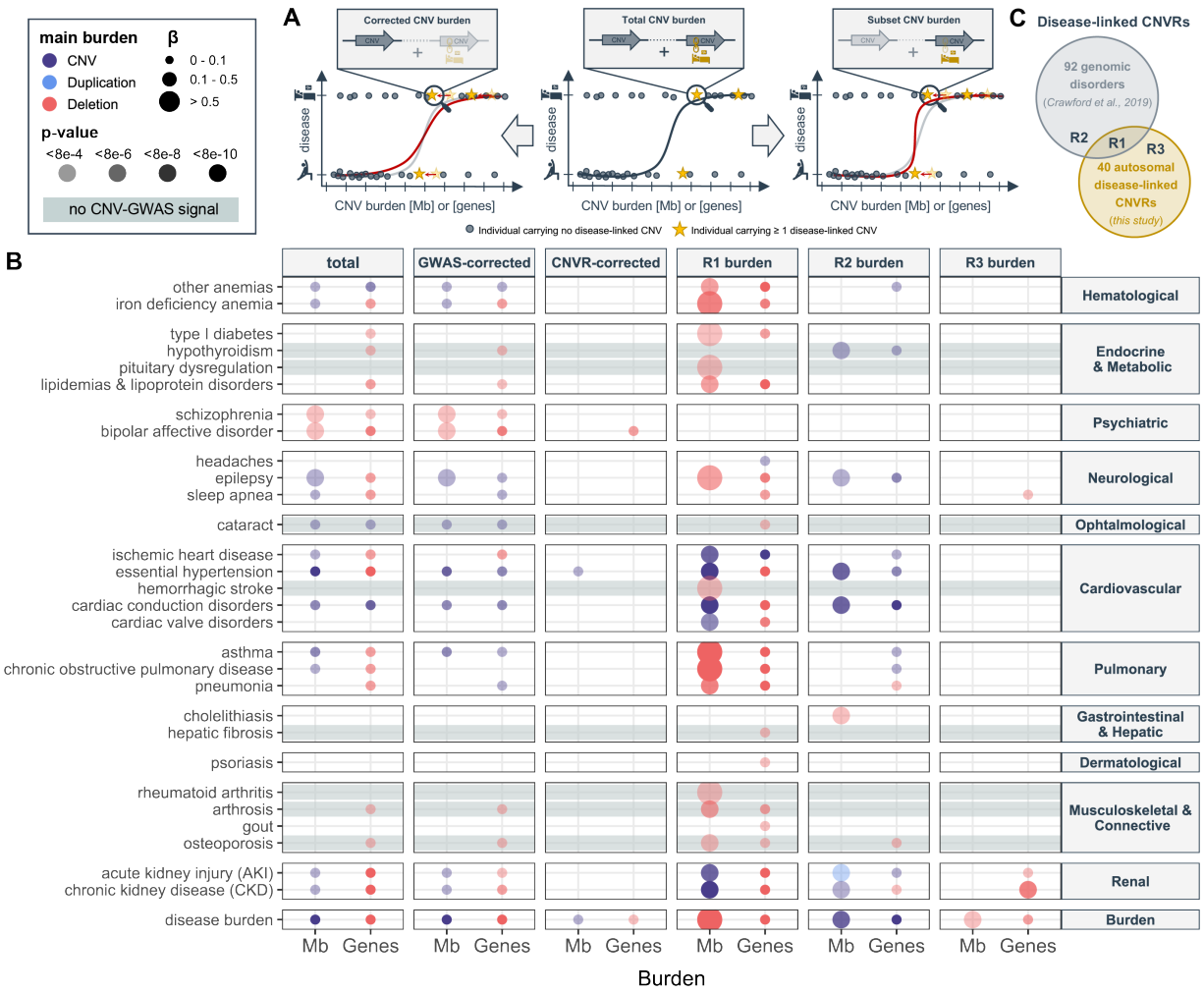


Figure 3.17: CNV burden at known genomic disorder CNVRs increases overall disease risk.
(**A**) Burden calculation. Middle: Total CNV (duplication + deletion), duplication, or deletion burdens are calculated by summing up the length (in affected Mb or genes) of all CNVs, duplications, or deletions in an individual, respectively. Burden values are used as a predictor for disease risk. Left: Corrected burdens are calculated by summing up the length of all CNVs, duplications, or deletions that do not overlap with regions listed in a given genomic partition. Right: Subset burdens are calculated by summing up the length of all CNVs, duplications, or deletions that overlap with regions listed in a given genomic partition. Both corrected and subset burden values are used to re-estimate the contribution of the CNV burden to disease risk (red curve). (**B**) Contribution of the total burden, CNV-GWAS signal- and CNVR-corrected burdens, and the R1, R2, and R3 subset burdens measured in number of affected Mb (x-axis; left) or genes (x-axis; right) to disease risk (y-axis). Only the effect of the most significantly associated of the CNV (purple), duplication (blue), or deletion (red) burdens, providing $p \leq 0.05/61 = 8.2 \times 10^{-4}$, is shown. Color indicates whether the CNV, duplication, or deletion burden was most significantly associated, with size and transparency being proportional to the effect size (β) and p-value, respectively. Gray horizontal bands mark traits with no CNV-GWAS signal. (**C**) Schematic representation of the R1, R2, and R3 partitions used to define the subset burdens in (B).

than the one of the R1 burden subset (Figure 3.17B; *R2 burden*; Table S3.7). Few associations were observed for the R3 CNV burden (Figure 3.17B; R3 burden; Table S3.7). Supporting these results, the CNVR (R1 + R3 partitions) and GD (R1 + R2 partitions) burden subsets strongly associated with 28 and 23 phenotypes, respectively (Figure 3.18; Table S3.7). A gradual loss of the number of associations was found when correcting the total CNV burden for the R3, R2, R1, GD, and CNVR partitions, with similar trends observed when requiring a more stringent overlap between CNVs and defined regions (Figure 3.18; Table S3.7). Overall, our results indicate that known GD CNVRs are the major drivers of the CNV burden's pathogenicity and hint at their currently underestimated pleiotropy.

Figure 3.18: Phenotype associations with genome-wide and partition total, corrected, and subset CNV burdens.
(**A**) Schematic representation of two ways to define overlap between a CNV and a CNV region (CNVR) in a given genomic partition. Left: Any overlap of ≥ 1 bp is sufficient to consider that the CNV overlaps the region. Right: a reciprocal 50% base pair overlap is required (stripes) to consider that the CNV overlaps the region, so that the CNV covers > 50% of the region defined by the partition and the region defined by the partition covers > 50% of the CNV. CNVs with blunted arrows extend over a range longer than the depicted CNVR. CNVs considered as overlapping are depicted in green and those that are not in red. Contribution of the total CNV burden, the GWAS-corrected burden, as well as the total CNV burden corrected for the five considered genomic partitions (i.e., R1, R2, R3, CNVR, GD) and the subset burden of the same five genomic partitions in number of affected Mb (x-axis; left) or genes (x-axis; right) to disease risk (y-axis) using (B) any or (**C**) a stringent approach to define overlap, as depicted in (A). Only the most significantly associated of the CNV (purple), duplication (blue), or deletion (red) burdens, providing $p \leq 0.05/61 = 8.2 \times 10^{-4}$, is shown. Color indicates whether the CNV (duplication + deletion), duplication, or deletion burden was most significantly associated, with size and transparency being proportional to the effect size (β) and p-value, respectively. Gray horizontal bands mark traits with no CNV-GWAS signal.

Discussion

Using an adapted GWAS framework, we provide a detailed investigation of the contribution of rare CNVs to the genetic architecture of 60 common diseases and showcase how the rich phenotypic data of the UKBB can be leveraged to gain new biological insights, highlighting the role of CNVs as modulators of common disease susceptibility in the general population.

Various strategies have been used to study CNV-disease associations in the UKBB. Focusing on diseases related to the ones assessed in the current study, we replicate 10 out of the 24 detected associations (at false discovery rate ≤ 0.1) with 54 likely pathogenic CNVs (*293*) and all four associations (at $p \leq 1 \times 10^{-9}$) in a recent CNV-GWAS investigating 757 diseases (*306*). Despite data originating from the same cohort, we often obtained p-values orders of magnitude smaller (e.g., 16p11.2 BP4-5 deletion and AKI: $p = 5.6 \times 10^{-20}$; $p = 6.0 \times 10^{-5}$ (*293*); $p = 3.3 \times 10^{-15}$ (*306*)). The increased power of our study might be explained by accruing case count from updated hospital records, careful case-control definition and statistical handling of the binary outcomes, probe-level association analysis, and usage of different association models to mimic various dosage mechanisms. We consequently identified previously unreported CNV-disease associations whose relevance was asserted by follow-up analyses. Only one signal – 17q12 CNVs increasing CKD risk – was backed by all approaches, emphasizing the importance of considering diverse lines of corroborative evidence, such as overlap with relevant SNP-GWAS signals and OMIM genes that indicate shared genetic mechanisms or both disease and disease-relevant biomarker associations mapping to the same CNVR. For instance, four (1q21.1–1q21.2, 15q13, 16p12.2, 16p11.2 BP4-5) out of six CNVRs decreasing forced vital capacity (*208*) were found to increase risk for pulmonary diseases, with the association between 15q13 and asthma replicating in the EstBB ($p = 6.2 \times 10^{-3}$) and 16p11.2 BP4-5 CNVs carriers being found to be enriched for "abnormal findings examination of lungs" in the Vanderbilt University Medical Center electronic health record database (*413*). This demonstrates that biomarkers are efficient proxies underlying (CNV-driven) pathological processes, often increasing the statistical power to detect associations due to their continuous nature. While we regressed covariates out of disease status to render the outcome quantitative, more sophisticated approaches have recently been developed for SNP-based GWASs that transform binary outcomes into continuous liability scores while borrowing information from age of disease onset, sex, and familial history (*139*). Future exploration is warranted to assess the benefit of this approach in the context of CNV-GWASs. By coupling a CNV-GWAS framework that accounts for challenges linked to disease CNV association studies in population cohorts to extensive validation, we generated a list of 73 CNV-disease pairs with various levels of supporting evidence that can inform follow-up studies.

Disease-associated CNVRs harbored genes under stronger evolutionary constraint than those lacking associations and their length correlated with their propensity for pleiotropy, indicating that as previously observed (*229*), both the number and the nature of genes affected by CNVs influence their pathogenicity. Consequently, large, multi-gene, recurrent CNVs exhibited the strongest pleiotropy. A longstanding question relates to the identification of causal genes whose altered dosage drives the phenotypic alterations observed in carriers. Models with various levels of complexity have been proposed, ranging from a single driver gene to multiple driver genes modulated by epistatic interactions with other genes in the CNVR (*458*). By analyzing disease prevalence in subsets of CNV carriers, association signals could be fine-mapped to narrower regions, pinpointing candidate drivers—such as *ABCC6* for kidney stones. In other cases, our data suggests that multiple subregions of the CNVR

contribute to increased risk for a given disease, as observed for 22q11.2 and ischemic heart disease or 16p13.11 and hypertension. Interestingly, the putative driver for phenotypes originally associated with a CNVR might not be driving our newly identified associations, as shown for the 15q13 CNVR, whose non-neurological phenotypes do not appear to be linked to altered dosage of *CHRNA7*. Beyond characterizing the pleiotropic pathological consequences of recurrent CNVRs, we demonstrate that dissection of CNV-GWAS signals can fine-map associations and provide mechanistic insights into their phenotypic expression.

We show that rare CNVs, such as the ones assessed in our study, only contribute marginally (0.02%) to the global disease burden in the general population. Still, from a personalized medicine perspective, these variants are highly relevant. Indeed, all detected CNV-disease associations pointed at CNVs increasing disease risk and leading to an earlier age of onset. Incorporating age of onset information has been shown to improve power to detect associations (*139*), and more importantly, represents proof of clinical relevance. Many signals mapped to regions whose genetic perturbation has been reported to be pathogenic in an autosomal dominant fashion. These include associations between well-described, clinically relevant gene-disease pairs – such as *BRCA1* and *LDLR* deletions increasing the risk for early-onset ovarian cancer and ischemic heart disease, respectively – but for which the role of CNVs in a large population cohort had not been previously investigated. CNVs in these genes have high penetrance but are extremely rare in the UKBB. Follow-up analyses based on the medical records, family history, medication use, and biomarkers could recapitulate additional clinical associations and establish that these deletions were most likely inherited. By recovering known gene-disease pairs typically studied in clinical cohorts, we showcase how the rich phenotypic data from biobanks can generate insights into the mechanisms, epidemiology, and comorbidities of these diseases, implicating CNVs as important genetic risk factors. We also highlight several examples where deviations by one copy number are linked to common diseases which share clinical features with rare Mendelian conditions caused by homozygous perturbations of the same genetic region. For instance, risk for kidney stones is gradually increased in carriers of partial vs full *ABBC6* deletions. Another intriguing example is the association between a relatively common CNV (frequency = 0.22%) affecting exon 2 and intron 2–3 of *PRKN* (MIM: 602544) – a gene causing juvenile autosomal recessive Parkinson's disease (MIM: 600116) – and sleep disorders such as insomnia and hypersomnia. As sleep disturbances are among the earliest symptoms of Parkinson's disease (*459*), follow-up studies should determine whether these individuals are more prone to develop Parkinson's disease. Overall, this argues against a dichotomic view on dominant vs recessive modes of inheritance and analogously to allelic series (*234*–*237*), suggests that Mendelian and common diseases represent different ends of the phenotypic spectrum caused by genetic variation at a given locus. We further show that nine CNVRs previously linked to pediatric GDs also increased risk for a broad spectrum of adult-onset common diseases. These associations were probably overlooked as the medical consequences in adulthood of these etiologies are often poorly characterized owing to ascertainment bias and difficulty to gather large cohorts. Importantly, 12 out of 24 associations mapping to a GD linked to altered BMI remained significant when accounting for the

latter. This indicates that while part of the increased disease risk among individuals with GDs represents a mere comorbidity of obesity, other BMI-independent mechanisms further contribute to the high disease burden observed in these individuals. In the future, it will be important to assess the role of other possible confounders, such as clinical biomarkers or socio-economic status, as such knowledge can guide preventive strategies and improve understanding of disease mechanisms. While awaiting validation in clinical cohorts of CNV carriers, we hope that these findings will improve clinical characterization of GDs, thereby facilitating diagnosis and allowing physicians to anticipate later-onset comorbidities. For instance, we found carriers of 16p13.11 deletions affecting *ABCC6*, the causal gene for pseudoxanthoma elasticum, to be at increased risk for kidney stones, paralleling reports from clinical cohorts showing that kidney stones represent an unrecognized feature of the disease (*448*–*450*). Awareness of this disease feature can mitigate kidney stone risk through adapted diet and sufficient water intake. Together, our results advocate for a complex model of variable CNV expressivity and penetrance that can result in a broad range of phenotypes along the rare-to-common disease spectrum and represent fertile ground for in-depth, phenome-wide studies aiming at better characterizing specific CNV regions (*411*, *412*).

Corroborating the deleterious impact of rare CNVs on an individual's health parameters, socio-economic status, and lifespan (*208*, *237*, *295*, *300*, *306*, *382*–*385*), we here speculate that the CNV burden acts on the latter by increasing risk for a broad range of common diseases beyond their known role in neuropsychiatric disorders (*335*, *386*, *387*, *389*). While both duplications and deletions contributed to increased disease risk, the deletion burden's impact was much stronger – especially for metabolic, psychiatric, pulmonary, and musculoskeletal diseases – in line with the commonly accepted view that deletions tend to be more deleterious. While only a marginal fraction of the CNV burden's contribution to disease risk was captured by CNV-GWAS signals, burden associations were mainly driven by known GDs. Only psychiatric disorders and the disease burden retained a significant association with the CNV burden when accounting for GDs, highlighting the polygenic CNV architecture of these traits. Illustrating the added value of the burden analysis, nine diseases showed a burden association despite lacking any CNV-GWAS signal. In some cases, such as for hypothyroidism, the burden signal originated from GDs that did not yield any significant CNV-GWAS associations, possibly because the involved regions did not pass the $\geq$ 0.01% CNV frequency filter. In other cases, such as for osteoporosis, the signal appeared to emanate from the CNVRs pick-up by the CNV-GWAS, indicating that we were likely underpowered to detect associations with any specific region. Overall, a total of 49 (82%) of the assessed diseases associated with CNVs either through CNV-GWAS or burden analysis, emphasizing the important role of this mutational class. While our burden analysis revealed that these associations mainly stem from known GDs, it also highlights that the latter are even more pleiotropic than what appears from our CNV-GWAS, implying that increased power will broaden the spectrum of common diseases associated with rare GDs.

A major limitation of our study is the reliance on microarray CNV calls, which allows us to assess only a fraction of the CNV landscape, i.e.,

mostly large CNVs or in regions with high probe coverage. Furthermore, as different population cohorts are genotyped with different arrays, partial probe overlap hinders replication power in external biobanks, as well as the ability to meta-analyze summary statistics. We speculate that small and/or multiallelic CNVs that can only be uncovered by sequencing will have a genetic architecture closer to the one of SNPs and indels, with higher frequencies and more subtle effect sizes. These effects, however, are more likely tagged by common variants, limiting novel discoveries. Furthermore, by detecting more events, sequencing-based studies require adapted and more stringent significance thresholds. Still, having improved breakpoint resolution, such CNV calls are also likely to enhance fine-mapping strategies. Microarray CNV calls also exhibit high false positive rates (*205*). By using stringent CNV selection criteria, we decrease the latter at the cost of decreasing power to detect true associations. This aspect is particularly relevant given that the type of CNVs we assess are rare and that the UKBB is not enriched for disease cases (*59*), resulting in low-powered GWASs. While we adopt strategies to counter the lack of power, our results are likely subject to Winner's curse, only capturing a fraction of the strongest, possibly overestimated effects. This phenomenon might be compensated by UKBB CNV carriers being at the milder end of the clinical spectrum, leading to effect underestimation. An interesting question will be to compare effect sizes from population-based studies to those emerging from clinical cohorts. In the future, longitudinal follow-up of UKBB participants will increase the number of cases – especially for late-onset diseases such as Alzheimer's or Parkinson's diseases – allowing better-powered CNV-GWASs. Larger and more diverse biobanks linking genotype to phenotype data (*8*, *460*, *461*) should both validate reported associations and identify new ones.

In conclusion, our study provides an in-depth analysis of the role of rare CNVs in modulating susceptibility to 60 common diseases in the general population, broadening our view on how this class of mutations impacts human health. Besides describing clinically relevant and actionable associations, we illustrate how complex pleiotropic patterns can be dissected to gain new insights into the pathological mechanisms of large recurrent CNVs, providing a framework that can be applied to an even larger spectrum of diseases.

Acknowledgments

We thank all biobank participants for sharing their data. UKBB and EstBB computations were performed on the JURA server (University of Lausanne) and the High-Performance Computing Center (University of Tartu), respectively. Open access funding was provided by the University of Lausanne. The study was funded by the Swiss National Science Foundation (31003A_182632, AR; 310030_189147, ZK), Horizon2020 Twinning projects (ePerMed 692145, AR), the Estonian Research Council (PRG687, MJ and RM), and the Department of Computational Biology (ZK) and the Center for Integrative Genomics (AR) from the University of Lausanne.

Declaration of interests

SO was an employee of MSD at the time of the submission; their contribution to the research occurred during affiliation at the University of Lausanne. The remaining authors declare that they have no competing interests.

Supplemental tables

Supplemental tables are available for download as a single Excel file.

- **Table S3.1** Disease epidemiology.
- **Table S3.2** Covariate selection and probe prefiltering.
- **Table S3.3** Genomic inflation of genotypic Fisher test p-value.
- **Table S3.4** Genome-wide significant CNV-GWAS associations.
- **Table S3.5** BMI adjustment for possibly confounded CNV-disease associations.
- **Table S3.6** CNV region characteristics.
- **Table S3.7** Impact of CNV burden on disease risk.
- **Table S3.8** Disease variance explained by the CNV burden.

Approaches to dissect the pleiotropy of recurrent genomic rearrangements

22q11.2 | 4

Remember, there are always two sides to every story. Understanding is a three edged sword. Your side, their side, and the truth in the middle. Get all the facts before you jump to conclusions.

– J. Michael Straczynski

4.1 Aims 127
4.2 Key Findings 127
4.3 Author Contributions 128
4.4 The impact of 22q11.2 copy-number variants on human traits in the general population 129

This chapter describes an extended version of *"The impact of 22q11.2 copy-number variants on human traits in the general population"* (411), which was published in the *American Journal of Human Genetics*. This work, led by Malú Zamariolli, was the first study to make use of the methodology developed in the two previous chapters to perform a detailed investigation of a single genomic disorder-associated locus, the 22q11.2 CNV region.

Malú visited our group for about a year, at the time when we were still developing the extension of the CNV-GWAS pipeline to binary traits. Given her interest and expertise in the 22q11.2 deletion syndrome, this project represented the perfect opportunity to apply and test some of the features that we were developing, before deploying them for genome-wide studies. Besides the many new insights in the genetics and biology of 22q11.2 itself, this study represents a testing ground for statistical approaches that can be used to gain deeper mechanistic insights into the relation between recurrent CNVs and phenotypes.

ARTICLE

The impact of 22q11.2 copy-number variants on human traits in the general population

Malú Zamariolli,[1,2] Chiara Auwerx,[2,3,4,5] Marie C. Sadler,[2,3,4] Adriaan van der Graaf,[2] Kaido Lepik,[2] Tabea Schoeler,[2,7] Mariana Moysés-Oliveira,[6] Anelisa G. Dantas,[1] Maria Isabel Melaragno,[1] and Zoltán Kutalik[2,3,4,*]

Summary

While extensively studied in clinical cohorts, the phenotypic consequences of 22q11.2 copy-number variants (CNVs) in the general population remain understudied. To address this gap, we performed a phenome-wide association scan in 405,324 unrelated UK Biobank (UKBB) participants by using CNV calls from genotyping array. We mapped 236 Human Phenotype Ontology terms linked to any of the 90 genes encompassed by the region to 170 UKBB traits and assessed the association between these traits and the copy-number state of 504 genotyping array probes in the region. We found significant associations for eight continuous and nine binary traits associated under different models (duplication-only, deletion-only, U-shape, and mirror models). The causal effect of the expression level of 22q11.2 genes on associated traits was assessed through transcriptome-wide Mendelian randomization (TWMR), revealing that increased expression of *ARVCF* increased BMI. Similarly, increased *DGCR6* expression causally reduced mean platelet volume, in line with the corresponding CNV effect. Furthermore, cross-trait multivariable Mendelian randomization (MVMR) suggested a predominant role of genuine (horizontal) pleiotropy in the CNV region. Our findings show that within the general population, 22q11.2 CNVs are associated with traits previously linked to genes in the region, and duplications and deletions act upon traits in different fashions. We also showed that gain or loss of distinct segments within 22q11.2 may impact a trait under different association models. Our results have provided new insights to help further the understanding of the complex 22q11.2 region.

Introduction

The 22q11.2 region is a structurally complex region of the genome because of the presence of segmental duplications or low-copy repeats (LCRs), named LCRA to LCRH, which predispose the region to genomic rearrangements, resulting in deletions or duplications of different segments. Specifically, deletions within the ~3 Mb segment from LCRA to LCRD represent the main cause of the 22q11.2 deletion syndrome (22q11.2DS [MIM: 188400]), the most frequent microdeletion syndrome in humans, with an estimated incidence between 1:3,000 and 1:6,000 live births.[1]

Studies in clinical cohorts have investigated the phenotypic consequences of the 22q11.2 deletion, which include cardiac defects; facial and palate alterations; immunodeficiencies; endocrine, genitourinary, and gastrointestinal alterations;[1,2] developmental delay, cognitive deficits; and psychiatric disorders, such as schizophrenia.[1] In contrast, the phenotypic consequences of the region's duplication (MIM: 608363) remain more elusive. Most of what is known is based on studies of a few individuals or families, but the findings indicate pleiotropy and variable consequences, similar to the deletion. Some features, such as heart defects, velopharyngeal insufficiency, and neurodevelopmental and psychiatric disorders, are shared with the 22q11.2DS.[3,4] Other 22q11.2 duplication carriers exhibit very mild or unnoticeable phenotypes,[5] suggesting variable expressivity and/or reduced penetrance. While many phenotypes are shared between duplication and deletion carriers, some may be gene dosage sensitive. The 22q11.2 deletion is a strong risk factor for schizophrenia; however, the reciprocal duplication seems to be less common and has been suggested as protective for this phenotype.[6] In addition, differential impact of duplications and deletions in psychosis-related traits[7] and brain structure[8] has been described.

Finally, rare single-nucleotide variants (SNVs) in genes encompassed by the region have been linked to various disorders, such as Bernard-Soulier syndrome (MIM: 231200), caused by SNVs in *GP1BB* (MIM: 138720),[9] or CEDNIK (MIM: 609528) syndrome, caused by SNVs in *SNAP29* (MIM: 604202).[10] Overall, the multitude of variants and phenotypes that have been linked to the 22q11.2 LCRA to LCRD region highlights its clinical relevance.

Because of their highly deleterious impact, 22q11.2 variants are often investigated in clinical settings. Studied cohorts are thus heavily biased toward individuals with severe phenotypic manifestation, leading to an incomplete and biased understanding of these variants' role in the human population. This is particularly relevant considering recent studies that have shown variable expressivity and

[1]Genetics Division, Universidade Federal de São Paulo, São Paulo, Brazil; [2]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland; [3]Swiss Institute of Bioinformatics, Lausanne, Switzerland; [4]University Center for Primary Care and Public Health, University of Lausanne, Lausanne, Switzerland; [5]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland; [6]Sleep Institute, São Paulo, Brazil; [7]Department of Clinical, Educational and Health Psychology, University College London, London, UK
*Correspondence: zoltan.kutalik@unil.ch
https://doi.org/10.1016/j.ajhg.2023.01.005.
© 2023 The Author(s). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

300 The American Journal of Human Genetics *110*, 300–313, February 2, 2023

Figure 4.1: Cover of *Zamariolli et al., 2023.*

Data & code availability:

- → Disease summary statistics
- → Quantitative trait summary statistics
- → GitHub

4.1 Aims

Deletions – and to a lesser extent duplications – of the 22q11.2 low-copy repeat (LCR) A-D region are among the most common etiologies of genomic disorders, whose phenotypic consequences have primarily been described in clinical cohorts. Increased awareness around the variable expressivity of pathogenic variants indicates that focusing solely on clinical cases might bias our understanding towards the more severe end of the phenotypic spectrum. As such, the study's aims were to:

1. Assess whether phenotypes previously linked to CNVs and SNVs mapping to the 22q11.2 region are altered in 22q11.2 CNV carriers from the UKBB general population.
2. Use four probe-level association models and causal inference strategies to gain mechanistic insights into the specific genes whose dosage is linked to particular phenotypic alterations in 22q11.2 CNV carriers.

4.2 Key Findings

A total of 631 Human Phenotype Ontology (HPO) terms were linked to 24 protein-coding genes within the 22q11.2 CNV region. These could be

mapped to 152 binary and 18 quantitative UKBB traits, among which, 9 (6%) and 8 (44%) were significantly associated with CNVs in the 22q11.2 region through at least one association model. Further analyses indicated that the majority of these associations reflect genuine direct pleiotropy, as opposed to indirect secondary consequences of the CNV's effect on one of the associated traits.

The U-shape model yielded the most associations (seven best associations; e.g., fluid intelligence or hearing loss), followed by the mirror model (six best associations; e.g., platelet count and volume). Intriguingly, we identified two independent associations with height: a U-shape association involving the LCR A-B region and a deletion-only effect of the LCR C-D region. Besides suggesting that multiple genes in the 22q11.2 regions are important in determining height, it showcases the importance of considering multiple dosage mechanisms.

Seventeen genes in the regions could be genetically instrumented to assess the causal impact of changes in their expression on analyzed traits through transcriptome-wide Mendelian randomization (TWMR). This identified two putative causal genes: increased *ARVCF* expression increased BMI, in line with the BMI-increasing effect of the duplication, while increased *DGCR6* decreased mean platelet volume, aligning with the increased mean platelet volume detected in deletion carriers. Relaxing our approach to nominally significant effects, we further detected overall directional concordance between TWMR and CNV association effects.

By showing that traits linked to 22q11.2 genes are affected in UKBB 22q11.2 CNV carriers, our study stresses the benefits of leveraging population cohorts to study rare CNV syndromes and supports a model of variable expressivity and incomplete penetrance for 22q11.2 CNVs. Furthermore, by leveraging tools typically used to study common variants, we show how we can gain mechanistic insights into these complex regions.

4.3 Author Contributions

I contributed to the design of this study, together with Malú Zamariolli, Mariana Moysés-Oliveira, Anelisa Dantas, Maria Isabel Melaragno, and Zoltán Kutalik. I further contributed by sharing the UKBB CNV calls, providing guidance for the CNV association analysis, and helping with data interpretation and manuscript revision.

The bulk of the analyses, as well as the drafting of the manuscript and designing of figures, was done by Malú Zamariolli, under the supervision of Zoltán Kutalik. The following co-authors contributed:

- Marie Sadler performed the TWMR analyses.
- Adriaan van der Graaf carried out the multivariable MR analyses.
- Kaido Lepik designed the web scraping approach used to map HPO terms to UKBB traits.
- Tabea Schoeler contributed to the statistical interpretation of potential biases in the study.

4.4 The impact of 22q11.2 copy-number variants on human traits in the general population

Malú Zamariolli [1,2], **Chiara Auwerx [2,3,4,5]**, Marie C. Sadler [2,3,4], Adriaan van der Graaf [2], Kaido Lepik [2], Tabea Schoeler [2,6], Mariana Moysés-Oliveira [7], Anelisa G. Dantas [1], Maria Isabel Melaragno [1], Zoltán Kutalik [2,3,4,*].

Abstract

While extensively studied in clinical cohorts, the phenotypic consequences of 22q11.2 copy-number variants (CNVs) in the general population remain understudied. To address this gap, we performed a phenome-wide association scan in 405,324 unrelated UK Biobank (UKBB) participants by using CNV calls from genotyping array. We mapped 236 Human Phenotype Ontology terms linked to any of the 90 genes encompassed by the region to 170 UKBB traits and assessed the association between these traits and the copy-number state of 504 genotyping array probes in the region. We found significant associations for eight continuous and nine binary traits associated under different models (duplication-only, deletion-only, U-shape, and mirror models). The causal effect of the expression level of 22q11.2 genes on associated traits was assessed through transcriptome-wide Mendelian randomization (TWMR), revealing that increased expression of *ARVCF* increased BMI. Similarly, increased *DGCR6* expression causally reduced mean platelet volume, in line with the corresponding CNV effect. Furthermore, cross-trait multivariable Mendelian randomization (MVMR) suggested a predominant role of genuine (horizontal) pleiotropy in the CNV region. Our findings show that within the general population, 22q11.2 CNVs are associated with traits previously linked to genes in the region, and duplications and deletions act upon traits in different fashions. We also showed that the gain or loss of distinct segments within 22q11.2 may impact a trait under different association models. Our results have provided new insights to help further the understanding of the complex 22q11.2 region.

[1] Genetics Division, Universidade Federal de São Paulo, São Paulo, Brazil; [2] Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland; [3] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; [4] University Center for Primary Care and Public Health, 1010 Lausanne, Switzerland; [5] Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland; [6] Department of Clinical, Educational and Health Psychology, University College London, London, UK; [7] Sleep Institute, São Paulo, Brazil; [*] **Correspondence**.

Introduction

The 22q11.2 region is a structurally complex region of the genome because of the presence of segmental duplications or low-copy repeats (LCRs), named LCR A-H, which predispose the region to genomic rearrangements, resulting in deletions or duplications of different segments. Specifically, deletions within the ~3 Mb segment from LCR A-D represent the main cause of the 22q11.2 deletion syndrome (22q11.2 DS [MIM: 188400]), the most frequent microdeletion syndrome in humans, with an estimated incidence between 1 in 3,000-6,000 live births (*453*).

Studies in clinical cohorts have investigated the phenotypic consequences of the 22q11.2 deletion, which include cardiac defects; facial and palate alterations; immunodeficiencies; endocrine, genitourinary and gastrointestinal alterations; developmental delay, cognitive deficits; and psychiatric disorders, such as schizophrenia (*453*, *462*). In contrast, the phenotypic consequences of the region's duplication (MIM: 608363) remain more elusive. Most of what is known is based on studies of a

few individuals or families, but the findings indicate pleiotropy and variable consequences, similar to the deletion. Some features, such as heart defects, velopharyngeal insufficiency, and neurodevelopmental and psychiatric disorders, are shared with the 22q11.2 DS (*463*, *464*). Other 22q11.2 duplication carriers exhibit very mild or unnoticeable phenotypes (*465*), suggesting variable expressivity and/or reduced penetrance. While many phenotypes are shared between duplication and deletion carriers, some may be gene dosage sensitive. The 22q11.2 deletion is a strong risk factor for schizophrenia; however, the reciprocal duplication seems to be less common and has been suggested as protective for this phenotype (*466*). In addition, the differential impact of duplications and deletions in psychosis-related traits (*467*) and brain structure (*468*) has been described.

Finally, rare single-nucleotide variants (SNVs) in genes encompassed by the region have been linked to various disorders, such as Bernard-Soulier syndrome (MIM: 231200), caused by SNVs in *GP1BB* (MIM: 138720) (*469*), or CEDNIK syndrome (MIM: 609528), caused by SNVs in *SNAP29* (MIM: 604202) (*470*). Overall, the multitude of variants and phenotypes that have been linked to the 22q11.2 LCR A-D region highlights its clinical relevance.

Because of their highly deleterious impact, 22q11.2 variants are often investigated in clinical settings. Studied cohorts are thus heavily biased toward individuals with severe phenotypic manifestation, leading to an incomplete and biased understanding of these variants' role in the human population. This is particularly relevant considering recent studies that have shown variable expressivity and incomplete penetrance of SNVs (*235*, *241*) and CNVs (*208*) that were previously believed to be highly pathogenic, including at the 22q11.2 LCR A-D locus (*291*). To address this gap, we performed a phenome-wide analysis in the UK Biobank (UKBB) (*61*), a populational cohort of ~500,000 individuals, to identify associations of 22q11.2 CNVs with traits previously implicated by their genetic content.

Materials and methods

Study material

Cohort description

Analyses were performed in the UK Biobank (UKBB), a volunteer-based cohort from the general UK adult population (*61*). Gender mismatched, related, and retracted samples (09/08/2021), as well as CNV outliers (see *CNV calling*) were excluded, resulting in a total of 405,324 participants (54% females) used for the analyses. Individuals were aged 40 to 69 years at recruitment. All participants signed a broad informed consent form and data were accessed through the UKBB application 16389.

Software versions:

- CNV calling: PennCNV v1.0.5 (*203*).
- CNV QC: (*206*).
- PLINK v1.9 and PLINK v2.0.26 (*88*).
- Statistical analyses: R v3.6.1.
- Graphs: R v4.2.0.

22q11.2 region definition

We defined the 22q11.2 region as chr22:18,630,000–21,910,000 based on the human genome reference build GRCh37/hg19 to encompass LCR A-D. The 90 NCBI RefSeq genes contained in the region were downloaded from the UCSC Table Browser.

Trait selection

Phenotypes linked to the 22q11.2 region's genetic content were identified with the Human Phenotype Ontology (HPO) mapping (*471*), an

ontology-based system that uses information from different medical sources, including OMIM and Orphanet. Genes and their most specific associated HPO term (i.e., not all ancestors) were downloaded from the HPO database (accessed 22/10/2021). Overall, 24 out of 90 genes in the 22q11.2 region, all protein-coding, were associated with at least one HPO term, yielding 631 associated HPO terms.

Mapping of HPO terms to UK Biobank phenotypes

Binary traits

To map HPO terms to binary UKBB traits, we used two complementary approaches. First, we used the online tool EMBL-EBI Ontology Xref Service (OxO) to map HPO terms to International Classification of Diseases, 10th Revision (ICD-10) codes, followed by manual curation and grouping of ICD-10 codes into broader phenotypes when appropriate according to the Phecode map (*309*). We mapped the remaining HPO terms to Phecode definitions by using manual curation (*472*). Mapping was manually curated and only phenotypes with $\geq$ 500 cases were retained. In addition, individuals with a related ICD-10 code or self-reported disease to the one studied were excluded from controls in a phenotype-specific fashion (Table S4.1). Overall, 218 HPO terms were mapped to 152 UKBB binary traits (Table S4.2). The number of individuals by phenotype is reported in Table S4.3.

Continuous traits

We developed an in-house web-scraping approach to map HPO terms to UKBB continuous traits. We used a list of 1,769 continuous UKBB measures as input on the HPO database to obtain the web page's results for each query. Results were filtered for HPO terms of interest, i.e., 631 terms linked to 22q11.2 genes. With this approach, 18 UKBB continuous traits were obtained from 18 HPO terms (Table S4.4). The number of individuals by trait is reported in Table S4.5.

22q11.2 CNV association scan

CNV calling

CNVs were called with PennCNV v.1.0.5 and underwent quality control as previously described (*208*). Briefly, a quality score (QS) reflecting the probability for the CNV to be a true positive was assigned to each call and used for filtering ($|QS| \geq 0.5$) (*206*). We excluded CNVs from samples genotyped on plates with a mean CNV count per sample > 100 or from samples with > 200 CNVs or a single CNV > 10 Mb to minimize batch effects, genotyping errors, or extreme chromosomal abnormalities. CNV calls were transformed into $probe \times sample$ matrices with copy-number state for each probe (deletion = -1; copy-neutral = 0; duplication = 1).

PLINK encoding and association models

We converted probe-level matrices to PLINK binary file sets, where copy-number states were encoded to accommodate analysis according to four different association models: duplication-only, deletion-only, mirror, and U-shape models (*82*). The duplication-only model assessed the impact of duplications disregarding deletions; the deletion-only model assessed the impact of deletions disregarding duplications; the mirror model assessed the additive effect of each additional copy of a probe (i.e., duplications

and deletions have opposing effects); the U-shape model assumes that duplications and deletions have the same effect direction (*208*).

CNV probe selection and number of effective tests
Probes with high genotype missingness (> 5%) were excluded, resulting in 864 CNV-proxy probes spanning chr22:18,630,000–21,910,000. We retained 504 CNV-proxy probes that are highly correlated ($r^2 \geq 0.999$) to at least ten other probes, allowing us to reduce the multiple testing burden while ensuring that selected probes adequately capture the CNV landscape of the region.

The number of effective probes (i.e., number of probes required to capture 99.5% of the variance in the $probe \times sample$ matrices) was calculated (*85*) based on the 504 CNV-proxy probes ($N_{probes} = 6$). We used the same approach to account for correlation among 18 continuous ($N_{cont} = 16$) and 152 binary traits ($N_{bin} = 113$). This resulted in 774 effective tests[1], setting the threshold for significance at $p \leq 0.05/774 = 6.5 \times 10^{-5}$.

1: $N_{eff} = N_{probes} \cdot (N_{cont} + N_{bin})$

Continuous traits
The 18 selected continuous traits were inverse normal transformed and corrected for covariates: age, age^2, sex, genotyping batch, and principal components (PCs) 1–40. Associations between the copy number (CN) of selected probes and normalized covariate-corrected traits were performed in PLINK v.2.0 according to all four association models with linear regression, as previously described (*208*). Significant associations ($p \leq 6.5 \times 10^{-5}$) were retained.

Binary traits
For each trait, covariates among age, age^2, sex, genotyping batch, and PCs 1–40 that were significantly associated with the trait ($p \leq 0.05$) were selected with logistic regression in R. Associations between the CN of selected probes and 152 binary selected traits were performed in PLINK v.2.0 according to all four association models with logistic regression and correcting for trait-specific selected covariates. Significant associations ($p \leq 6.5 \times 10^{-5}$) were retained.

Stepwise conditional analysis
The number of independent signals per trait and association model was determined by stepwise conditional analysis (*208*), i.e., the CNV status of the lead probe was regressed out from the trait and association scan was conducted again until no more significantly associated probes remained.

Sensitivity analysis
Due to the low frequency of CNVs within the 22q11.2 region, alternative tests were performed to ensure the confidence of significant associations. For significant associations with continuous traits, we performed a Wilcoxon rank-sum test as a sensitivity analysis to assess agreement with linear regression. Significant associations with binary traits were retained only when confirmed by at least one of two approaches: i) Fisher's exact test ($p \leq 0.005$) for the duplication-only, deletion-only, and U-shape models and Cochran-Armitage test ($p \leq 0.0005$) for the mirror model; ii) linear regression ($p \leq 0.005$) of the inverse-normal-quantile-transformed trait residuals obtained from the logistic regression model of the binary outcome on the selected covariates.

Enrichment analysis

For each gene, two groups of traits were defined: traits linked to the focal gene implicated by HPO versus other traits related to other genes in the 22q11.2 region but not to the focal gene. Association p-values for each probe within the gene (± 10 kb) and each association model were compared between traits in the two groups with a one-sided Wilcoxon rank-sum test (i.e., H_a: unrelated traits have higher association p-values with the focal gene than related ones). We calculated the number of effective tests for each gene and used this to define gene-specific significance thresholds. Genes were considered significant when the probe with the smallest p-value reached that threshold. We only performed the comparison for genes with at least four continuous traits and ten binary traits in each group to avoid selecting genes associated with very few traits that would not have sufficient statistical power to test for enrichment. We performed a binominal enrichment to establish whether the number of genes significant in the Wilcoxon rank-sum test was higher than expected by chance with `pbinom()` in R.

Transcriptome-wide Mendelian randomization

Transcriptome-wide Mendelian randomization (TWMR) was conducted as previously described (*173*) to identify changes in transcript levels of genes in the 22q11.2 region that causally modulate traits found to be associated with 22q11.2 CNVs by our association scan and, if this was the case, in which direction (i.e., whether increased gene expression associates with increased or decreased phenotype value). Briefly, the exposure (i.e., transcript level) and outcome (i.e., trait) are instrumented with independent genetic variants (instrumental variables [IVs]; $r^2 < 0.01$). Given their genetic effect sizes on these two quantities, a causal effect of the exposure on the outcome can be estimated with two-sample Mendelian randomization (MR). Genetic effect sizes on transcript levels originate from whole blood expression quantitative trait loci (eQTLs) provided by the eQTLGen consortium (*cis*-eQTLs at false discovery rate < 0.05, two-cohort filter) (*154*). Effect sizes on the traits stem from genome-wide association study (GWAS) summary statistics conducted on the UK Biobank (Neale's lab; Pan-UKBB; Table S4.6). Prior to the analysis, eQTL and GWAS data were harmonized and palindromic SNPs were removed, as well as SNPs with an allele frequency difference > 0.05 between datasets. For increased robustness of the estimated causal effects, ≥ 5 (independent) IVs were required. MR estimates were considered significant when $p \leq 0.05/17 = 0.003$ to account for the testing of 17 transcripts with ≥ 5 IVs and only significant genes overlapped by the CNV-association signal were reported.

TWMR results were used for validation of the mirror model associations. It is expected that TWMR and mirror model effects are directionally concordant, i.e., increase/decrease in copy number has the same direction of effect on a trait as an increase/decrease in gene expression. For this purpose, nominally significant ($p < 0.05$) TWMR effects were retained and their direction was compared to the direction of the probe with the smallest nominally significant p-value ($p < 0.05$) in the mirror association model for the corresponding gene (± 10 kb) and trait.

Multivariable Mendelian randomization

We performed multivariable Mendelian randomization (MVMR) to assess the causal relationship between significantly associated traits and compute a phenotype network. IVs were obtained from the Neale lab and Pan-UKBB GWAS summary statistics for all eight significant continuous traits and nine significant binary traits (Table S4.6). Data were harmonized with genetic variants in the UK10K reference dataset and variants with minor allele frequency (MAF) ≤ 0.01 were filtered out. Genetic variants were clumped at $r^2 = 0.001$ with UK10K as a reference panel in PLINK v.1.9. MR analysis was performed in two steps. First, potentially causal effects were identified with a univariable inverse-variance weighted (IVW) MR for all exposure-outcome combinations (i.e., pairs of associated traits). Second, all exposures with nominally significant IVW causal effect estimates for a given outcome were included in an MVMR analysis as exposures. To reduce bias due to potential reverse causation, we performed Steiger filtering in all MR analyses ($p < 5 \times 10^{-3}$).

MVMR established the causal relationships among assessed traits by using genetic variants as IVs. To infer whether the pleiotropic effect of CNVs is vertical (indirect) or horizontal (genuine), we estimated what would be the expected CNV effect on the outcome trait ($\beta_{\text{expected outcome}}$) if that outcome is a downstream result of the exposure trait as suggested by the MVMR analysis (vertical pleiotropy). $\beta_{\text{expected outcome}}$ was determined as $\beta_{\text{exposure}} \times \beta_{\text{IVW}}$, where β_{exposure} is the effect size of the best probe in the mirror model for each exposure (i.e., observed CNV-exposure trait association) and β_{IVW} is the causal estimate for each exposure-outcome pair obtained from IVW MR. We then compared $\beta_{\text{expected outcome}}$ with the observed CNV effect on the outcome trait ($\beta_{\text{observed outcome}}$) obtained from the mirror association model.

Results

22q11.2 CNVs in the UKBB

After CNV calling and quality control in 405,324 unrelated individuals of the UKBB, we identified 1,127 individuals with a duplication and 694 individuals with a deletion overlapping the 22q11.2 LCR A-D region (Figure 4.2A). CNVs varied in size: duplication length ranged between 71 kb and 8.8 Mb (i.e., breakpoints extending beyond the defined region) with a median of 132 kb, while deletion length ranged between 80 kb and 2.8 Mb also with a median of 132 kb.

To assess whether individuals with these CNVs (mean number of diagnoses = 8.6) had a higher disease burden than individuals who are copy neutral within this region (mean number of diagnoses = 8), we compared the reported number of ICD-10 codes and identified no statistical difference (two-sided Wilcoxon rank-sum test: $p_{DEL} = 0.44$; $p_{DUP} = 0.053$) (Figure 4.2B).

CNVs were classified according to their localization as defined by LCR A-D. Between LCRs A-B, duplications were identified at a frequency of 0.01% and deletions at 0.002%; CNVs from LCR A-D had a frequency of 0.06% and 0.001% for duplications and deletions, respectively; from LCR B-D, duplications had a frequency of 0.002% and no deletions were

identified; between LCRs C and D, duplications were identified at a frequency of 0.04% while deletions were identified at 0.008%. CNVs that did not fall into these categories were considered atypical and had a frequency of 0.16% for both duplications and deletions (Figure 4.2A).

To account for all CNVs and bypass issues related to breakpoint variability, CNV calls were converted into $probe \times sample$ matrices for the CNV association scan. Probe-level CNV frequency after excluding LCR A probes (mean duplication frequency: 0.07%; mean deletion frequency: 0.004%) ranged between 0.004% and 0.1% and 0.001% and 0.01% for duplications and deletions, respectively (Figure 4.2C).

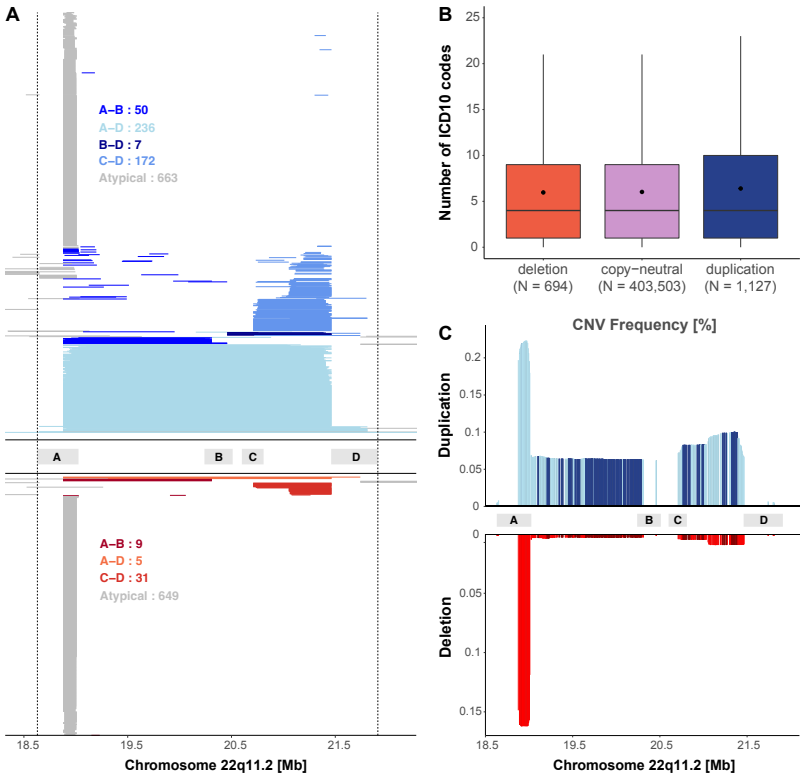


Figure 4.2: 22q11.2 CNVs landscape. (**A**) Each UKBB CNV carrier is displayed through a segment that spans the genomic coordinates of the CNV. Duplications are represented in the top part of the graph, while deletions are at the bottom. Shades of blue and red represent different duplication and deletion categories, respectively, according to their localization in reference to the LCR A-D. The number of duplications and deletions for each category is displayed in the boxes. (**B**) Boxplot representing the number of ICD-10 codes reported in individuals grouped according to their copy-number state in the 22q11.2 region; dots show the mean; outliers are not shown. N indicates the sample size for each category. (**C**) Probe-level duplication (top, blue) and deletion (bottom, red) frequencies (%) for 864 probes plotted against the 22q11.2 genomic region. Frequency was calculated as the number of duplications or deletions divided by the total number of individuals assessed for the probe.

Associated traits

CNV association scan revealed significant links for eight continuous (Figure 4.3A; Table 4.1A) and nine binary traits (Figure 4.3B; Table 4.1B), which were associated under different association models. Seven associations (two continuous and five binary traits) were associated most significantly under the U-shape model, six (four continuous and two binary traits) did so under the mirror model, three (two continuous and one binary trait) associated most significantly under the deletion-only model, and two (one continuous and one binary trait) under the duplication-only model, highlighting the importance of testing models mimicking different dosage mechanisms.

Among the identified continuous traits, body mass index (BMI) was found associated under the U-shape model ($\beta = 1.56\,kg/m^2$, $p = 4.9 \times 10^{-10}$) throughout LCR A-D (Figure 4.4A), indicating that both duplications and deletions increase BMI level (Figure 4.4B). TWMR analysis showed that increased expression of *ARVCF* (MIM: 602269) increases BMI (β =

0.05, p = 10^{-4}), concordantly with the positive association found by the mirror CNV association scan (Figure 4.4C).

Mean platelet volume was found associated under the mirror model (β = -0.58 fL, p = 1.3 × 10^{-18}), and the strongest association occurred in the LCR A-B region (Figure 4.5A). The signal replicated in both the duplication-only (β = -0.54 fL, p = 1.16 × 10^{-15}) and deletion-only (β = 1.66 fL, p = 1.13 × 10^{-6}) models, providing further evidence of a true mirror effect, despite the deletion effect's being slightly stronger than the duplication one (Figure 4.5B). In line with this effect, TWMR revealed that increased *DGCR6* (MIM: 601279) expression causally reduces mean platelet volume (β = -0.03, p = 0.001) (Figure 4.5C). It is worth noting that this trait is negatively correlated with platelet count (also significant under the mirror model, β = 19.86 10^9 cells/L, p = 2.5 × 10^{-8}). As expected, MVMR showed bidirectional causality between both traits, highlighting the challenges of interpreting their association separately.

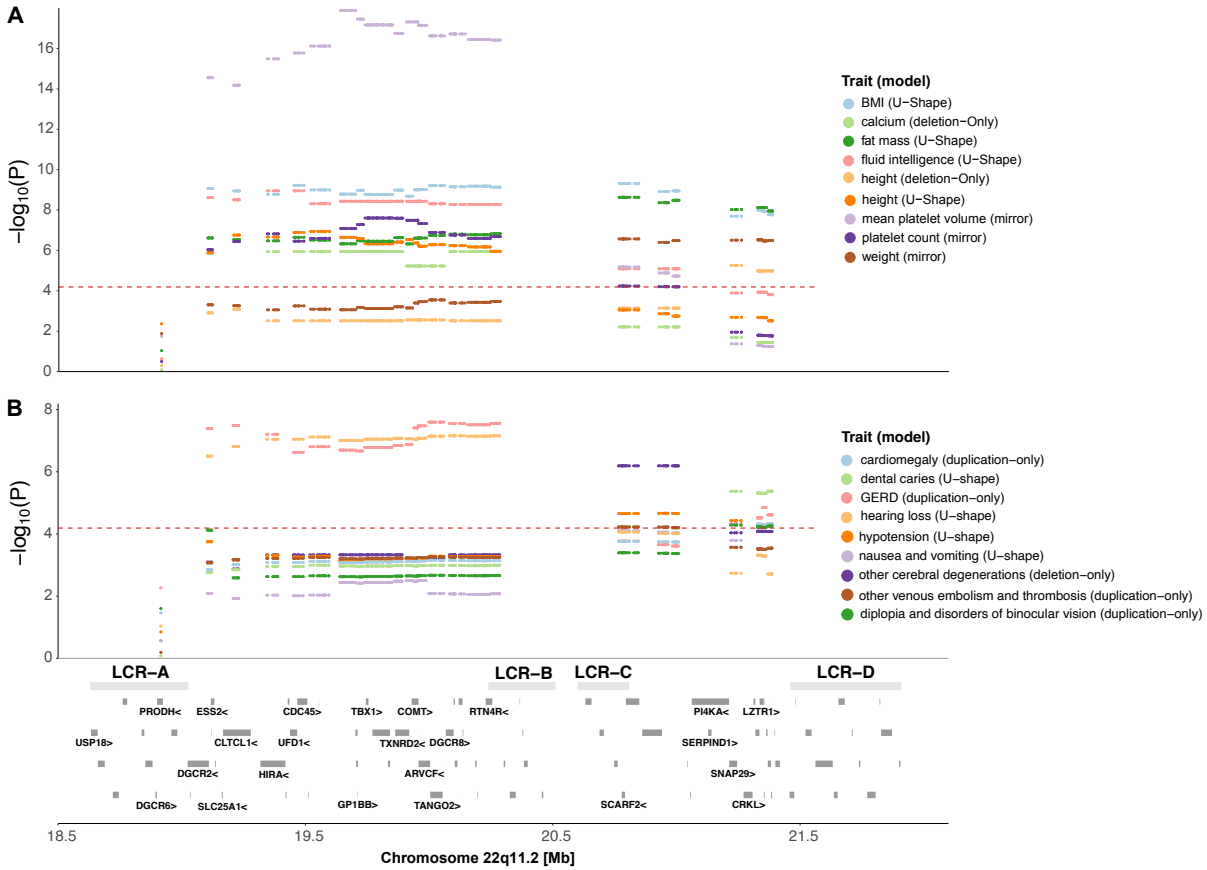


Figure 4.3: Phenotype associations with 22q11.2 CNVs.
(**A**) Continuous and (**B**) binary traits that are significantly associated with 22q11.2 CNVs according to the best association model (indicated in parenthesis). Each dot represents a CNV-proxy probe, with the negative logarithm of the association p-value (y-axis) plotted against the genomic coordinated (x-axis), colored according to the phenotype. The red dashed line indicates significance threshold (p < 6.5 × 10^{-5}). Bottom: Gray bars represent low copy-repeat region (LCR) A-D, as well as the 90 genes contained in the region. The 24 genes used for trait selection are labeled in black.

Table 4.1: Traits associated with CNVs in the 22q11.2 region with different models.
Reported effect sizes and p-values for each model refer to the lead signal of the most significant model for each trait (in bold).

		Duplication-only			Deletion-only			U-shape			Mirror		
A. Quantitative traits	**Genomic Position**	β	**95% CI**	**P**	β	**95% CI**	**P**	β	**95% CI**	**P**	β	**95% CI**	**P**
Mean platelet volume [fL]	chr22:19,639,383	-0.54	[-0.67, -0.41]	1.16×10^{-15}	1.66	[0.99, 2.32]	1.13×10^{-6}	-0.46	[-0.59, -0.33]	4.97×10^{-12}	**-0.58**	**[-0.71, -0.45]**	$\mathbf{1.31 \times 10^{-18}}$
BMI [kg/m^2]	chr22:20,765,989	**1.65**	**[1.15, 2.16]**	$\mathbf{1.55 \times 10^{-10}}$	-0.06	[-2.23, 2.12]	0.96	1.56	[1.07, 2.06]	4.9×10^{-10}	1.57	[1.08, 2.06]	4.23×10^{-10}
Whole body fat mass [kg]	chr22:20,765,989	3.17	[2.18, 4.16]	3.70×10^{-10}	-1.74	[-6.37, 2.88]	0.46	2.95	[1.98, 3.92]	2.33×10^{-9}	**3.11**	**[2.14, 4.07]**	$\mathbf{3.35 \times 10^{-10}}$
Fluid intelligence score [pt]	chr22:19,343,881	-1.21	[-1.64, -0.79]	2.25×10^{-8}	-3.76	[-6.04, -1.49]	0.001	**-1.3**	**[-1.72, -0.88]**	$\mathbf{1.12 \times 10^{-9}}$	-1.04	[-1.46, -0.63]	9.54×10^{-7}
Weight [kg]	chr22:20,765,989	3.83	[2.33, 5.32]	5.63×10^{-7}	-4.28	[-11.28, 2.73]	0.231	3.47	[2.01,4.94]	3.44×10^{-6}	**3.85**	**[2.38,5.31]**	$\mathbf{2.70 \times 10^{-7}}$
Height [cm]	chr22:21,219,710	-0.60	[-1.23, 0.03]	0.064	**-4.86**	**[-6.96,-2.77]**	$\mathbf{5.51 \times 10^{-6}}$	-0.95	[-1.56,-0.35]	0.002	-0.14	[-0.75, 0.46]	0.64
Height [cm]	chr22:19,518,079	-1.94	[-2.72, -1.15]	1.43×10^{-6}	-6.02	[-10, -2.04]	0.003	**-2.09**	**[-2.86, -1.32]**	$\mathbf{1.14 \times 10^{-7}}$	-1.64	[-2.41, -0.86]	3.26×10^{-5}
Platelet count [10^9 cells/L]	chr22:19,738,355	16.68	[9.56, 23.8]	4.43×10^{-6}	-100.09	[-135.83, -64.35]	4.05×10^{-8}	12.22	[5.24, 19.21]	0.0006	**19.86**	**[12.88, 26.85]**	$\mathbf{2.48 \times 10^{-8}}$
Calcium level [mmol/L]	chr22:19,207,491	0.01	[0, 0.02]	0.089	**-0.13**	**[-0.18, -0.08]**	$\mathbf{2.86 \times 10^{-7}}$	0.003	[-0.01, 0.01]	0.64	0.02	[0.01, 0.03]	0.004
B. Binary traits	**Genomic Position**	**OR**	**95% CI**	**P**	**OR**	**95% CI**	**P**	**OR**	**95% CI**	**P**	**OR**	**95% CI**	**P**
Gastroesophageal reflux disease	chr22:19,998,655	**2.72**	**[1.91, 3.88]**	$\mathbf{2.53 \times 10^{-8}}$	1.65	[0.21, 13.01]	0.63	2.68	[1.89, 3.79]	2.69×10^{-8}	2.66	[1.87, 3.79]	6.23×10^{-8}
Hearing loss	chr22:20,082,293	4.47	[2.49, 8.02]	5.32×10^{-7}	12.9	[1.58, 105.24]	0.017	**4.71**	**[2.68, 8.27]**	$\mathbf{6.95 \times 10^{-8}}$	4.08	[2.22, 7.5]	5.87×10^{-6}
Cardiomegaly	chr22:21,370,246	3.53	[1.92, 6.47]	4.69×10^{-5}	4.78	[0.64, 35.95]	0.13	**3.6**	**[2.02, 6.45]**	$\mathbf{1.53 \times 10^{-5}}$	3.21	[1.71, 6.03]	0.0003
Dental caries	chr22:21,370,246	3.29	[1.85,5.85]	5.21×10^{-5}	5.94	[1.4, 25.12]	0.015	**3.51**	**[2.06, 5.99]**	$\mathbf{4.21 \times 10^{-6}}$	2.76	[1.48, 5.13]	0.001
Diplopia and disorders of binocular vision	chr22:21,219,710	6.23	[2.57, 15.09]	5.18×10^{-5}	7.31	[0.43, 124.7]	0.17	5.74	[2.37, 13.92]	0.0001	**6.24**	**[2.58, 15.1]**	$\mathbf{4.89 \times 10^{-5}}$
Other venous embolism and thrombosis	chr22:20,765,989	7.6	[2.82, 20.46]	6×10^{-5}	18.57	[1.01, 340.01]	0.049	7.24	[2.69, 19.49]	8.9×10^{-5}	**7.61**	**[2.83, 20.46]**	$\mathbf{5.86 \times 10^{-5}}$
Other cerebral degenerations	chr22:20,927,716	1.76	[0.44, 7.07]	0.428	**45**	**[10.05, 201.43]**	$\mathbf{6.43 \times 10^{-7}}$	3.38	[1.26, 9.1]	0.016	0.21	[0.01, 3.57]	0.28
Hypotension	chr22:20,927,716	3.16	[1.79, 5.6]	7.7×10^{-5}	7.39	[0.89, 61.65]	0.065	**3.30**	**[1.9, 5.73]**	$\mathbf{2.16 \times 10^{-5}}$	2.91	[1.61, 5.25]	0.0004
Nausea and vomiting	chr22:21,370,246	2.17	[1.42, 3.31]	0.0003	3.67	[1.09, 12.37]	0.036	**2.28**	**[1.53, 3.39]**	$\mathbf{5.34 \times 10^{-5}}$	1.92	[1.24, 2.98]	0.003

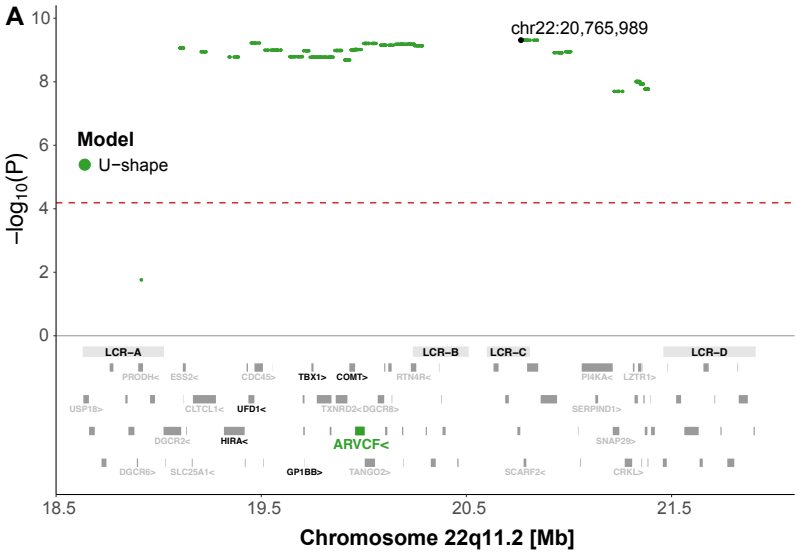


Figure 4.4: 22q11.2 CNVs and body mass index.
(**A**) Top: the negative logarithm of the association p-value for the U-shape CNV-body mass index (BMI) association scan (y-axis) is plotted against the 22q11.2 genomic region (x-axis). Each point represents a CNV-proxy probe and the lead signal (chr22:20,765,989) is shown in black. The red dashed line indicates significance threshold ($p < 6.5 \times 10^{-5}$). Bottom: low-copy repeat (LCR) A–D region as well as the 90 genes contained in the region. The 24 genes linked to traits according to HPO are labeled and genes linked to BMI through HPO are labeled in black. *ARVCF* expression was found to causally influence BMI through TWMR and is shown in green. (**B**) Boxplots representing BMI in individuals grouped according to their copy-number state of the lead signal probe; dots show the mean; outliers are not shown. N indicates the sample size for each category. (**C**) Representation of the TWMR analysis showing SNPs as instrumental variables (IVs), *ARVCF* gene expression as exposure, and its causal effect size ($\beta = 0.05$) on BMI.

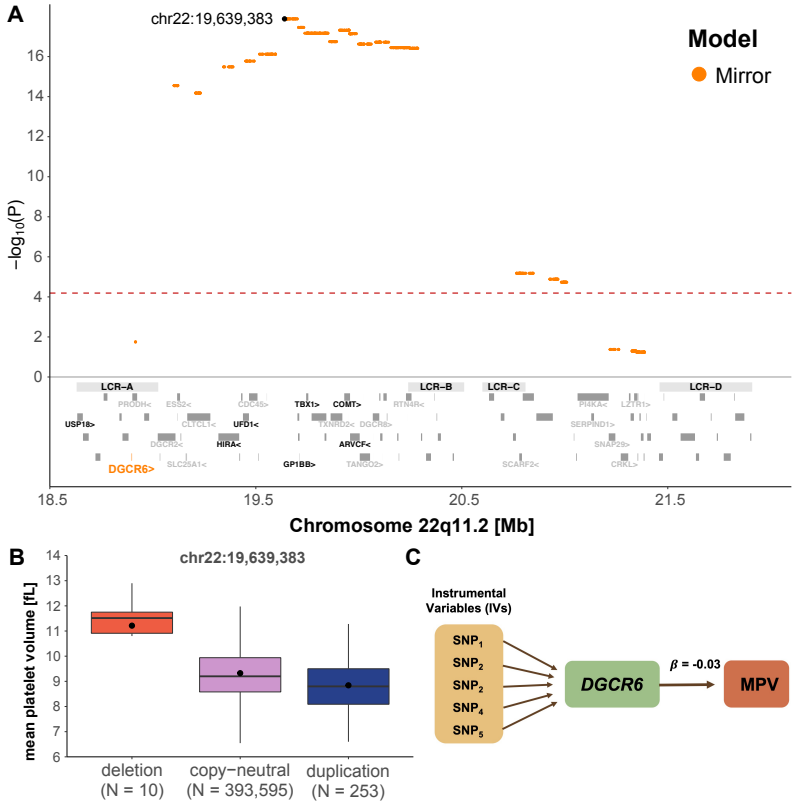


Figure 4.5: 22q11.2 CNVs and mean platelet volume.
(**A**) Top: the negative logarithm of the mirror association p-value for the CNV-mean platelet volume (MPV) association (y-axis) is plotted against the 22q11.2 genomic region (x-axis). Each point represents a CNV-proxy probe and the lead signal (chr22:19,639,383) is shown in black. The red dashed line indicates significance threshold ($p < 6.5 \times 10^{-5}$). Bottom: low-copy repeat (LCR) A–D region as well as the 90 genes contained in the region. The 24 genes linked to traits according to HPO are labeled and genes linked to mean platelet volume through HPO are labeled in black. *DGCR6* expression was found to causally influence mean platelet volume through TWMR and is shown in orange. (**B**) Boxplots representing mean platelet volume in individuals grouped according to their copy-number state for the lead signal probe; dots show the mean; outliers are not shown. N indicates the sample size for each category. (**C**) Representation of the TWMR analysis showing SNPs as instrumental variables (IVs), *DGCR6* gene and its causal effect size (β = -0.03) on MPV.

Unlike other phenotypes, height was associated under different models in distinct regions. The U-shape model appeared as the most significant model in the region spanning LCR A-B (β = -2.09 cm, p = 1.1×10^{-7}), while the deletion-only model was the only significant one at the distal portion between LCR C-D (β = -4.86 cm, p = 5.5×10^{-6}) (Figure 4.6A). Given this unexpected pattern, we stratified CNVs according to LCR categories (Figure 4.2A) to inspect their impact on height. Within LCR A-B and LCR A-D (Figure 4.6B-C), both duplications and deletions were associated with a height decrease in concordance with the U-shape model. However, duplications and deletions within LCR C-D had opposing effects on height, in line with a mirror model, which was confirmed by linear regression (β = 0.17 cm, p = 0.0003) (Figure 4.6D).

Given the low number of deletion carriers affected by binary outcomes (0–3 carriers) (Table S4.7), associations found under the U-shape or mirror models often reflect the effect of duplications (i.e., the most common CNV type) in these phenotypes. One example is gastroesophageal reflux disease (MIM: 109350), which was found to be associated under the duplication-only model (OR = 2.72, p = 2.53×10^{-8}) and had a stronger association occurring in the LCR A-B region (Figure 4.7A), indicating an increased prevalence of gastroesophageal reflux disease among duplication carriers (Figure 4.7B).

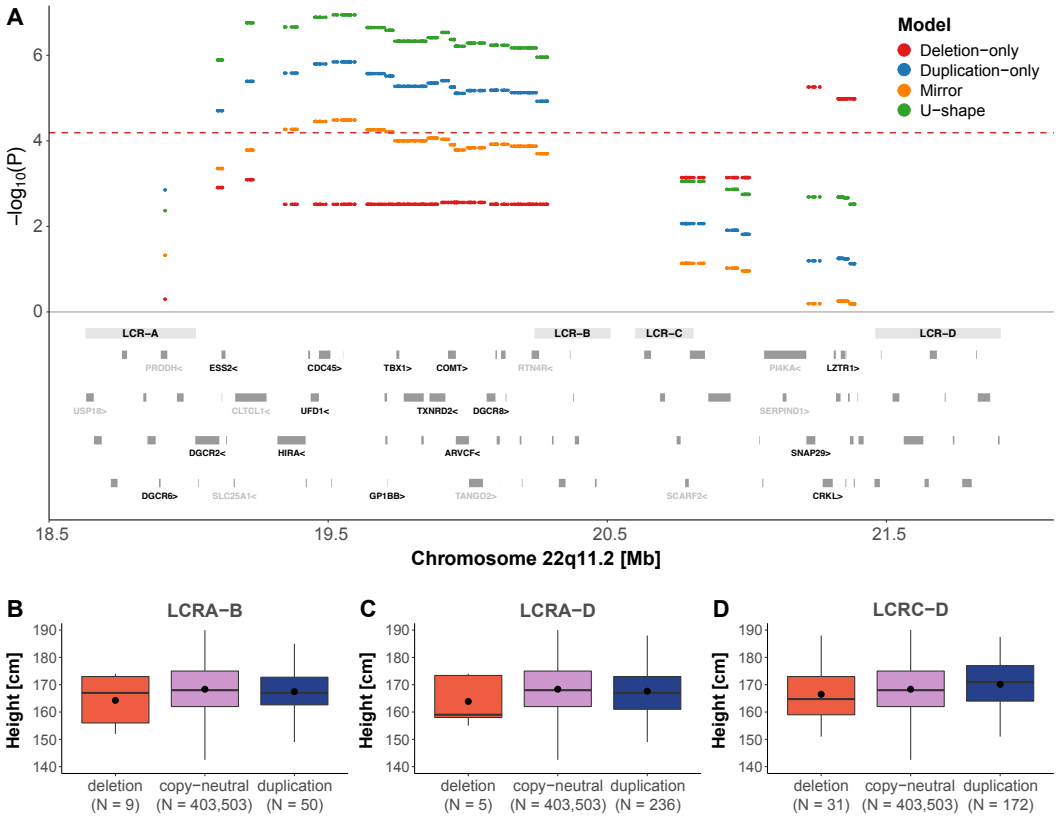


Figure 4.6: 22q11.2 CNVs and height.
(**A**) Top: the negative logarithm of the association p-value for the CNV-height association according to a deletion-only (red), duplication-only (blue), mirror (orange), and U-shape (green) is plotted against the 22q11.2 genomic region (x-axis). The red dashed line indicates significance threshold ($p < 6.5 \times 10^{-5}$). Bottom: low-copy repeat (LCR) A–D region as well as the 90 genes contained in the region. The 24 genes linked to traits according to HPO are labeled and genes linked to height through HPO are labeled in black. Boxplots representing height in individuals with (**B**) LCR A-B, (**C**) LCR A-D, and (**D**) LCR C-D CNVs grouped according to their copy-number state; dots show the mean; outliers are not shown. N indicates the sample size for each category.

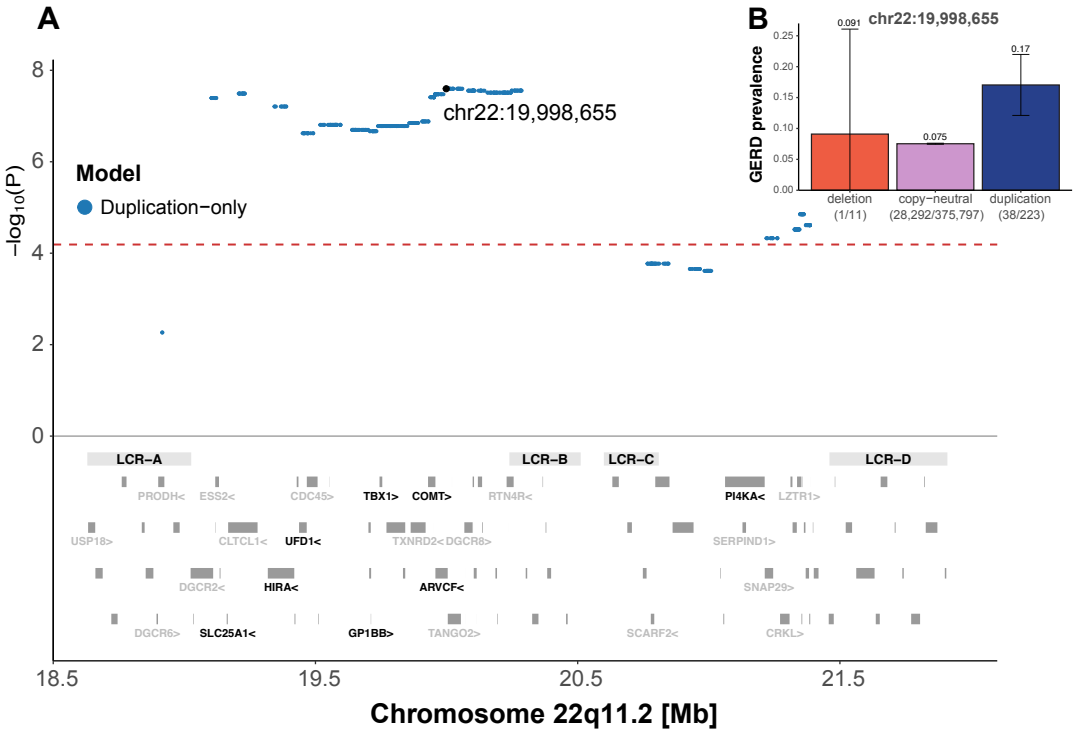


Figure 4.7: 22q11.2 CNVs and gastroesophageal reflux disease.
(**A**) Top: the negative logarithm of the duplication-only association p-value for the CNV-gastroesophageal reflux disease (GERD) association (y-axis) is plotted against the 22q11.2 genomic region (x-axis). Each point represents a CNV-proxy probe and the lead signal (chr22:19,998,655). The red dashed line indicates significance threshold ($p < 6.5 \times 10^{-5}$). Bottom: low-copy repeat (LCR) A–D region as well as the 90 genes contained in the region. The 24 genes linked to traits according to HPO are labeled and genes linked to mean platelet volume through HPO are labeled in black. (**B**) Bar plot representing prevalence (cases/total) of GERD grouped according to copy-number state for the lead signal probe. 95% confidence interval is depicted and for deletion is truncated at zero.

Enrichment analysis

For continuous traits, six out of eight assessed genes were found to have significantly greater association p-values for the group of unrelated traits compared to the group of linked traits for all association models. Binomial enrichment analysis indicated that CNV probes in genes linked to a given HPO term are 15 times more likely ($p < 6 \times 10^{-9}$) to show stronger association with the corresponding UKBB continuous trait. For the binary traits, however, only two out of 19 assessed genes were significant in the mirror model, which does not indicate an enrichment ($p = 0.07$).

Concordance in the direction of the effect between association scan and TWMR

Besides showing that differential expression of two 22q11.2 genes (*ARVCF* and *DGCR6*) causally affects two associated traits (BMI and mean platelet volume), TWMR results were also used to reinforce reliability of CNV associations. We evaluated concordance in the direction of effect sizes from nominally significant ($p < 0.05$) results of the mirror CNV association scan and nominally significant ($p < 0.05$) TWMR results (Table S4.8). As expected, we observed a significant agreement in effect size directions between both when fitting a linear regression line ($\beta = 1.6$, $p = 0.01$; Figure 4.8).

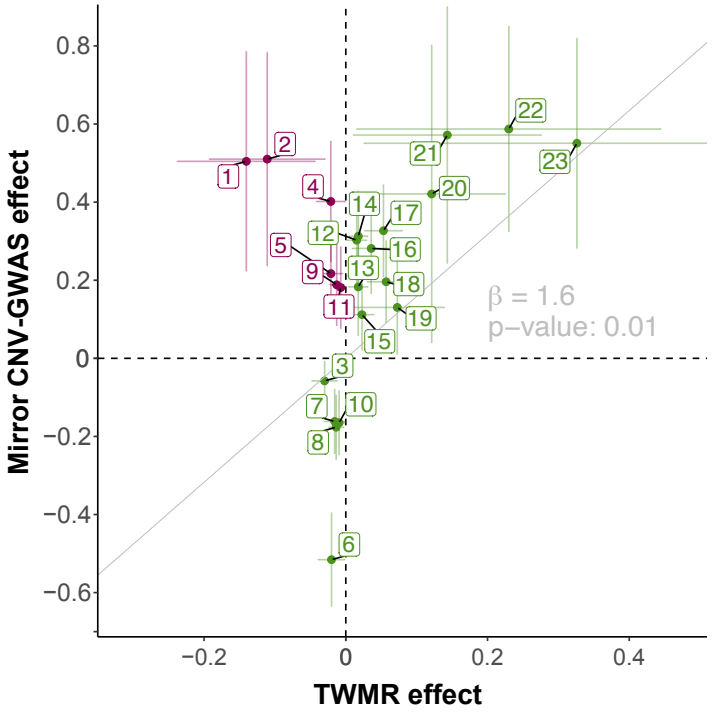


Figure 4.8: Concordance between TWMR and CNV effect sizes. Scatterplot depicting mirror CNV association scan (y-axis) versus transcriptome-wide Mendelian randomization (TWMR; x-axis) effect sizes. Vertical and horizontal bars represent the 95% confidence intervals. The (zero-intercept) regression line and the corresponding slope are in gray. For association scan effect sizes, the probe with the smallest p-value in the mirror model located in the TWMR gene was selected. Trait-gene pairs with agreeing direction between TWMR and CNV association scan are in green and trait-gene pairs with opposite directions are in pink. Labels indicate the following: *GNB1L* → (1) hypotension; (4) gastroesophageal reflux disease; (10) height; (11) weight. *P2RX6* → (2) cardiomegaly; (5) weight; (15) platelet count. *DGCR6* → (3) mean platelet volume; (19) nausea and vomiting; (20) diplopia and disorders of binocular vision. *CLDN5* → (6) mean platelet volume; (8) height; (9) weight; (14) platelet count. *TANGO2* → (7) height. *SLC25A1* → (12) BMI; (22) hearing loss. *CLTCL1* → (13) calcium levels. *ARVCF* → (16) whole body fat mass; (17) BMI; (18) weight; (21) cardiomegaly. *DGCR2* → (23) hypotension.

Causal links between traits and CNV pleiotropy

Cross-trait MVMR was performed for all 17 significantly associated traits. Out of a total of 289 trait-pair combinations, we identified 48 pairs that are causally linked to each other at nominal significance ($p < 0.05$) by using the IVW MR method. MVMR was then applied to these 48 combinations and 17 trait-pairs were significant after Bonferroni correction ($p < 0.05/289 = 0.0002$) (Figure 4.9A). Most traits were associated in a bidirectional manner, indicating that many (closely related) traits are mutually related to each other, most likely because of high genetic correlation. To distinguish between horizontal and vertical pleiotropy, we plotted the CNV effect on the outcome expected under vertical pleiotropy ($\beta_{\text{expected outcome}}$) against the effect observed in the association scan ($\beta_{\text{observed outcome}}$) to examine the concordance in effect direction (Figure 4.9B). This analysis revealed agreement only for very closely related trait pairs (driven by strong genetic correlation), such as platelet count - mean platelet volume, and indicated that, in general, pleiotropic CNV associations are not due to vertical but rather due to genuine horizontal pleiotropy.

Discussion

Most of our knowledge regarding the impact of CNVs in the 22q11.2 region in the general population stems from genome-wide studies (*208*, *292*–*294*, *300*, *302*). Here, we focused on this region specifically and developed a tailored set of analyses with more lenient, yet appropriate, significance threshold and in-depth follow-up analyses that allowed us to detect plausible associations missed by genome-wide studies (i.e., hearing loss, cardiomegaly, diplopia, and disorders of binocular vision). Our findings show that 22q11.2 CNV carriers in the general population that are likely on the milder end of the phenotypic spectrum are associated

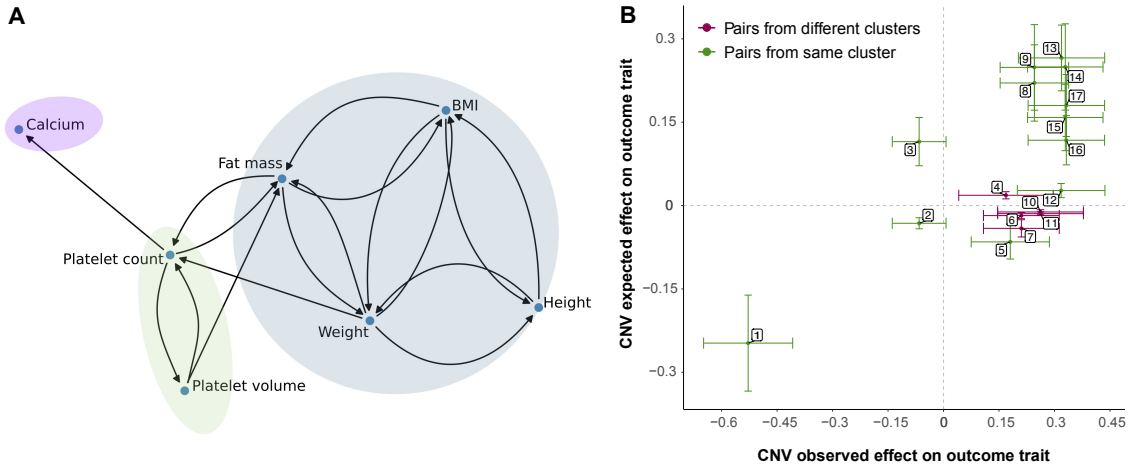


Figure 4.9: Concordance between CNV expected and observed effect on outcome trait.
(**A**) Causal links identified in the MVMR analysis. Colored clusters of traits are grouped based on phenotypic correlation ($r > |0.45|$). (**B**) Scatterplot depicting estimated CNV expected (y-axis) versus observed effect on outcome (x-axis) for each trait pair. Trait pairs from the same cluster in (A) are in green, while those from different clusters are in pink. The vertical and horizontal bars represent the 95% confidence intervals. Labels indicate exposure-outcome pairs: (1) platelet count-mean platelet volume; (2) BMI-height; (3) weight-height; (4) platelet count-calcium levels; (5) height-weight; (6) fat mass-platelet count; (7) weight-platelet count; (8) BMI-weight; (9) fat mass-weight; (10) platelet count-fat mass; (11) mean platelet volume-fat mass; (12) height-BMI; (13) mean platelet volume-platelet count; (14) BMI-fat mass; (15) weight-fat mass; (16) weight-BMI; (17) fat mass-BMI.

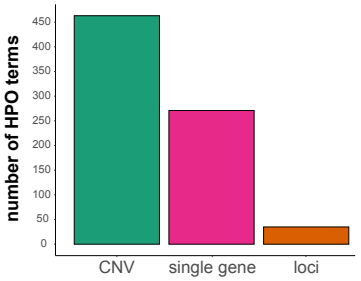


Figure 4.10: HPO terms linked to 22q11.2.
Barplot showing the number of HPO terms used in this study linked to conditions caused by 22q11.2 CNVs (green), genetic variants affecting a single gene in the region (pink) or linked to the 22q11.2 locus (orange).

with traits previously implicated by genes in the region, shedding light on the variable expressivity and penetrance of CNVs impacting this complex genomic region.

Assessed traits linked to 22q11.2 genes have been previously identified in different contexts including the 22q11.2 deletion and duplication syndromes, clinical conditions caused by variants in a single gene, and complex conditions associated with the locus (Figure 4.10). Therefore, using the HPO database to select investigated traits allowed us to leverage information from different genetic variants in a clinical context (*471*) to identify associations in the general population. For instance, we show that CNVs can impact traits previously known to be associated with individual genes in the region, such as cardiomegaly (*LZTR1* [MIM: 600574, 616564]) and other venous embolism and thrombosis (*SERPIND1* [MIM: 142360, 612356]), that were both associated under the duplication model in the distal region between LCR C-D, which harbors these genes.

Our enrichment analysis showed that for continuous – but not binary – traits, leveraging the HPO database for trait selection was an effective approach. This observation may stem from the fact that continuous traits are better powered and more accurate than binary traits, which may ignore underlying continuous phenomena. In addition, because association p-values for binary traits are closer to the multiple testing threshold, we expect weaker enrichment p-values.

Our results validated several known associations and shed light on traits that have not yet been extensively studied in the context of 22q11.2 CNVs. For instance, gastroesophageal reflux disease is not a vastly explored clinical feature in 22q11.2 deletion or duplication syndromes. While LCR A-D duplications have been previously associated with this trait in the UKBB cohort (*293*), replication of the association in our study emphasizes its relevance in 22q11.2 CNV carriers. Another relevant association identified in our study is with adult BMI. Obesity (MIM:

601665) (BMI > 30) has been previously described in adults with 22q11.2 DS (*473*). Even though this phenotype is not well described in clinical studies characterizing the 22q11.2 duplication syndrome, an increase in BMI has been associated with duplications in other studies assessing the UKBB (*208*, *294*). We further showed a causal effect of differential expression of *ARVCF* – a gene whose product is part of the catenin family and is involved in protein-protein interactions at adherent junctions – on BMI. Recently, a rare *ARVCF* missense variant of unknown significance has been identified in an individual with early-onset severe obesity (*474*), suggesting that *ARVCF* may play a role in the etiology of obesity.

Besides validating the link between CNVs in the 22q11.2 region and platelet count (*208*), we revealed a new association with mean platelet volume, which exhibits a true mirror effect, reinforcing the role of this genomic region in phenotypes such as thrombocytopenia. Thrombocytopenia (MIM: 313900) is a well-known clinical hallmark in 22q11.2 DS (*453*) but is not yet recognized as a clinical feature of the 22q11.2 duplication syndrome. *GP1BB* represents a top candidate to explain the observed platelet phenotypes as bi-allelic loss of function variants in the gene are responsible for Bernard-Soulier syndrome, a platelet disorder, and inclusion of *GP1BB* in the deleted region has been implicated in reduction of platelet count levels in 22q11.2 DS-affected individuals (*475*). Because of the lack of sufficient IVs, *GP1BB* could not be assessed by TWMR analysis, which instead revealed a causal effect of *DGCR6* differential expression on mean platelet volume. While *DGCR6*'s function is not yet clearly defined, it has been implicated in regulating other genes in the 22q11.2 region (*476*), suggesting that multiple genes in the region influence platelet phenotypes.

Usage of four different association models allowed for the identification of deletion-specific effects (e.g., calcium level) as well as traits in which duplications and deletions act in the same or opposite directions. By performing association scans at the probe level, we also showed that gain or loss of distinct segments within 22q11.2 may impact a trait following different association models, as was seen for height. Short stature has been identified for the 22q11.2 DS (*453*) but variable height measures have been described for the 22q11.1 duplication syndrome (*464*, *477*, *478*). In concordance with our study, both duplications and deletions (LCR A-D) have been previously associated with a decrease in height in the UKBB cohort (*294*). However, our study shows a mirror behavior involving the LCR C-D region. The impact of CNVs in the LCR C-D region is often overlooked or considered in combination with LCR A-B. However, the unexpectedly distinct impact of CNVs in this region on height, as well as certain traits that were only significant in this region (such as weight, cardiomegaly, other venous embolism and thrombosis, or dental caries), reveals the value of a more refined study of CNVs overlapping this complex region. It is important to note that gene dosage might not be the only mechanism underlying the CNVs' clinical impact, and gain/loss of different segments within 22q11.2 region could have distinct impacts over regulatory contacts, with diverse positional effect outcomes (*479*), adding complexity to the functional interpretation of the association models here described.

A drawback of studying pathogenic CNVs in a general population such as the UKBB is that the number of affected participants is low, as carriers of

22q11.2 CNVs with larger phenotypic impact are less likely to participate, a phenomenon often described as the healthy volunteer selection bias (*59*). As such, frequencies of the 22q11.2 deletions and duplications have not been precisely estimated outside of clinical cohorts. This task is complicated by the low frequency of 22q11.2 CNVs, which means that very large sample sizes are required to obtain precise estimates. For instance, a population-based French-Canadian cohort (N = 6,813) did not identify any 22q11.2 deletion carriers and only six duplication carriers (*480*), while a slightly larger study conducted in the Norwegian MoBa population-based cohort (N = 12,252) identified one 22q11.2 deletion carrier and six duplication carriers, resulting in frequency estimates of 0.008% and 0.05%, respectively, considering CNVs that overlapped in at least 50% with the region between LCR A-D (*481*). In the general population, using different available datasets, frequency of deletions and duplications encompassing the LCR A-B region have been estimated at 0.02% and 0.08%, respectively (*482*). Another study, in a population-based Danish cohort (N = 76,128), estimated a frequency of 0.03% for deletions and 0.07% for duplications considering the typical 3 Mb and 1.5 Mb CNVs (*483*). In our work, the frequency of CNVs in LCR A-B and LCR A-D is 0.07% for duplications and 0.003% for deletions. It is worth noting that we consider smaller nested CNVs between LCR A-B that were not appreciated in previous studies, indicating that if we applied similar definitions to these works, our frequency estimates would be lower. Although clinically ascertained cohorts overestimate the 22q11.2 carrier frequency, our study, because of healthy volunteer bias, underestimates it. However, adjusting carrier frequency estimates for such ascertainment is very difficult because the estimated frequency is very low, and we lack population reference data with variables relevant to the presence of 22q11.2.

While the absolute number of CNV carriers considered in our study is still larger than the sample size of some clinical cohorts, these individuals tend to exhibit milder phenotypes. This hampers statistical power to detect associations, especially for binary outcomes for which trait definition through grouping of ICD-10 codes is imperfect and arbitrary and case number can be extremely low. We offer corroborating evidence of our findings' reliability by performing sensitivity analyses and examining the concordance of CNV findings with TWMR effects. Importantly, effects observed in our study are potentially smaller than the ones observed in clinical cohorts (*237*), as they are mainly derived from CNV carriers with sub-clinical phenotypes and thus represent lower bound estimates. While in theory estimates from clinical cohorts might offer upper bound estimates, their poor and unstandardized reporting makes it difficult to establish accurate comparisons. Still, we hope that our study offers a better understanding of the spectrum of phenotypic consequences exerted by 22q11.2 and will improve diagnostic rates in individuals with low expressed phenotypes, as molecular diagnostic of genomic syndromes still often relies on recognition of characteristic signs to guide genetic testing. The co-occurrence of a series of sub-clinical signs in the same individual should increase the support for a diagnosis of a genomic imbalance at 22q11.2. In addition to diagnostic improvement, as the genotyping-first approach becomes more common in clinical practice, the accurate description of the phenotypes associated with 22q11.2 variants can benefit the prognosis of individuals in which a

genomic variant was already detected.

In conclusion, we found that 22q11.2 CNVs affect traits compatible with clinical manifestations seen in the genomic disorders within the general population. The probe-level association scan revealed that dosage of different segments within the 22q11.2 region may impact a trait through different mechanisms, as illustrated with height. Besides yielding further insights into the complex 22q11.2 region, our study provides a framework that can be adapted to study the phenotypic consequences of other clinically relevant genomic regions.

Acknowledgments

We thank all UK Biobank participants for sharing their data. All computations have been performed on the JURA computer infrastructure of the University of Lausanne. This work was supported by funding from the Department of Computational Biology (ZK) and the Swiss National Science Foundation (310030_189147) as well as financial support from Fundação de Amparo à Pesquisa do Estado de São Paulo (2020/11241–2, MZ; 2019/21644–0 MIM) and from the Coordenaçãde Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES).

Declarations of interest

The authors declare no competing interests.

Supplemental tables

Supplemental tables are available for download as a single Excel file.

- **Table S4.1** ICD-10 codes or self-reported diseases excluded from each phenotype.
- **Table S4.2** Mapping of HPO terms to UKBB binary traits.
- **Table S4.3** Sample size of UKBB binary traits.
- **Table S4.4** Mapping of HPO terms to UKBB continuous traits.
- **Table S4.5** Description of analyzed UKBB continuous traits.
- **Table S4.6** Source of SNP-GWAS summary statistics.
- **Table S4.7** Number of individuals of each significant binary trait.
- **Table S4.8** Nominal TWMR results.

16p11.2 BP2-3 | 5

One can state, without exaggeration, that the observation of and the search for similarities and differences are the basis of all human knowledge.

– Alfred Nobel

5.1 Aims 147
5.2 Key Findings 147
5.3 Author Contributions 148
5.4 Chromosomal deletions on 16p11.2 encompassing *SH2B1* are associated with accelerated metabolic disease 149

This chapter describes *"Chromosomal deletions on 16p11.2 encompassing SH2B1 are associated with accelerated metabolic disease"* (412), which was published in *Cell Reports Medicine*. The version presented in the dissertation includes supplemental material.

This translational project is the fruit of a collaboration with the team of Sadaf Farooqi at the University of Cambridge, UK. Increasingly, common complex diseases are seen as aggregates of multiple rarer conditions whose pathogenic mechanisms converge onto a similar outcome. Knowledge about these mechanisms can guide targeted therapeutic approaches. Expertise in the genetics of obesity of the Farooqi lab, complemented by our experience in studying CNVs in population biobanks allowed us to to characterize the metabolic disease of a form of obesity caused by a rare deletion at 16p11.2 BP2-3. Results of this study will hopefully inform clinical guidelines.

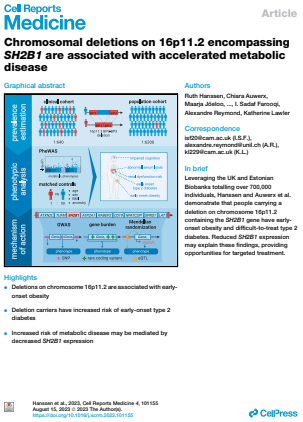


Figure 5.1: Cover of *Hanssen & Auwerx et al., 2023.*

5.1 Aims

SH2B1 deficiency is linked to severe obesity and an ongoing phase III clinical trial is testing the efficacy of the melanocortin-4 receptor (MC4R) agonist Setmelanotide as a weight loss therapy in individuals with genetic mutations in the leptin-melanocortin pathway genes (Figure 5.2), which includes the adaptor protein SH2B1. Because 16p11.2 BP2-3 deletions represent a common form of SH2B1 deficiency, our study aimed to assess whether these individuals are good candidates to benefit from this therapy. Specifically, we aimed to:

1. Characterize the spectrum of metabolic alterations and disease trajectories in adult 16p11.2 BP2-3 deletion carriers in the UKBB, compared to the general population and BMI-matched controls.
2. Evaluate the individual contribution of the nine genes encompassed in the 16p11.2 BP2-3 region to metabolic phenotypes through various statistical approaches.
3. Compare findings to previously reported phenotypes linked to SH2B1 deficiency in clinical patient cohorts and mouse models.

5.2 Key Findings

We identified 59 unrelated 16p11.2 BP2-3 deletion carriers in the UK Biobank. These individuals suffered from a complex spectrum of metabolic conditions that included severe obesity with childhood onset. Compared

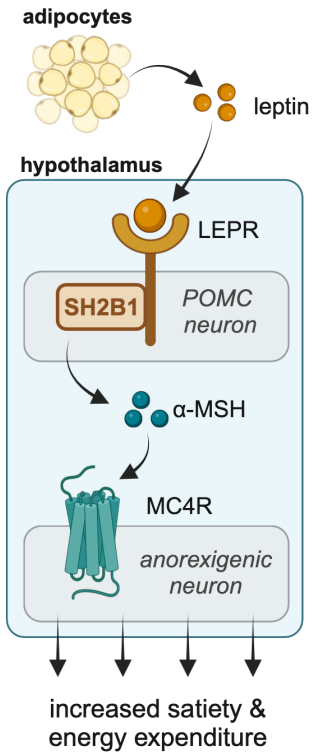


Figure 5.2: Leptin satiety pathway. Schematic representation of the leptin-melanocortin pathway of which SH2B1 is part. Briefly, in the fed state, adipocytes secrete the hormone leptin (orange). In the hypothalamus, leptin binds the leptin receptor (LEPR) at the surface of pro-opiomelanocortin (POMC) neurons. SH2B1 acts as an adaptor molecule that mediates intra-cellular signaling in POMC neurons upon binding of leptin to the LEPR. This results in the release of α-melanocyte-stimulating hormone (α-MSH; blue), which binds the melanocortin-4 receptor (MC4R) expressed on anorexigenic neurons, which in turn will signal the organism to decrease energy intake by inducing satiety and will stimulate energy expenditure. Mutations in key genes of this pathway will abolish the latter response, leading to continuous energy intake and severe obesity. Drawn based on (*484*).

to BMI-matched controls, deletion carriers exhibited higher prevalence, earlier-onset, and more difficult-to-treat type 2 diabetes, as well as increased inflammation and kidney damage, indicating that these traits are not a mere secondary consequence of the weight excess induced by the deletion.

Other key features of the metabolic syndrome were not affected beyond what is expected from the increased weight, in line with a monogenic etiology of obesity. Specifically, and compared to BMI-matched controls, deletion carriers were not at increased risk for cardiovascular events, had a slightly lower diastolic blood pressure despite no difference in hypertension risk, and had globally reduced serum lipid levels, except for triglycerides, whose levels were not distinguishable from controls.

More unclear is the impact of the deletion on the liver. Hepatic enzymes were increased in a BMI-dependent fashion but only levels of alkaline phosphatase and total bilirubin remained elevated after controlling for BMI, possibly indicating gallbladder disorders. Further investigation is needed to clarify whether deletion carriers suffer from hepatic steatosis.

Finally, transcriptome-wide Mendelian randomization (TWMR) highlighted decreased expression of *SH2B1* in the brain, adipose tissue, and muscle as a potential mediator of the increased type 2 diabetes risk observed in deletion carriers. Still, available data were insufficient to unequivocally affirm the causal role of *SH2B1* and exclude a contribution of other genes in the deleted region, despite strong evidence of the gene's role in metabolic phenotypes from clinical and mouse studies.

5.3 Author Contributions

This study was originally conceptualized by Ruth Hanssen, Katherine Lawler, and Sadaf Farooqi, who approached Alexandre Reymond with a study plan.

Ruth Hanssen and I carried out the majority of the analyses. Specifically, I provided the UKBB CNV calls, which were manually reviewed by Katherine Lawler. I also performed the UKBB phenome-wide association scan, gathered evidence from rare protein-coding variant burden tests and common variant association studies, carried out the colocalization analysis, and coordinated the EstBB replication executed by Maarja Jõeloo and the Mendelian randomization analyses executed by Marie Sadler. Ruth Hanssen carried out the matched control analysis. Statistical analyses were supervised by Katherine Lawler and Zoltán Kutalik.

Results were interpreted by Ruth Hanssen, Katherine Lawler, Sadaf Farooqi, Alexandre Reymond, Zoltán Kutalik, and myself. Ruth Hanssen, Katherine Lawler, Sadaf Farooqi and I designed the figures and drafted the manuscript, with critical revisions made by Zoltán Kutalik and Alexandre Reymond.

5.4 Chromosomal deletions on 16p11.2 encompassing *SH2B1* are associated with accelerated metabolic disease

Ruth Hanssen [1,†], **Chiara Auwerx [2,3,4,5,†]**, Maarja Jõeloo [6,7], Marie C. Sadler [3,4,5], Estonian Biobank Research Team [7], Elana Henning [1], Julia Keogh [1], Rebecca Bounds [1], Miriam Smith [1], Helen V. Firth [8], Zoltán Kutalik [3,4,5], I. Sadaf Farooqi [1,†,*,○], Alexandre Reymond [2,†,*], and Katherine Lawler [1,*,§].

Abstract

New approaches are needed to treat people whose obesity and type 2 diabetes (T2D) are driven by specific mechanisms. We investigate a deletion on chromosome 16p11.2 (breakpoint 2–3 [BP2-3]) encompassing *SH2B1*, a mediator of leptin and insulin signaling. Phenome-wide association scans in the UK (N = 502,399) and Estonian (N = 208,360) biobanks show that deletion carriers have increased body mass index (BMI; $p = 1.3 \times 10^{-10}$) and increased rates of T2D. Compared with BMI-matched controls, deletion carriers have an earlier onset of T2D, with poorer glycemic control despite higher medication usage. Cystatin C, a biomarker of kidney function, is significantly elevated in deletion carriers, suggesting increased risk of renal impairment. In a Mendelian randomization study, decreased *SH2B1* expression increases T2D risk ($p = 8.1 \times 10^{-6}$). We conclude that people with 16p11.2 BP2-3 deletions have early, complex obesity and T2D and may benefit from therapies that enhance leptin and insulin signaling.

[1] University of Cambridge Metabolic Research Laboratories, Wellcome-MRC Institute of Metabolic Science and NIHR Cambridge Biomedical Research Centre, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK; [2] Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland; [3] Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland; [4] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; [5] University Center for Primary Care and Public Health, 1010 Lausanne, Switzerland; [6] Institute of Molecular and Cell Biology, University of Tartu, 51010 Tartu, Estonia; [7] Estonian Genome Centre, Institute of Genomics, University of Tartu, 51010 Tartu, Estonia; [8] Department of Clinical Genetics, Cambridge University Hospitals NHS Foundation Trust & Wellcome Sanger Institute, Cambridge, UK; **[†]Authors contributed equally; [*]Correspondence; [§] Lead contact**.

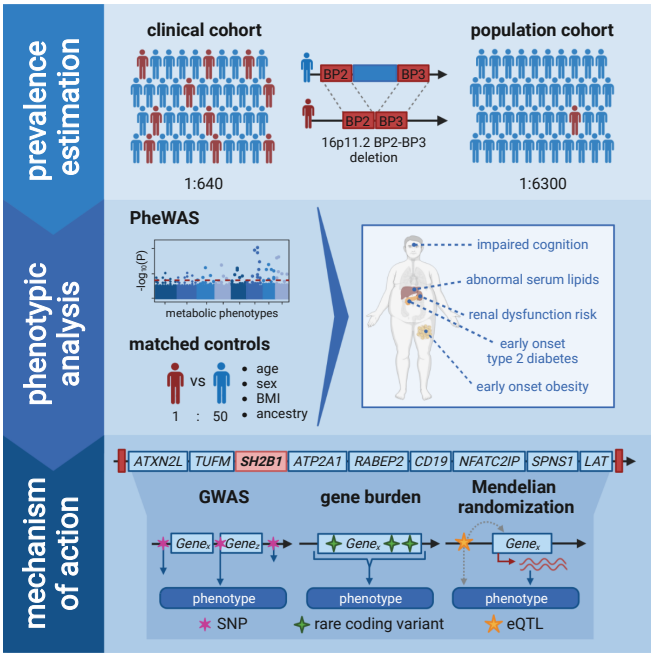


Figure 5.3: Graphical abstract of *Hanssen & Auwerx et al., 2023.*

Introduction

Obesity and type 2 diabetes (T2D) are highly prevalent, heterogeneous conditions associated with significant morbidity and mortality (*485*). The identification of subgroups of people whose metabolic disease is driven

by shared pathogenic mechanisms can inform approaches to treatment. This is exemplified by monogenic forms of obesity due to penetrant rare variants affecting the development and/or function of the hypothalamic leptin-melanocortin pathway (*486*). Some of these disorders can now be treated with licensed therapies, such as recombinant leptin for congenital leptin deficiency or the melanocortin-4 receptor (MC4R) agonist Setmelanotide for Leptin receptor (LEPR) (OMIM: 601007), POMC (OMIM: 176830), and PCSK1 (OMIM: 162150) deficiencies (*487*–*489*). *SH2B1* (Sarcoma homology 2 [SH2] B adaptor protein 1) (OMIM: 608937) acts as an intracellular adaptor that supports the assembly of proteins involved in leptin, insulin, and brain-derived neurotrophic factor (BDNF) signaling (*490*). *Sh2b1* knockout mice develop obesity, hyperglycemia, hepatic steatosis, and lipid accumulation in skeletal muscle (*491*–*493*). In humans, rare heterozygous loss-of-function mutations in *SH2B1* have been identified in children with hyperphagia, severe obesity, hyperinsulinemia, and maladaptive behavior (*494*–*496*). However, the trajectory of their metabolic disease in adulthood remains unclear.

Chromosome 16p11.2 contains five clusters of segmental duplications that increase the risk of recurrent copy-number changes at this locus through non-allelic homologous recombination (*497*) (Figure 5.4). Copy-number variants (CNVs; duplications or deletions) with breakpoints (BPs) at these clusters have been reported in clinical (*325*, *327*, *498*, *499*) and population-based cohorts (*208*, *292*, *293*, *295*). Rearrangement of the 600-kb proximal region (BP4–5) encompassing 33 genes (chr16:29.6–30.2 Mb; GRCh37) (OMIM: 611913) represents the most common deletion at the locus and has been associated with developmental delay, autism spectrum disorder (ASD), obesity, macrocephaly, and younger age at menarche (*208*, *292*, *293*, *295*, *421*, *422*, *426*, *500*). A smaller, 220-kb distal deletion (BP2-3; chr16:28.82–29.04 Mb; GRCh37) has been associated with early-onset obesity, macrocephaly, ASD, and schizophrenia (*279*, *325*, *327*, *501*), and increased rate of obesity and T2D in population-based cohorts (*208*, *292*, *293*, *295*). The latter interval encompasses *SH2B1* and eight other protein-coding genes (Figure 5.4).

In this study, we characterized the clinical spectrum associated with the 16p11.2 BP2-3 deletion in adults from two population-based cohorts,

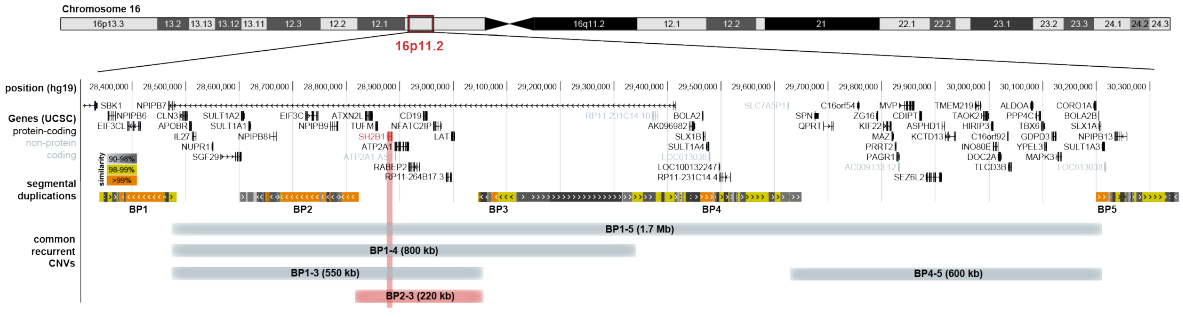


Figure 5.4: *SH2B1* encompassing 16p11.2 BP 2–3 deletions.
UCSC Genome Browser view of the 16p11.2 region (GRCh37/hg19). Upper track: exonic structure of genes in black (protein-coding) or gray (non-protein-coding). Middle track: segmental duplications forming the five breakpoint regions giving rise to recurrent copy-number variants (CNVs) at the 16p11.2 locus are colored according to the degree of similarity (light gray [90%] to orange [> 99%]). Lower track: minimally deleted or duplicated region encompassed by the most common CNVs in the region. Recurrent CNVs are named after the BP regions that frame them (approximate size). Exact breakpoints occur at variable locations within the breakpoint region so that exact genomic coordinates and CNV length may differ between individuals. The 16p11.2 BP2-3 region, which represents the minimal and most common *SH2B1* encompassing deleted region, is highlighted in red.

the UK Biobank (UKBB) and Estonian Biobank (EstBB). Individuals recruited to population-based cohorts are typically older and healthier than individuals in clinically ascertained cohorts, allowing us to test hypotheses about the development, severity, and treatment of diseases and their complications.

Materials and methods

Study material

UK Biobank

The UK Biobank (UKBB) is a voluntary-based cohort of 502,399 individuals (54% females) from the general UK population that were recruited at age 40–69 years (*61*). Participants signed a broad informed consent form for the usage of their data. This research was conducted using the UK Biobank resource under application numbers 16389 and 53821.

Phenome-wide association scan

Primary phenome-wide association scan (PheWAS) was carried out on a set of 404,977 individuals of mixed ancestry retained after filtering out i) related samples ($\leq 3^{rd}$ degree, preference given to 16p11.2 BP 2-3 deletion carriers), ii) CNV outliers (i.e., individuals genotyped on plates with an average CNV count/plate > 100 and individuals with > 200 CNVs or a single CNV > 10 Mb) (*208*), and iii) individuals with a duplication or non-manually validated deletion encompassed within chr16:28.6–29.2Mb. Among these, 59 unrelated ($\leq 1^{st}$ degree) 16p11.2 BP2-3 deletion carriers were retained (Figure 5.5). For all participants, self-reported gender and chromosomal sex were concordant. Participant characteristics are summarized in Table 5.1. Sensitivity analyses were carried out on a restricted set of 335,656 individuals of white British ancestry (`in.white.British.ancestry.subset = 1` in `ukb_sqc_v2.txt`) which comprised 52 deletion carriers.

Matched cohort study

We aimed to identify 50 body mass index (BMI)-matched UKBB participants for each of the 59 deletion carriers (Figure 5.5). Matched participants[1] were drawn randomly without replacement after excluding i) related UKBB participants ($\leq 3^{rd}$ degree) and ii) individuals with 16p11.2 BP2-3 deletion. We could not identify 50 matched participants for one deletion carrier of Bangladeshi ethnicity, who was therefore excluded. The final matched cohort analysis was performed on 58 deletion carriers and 2,900 matched control individuals. Participant characteristics are summarized in Table 5.1.

Software versions:

- CNV calling: PennCNV v1.0.5 (*203*).
- CNV QC: (*206*).
- UCSC Browser LiftOver (*414*).
- TWMR: smrivw v1.1 (*175*).
- Statistical analyses: R v3.6.1 & v4.1.1.
- Graphs: R v4.1.3.

1: **Matching criteria:**

- BMI (#21001): ± 2.5 kg/m^2.
- Age (#21003): ± 3.5 years.
- Sex (#31): identical.
- Ethnicity (#21000): identical.

Table 5.1: Characteristics of study participants.
Sample size, sex ratio (counts and percentage), and mean (± standard error [SE]) BMI and age for individuals studied in the PheWAS and matched cohort analysis. Deletion carriers are compared against non-carriers in the whole UKBB cohort (PheWAS) or BMI-matched controls (matched cohort analysis). Differences between the two groups were assessed through a chi-squared test (sex ratio) or Wilcoxon test (BMI and age) with the respective p-value displayed.

	PheWAS			**Matched cohort analysis**		
	Deletion carriers	**UKBB**	**P**	**Deletion carriers**	**Matched controls**	**P**
Sample size	59	404,918	-	58	2,900	-
Sex, male: female (%)	32:27 (54:46)	186,415:218,503 (46:54)	0.257	31:27 (53:47)	1,550:1,350 (53:47)	1
BMI [kg/m^2]	31.67 ± 0.72	27.40 ± 0.01	1.3×10^{-9}	31.66 ± 0.74	31.65 ± 0.10	0.991
Age [years]	54.54 ± 0.97	56.47 ± 0.01	0.046	54.71 ± 0.97	54.39 ± 1.35	0.752

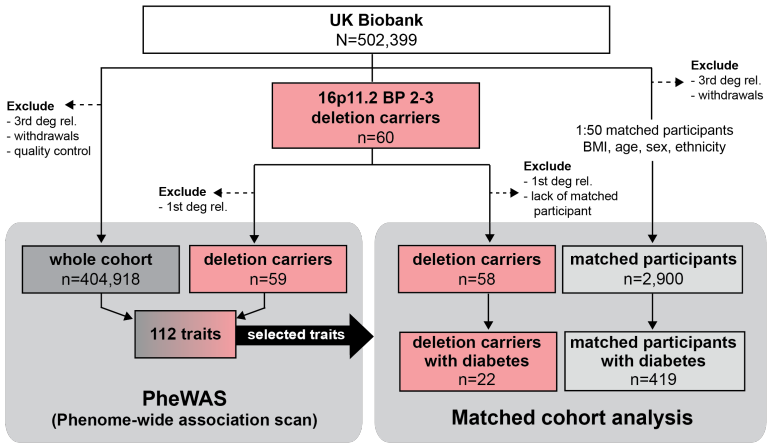


Figure 5.5: Study design.
Flow diagram (according to Consolidated Standards of Reporting Trials [CONSORT] principles) illustrating the detection of 16p11.2 BP2-3 deletion carriers in UKBB and the exclusion and inclusion criteria used to define the set of control individuals included in both the phenome-wide association scan (PheWAS) and matched cohort analysis. N represents the sample size of the whole UKBB, and n represents the subsets of individuals considered at various steps in the analysis. BMI = body mass index; deg. rel. = degree relatives.

Estonian Biobank

The Estonian Biobank (EstBB) is a population-based cohort encompassing ~20% of Estonia's adult population, with 208,360 individuals (65% females) in the data freeze 2022v01 (12/04/2022) (*62*). The activities of the EstBB are regulated by the Human Genes Research Act, which was adopted in 2000 specifically for the operations of the EstBB. Individual level data analysis in the EstBB was carried out under ethical approval 1.1–12/624 from the Estonian Committee on Bioethics and Human Research (Estonian Ministry of Social Affairs), using data according to release application 6–7/GI/2018 [2023/01/18] from the EstBB. All participants signed an informed consent form. Replication of association signals was carried out in a subset of 90,211 unrelated individuals of European ancestry after genotype/CNV quality control and pruning of related individuals (KING kinship coefficient > 0.0884) and preferentially including i) deletion carriers and ii) individuals with phenotypic measurements. Among these, 19 deletion carriers were retained.

Detection of 16p11.2 BP2-3 deletion

UK Biobank

Samples in the UKBB have been genotyped with either the Applied Biosystems UK Biobank Axiom Array or the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix, which share 95% probe overlap (*61*). We used two orthogonal approaches to identify high-confidence 16p11.2 BP2-3 deletion carriers: fully automated CNV-calling with quality scoring pipeline and manual review of the genotyping fluorescence signal across the 16p11.2 BP1-5 region. Data was acquired in GRCh37/hg19 and unless specified otherwise, genomic coordinates follow this reference build.

We performed fully automated deletion detection and quality scoring, as previously described for genome-wide CNV detection in UKBB (*208*) to detect CNVs fully contained in chr16:28.6–29.2 Mb. This pipeline is based on PennCNV (*203*) calls and offers the advantage of estimating breakpoints and assigning a probabilistic confidence quality score to each called deletion (*206*). To avoid using an arbitrary quality score cutoff to select deletion carriers, we manually reviewed fluorescent signal intensities (log R ratio, LRR) and B-allele frequency (BAF) in the 16p11.2 BP1-5 region (chr16:27-31 Mb) for the 272 deletion carriers identified

through our automated pipeline, ensuring that each of them had a median probe-level LRR < -0.2 in 4 adjacent 16p11.2 BP2-3 regions covered by 20 probes (chr16:28,835,900-28,881,001; chr16:28,883,241-28,914,162; chr16:28,914,458-28,9668,35; chr16:28,970,852-29,001,460). We identified 60 unambiguous 16p11.2 BP2-3 deletion carriers (i.e., with no evidence of other CNV in the BP1-5 region). We established that 51 (85%) of these 60 individuals had a quality score ≤ -0.5 (i.e., stringent cutoff used for genome-wide studies) and all samples harboring a deletion with a quality score ≤ -0.5 were retained by manual review. This indicates that while an automated approach represents a good alternative when manual review is not feasible, the latter allows to boost power by retaining a few additional deletion carriers. The 60 identified deletion carriers included one pair of first-degree relatives (i.e., likely inherited deletion) of which the parent was excluded so that a total of 59 unrelated deletion carriers were taken forward for analyses (Figure 5.6; Table S5.1). Individuals having a duplication or low confidence deletion (i.e., not validated by manual review) were excluded from the PheWAS (Figure 5.5).

Estonian Biobank
Illumina Global Screening Array genotype data was available for 193,844 individuals included in the single nucleotide polymorphism (SNP) imputation pipeline with matching genotype-phenotype identifiers and inferred versus reported sex, as well as a SNP-call rate ≥ 98%. Autosomal CNVs were called and quality-controlled as previously described, including exclusion of CNV outliers (*208*). Breakpoints of CNVs fully encompassed in chr16:28.8–29.1 Mb were visually inspected and retained if the distal coordinate of the deleted region encompassed or truncated *SH2B1* (chr16:28,857,921-28,885,526) and the proximal coordinate fully encompassed *LAT* (chr16:28,996,147-29,002,104). This resulted in 19 deletion carriers (63% females), among which 3 individuals had a fragmented deletion call.

Prevalence estimation of the 16p11.2 deletion

Prevalence of the 16p11.2 BP2-3 deletion in clinical and population cohorts was estimated based on literature review and data generated in this study (UKBB and EstBB estimates; Table S5.2). Prevalence in percentage was defined as the number of deletion carriers divided by the number of assessed individuals. To obtain estimates from the clinically ascertained DECIPHER database (accessed 27/05/2022) (*265*), we searched for CNVs affecting *SH2B1*, filtered for "Loss" to obtain deletions and retrieved 150 *SH2B1*-containing deletions. Deletions were further categorized according breakpoints by assigning the reported start and end of the deleted region to the closest segmental duplication obtained from UCSC segmental duplication track (accessed 06/07/2022; downloaded table: `genomicSuperDups` for chr16:21,000,000–34,800,000 (GRCh38), to match DECIPHER coordinates in GRCh38) (*502*, *503*). Prevalence of the 16p11.2 BP2-3 deletion was calculated as a proportion of the total number of patients reported in DECIPHER (N ≈ 45,700).

Phenome-wide association scan

Phenotype definition
Hundred and twelve traits, with an emphasis on metabolically relevant

phenotypes, were selected for association study with the 16p11.2 BP2-3 deletion carrier status. For all traits, entries encoded as "do not know" or "prefer not to answer" were set as missing. Exact definitions of these traits and summary statistics are provided in Tables S5.3-6.

2: **Physical measurements:**

- Adiposity (n = 11).
- Height (n = 2).
- Childhood/puberty (n = 7).
- Cardiovascular (n = 2).
- Cognitive/behavioral (n = 3).
- Physical activity (n = 5).
- Sleep-related (n = 3).

Thirty-three physical measurements[2] were treated as continuous variables (ordinal traits were re-coded as increasing continuous traits). Among these, four represent custom traits derived from existing data fields: systolic/diastolic blood pressures were inferred by completing *automated reading* (#4080/#4079) with *manual readings* (#93/#94) when the former was missing and waist-to-hip ratio (WHR) and WHR adjusted for BMI (WHRadjBMI) were calculated by dividing *waist circumference* (#48) by *hip circumference* (#49) and regressing out the effect of BMI and its interaction with sex for WHRadjBMI. We further assessed 58 biomarkers[3]. We included both raw and normalized (by *total fatty acids*; #23442) nuclear magnetic resonance (NMR) values for six fatty acid measurements. Continuous traits were inverse normal transformed before regressing out the effect of sex, age, age^2, genotyping batch, and principal components (PCs) 1–40. For blood measurements, we further corrected for *fasting time* (#74), as well as fasting time squared if the latter parameter was significantly ($p \leq 0.01$) impacting the trait when modeling $phenotype \sim fasting\ time + fasting\ time^2$.

3: **Biomarkers:**

- Blood biochemistry (n = 26).
- Urine assay (n = 2).
- NMR (n = 30).

4: **Binary traits:**

- ICD-10 codes (n = 13).
- Mental health conditions* (n = 4).
- Medication usage* (n = 4).

* = self-reported.

Twenty-one binary traits were evaluated[4]. For International Classification of Diseases, 10th Revision (ICD-10)-based diagnoses, age at diagnosis was computed by subtracting matching date at *first in-patient diagnosis – ICD10* (#41280) from the birth date, calculated from the individual's *month* (#52) and *year* (#34) *of birth* (birthday assumed on average to be the 15th). Results were converted in years by dividing by 365.25 to account for leap years.

Association study

Association between the 16p11.2 BP2-3 deletion carrier status (1 = deletion carrier; 0 = copy neutral; NA = duplication or non-manually validated deletion) and normalized, covariate-corrected continuous traits (i.e., physical and blood measurements) were assessed through linear regression (`lm()`). For binary traits, logistic regressions (`glm(family = binomial(link = "logit"))`) were used to model the effect of deletion carrier status on disease/phenotype risk. As no correction for covariates was performed on binary traits, sex, age, age^2, genotyping batch, and PC1-40 were included in the model. Model details are displayed in Tables S5.3-6.

Time-to-event analysis

To assess whether 16p11.2 BP2-3 deletion carrier status also influenced the age of onset of ICD-10-based diseases we used Cox proportional hazards models implemented in the `survival` R package (*416*). For this purpose, we used the earliest documented disease onset for cases and the date of the last reported diagnosis across all individuals (30/09/2021) minus the birth date converted in years for controls. Sex, age, age^2, genotyping batch, and PC1-40 were included in the regression model.

Multiple testing correction

Some of the 112 assessed traits are highly correlated and thus not independent. We accounted for this in our multiple testing strategy by calculating the number of effective tests, i.e., the number of tests required to explain

99.5% of the variance in the phenotypic dataset (*85*). This number was estimated to eighty-eight, both when considering all ancestries or only the subset of white British individuals, setting the strict threshold for genome-wide significance at $p \leq 0.05/88 = 4.7 \times 10^{-4}$ for the PheWAS. Nominal significance refers to $p \leq 0.05$.

Replication in the Estonian Biobank

Phenotype definition
Height, weight, and BMI were collected at recruitment. Traits were inverse normal transformed and corrected for sex, year of birth, genotyping batch (1–11), and PCs 1–20. Disease diagnoses are available as ICD-10 codes through cross-linking with national and hospital databases (last updated end of 2021) and were used to replicate the association with diabetes, defined as any of the E10-E14 codes. Exact definitions and summary statistics are found in Tables S5.3-4.

Association study
Association between the 16p11.2 BP2-3 deletion carrier status and normalized, covariate-corrected continuous traits (i.e., BMI, weight, height) and binary outcomes (i.e., diabetes) were performed using linear and logistic regressions, respectively, following the same procedure as described for UKBB. Sex, year of birth, genotyping batch (1–11), and PC1-20 were included as covariates for the association with diabetes. As all replicated signals were concordant in direction, we reported one-sided p-values, which were deemed significant at $p \leq 0.05/4 = 0.0125$ to account for the 4 performed tests.

Matched cohort study

Phenotype definition
Selected traits showing statistically significant or suggestive results in the PheWAS were followed up upon in our BMI-matched cohort study using curated phenotype definitions. Exact definitions and summary statistics are provided in Tables S5.7 and S5.8-9, respectively. Briefly, case definitions were obtained by combining ICD-10 codes (#41270) and information from self-reported diseases (#20002), disease-specific medication (#20003), and physical measurements or blood biomarkers at instance 0. Earliest documented age of onset was deduced from *date at first in-patient diagnosis – ICD10* (#41280), the age of onset of a self-reported condition, or *age when attended assessment center* (#21003) for physical measurements or blood biomarkers. Age at diagnosis was computed by adding the *age when attended assessment center* (#21003) to the difference between the *date of attending assessment center* (#53) and the date at diagnosis converted in years. Traits with no specific indication in Table S5.8 used the same definition as for the PheWAS.

Association study
Detailed methodology including covariates, statistical tests, and results are reported for each trait in the main text or in Tables S5.8-9. For continuous traits, linear models were implemented with `lm()` and `cohens_f()` from the package `effect size` v0.8.2. We considered the main effect (i.e., effect of the deletion compared to matched non-carriers as a baseline) and interactions with relevant covariates (e.g., lipid-lowering drug,

when assessing cholesterol levels). If continuous traits were not normally distributed, Wilcoxon rank-sum was applied (`wilcox.test()`) and effect sizes were estimated with `rFromWilcox()` (*504*). All *post hoc* analyses were performed using Tukey's procedure from the `lsmeans` package v2.30-0 (*505*, *506*). Nominal traits were assessed with logistic regression (`glm(family = binomial(link = "logit"))`) or with Fisher's exact test (`fisher.test()`) for which effect sizes were estimated as odds ratios (OR).

Time-to-event analysis
Association analyses between deletion carrier status and age at condition onset were implemented as previously described. We used the earliest documented age at disease onset for cases and the last documented age without diagnoses otherwise. To determine the latter, *age when attended assessment center* (#21003; instance 0 for physical measurements or blood biomarkers) and age of last documented ICD-10 diagnosis were considered. The age of the last documented ICD-10 diagnosis was determined by subtracting the *date of attending assessment center* (#53) from the last date of all *date at first in-patient diagnosis – ICD10* (#41280), converting the result in years by dividing trough 365.25 to account for leap years and adding it to the *age when attended assessment center* (#21003). Of the age when attended the assessment center (#21003) and the age of the last documented ICD-10 diagnosis, the oldest age was defined as the last documented age without diagnosis. Results were plotted with Kaplan-Meier curves.

Multiple testing correction
Reported p-values are nominal and two-sided. Bonferroni threshold for testing ~40 traits is $0.05/40 = 0.00125$.

Rare protein-coding variant burden tests

We used gene burden test results previously computed from 454,787 whole exome sequencing of the UKBB (*327*). Briefly, the study performed burden tests between ~18,800 genes and ~4,000 health-related traits using masks on variant function (i.e., predicted loss-of-function (pLoF)-only or pLoF and predicted deleterious missense variants) and minor allele frequency (MAF; i.e., MAF $\leq$ 1%, 0.01%, 0.001%, 0.0001%, or singletons). Association data with BMI, lipids, and T2D (defined as E11 ICD-10 code) were extracted for the nine genes in the 16p11.2 BP2-3 interval for all different test combinations and filtered for nominal significance ($p \leq 0.05$).

Common variant associations at 16p11.2 BP2-3

GWAS Catalog data
To determine whether common genetic variants in the 16p11.2 BP2-3 region had previously been found to impact traits we identified to be associated with the region's deletion, we used the 16p11.2 BP2-3 coordinates ± 50 kb (chr16:28,811,314-29,035,178 in GRCh38; (*507*)) and retrieved all mapped associations from the GWAS Catalog (accessed 22/12/2022) (*78*). Coordinates of retrieved associations were converted to GRCh37 with the UCSC LiftOver tool and involved traits were manually annotated with one of twelve trait categories.

Recombination rate estimation
Recombination rate was calculated by dividing the local difference in centimorgans (cM) by the local difference in Mb, using data from the HapMap (*508*) (Phase II) lifted over to GRCh37 (*508*).

Transcript Mendelian randomization

Transcriptome-wide Mendelian randomization (TWMR) was conducted following previously described methodology (*173*, *175*) to determine whether changes in transcript levels of genes in the deleted 16p11.2 BP2-3 region causally modulate T2D risk. Exposures (i.e., transcript levels) were instrumented with independent genetic variants ($r^2 < 0.01$), i.e., expression quantitative loci (eQTLs) for the gene of interest. Briefly, for the 6 genes with at least one eQTL (i.e., *ATXN2L*, *TUFM*, *SH2B1*, *AP2A1*, *NFATC2IP*, *SPNS1*), the effect of selected eQTLs on exposure (i.e., gene expression) and outcome (i.e., T2D risk) were used to estimate the causal effect of the former on the latter by inverse-weighted variance two-sample Mendelian randomization. Genetic effect sizes on transcript levels ($p \leq 1 \times 10^{-6}$) originate from either whole blood *cis*-eQTLs from the eQTLGen (*154*) or tissue-specific *cis*-eQTLs from the GTEx project (*151*) while those on T2D risk stem from a T2D genome-wide association study (GWAS) (*509*). Prior to the analysis, datasets were harmonized and palindromic variants or those that had an allele frequency difference > 0.05 between the datasets were removed. TWMR estimates were considered significant when $p \leq 0.05/9 = 5.6 \times 10^{-3}$ to account for the nine genes in the 16p11.2 BP2-3 interval. We used standardized genetic effect sizes, therefore TWMR estimates can be interpreted as the phenotypic impact of one standard deviation increase in expression. Since we expect the deletion to decrease expression, negative TWMR effects (i.e., increased expression decreases T2D risk) were considered directionally concordant with the association study results (i.e., deletion increases T2D risk).

Colocalization analysis

Genetic colocalization analysis was performed to determine whether genetically determined expression levels of the genes found to have a significant causal effect on T2D through TWMR (i.e., *TUFM*, *SH2B1*, *AP2A1*, *SPNS1*) shared a common genetic causal variant with the T2D GWAS signal. The same eQTL (*154*) and GWAS (*509*) summary statistics were used as in the TWMR analysis. Colocalization was performed with `coloc.abf()` from the R `coloc` package v5.1.0.1 (*167*), using a ± 250 kb window around the lead T2D GWAS signal (rs8046545; chr16:28,915,217; GRCh37) and following standard protocol. For each tested gene, coloc outputs the posterior probability supporting five different scenarios (*167*). Evidence for shared causal genetic signal from the eQTL and GWAS data (i.e., scenario H4) was considered when the posterior probability for that hypothesis was PP_H4 > 0.8.

Results

Prevalence of *SH2B1* encompassing 16p11.2 deletions

The UKBB is a cohort of 502,399 individuals (54% female) aged between 40 and 69 years at recruitment (*61*). To identify 16p11.2 BP2-3 deletion

carriers (DELs), we used an automated CNV calling pipeline (*208*) that feeds genotype microarray data to PennCNV (*203*) and attributes a probabilistic quality score (*206*) to each of the 272 deletions and 157 duplications identified across chr16:28.6–29.2 Mb (GRCh37). To avoid using an arbitrary quality score cutoff, fluorescent signal intensities (LRR) and BAF were manually reviewed in candidate deletion carriers, resulting in the detection of 60 unambiguous heterozygous deletion carriers with no other CNV in the 16p11.2 region. Of these, 51 (85%) had a quality score meeting the stringent cutoff (≤ -0.5) previously used in genome-wide studies with no manual validation of CNV calls (*208*). After excluding one individual from a pair of first-degree relatives, we retained 59 unrelated deletion carriers for further analysis (Figure 5.5 and Figure 5.6A; Table S5.1). These individuals comprised a similar proportion of males (DEL = 54%; UKBB = 46%; $p_{\chi^2} = 0.257$) and were slightly younger ($mean_{DEL}$ = 54.5 years; $mean_{UKBB}$ = 56.5 years; $p_{Wilcoxon} = 0.046$) than the whole UKBB cohort, with 52 (88.1%) individuals of self-reported and genetically estimated white British ancestry (Table 5.1; Figure 5.6B). In parallel and using a similar approach, we identified 19 unrelated deletion carriers in the EstBB, a population-based cohort coupled to the national health system that encompasses 208,360 Estonians (65% females) aged between 18 and 103 years.

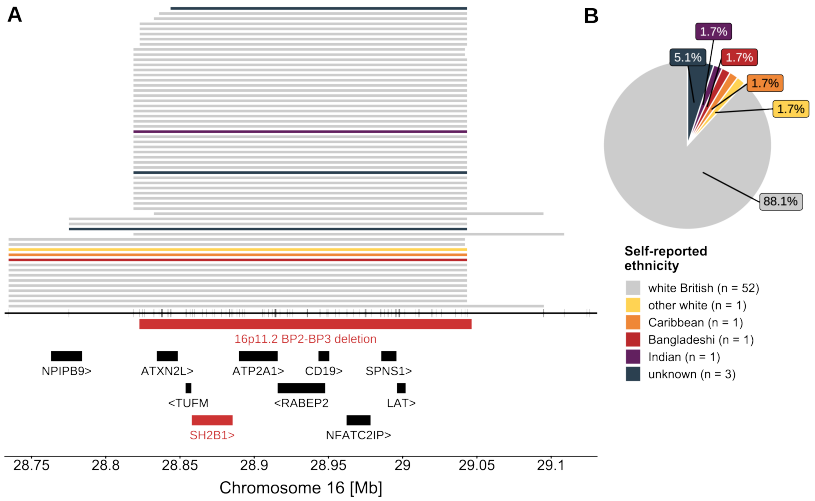


Figure 5.6: Characteristics of 16p11.2 BP2-3 deletion carriers in UK Biobank. (**A**) Breakpoints of the 59 unrelated 16p11.2 BP2-3 deletion carriers included in the phenome-wide association scan (PheWAS) determined through an automated CNV calling pipeline. Each line represents one individual according to self-reported ethnic background (legend in B). Vertical ticks indicate the location of genotyping probes on the microarray from which deletions were called (middle). Genomic location and orientation of the recurrently deleted BP2-3 region including *SH2B1* in red, along with other genes in the region in black. (**B**) Percentage of 59 unrelated deletion carriers belonging to each ethnic group; sample size indicated in the legend (n).

We estimated the BP2-3 deletion frequency in UKBB as 1 in 6,868 (0.016%), which is concordant with previous estimates in UKBB (*208*, *292*, *293*, *295*) and other population-based cohorts such as deCODE (*510*) (Table S5.2). The slightly higher prevalence in the EstBB of 1 in 4,748 (0.021%) is likely due to differences in enrollment criteria. In comparison, estimates from clinical cohorts of children ascertained for various conditions, including developmental delay, were about 10-fold higher (1 in 642; 0.156%) (Table S5.2). Among considered cohorts, DECIPHER had the highest prevalence of deletion carriers, with estimates of 1 in 435 (0.230%). This online repository provides both genetic and phenotypic description of ~45,700 patients with CNVs contributed by an international consortium of > 200 academic clinical centers and ≥ 1,600 clinical geneticists and diagnostic laboratory scientists (*265*). Specifically, 105 individuals carried the distal BP2-3 deletion; 24% of the 66 individuals on whom clinical information was available were reported to have obesity. Overall, our estimates are in line with results from a meta-analysis of 17 clinical and

population-based cohorts that found a 16p11.2 BP2-3 deletion prevalence of 1 in 613 (0.163%) and 1 in 7,343 (0.014%) among individuals with or without any of the 54 diseases investigated by the study, respectively (*229*).

PheWAS in 16p11.2 BP2-3 deletion carriers in UKBB

To gain insights into the clinical characteristics of 16p11.2 BP2-3 deletion carriers, we performed a PheWAS as a primary analysis, assessing 112 complex traits and hospital-diagnosed diseases (ICD-10 codes) in 59 deletion carriers versus 404,977 unrelated UKBB non-carriers (Figure 5.7; Tables S5.3-6). Estimating that the 112 traits correspond to 88 independent tests, we identified 23 strictly significant associations ($p \leq 0.05/88 = 4.7 \times 10^{-4}$) with deletion carrier status and 21 further nominally significant ones ($p \leq 0.05$). As a sensitivity analysis to ensure that results were not affected by population stratification, we repeated the PheWAS on 52 deletion carriers versus 335,656 unrelated non-carriers of white British ancestry. Estimates were in high agreement with those of the whole cohort (Pearson correlation = 0.987; $p < 2.2 \times 10^{-16}$) supporting the robustness of our findings (Figure 5.8; Tables S5.3-6).

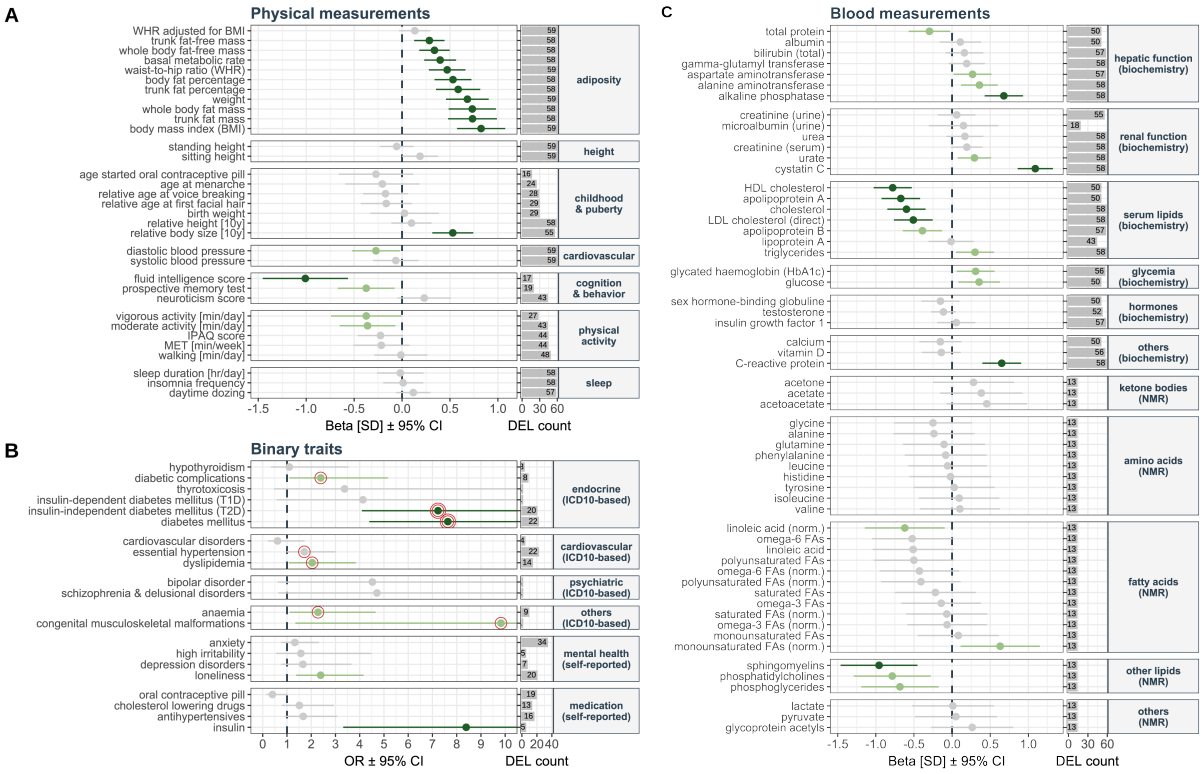


Figure 5.7: Phenome-wide association scan in carriers versus non-carriers of 16p11.2 BP2-3 deletions.
Results of the PheWAS for (**A**) 33 physical measurements, (**B**) 21 binary traits, and (**C**) 58 blood measurements according to trait category (y-axis). (A and C) Left panel, x-axis shows the effect of the deletion (beta) on each trait in standard deviations (SDs) with error bars representing 95% confidence intervals (CIs). (B) Left panel, x-axis shows the odds ratio (OR) with error bars representing the 95% CI. Upper range of the CI is truncated for some traits to facilitate visualization. Color indicates level of statistical significance: dark green ($p \leq 0.05/88 = 4.7 \times 10^{-4}$), light green ($p \leq 0.05$), and gray (non-significant). ICD-10-based diagnoses were assessed with a Cox proportional hazards model and strictly ($p \leq 0.05/88 = 4.7 \times 10^{-4}$) and nominally ($p \leq 0.05$) significant associations between deletion carrier status and early onset of the disease are indicated by a double or single red circle surrounding the OR, respectively. The vertical dashed line represents a null effect size. Right panel, x-axis indicates the number of deletion carriers (DEL, maximum n = 59) in whom the trait was measured (A and C) or the number of cases within deletion carriers (B).

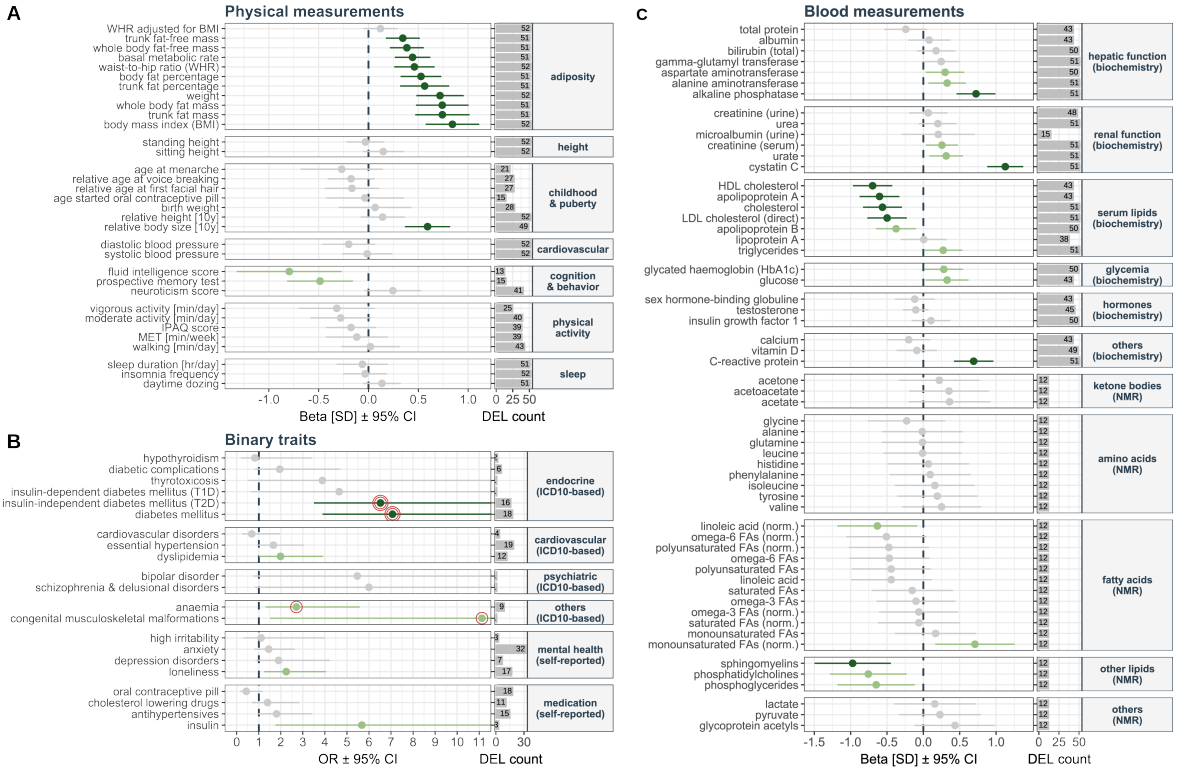


Figure 5.8: Sensitivity phenome-wide association scan in participants of white British ancestry.
Results of the PheWAS for (**A**) 33 physical measurements, (**B**) 21 binary traits, and (**C**) 58 blood measurements according to trait category (y-axis). (A and C) Left panel, x-axis shows the effect of the deletion (beta) on each trait in standard deviations (SD) with error bars representing 95% confidence intervals (CI). (B) Left panel, x-axis shows the odds ratio (OR) with error bars representing the 95% CI. Upper range of the CI was truncated for some traits to facilitate visualization. Color indicates level of statistical significance, dark green ($p \leq 0.05/88 = 4.7 \times 10^{-4}$), light green ($p \leq 0.05$), and gray (non-significant). ICD-10-based diagnoses were assessed with a Cox proportional hazards model and strictly ($p \leq 0.05/88 = 4.7 \times 10^{-4}$) and nominally ($p \leq 0.05$) significant associations between deletion carrier status and early onset of the disease are indicated by a double or single red circle surrounding the OR, respectively. The vertical dashed line represents a null effect size. Right panel, x-axis indicates the number of deletion carriers (DEL, maximum = 52) in whom the trait was measured (A, C) or the number of cases (B).

16p11.2 BP2-3 deletions have increased adiposity

We found that 16p11.2 BP2-3 deletion carriers were significantly more likely to have a higher BMI (β = 3.9 kg/m^2; p = 1.3×10^{-10}), weight (β = 10.8 kg; p = 2.0×10^{-9}), whole-body fat mass (β = 7.0 kg; p = 5.9×10^{-9}), and percentage fat mass (β = 4.5%; p = 5.9×10^{-8}) (Figure 5.7A). While waist-to-hip ratio appeared increased in deletion carriers (β = 0.47 standard deviation [SD]; p = 1.4×10^{-6}), the effect disappeared upon correction for BMI (β = 0.13 SD; p = 0.109), suggesting no difference in fat distribution. Increased adiposity appeared in childhood, with 41.4% of deletion carriers self-reporting to be "plumper at age 10" compared with 15.5% in the whole UKBB (p = 1.2×10^{-6}). Neither childhood (p = 0.359) nor adult (p = 0.531) height was significantly associated with deletion carrier status. These results were replicated in the EstBB, where we found a significant increase in BMI (β = 3.7 kg/m^2; p = 6.3×10^{-4}) and weight (β = 10.0 kg; p = 2.2×10^{-3}) among the 19 deletion carriers but no effect on height (Table S5.3).

16p11.2 BP2-3 deletion carriers have early-onset T2D that is difficult to treat

Our PheWAS indicated that 16p11.2 BP2-3 deletion carriers were at

significantly increased risk for T2D (odds ratio [OR] = 7.2; $p = 1.0 \times 10^{-11}$) with considerably earlier disease onset (hazards ratio [HR] = 6.1; $p_{Cox\text{-}PH} = 8.4 \times 10^{-16}$) and were more likely to receive insulin treatment (OR = 8.4; $p = 6.9 \times 10^{-6}$). They had nominally increased levels of glycated hemoglobin (HbA1c; β = 2.1 mmol/mol; p = 0.015) and random serum glucose (β = 0.4 mmol/L; p = 0.011) (Figure 5.7B-C). The increased risk of T2D among deletion carriers was replicated in the EstBB (OR = 7.3; $p = 2.5 \times 10^{-4}$; Table S5.4). To test whether these results were driven by increased adiposity in deletion carriers, we selected 50 controls (ctrl; unrelated non-carriers; i.e., UKBB participants without the deletion) matched for BMI, age, sex, and self-reported ethnicity for 58 of the 59 deletion carriers (excluding one individual with < 50 ethnicity-matched participants), amounting to a total of 2,900 matched non-carriers (Table 5.1). Disease cases were defined using additional curation of self-reported clinical data, medication usage, biomarker levels, and physical measurements in addition to ICD-10 codes (Table S5.7). Even after matching for adult BMI (Figure 5.9A), deletion carriers more frequently reported to be "plumper at age 10" (DEL = 41%; ctrl = 23%; p = 0.002; Figure 5.9B; Table 5.2), consistent with earlier onset of obesity. T2D prevalence was increased 2.7-fold (DEL = 38%; ctrl = 14%; p = 0.004; Figure 5.9C; Table 5.2) irrespective of body size at age 10 (all interactions DEL × body size have p > 0.27). Deletion carriers developed T2D at an earlier age than BMI-matched non-carriers (HR = 4.0; $p_{Cox\text{-}PH} = 1.6 \times 10^{-7}$; Figure 5.9D; Table 5.2). A higher proportion of the 22 deletion carriers with T2D reported usage of antidiabetic drugs compared with the 419 diabetic matched non-carriers (DEL = 59%; ctrl = 36%; p = 0.033; Figure 5.9E; Table 5.2) and they were prescribed a larger number of medications (p = 0.022; Figure 5.9E; Table 5.2). Despite higher antidiabetic medication usage, glycemic control measured by random serum glucose was worse in deletion carriers than in matched non-carriers with T2D ($p_{T2D \times DEL} = 0.006$; *post hoc* analysis among cases, $mean_{DEL}$ = 8.39 mmol/L; $mean_{ctrl}$ = 6.97 mmol/L; p = 0.018; Figure 5.9F; Table 5.2). A similar trend was observed for HbA1c levels ($p_{T2D \times DEL} = 0.002$; *post hoc* analysis, $mean_{DEL}$ = 53.3 mmol/mol; $mean_{ctrl}$ = 48.7 mmol/mol; p = 0.080; Figure 5.9G; Table 5.2).

Table 5.2: Metabolic characteristics of deletion carriers and BMI-matched controls.
Descriptive statistics reporting the prevalence or mean value (± standard error) for key metabolic phenotypes in deletion carriers and BMI-matched controls. Statistical significance of the difference between the two group is reported as a p-value. [a]Among people with documented diabetes.

Category	Trait	Deletion carriers	Matched controls	P
Adiposity	Prevalence of *plumper* at age 10 (%)	41.4	23.3	0.002
Glycemia	Prevalence of type 2 diabetes (%)	37.9	14.4	$<2 \times 10^{-16}$
	Age of onset of type 2 diabetes	51.1 ± 2.4	54.8 ± 0.5	9.1×10^{-8}
	Prevalence of diabetes treatment (%)	59.1	35.8	0.033
	Number of antidiabetic drugs[a]	1.69 ± 0.13	1.37 ± 0.04	0.022
	Glucose[a] [mmol/L]	8.39 ± 1.18	6.97 ± 0.17	0.018
	Glycated hemoglobin (HbA1c)[a] [mmol/mol]	53.3 ± 3.9	48.7 ± 0.7	0.080
	Prevalence of diabetes with complications (%)	31.8	25.8	0.534
Renal function	Cystatin C [mg/L]	1.077 ± 0.028	0.929 ± 0.003	6.0×10^{-14}
Inflammation	C-reactive protein (CRP) [mg/L]	4.84 ± 0.73	3.49 ± 0.09	0.015
Serum lipids	Total cholesterol [mmol/L]	5.04 ± 0.14	5.62 ± 0.02	5.8×10^{-5}
	Triglycerides [mmol/L]	2.10 ± 0.17	1.97 ± 0.02	0.926
Cardiovascular system	Prevalence of hypertension (%)	60.3	66.0	0.373
	Diastolic blood pressure [mmHg]	79.8 ±1.5	84.6 ± 1.9	2.8×10^{-4}
	Prevalence of cardiovascular diseases (%)	10.3	14.7	0.357

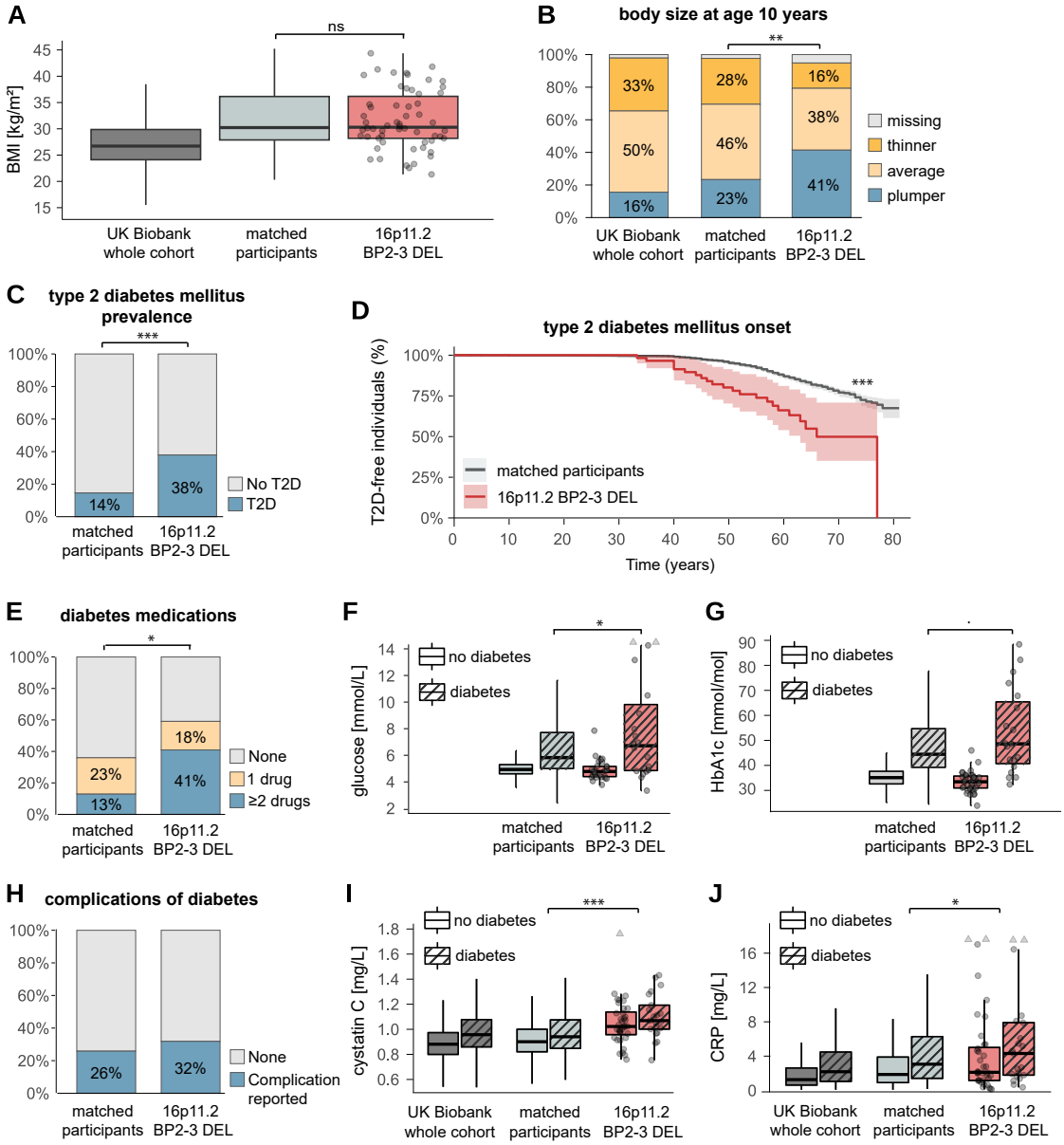


Figure 5.9: 16p11.2 BP2-3 deletion carriers are at increased risk for early-onset T2D compared with BMI-matched non-carriers. (**A**) BMI (kg/m^2) of deletion carriers (16p11.2 BP2-3 DEL; red; n = 59) compared with UKBB whole cohort (dark gray; n = 403,280) and BMI-matched participants (light gray; n = 2,900). (**B**) Proportion (%) of individuals self-reporting their comparative body size at age 10 as *plumper* (blue), *average* (light yellow), or *thinner* (dark yellow); missing data (gray) among UKBB whole cohort (n = 396,450), matched participants (n = 2,900), and deletion carriers (16p11.2 BP2-3 DEL; n = 59). (**C**) Prevalence (%) of T2D among deletion carriers (16p11.2 BP2-3 DEL; n = 58) and matched participants (n = 2,900). (**D**) Kaplan-Meier curves illustrating the proportion of T2D-free individuals (%) over time (years) among deletion carriers (16p11.2 BP2-3 DEL; red; n = 58) and matched participants (gray; n = 2,900). Shaded areas represent 95% CIs. (**E**) Proportion (%) of individuals taking no (gray), one (yellow), or several (blue) antidiabetic drugs among deletion carriers (16p11.2 BP2-3 DEL; n = 22) and matched participants with diabetes (n = 406). (**F**) Glucose (mmol/L) and (**G**) glycated hemoglobin (HbA1c) (mmol/mol) levels among deletion carriers (16p11.2 BP2-3 DEL; red; glucose n = 49; HbA1c n = 55) and matched participants (light gray; glucose n = 2,490; HbA1c n = 2,727) according to diabetic status. (**H**) Prevalence (%) of reported diabetic complications among deletion carriers (16p11.2 BP2-3 DEL; n = 22) and matched participants with diabetes (n = 406). (**I**) Cystatin C (mg/L) levels according to diabetic status in UKBB whole cohort (dark gray; n = 385,797), matched participants (light gray; n = 2,698), and deletion carriers (16p11.2 BP2-3 DEL; red; n = 58). (**J**) C-reactive protein (CRP) (mg/L) in UKBB whole cohort (dark gray; n = 384,965), matched participants (light gray; n = 2,691), and deletion carriers (16p11.2 BP2-3 DEL; red; n = 58). Boxplot outliers are not shown for the whole cohort and matched participants. Data points depicted for deletion carriers (circles; triangles indicate values cropped at the maximum of the depicted range); ns, $p > 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$, $^{***}p < 0.001$.

16p11.2 BP2-3 deletion carriers have increased risk of renal impairment

Although the overall occurrence of known diabetic complications (retinopathy, kidney failure, polyneuropathy; Table S5.7) was comparable in 16p11.2 BP2-3 deletion carriers and matched controls (Figure 5.9H; Table 5.2), levels of cystatin C, an early biomarker of kidney dysfunction, were significantly elevated in deletion carriers compared with both the whole UKBB cohort (β = 0.19 mg/L; p = 2.0×10^{-20}; Figure 5.7C) and matched non-carriers ($mean_{DEL}$ = 1.08 mg/L; $mean_{ctrl}$ = 0.93 mg/L; p = 6.0×10^{-14}; Figure 5.9I; Table 5.2), indicating that deletion carriers may be at increased risk of developing chronic kidney disease. Levels of C-reactive protein, a marker of chronic inflammation, were also increased in deletion carriers in both PheWAS (β = 2.8 mg/L; p = 5.1×10^{-7}; Figure 5.7C) and matched control analyses ($mean_{DEL}$ = 4.84 mg/dL; $mean_{ctrl}$ = 3.49 mg/dL; p = 0.015; Figure 5.9J; Table 5.2).

Hepatic steatosis is a common complication of obesity and T2D. Our PheWAS revealed increased serum levels of hepatic enzymes in deletion carriers (Figure 5.7C;) with significantly increased levels of alkaline phosphatase (ALP; β = 17.9 U/L; p = 1.2×10^{-7}) and nominally increased levels of alanine (ALT; β = 5.1 U/L; p = 3.6×10^{-3}) and aspartate (AST; β = 2.9 U/L; p = 0.034) aminotransferases. After controlling for alcohol consumption, diabetes, and lipid-lowering drugs, only ALP (p = 1.9×10^{-4}) and total bilirubin (p = 0.049) levels were increased in deletion carriers compared with BMI-matched non-carriers, while ALT, AST, and gamma-glutamyl transferase (GGT) levels did not differ between the groups. Very few ICD-10-documented cases of non-alcoholic fatty liver disease are reported in UKBB; accordingly, no association with deletion carrier status could be detected. Considering all liver diagnoses (K70–77), a higher proportion of deletion carriers was affected compared with matched non-carriers (p = 0.005). Specifically, deletion carriers had hepatic steatosis and cirrhosis diagnoses (mean age of onset = 64 years), possibly representing end-stage metabolic liver disease, which is often not accompanied by elevated liver enzymes.

To study dyslipidemia in the matched cohort setting, we considered ICD-10-coded and self-reported dyslipidemia, as well as blood-panel-derived cases (Table S5.7). Prevalence of dyslipidemia in deletion carriers was not increased after accounting for BMI (Figure 5.10A). However, the proportion of individuals with hypertriglyceridemia only or mixed dyslipidemia was increased in deletion carriers (DEL = 17%; ctrl = 9%, p = 0.029; Figure 5.10A), findings that may be explained by their suboptimal glycemic control. We observed that triglyceride levels were comparable between deletion carriers and matched non-carriers (Figure 5.10B; Table 5.2), while low-density lipoprotein (LDL)-cholesterol, total cholesterol, and apolipoproteins A and B levels were significantly decreased in deletion carriers compared with the whole UKBB cohort (Figure 5.7C) and matched non-carriers (all p < 0.003;Figure 5.10C). High-density lipoprotein (HDL)-cholesterol levels followed the same trend and were decreased compared with both the UKBB cohort (β = -1.13 mmol/L; p = 9.2×10^{-10}; Figure 5.7C) and matched non-carriers ($mean_{DEL}$ = 1.17 mmol/L; $mean_{ctrl}$ = 1.32 mmol/L; p = 4.8×10^{-9}; Figure 5.10D). There was no increase in the use of cholesterol-lowering drugs in deletion carriers in the PheWAS or matched cohort analysis (Figure 5.7B).

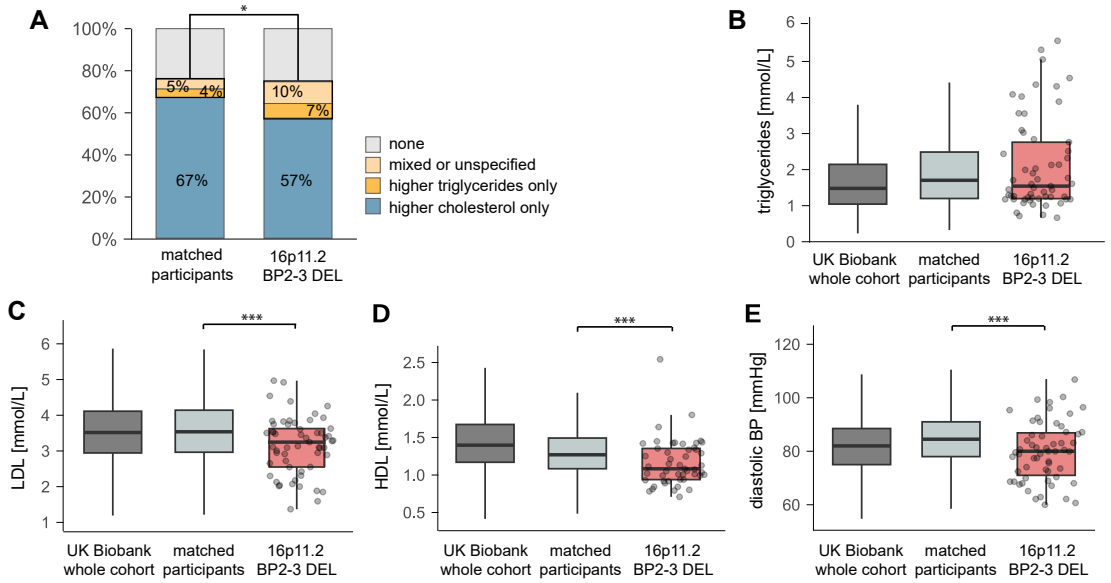


Figure 5.10: Cardiovascular risk factors in 16p11.2 BP2-3 deletion carriers compared with BMI-matched non-carriers.
(**A**) Proportion (%) of individuals with hypertriglyceridemia only (dark yellow), mixed or unspecified dyslipidemia (light yellow), hypercholesterolemia only (blue), or no dyslipidemia (gray) among deletion carriers (16p11.2 BP2-3 DEL; n = 58) and BMI-matched participants (n = 2,900). Star indicates significance for the comparison of hypertriglyceridemia and mixed/unspecified dyslipidemia between 16p11.2 BP2-3 DEL and matched participants. (**B**) Serum triglycerides (mmol/L) in UKBB whole cohort (dark gray; n = 385,495), matched participants (light gray; n = 2,695), and deletion carriers (16p11.2 BP2-3 DEL; red; n = 57). (**C**) LDL-cholesterol levels (mmol/L) in UKBB whole cohort (dark gray; n = 385,079), matched participants (light gray; n = 2,692), and deletion carriers (16p11.2 BP2-3 DEL; red; n = 57). (**D**) HDL-cholesterol (mmol/L) in UKBB whole cohort (dark gray; n = 353,195), matched participants (light gray; n = 2,496), deletion carriers (16p11.2 BP2-3 DEL; red; n = 49). (**E**) Diastolic blood pressure (BP) (mmHg) levels in UKBB whole cohort (dark gray; n = 404,478), matched participants (light gray), and deletion carriers (16p11.2 BP2-3 DEL; red; n = 58). Boxplot outliers are not shown for the whole cohort and matched participants. Data points depicted for deletion carriers (circles). $^{*}p < 0.05$; $^{***}p < 0.001$.

NMR spectroscopy revealed that serum levels of linoleic acid, sphingomyelins, phosphatidylcholines, and phosphoglycerines were significantly reduced in deletion carriers compared with the UKBB cohort (Figure 5.7C; Table S5.6) despite availability of these measurements in only 13 deletion carriers. Cross-sectional and longitudinal studies have shown that higher levels of linoleic acid are associated with decreased incidence of T2D (*511*) which aligns with deletion carriers having both lower levels of the metabolite and increased incidence of T2D. Furthermore, these results are concordant with a previous study of patients with obesity with T2D who were found to have lower levels of sphingomyelin, an abundant sphingolipid involved in ceramide metabolism, compared with people with obesity without T2D (*512*).

Although the prevalence and age of onset of hypertension were not significantly different between deletion carriers and matched non-carriers, diastolic blood pressure was lower in deletion carriers compared with the whole UKBB cohort (β = -2.8 mmHg; p = 0.033; Figure 5.7A). This trend was preserved in comparison to BMI-matched non-carriers, irrespective of the use of antihypertensive medication ($mean_{DEL}$ = 79.8 mmHg; $mean_{ctrl}$ = 84.6 mmHg; $p = 2.8 \times 10^{-4}$; Figure 5.10E). Neither the PheWAS (Figure 5.7B) nor the matched participant analysis (Table 5.2) found deletion carriers to be at increased risk for cardiovascular disease.

16p11.2 BP2-3 deletions are associated with additional non-metabolic phenotypes

ASD and developmental delay have previously been associated with 16p11.2 BP2-3 deletions (*279*). However, UKBB individuals present with a lower disease burden compared with the general UK population (*59*) and ASD prevalence in UKBB is about 0.05%, compared with a recent estimate of 1.76% across 7 million English school children (*513*). Accordingly, none of the UKBB deletion carriers were diagnosed with ASD, suggesting that carriers from the general population are at the milder end of the phenotypic range, paralleling what has been shown for other CNVs (*384*, *435*). Self-reported behaviors can indicate features that lie at the mild end of the clinical spectrum. The PheWAS indicated that deletion carriers report higher rates of loneliness (OR = 2.4; p = 0.002; Figure 5.7B), a trend maintained in the matched cohort analysis (DEL = 34%; ctrl = 21%; p = 0.036; Figure 5.11A). We found no significant differences in prevalence of anxiety, irritability, or depressive disorders in deletion carriers compared with the whole UKBB cohort and matched non-carriers (Figure 5.7B), but cognitive ability seemed to be impaired among deletion carriers, who performed worse on both fluid intelligence ($p_{PheWAS} = 8.6 \times 10^{-6}$; Figure 5.7A; $p_{matched\ control} = 8.1 \times 10^{-4}$; Figure 5.11B) and prospective memory tests ($p_{PheWAS} = 0.013$; Figure 5.7A; $p_{matched\ control} = 0.047$).

The PheWAS also revealed a nominally significant increased risk (OR = 2.3; p = 0.024) and earlier onset (HR = 2.1; p = 0.022) of anemia among 16p11.2 BP2-3 deletion carriers (Figure 5.7B). Similarly, anemia was more prevalent in deletion carriers than in matched non-carriers (Figure 5.11C). Hemoglobin, hematocrit, mean corpuscular hemoglobin and volume, and reticulocyte count were all higher in deletion carriers compared with matched non-carriers.

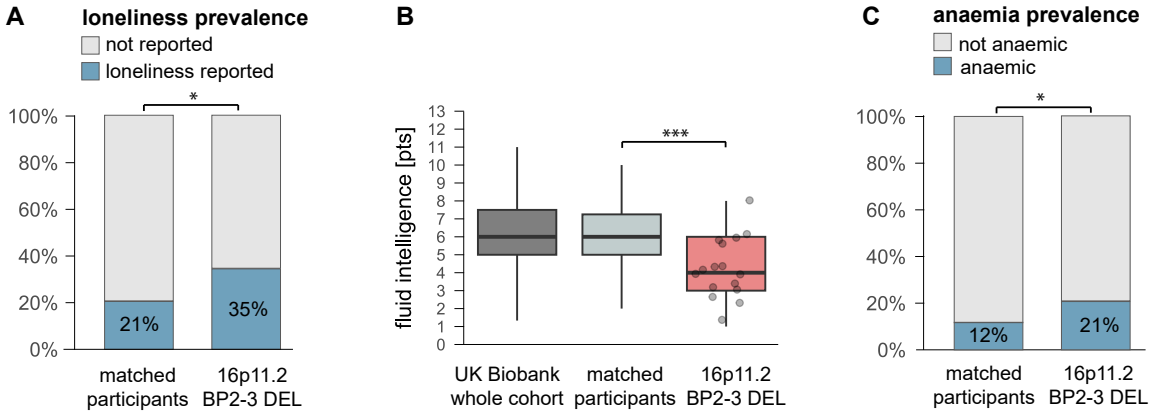


Figure 5.11: 16p11.2 BP2-3 deletion carriers have cognitive impairment.
(**A**) Prevalence of self-reported loneliness among deletion carriers (16p11.2 BP2-3 DEL) and BMI-matched participants. (**B**) Fluid intelligence score [points] on a scale from 0 to 13 in control individuals from the phenome-wide association scan (PheWAS; UK Biobank whole cohort; dark gray), matched participants (light gray), and deletion carriers (16p11.2 BP2-3 DEL; red). Data points are depicted only for deletion carriers (n = 16 with available fluid intelligence score). (**C**) Prevalence of anemia among matched participants and deletion carriers (16p11.2 BP2-3 DEL). * = p < 0.05; *** = p < 0.001.

Mechanism of action of 16p11.2 BP2-3 deletions

We investigated whether haploinsufficiency of the nine genes mapping to the 16p11.2 BP2-3 interval could corroborate the increased BMI and T2D risk observed in deletion carriers. We explored rare variant gene burden association summary statistics for BMI and T2D performed in 454,787 whole exomes of the UKBB using different masks on variant function and minor allele frequency (MAF) (*6*). Rare (MAF ≤ 0.001%) pLoF variants in *NFATC2IP* were associated with increased BMI at nominal significance (β = 0.32; p = 0.012). Interestingly, while the burden of pLoF and predicted deleterious missense variants in *ATXN2L* (OR = 0.76; p = 0.011) and *SPNS1* (OR = 0.89; p = 0.032) nominally decreased T2D risk, the singleton burden in *SH2B1* nominally increased it (OR = 2.5; p = 0.028) (Table S5.10). Similarly, we investigated whether gene burden test results supported the unusual pattern in serum lipid levels observed among deletion carriers, characterized by a reduction in both LDL and HDL levels, compared with BMI-matched controls. Concordantly, singleton LoF burden in *SH2B1* decreased both total cholesterol (β = -0.63; p = 0.002) and LDL (β = -0.58; p = 0.005) levels, and while rare variants (MAF ≤ 0.01%) in *SH2B1* also decreased HDL levels (β = -0.19; p = 0.022), more significant HDL-decreasing (*ATP2A1*, p = 0.002; *LAT*, p = 0.010) and -increasing (*RABEP2*, p = 0.013) effects were observed for other genes (Table S5.10).

Next, we assessed whether common SNPs in the 16p11.2 BP2-3 interval ± 50 kb were associated with traits affected by the deletion. We retrieved 287 association signals ($p < 9 \times 10^{-6}$) from the GWAS catalog (*78*) (Table S5.11), including signals related to adiposity (n = 95), cognitive function (n = 38), anemia (n = 17), serum lipid levels (n = 5), renal function (n = 4), diabetes (n = 3), physical activity (n = 2), and hepatic function (n = 2) (Figure 5.12A). Other signals were related to traits not assessed by our PheWAS, e.g., related to the reward system, immunity, autoimmunity, or brain morphology, and represent interesting leads for future investigation. About half of the reported signals mapped to *ATXN2L* (n = 85) and *SH2B1* (n = 66), the two genes in the region under the strongest evolutionary constraint according to the genome aggregation database (GnomAD; probability of LoF intolerance [pLi] = 1; LoF observed over expected upper bound fraction [LOEUF] < 0.23) (*29*). Focusing on the 95 adiposity-related signals, 30 and 20 were reported to map to *SH2B1* and *ATP2A1/SH2B1*, respectively. However, the low recombination rate of the region prevents accurate fine mapping of GWAS signals (Figure 5.12A).

To gain further resolution, we used transcriptome-wide Mendelian randomization (TWMR) (*173*), a causal inference approach that aims at identifying statistical causal links between changes in gene expression levels and an outcome, here T2D risk (Figure 5.12B). We could evaluate the causal impact of expression changes on T2D risk for six out of the nine 16p11.2 BP2-3 genes that had at least one eQTL variant in blood (*154*) (Figure 5.12C; Table S5.12). Among the four genes with a significant TWMR effect ($p \leq 0.05/9 = 5.6 \times 10^{-3}$), only *SH2B1* had a directionally concordant effect (α = -0.23; $p = 8.1 \times 10^{-6}$) with the one observed in our CNV association study, i.e., increased *SH2B1* expression decreased T2D risk, which is compatible with the deletion reducing the gene's expression and increasing T2D risk. While blood offers the largest eQTL datasets, this tissue is unlikely to mediate metabolic phenotypes. We repeated this analysis using smaller-sized tissue-specific eQTLs from the

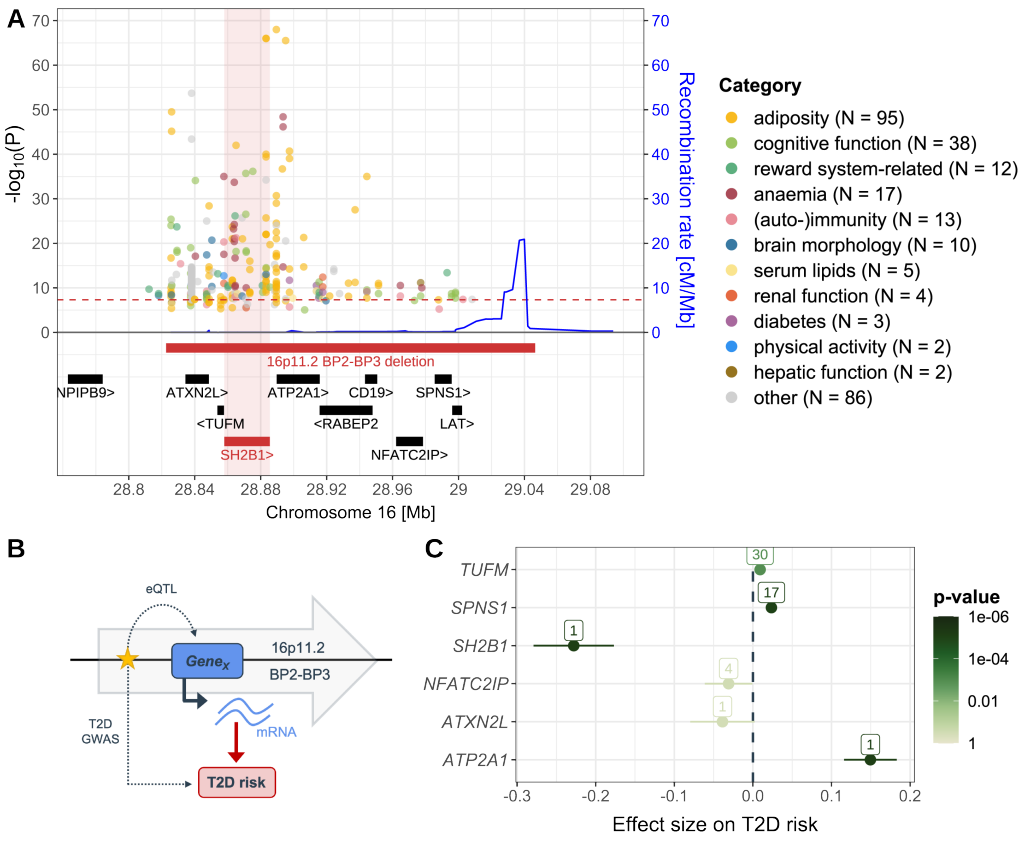


Figure 5.12: Common variant associations and transcriptome-wide Mendelian randomization effects at the 16p11.2 BP2-3 region. (**A**) Single-nucleotide polymorphism (SNP)-genome-wide association study (GWAS) signals retrieved from the GWAS Catalog for the 16p11.2 BP2-3 region ± 50 kb. The x-axis represents the genomic coordinates (GRCh37/hg19). Top: Left y-axis indicates the negative logarithm of reported association p-values, with each signal colored according to a manually assigned broader trait category. Number of signals per category is indicated (N). Right y-axis indicates the local recombination rate in cM/Mb and is represented as a blue line. The dashed horizontal red line indicates the commonly accepted threshold for GWAS genome-wide significance at $p \leq 5 \times 10^{-8}$. Bottom: Genomic location and orientation of the recurrently deleted region and *SH2B1* in red, along with other genes in the region in black. (**B**) Schematic representation of the transcriptome-wide Mendelian randomization (TWMR) approach that was applied to six 16p11.2 BP2-3 genes with at least one expression quantitative trait locus (eQTL). First, eQTLGen data (*154*) was used to identify independent *cis*-eQTLs (yellow star) for the assessed gene (blue box). Next, the effect of these variants on the expression of the gene was retrieved (dotted arrow labeled "eQTL"). Next, the effect of the same variants on type 2 diabetes (T2D) risk was assessed based on T2D GWAS summary statistics (*509*) (dotted arrow labeled "T2D GWAS"). These quantities were used to estimate the causal impact of one standard deviation increase in the expression of the assessed gene on T2D risk (red arrow) based on inverse-weighted variance two-sample Mendelian randomization. (**C**) TWMR estimates with standard error (x-axis) representing the causal effect of changes in expression of six 16p11.2 BP2-3 genes with at least one eQTL (y-axis) on T2D risk. Estimates are colored according to the p-value, with the threshold for significance at $p \leq 0.05/9 = 5.6 \times 10^{-3}$. Labels indicate the number of eQTLs used to estimate TWMR effects.

Genotype-Tissue Expression (GTEx) project (*151*) available for six out of nine genes (Table S5.13). Results were consistent across tissues, with increased expression of *ATP2A1*, *NFATC2IP*, *SPNS1*, and *TUFM* increasing T2D risk, and increased expression of *SH2B1* and *ATXN2L* decreasing risk for T2D, even if for the latter the effect was only found in whole blood. These results align with results obtained from the eQTLGen dataset and highlight *SH2B1* as the best candidate gene for the increased T2D risk observed in deletion carriers, involving brain, adipose tissue, and muscle as plausible effector tissues. One caveat is that all but one TWMR estimate for *SH2B1* relies on a single eQTL. Seeking further evidence that changes in *SH2B1* expression affect T2D, we performed colocalization analysis (*167*) between the T2D GWAS signal and expression levels of the four genes with a significant TWMR effect but could not find any evidence of a shared causal variant (posterior probability of signal colocalization [PP_H4] < 0.387) (Table S5.14).

Discussion

We show that people who are heterozygous carriers of 16p11.2 BP2-3 deletions have a higher rate of obesity, which is typically earlier in onset and associated with an accelerated form of metabolic disease characterized by early and more difficult-to-treat T2D. Experimental studies in animals will be needed to test whether disruption of *SH2B1* and/or other genes in this locus cause accelerated chronic liver disease, as suggested by our findings.

These findings have direct clinical relevance as current clinical guidelines recommend that people with severe, early-onset obesity ($\leq$ 5 years) should be offered genetic testing (*514*). While targeted gene panels or whole-exome sequencing are the most frequently offered investigations, they are often blind to chromosomal rearrangements unless the diagnosis pipeline uses depth-of-coverage maps to identify deleted exons and CNVs. The latter approach, or alternatively array CGH (comparative genomic hybridization) or MLPA (multiplex ligation-dependent probe amplification), should be considered to detect 16p11.2 BP2-3 deletions in children and young adults presenting with obesity and features of insulin resistance and/or early or difficult-to-treat T2D. Deletions involving 16p11.2 BP2-3 may be identified by a range of physicians who organize genetic testing to investigate developmental delay and ASD. Diagnosed individuals must be also reviewed by endocrinologists so that weight loss therapies, insulin sensitizers, and other glucose-lowering agents can be started at a young age to limit the impact of poor glycemic control and prevent the complications of accelerated metabolic disease.

To examine potential mechanisms underlying the observed associations, we investigated the individual contribution of the nine genes in the 16p11.2 BP2-3 interval to associated phenotypes. Among these, four genes are associated with autosomal recessive disorders: *ATP2A1* with Brody myopathy (OMIM: 601003), *TUFM* with combined oxidative phosphorylation deficiency 4 (OMIM: 610678), and both *CD19* and *LAT* with common variable immunodeficiency 3 (OMIM: 613493) and immunodeficiency 52 (OMIM: 617514), respectively. Heterozygosity of the latter was also proposed to drive increased head circumference in deletion carriers (*280*). Furthermore, experiments in mice have shown that homozygous ablation of *Atxn2l* causes lethal *in utero* brain lamination defects (*515*). The International Mouse Phenotyping Consortium found that heterozygous deletion of *Spns1* leads to increase in both total body fat and lean body mass (*516*), and a recent study demonstrated the role of the encoded protein in lysosomal lysophospholipid efflux (*517*), warranting further investigation to determine whether the gene is involved in the reduced levels of phosphatidylcholines, phosphoglycerides, and sphingomyelins observed in deletion carriers. As people carrying rare dominant mutations in *SH2B1* and *Sh2b1* knockout mice have obesity and insulin resistance (*491*–*493*, *496*), *SH2B1* appears to be the most likely candidate gene for the metabolic phenotype observed in 16p11.2 BP2-3 deletion carriers. These results are supported by our tissue-specific TWMR analysis, which suggests the importance of *SH2B1* expression in the brain, adipose tissue, and muscle in mediating T2D susceptibility. However, it remains unclear whether epistatic interactions resulting from the deletion of multiple genes could contribute to phenotypes unique to 16p11.2 BP2-3 deletion carriers.

Our clinical description of a large cohort of adult 16p11.2 BP2-3 deletion carriers indicates phenotypes that overlap with previous reports of people with SH2B1 deficiency. For instance, leptin couples changes in weight to changes in blood pressure so that mice and humans lacking leptin or its receptor have low blood pressures, despite severe obesity (*518*), in line with the reduced diastolic blood pressure seen in deletion carriers compared with BMI-matched non-carriers. Furthermore, studies in mice and humans have suggested that leptin stimulates hepatic triglyceride export via the brain-vagus nerve-liver axis (*519*), which may explain the increased levels of hepatic biomarkers and lower lipid levels seen in deletion carriers. The lack of reduction of triglyceride levels in deletion carriers may be explained by the poorer glycemic control seen in deletion carriers. In the brain, SH2B1 mediates BDNF signaling (*520*). In humans, loss of function of BDNF and its receptor TrkB, as well as SH2B1 deficiency, have been associated with speech and language delay, behavioral abnormalities, and memory impairment (*494*, *520*, *521*), features overlapping the behavioral and cognitive phenotypes seen in deletion carriers. Finally, SH2B1 acts as a negative regulator of erythropoietin receptor-mediated signaling (*522*), which may in part explain the increased blood count values seen in deletion carriers. These findings require further investigation to delineate the underlying mechanisms.

Our study has several limitations. First, population-based cohorts suffer from ascertainment bias as individuals with a high disease burden, such as 16p.11.2 BP2-3 deletion carriers, are less likely to volunteer for research studies. This decreases the case number of an already rare genetic alteration, limiting the statistical power to dissect the health consequences of the 16p11.2 BP2-3 deletion. Power is further limited as carriers present in the cohort have milder clinical phenotypes. A second limitation is the lack of advanced clinical measurements of insulin sensitivity, or the inability to recall individuals based on their genotype to perform additional investigations (e.g., hyperinsulinemic-euglycemic clamps), which would allow a more detailed understanding of the metabolic consequences of the deletion. Finally, our attempt at pinpointing individual genes responsible for the phenotypic associations is limited by several factors, including i) the lack of sufficiently variable CNV breakpoints in the region (*208*), ii) the low frequency of pLoF variants in evolutionary constrained genes in the region, iii) the low recombination rate that hinders fine-mapping of common variant association signals, and iv) the lack of sufficient eQTLs to robustly instrument TWMR analyses. The latter is particularly relevant as it makes our analysis susceptible to violation of MR assumptions. Indeed, while colocalization did not unambiguously favor any scenario, highest support was given to H3 (PP_H3: 0.60–0.76). This possibly indicates that different variants underlie the change in gene expression and T2D risk, violating the second MR assumption through linkage-disequilibrium-induced horizontal pleiotropy. However, the high probability of H3 may only reflect that there are multiple underlying signals for both traits, violating the assumption of the colocalization method, hence it is inconclusive regarding the MR assumption violation. Although there are substantial experimental data to support the role of SH2B1 in mediating the phenotypes of obesity, T2D, and fatty liver disease, further studies are needed to examine the potential phenotypic contribution of other coding genes and noncoding RNAs affected by the 16p11.2 BP2-3 deletion. In the future, availability of large, longitudinal

clinical and population cohorts with detailed phenotypic data should mitigate these hurdles.

In conclusion, 16p11.2 BP2-3 deletion carriers have a subtype of obesity that is characterized by early onset of metabolic complications including T2D. People with this disorder should be considered for early intervention with weight-loss therapies. The results of ongoing phase 3 clinical trials of Setmelanotide, an MC4R agonist in genetic obesity syndromes (ClinicalTrials.gov: NCT05093634) will provide critical information as to whether people with pathogenic mutations in *SH2B1* and with 16p11.2 BP2-3 deletions may benefit from treatment with drugs that improve signaling through the leptin-melanocortin pathway (*488*). Indeed, if the clinical trial demonstrates that 16p11.2 BP2-3 deletion carriers lose a significant amount of weight, this will provide orthogonal evidence of the contribution of SH2B1 to the obesity of deletion carriers, as people with common obesity are unlikely to respond to MC4R agonism. Collectively, these findings highlight the growing importance of mechanism-based approaches to the treatment of patients with subtypes of severe obesity.

Acknowledgments

We thank all biobank participants for sharing their data. UKBB computations were carried out on the JURA server (University of Lausanne) and on the UKBB Research Analysis Platform. EstBB computations were performed on the High-Performance Computing Center (University of Tartu). Graphical abstract was created with BioRender. This study makes use of data generated by the DECIPHER community; those who carried out the original analysis and collection of DECIPHER data bear no responsibility for the further analysis or interpretation of the data. A full list of centers that contributed to the generation of the data is available via e-mail from contact@deciphergenomics.org. The DECIPHER project was funded by Wellcome (grant no. WT223718/Z/21/Z). This study was supported by a Wellcome Principal Research Fellowship (207462/Z/17/Z), National Institute for Health and Care Research (NIHR) Cambridge Biomedical Research Centre, Botnar Fondation, Bernard Wolfe Health Neuroscience Endowment, and NIHR Senior Investigator Award (ISF); funding from the Department of Computational Biology (ZK) and the Center for Integrative Genomics (AR) from the University of Lausanne; as well as grants from the Swiss National Science Foundation (310030-189147, ZK; 31003A_182632, AR) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) (496538063, RH).

Declaration of interests

ISF has consulted for a number of companies developing weight-loss drugs, including Eli Lilly, Novo Nordisk, and Rhythm Pharmaceuticals. The remaining authors declare that they have no competing interests.

Supplemental tables

Supplemental tables are available for download as individual Excel files.

- **Table S5.1** Breakpoint and quality score of the 59 UKBB 16p11.2 BP2-3 deletion carriers.

- **Table S5.2** Prevalence of the 16p11.2 BP2-3 deletion in clinical and population cohorts.
- **Table S5.3** Summary statistics of the 16p11.2 BP2-3 deletion PheWAS, physical measurements.
- **Table S5.4** Summary statistics of the 16p11.2 BP2-3 deletion PheWAS, ICD-10 codes.
- **Table S5.5** Summary statistics of the 16p11.2 BP2-3 deletion PheWAS, other binary traits.
- **Table S5.6** Summary statistics of the 16p11.2 BP2-3 deletion PheWAS, blood measurements.
- **Table S5.7** Disease definitions for matched cohort analyses.
- **Table S5.8** Summary statistics for the matched cohort analyses.
- **Table S5.9** Statistical tests for the matched cohort analyses.
- **Table S5.10** Gene burden tests between BMI, plasma lipids, and T2D for the genes encompassed in 16p11.2 BP2-3.
- **Table S5.11** GWAS catalog SNP-GWAS lead signals for the 16p11.2 BP2-3 region.
- **Table S5.12** Transcript Mendelian randomization estimating the causal effect of changes in the expression of 16p11.2 BP2-3 genes on T2D risk in whole blood (eQTLGen).
- **Table S5.13** Tissue-specific transcript Mendelian randomization estimating the causal effect of changes in the expression of 16p11.2 BP2-3 genes on T2D (GTEx).
- **Table S5.14** Colocalization analysis between the expression of 16p11.2 BP2-3 genes and T2D GWAS signal.

16p11.2 BP4-5 | 6

E pluribus unum... or *Ex uno multis*?

6.1 Aims 173
6.2 Key Findings 173
6.3 Author Contributions 174
6.4 Disentangling mechanisms behind the pleiotropic effects of proximal 16p11.2 CNVs . . 175
6.5 The pleiotropic spectrum of proximal 16p11.2 CNVs 194

This chapter describes a research article entitled *"Disentangling mechanisms behind the pleiotropic effects of proximal 16p11.2 CNVs"*, as well as a review entitled *"The pleiotropic spectrum of proximal 16p11.2 CNVs"*. The research article is available as a preprint on medRxiv and both manuscripts have been jointly submitted to the *American Journal of Human Genetics*, where they are currently under review.

In the studies presented in Chapters 2 and 3, 16p11.2 BP4-5 CNVs were by far the most pleiotropic rearrangements, totaling over 40 distinct associations. Yet, the sheer volume of findings prevented us from duly discussing these relations, prompting us to write a review that integrates evidence from both clinical and population cohorts to describe the full phenotypic spectrum associated with 16p11.2 BP4-5 rearrangements. Writing this review, one question kept resurfacing: Is the 16p11.2 BP4-5 pleiotropy genuine or simply a reflection of secondary consequences of the CNV's impact on a few mediatory traits? Attempting to answer that latter question was the starting point for the companion article presented in this chapter.


medRχiv
THE PREPRINT SERVER FOR HEALTH SCIENCES
BMJ Yale
Follow this preprint
Disentangling mechanisms behind the pleiotropic effects of proximal 16p11.2 BP4-5 CNVs
Chiara Auwerx, Samuel Moix, Zoltán Kutalik, Alexandre Reymond
doi: https://doi.org/10.1101/2024.03.20.24304613
This article is a preprint and has not been certified by peer review [what does this mean?]. It reports new medical research that has yet to be evaluated and so should not be used to guide clinical practice.
Abstract
Info/History
Metrics
Preview PDF
Abstract
Whereas 16p11.2 BP4-5 copy-number variants (CNVs) represent one of the most pleiotropic etiologies of genomic syndromes in both clinical and population cohorts, the mechanisms leading to such pleiotropy remain understudied. Identifying 73 deletion and 89 duplication carriers among unrelated white British UK Biobank participants, we performed a phenome-wide association study between the region's copy number and 117 complex traits and diseases, mimicking four dosage models. Forty-six phenotypes (39%) were affected by 16p11.2 BP4-5 CNVs, with the deletion-only, mirror, U-shape, and duplication-only models being the best fit for thirty, ten, four, and two phenotypes, respectively, aligning with the stronger deleteriousness of the deletion. Upon individually adjusting CNV effects for either body mass index (BMI), height, cognitive function, or socio-economic status as potential mediators, we found that sixteen testable deletion-driven associations (61%) – primarily with cardiovascular and metabolic traits – were BMI-dependent, with other mediators playing a more subtle role. Bidirectional Mendelian randomization supported that 13 out of these 16 associations (81%) were secondary consequences of the CNV's impact on BMI. For the 22 traits that remained significantly associated upon individual adjustment for mediators, matched-control analyses found that eleven phenotypes, including musculoskeletal traits, liver enzymes, fluid intelligence, platelet count, pulmonary capacity, pneumonia, and acute kidney injury, remained associated under strict Bonferroni correction, with eight additional nominally significant associations. These results paint a complex picture of 16p11.2 BP4-5's pleiotropic pattern that involves direct effects on multiple physiological systems and indirect co-morbidities consequential to the CNV's impact on BMI and cognition, acting through trait-specific dosage mechanisms.

Figure 6.1: Preprint *Auwerx et al., 2024.*

Data & code availability:

→ GitHub

6.1 Aims

This work aimed to shed light on the pleiotropy of 16p11.2 BP4-5 CNVs, unravelled in the previous chaoters. To achieve this, we:

1. Perform a homogenous re-analysis of the association between 16p11.2 BP4-5 CNVs and 117 complex traits and diseases in the UKBB according to four dosage models.
2. Evaluate the role of adiposity, height, cognition, and socio-economic status as mediators of the 16p11.2 BP4-5 pleiotropy.
3. Review over 950 publications in PubMed matching the search term "16p11" and integrate literature findings to the results of our UKBB phenome-wide association scan (PheWAS) to provide a comprehensive overview of the clinical alterations observed in 16p11.2 BP4-5 CNV carriers, with a focus on physiological systems for which the role of the CNV is less appreciated.
4. Provide ways forward to better understand pleiotropy and phenotypic heterogeneity by focusing on diversity.

6.2 Key Findings

PheWAS in the UKBB found that 46 out of 117 tested phenotypes were associated with the 16p11.2 BP4-5 CNV status. This reaffirms the ex-

treme pleiotropy of these rearrangements, which extends far beyond neuropsychiatric conditions and obesity. These results are compatible with a model wherein the region's deletion is more pathogenic than the duplication, even though different dosage-to-phenotype relations apply for different traits, making this conclusion phenotype-dependent.

Assessing the role of multiple potential mediators we estimated that the CNV has a genuine, direct pleiotropic effect on numerous physiological functions, including the musculoskeletal, adipose, hepatic, hematologic, renal, and pulmonary/immune systems. What remains to be determined is whether this pleiotropy is independent at the molecular level, i.e., which specific gene(s) is/are linked to which phenotype.

Our analysis of 16p11.2 BP4-5 CNV frequency across various study types revealed the impact of ascertainment on prevalence estimates. Yet, a recurring theme is the convergence of findings from clinical and population cohorts onto similar physiological systems but with variable degrees of severity, in line with a model of variable expressivity. This stresses the importance of incorporating data from both sources to gain a deeper understanding of the true phenotypic expression of 16p11.2 BP4-5 CNVs. Population cohorts are particularly well-suited to uncover adult-onset phenotypic alterations, in often more mildly affected carriers. We found this to be especially true for metabolic phenotypes, for which detailed analysis of disease onset and evolution is often lacking in clinical studies. We show that many associations with metabolic traits represent secondary consequences of early-onset obesity, emphasizing the importance of weight management in deletion carriers to avoid adult-onset obesity-related comorbidities.

Through its comprehensiveness, our review highlights open questions and areas requiring further research. We propose several approaches to address these knowledge gaps and bring forward the idea that diversity – in terms of data source, ascertainment strategy, and analytic and experimental approaches – will be key to characterizing and understanding the pleiotropy and variable expressivity of the region. This in turn might open the door for the development of treatment and prevention strategies.

6.3 Author Contributions

The idea of writing a review came from Alexandre Reymond and myself, based on which I outlined a synopsis. I created the illustrations and wrote the first draft of the review, except for the *autism* and *experimental approaches* sections, which were written by Alexandre Reymond. Zoltán Kutalik provided critical feedback.

The research study was conceptualized by Alexandre Reymond, Zoltán Kutalik, and myself. I carried out all computations, except for the Mendelian randomization analyses that were performed by Samuel Moix, under the supervision of Zoltán Kutalik. Alexandre Reymond, Zoltán Kutalik, and I interpreted the results. I drafted the manuscript and generated all figures, with critical revisions made by Alexandre Reymond and Zoltán Kutalik.

6.4 Disentangling mechanisms behind the pleiotropic effects of proximal 16p11.2 CNVs

Chiara Auwerx [1,2,3,4,*], Samuel Moix [2,3,4], Zoltán Kutalik [2,3,4,*], Alexandre Reymond [1,*]

Abstract

Whereas 16p11.2 BP4-5 copy-number variants (CNVs) represent one of the most pleiotropic etiologies of genomic syndromes in both clinical and population cohorts, the mechanisms leading to such pleiotropy remain understudied. Identifying 73 deletion and 89 duplication carriers among unrelated white British UK Biobank participants, we performed a phenome-wide association study between the region's copy number and 117 complex traits and diseases, mimicking four dosage models. Forty-six phenotypes (39%) were affected by 16p11.2 BP4-5 CNVs, with the deletion-only, mirror, U-shape, and duplication-only models being the best fit for thirty, ten, four, and two phenotypes, respectively, aligning with the stronger deleteriousness of the deletion. Upon individually adjusting CNV effects for either body mass index (BMI), height, cognitive function, or socio-economic status as potential mediators, we found that sixteen testable deletion-driven associations (61%) – primarily with cardiovascular and metabolic traits – were BMI-dependent, with other mediators playing a more subtle role. Bidirectional Mendelian randomization supported that 13 out of these 16 associations (81%) were secondary consequences of the CNV's impact on BMI. For the 22 traits that remained significantly associated upon individual adjustment for mediators, matched-control analyses found that eleven phenotypes, including musculoskeletal traits, liver enzymes, fluid intelligence, platelet count, pulmonary capacity, pneumonia, and acute kidney injury, remained associated under strict Bonferroni correction, with eight additional nominally significant associations. These results paint a complex picture of 16p11.2 BP4-5's pleiotropic pattern that involves direct effects on multiple physiological systems and indirect comorbidities consequential to the CNV's impact on BMI and cognition, acting through trait-specific dosage mechanisms.

[1] Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland; [2] Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland; [3] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; [4] University Center for Primary Care and Public Health, 1010 Lausanne, Switzerland; [*] **Correspondence**.

Introduction

Genomic disorders are caused by recurrent genomic rearrangements that lead to the gain (duplication) or loss (deletion) of large, multi-kilobase pair (kb) DNA fragments. The proximal 16p11.2 rearrangement spans a region of ~600 kb between recurrent breakpoints (BP) 4 and 5 and includes 27 unique protein-coding genes. Copy-number variants (CNVs) of the region represent one of the most common genomic disorders, with population prevalence estimates of 1 in 3,000 and 1 in 2,800 for the deletion (MIM: 611913) and duplication (MIM: 614671), respectively[1]. Prevalence in clinical cohorts is about eight-fold higher, with a particularly strong enrichment in individuals ascertained for intellectual disability and developmental delay (*335*, *523*, *524*) or autism spectrum disorder (*386*, *422*, *525*, *526*), the first phenotypes associated with the CNV. Other hallmark features include a negative dosage effect on body mass index (BMI) (*325*, *326*, *426*) and head circumference (*421*, *425*), a predisposition for seizure disorders

1: These estimates are derived from the review presented later in this Chapter, in section 6.5.

(389, 421, 523, 524), and a duplication-specific increased susceptibility to schizophrenia and other psychiatric conditions (387, 423, 425, 500, 527–529). The recent establishment of large biobanks coupling genetic information to phenotypic data such as physical measurements, blood biomarkers, and electronic health records, has allowed to study the phenotypic expression of 16p11.2 BP4-5 rearrangements in individuals that are typically older and less severely affected than those recruited in pediatric clinical cohorts (82, 208, 229, 292–295, 305, 306, 402, 530). Results of these studies often converge onto similar pathophysiological processes than those highlighted by clinical studies but also report associations with biomarkers and common diseases that are typically overlooked or not assessed in clinical cohorts.

If the pleiotropic nature of 16p11.2 BP4-5 rearrangements is now well-established, the mechanisms through which CNVs in the region affect such diversity of traits remain poorly studied. Under a model of direct (or horizontal) pleiotropy, the CNV causally impacts associated phenotypes through independent mechanisms (Figure 6.2A). Conversely, indirect (or vertical) pleiotropy implies that the CNV causally impacts a mediatory trait, which in turn causally impacts other traits that will appear as linked with the CNV in association studies (Figure 6.2B). These models are not mutually exclusive, and a fraction of the associations might result from direct effects while others might be secondary consequences. This question is particularly relevant given the BMI-modulating role of 16p11.2 BP4-5 CNVs (208, 292, 294, 295, 325, 326, 426). Indeed, BMI represents a strong risk factor for other diseases and knowledge about which associations are consequential to altered BMI could therefore inform epidemiology of associated comorbidities and clinical practice. To address this knowledge gap, we re-analyzed two recent UK Biobank (UKBB) studies that assessed the impact of 16p11.2 BP4-5 rearrangements on 117 complex traits and common diseases (82, 208) with the aims to i) determine the most likely dosage mechanism for different traits and ii) estimate the fraction and nature of associations that are mediated by primary changes in anthropometric measurements, cognitive ability, and socio-economic status (Figure 6.2C).

Materials and methods

Study material

Cohort description & sample selection

Analyses were carried out in the UKBB, a volunteer-based UK population cohort of about half a million individuals (54% females) aged 40-69 years at recruitment, who signed a broad informed consent form (61). Available data include microarray genotype data acquired in GRCh37/hg19 from two similar arrays, as well as rich phenotypic data, including anthropometric measurements, vital signs, blood biomarker levels, life history and lifestyle questionnaire data, hospital-based International Classification of Diseases, 10th Revision (ICD-10) codes (up to September 2021), and self-reported conditions. Analyses conducted in this study focus on 331,522 unrelated individuals from the "white British" UKBB subset (54% females) that were filtered to exclude samples with abnormal CNV profiles and/or a report of blood malignancy. Filtering criteria to obtain this set are described elsewhere (82).

Software versions:

- MR pipeline: R v4.2.1.
- TwoSampleMR v0.5.7 (531).
- PLINK v1.9 (88).
- Statistical analyses: R v4.3.1.
- Graphs: R v4.3.1.

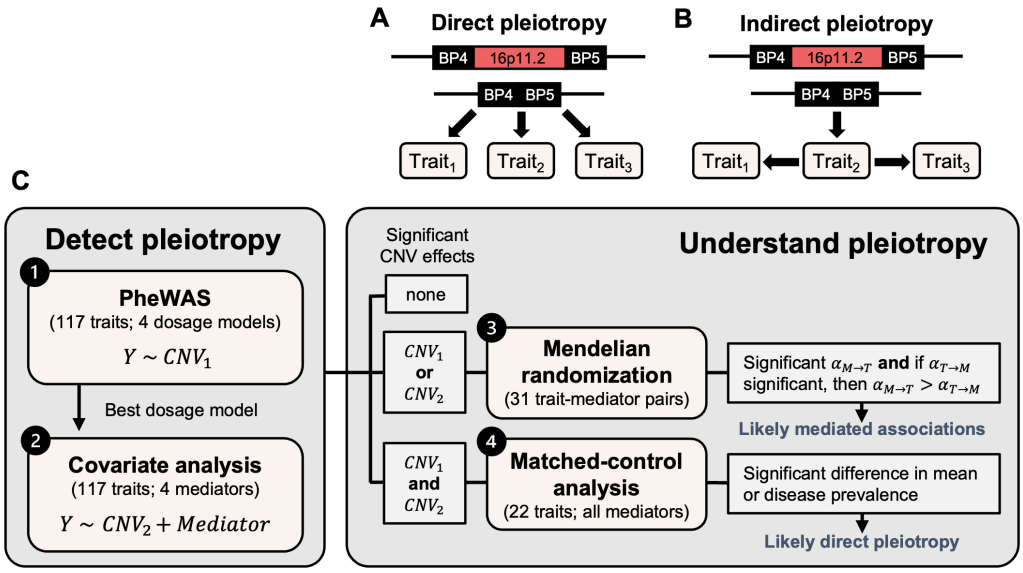


Figure 6.2: Study workflow.
(**A**-**B**) Schematic representation of pleiotropy mechanisms. For illustration, the 16p11.2 BP4-5 deletion is depicted but the same concept applies to the duplication. (A) Direct (horizontal) pleiotropy: The CNV causally affects multiple traits – here Trait_1, Trait_2, and Trait_3 – through independent mechanisms. (B) Indirect (vertical) pleiotropy: The CNV causally impacts Trait_2, which in turn causally affects Trait_1 and Trait_3. The impact of the deletion on Trait_1 and Trait_3 is thus indirect and mediated by a shared mechanism, i.e., Trait_2. (**C**) Overview of the study. The first two analyses aim at detecting and characterizing the pleiotropy of 16p11.2 BP4-5 CNVs through four distinct dosage models that estimate the effect of the CNV on the trait (Y) either (**1**) without or (**2**) with adjustment for one of four covariates that could potentially mediate the CNV-phenotype association. The second part of the study aims at understanding the mechanisms through which pleiotropy arises. (**3**) Bidirectional Mendelian randomization was used to investigate the causal relationship between trait-mediator pairs for which the significance of the CNV effect on the trait was affected by adjustment for the mediator. Support for mediation was claimed when the forward MR effect of the mediator on the trait ($\alpha_{M\to T}$) is significant and larger than the reverse effect of the trait on the mediator ($\alpha_{T\to M}$), providing the latter is significant. (**4**) For traits that showed a significant association with the CNV regardless of covariate adjustment, we performed a matched-control analysis that allowed us to adjust for all possible mediators simultaneously and detect genuine, direct pleiotropic associations. PheWAS = phenome-wide association study.

CNV carrier identification

CNV calls from a previous study were used (*208*). Briefly, CNV calling was done based on the UKBB microarray data using standard PennCNV v1.0.5 settings (*203*). Each call was attributed a quality score ranging from −1 (likely deletion) to 1 (likely duplication) reflecting the probability for the CNV to be a consensus call across three algorithms and thus a true positive (*206*). 16p11.2 BP4-5 deletion and duplication carriers were identified as carrying a high-confidence CNV call (quality score < -0.5 for deletions; quality score > 0.5 for duplications) on chromosome 16 with start and end site within 29.4-29.8 Mb and 30.05-30.4 Mb, respectively. Individuals with a low-quality 16p11.2 BP4-5 CNV were excluded from copy-neutral controls. CNV genotype vectors were then encoded to allow the fitting of regression models according to four dosage mechanisms (*82*).

Phenotype selection

We analyzed the same 117 phenotypes as defined in previous studies (*82*, *208*). This includes 57 quantitative traits that were inverse normal transformed before being corrected for sex (except for sex-specific traits), age (UKBB field identifier #21003), age^2, genotyping batch, and principal components 1-40 (*208*). We further include 60 common diseases based on ICD-10 clinical diagnoses using a case-control definition procedure that excludes from controls individuals with a condition related to the one under investigation (*82*).

Mediator selection
We tested the role of four factors that could potentially mediate associations between 16p11.2 BP4-5 CNVs and the assessed phenotypes:

1. **Body mass index (BMI):** average over available instances of BMI (#21001).
2. **Educational attainment (EA):** age at which full-time education was completed (#845). Values matching "prefer not to answer", "never went to school", and "do not know" were set as missing, and average over available instances was calculated. Individuals for which average age at which full-time education was completed was below 14 years or over 19 years were set to 14 years and 19 years, respectively. Individuals reporting a "college or university degree" in their qualifications (#6138) were set to 19 years.
3. **Townsend deprivation index** at recruitment (TDI; #22189).
4. **Height:** average over available instances of standing height (#50).

GWAS summary statistics
Mendelian randomization (MR) studies rely on publicly available genome-wide association studies (GWAS) summary statistics for both sexes and individuals of European ancestry. For mediators, summary statistics from Pan-UK Biobank (manifest updated 01/03/2023) were used for BMI, TDI, and height (*532*). For EA, summary statistics from a large meta-analysis by Okbay *et al.* were used (excluding 23andMe data) (*533*). For other phenotypes, summary statistics from the Neale group (released 07/2018) were used. These summary statistics were favored over those of large disease-specific consortia as summary statistics for binary traits were calculated through linear regression, allowing comparison of forward and reverse effects. For diseases, we used the closest possible match to our phenotype definition, i.e., phenotype code: E10 for "T1D" (type 1 diabetes); G47 for "sleep" (sleep apnea); I10 for "HTN_essential" (essential hypertension); I35 for "valves" (cardiac valve disorders); I44 for "conduction" (cardiac conduction disorders); J45 for "asthma"; M19 for "OA" (arthrosis); N18 for "CKD" (chronic kidney disease); 20002_1473 for "lipid" (lipidemias & lipoprotein disorders). Summary statistics for autosomal chromosomes were harmonized to the UK10K reference panel (*534*). After excluding palindromic single-nucleotide polymorphisms (SNPs) and adjusting strand-flipped SNPs, effect sizes were standardized to represent the square root of the explained variance.

16p11.2 BP4-5 association studies

Phenome-wide association study
For the phenome-wide association study (PheWAS), regression analysis was performed to estimate the effect of the CNV genotype – encoded according to either of the four models – and the 117 selected phenotypes. For quantitative traits, linear regressions (`lm()` in R) were used and 95% confidence intervals (CI) were calculated as beta ± 1.96 · standard error (SE). For binary traits, Firth's bias-reduced penalized-likelihood logistic regression was used (`logistf(plconf = 2, maxit = 100, maxstep = 10)` from the `logistf` package v1.26.0 in R) to account for the fact that both CNV carriers and disease cases are rare. The same function also produces estimates for the 95% CIs. As disease diagnoses were defined as binary variables and could not be adjusted beforehand, sex (except for sex-specific traits), age, genotyping array, and principal components 1-40

were included as covariates. For each trait, the dosage model yielding the lowest p-value for the CNV effect was retained and effects were defined as strictly significant under Bonferroni correction criteria ($p \leq 0.05/117 = 4.3 \times 10^{-4}$).

Covariate analysis

For all phenotype-mediator pairs, including those involving phenotypes that did not significantly associate with the CNV status in our original PheWAS, we estimated the Pearson correlation (`cor(use = "pairwise.complete.obs")` in R), as well as the effect of the mediator on the phenotype in a linear/Firth regression model without covariates, as previously described. For pairs with Pearson correlation < 0.5 and effect of the mediator on the trait $p \leq 0.05/117 = 4.3 \times 10^{-4}$, we estimated the effect of the CNV carrier status encoded according to the best PheWAS model. Regressions were implemented as previously described, adding the mediator as an additional covariate. Adjusted effects were defined as strictly significant when meeting Bonferroni correction criteria ($p \leq 0.05/117 = 4.3 \times 10^{-4}$). We additionally compared effect estimates with ($\beta_{adjusted}$) and without (β) mediator adjustment based on a t-statistic[2]. Two-sided p-values were calculated (`2*pnorm(-abs(t), mean = 0, sd = 1)` in R). The difference in correlation between BMI-dependent and BMI-independent traits with BMI was assessed with a two-sided t-test.

2: t-statistic: $t = \frac{\beta - \beta_{adjusted}}{\sqrt{SE^2 + SE^2_{adjusted}}}$ where SEs are the standard errors of the effects

Mendelian randomization

GWAS summary statistics were used to conduct bidirectional MR according to a previously published pipeline (*175*, *535*) for 31 mediator-trait pairs for which the CNV-trait association either gained or lost significance upon adjusting for that mediator. Concretely, the forward effect of the mediator (exposure) on the trait (outcome) and the reverse effect of the trait (exposure) on the mediator (outcome) were estimated. Harmonized SNPs significantly ($p < 5 \times 10^{-8}$) associating with the exposure were clumped with PLINK v1.9 (`p1 = 0.0001, p2 = 0.01, kb = 250, and r2 = 0.01`) and retained as instrumental variables. Instrumental variables mapping to the extended HLA region (chr6:25-37 Mb; GRCh37/hg19) were excluded, as well as those with a difference in allele frequency (≥ 0.05) between the outcome and exposure summary statistic. Steiger filtering was applied ($Z \leq -1.96$) to ensure that the effect of the selected variants on the exposure was stronger than their effect on the outcome. Bidirectional inverse variance weighted MR analyses were carried out with the `TwoSampleMR` R package when at least two instrumental variables were available. MR effects were called significant under Bonferroni correction criteria, when $p \leq 0.05/62 = 8.1 \times 10^{-4}$, to account for the 31 bidirectional tests performed.

Matched-control analysis

For each CNV carrier, we identified all copy-neutral unrelated individuals from the "white British" subset of UKBB participants that were matching based on sex (identical), age (± 2.5 years), BMI (± 2.5 kg/m^2), TDI (± 2), average household income before tax (#738) averaged over instances (identical category), and EA (± 1 year). Fifty-eight deletion and sixty-one duplication carriers had no missing data and qualified for the matching procedure. The number of identified matching controls per carrier ranged from 1 to 918 and 12 to 1,590 for deletion and duplication carriers, respectively, with 49 deletion and 60 duplication carriers having at least 25 matching controls. When more than 25 matched controls were

available, the ones used for the analysis were selected randomly (`sample_n()` in R), without replacement. For quantitative traits, we compared mean phenotypic values between deletion and duplication carriers and the respective control groups through a two-sided t-test. For binary traits, disease prevalence was compared between the same groups based on a two-sided Fisher test[3]. Sample sizes vary between phenotypes due to missing data. We define significant associations based on a Bonferroni correction that accounts for the 22 traits of interest in this analysis ($p \leq 0.05/22 = 2.3 \times 10^{-3}$), i.e., phenotypes that remained associated with the CNV under strict Bonferroni correction when adjusting for BMI, height, EA, or TDI individually. We report all nominally significant ($p < 0.05$) associations on figures.

In a related analysis aiming at assessing the consequences of losing samples for the matched-control analysis, we used the same statistical framework to compare mean phenotypic value and disease prevalence between deletion and duplication carriers that were included in the matched-control analysis versus those that were not due to missing data or lack of sufficient controls.

3: Prevalence standard error:

$SE = \sqrt{\frac{q \cdot (1-q)}{n}}$

where q is the disease prevalence, and n the sample size.

Results

16p11.2 BP4-5 phenome-wide association study

Using previously published high confidence CNV calls for 331,522 unrelated, white British UKBB participants (*82*, *208*), we identified 73 and 89 individuals with a 16p11.2 BP4-5 (start: chr16:29.40-29.80 Mb; end: chr16:30.05-30.40 Mb) deletion and duplication, respectively. CNV genotypes were encoded to allow testing of four dosage mechanisms, namely a mirror model assessing the additive impact of each additional copy, a U-shape model testing the same-direction impact of any deviation from the copy-neutral state, and duplication- and deletion-only models that assess the separate impact of duplications and deletions, respectively. Next, we evaluated the association between an individual's CNV carrier status and 117 phenotypes – that comprise 57 quantitative variables including anthropometric measurements, vital signs, biomarker levels, life history events, and 60 common diseases – while correcting for sex, age, age^2, genotyping array, and population stratification (Figure 6.3; Table S6.1).

Overall, 46 (39%) traits, including 16 diseases, were associated with the CNV carrier status under at least one association model (Bonferroni correction: $p \leq 0.05/117 = 4.3 \times 10^{-4}$; Table 6.1), with an additional 32 (27%) showing a trend for association (nominal significance: $p \leq 0.05$). Specifically, 10 and 38 traits showed a significant association through the duplication-only and deletion-only models, respectively, indicating a stronger propensity for pleiotropy and deleteriousness of the deletion, compared to the duplication. Exceptions are recurrent depression and bipolar disorder, the two only traits for which the duplication-only model yielded the most significant result. This is in line with the duplication representing a strong susceptibility factor for psychiatric conditions (*387*, *423*, *425*, *528*, *529*). Similarly, the risk for schizophrenia was strongly increased by the duplication, even if our analysis finds that the relation is better described by a U-shape model wherein the deletion also tends to increase schizophrenia risk. Surprisingly, the CNV did not associate with neuroticism score, despite the high genetic correlation between

neuroticism and psychiatric conditions (*536*). Three other traits, namely fluid intelligence, vitamin D levels, and waist-to-hip ratio adjusted for BMI (WHRadjBMI), were also most significantly associated through a U-shape effect, while grip strength was decreased in both deletion and duplication carriers, but more strongly so in the former. Conversely, ten traits were most significantly associated through a mirror model, including multiple hepatic biomarkers, platelet count, and traits related to sexual characteristics such as puberty timing and sex hormone binding globulin (SHBG) levels. Finally, the deletion-only model was the most significant fit for 30 phenotypes, including mostly pulmonary, cardiovascular, metabolic, and renal traits.

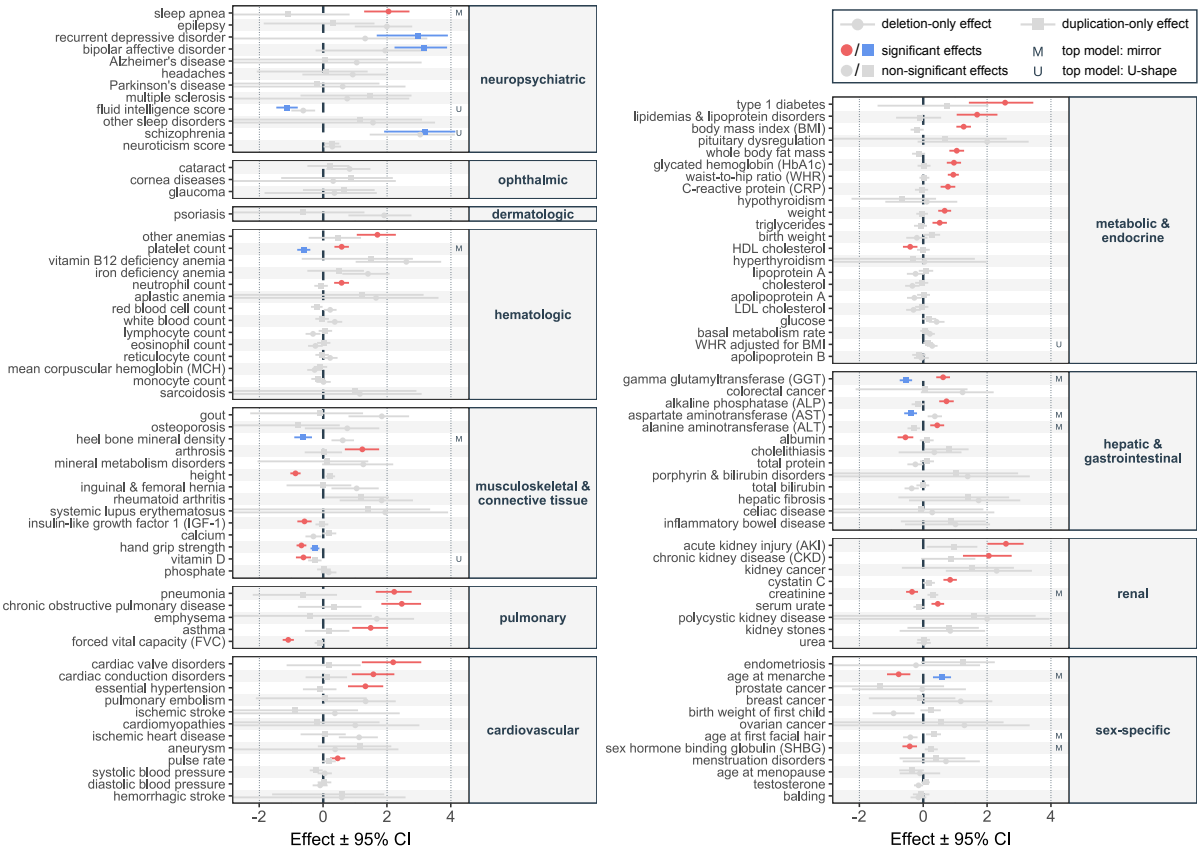


Figure 6.3: 16p11.2 BP4-5 phenome-wide association study.
Effect sizes (beta; x-axis) with 95% confidence interval (CI) of the 16p11.2 BP4-5 deletion (circle) and duplication (square) on 117 complex traits and diseases, ordered by physiological system (y-axis). Effect sizes are in standard deviation units of the outcome (quantitative traits) or logarithms of the odds ratio of a logistic regression (disease traits). Deletion- and duplication-only effects that are significant under Bonferroni correction ($p \leq 0.05/117 = 4.3 \times 10^{-4}$) are in blue and red, respectively, while non-significant effects are in gray. If the most significant among the four tested association models was the mirror or U-shape model, it is denoted with an "M" or "U", respectively (right).

Covariate analysis

Having characterized the pleiotropic nature of 16p11.2 BP4-5 rearrangements, we next sought to establish whether some of these associations might be secondary to the CNV affecting core mediatory phenotypes, i.e., reflect indirect pleiotropy (Figure 6.2). We focus on four traits that proxy hallmark features of the 16p11.2 BP4-5 rearrangement and have the potential to influence other associated traits: i) BMI, which characterizes the negative correlation between dosage and adiposity (*208*, *295*, *325*, *326*, *426*, *500*) and represents a major risk factor for many common diseases;

Table 6.1: Traits significantly associated with 16p11.2 BP4-5 CNVs.
Traits that are significantly ($p \leq 0.05/117 = 4.3 \times 10^{-4}$) associated with 16p11.2 BP4-5 CNVs through at least one of the four assessed association models, following the ordering of Figure 6.3. (**A**) For quantitative traits, the mean value of the traits in copy-neutral individuals (controls) is provided along with the mean value and standard error (SE) among duplication and deletion carriers. The number of duplication and deletion carriers with available data is specified as N. Values are given in the indicated unit. (**B**) For binary disease traits, prevalence in percentage among copy-neutral individuals is provided along with prevalence and SE among duplication and deletion carriers. Diseased (case) and total (N) number of duplication and deletion carriers are indicated.

CNV status		**Controls**	**Deletion carriers**		**Duplication carriers**	
A. Quantitative traits						
	unit	**mean**	**N**	**mean ± SE**	**N**	**mean ± SE**
Fluid intelligence score	points	6.24	26	5.15 ± 0.39	32	4.05 ± 0.37
Platelet count	109 cells/L	252.8	70	286.4 ± 7.5	85	219.4 ± 5.7
Neutrophil count	109 cells/L	4.25	70	5.00 ± 0.15	85	4.14 ± 0.15
Heel bone mineral density	g/cm2	0.54	29	0.66 ± 0.04	47	0.46 ± 0.01
Height	cm	168.8	73	163.3 ± 1.0	89	171.1 ± 1.1
Insulin-like growth factor 1 (IGF-1)	nmol/L	21.4	71	18.9 ± 0.7	88	21.4 ± 0.6
Hand grip strength	kg	30.7	73	26.7 ± 1.0	89	28.6 ± 1.2
Vitamin D	nmol/L	49.8	68	36.0 ± 2.1	86	44.7 ± 2.2
Forced vital capacity (FVC)	L	3.63	64	2.94 ± 0.11	72	3.64 ± 0.12
Pulse rate	bpm	69.3	67	74.3 ± 1.5	81	71.3 ± 1.3
Body mass index (BMI)	kg/m2	27.4	72	35.0 ± 0.9	89	26.3 ± 0.4
Whole body fat mass	kg	24.9	67	35.5 ± 1.6	85	23.2 ± 0.9
Glycated hemoglobin (HbA1c)	mmol/mol	36.0	70	43.3 ± 1.8	85	35.7 ± 0.5
Waist-to-hip ratio (WHR)	-	0.87	72	0.97 ± 0.01	89	0.87 ± 0.01
C-reactive protein (CRP)	m/L	2.57	71	5.50 ± 0.76	88	2.03 ± 0.23
Weight	kg	78.3	72	93.3 ± 2.7	89	77.1 ± 1.4
Triglycerides	mmol/L	1.75	71	2.30 ± 0.14	87	1.60 ± 0.09
HDL cholesterol	mmol/L	1.46	61	1.26 ± 0.04	79	1.45 ± 0.04
WHR adjusted for BMI	-	0.00	71	0.03 ± 0.01	89	0.01 ± 0.01
Gamma-glutamyltransferase (GGT)	U/L	37.3	71	64.1 ± 11.3	87	27.7 ± 2.8
Alkaline phosphatase (ALP)	U/L	83.6	71	99.8 ± 3.4	88	79.8 ± 2.6
Aspartate aminotransferase (AST)	U/L	26.2	70	35.8 ± 6.7	87	23.4 ± 0.7
Alanine aminotransferase (ALT)	U/L	23.5	71	31.2 ± 2.1	88	20.8 ± 1.3
Albumin	g/L	45.3	61	44.0 ± 0.4	79	45.6 ± 0.3
Cystatin C	mg/L	0.91	71	1.04 ± 0.02	88	0.93 ± 0.02
Creatinine	mmol/L	72.4	71	68.8 ± 1.6	87	76.9 ± 2.0
Urate	mmol/L	309.4	70	355.2 ± 8.1	88	298.8 ± 9.1
Age at menarche	years	12.9	27	11.7 ± 0.3	45	14.0 ± 0.4
Relative age at first facial hair (group 1-3)	-	2.06	41	1.83 ± 0.06	38	2.26 ± 0.08
Sex hormone binding globulin (SHBG)	nmol/L	51.9	61	38.4 ± 2.4	77	58.8 ± 3.4
B. Disease Prevalence						
		prevalence	**case/N**	**prevalence ± SE**	**case/N**	**prevalence ± SE**
Sleep apnea		2.1%	9/60	15.0 ± 4.6%	0/70	0.0 ± 0.0%
Recurrent depressive disorder		0.3%	0/49	0.0 ± 0.0%	3/68	4.4 ± 2.5%
Bipolar affective disorder		0.4%	1/50	2.0 ± 2.0%	6/71	8.5 ± 3.3%
Schizophrenia		0.2%	2/51	3.9 ± 2.7%	3/68	4.4 ± 2.5%
Other anemias		5.5%	13/67	19.4 ± 4.8%	6/84	7.1 ± 2.8%
Arthrosis (OA)		21.1%	23/62	37.1 ± 6.1%	15/74	20.3 ± 4.7%
Pneumonia		5.9%	19/61	31.1 ± 5.9%	2/83	2.4 ± 1.7%
Chronic obstructive pulmonary disease (COPD)		4.9%	17/55	30.9 ± 6.2%	4/76	5.3 ± 2.6%
Asthma		12.1%	19/55	34.5 ± 6.4%	9/69	13.0 ± 4.1%
Cardiac valve disorders		5.0%	7/33	21.2 ± 7.1%	3/54	5.6 ± 3.1%
Cardiac conduction disorders		19.5%	18/44	40.9 ± 7.4%	13/64	20.3 ± 5%
Essential hypertension (HTN)		35.3%	36/62	58.1 ± 6.3%	23/74	31.1 ± 5.4%
Type 1 diabetes (T1D)		1.0%	4/42	9.5 ± 4.5%	1/75	1.3 ± 1.3%
Lipidemias & lipoprotein disorders		22.0%	23/48	47.9 ± 7.2%	11/57	19.3 ± 5.2%
Acute kidney injury (AKI)		4.7%	20/65	30.8 ± 5.7%	7/71	9.9 ± 3.5%
Chronic kidney disease (CKD)		4.4%	9/54	16.7 ± 5.1%	6/70	8.6 ± 3.3%

ii) Height, which is reduced in deletion carriers (*208*, *295*, *500*) and can influence musculoskeletal phenotypes; iii) Educational attainment (EA) proxied by age at which an individual completed their education. This variable offers the advantage of being available for the near totality of the UKBB cohort while strongly correlating with fluid intelligence score that is limited to about half of its participants (Pearson correlation = 0.42), thereby reflecting the decreased cognitive function observed in both duplication and deletion carriers (*208*, *423*, *500*, *523*, *524*); and iv) Townsend deprivation index (TDI) as a measure of SES, which we expect to be reduced as a corollary of the health burden imposed by the CNV (*292*). Of note, while TDI specifically aims at assessing SES, BMI, height, and EA also partly capture socio-economic status (*537*). For the association between CNV and phenotype to be mediated by one of these factors, the mediator needs to significantly ($p \leq 0.05/117 = 4.3 \times 10^{-4}$) associate with the tested phenotype. Furthermore, phenotypes cannot be too correlated with the mediator (Pearson's correlation > 0.5), as in such situations distinguishing mediator and outcome would be particularly difficult. For all mediator-trait pairs that fulfilled these criteria, we tested the impact of adjusting the CNV-trait effect for mediatory factors by including them individually in the regression model yielding the most significant CNV-trait effect (Figure 6.4A; Table S6.2).

Upon adjustment for BMI, TDI, EA, and height, nineteen, four, four, and zero CNV-trait associations fell below the significance cutoff ($p \leq 0.05/117 = 4.3 \times 10^{-4}$), respectively. Comparing effect sizes, only the mirror association with sleep apnea was nominally significantly reduced upon adjustment for BMI ($p = 0.04$). Remarkably, the association with basal metabolic rate (deletion-only) became significant upon adjustment for height, while the one with diastolic blood pressure (mirror), eosinophil count (deletion-only), and lymphocyte count (deletion-only) became so upon adjustment for BMI (Figure 6.4B), even though the change in effect size were not significant ($p > 0.45$). The impact of adjusting for BMI was most striking on deletion-driven associations, for which 61% (16/26) of the associations fell below the significance threshold (Figure 6.4C). In line with expectations, BMI-dependent traits tended to have a stronger correlation with BMI than those that remained significant upon adjustment for BMI ($p = 0.05$) (Figure 6.4D). Among the lost associations, we find nine out of the ten metabolic and cardiovascular traits associated with the deletion. These associations likely reflect secondary consequences of the propensity for obesity of deletion carriers as they include levels of serum lipid and the inflammation biomarker C-reactive protein (CRP), cardiac valve and conduction disorders, and hypertension. The effect of BMI on musculoskeletal, pulmonary, or renal traits is more balanced, with some associations, such as the ones with arthritis (OA), asthma, or urate and chronic kidney disease (CKD), appearing to be driven by BMI, while others, such as grip strength, chronic obstructive pulmonary disease (COPD), or cystatin C and acute kidney injury (AKI), remaining significant upon BMI adjustment. The mediating role of TDI and EA was much milder, as only four associations were lost upon adjustment for either variable – including a shared association with WHR adjusted for BMI, heart rate, and high-density lipoprotein (HDL) cholesterol – suggesting that TDI and EA capture partially overlapping mediatory processes. Surprisingly, associations with psychiatric disorders were not affected by EA, suggesting that cognition and psychiatric diseases

are regulated by (at least partially) independent pathways. Finally, the observation that no associations were affected by adjusting for height confirms that the decrease in traits such as grip strength and forced vital capacity among deletion carriers is not driven solely by their short stature.

Mendelian randomization

One caveat of our analysis is that it cannot distinguish whether changes in CNV-trait associations are indeed secondary effects of the mediator on the trait. At least three scenarios could result in the loss (or gain) of a CNV-trait effect upon covariate adjustment (Figure 6.5A). The first one is mediation, wherein the CNV affects the trait through the mediator, resulting in a dominant causal effect of the mediator on the trait. The second scenario is when the variable we adjusted for turns out to be a collider of the CNV and the trait, in which case we expect a dominant causal effect from the trait to the "mediator". Finally, data could be explained by an unobserved confounder that affects both the adjustment variable and the trait, in which case we do not expect any causal link between trait and mediator. Of note, in the latter scenario, we further distinguish between whether the CNV has an impact on the confounder, the "mediator", the trait, or a combination thereof. Importantly, adjusting for the mediator in the regression model is an appropriate solution to obtain meaningful direct CNV-trait effects (i.e., genuine direct pleiotropy) only in the i) mediator scenario or ii) the confounder scenario where the CNV has a direct effect on the trait, in which case adjustment for the mediator could result in a gain of power (Figure 6.5A). To identify cases where mediation is a likely scenario, we resorted to bidirectional Mendelian randomization (MR), a causal inference approach that allows to estimate the genetically determined causal effect of an exposure on an outcome (Figure 6.5B; Table S6.3). Firstly, we estimated the forward mediator-to-trait effect for all 31 mediator-trait pairs that either gained (N = 4; Figure 6.4B) or lost (N = 27; Figure 6.4A) significance upon adjustment for the mediator. Except for the four TDI-dependent associations which had large confidence intervals due to the lack of good genetic instruments for TDI and the effect of BMI on hypertension, type 1 diabetes, and cardiac conduction disorders, all effects were significant ($p \leq 0.05/62 = 8.1 \times 10^{-4}$), confirming that the mediators can causally influence the involved traits. Secondly, we estimated the reverse trait-to-mediator causal effects. Ten reverse effects were significant and thus represent mediator-trait pairs at risk for collider bias. Yet, for nine of them, the forward effect had a larger magnitude, making the mediator-to-trait causal path more likely. The only exception is the association between the deletion and basal metabolic rate that became significant upon adjustment for height and for which the reverse effect was stronger than the forward effect. This suggests that height could act as a collider and adjustment for it could bias estimates. Hence, we conclude based on the unadjusted effect that the association between the deletion and basal metabolic rate is non-significant. It is also worth noting that six out of seven associations lacking a significant forward effect also lacked a significant reverse effect, possibly indicating presence of an unobserved confounder. This is particularly likely for the BMI effect on hypertension, type 1 diabetes, and cardiac conduction disorders, where estimates are close to null despite being well-instrumented (≥ 50 instruments). Globally, these analyses support that a large fraction (74%) of the flagged associations are likely indirect consequences of the CNV's

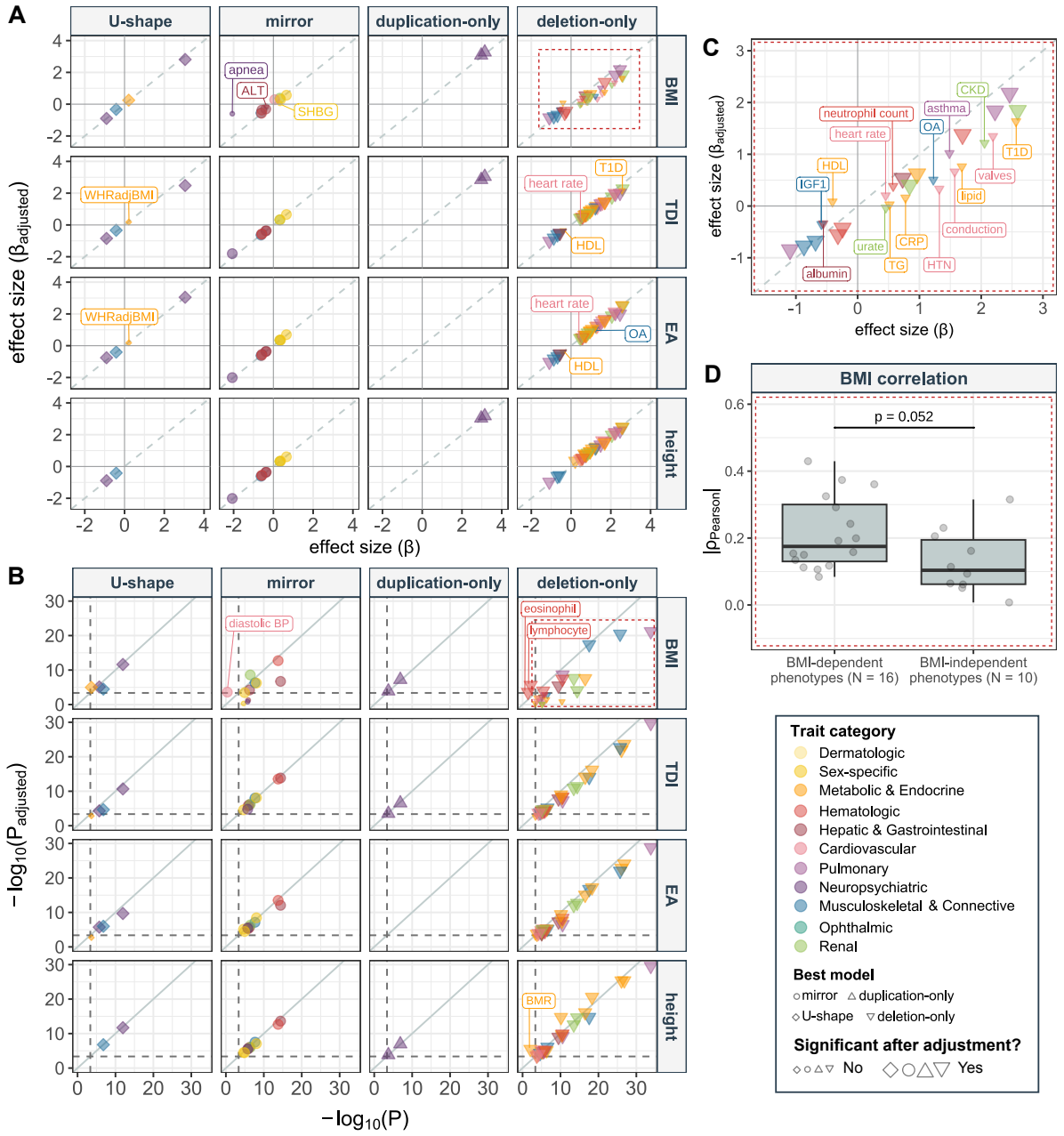


Figure 6.4: Adjustment for potential mediators of 16p11.2 BP4-5 pleiotropy.
(**A**) Effects (beta) of 16p11.2 BP4-5 CNVs on traits with adjustment for potential mediators (y-axis) – i.e., body mass index (BMI), Townsend deprivation index (TDI), age at end of education (EA), and height (rows, right) – against those without adjustment (x-axis), stratified (columns, top) according to the best (i.e., most significant) association model (shape). Only associations that were significant prior to or become significant after adjustment are plotted. Traits are colored according to physiological system. Size reflects whether the effect is Bonferroni significant (large) or not (small) after adjusting for the potential mediator. Traits losing significance upon adjustment are labeled. Gray dashed diagonal represents the identity line. (**B**) Negative logarithm of p-values of 16p11.2 BP4-5 CNV effects depicted in (A), following the same legend. Traits that become Bonferroni significant after adjustment are labeled. Gray diagonal represents the identity line; Dark gray dashed lines represent the Bonferroni threshold of $p \leq 0.05/117 = 4.3 \times 10^{-4}$. (**C**) Enlargement of the area delimited by a red dashed rectangle in (A), showing the effect of BMI adjustment for deletion-driven association, using the same legend as in (A). (**D**) Pearson correlation of BMI with traits that are significantly associated with the deletion (red dashed square in (B)), stratified according to whether the association with the deletion is lost ("BMI-dependent") or not ("BMI-independent") after adjustment for BMI. The P-value compares the two groups with a two-sided t-test. Number of traits in the two groups is indicated as N. ALT = alanine aminotransferase; BMR = basal metabolic rate; BP = blood pressure; CKD = chronic kidney disease; CRP = C-reactive protein; eosinophil = eosinophil count; HDL = high-density lipoprotein cholesterol; HTN = essential hypertension; IGF-1 = insulin-like growth factor 1; lymphocyte = lymphocyte count; OA = arthritis; SHBG = sex hormone binding globulin; T1D = type 1 diabetes; TG = triglycerides; WHRadjBMI = waist-to-hip ratio adjusted for BMI.

effect on our selected mediators.

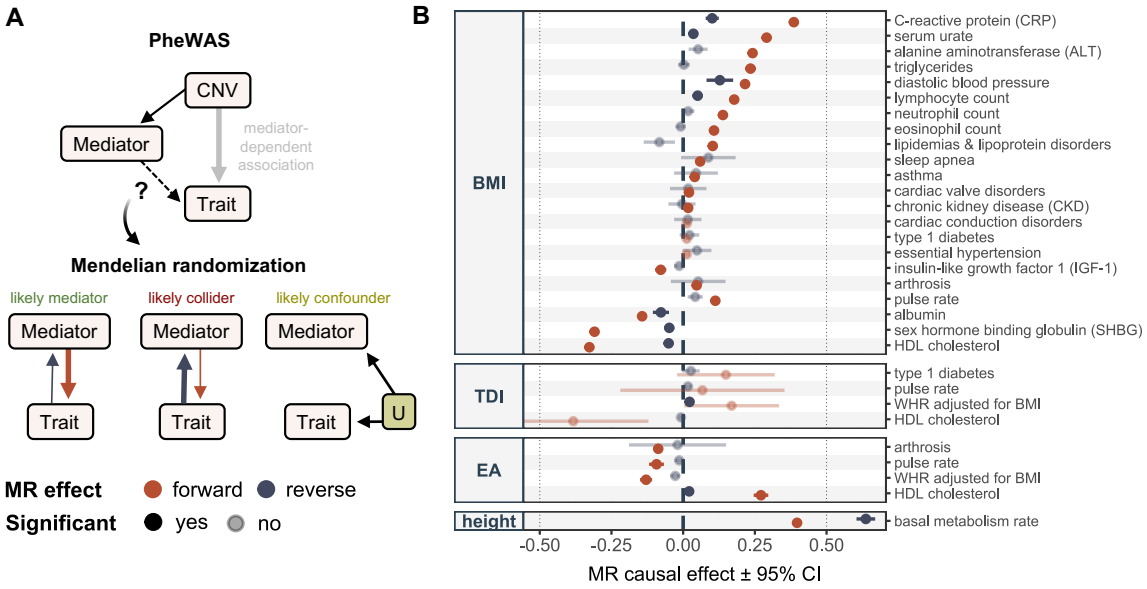


Figure 6.5: Bidirectional Mendelian randomization.
(**A**) Schematic of the links between copy-number variant (CNV), potential mediators, and assessed traits. Covariate-adjusted phenome-wide association studies (PheWAS) identified CNV-trait associations that are dependent on either of the four tested factors (thick gray arrow) in (A). This scenario can be explained through mediation, collider bias, or confounding. We used Mendelian randomization (MR) to assess the genetically determined causal effect of the putative mediator on the trait (forward effect, red arrow) and of the trait on the mediator (reverse effect; dark blue arrow). MR effect arrows are proportional to causal effect sizes. When the forward effect is larger than the reverse one, mediation is a likely scenario; when the reverse effect is larger, the putative mediator likely acts as a collider; absence of causal effects likely indicates presence of an unobserved confounder, U. Depending on the scenario, adjustment for the mediator in the regression analysis might (green) or might not (red) be appropriate, as reflected by the color of each scenario's title. (**B**) Bidirectional forward (red) and reverse (dark blue) MR effects with 95% confidence interval (CI; x-axis truncated on the right) of potential mediators (left y-axis) on traits (right y-axis) for all mediator-trait pairs that either gained or lost significance upon adjustment for the mediator. Non-significant effects ($p > 0.05/62 = 8.1 \times 10^{-4}$) are semi-transparent.

Matched-control analysis

Next, we focused on the 22 traits whose association with the CNV remained significant after adjusting for BMI, height, TDI, or EA. To confirm that these represent cases of genuine direct pleiotropy, we used a matched-control approach that offers the advantage of allowing adjustment for multiple mediatory variables at once but at the cost of losing some statistical power. Specifically, for each of the 58 deletion and 61 duplication carriers with sufficient data to carry out the matching, we identified individuals with matched age (± 2.5 years), sex (identical), BMI (± 2.5 kg/m^2), TDI (± 2), income class (identical), and EA (± 1 year) among a pool of copy-neutral, unrelated, white British UKBB participants (Figure 6.6). For 49 deletion and 60 duplication carriers, at least 25 matched controls could be identified, and phenotype mean or disease prevalence between the two CNV groups and their respective controls were compared (Figure 6.7; Tables S6.4-5). Eleven traits (50%) retained a strictly significant effect ($p \leq 0.05/22 = 2.3 \times 10^{-3}$), affecting six independent physiological systems: musculoskeletal, neuropsychiatric, pulmonary/immune, renal, hepatic, and hematological. Specifically, deletion carriers presented with decreased hand grip strength ($p = 1.4 \times 10^{-3}$; Figure 6.7A), shorter stature ($p = 1.2 \times 10^{-5}$; Figure 6.7B), increased alkaline phosphatase (ALP; $p = 1.8 \times 10^{-3}$; Figure 6.7G), decreased forced vital capacity (FVC; $p = 2.2 \times 10^{-3}$; Figure 6.7R), and increased risk for pneumonia ($p = 3.8 \times 10^{-4}$; Figure 6.7Q) and AKI ($p = 2.9 \times 10^{-4}$; Figure 6.7T). Duplication carriers showed decreased bone mineral density (p

= 6.3 × 10^{-4}; Figure 6.7C), lower aspartate aminotransferase (AST; p = 1.5 × 10^{-3}; Figure 6.7E) and gamma-glutamyltransferase (GGT; p = 2.2 × 10^{-4}; Figure 6.7F) levels, and reduced fluid intelligence (p = 1.6 × 10^{-3}; Figure 6.7I). Noteworthy is the strong mirror effect on platelet count (Figure 6.7P), with higher (p = 1.9 × 10^{-3}) and lower (p = 3.4 × 10^{-4}) counts observed in deletion and duplication carriers, respectively. Whereas for the other phenotypes the other CNV type did not meet strict significance criteria, all effects showed a trend for a mirror effect, except for fluid intelligence and AKI, which followed a U-shape trend. Besides reinforcing its long-established consequence on cognitive function, our results assert the role of the hepatic, musculoskeletal, and pulmonary systems in the 16p11.2 BP4-5 pathophysiology through mechanisms that are independent of the CNV's impact on anthropometric traits and SES.

Finally, we performed sensitivity analyses to validate the robustness of our conclusions. As a negative control, we performed the matched-control analysis for the 24 traits that were significantly associated with 16p11.2 BP4-5 CNVs in our PheWAS but whose association was dependent on adjustment for mediators or that could not be tested in the covariate analysis due to high trait-mediator correlation (Figure 6.8; Tables S6.4-5). In line with these associations being secondary consequences to the effect of the CNV on factors on which the matching was performed, only three traits had a nominally significant CNV association, and none survived Bonferroni correction. This strongly contrasts with our main matched-control analysis, where only three traits lacked a nominally significant effect: recurrent depression (Figure 6.7L), anemia (Figure 6.7O), and cystatin C (Figure 6.7U). This absence of results could either be the result of a loss in statistical power resulting from CNV carrier subsampling or by these associations being driven by a combination of factors on which the matching was performed. The former could be exacerbated by the fact that CNV carriers with the more extreme phenotypes were less likely to have 25 matched controls in the UKBB. To explore this hypothesis, we compared mean trait value or disease prevalence between the subset of CNV carriers used for the matched-control analysis and the one excluded due to missing data or lack of a sufficient number of matched controls (Figure 6.9; Tables S6.4-5). Except for recurrent depression and FVC, all comparisons were non-significant (p ≥ 0.05), indicating that subsampling does not strongly impact our results. For recurrent depression, the only three duplication carriers diagnosed with the disease were not included in the matched-control analysis (p = 0.03; Figure 6.9L), indicating that the non-significant effect of the duplication on recurrent depression (Figure 6.7L) is likely caused by subsampling. For FVC, excluded deletion carriers exhibited a more pronounced phenotypic decrease than the ones retained for the matched-control analysis (p = 0.02; Figure 6.9R), suggesting that an even more extreme difference would have been observed if these individuals had been included in the matched-control analysis (Figure 6.7R). Conversely, the role of the CNV on anemia risk and cystatin C is likely driven by the effect of the CNV on adiposity and SES.

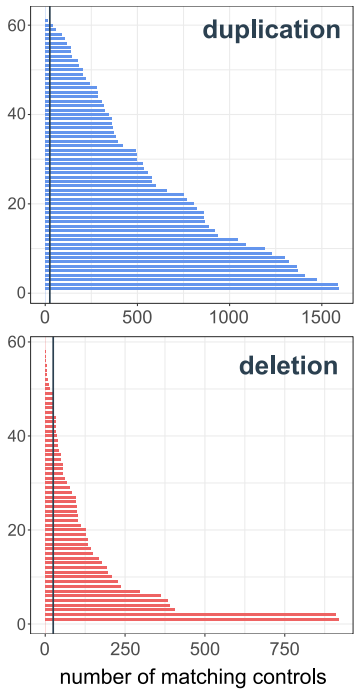


Figure 6.6: Number of matched controls per 16p11.2 BP4-5 CNV carrier.
Total number of identified matched controls (x-axis) per 16p11.2 BP4-5 duplication (N = 61; blue) and deletion (N = 58; red) carrier (y-axis). The black vertical line represents the cutoff of 25 randomly sampled matched controls per CNV carrier. In total 60 duplication and 49 deletion carriers passed this threshold.

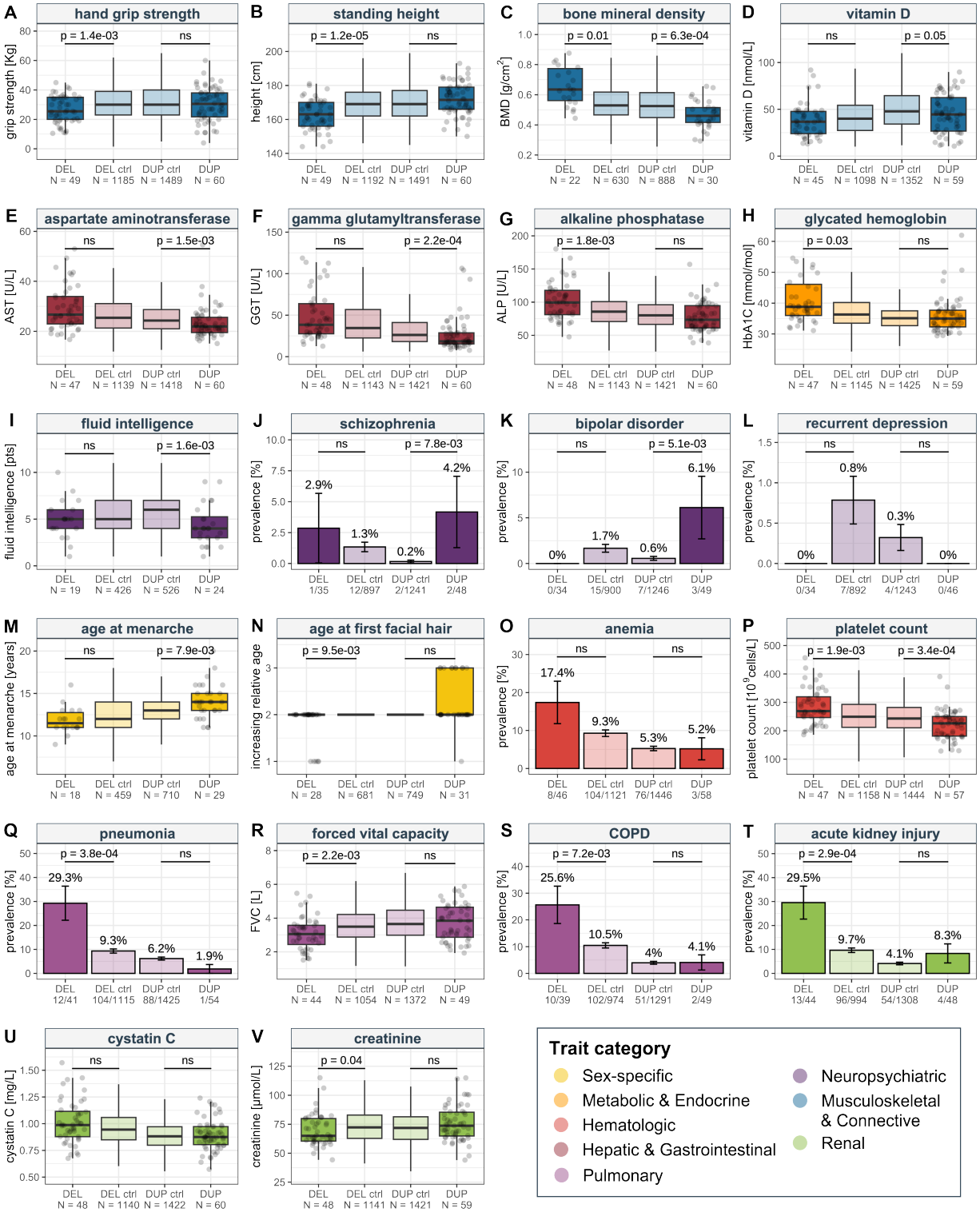


Figure 6.7: 16p11.2 BP4-5 CNV carriers matched-control analyses.
(**A-V**) Comparison between deletion (DEL) and duplication (DUP) carriers (darker shade) and their respective matched controls (DEL or DUP ctrl; lighter shade) for 22 traits that remained Bonferroni-significant after individually adjusting for body mass index (BMI), height, Townsend deprivation index (TDI), and age at end of education (EA). For quantitative traits, data are represented as boxplots without outliers and data points for CNV carriers are shown as gray dots. Sample size of each group is indicated as N. P-values of two-sided t-tests comparing CNV carriers to matched controls are indicated. For binary traits, bars represent disease prevalence in percentage and error bars represent the standard error. Number of cases and total sample size for each group is indicated. P-values of two-sided Fisher tests comparing CNV carriers to matched controls are indicated. "ns" indicates $p > 0.05$. Traits are colored according to physiological systems. COPD = chronic obstructive pulmonary disease.

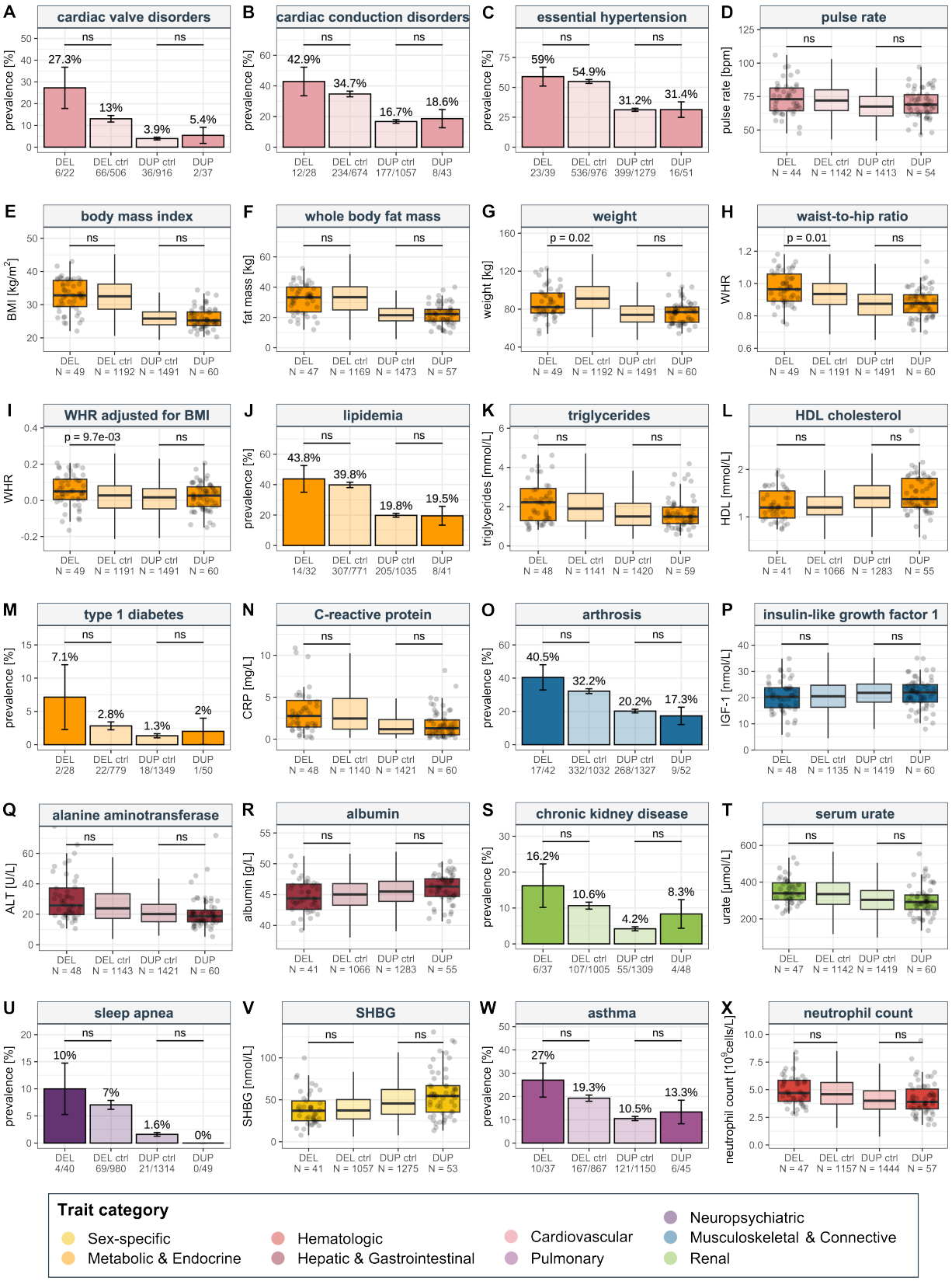


Figure 6.8: 16p11.2 BP4-5 CNV carriers matched-control analyses negative control.
(**A-X**) Comparison between deletion (DEL) and duplication (DUP) carriers (dark shade) and their respective matched controls (DEL or DUP ctrl; lighter shade) for 24 traits that were significantly associated with 16p11.2 BP4-5 CNVs in our PheWAS but whose association was dependent on adjustment for mediators or that could not be tested in the covariate analysis due to high trait-mediator correlation. For quantitative traits, data are represented as boxplots without outliers and data points for CNV carriers are shown as gray dots. Sample size of each group is indicated as N. P-values of a two-sided t-test comparing CNV carriers to matched controls are indicated. For binary traits, bars represent disease prevalence in percentage and error bars represent the standard error. Number of cases and total sample size for each group are indicated. P-values of two-sided Fisher tests comparing CNV carriers to matched controls are indicated. "ns" indicates $p > 0.05$. Traits are colored according to physiological systems. SHBG = sex hormone binding globulin.

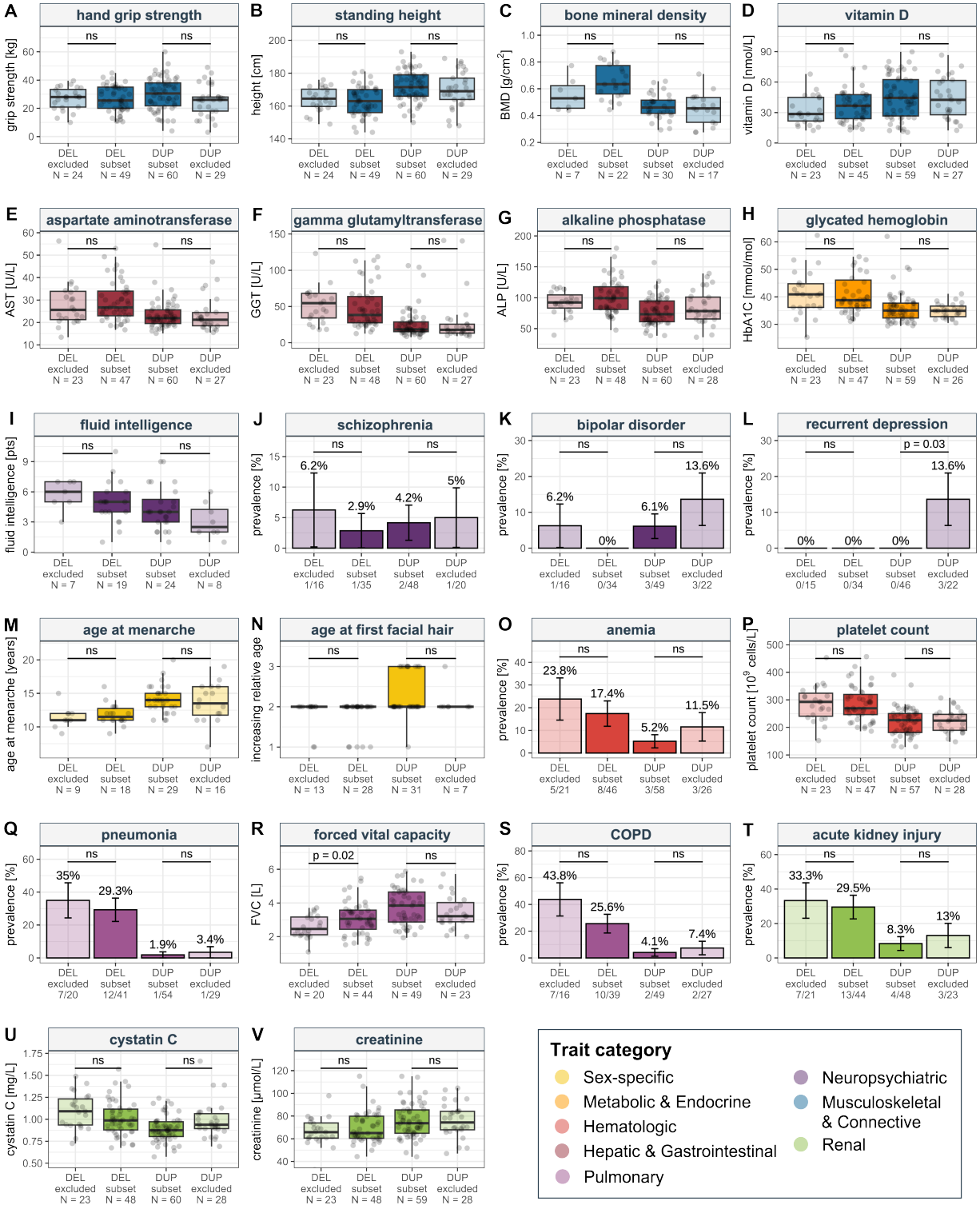


Figure 6.9: Impact of CNV carriers subsampling on matched-control analyses.
(**A**-**V**) Comparison between deletion (DEL; N_{max} = 49) or duplication (DUP; N_{max} = 60) carriers that were in the subset (subset; darker shade) used for the matched-control analyses against deletion (DEL; N_{max} = 24) or duplication (DUP; N_{max} = 29) carriers that were not included due to lack of data (excluded; lighter shade) for 22 traits which remained Bonferroni-significant after adjusting for body mass index (BMI), height, Townsend deprivation index (TDI), and age at end of education (EA). For quantitative traits, data are represented as boxplots without outliers and data points for CNV carriers are shown as gray dots. Sample size of each group is indicated as N. P-values of two-sided t-test comparing CNV carriers to matched controls are indicated. For binary traits, bars represent disease prevalence in percentage and error bars represent the standard error. Number of cases and total sample size for each group is indicated. P-values of two-sided Fisher tests comparing CNV carriers to matched controls are indicated. "ns" indicates $p > 0.05$. Traits are colored according to physiological systems. COPD = chronic obstructive pulmonary disease.

Discussion

In this study, we perform a comprehensive PheWAS assessing the relation between 16p11.2 BP4-5 CNVs and 117 complex traits and diseases in the general population through four dosage mechanisms of action. Our results confirm the extreme pleiotropy of 16p11.2 BP4-5 rearrangements, with 46 traits associating with the CNV. In line with the more deleterious nature of the deletion, haploinsufficiency associated with 38 unique traits, while only 10 traits associated with the region's duplication. Further emphasizing how the same genetic region can affect different traits through different dosage mechanisms, we identify traits for which the loss and gain of a copy had an opposite (e.g., BMI or platelet count) or alternatively, a similar (e.g., grip strength or fluid intelligence) consequences on the phenotype. Besides assessing the role of dosage in pleiotropy, we also estimated the fraction of associations that are likely to be secondary to some hallmark features of the CNV and validated through bidirectional MR that mediation is a likely scenario. While height did not mediate any associations, sixteen (61%) of the deletion-driven associations were found to be BMI-dependent, thirteen of which (81%) received support from MR for a scenario wherein the association is consequential to an initial increase in BMI. Conversely, the role of EA and TDI was more subtle, with only five associations showing confounding by these factors. Importantly, some associations were found to be independent of all the tested mediators, suggesting genuine direct pleiotropy of the region on musculoskeletal, hepatic, metabolic, neuropsychiatric, reproductive, hematological, pulmonary, immune, and renal function.

Our findings have far-reaching consequences for clinical practice and highlight knowledge gaps. First, our results show that increased BMI in deletion carriers drives numerous adult-onset comorbidities. Studies have shown that weight gain in 16p11.2 BP4-5 deletion carriers starts during early childhood, rapidly progressing to obesity (*426*, *500*, *538*–*540*). This emphasizes the importance of following pediatric cases by a dedicated team of endocrinologists and nutritionists who can implement a weight control strategy at an early age to attenuate ensuing adult comorbidities. Second, we show that some other traits are affected independently of the CNV's effect on BMI, cognition, and SES. Besides recapitulating well-established hallmark features, such as the CNV's negative impact on cognitive ability or the duplication-specific risk of bipolar disorder or depression, we also link the CNV with milder afflictions of systems that had previously been implicated in clinical cohorts. For instance, increased risk for AKI might be the consequence of subclinical structural defects of the kidney that could affect renal function in the long term, paralleling the predisposition of deletion carriers to congenital anomalies of the kidney and urinary tract (*277*, *541*, *542*). Similarly, increased risk for pneumonia might reflect an impaired immune system that is exacerbated into a full-blown immunodeficiency in deletion carriers that also present with a loss-of-function variant in *CORO1A* (*543*) (MIM: 605000). Other traits that are affected through BMI-independent mechanisms, such as bone mineral density, platelet counts, pulmonary function, and liver enzymes have not been linked with the CNV in clinical cohorts and future research should establish how often these traits are altered in carriers and which are the molecular mechanisms that mediate this pleiotropy. These could be explored by gene-to-trait mapping strategies

such as rare variant gene burden tests (*6*, *145*), as well as MR (*173*) or colocalization (*167*) that integrate association signals from common SNP-GWAS with transcriptomic and proteomic data to pinpoint genes linked with specific phenotypes. These data could also be leveraged to generate gene-by-trait association matrices whose clustering may reveal groups of traits with shared underlying genetic influences and for which CNV associations are more likely to disappear upon adjustment for one another. Thirdly, our results expose intriguing findings, casting light on questions that remain unanswered by the current study. For instance, the BMI-dependent association of the deletion with type 1 diabetes could be driven by misdiagnosing type 2 diabetes as type 1 due to early-onset diabetes following early-onset obesity. We also identify an association between the deletion and decreased creatinine levels. Creatinine levels are typically elevated in patients with renal dysfunction, as is the case for many deletion carriers. We speculate that these results could be the consequences of reduced hepatic function or muscle mass, both of which are present among deletion carriers. Similarly, it remains unclear whether elevated levels of ALP – for which levels of specific isoenzymes were not determined in UKBB – reflect hepatic, renal, or skeletal dysfunction. Validation of these hypotheses requires in-depth phenotypic characterization of carriers' medical records but will be crucial to better define the molecular pathophysiology of 16p11.2 BP4-5 CNV carriers and hopefully lead to actionable insights related to the management of the condition's comorbidities.

Our study is not without limitations. First, by assessing a relatively homogenous cohort, our study likely misses pleiotropic consequences that are only expressed in certain genetic or environmental backgrounds, a phenomenon exacerbated by the relatively small absolute number of CNV carriers which hinders our statistical power. Future studies are needed to confirm trends that we observe at a sub-significant level. Second, we decided to focus on only four covariates, which based on the literature, represent strong candidates to mediate indirect pleiotropic consequences of the region's rearrangement. While height and BMI can be measured with relatively high accuracy, EA and TDI only offer rough and imperfect proxies for complex characteristics such as cognitive function and SES, possibly explaining their weaker mediatory role. Other factors that we did not assess might mediate the relation between 16p11.2 BP4-5 CNVs and some of the associated traits. Third, the conducted MR analysis comes with its own limitations, namely violation of the exclusion-restriction assumption via correlated pleiotropy, which may have resulted in false positive mediator-to-trait causal effects (*544*, *545*). Still, if both adjusted and unadjusted regression analyses show a significant CNV effect, we can convincingly suggest that independent pleiotropic mechanisms are at play. Finally, while our study brings us a step closer to understanding the pleiotropy of the region, it fails to provide molecular insights into mechanisms of pleiotropy, for which experimental approaches and leveraging of other mutational classes offer promising avenues.

In conclusion, our study provides a framework to start disentangling the complex pleiotropic patterns associated with genomic disorders. For 16p11.2 BP4-5, the latter appears to be a mixture of indirect effects mediated by the impact of the CNV on adiposity and cognition, and direct effects on a broad range of physiological systems. This suggests that

independent molecular mechanisms are involved in translating dosage changes into the many comorbidities linked to the genomic disorder.

Acknowledgments

We thank UKBB biobank participants for sharing their data. Computations were performed on the Urblauna server from the University of Lausanne. The study was funded by the Swiss National Science Foundation (31003A_182632, AR; 310030_189147, ZK), Horizon2020 Twinning projects (ePerMed 692145, AR), and the Department of Computational Biology (ZK) and the Center for Integrative Genomics (AR) from the University of Lausanne.

Declaration of interests

The authors declare no competing interests.

Supplemental tables

Supplemental tables are available for download as a single Excel file.

- **Table S6.6.1** 16p11.2 BP4-5 CNV phenome-wide association study (PheWAS).
- **Table S6.6.2** Covariate regression analysis.
- **Table S6.6.3** Bidirectional Mendelian randomization (MR) analysis.
- **Table S6.6.4** Matched-control analysis – quantitative traits.
- **Table S6.6.5** Matched-control analysis – binary disease traits.

6.5 The pleiotropic spectrum of proximal 16p11.2 CNVs

Chiara Auwerx [1,2,3,4], Zoltán Kutalik [2,3,4], Alexandre Reymond [1,*]

[1] Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland; [2] Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland; [3] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; [4] University Center for Primary Care and Public Health, 1010 Lausanne, Switzerland; [*] **Correspondence**.

Abstract

Recurrent genomic rearrangements at 16p11.2 BP4-5 represent one of the most common causes of genomic disorders. Originally associated with increased risk for autism spectrum disorder, schizophrenia, intellectual disability, adiposity, and head circumference, these CNVs have since been associated with a plethora of phenotypic alterations, albeit with high variability in expressivity and penetrance. Here, we comprehensively review the pleiotropy associated with 16p11.2 BP4-5 rearrangements to shine light on its full phenotypic spectrum. Illustrating this phenotypic heterogeneity, we find many parallels between findings gathered from clinical versus population-based cohorts, which often point to the same physiological systems, and emphasize the role of the CNV beyond neuropsychiatric and anthropometric traits. Revealing the complex and variable clinical manifestations of this CNV is crucial for accurate diagnosis and personalized treatment strategies for carriers. In a second time, we discuss areas of research that will be key to identifying factors contributing to phenotypic heterogeneity and gaining mechanistic insights into the molecular pathways underlying observed associations, while demonstrating how diversity in patients, cohorts, experimental models, and analytical approaches can catalyze discoveries.

Hallmarks of 16p11.2 BP4-5 rearrangements

Chromosome 16 is particularly rich in segmental duplications, which are typically defined as clusters of repeated sequences larger than 1 kb (*502*, *546*). Due to their high sequence similarity (≥ 90%), segmental duplications promote **non-allelic homologous recombination (NAHR)**[4] and form the breakpoints (BPs) of recurrent genomic rearrangements. These rearrangements are at the origin of genomic disorders through the deletion and/or reciprocal duplication of one or more **dosage-sensitive genes**[5] (*180*). The 16p11.2 **cytoband**[6] comprises five segmental duplication clusters termed BP1-5 (Figure 6.10), two of which (BP4 and BP5) underwent a rapid, *Homo sapiens*-specific expansion that favors the creation of proximal 16p11.2 copy-number variations (CNVs; MIM: 611913; 614671) (*41*). While the exact breakpoints of the 16p11.2 BP4-5 rearrangements might vary between individuals, the recurrent CNV encompasses a core region of ~600 kb (Table 6.3), which overlaps 27 unique protein-coding genes, as well as 4 multi-copy genes mapping to the low-copy repeat flanking regions. Expression of 16p11.2 BP4-5 genes is positively correlated with the CNV dosage (*547*, *548*), with no dosage compensation. Hinting at the deleterious potential of 16p11.2 BP4-5 rearrangements, some of these genes are **evolutionarily constrained**[7] and/or have been linked to Mendelian disorders in the Online Mendelian Inheritance in Man (OMIM) (Figure 6.10). Accordingly, multiple mice models deleted for single 16p11.2 BP4-5 orthologs show embryonic or preweaning lethality (Box 1; Figure 6.11). While no homozygous 16p11.2 BP4-5 deletion has been reported, suggesting lethality of such a rearrangement, triplication

4: Homologous recombination between two DNA regions that typically show high similarity but are not alleles. NAHR is the main mechanism behind *de novo* 16p11.2 BP4-5 CNV generation.

5: Dosage-sensitive genes will lead to pathogenic consequences when present in more or less than two functional autosomal copies, in which case they are specifically referred to as triplosensitive and haploinsufficient, respectively.

6: Approximate chromosomal locations defined based on bands produced by Giesma-staining. Cytobands are termed by chromosome number (e.g., "16"), arm (e.g., "p"), followed by region (e.g., "1"), band (e.g., "1"), and sub-band (e.g., "2"), the three last ones being numbers of increasing value from centromere to telomere, to describe a genomic location (e.g., 16p11.2).

7: Constrained genetic regions are depleted of deleterious genetic variants. Such constraint indicates functionality, under the assumption that mutations have pathogenic consequences and will be purged by natural selection.

– either in tandem (549) or due to biparental inheritance (550, 551) – has been reported in four individuals to date. More common is the loss or gain of a single copy, resulting in a heterozygous deletion and duplication (Figure 6.12A), which will represent the focus of this review.

Table 6.3: Genomic coordinates of the 16p11.2 BP4-5 rearrangement.
ClinGen coordinates for the minimal region affected by the 16p11.2 BP4-5 rearrangements in three human reference genome builds. Coordinates in GRCh37 were lifted over with the University of California Santa Cruz (UCSC) LiftOver tool. Because breakpoints might occur at several locations within the segmental duplication region, exact genomic coordinates and length might vary across individuals.

	chromosome	start [bp]	end [bp]
GRCh37 (hg19)	16	29,649,997	30,199,852
GRCh38 (hg38)	16	29,638,676	30,188,531
T2T-CHM13	16	29,920,721	30,473,113

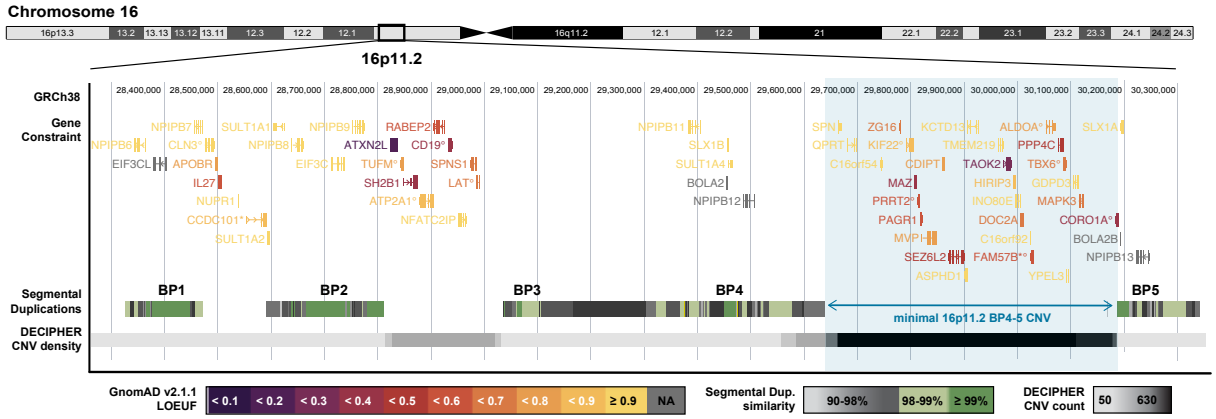


Figure 6.10: Genomic landscape of the 16p11.2 region.
Overview of 16p11.2 cytoband (GRCh38), with the minimal 16p11.2 BP4-5 region highlighted in blue. Upper track: exonic structure of protein-coding genes overlapping the region colored according to GnomAD v2.1.1 loss-of-function observed over expected upper bound fraction (LOEUF) score. Small LOEUF (< 0.35) indicates selection against loss-of-function variants in the gene, i.e., constraint. Genes with no LOEUF score are in gray. Tagged genes: °Indicates OMIM morbid genes; *Have a new HGNC symbol since the GnomAD v2.1.1 release (*CCDC101* = *SGF29*; *FAM57B* = *TLCD3B*). Middle track: segmental duplications colored according to similarity degree, ranging from 90% to ≥ 99%. These form the breakpoints (BP) for recurrent copy-number variants (CNVs). Lower track: Density of CNVs reported in the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources DECIPHER (06/12/2020) (265) colored according to CNV count. While rearrangements of the BP4-5 interval are the most common, rearrangements between other BPs have been described, in particular the 220 kb interval between the BP2-3 (MIM: 613444), the second most common CNV in the region.

Box 1. Animal models of 16p11.2 BP4-5 rearrangements

Three different series of CNV mice models approximating the 16p11.2 deletion (*Del*/+) or reciprocal duplication (*Dup*/+) have been engineered (552–554). The oldest mouse models' rearrangement extends on the 7qF3 mouse chromosome beyond the region **syntenic**[8] to the 16p11.2 BP4-5 interval orthologous to the single copy genes of the BP4-5 interval. Specifically, it ranges from *Slx1b* to *Sept1*, while at the same time excluding *Sult1a1*, one of the multi-copy genes in the breakpoint region (552). That gene is also excluded from the second deletion mouse model (553). The third set of models modifies the number of copies of all syntenic genes orthologous to unique genes of the BP4-5 interval (Figure 6.10), i.e., *Sult1a1* to *Spn* (554). Importantly, none of these models is fully representative of the human rearrangements as the segmental duplication regions forming BP4 and 5 are *Homo sapiens*-specific (41) and human deletion carriers can retain multiple copies of *BOLA2A/B* (MIM: 613182), *SLX1A/B* (MIM: 615822; 615823) and *SULT1A3/4* (MIM: 600641; 615819), while

8: Genetic region with conserved gene order across species.

duplication carriers have an even higher number copy of these genes. For instance, human deletion carriers have a mode of four copies of *BOLA2*, compared to six for healthy controls (*40*). Compounded by the poor reproducibility of mouse behavioral tests often used to proxy human 16p11.2 BP4-5 phenotypes, differences in model engineering and/or genetic background can lead to artefactual findings. To mitigate this, a consortium of laboratories recently set out to replicate their findings across the three deletion models, highlighting divergences in results across models despite globally concordant conclusions (*555*). The recent engineering of two series of rat models that delete and reciprocally duplicate the *Sult1a1-Spn* interval opens the possibility of studying the 16p11.2 CNVs in outbred models (Sprague Dawley and Long Evans) (*272*, *556*).

Another approach is to target individual genes. The International Mouse Phenotyping Consortium (IMPC) (*557*) has produced knockout mice for 24 genes spanning the region and flanking breakpoints, for which broad phenotyping is available (Figure 6.11). Detailed neuroanatomical phenotypes were assessed for 20 of them (*274*) and a similar screen in zebrafish (*558*) revealed that most genes in the regions are required for proper nervous system development. While a comprehensive description of all animal models individually knocked down for 16p11.2 BP4-5 orthologs falls out of the scope of this review, many single genes models partially replicate phenotypes observed in 16p11.2 BP4-5 CNV carriers. Furthermore, multiple studies explored double/triple hemi-deletion and their reciprocal triplosensitivity in *Drosophila* (*559*), zebrafish (*266*, *272*, *328*, *560*), and mice (*273*, *274*, *561*).

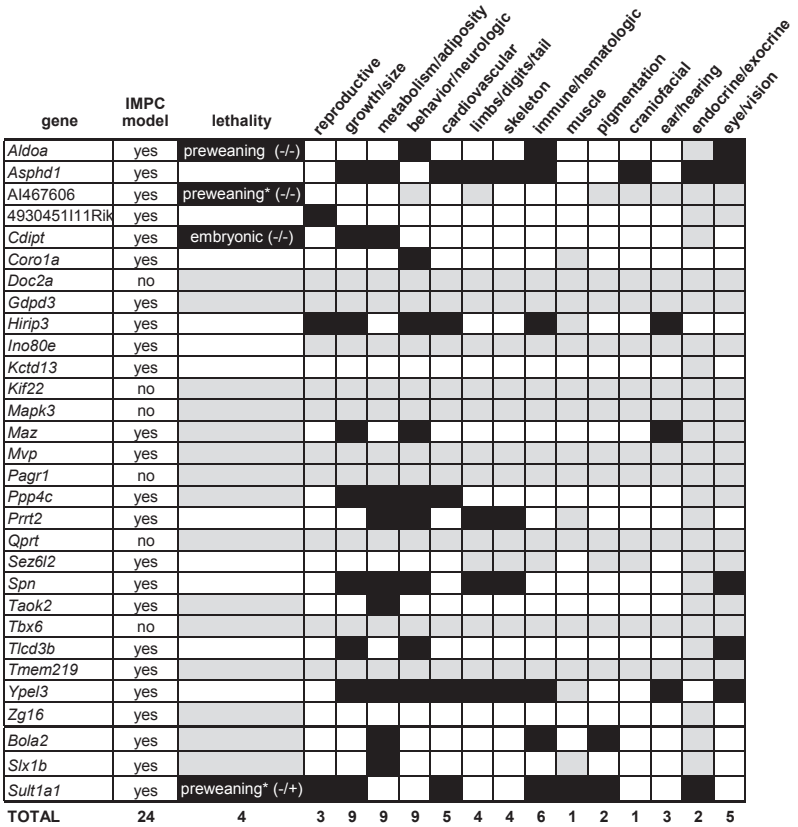


Figure 6.11: International Mouse Phenotyping Consortium 16p11.2 BP4-5 mouse models.
Alphabetical list of mouse genes orthologous to human 16p11.2 BP4-5 genes and whether a knockout was generated by the International Mouse Phenotyping Consortium (IMPC). Subsequent columns indicate phenotypes that were assessed, with black, white, and gray cells indicating whether the system was significantly affected, not affected, or not assessed, respectively. All systems for which at least one model exhibited a phenotype are shown. The total number of affected models is indicated in the last row. For lethality, the stage is indicated, along with whether observed in homozygous (-/-) or heterozygous (+/-) models and if fully penetrant or not (*). Genes in the flanking breakpoint regions are at the bottom.

Studies in clinical cohorts allow to estimate the prevalence of 16p11.2 BP4-5 rearrangements to 1 in 360 and 1 in 390 for the deletion and reciprocal duplication, respectively (Table 6.4). The slightly higher deletion prevalence hints at their stronger deleteriousness, which is reflected by a higher global **penetrance**[9] (47%) compared to duplications (28%) (*499*). In line with their higher pathogenicity, deletions have a higher *de novo* rate, with estimates ranging between 60% and 90%, compared to 25% for duplication carriers (*193*, *286*, *499*). Unlike other CNVs linked to genomic disorder, which tend to occur more frequently on the paternal haplotype, *de novo* 16p11.2 BP4-5 CNVs exhibit up to 90% maternal transmission bias which can neither be explained by older maternal age nor by imprinting, suggesting that the 16p11.2 BP4-5 is a female-specific recombination hotspot (*193*, *286*). Specifically, 16p11.2 BP4-5 CNVs were established as an important susceptibility risk factor for autism spectrum disorders (ASD) (*386*, *422*, *525*, *526*), developmental delay and intellectual disability (DD/ID) (*523*, *524*, *562*), schizophrenia (SCZ) (*387*, *425*), and seizure disorders (*389*, *421*, *562*). Additionally, mirror effects on body mass index (BMI) (*325*, *326*, *426*) and head circumference (*421*) were described, with deletion carriers presenting with obesity and macrocephaly, while duplication carriers tended to be underweight and microcephalic.

9: The penetrance of variant A for a binary trait B describes the fraction of individuals carrying the A genotype who will present with trait B. If penetrance is incomplete (e.g., 60%), then not all individuals will present with the phenotype (i.e., 6 out of 10 carriers). Note that if diseases are considered on a liability scale or in terms of severity of clinical presentation, they can be described through expressivity.

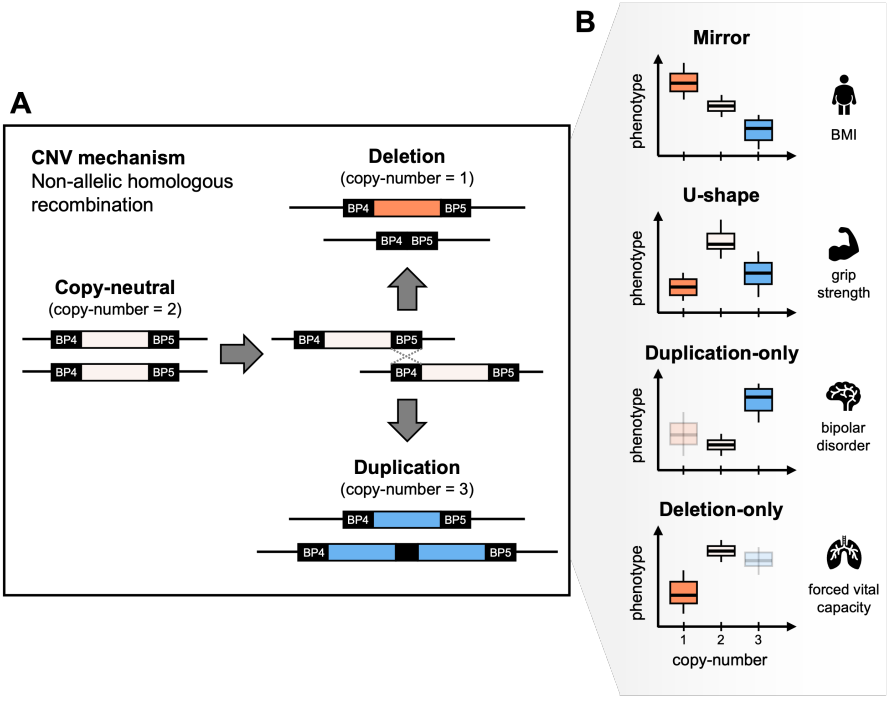


Figure 6.12: Models of CNV dosage mechanisms.
(**A**) Most common copy-number states for the 16p11.2 BP5-4 locus, including the copy-neutral state (2 copies; white), deletion (1 copy; red), and duplication (3 copies; blue), which typically arise through non-allelic homologous recombination (NAHR). (**B**) Phenotypic distribution, shown as boxplots, of individuals with different 16p11.2 BP5-4 copy-number states according to four dosage mechanisms: an additive mirror mechanism wherein deletion and duplication affect the phenotype in the opposite direction, a U-shape mechanism wherein any deviation from the copy-neutral state affects the phenotype in the same direction, and a duplication-only or deletion-only mechanism wherein only duplication or deletion carriers deviate from the copy-neutral phenotypic distribution, respectively. For the two last models, deletion and duplication carriers (semi-transparent) are not assessed to obtain the effect of the duplication and deletion, respectively. Different traits can follow distinct dosage models. At the right is one example trait behaving according to the adjacent mechanism (*563*).

Table 6.4: Prevalence estimates of 16p11.2 BP4-5 CNVs in clinical cohorts.
Prevalence of 16p11.2 BP4-5 deletion and duplication estimated from non-overlapping clinical cohorts, ascertained for various phenotypes. The cohort description includes the cohort's name, sample size (N), the predominant age group, and the proportion of females (♀). Country reflects where samples were recruited; The predominant ancestry group usually matches the most common ancestry group of the recruitment country. Relatives specifies if there are relatives present (Yes) or not (No) in the cohort. The number of carriers (N) and prevalence (Prev) of the deletion (DEL) and duplication (DUP) are reported. Empty cells reflect data that were not reported. Average sample size and prevalence are calculated and put in comparison to the prevalence of carriers among individuals with at least one of the 54 diseases considered in a large meta-analysis (with sample overlap) (*229*). ADHD = attention-deficit hyperactivity disorder; ASD = autism spectrum disorder; BP = Bipolar disorder; CA = congenital anomalies; CAKUT = Congenital anomalies of the kidney and urinary tract; DD/ID = Developmental delay/intellectual disability; Int. = International; SCZ = schizophrenia; YA = young adults.

Cohort	N	Age	♀	Country	Relatives	Disease	N_{DEL}	$Prev_{DEL}$	N_{DUP}	$Prev_{DUP}$
Baylor Genetics Laboratories (*564*)	54,407	pediatric		USA	No	DD/ID, ASD, CA	186	0.342% (1/290)	136	0.250% (1/400)
iPSYCH2012 (*408*)	35,955	pediatric & YA	43%	Denmark	Yes	Psychiatry (1981-2005)	28	0.078% (1/1,300)	88	0.245% (1/410)
Signature Genomics Laboratories (*499*)	33,226	pediatric		USA		DD/ID, ASD, epilepsy, CA	146	0.439% (1/230)	93	0.280% (1/360)
Epi25 (*410*)	26,699			Int.		Seizures	44	0.165% (1/610)	34	0.127% (1/790)
PGC - SCZ (*466*)	21,094	adult		Int.	No	SCZ			70	0.332% (1/300)
SCZ cohorts (*565*)	9,384	adult	44%	China		SCZ			26	0.277% (1/360)
ADHD cohorts (*566*)	8,883	pediatric & adults	43%	Iceland, Norway	Yes	ADHD	7	0.079% (1/1,270)	17	0.191% (1/520)
CLOZUK1+2 (*567*)	6,934	adults	29%	UK		SCZ	4	0.058% (1/1,700)	47	0.678% (1/150)
DD/ID cohorts (*523*)	4,284	pediatric		Europe	Yes	DD/ID, CA	22	0.514% (1/200)		
Children's Hospital Boston (*568*)	3,450	pediatric		USA		DD/ID, ASD, CA	20	0.580% (1/170)		
Obesity cohorts (*326*)	3,103	pediatric & adults		Europe		Obesity	26	0.838% (1/120)	0	0%
KIMONO + CKiD (*277*)	2,824	pediatric & YA	43%	Int.	No	CAKUT	7	0.248% (1/400)	1	0.035% (1/2,800)
BDRN (*569*)	2,591	adults	69%	UK	No	BP			3	0.116% (1/860)
AGRE + ACC (*570*)	2,195	pediatric & YA	20%	USA	Yes	ASD	9	0.410% (1/240)	8	0.364% (1/270)
ASD cohort (*571*)	1,132	pediatric & YA	22%	Japan		ASD	1	0.088% (1/1,100)	4	0.353% (1/280)
SSCs (*572*)	1,124	pediatric & YA	14%	USA	No	ASD	8	0.712% 1/140)	6	0.534% (1/190)
TOTAL	**217,285**						**508**	**0.276% (1/360)**	**533**	**0.254% (1/390)**
Meta-analysis (*229*)								**0.264%**		**0.153%**

Table 6.5: Prevalence estimates of 16p11.2 BP4-5 CNVs in population cohorts.
Prevalence of 16p11.2 BP4-5 deletion and duplication estimated from non-overlapping population cohorts. The cohort description includes the cohort's name, sample size (N), the predominant age group, and the proportion of females (♀). Country reflects where samples were recruited; The predominant ancestry group usually matches the most common ancestry group of the recruitment country. Relatives specifies if there are relatives present (Yes) or not (No) in the cohort. Recruitment is described, with years of birth in parentheses. The number of carriers (N) and prevalence (Prev) of the deletion (DEL) and duplication (DUP) are reported. Empty cells reflect data that were not reported. Average sample size and prevalence are calculated and put in comparison to the prevalence of carriers among individuals with none of the 54 diseases considered in a large meta-analysis (with sample overlap)(*229*). NB = newborns; Int. = International YA = young adults.

Cohort	N	Age	♀	Country	Relatives	Recruitment	N_{DEL}	$Prev_{DEL}$	N_{DUP}	$Prev_{DUP}$
UKBB (*61*)	331,522	adults	54%	UK	No	Invitation (1936-1970)	73	0.022% (1/4,500)	89	0.027% (1/3,700)
deCODE (*566*)	155,122	adults	54%	Iceland	Yes	General population	56	0.036% (1/2,800)	69	0.045% (1/2,200)
DiscovEHR (*402*)	90,595	adults	61%	USA	Yes	Health care system	59	0.065% (1/1,500)	63	0.070% (1/1,400)
Estonian Biobank (*62*)	89,516	adults	66%	Estonia	No	Health care system	14	0.016% (1/6,400)	11	0.012% (1/8,100)
Bio*Me* (*530*)	24,877	adults	59%	USA	Yes	Health care system	15	0.060% (1/1,700)	4	0.016% (1/6,300)
FINRISK (*385*)	23,053	adults	53%	Finland	No	2,000 samples per Finnish region every 5 years (1992-2012)	6	0.026% (1/3,800)	5	0.022% (1/4,600)
Rosenfeld 2013 controls (*499*)	22,246	adults		Int.		Neurologically normal adults	6	0.027% (1/3,700)	9	0.040% (1/2,500)
iPSYCH2012 controls (*408*)	19,169	pediatric & YA	49%	Denmark	Yes	Random sample (1981-2005)	10	0.052% (1/1,900)	21	0.110% (1/910)
MoBA (*481*)	12,252	NB		Norway	Yes	Children from women attending routine ultrasound (1999-2009)	6	0.049% (1/2,000)	5	0.041% (1/2,400)
NFBC1966 (*385*)	4,895	NB	49%	Finland	No	All children Oulu/Lappland (1966)	3	0.061% (1/1,600)	3	0.061% (1/1,600)
TOTAL	**773,247**						**248**	**0.032% (1/3,100)**	**279**	**0.036% (1/2,800)**
Meta-analysis (*229*)								**0.026%**		**0.032%**

Table 6.6: Prevalence estimates of 16p11.2 BP4-5 CNVs in prenatal cohorts.
Prevalence of 16p11.2 BP4-5 deletion and duplication estimated from non-overlapping prenatal cohorts of pregnant women undergoing amniocentesis due to abnormal ultrasound, high-risk pregnancy, or family history of developmental delay/intellectual disability. The cohort description includes the sample origin, sample size (N), predominant age group, and proportion of females (♀). Country reflects where samples were recruited. Relatives specifies if there are relatives present (Yes) or not (No) in the cohort. Ascertainment describes how participants were recruited. The number of carriers (N) and prevalence (Prev) of the deletion (DEL) and duplication (DUP) are reported. Empty cells reflect data that were not reported. Average sample size and prevalence are calculated.

Cohort	N	Age	♀	Country	Relatives	N_{DEL}	$Prev_{DEL}$	N_{DUP}	$Prev_{DUP}$
West China Second University Hospital (573)	86,035	prenatal		China		55	0.064% (1/1,600)		
Maternal and Child Health Hospital of Hubei (573)	8,578	prenatal		China	Yes	17	0.198% (1/500)	4	0.047% (1/2,100)
Chengdu Women's and Children's Central Hospital (574)	7,078	prenatal		China		3	0.042% (1/2,400)	4	0.057% (1/1,800)
TOTAL	**101,691**					**75**	**0.073% (1/1,400)**	**8**	**0.051% (1/2,000)**

Beyond clinical cohorts

Large biobanks encompassing thousands of individuals allowed estimating the prevalence of 16p11.2 BP4-5 deletions and duplications in the general population to 1 in 3,100 and 1 in 2,800, respectively (Table 6.5), which corresponds to approximatively eight-fold lower estimates than in clinically ascertained cohorts. Interestingly, both our clinical and population prevalence estimates align with the ones obtained for individuals suffering from any or none of the 54 diseases assessed, respectively, in the largest CNV meta-analysis to date (*229*) (Table 6.4; Table 6.5). Furthermore, our deletion population estimate closely matches the one predicted by another study (1 in 3,021), based on clinical and epidemiological data (*564*). Some population cohorts, such as Bio*Me* (*530*), exhibit stronger discrepancies in deletion versus duplication prevalence that might be attributed to slight enrichment for diseased individuals due to their enrollment protocol. These observations showcase the role of **ascertainment bias**[10] in obtaining accurate prevalence estimates. While clinical cohorts are enriched for 16p11.2 BP4-5 CNV carriers, population studies suffer from a **healthy volunteer bias**[11] (*59*), leading to prevalence underestimation. In line with that, prenatal cohorts, which have lower ascertainment, yield intermediate prevalence estimates (Table 6.6), suggesting that true prevalence lies somewhere in between estimates from clinical and population cohorts. Nevertheless, the presence of carriers in cohorts largely considered to be healthy reinforces a model of variable **expressivity**[12] and penetrance. Because biobanks are typically coupled with comprehensive phenotypic assessment and electronic health records, they offer the opportunity to evaluate the consequences of 16p11.2 BP4-5 CNVs in a population that is not ascertained for severe clinical conditions and likely at the milder end of the phenotypic spectrum.

Besides replicating core features associated with 16p11.2 BP4-5 CNV carriers, such as decreased cognitive ability (*300*, *384*, *435*) or the mirror effect on BMI (*295*), phenome-wide analyses consistently highlighted 16p11.2 BP4-5 as one of the most **pleiotropic**[13] structural rearrangements genome-wide (*82*, *208*, *229*, *292*–*294*, *305*, *530*). We recently developed a framework to perform CNV genome-wide association studies (GWAS) in the UK Biobank (UKBB) (*208*, *563*), allowing us to assess the impact of

10: Sampling bias that will lead some individuals to be more or less likely to be included in a study or cohort, so that the resulting sample is not fully representative of the targeted population.

11: Type of ascertainment bias wherein individuals who volunteer to participate in a study tend to be healthier (and often from a higher socio-economic background) than the general population from which they originate. This will affect phenotype prevalence estimates and bias estimates of genetic effect sizes.

12: Degree to which a variant A will impact a typically quantitative trait B. Variable expressivity indicates that not all individuals with variant A will show the same levels of B.

13: Phenomenon through which a single genetic variant or locus is associated with multiple traits.

16p11.2 BP4-5 CNVs on 117 complex traits and diseases according to four dosage mechanisms: an additive mirror model, a U-shape model, and two models assessing the impact of duplications and deletions independently (Figure 6.12B). A total of 46 traits were significantly affected by CNVs in the region ($p \leq 0.05/117 = 4.3 \times 10^{-4}$) (*563*). Deletions were more deleterious, leading on average to 2.8 additional disease diagnoses ($p = 2.6 \times 10^{-24}$), as opposed to 0.3 for duplication carriers ($p = 0.183$). About 9% of the signals were better captured by a U-shape model, including those related to cognitive function and grip strength. Conversely, 22% of the associations exhibited a mirror effect with a positive dosage correlation with puberty timing, liver enzymes, bone mineral density, or sleep apnea risk. The marked difference between the U-shaped and mirror models indicates that disparate evolutionary forces (directional vs stabilizing selection) may act on the expression level of certain genes in the region. Most importantly, and in line with the syndromic nature of 16p11.2 BP4-5 rearrangements, associations involved a broad spectrum of physiological systems (Figure 6.13), even after accounting for potential confounders such as adiposity or cognition (*563*).

Here, we review evidence from both clinical and population studies to describe the full phenotypic spectrum associated with 16p11.2 BP4-5 CNVs. Highlighting the complementarity of these approaches, we discuss the importance of awareness around phenotypic heterogeneity and adoption of diverse data sources and analytic strategies to better diagnose, monitor, and possibly prevent 16p11.2 BP4-5-associated comorbidities.

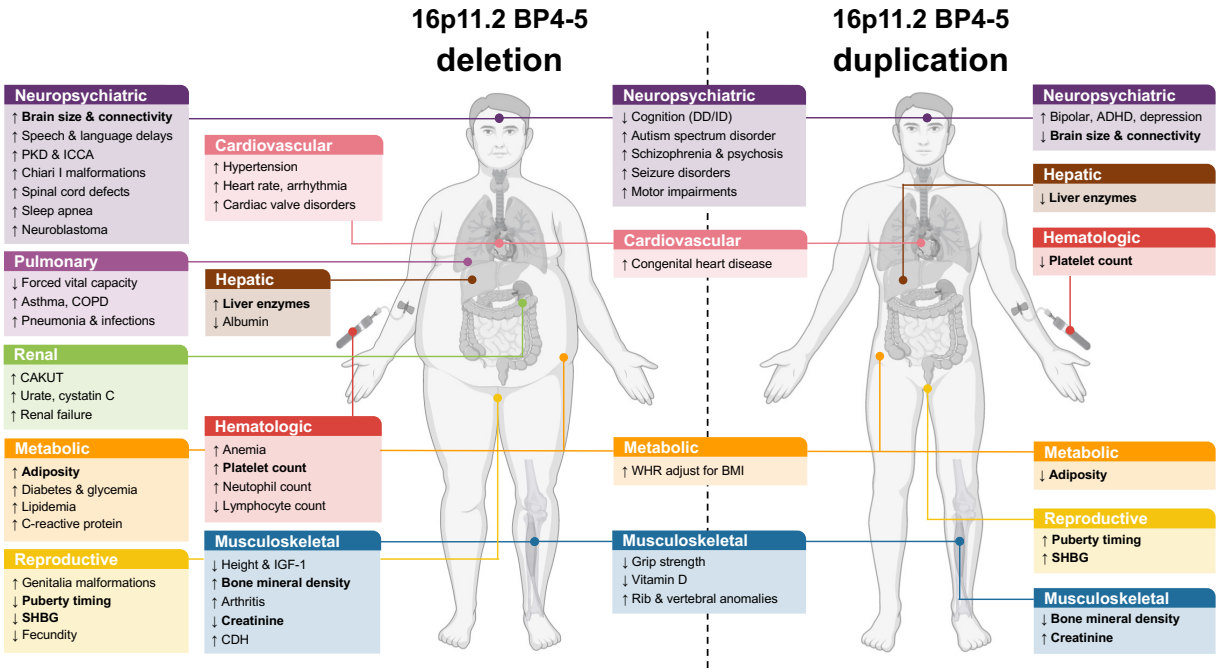


Figure 6.13: Pleiotropy of the 16p11.2 BP4-5 region.
Overview of phenotypes associated with the 16p11.2 BP4-5 deletion (left) and duplication (right) across clinical and population cohorts, organized by physiological system. For each trait, an arrow indicates if the phenotype is increased (upwards) or decreased (downwards) in the CNV carriers, compared to copy-neutral individuals. Phenotypes listed in the middle are shared between deletion and duplication carriers and follow a U-shape model. Phenotypes on the left and right are either specific to deletion or duplication carriers or are affected in opposite directions, in which case they are reported in bold. ADHD = attention-deficit hyperactivity disorder; BMI = body mass index; CAKUT = congenital anomalies of the kidney and urinary tract; CDH = congenital diaphragmatic hernias; COPD = chronic obstructive pulmonary disorder; DD/ID = developmental delay/intellectual disability; ICCA = infantile convulsion with choreoathetosis; IGF-1 = insulin-like growth factor 1; PKD = paroxysmal kinesigenic dyskinesia; SHBG = sex hormone binding globulin; WHR = waist-to-hip ratio.

Pleiotropy of 16p11.2 rearrangements

Neurology

Developmental delay & intellectual disability

DD/ID was present in virtually all probands of early descriptive studies of 16p11.2 BP4-5 duplication and deletion carriers (*421*, *524*, *562*) and studies of clinical cohorts ascertained for DD/ID systematically identified enrichment for 16p11.2 BP4-5 CNV, particularly deletion, carriers (*335*, *523*). Compared to non-carrier parents, deletion probands have an average reduction of 25-35 full-scale intelligence quotient (IQ) points (*500*, *575*), with similar findings for duplication carriers (*423*, *576*). Accordingly, about one third of clinically ascertained CNV carriers meet ID criteria (*529*). Duplication carriers exhibit higher variation in full-scale IQ, with an almost 20-fold enrichment for individuals with extremely low values, compared to deletion carriers (*423*). Reduced cognitive performance among 16p11.2 BP4-5 CNV carriers was replicated in multiple population cohorts (*300*, *384*, *435*), including in our UKBB analysis that found a highly significant U-shape effect on fluid intelligence score with a slightly stronger effect in duplication carriers (*208*, *563*). Together, this makes DD/ID one of the most consistently associated traits with the region's rearrangement.

One crucial component of DD in 16p11.2 BP4-5 CNV carriers is **language and speech impairments**[14], which have systematically been reported (*523*, *524*, *526*, *527*, *562*), and manifest through lower verbal, as opposed to non-verbal IQ, and high (83%) rates of speech and language therapy during childhood among deletion carriers (*500*). All language components (i.e., phonology, lexicon, syntax, semantics, and pragmatics) are negatively impacted in deletion carriers, with milder evidence in duplication carriers, which even tended to outperform familial controls with similar IQ for verbal memory skills (*577*, *578*). Motor speech disorders are also common, with 79% of deletion and 30% of duplication carriers suffering from speech articulation defects (*498*), possibly due to reduced sensorimotor adaptation (*579*). There is evidence that 16p11.2 BP4-5 deletions predispose to childhood apraxia of speech (*580*–*582*), which often co-occurs with receptive (73%) and expressive (70%) language disorders, as well as mild-to-moderate speech impairments (89%) in deletion carriers (*582*). If about three quarter of children carrying a deletion meet childhood apraxia of speech diagnostic criteria, about two thirds of them go undiagnosed (*582*) and prevalence estimates among duplication carriers are currently lacking. While the presence of cognitive delay or ASD diagnosis exacerbates speech and language impairments, they cannot fully account for them, indicating that the latter represent core features of 16p11.2 BP4-5 rearrangements, with exacerbated penetrance in deletion carriers (*577*, *578*, *582*).

14: Language disorders describe difficulties in understanding (receptive language) or getting across (expressive language) a message, which is distinct from speech disorders wherein the individual struggles with forming the sounds necessary to communicate.

Developmental trajectories in childhood are globally similar between deletion and duplication carriers, with an increase in verbal IQ over time (*583*). Yet, deletion carriers showed a decrease in motor and social function over the same period, so that 67% of deletion (and 56% of duplication) carriers end up being diagnosed with developmental coordination disorder at age 6-8 years (*583*). Motor delays include feeding difficulties in newborns (*421*), hypotonia (*498*, *524*), hyporeflexia (*498*), poor agility

(*498*), and impaired balance, speed, and endurance in locomotion tests (*584*). If most of these are observed in both deletion and duplication carriers, the latter showed stronger impairments with additional features such as hyperreflexia (32%) and tremors (43%)(*498*), as well as very late onset walking (*423*). This is paralleled by the finding that duplication carriers had worse accuracy and speed in a battery of neurocognitive assessments evaluating executive function, episodic memory, complex and social cognition, and psychomotor speed, compared to deletion carriers (*585*). Furthermore, diagnosed deletion (N = 48) and duplication (N = 48) carriers in the Vanderbilt University Medical Center's biobank (BioVU) showed increased rates of "abnormal movement and developmental delay" (CNV carriers), "muscle weakness" (deletion carriers), and "speech and language disorders" (duplication carriers) in their electronic health records, even though age of the CNV carrier and age at diagnosis were not reported (*413*). Adult populations are not ideally suited to study language, speech, and motor impairment, and only very few diagnosed cases of language and speech, scholastic skills, and motor impairment are present in UKBB. Yet, decreased grip strength was observed in both deletion and duplication carriers (*208*), suggesting that impaired motor function persists in adulthood.

Structural alterations of the nervous system

Recent efforts have concentrated on identifying brain alterations that could explain the strong predisposition of 16p11.2 BP4-5 CNV carriers for neurodevelopmental and psychiatric disorders. One striking feature includes the global increase of brain size – including total intracranial, white matter, and gray matter volumes – among deletion carriers, which opposes the pervasive size reduction observed among duplication carriers (*586*–*589*) and aligns with the previously described macrocephaly and microcephaly phenotypes observed in deletion and duplication carriers, respectively (*421*, *425*) (see *Craniofacial features*). Changes in brain volume have been modeled in cellular models and cortical **organoids**[15], where dosage negatively correlates with neuron size, dendrite length, and neuronal differentiation (*590*–*592*). Focal cortical anomalies are widespread among CNV carriers. They correlate negatively with full-scale IQ, with duplication carriers exhibiting an increased number of abnormally thin cortex areas, while deletion carriers rather exhibit increased cortical thickness (*593*). Up to a quarter of duplication carriers present with increased ventricular volume (*589*, *594*) and cerebellar tonsillar ectopia/**Chiari type I malformations**[16] (MIM: 118420) have been reported in up to a third of deletion carriers (*423*, *498*, *500*, *594*–*596*). Other defects of the spinal cord such as syringomyelia (MIM: 186700) (*276*, *423*, *500*, *595*) or spina bifida (*276*, *423*, *500*, *597*–*600*) have been reported, especially among deletion carriers. Whereas the precise onset of these structural alterations is unclear, they are often already present at age 5 and remain stable until adulthood (*589*, *594*). Dosage effect of white matter microstructure was also identified (*601*), with deletion carriers consistently showing increased diffusivity that could reflect decreased myelin or axonal density (*601*–*604*). Importantly, anomalies often involve regions involved in auditory, language, speech, and social function (*588*, *593*, *603*, *604*), the reward system (*587*, *588*), or the cerebellum (*586*, *588*), all of which play crucial roles in the etiology of phenotypes commonly observed among 16p11.2 BP4-5 CNV carriers. At the molecular level, neuroanatomical changes have been reported for over 14 mouse models with individual

15: Cells derived from (human) embryonic stem cells or induced pluripotent stem cells cultured into 3D structures that aim at partially recapitulating brain structure and organization, and *in vivo* cell interactions.

16: Cerebellar herniation in the spinal canal due to malformation (or small) skull. Type I malformations are the least severe ones.

16p11.2 BP4-5 gene deletion – including *Mapk3* (*Erk1*) (605), *Taok2* (606), *Mvp* (274), and *Doc2a* (607) – often resulting in cognitive or behavioral deficits. This emphasizes that brain morphology is highly polygenic and regulated by multiple genes of the region.

Aligning with the idea that brain structure correlates with function, impaired prefrontal connectivity was found in human and mouse 16p11.2 BP4-5 deletion carriers (608), with global reinforcement of functional connectivity among deletion carriers and a trend for lower connectivity among duplication carriers, suggesting a dosage effect (609). Specifically, pervasive increase in intra-axonal volume in multiple white matter tracts is already visible at an early age (2 years) in deletion carriers (610). In parallel, several studies have reported atypical neural activity upon auditory (611–613), visual (614, 615), or social (616) stimuli, as well as during preparation of overt speech and hand movement (617), with left hemispheric language specialization being decreased among deletion carriers (617). The deletion mice model correspondingly showed abnormally high activity in the motor cortex during learning in males (618). If these studies identify clear alterations in brain signal processing in deletion carriers, the effect of the duplication remains less clear. Neurophysiological differences might translate into the broad spectrum of phenotypic alterations observed in 16p11.2 BP4-5 CNVs. Indeed, affected brain areas overlap with the ones altered in idiopathic psychiatric cases (588) – with a particularly strong correlation between the effect of the region's deletion and ASD (619) – but also harbor some unique features (586). Importantly, 16p11.2 BP4-5 CNVs exert a stronger effect on overall brain structure (619) and connectivity (609) than idiopathic cases of ASD or SCZ, motivating genotype-first approaches to elucidate the pathological mechanisms of these diseases. This is the approach followed by Enhancing NeuroImaging Genetics through Meta-Analysis CNV (ENIGMA-CNV), which aims to meta-analyze brain imaging data from both population and clinical CNV carriers originating from 38 worldwide research and data collection sites (620). While no ENIGMA study focusing on 16p11.2 BP4-5 has been released to date, others have shown that brain structure profiles defined from clinically ascertained CNV carriers mimicked those of seven duplication and four deletion carriers with available brain imaging in the UKBB (404). Duplication and deletion profiles further associated with 55 and 34 traits, respectively, linking them more broadly to the human phenome (404). The low frequency of 16p11.2 BP4-5 CNVs compounded by the even smaller number of carriers with available brain imaging, will make collaborative approaches crucial to accurately establish the impact of the region's dosage on brain structure and connectivity and interpret the functional consequences of these differences.

Seizure disorders

Early prevalence estimates for seizure disorders and epilepsy from clinically ascertained 16p11.2 BP4-5 cohorts range between 10 and 30% for both duplication and deletion carriers (421, 423, 424, 498, 500, 523, 524, 562, 621), making it one of the phenotypes systematically associated with these CNVs, albeit at lower prevalence than other neurodevelopmental and psychiatric disorders. The association was reiterated by a case-control study that found that both deletion and duplication carriers were enriched in over 26,000 individuals diagnosed with epilepsy and seizures

(*410*). Duplication and deletions are associated with both severe and milder epilepsies. Among 315 cases with **developmental and epileptic encephalopathies (DEE)**[17], 7.9% harbored a rare CNV, including two 16p11.2 BP4-5 duplication carriers with West syndrome and multifocal epileptic encephalopathy (*390*). These results are paralleled by case reports of duplication carriers with epilepsy of infancy with migrating focal seizures (*622*), Landau-Kleffner syndrome (*424*), and epileptic encephalopathy with continuous spike and wave in sleep (*621*), as well as a deletion carrier with West syndrome (*623*). West syndrome was also diagnosed in 0.5% of 390 deletion and 1.1% of 270 duplication carriers (*423*). Milder epilepsies typically resolve without disrupting developmental progress. **Childhood epilepsy with centrotemporal spikes**[18] (MIM: 117100) was diagnosed in 1.5% of duplication carriers (*424*), a finding supported by a smaller study identifying two duplication carriers among 47 cases (*624*). This association was specific to duplication carriers, who were not enriched for other specific epilepsy types (*424*). Conversely, **absence epilepsy**[19] was observed in 33% of deletion but only 5% of duplication carriers (*498*). A recent systematic characterization of seizure disorders among 16p11.2 BP4-5 CNV carriers found that **self-limited familial and non-familial infantile epilepsy (SeLIE)**[20] (MIM: 605751) was the most common seizure disorder among deletion carriers, accounting for 42% of epilepsies (*621*) and was wound in 3 out of 33 deletion carriers in a Dutch study (*625*). SeLIE accounted for only 13% of epilepsies among duplication carriers, which presented with a more heterogeneous disease spectrum (*621*). While we previously reported increased epilepsy risk among UKBB deletion carriers (*82*), the association falls below the threshold for significance in our new re-analysis (*563*). Overall, these results highlight the contribution of 16p11.2 BP4-5 CNVs to a broad spectrum of epileptic disorders with varying severity degrees and suggest that the region's dosage affects epilepsy subtype. Consistent with this hypothesis, the 16p11.2 BP4-5 genes *PRRT2* (MIM: 614386) and *SEZ6L2* (MIM: 616667) act as hubs in an epilepsy protein subnetwork dysregulated in a duplication mouse model, and correcting the dosage of the former gene rescued seizure susceptibility in these mice (*626*). In zebrafish, an epistatic contribution to seizure susceptibility has been reported in double $doc2a^{+/-}fam57b4^{+/-}$ knockdowns (*560*), suggesting an **oligogenic**[21] contribution to the phenotype.

17: Group of rare and severe epileptic syndromes characterized by severe seizures and epileptic activity that leads to cognitive impairment/regression. DEE are often refractory to treatment and associated with early age of onset.

18: Formerly known as Rolandic epilepsy, it is the most common form of epilepsy in childhood and is characterized by seizures originating in the Rolandic area of the brain. Seizures usually disappear in adolescence.

19: More frequent in children, the generalized onset seizures of absence epilepsy are characterized by very brief, sudden-onset periods of "blanking out" and often disappear in adolescence.

20: Formerly known as benign infantile seizures, SeLIE seizures typically start around 6 months and remit within one year of onset, without disrupting developmental progress.

21: Model of genetic architecture wherein a trait is under the control of few genetic loci or genes.

Movement disorders

Paroxysmal kinesigenic dyskinesia (PKD; MIM: 128200), a rare movement disorder characterized by brief and recurrent involuntary movement attacks, has been associated with 16p11.2 BP4-5 deletions (*627*–*631*). PKD can co-occur with SeLIE, a combination of features referred to as infantile convulsion with choreoathetosis syndrome (ICCA; MIM: 602066). These disorders were shown to be caused by heterozygous variants in the epilepsy-hub gene *PRRT2* (*626*, *632*). In a review of 1,444 published cases with 70 distinct *PRRT2* mutations, 42%, 39%, and 14% of affected individuals were diagnosed with SeLIE, PKD, and ICCA, respectively, with the remaining cases suffering from various disorders, including seizures and headache disorders (*633*). Importantly, *PRRT2* mutations can lead to different disorders within the same family (*634*). Single cases of deletion carriers with benign nocturnal alternating hemiplegia of childhood (*635*) and hemiplegic migraine (*636*) suggested that **haploinsufficiency**[22] of *PRRT2* might also be associated with these related pathologies. Yet, UKBB

22: Type of dosage sensitivity wherein a single wildtype copy of a gene is not sufficient to produce the wildtype phenotype. Haploinsufficient genes are therefore intolerant to heterozygous loss-of-function mutations.

deletion carriers were not more prone to migraines and headaches (*82*). Interestingly, a female carrier of a heterozygous *PRRT2* mutation with SeLIE was found to experience sudden and extreme autistic regression at 15 months (*637*). While SeLIE is typically not associated with poor neurodevelopment outcomes, this case highlights a possible contribution of *PRRT2* to the autism phenotype in 16p11.2 BP4-5 CNV carriers. While the pleiotropy and variable expressivity of *PRRT2* haploinsufficiency are well established, further research is required to ascertain the full phenotypic range linked to 16p11.2 BP4-5 deletions.

The first case of PKD among a 16p11.2 BP4-5 deletion carrier presented with dopa-responsive parkinsonism (*627*), and a later case report described a duplication carrier with levodopa-non-responsive early-onset parkinsonism (*638*). This increased rate of tremors and dysrhythmia in duplication carriers and the reduced nimbleness of CNV carriers in general (*498*) suggests that 16p11.2 BP4-5 CNV carriers could have an increased liability for Parkinson's disease. This hypothesis could not be cross-validated in the UKBB (*82*), possibly because UKBB volunteers, like participants of clinical cohorts, are typically too young to assess associations with late-onset diseases such as Parkinson's disease.

Sleep disorders

Based on questionnaire data and compared to familial controls, 16p11.2 BP4-5 CNV carriers have higher rates of sleep disturbance in childhood and adolescence and medical sleep concerns and insomnia in adulthood, even if no difference in sleep duration was observed (*639*). Obesity, which is common among deletion carriers, represents a key risk factor for obstructive sleep apnea, a comorbidity reported by several descriptive studies (*421*, *423*, *599*, *640*). In line with this, increased prevalence of sleep apnea was identified in deletion carriers of both BioVU (*413*) and UKBB (*82*), where the association was demonstrated to be driven by increased BMI (*563*). Conversely, associations with other sleep-related disorders such as insomnia, hypersomnia, or narcolepsy were not identified in UKBB (*82*), despite differential sleep architecture and increased wake time being reported in mouse models of the deletion (*641*–*643*). Studies investigating sleep quality through objective approaches, such as polysomnography, will be required to gauge the extent to which these mouse findings translate to human CNV carriers.

Psychiatry

Autism spectrum disorder

CNVs at 16p11.2 BP4-5 were associated with ASD early on. With about 1% of individuals suffering from ASD across various cohorts (e.g., Autism Genetic Resource Exchange (AGRE), Icelandic cohort, Simon Simplex Cohort (SSC)) being deletion carriers, and another 1% carrying the duplication (*315*, *386*, *422*, *525*, *526*), 16p11.2 BP4-5 CNVs represent one of the strongest genetic risk factors for ASD and are commonly used as a model to study the disease (*644*). In line with that, about 20% of individuals carrying a 16p11.2 BP4-5 CNVs showed autistic features (*423*, *500*, *528*, *529*). For example, 22% and 26% of 294 and 146 16p11.2 deletion and duplication carriers, respectively, presented with ASD with a wider variation for any psychiatric disorder for duplication carriers (*529*). The short arm of chromosome 16 encompasses multiple ASD-associated loci, as rearrangements of the nearby 16p11.2 BP2-3 interval

were similarly associated with ASD (279) (Figure 6.10), and as the entire 33 Mb of this chromosomal arm presents with the greatest excess of autism's common **polygenic**[23] influence (282). Response to social cues was correspondingly altered in mouse (645, 646) and rat (556) deletion models, while duplication models presented an increase in electrophysiological perturbations in regions of the brain critical for social and cognitive functions (647, 648).

23: Model of genetic architecture wherein a trait is under the control of many genetic loci or genes.

Schizophrenia & psychosis

Shortly after describing the association with ASD, the 16p11.2 BP4-5 duplication – but not its deletion – was identified as a major risk factor for SCZ (425). This association was replicated multiple times (466, 567, 649), including in individuals of Han Chinese ancestry (565), leading to a 10-fold increase in SCZ risk with a penetrance of 6.9% (650). These results contrast with recent results from the Danish Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), which did not find any significant effect of 16p11.2 BP4-5 CNVs on SCZ (407). This study also found a damped effect for other SCZ CNVs, such as the 22q11.2 deletion, suggesting that these results might stem from differences in ascertainment. With 4.1% of deletion and 4.6% of duplication carriers being diagnosed with SCZ in UKBB (82), our analysis is compatible with a model wherein both the region's deletion and duplication increase risk for SCZ. While never meeting criteria for significance, deletion carriers have been identified in SCZ clinical cohorts (425, 466, 567, 649). We hypothesize that the milder affliction of deletion carriers in population cohorts unmasks SCZ, whose diagnosis might be impaired in clinically ascertained deletion carriers with severe DD/ID.

Duplication carrier status also increases risk for psychotic symptoms (651), which represents a hallmark of SCZ. Psychosis is common in Alzheimer's disease and shared mechanisms between Alzheimer's with psychosis and SCZ have been hypothesized (652). In line with this, two duplication carriers have been identified among 440 cases of severe Alzheimer's disease with psychosis, while none were found among 729 cases with mild/no psychosis (653). While this enrichment was not significant, further studies assessing the possibility that 16p11.2 BP4-5 CNV carriers are more prone to develop Alzheimer's disease as they age are warranted.

Bipolar disorder

While frequency of *de novo* CNVs is increased in bipolar disorder (654), their role in the disease's etiology is less clear than for SCZ and remains debated (655, 656). The 16p11.2 BP4-5 duplication represents the only CNV robustly associated with the risk of bipolar disorder (425, 569). Importantly, duplications were confirmed to represent a risk factor for bipolar disorder in UKBB (82), with at least 9% of duplication carriers being diagnosed with the condition, even though comorbidity with other psychiatric disorders such as SCZ was not assessed.

Attention-deficit hyperactivity disorder

Children with attention-deficit hyperactivity disorder (ADHD) present with an excess of both large duplications and deletions, irrespective of the concurrent presence of an ID/DD diagnosis (657). ADHD was consistently reported in descriptive studies of clinically ascertained 16p11.2 BP4-5 duplication and deletion carriers, with higher prevalence among duplication carriers (421, 524, 527–529, 583). This association

was confirmed in 8,883 cases of Icelandic and Norwegian origin, which identified the region's duplication as a risk factor for ADHD (*566*), an observation later confirmed in iPSYCH (*407*, *408*). Importantly, a nominally significant association with ADHD remains upon exclusion of ASD and SCZ cases (*566*), indicating that the condition can arise independently of the latter diagnoses. As ADHD is predominantly diagnosed in children, the number of cases in the UKBB is low, preempting association analysis.

Depression

Early literature investigating the role of CNVs in depression reported conflicting results (*658*–*660*) with a single study reporting a nominally significant enrichment for 16p11.2 BP4-5 CNV carriers among 604 patients suffering from major depressive disorder (*661*). After excluding individuals with ASD, SCZ, bipolar disorder, ADHD, and ID, a UKBB study identified the 16p11.2 BP4-5 duplication as one of three recurrent CNVs associated at Bonferroni significance level with self-reported depression (*302*). These results were replicated based on hospital-diagnosed cases in UKBB (*82*) but not in iPSYCH (*407*, *408*), paralleling the dampened effect size observed for SCZ.

Other psychiatric conditions

Over the last 15 years, the pleiotropic effect of 16p11.2 BP4-5 CNVs on psychiatric conditions became increasingly evident (*408*, *423*, *500*, *528*, *529*) and it is not uncommon for CNV carriers to be diagnosed with multiple conditions: on average, clinically ascertained deletion carriers were diagnosed with 2.9 psychiatric conditions, a 10-fold increase compared to familial controls (*527*). These results remain significant when accounting for ASD diagnosis, indicating that they are not solely driven by the latter. Indeed, 16p11.2 BP4-5 CNVs have been linked to additional psychiatric conditions, such as anxiety, disruptive behavior, tic disorders, and obsessive-compulsive disorders (*402*, *498*, *500*, *527*–*529*, *651*, *662*), although with more limited evidence. Perplexingly, neuroticism, which strongly correlates with several psychiatric conditions (*663*), was associated with neither duplication nor deletion carrier status in UKBB (*82*, *208*). While both duplications and deletions are now recognized as important risk factors for psychiatric conditions, current evidence suggests higher prevalence and heterogeneity in diagnoses among duplication carriers (*528*, *529*). In line with this, psychiatric conditions are the only disease type primarily driven by the region's duplication in UKBB. Further research is required to better delineate the precise nature and penetrance of various psychiatric disorders linked to 16p11.2 BP4-5 rearrangements and determine the extent of shared disease mechanisms among them.

Endocrinology & Metabolism

Obesity

Despite obesity being frequent among the first described 16p11.2 BP4-5 deletion carriers (*422*, *523*, *526*, *562*, *664*), obesity was only recognized as a core feature of the rearrangement when 1-3% of individuals suffering from severe obesity were found to carry the deletion (*325*, *426*), an association later reproduced in large deletion clinical cohorts (*500*, *538*). While feeding difficulties and failure to thrive have been reported early in life (*538*), BMI was consistently found to increase at around 4-6 years and rapidly

progresses to obesity (*426*, *538*–*540*), leading to a penetrance of 70% in adulthood (*500*). Conversely, duplication carriers were found to be at increased risk for being underweight, establishing a negative correlation between the region's copy number and BMI and demonstrating for the first time that overweight and underweight could have the same etiology (*326*, *423*). This mirror effect was replicated in UKBB, with the deletion and duplication leading to a BMI increase and decrease of 6.2 kg/m^2 and 1.8 kg/m^2, respectively (*295*). Similar findings in population cohorts have since been reported for continuous measures of adiposity such as BMI, weight, or body fat mass (*208*, *292*, *294*, *306*), as well as binary diagnosis of obesity (*293*, *306*), with the deletion exerting a stronger phenotypic effect than the duplication. Multiple studies also reported an association with waist-to-hip ratio (*208*, *292*, *294*), the latter being indicative of a shift from subcutaneous to visceral adiposity that has been linked to adverse health outcomes, even if this association is strongly attenuated upon adjustment for BMI (*208*). The abundance of evidence from clinical and population cohorts means that the dosage effect on BMI is one of the most striking and robust features associated with 16p11.2 BP4-5 CNVs. Mechanistically, hyperphagia is prevalent among deletion carriers, especially those suffering from obesity (*426*, *500*, *526*), and deletion carriers were found to exhibit altered satiety response preceding obesity onset (*665*), as well as structural changes in parts of the brain associated with reward mechanisms. Consistent with this observation, deletion carriers are prone to disinhibiting eating disorders leading to eating in the absence of hunger when they see others eat or are bored (*539*). These behaviors likely do not fully account for BMI increase (*539*), suggesting that other mechanisms, such as reduced energy expenditure, might be at play. Importantly, several studies have suggested that obesity is independent of the neuropsychiatric phenotypes frequently observed among CNV carriers (*423*, *426*, *500*, *665*).

Diabetes & pancreas disorders

Obesity is common among 16p11.2 BP4-5 deletion carriers and represents a major risk factor for type II diabetes (*666*). Still, prevalence of the latter has not been systematically studied in clinical cohorts of deletions carriers. Population studies have shown that deletion carriers are at increased risk for type II diabetes (*292*, *293*, *306*) and exhibit higher levels of glycated hemoglobin (*208*, *292*, *305*), which are at least partially independent of BMI (*563*). Deletions were also found to increase risk for type I diabetes (*82*) an autoimmune disease caused by insulin deficiency, as opposed to insulin resistance. While this disease is not strongly associated with obesity, this association is lost upon adjusting for BMI (*563*), suggesting that the observed association might result from early-onset type II diabetes cases misdiagnosed as type I, which often has a childhood onset (*667*). Independent reports identified two deletion carriers with neonatal hyperinsulinemic hypoglycemia (*668*, *669*), and a third presenting with hypoglycemic coma with fluctuating blood glucose levels (*640*). These cases suggest broad, early-onset insulin dysregulation as a feature of the deletion. This highlights the need for better assessment of glycemia and insulinemia in pediatric cohorts, allowing improved characterization of the type, severity, and age of onset of different forms of diabetes and the adoption of adequate treatment strategies.

Other features of the metabolic syndrome

Features of the metabolic syndrome in 16p11.2 BP4-5 CNV carriers have primarily been assessed in adult population cohorts. Besides previously described increased rates of obesity and poor glycemic control, UKBB deletion carriers are at increased risk for essential hypertension (*82*, *292*, *293*). This association is lost when accounting for BMI, revealing that deletion carriers have lower diastolic blood pressure compared to BMI-matched copy-neutral individuals (*563*). Similarly, UKBB deletion carriers have lower levels of high-density lipoprotein cholesterol and elevated triglycerides (*208*), putting them at increased risk for hyperlipidemia (*563*), even though these associations are largely driven by the increase in BMI. Despite these observations, deletion carriers were not at a significantly increased risk for ischemic heart disease, regardless if accounting for BMI or not (*82*, *563*).

Characterized by hepatic fat accumulation, onset of metabolic dysfunction-associated steatotic liver disease – previously referred to as non-alcoholic fatty liver disease – is precipitated by the metabolic syndrome and typically evolves towards hepatic steatohepatitis and fibrosis, which further contributes to the metabolic syndrome through over-secretion of triglycerides and glucose in the plasma (*670*, *671*). Elevated liver enzymes are an indicator of hepatic dysfunction and even though liver biopsy remains the diagnostic gold standard, the condition should be suspected in presence of metabolic syndrome features. Accordingly, key liver enzymes, i.e., alanine aminotransferase (ALT), aspartate aminotransferase (AST), and gamma-glutamyltransferase (GGT), negatively correlated with 16p11.2 BP4-5 dosage (*208*). Furthermore, alkaline phosphatase (ALP) (*208*, *306*) and albumin (*208*) levels were increased and decreased among deletion carriers, respectively. The effects on AST, GGT, and ALP remain significant when accounting for BMI, even though no increased risk for hepatic fibrosis was observed (*82*), possibly due to the condition being underdiagnosed (*672*). Finally, C-reactive protein, a nonspecific marker of inflammation, was increased among deletion carriers (*208*, *292*, *305*) in a BMI-dependent fashion (*563*).

Epidemiologic data on the age of onset of metabolic phenotypes, as well as estimates of prevalence and efficacy of medication and lifestyle modifications, remain scarce. This is particularly relevant as other comorbidities could alter adherence to treatment strategies, e.g., motor delays and the slower walking pace of deletion carriers (*292*) could impair capacity to exercise. Current data suggest that while a large fraction of metabolic alterations is consequential of increased BMI, some, including changes in glycemic and hepatic biomarkers, are driven by independent pathways (*563*). This parallels findings for the adjacent 16p11.2 BP2-3 deletion (*412*) (Figure 6.10), but more research is needed to decipher underlying molecular mechanisms. Interestingly, in mouse models for the CNV, the mirror effect is reversed (*552*–*554*), with animals carrying the deletion having small body size and altered basal metabolism (*673*), while animals harboring the duplication are severely overweight and exhibit hepatic steatosis, hyperlipidemia, and hyperinsulinemia (*674*). Mice with the deletion further exhibit altered brain metabolism and a reduced number of mitochondria in brain endothelial cells (*675*). Using human and zebrafish models, haploinsufficiency of the ceramide synthase modulator *FAM57B* (now *TLCD3B*; MIM: 615175) was shown to disrupt sphingolipid and glycerolipid homeostasis in the brain, leading to defects in synapto-

genesis, brain activity, and behavior (*676*). Together, these studies start to establish a link between the metabolic and neurologic phenotypes observed in CNV carriers.

Recently, long-term follow-up of two deletion carriers treated with liraglutide (glucagon-like peptide 1 analog) demonstrated effective weight loss accompanied by improved glycemia, lipidemia, and overall life quality (*677*). Offering promising perspectives, replication studies are required to establish the safety and efficacy of these therapies in deletion carriers.

Reproduction
Robust evidence shows that dosage of 16p11.2 BP4-5 correlates with age at menarche in both clinical (*328*) and population cohorts (*208*, *306*, *328*), with deletion and duplication carriers experiencing menarche on average 1.5 years earlier and later than controls, respectively. As for other metabolic phenotypes, mice models of the CNV exhibit a reversed mirror effect on female sexual maturation, with duplication and deletion carriers experiencing earlier and delayed first ovulation, respectively (*328*). While childhood obesity causally lowers age at menarche (*678*), in humans the mirror effect was robust to correction for adult BMI (*328*). A similar effect is observed on relative age at first facial hair (*208*, *328*), suggesting that puberty timing is affected in both males and females. Conversely, age at menopause and balding are not affected (*208*). Despite lowering puberty timing, an Icelandic study found that deletion carriers exhibited markedly reduced fecundity – measured as the number of children in individuals over 45 years – while no effect was observed for the duplication carriers (*435*). Males were more affected than females, an observation that has since been generalized to a broader spectrum of rare deleterious mutations with potential explanations including infertility, congenital malformations, and increased burden of neuropsychiatric disorders and other health outcomes that make it less likely to find a partner (*679*). In support of the former, sex hormone binding globulin levels, which regulate the amount of bioavailable testosterone, were reduced in UKBB deletion carriers (*82*), even though this association was driven by increased BMI (*563*). Susceptibility for congenital malformations of the genital system, discussed later in this review, along with the high prevalence of neuropsychiatric conditions, discussed earlier, could additionally contribute to the reduced fecundity and further research should disentangle the contribution of these factors.

Cardiac

Case reports have identified multiple congenital heart defects among 16p11.2 BP4-5 CNV carriers (*524*, *562*, *680*–*690*). Within a study of 1,118 fetuses with congenital heart defects, 16p11.2 BP4-5 deletions were the second most common chromosomal alteration found in 0.9% of cases (*690*). Penetrance of congenital heart defects among deletion carriers is low, with estimates consistently ranging between 5-10% (*423*, *500*, *691*) and it is unclear if dosage of the interval is causal for these anomalies. Arguing in favor of a causal role, mouse models for the deletion present with subtle heterogenous alterations in cardiac structure and function (*692*). Furthermore, over 5% of BioVU CNV carriers had cardiac findings in their electronic health records, with enrichment for “cardiac

dysrhythmias" among deletion carriers, and various congenital anomalies of the heart, cardiomegaly, and cardiac interventions (e.g., "heart transplant/surgery") among duplication carriers (413). In the UKBB increased risk for cardiac valve disorders and arrhythmias were observed among deletion carriers (82), but associations were lost when adjusting for BMI (563). Hence, congenital heart anomalies represent a rare but consequential feature of the 16p11.2 BP4-5 rearrangement with milder defects potentially contributing to cardiac diseases in adulthood.

Pulmonary

Thorough investigation of pulmonary phenotypes is lacking in clinical cohorts, despite isolated reports of infancy or childhood onset of asthma (421, 640, 687). In UKBB, deletion carriers have strongly reduced pulmonary function (208, 292, 294, 306), as well as increased risk for asthma (82, 293), chronic obstructive pulmonary disease (COPD) (82, 306), and respiratory failure (292). Similarly, BioVU CNV carriers frequently presented with "abnormal findings during examination of lungs" (413). Importantly, while asthma risk was driven by an increase in BMI, a well-known risk factor for the disease, this was not the case for COPD and forced vital capacity (563). Recurrent pulmonary infections (see *Hematological & Immune system*) and environmental factors such as smoking, air pollution, occupational or residential exposure to allergens, chemicals, dust, fumes, or molds represent major risk factors for lung diseases. Except for tobacco smoking, whose rates were found to be increased among UKBB CNV carriers (292), very little is known about whether 16p11.2 CNV carriers are differentially exposed to such factors and how these exposures affect expressivity of the rearrangement.

Musculoskeletal & connective tissue

Global musculoskeletal features

16p11.2 BP4-5 CNV carriers present with global musculoskeletal alterations. Unlike the mirror effect on BMI, the effect of the region's dosage on height is more subtle, with shorter stature in deletion carriers reported in both clinical (500) and population (208, 292, 294, 306) cohorts but only a fraction of population studies reporting a taller stature in duplication carriers (294, 306). As a consequence of increased BMI (563), adult levels of insulin-like growth factor 1 (IGF-1), which mediates the effect of growth hormone, are decreased in UKBB deletion carriers (208, 292). Future studies should establish whether decreased levels of IGF-1 are already present during childhood, as this could explain the short stature of deletion carriers. Bone composition is also affected, with the region's dosage negatively correlating with heel bone mineral density (208, 294, 306). Even though obesity correlates with high bone mineral density (693) the dosage effect is robust to BMI correction. The observed increased risk for arthrosis among UKBB deletion carriers (82) appears to be caused by excess adiposity (82), even though other mechanisms, such as structural anomalies of the joints, cannot be excluded. Indeed, several reports of joint hypermobility among clinically ascertained CNV carriers have been described (421, 423, 523, 524). Joint laxity – along with short stature, limb malalignment, and spinal deformity – is a hallmark feature of spondyloepimetaphyseal dysplasia with joint laxity type 2 (MIM: 603546), and autosomal dominant disorder caused by mutations in the 16p11.2 gene

KIF22 (MIM: 603213) that often leads to early-onset arthrosis (*694*). To date, prevalence of this disorder among deletion carriers has not been assessed. UKBB CNV carriers also exhibited a strong decrease in hand grip strength (*208*, *294*), paralleled by high rates of "muscle weakness" in BioVU deletion carriers (*413*). While decreased muscle strength could not be explained by increased BMI and shorter stature, possible mechanisms might include low IGF-1 levels, reduced physical activity, or neurological defects leading to hypotonia and muscle weakness.

Craniofacial features

The mirror effect on head circumference – making deletion and duplication carriers more prone to macro- and microcephaly, respectively – represents one of the first described hallmarks of the 16p11.2 BP4-5 rearrangement (*421*, *423*, *425*, *498*, *500*) and was later paralleled by changes in brain volume (*586*–*589*). Importantly, head circumference positively correlates with BMI, with about one third of obese deletion carriers being macrocephalic (*423*, *500*). Mechanistically, modulating expression of the ortholog of the 16p11.2 gene *KCTD13* (MIM: 608947) in zebrafish was found to recapitulate the head size phenotype through perturbation of RhoA signaling (*266*, *273*, *695*, *696*). *kctd13* expression negatively correlated with proliferation of neuronal progenitor as well as increasing apoptosis upon overexpression (*266*). While *kctd13* was sufficient to establish the neuroanatomical changes, expressivity was increased by simultaneously altering the expression of two other genes in the region, *mvp* and *mapk* (*266*), suggesting *cis*-epistatic interactions. Concordantly, dysregulation of the ERK signaling cascade – of which MAPK3 is part – was suggested to play a role in the increase in progenitor proliferation and decrease in hippocampal synaptic protein synthesis in a mouse model of the deletion (*697*, *698*). Increased dendritic arborization in a duplication mouse model was linked to the same kinase cascade (*699*). Another study investigating global craniofacial features in 16p11.2 BP4-5 CNV carriers found that individual overexpression of seven 16p11.2 genes in zebrafish could induce an analogous phenotype to the lower jaw protrusion phenotype observed in human duplication carriers (*272*). Simultaneous overexpression of *kctd13*, *mapk3*, and *mvp* yielded an even stronger jaw protrusion phenotype, despite no effect of individual gene overexpression (*272*). In addition to jaw protrusion, a positive and negative dosage effect on the nasal and frontal regions, respectively, were identified from 3D morphometric facial imaging, even if these effects remain small and variable (*272*). These findings align with the frequently reported facial features of CNV carriers - broad forehead, micrognathias, or flattened profile – despite no recognizable facial gestalt (*421*, *500*, *523*, *524*, *687*). Dysmorphic features can result from skull deformities, such as **craniosynostosis**[24], which has been reported in deletion carriers (*500*, *562*, *600*), with a prevalence estimate of 1.3% (*423*). Syndromic and multisuture craniosynostosis can also lead to Chiari type 1 malformations, which are frequent among deletion carriers (*423*, *498*, *500*, *594*–*596*) (see *Structural alterations of the nervous system*). In rarer cases, more severe malformations of the **posterior fossa**[25] have been reported (*700*, *701*). To conclude, 16p11.2 BP4-5 dosage negatively correlates with head circumference and predisposes to a broad range of usually mild dysmorphic features and cranial anomalies. The latter have low penetrance, especially among duplication carriers and non-medically ascertained deletion carriers (*500*).

24: Rare birth defect characterized by premature fusion of skull bones that can affect brain development.

25: Small cavity in the skull in which the cerebellum and part of the brain stem are located. Malformations typically affect cerebellum development and are classified depending on whether the fossa is enlarged (e.g., Dandy-Walker malformation) or too small (e.g., rhombencephalosynapsis).

Spine and thoracic cage deformities

Deformities of the spine and thoracic cage, such as pectus excavatum (*421*, *423*, *500*, *702*, *703*), or idiopathic scoliosis and vertebral anomalies (*500*, *523*, *681*, *703*, *704*), represent recurrent phenotypes of the 16p11.2 BP4-5 deletion. Interestingly, carriers of the deletion or a loss-of-function variant in the 16p11.2 *TBX6* (MIM: 602427) gene in combination with a **hypomorphic**[26] *TBX6* allele explained up to 11% of congenital scoliosis cases in a Chinese population (*142*). Highlighting *TBX6* as the causal gene for spinal malformations, these results were replicated in additional cohorts (*598*, *705*). Further research showed that *TBX6*-associated congenital scoliosis presents with distinguishable endophenotypes including earlier onset, increased prevalence of hemivertebrae and rib anomalies, and lower rates of spinal cord defect (*706*). *TBX6* **compound inheritance**[27] was also shown to be associated with a broad spectrum of disorders of vertebral development and segmentation – ranging in severity from scoliosis or **kyphosis**[28] to generalized defects such as **spondylocostal dysostosis**[29] (MIM: 122600) (*683*) – as well as a cooccurrence of structural defects of the vertebra, ribs, and kidney (*275*), in line with the role of *TBX6* in development of the genitourinary tract (*277*, *542*) (see *Congenital anomalies of the genitourinary tract*). Effects of increased dosage of *TBX6* are less well defined, although duplication carriers have been reported to suffer from congenital vertebral malformations (*421*, *664*, *707*). Interestingly, duplication carriers tend to be more affected by upper spine (i.e., cervical vertebra) defects (*707*), in contrast with the higher predisposition to lower spine defects (i.e., thoracic and lumbar vertebra) in deletion carriers (*276*, *706*). Note that *KIF22*-associated spondyloepimetaphyseal dysplasia with joint laxity type 2 is also characterized by spinal deformities (*694*), so that additive or epistatic interactions between 16p11.2 genes could contribute to heterogeneity in skeletal phenotypes. While we did not identify an association with scoliosis in UKBB (unpublished data), other reported increased risk for sciatica among duplication carriers (*293*), while BioVU deletion carriers had higher diagnostic rates of "congenital musculoskeletal deformities of the spine" (*413*).

Hernia

Increased risk for abdominal hernia among UKBB 16p11.2 BP4-5 CNV has been reported (*292*), even though others did not find the effect to be significant (*82*, *293*). Inguinal and umbilical hernias account for about 85% of repaired abdominal hernias (*708*) and have at times been reported among clinically ascertained 16p11.2 BP4-5 CNV carriers (*423*, *562*, *687*, *704*). More striking are the multiple reports of the much more rare and severe congenital diaphragmatic hernias (*421*, *423*, *524*, *664*). Across two studies totaling 120 congenital diaphragmatic hernia cases, 2.5% were deletion carriers (*709*, *710*). Further research is needed to elucidate what predisposes CNV carriers to different types of congenital and acquired hernia, with possible explanations including weakness of the connective tissue, increased pressure on abdominal organs due to spinal and thoracic cage deformities, **cryptorchidism**[30], or obesity.

Genitourinary

Congenital anomalies of the genitourinary tract

With a 6.3% fraction of the cases, 16p11.2 BP4-5 deletions were found to be enriched in a **Müllerian aplasia**[31] (MIM: 277000; 601076) cohort

26: Partial loss-of-function allele that results in reduced production, function, or stability of the wildtype allele.

27: A genetic variant is compounded when another variant impacting the function of the same gene is present on the other allele. Compound heterozygotes will carry two different mutations, one on each allele, which can result in a recessive disorder.

28: Abnormal forward rounding of the spine ("hunchback"), in opposition to scoliosis, which is defined by an abnormal sideways curvature of the spine.

29: Rare disorder characterized by severe, congenital deformities of the spine and ribs that cause short-trunk dwarfism. Deformities increase risk of breathing problems, hernia, spina bifida, and Chiari malformations.

30: Relatively common male congenital birth defect in premature infants characterized by at least one testis not fully descended into the scrotum.

31: Mayer-Rokitansky-Küster-Hauser syndrome or Müllerian aplasia is a rare congenital defect of the female reproductive system characterized by aplasia of the uterus, cervix, and vagina, leading to infertility. It can be accompanied by malformations of the Fallopian tubes, ovaries, urinary tract, and spine, in which case it is referred to as Müllerian-renal-cervicothoracic somite dysplasia.

(*278*), while haploinsufficiency of *TBX6* – through gene deletion or point mutations – was identified in 23 out 112 cases of Müllerian aplasia (*711*), as well as in one patient with distal vaginal atresia but normally developed uterus and cervix (*712*). The role of the deletion was consolidated by two studies in individuals of Chinese ancestry, where it accounted for 0.9-1.4% of the cases (*713*, *714*). These findings echo the increased rate of female reproductive tract disorders observed among Estonian Biobank CNV carriers, which have been proposed to be driven by dosage of *ASPHD1* and *KCTD13* based on Mendelian randomization and single-gene dosage modulation in zebrafish (*328*). Anomalies of the male genitalia, including cryptorchidism, **hypospadias**[32], and micropenis, have also been reported in 16p11.2 BP4-5 deletion (*421*, *523*, *664*) and less frequently in duplication (*326*, *421*) carriers. After identifying an enrichment of 16p11.2 BP4-5 overlapping deletions among individuals with genitourinary defects, two studies used mice models to show that decreased dosage of *KCTD13* associated with penile and testicular anomalies (*715*), while reduced dosage of *MAZ* (MIM: 600999) led to defects of the upper genitourinary tract and high penetrance of congenital anomalies of the kidney and urinary tract (CAKUT) (*716*).

32: Male congenital birth defect causing the opening of the urethra not to be located at the tip of the penis.

CAKUT describes a broad spectrum of phenotypes, including kidney anomalies, ectopic or horseshoe kidneys, obstructive uropathies, and **vesicoureteral reflux**[33]. It was reported in a small fraction of early descriptive studies of 16p11.2 BP4-5 CNV carriers (*421*, *524*). Echoing the finding that 0.5% of fetuses with ultrasound renal anomalies harbored a 16p11.2 BP4-5 deletion (*717*), deletion carriers were enriched in a cohort of 2,800 CAKUT cases (*277*). Unlike other recurrent CNVs, 16p11.2 BP4-5 deletions induced a broad spectrum of genitourinary defects (*277*, *541*). Using a genotype-first approach, another study found that 13 of 52 deletion carriers presented with defects of the urinary tract (*542*), establishing the deletion as an important risk factor for CAKUT. Frequent co-occurrence of skeletal and genitourinary malformations (*275*) has led to the hypothesis that haploinsufficiency of *TBX6* is at the origin of both phenotypes. Concordantly, mouse models with various degrees of reduced *Tbx6* expression exhibited CAKUT phenotypes (*277*, *542*). This implicates *TBX6* dosage as the driver of both skeletal and genitourinary phenotypes, with dosage of *cis*-genes (i.e., *KIF22*, *KCTD13*, *MAZ*, and *ASPHD1*) likely contributing to phenotypic variability.

33: Abnormal flow of urine from the bladder back up the ureters towards the kidneys, which increases infection risk and can cause renal damage.

Renal function

16p11.2 BP4-5 deletion carriers were enriched in a cohort of 6,679 chronic kidney disease cases (*409*). Paralleling these findings, UKBB CNV carriers, and in particular deletion carriers, had increased levels of the renal biomarker cystatin C (*208*, *292*, *305*) and were at increased risk for both chronic kidney disease and acute kidney injury (*82*, *293*, *306*). The region's dosage was also found to positively correlate with serum creatinine levels (*208*, *306*). While impaired kidney function is typically associated with high levels of creatinine, we cannot exclude that muscle wasting or liver diseases, both observed in deletion carriers, play a role in the lowered creatinine levels. Importantly, the U-shape effect on cystatin C and acute kidney injury and the mirror effect on creatinine were robust to BMI adjustment, putting forward the hypothesis that other mechanisms, such as presence of subclinical structural renal alterations, affect renal function in the long term. Deletion carriers also exhibited high levels of serum

uric acid (208). Hyperuricemia is linked to the metabolic syndrome and represents a strong risk factor for gout, whose prevalence was increased, albeit not significantly, among deletion carriers (82). Overall, there is clear evidence for impairment of renal function among adult CNV carriers.

Hematological & Immune system

16p11.2 BP4-5 deletion carriers are at increased risk for anemia, and in particular iron deficiency anemia (40, 82, 293). Anemia risk was associated with the number of copies of *BOLA2*, a gene involved in cellular iron homeostasis mapping within the 16p11.2 BP4-5 flanking breakpoints (Figure 6.10) and present in 3 to 8 copies in humans (40, 41). These *Homo sapiens*-specific copy number polymorphic duplications of *BOLA2* are under positive selection and were suggested to provide an adaptive role in protecting against iron deficiency (40, 41). Other findings related to red blood cells include increased mean reticulocyte volume and decreased high light scatter reticulocyte count in deletion carriers, as well as increased immature reticulocyte fraction in duplication carriers (306). There is also evidence that the myeloid and lymphoid lineages are compromised in deletion carriers. Severe combined immunodeficiency 8 (MIM: 615401) with T lymphocytopenia has been reported in deletion carriers compounded with mutations in the 16p11.2 T-cell mediated immunity gene *CORO1A* (MIM: 605000) (543, 718), while immune deficiency was suspected in three independent deletion carriers due to severe pneumonia or low immunoglobulins (562, 640, 687). Retrospective analysis of 170 deletion carriers ascertained for ASD revealed that 81% had a history of significant infection, including recurrent otitis (28%), chronic bronchitis (4%), or pneumonia (26%) (719). Low lymphocyte levels have been reported in UKBB deletion carriers (306, 720), along with an increased risk for pneumonia (82, 306), with recent evidence suggesting that the 16p-associated impaired immunity is not secondary to increased adiposity (563). More controversial are findings related to neutrophil count, which were increased in UKBB deletion carriers (208, 306, 720) despite cases of neutropenia in clinical cohorts (720) or platelet count, which in UKBB show a strong, BMI-independent, negative correlation with the region's dosage (208, 306) despite a case report of a deletion carrier with thrombocytopenia (687). Recently, the immunity gene *CORO1A* has also been shown to play an important role in platelet biology (718, 721), while the encoded protein is part of the AP2-mediated clathrin-coated pit subcomplex within the atlas of autism protein interactions (722). While some of these associations remain to be clarified, current evidence indicates that the 16p11.2 BP4-5 deletion represents an important risk factor for both anemia and recurrent infection.

Sensory organs

Overall sensory processing is affected in 16p11.2 BP4-5 CNV carriers (641, 723, 724). Ophthalmologic findings have sporadically been documented in 16p11.2 BP4-5 CNV carriers and their frequency in 43 deletion carriers was recently reviewed, highlighting abnormal palpebral fissures (41.9%), deep-set eyes (20.9%), ptosis (18.6%), and hypertelorism (18.6%) as the most common findings (596). Strabismus and refractive errors were reported in 6-8% of CNV carriers (423, 524, 681), while 11% and 30% of

deletion and duplication carriers were diagnosed with abnormal eye convergence, respectively (*498*). More severe defects, such as microphthalmia and optic nerve **coloboma**[34] have been described (*702*), and prevalence of major ophthalmic malformations or blindness was estimated at 2.6% and 1.5% among deletion and duplication carriers, respectively, while it is estimated at 0.5% in the general population (*423*). Unfortunately, comprehensive ophthalmologic examination of CNV carriers is often lacking. In the UKBB, no significant increase in cataract, glaucoma, and cornea disorders was observed (*82*). Auditory dysfunction was reported in about 9.5% of deletion (*421*, *500*, *524*, *562*) and 3.7% of duplication (*423*) carriers. However, the same caveats apply as for ophthalmologic findings, and a more detailed and systematic characterization of auditory dysfunctions is required. Future studies should investigate associations with UKBB refractometer, intraocular pressure, and visual acuity measurements, as well as hearing test results to assess presence of subtle ophthalmologic and/or auditory defects and establish whether CNV carriers are at increased risk for sensory impairment.

34: Rare congenital ophthalmic defect characterized by the absence of eye tissue (e.g., iris, retina, or optic nerve).

Cancer

Tumors and cancers have rarely been reported in 16p11.2 BP4-5 CNV carriers (*423*, *725*) and no significant associations were reported in general populational cohorts, even if cancers were not assessed as thoroughly as other phenotypes. The only robustly associated cancer is neuroblastoma, with 22 deletion carriers identified out of 5,585 cases (*726*). None of the neuroblastomas among deletion carriers harbored a *MYC* (MIM: 190080) amplification, so that two thirds were anticipated to have a low risk of relapse (*726*), suggesting that the deletion is associated with less aggressive neuroblastoma.

Embrace diversity to better understand phenotypic heterogeneity

Over the years, numerous studies have demonstrated the extensive pleiotropy of 16p11.2 BP4-5 CNVs, establishing the rearrangement as an important susceptibility locus for a wide range of disorders. As such, diagnostic identification of the CNV is typically disclosed to patients. Yet, in 90,000 patients from the Geisinger MyCode Community Health Initiative health system, less than 10% of carriers of a CNV associated with a genomic disorder had received a clinical diagnosis, despite exhibiting clinical features associated with the condition (*402*). From a personalized medicine perspective, this emphasizes the importance of adopting a holistic approach, allowing diagnosis of individuals with milder and/or atypical presentation, as well as follow-up by multiple specialists to anticipate and potentially treat and/or prevent future complications. This is particularly relevant, given the highly heterogeneous clinical manifestation of the CNV (Figure 6.14). To achieve this, it is imperative to i) define and bring awareness to clinicians of the spectrum of possible manifestations of the rearrangement, ii) understand factors contributing to phenotypic heterogeneity, and iii) gain mechanistic insights into the molecular pathways connecting altered gene dosage to phenotype. Here, we discuss some key areas that might allow filling missing knowledge gaps, while emphasizing how diversity in patients, experimental models,

and analytical approaches can catalyze discoveries.

Diversity in ascertainment and demographics

As extensively described in this review, results from clinical and population studies often converge onto similar physiological systems. Both types of studies suffer from ascertainment biases, leading to over- and under-estimation of the CNV's effect, respectively, while still exposing the two extremities of the same phenotypic continuum. The latter ranges from subtle subclinical alterations – as often seen in transmitting parents of clinically ascertained carriers identified by cascade testing (*287*, *727*) or carriers from population cohorts – to severe medical conditions observed in probands from clinical cohorts. Hence, results of clinical and population cohorts should be seen as complementary approaches investigating the same question but from a different angle.

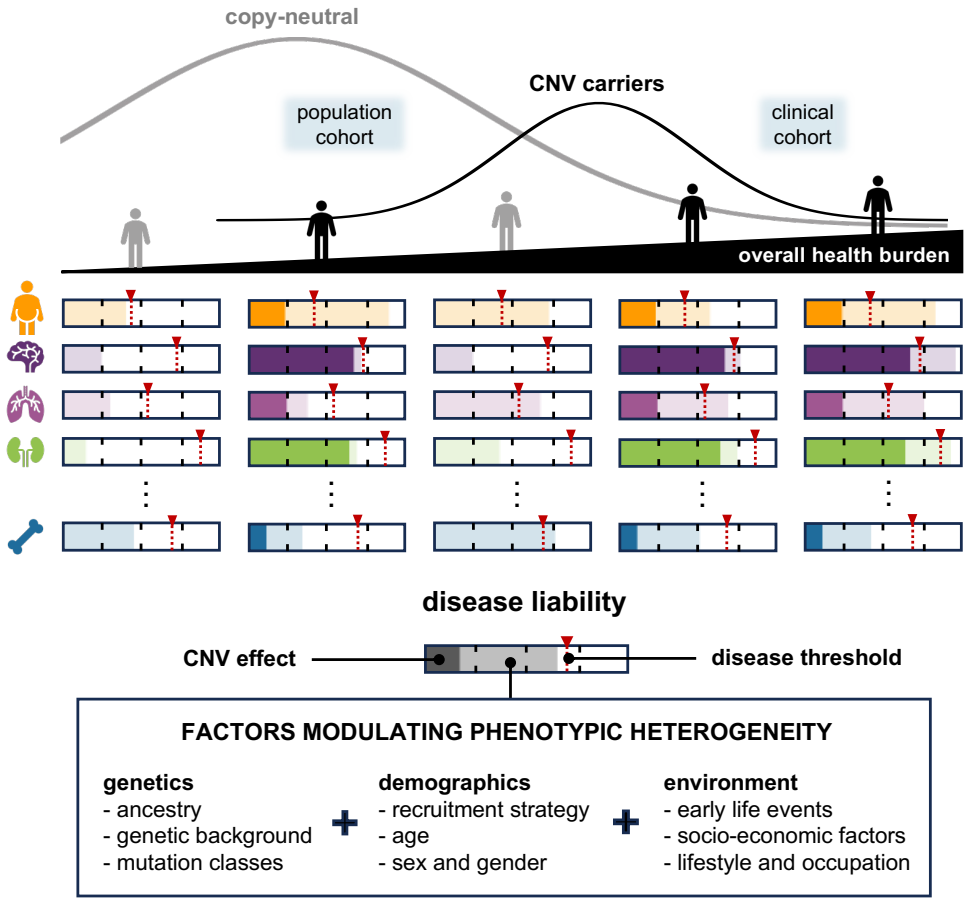


Figure 6.14: Model of phenotypic variability among CNV carriers.
Schematic view on a holistic approach to understanding phenotypic heterogeneity. Top: Distribution of the global health burden among copy-neutral (gray) and CNV carriers (black). CNV carriers from population cohorts tend to be sampled from the left side of the CNV carrier distribution, while CNV carriers from clinical cohorts tend to be sampled from the right side of that distribution. Bottom: Liability to diseases affecting different physiological systems for five individuals sampled from the above distributions. The red mark represents the liability threshold that needs to be exceeded for an individual to be diagnosed with a disease. The threshold is lower for common diseases and individuals that are near the threshold might present with subclinical features, e.g., the first individual is overweight without meeting diagnostic criteria for obesity (orange). The dark-colored area represents the true contribution of the CNV to disease liability, which in the absence of epistasis or gene-environment interactions is constant across CNV carriers but variable across diseases. Typically, contribution is stronger for rarer disorders, but usually not sufficient to pass the disease threshold. The light-colored area represents the contribution of various other genetic, demographic, and environmental factors to disease liability, which will determine whether the individual reaches the disease threshold or not. Importantly, contribution of these factors is variable across both diseases and individuals, resulting in phenotypic heterogeneity across CNV carriers.

Differences in ascertainment also mean that clinical and population cohorts have different demographics. Clinical cohorts tend to be enriched for severe pediatric cases, while population cohorts are usually composed of adults. The latter often provide longitudinal follow-up data, which offers the opportunity to investigate age at disease onset and clinical trajectory in adulthood, especially for phenotypes expressed only later in life, which are often less well characterized among carriers of syndromic CNVs. Indeed, not only do 16p11.2 BP4-5 CNV carriers suffer from increased risk for a broad range of common diseases, but they also suffer from earlier age of onset, compared to individuals lacking the CNV (*82*). This information can be tapped by physicians to establish preventive measures to anticipate and attenuate later-onset comorbidities. For instance, given the important neurodevelopmental phenotypes associated with 16p11.2 BP4-5 CNVs, we should assess whether carriers also have altered risk for neurodegenerative disorders at older ages.

Another important consideration relates to sex. Because clinical cohorts are typically recruited with a phenotype-first approach and as many traits exhibited skewed male-female ratios, this could impact sex representation and lead to biases in the clinical description of comorbidities. For instance, ASD, a hallmark feature of the 16p11.2 BP4-5 rearrangement, has an estimated male-to-female ratio of 3:1 (*728*). While true differences in disease prevalence, behavioral symptoms, and neurobiological profiles between sexes exist (*729*), evidence suggests that we underdiagnose ASD among females due to differences in clinical presentation and/or societal stereotypes (*728*, *730*). Furthermore, factors such as comorbidities and genetic etiology impact sex ratio estimates (*731*). Interestingly, the sex ratio across 16p11.2 BP4-5 deletion carriers appears stable – about 1.5 male carriers per female carrier – across different ascertainment strategies, whereas for the duplication, there are about twice as many male carriers in clinically ascertained cohorts, compared to an almost equal sex ratio in population cohorts (Table 6.7). Surprisingly, and unlike what would be expected from a female-protective effect (*312*), UKBB is significantly depleted from female deletion carriers (Table 6.7). One explanation for this could be sex-specific differences in participation compounded over multiple traits affected by the deletion, as suggested by the widespread genetic correlation between sex and adipose or psychiatric traits (*732*). This would mean that females with hallmark features of the deletion, such as increased BMI and decreased cognitive ability, might be less likely to participate. In line with this hypothesis, the BMI-increasing *FTO* (MIM: 610966) allele was present at a higher frequency in male UKBB participants (*732*), suggesting that obese females are less likely to enroll in biobanks. Differences in prevalence across sexes might reflect the existence of **genetic interaction**[35] with sex. Little is known about single-sex or sex differential effects of 16p11.2 BP4-5 rearrangements. Rodents deleted for the 16p11.2 BP4-5 syntenic region show broad sex- and age-specific behavioral differences (*556*, *733*), as well as male-specific sleep (*643*), reward-learning (*734*), neovascularization (*735*), and vocal communication (*736*) impairments, while females exhibit increased levels of anxiety (*737*). Similarly, the social behavior and the reaction to novel objects were more affected in 16p11.2 male rat models (*556*). In humans, a significantly stronger reduction in fecundity was observed in male deletion carriers (*435*), while another study found that female CNV carriers ascertained for DD/ID experienced a larger number of comorbidities

35: Genetic interactions indicate that the relation between a genotype and a phenotype will depend on another factor, such as an individual's sex, environmental exposures, or other genetic variants. The latter is referred to as epistasis, in opposition to additive genetic effects.

Table 6.7: Sex ratio among 16p11.2 BP4-5 CNV carriers with different ascertainments.
Number and percentage of male and female 16p11.2 BP4-5 deletion and duplication carriers for two clinical cohorts ascertained for i) autism spectrum disorder (ASD) and ii) developmental delay/intellectual disability (DD/ID), as well as the general population cohort UK Biobank (UKBB). Sex distribution of each cohort is indicated as a third row with gray background. Sex ratio is provided as the number of males per female in the considered sample. P-values of two-sided Fisher tests are reported, assessing differences in sex ratio among CNV carriers, compared to the entire cohort. Significant results ($p \leq 0.05$) are in bold.

Ascertainment	16p11.2BP4-5 status	Male (%)	Female (%)	Ratio	P
ASD (*731*)	deletion	9 (56%)	7 (44%)	1.3:1	0.061
	duplication	6 (67%)	3 (33%)	2:1	0.420
	cohort	4,588 (78%)	1,284 (22%)	3.6:1	
DD/ID (*731*)	deletion	45 (61%)	29 (39%)	1.6:1	0.910
	duplication	29 (64%)	16 (36%)	1.8:1	0.547
	cohort	17,061 (60%)	11,492 (40%)	1.5:1	
Population cohort (UKBB) (*563*)	deletion	45 (62%)	28 (38%)	1.6:1	**0.009**
	duplication	41 (46%)	48 (54%)	0.9:1	1
	cohort	152,967 (46%)	178,555 (54%)	0.9:1	

(*731*). In the future, sex differences should be investigated for a broad range of traits associated with 16p11.2 BP4-5 rearrangements.

Diversity in genetic background

There is a significant correlation between a CNV carrier's cognitive and social skills and those of non-carrier first-degree relatives, indicating that *familial background* modulates phenotypic expressivity with similar effects in deletion and duplication carriers (*500*, *575*, *576*). An individual's familial background encompasses many genetic variants that can be grouped depending on their frequency and phenotypic impact. Early studies hypothesized that additional rare variants sensitize genomes, leading to differential phenotypic expression of the same 16p11.2 BP4-5 rearrangement (*285*). Validating this "two-hit" theory, 16% of duplication and 8% of deletion carriers were found to harbor a second large CNV and these individuals exhibited a more severe and diverse phenotype (*286*). A later study found that 70% of clinically ascertained 16p11.2 BP4-5 CNV carriers harbored a rare secondary CNV and that there was a strong maternal transmission bias for pathogenic secondary deletions (*193*). Similarly, the number of secondary rare and predicted-to-be pathogenic variants was shown to be negatively correlated with cognitive function and head circumference in 16p11.2 BP4-5 deletion carriers (*287*). In some cases, the second hit is linked to known genetic disorders, such as severe combined immunodeficiency (*543*, *718*), Cohen syndrome (MIM: 216550) (*738*), Mowat-Wilson syndrome (MIM: 235730) (*739*), Zellweger spectrum disorders (MIM: 614862) (*669*), or Friedreich ataxia (MIM: 229300) (*740*), leading to more severe cases with atypical presentation, highlighting dual diagnosis as a possible explanation for phenotypic heterogeneity (*741*). While syndrome coexistence or bi-parental inheritance of duplication (*550*, *551*) might occur at random, the phenomenon might be fostered by cross-disorder assortative mating (*727*). Only a few studies (*193*, *727*) have investigated the interplay between assortative mating, CNV inheritance mode, and parent-of-origin effects and future work should aim at characterizing the interaction between these processes to better determine how phenotype severity and heterogeneity is compounded over multiple generations.

The "two-hit" model is paralleled by experimental findings in various model organisms. Pairwise gene knockdown experiments (Box 1) revealed intraregional epistatic interactions. Interestingly, a mouse model hemi-deleted for three genes (*Taok2*, *Sez6l2*, and *Mvp*) recapitulates the male-specific behavioral alterations observed in 16p11.2$^{Del/+}$ mice, while the additional hemi-deletion of *Mapk3* decreased phenotypic similarities (*561*). Yet, another mouse model hemi-deleted for *Mapk3* and *Mvp* leads to altered behavior in male mice (*274*). This suggests that phenotypes linked to the CNVs can be recapitulated through perturbation of various gene combinations, implying redundancy. Combinatorial knockdown and overexpression experiments, as well as transcriptome-wide studies of the gene expression dysregulation induced by the rearrangement also revealed widespread interactions between 16p11.2 BP4-5 homologs and other DD/ID, genomic disorder, and ciliopathy genes (*547*, *559*). Long-range interactions involve homologs of the distal 16p11.2 BP2-3 CNV region (*559*) (Figure 6.10), whose rearrangement in humans generates similar phenotypes to BP4-5 rearrangements, including increased risk for ASD and a mirror effect on BMI and head circumference (*279*). Specifically, mouse and zebrafish studies found that the ortholog of the BP2-3 *LAT* (MIM: 602354) gene acted in concert with *KCTD13* – the major BP4-5 driver of head circumference (*266*) – to modulate brain size, with additional contributions of *MVP* (MIM: 605088) and *MAPK3* (MIM: 601795) (*273*, *280*). Chromatin conformation assays have further demonstrated high levels of evolutionary conserved interaction between BP2-3 and BP4-5 (*279*, *742*), as well as the entire short arm of chromosome 16 (16p) (*282*), suggesting that CNVs in the region lead to broad disruption of local 3D genomic structure. This could explain why 16p11.2 BP4-5 deletions induce global downregulation of neuronally expressed 16p genes (*282*). This observation is exemplified by the downregulation of genes linked to DD/ID and psychiatric conditions in human cortical organoids derived from deletion carriers, including the 16p mRNA splicing regulator *RBFOX1* (MIM: 605104) (*743*). Intriguingly, increased local 16p **polygenic score (PGS)**[36] for ASD exerted a similar impact on gene expression (*282*), reconciling the rare and common component of an individual's familial background.

36: Quantity reflecting the contribution of a group of variants to a given phenotype in a single individual. Most often, PGS capture the additive effects of thousands of genome-wide single polymorphisms, even though they can also be built to account for other mutation types or be restricted to specific genomic regions.

More commonly, PGSs are assessed genome-wide and there is emerging evidence that the latter act additively to CNVs. For instance, 16p11.2 BP4-5 duplication carriers with a PGS predisposing to high BMI tend to exhibit a less severe reduction in BMI than those with a PGS predisposing to low BMI, with opposite trends observed among deletion carriers (*289*). Another study showed that SCZ cases carrying an SCZ-associated CNV had lower SCZ PGS than those that did not (*290*). Because PGS reduction was proportional to the CNV's effect on SCZ and the 16p11.2 BP4-5 duplication substantially contributes to SCZ risk, SCZ PGS was not a significant predictor among duplication carriers (*290*). These studies suggest that expression of 16p11.2 BP4-5 rearrangements is modulated by multiple common variants with minute effects, extending the "two-hit" model to a polygenic one. Studying the contribution of the polygenic background is complicated by healthy volunteer bias and assortative mating. While more work is required to gain an understanding of how different mutations act in concert to determine an individual's disease liability, these studies highlight the importance of evaluating the genomic context in which CNVs occur if we ought to understand phenotypic

expressivity.

Diversity in ancestry

An important source of diversity stems from genetic ancestry. A first key question is whether the frequency of the 16p11.2 BP4-5 rearrangements varies across ancestries. Deleterious CNVs were found to be less prevalent in UKBB samples of non-European ancestry (*744*). This could be explained by some haplotypes, e.g., at cytobands 17q21.31 and 16p12.1, favoring genetic rearrangements due to the size and/or orientation of encompassed segmental duplication blocks (*285*). This in turn could make some populations more susceptible to *de novo* CNVs (*285*) but does not seem to be the case for 16p11.2 BP4-5 (*41*), despite archaic introgression in this cytoband in some populations (*745*). Accordingly, neither the ASD Simons Foundation Powering Autism Research for Knowledge cohort (SPARK; N = 58,419; 20% non-European) (*744*) nor the healthcare cohort Bio*Me* (N = 24,877; 68% non-European) (*530*) identified a significant divergence in 16p11.2 BP4-5 CNV prevalence across ancestries, even though estimates are limited by the relatively small sample size of each ancestry group. Alternatively, differences in frequency of other mutations might modulate the CNV expressivity, leading to certain phenotypes being more frequently expressed in specific populations. For instance, autosomal recessive phenotypes might be more frequent in carriers from a population in which loss-of-function alleles are widespread, as illustrated by the high prevalence of congenital scoliosis in deletion carriers of Asian ancestry due to the high prevalence (44%) of a *TBX6* hypomorphic haplotype, which is rarer in individuals of European (33%) and African (<1%) ancestries (*276*). Similarly, compounding the deletion with a haplotype associated with reduced expression of MAPK3 affects early neuronal development (*746*). Conversely, one could expect phenotypes to become apparent in populations of duplication carriers in which relevant **hypermorphic**[37] alleles are widespread, even though no such example has been reported for the 16p11.2 BP4-5 rearrangement to our knowledge.

37: Allele with mutation that increases production, function, or stability of the wild type gene product.

A major limitation is that except for a few large studies in individuals of Asian ancestry, the bulk of current knowledge stems from investigating CNV carriers of European ancestry, specifically the Simons Variation in Individuals Project (Simons VIP) (*644*) and the 16p11.2 European Consortium for clinically ascertained cohorts and the UKBB (*61*) and Estonian Biobank (*62*) for general-population biobanks. While these cohorts have pioneered the field, more diverse population cohorts have been set up in recent years (*8*, *9*, *65*, *461*), which should allow to better grasp the extent to which frequency and phenotypic expression depend on an individual's ancestral background.

Diversity in environment

Environmental exposures are potent phenotypic modulators and could account for part of the phenotypic heterogeneity observed among 16p11.2 BP4-5 CNV carriers. Yet, their role remains unexplored. So far, one study found that an increased number of perinatal events (e.g., preterm birth, abnormal presentation, low birthweight, or respiratory distress), but not prenatal events, led to an increase in ASD symptomatology among

deletion carriers (*747*). In mice modeling the duplication, adolescent exposure to the psychoactive constituent of cannabis exacerbated deficits in social memory in adulthood (*748*), while *Mapk3* knockout mice are hypersensitive to the rewarding properties of morphine (*605*). Many other exposures during childhood, adolescence, and adulthood, including diet, smoking habitat, alcohol consumption, physical activity, sleep hygiene, medication usage, exposure to pollutants, occupation, socio-economic status, and access to medical care could impact penetrance, expressivity, and age of onset of diseases associated with the rearrangements. Future studies should systematically assess whether environmental factors exacerbate (or mitigate) clinical features beyond simple additive effects, i.e., through CNV-environment interactions. For common variants, it was demonstrated that genetic effects are modulated by different environments between populations, more so than by true differences in causal effects across ancestries (*749*). Hence, it would be useful to identify environmental exposures that prompt more severe expression of certain phenotypes, as well as factors, such as early genetic diagnosis or follow-up by multidisciplinary teams, that have the potential to effectively alleviate the symptomatologic burden.

Diversity in mutation classes

A longstanding challenge relates to linking genetic content to specific phenotypic features of 16p11.2 BP4-5 rearrangement. Besides the experimental approaches described below, it is possible to leverage the existing genetic diversity at the locus to gain functional insights. For instance, larger rearrangements, e.g., between BP1 and BP5 (Figure 6.10), demonstrated the additive contribution of BP2-3 and BP4-5 to BMI and head circumference (*280*), while a smaller 118 kb deletion encompassing only *MVP*, *CDIPT* (MIM: 605893), *SEZ6L2*, *ASPHD1*, and *KCTD13* was found to segregate with ASD features over three generations (*750*). Due to the absence of segmental duplications between BP4 and BP5, reports of recurrent partial rearrangements are sparse. Alternatively, rare protein-coding variants can provide insights into the function of some genes, as exemplified for *PRRT2* and *TBX6*, the causal genes for PKD (*632*) and congenital scoliosis (*142*), respectively. While these examples have been elucidated through family studies, an alternative approach in population cohorts are **burden tests**[38] or more elaborated variance component and combination tests (e.g., optimal sequence kernel association test (SKAT-O)) (*143*). These have been performed in the UKBB for a wide spectrum of traits (*6*, *145*) but did not yield any strictly significant association for 16p11.2 BP4-5 genes, except for an association between *SLX1A* and cannabis usage (*145*). The weakness of these tests, however, is that rare variants account for only a small fraction of **heritability**[39], which is concentrated on a few, highly constrained genes (*106*).

Common variants account for a much larger fraction of trait heritability and can also be leveraged to increase confidence in the causal role of the locus. For instance, early GWASs found the BP4-5 variant rs4583255[T] to increase risk for psychosis while decreasing BMI, mimicking two hallmark phenotypes of the duplication (*752*). Since then, 290 association signals with single nucleotide variants mapping to the CNV region have been reported in the NHGRI-EBI GWAS Catalog (*78*) (Figure 6.15). Paralleling the phenotypes observed in CNV carriers, most signals relate

38: Joint analysis of multiple rare variants meeting certain criteria that are grouped into a single analysis unit, typically a gene, to perform an association study with a selected phenotype. SKAT-O is one of the most common approaches, providing a computationally efficient test that can handle scenarios wherein variants have effects in opposite directions and only a fraction of them is causal.

39: Fraction of phenotypic variance explained by genetic variance. Heritability can be calculated for specific sets of variants, such as rare vs common variants, variants mapping to a specific genetic region, or belonging to a particular mutational class, to assess their contribution to phenotypes.

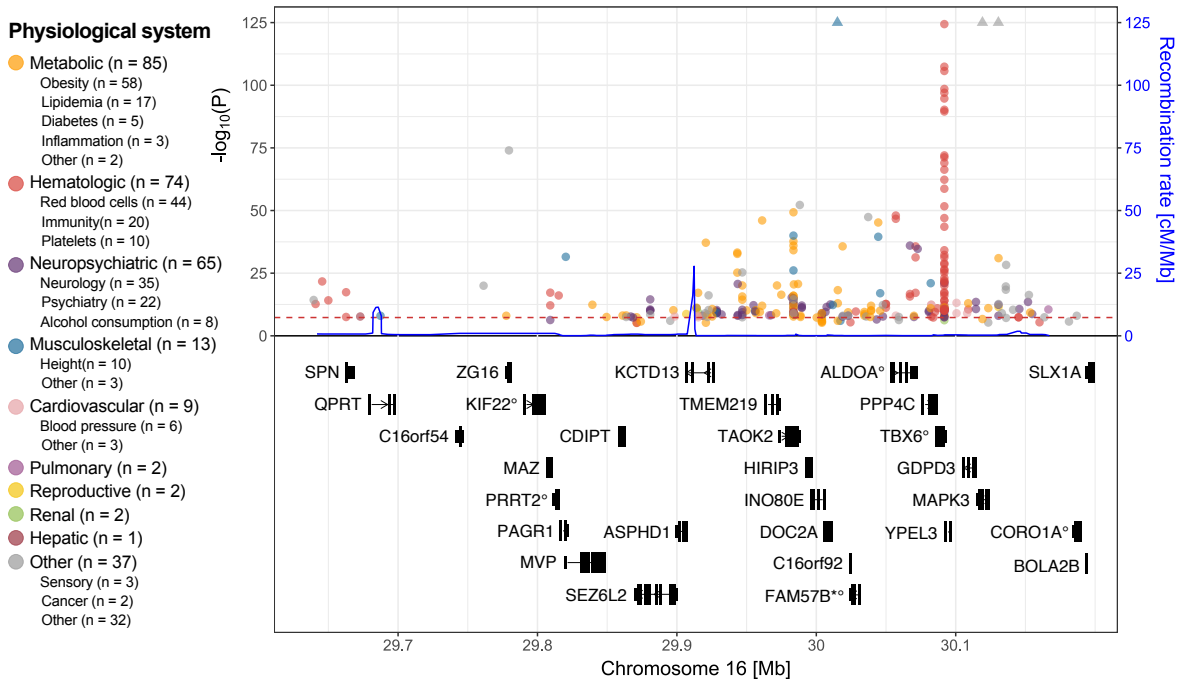


Figure 6.15: GWAS Catalog associations at 16p11.2 BP4-5.
Top: 290 single nucleotide variants (SNVs) associations mapping to the 16p11.2 BP4-5 CNV region (GRCh38) reported in the GWAS Catalog (*78*) (accessed 14/03/2024). The negative logarithm of the association p-value (left y-axis) is plotted against the genomic position (x-axis). The dashed red line represents the threshold for genome-wide significance, at $-\log_{10}(5 \times 10^{-8})$. Associations are plotted from the suggestive p-value of 7×10^{-6}. P-values for three signals, depicted as upward-facing triangles, were truncated. Associations are colored according to physiological system, using the same color scheme as in Figure 6.13. Number of signals for each category and subcategory is reported (n). The GRCh38 recombination rate in cM/Mb is depicted in blue (right y-axis) and was downloaded from Eagle (*751*). Bottom: Exonic structure of protein-coding genes overlapping the region. ° indicates OMIM morbid genes. * indicates genes that have a new HGNC symbol since the GnomAD v2.1.1 release: *FAM57B = TLCD3B*.

to metabolic, hematologic/immune, and neuropsychiatric phenotypes. Yet, associations with other traits, such as platelet count or diabetes, that have been associated with 16p11.2 BP4-5 CNVs only more recently, are also reported. This supports CNV findings, through independent genetic perturbations converging onto the same phenotypic changes, although the significance of the observed trait overlap has not been rigorously assessed via statistical tests. One caveat of single nucleotide variants GWAS results is that the lack of recombination prevents accurate mapping of these signals to specific causal genes. Strategies to contend with this include the incorporation of molecular data such as transcriptomics or proteomics. For instance, **expression quantitative loci (eQTL)**[40] were used to estimate the impact of changes in expression of 16p11.2 BP4-5 genes on hematological traits using **Mendelian randomization**[41] (*720*). The approach identified decreased expression of *CORO1A*, *KIF22*, and *BOLA2-SMG1P6* as causally decreasing lymphocyte count, thereby mimicking both the decreased gene expression expected from the region's deletion and the decreased lymphocyte count observed in deletion carriers (*720*). While few studies have successfully incorporated other mutation classes to gain functional insights, this strategy has not been systematically explored. Of course, this approach assumes that a single, or maybe a few genes in the region are responsible for a given phenotype (*458*). While this model might be true for some traits, other phenotypes might have a more polygenic basis, possibly involving interactions with genes beyond 16p11.2 BP4-5 (*273*, *279*, *280*, *282*, *547*, *559*).

40: Genetic variant associated with a given gene's expression (i.e., transcript levels).

41: Causal inference approach used in genetic epidemiology to identify causal relationships between two traits, i.e., from an exposure to an outcome. It does so by assessing the impact of genetic variants associated with the exposure trait on the outcome trait.

Diversity in experimental approaches

Establishment of the core features of 16p11.2 BP4-5 CNV carriers prompted the study of the region in controlled experimental settings through animal and human cellular models (e.g., using induced pluripotent stem cells) to gain mechanistic insights into the molecular pathways that connect altered dosage to disease features. They can be broadly divided into models that study the impact of the entire CNV versus those that independently assess the function of each of the genes mapping to the interval, sometimes using combinatorial approaches (Box 1). By controlling environmental variables and allowing engineering of precise genetic alterations, model organisms allow dissection of the individual contribution of the different genes at the locus. These experiments catalyze the development of pharmacological interventions – often targeting the GABAergic (*555*, *753*, *754*) and serotonin (*755*–*758*) systems – that result in improved cognitive and behavioral responses in mice models of the CNV. Similarly, inhibition of RhoA signaling partially restored neuronal morphology and migration, as well as functional and cognitive deficits in cellular and mouse models of the CNV (*590*, *591*, *759*), while inhibiting the ERK pathway rescued anatomical and behavioral deficits in cellular models of the duplication (*699*) and a mouse model of the deletion (*760*), respectively. Because all these models present their own limitations, it is important to replicate results across multiple experimental strategies and validate findings in humans to ensure their robustness and clinical utility (*761*). Indeed, a recent study performing transcriptional and functional profiling across various mouse tissues and human-derived cellular models emphasized the strong context dependency of transcriptomic, morphological, electrophysiological, and cell-fate signatures of the 16p11.2 CNV models (*548*).

Conclusions

The 16p11.2 BP4-5 rearrangement represents one of the most common etiologies of genomic disorders, leading to a broad and variable spectrum of phenotypes that extends far beyond neurodevelopmental disorders. Poor awareness around the syndrome and presence of varied phenotypes that require personalized solutions have been described as a challenge to access adequate and continued support after diagnosis by parents of children affected by 16p11.2 BP4-5 CNVs (*762*). To ensure equity in terms of diagnosis and provide personalized treatment plans, physicians must be aware of the different clinical presentations of these CNVs and assemble multidisciplinary teams of specialists who can anticipate and manage the different associated comorbidities (*507*, *763*). This task is complicated by our lack of understanding of the specific genetic and environmental factors that contribute to phenotypic heterogeneity. Yet, surveys of both parents of pediatric 16p11.2 BP4-5 deletion carriers and adults with incidental findings of a 16p11.2 BP4-5 CNV consecutive to their participation in a biobank reported that overall, they felt empowered and positively valued the diagnosis (*402*, *764*, *765*). In this review, we emphasize how integrating results from diverse data sources in terms of ascertainment, demographics, and ancestry, as well as analytical approach and experimental setting, can help fill current knowledge gaps and deepen our understanding of the mechanisms underlying variability

in expressivity and penetrance, with the hope that this will guide the development of personalized prevention and treatment strategies.

Deleterious enough to be enriched in clinical cohorts but not enough so to be absent from population cohorts, 16p11.2 BP4-5 is an ideal showcase example of pleiotropy, but we envision that the approaches described in this review can be adapted to better delineate and understand the pleiotropic spectrum of other recurrent CNVs and structural variants.

Acknowledgments

This study was funded by the Swiss National Science Foundation (31003A_182632, AR; 310030_189147, ZK), Horizon2020 Twinning projects (ePerMed 692145, AR), and the Department of Computational Biology (ZK) and the Center for Integrative Genomics (AR) from the University of Lausanne. It makes use of data generated by the DECIPHER community. Funding for the DECIPHER project was provided by the Wellcome Trust (WT223718/Z/21/Z).

Declaration of interests

The authors declare no competing interests.

Discussion

Discussion 7

The greatest danger for most of us is not that our aim is too high and we miss it, but that it is too low and we reach it.

– attributed to Michelangelo

7.1 Lessons learned from CNV-GWAS 229
7.1.1 Methodological advances 229
7.1.2 Beyond CNV-GWAS . . . 230
7.1.3 The future of CNV-GWAS 231
7.2 From global patterns to translational knowledge . . 231
7.2.1 Pleiotropy 232
7.2.2 Molecular mechanisms . . 232
7.2.3 Variable expressivity . . . 233
7.3 Perspectives 233
7.3.1 Mechanisms of CNV action 233
7.3.2 Modulators of CNV impact 236
7.3.3 Clinical translation 240
7.4 Conclusions 241

7.1 Lessons learned from CNV-GWAS

My thesis aimed to establish a framework to study the phenotypic consequences of CNVs in the general population. In Chapters 2 and 3, I present two studies in which this framework is described, with the first study establishing the foundations, and the second study providing numerous extensions that allow to accommodate binary traits. The framework was originally developed for microarray-based CNV calls. At the time when I started my PhD, the UK (~500k) and Estonian (~200k) Biobanks were among the largest publicly accessible cohorts with genetic information linked to phenotypic data. Hence, we applied our framework to these datasets. Other population cohorts have reached similar sample sizes since (Table 1.3) and represent prime candidates to deploy our framework.

7.1.1 Methodological advances

One of the main challenges when developing the CNV-GWAS framework was to deal with the low frequency of the accessed variants, which compounded with the low disease prevalence in UKBB severely hampered our statistical power. We used several strategies to mitigate the consequences of low power. First, we put great care into defining cases and controls, aggregating related diagnoses, and excluding individuals with uncertain disease status. By reducing noise in the phenotype, we improved statistical power. In Chapter 5, we further integrate continuous blood biomarker information to define cases and controls, illustrating how this approach can be tailored to specific research questions. Second, we adapted our multiple testing correction to account for the true number of independent tests performed. Still, relying on an arbitrary p-value cut-off was particularly subjective in the low-power setting we were working in. Indeed, different statistical tests would highlight different subsets of associations and very few associations reached statistical significance through all approaches, despite strong literature evidence supporting the finding. We found that a transparent way to deal with this uncertainty was to stratify our results into confidence tiers. By reporting these associations, along with all evidence gathered in favor of them, others might (or might not) validate our findings in the future. While not dealing with power *per se*, a third challenge was linked to the high computation time for logistic regressions in a setting where both the genetic event and outcome are rare. We pre-tested all genetic positions with computationally fast statistical models, allowing to eliminate the ones that do not show any

sign of association with the phenotype. More accurate – but also more time-consuming – statistical models can then be run on selected variants to obtain effect size estimates and p-values. This allowed to speed up computation time without compromising on accuracy. Importantly, the described strategies can be applied to other scenarios, providing a wider usage to the statistical genetics community.

Other methodological aspects, such as the choice of testing unit, are specific to CNV studies. A naïve approach is to group CNVs based on breakpoints, allowing some flexibility in absolute (i.e., fixed range) or relative (i.e., proportional to the CNV length) terms, to account for biological and technical variability in breakpoints. This tolerance parameters is arbitrary and introduces noise in the genotype definition, thereby reducing statistical power. Many of the studies presented in this dissertation use the copy-number state of the genotyping probe as the testing unit, bypassing the need for arbitrary CNV grouping. By leveraging correlation across the CNV genotype matrix, redundant probes at the core region of the CNV are grouped, reducing computation time while retaining variability around the breakpoint regions and further enabling selection of an adequate multiple testing correction threshold. Alternative strategies include using the gene or a sliding window as the testing unit (*34*, *306*). Another specificity of CNV-GWAS are the association models. Despite the existence of multiple SNP-GWAS models (Table 1.5), most studies rely on an additive model, which was shown to capture the bulk of h^2_{SNP} (*89*, *90*). In CNV-GWAS, the advantage of one model over the others is less pronounced. There is an overall trend for deletions to be more deleterious but whether the duplication will generate a similar, opposite, or no phenotype is highly dependent on the genomic region. Even at a single locus, different traits might associate through distinct models, making it crucial to test various dosage mechanisms. Yet, an additional complication is that the effect size of a CNV tends to negatively correlate with its frequency, especially in population biobanks subject to health cohort bias. This can lead to differences in power across models, warranting caution in effect interpretation.

7.1.2 Beyond CNV-GWAS

The two first chapters also propose developments that go beyond simple GWASs, allowing to gain a deeper understanding of the global impact of CNVs on human health. First, we propose a time-to-event framework that we applied to demonstrate that cases of common disease caused by rare CNVs tend to have an earlier onset than sporadic cases, regardless of the considered disease. These findings related to the concepts discussed in the introduction stating that i) common diseases represent aggregates of rarer conditions and ii) diseases with a strong genetic etiology tend to have an earlier onset. Second, CNV burden analyses increase our power to assess the role of this mutational class as a whole. Our results unanimously showed that CNVs negatively impact human health and that these consequences extend to global aspects of an individual's life, such as socio-economic status and lifespan. We hypothesize that the negative impact on lifespan results from an increased disease burden, despite not formally testing this premise. Partitioning the CNV burden to identify which genomic regions most strongly contribute to the disease burden

revealed that the majority of the detected signal originates from regions linked to known genomic disorders, including a small contribution from regions that did not yield any CNV-GWAS signal. Importantly, this analysis emphasized that the pleiotropy of these regions remains underestimated. Nonetheless, the CNV burden only explains a marginal fraction of the overall disease burden (~0.02%) and differences across diseases reflect the prominent role CNVs play in the genetic architecture of psychiatric conditions (*766*). These results are unsurprising given that the CNVs considered in our studies are rare and thus cannot explain many disease cases. One limitation is the lack of tools to properly estimate CNV-based heritability as existing methodology relies on assumptions about SNP genetic architecture that do not hold for CNVs. While such tools should provide more accurate estimates, it is unlikely that they will reveal large contributions of rare, large CNVs to the heritability of complex traits.

7.1.3 The future of CNV-GWAS

Presented conclusions only hold in the specific context of CNVs detectable from SNP microarrays. In section 1.4.2, I described modern technologies such as short- and long-read WGS that enable detection of the full spectrum of SVs, including smaller events missed by array-based technologies. Based on the observation that mutation size negatively correlates with frequency (*21*, *34*, *35*), it can be speculated that newly detected SVs will share characteristics with SNPs, so that a larger number of common SVs contribute to the genetic architecture of complex traits, possibly encapsulating some of the missing heritability. Because of their lower pathogenicity, these SVs might be more frequently inherited and thus in LD with (i.e., "**tagged**") by SNPs. This opens the possibility to impute SVs, a currently under-developed field. Supporting this view, haplotype sharing[1] can help to more sensitively detect small CNVs from both microarrays and WES data (*211*, *306*). While associations with common SVs are less likely to be novel (as tagged by SNPs), they might provide functional insights when the SV represent the casual variant. The release of hundreds of thousands of genomes over the last couple of years will allow to determine whether these speculations hold. While there is no doubt that these technologies will revolutionize our knowledge about SVs, it is worth mentioning that many of the methodological aspects developed throughout my PhD can be adapted to accommodate this new input data.

1: Haplotypes are blocks of genetic sequence inherited from a single parent as not broken up by recombination during gamete formation. They are defined through genetic linkage. Haplotype information can be obtained through **phasing**, which is key to e.g., detecting compound heterozygotes.

7.2 From global patterns to translational knowledge

Besides informing on general patterns that describe the relation between rare CNVs and the human phenome, the studies presented in Chapters 2 and 3 also form the foundation for in-depth analyses of specific CNVs, as presented in Chapters 4 through 6. These studies focus on providing mechanistic insights for specific CNVs, generating clinically valuable knowledge about these regions.

7.2.1 Pleiotropy

In line with the single gene model described in Figure 1.21A, the few single-gene CNVs uncovered by our CNV-GWAS (e.g., *LDLR* and *BRCA1* deletions) tended to manifest themselves through disruption of a single physiological system. These events amount to detecting human knock-downs and can be highly informative about a gene's function and the consequences of its disruption. Leveraging the rich phenotypic data that is available for carriers can generate new insights about the epidemiology and comorbidities of these gene-disease pairs, beyond the associated trait. A particularly exciting perspective about the increasingly common availability of sequencing-based CNV calls is the ability to detect more broadly small SV disrupting single exons/genes (*209*–*211*).

Yet, most trait-associated genetic regions overlapped multiple genes and were highly pleiotropic. Understanding this pleiotropy is key to understanding how these regions exert their pathogenic consequences, and in turn, develop treatment strategies. We use a combination of MR, covariate analysis, and matched-control approaches to elucidate whether the pleiotropy of 22q11.2, 16p11.2 BP2-3, and 16p11.2 BP4-5 CNVs is horizontal or vertical (Figure 1.13). For instance, deletions of the two 16p11.2 loci associated with severe obesity. We show that a subset of associations with these regions is secondary to the deletion's impact on adiposity. Yet, a dozen associations across a broad range of physiological systems were independent of the effect of 16p11.2 BP4-5 CNVs on BMI and other confounders, demonstrating the genuine pleiotropy of the region. The 16p11.2 BP2-3 study focused on metabolic traits and identified increased risk for early-onset type 2 diabetes, renal impairment, and inflammation as BMI-independent consequences of the deletion, suggesting a distinct form of metabolic disease. Together, this indicates that the phenotypic expression of genomic disorder CNVs results from a combination of direct and indirect pleiotropic effects. Of note, while BMI can be measured accurately and can be well-instrumented for MR studies, this is not the case for other potential confounding factors. Future research should aim at better capturing the mediatory role of a broader set of traits by incorporating CNV-specific information.

7.2.2 Molecular mechanisms

Mechanisms of pleiotropy can also be elucidated at the molecular level, by identifying driver and modifier genes (Figure 1.21). As discussed in the review presented in Chapter 6, smaller mutations can pinpoint genes whose disruption phenocopy the CNV, e.g., by leveraging SNP-GWASs and rare protein-coding burden test results, as well as molecular QTL data in an MR or colocalization framework. Phenotypic convergence across mutation types provides strong evidence for the causal involvement of a gene. A particularly interesting type of phenotypic convergence is described through the phenomenon of **allelic series**, wherein a collection of variants with variable consequences on a gene's product results in a graded phenotype (e.g., *SLCO1B1/SLCO1B3*). When multiple groups of recurrent CNVs are present, pleiotropic patterns can be dissected to narrow the putative causal region, sometimes down to a single gene (e.g., *ABCC6* and kidney stones). Importantly, different genes might drive

distinct phenotypes (e.g., 15q13.3) and multiple genes might drive the same trait (e.g., 22q11.2). Despite exploring the usage of other mutation classes throughout most of the studies presented in my thesis, more systematic and comprehensive work in that direction is needed. I discuss a few approaches in the ensuing Perspectives.

7.2.3 Variable expressivity

Another phenomenon that became increasingly apparent is the variable expressivity of recurrent CNVs. We identified multiple examples where a CNV linked to a severe clinical outcome was present in biobank participants exhibiting subclinical phenotypes. For instance, only about a third of carriers of the 17p12 duplication were diagnosed with Charcot-Marie-Tooth, despite non-diagnosed individuals having lower grip strength. Related to that, CNVs at loci linked with autosomal recessive Mendelian disorders were found to lead to mild phenotypic alterations reminiscent of the linked disorder, as exemplified by the allelic series at *SLCO1B1/SLCO1B3* locus, which causes Rotor syndrome in a digenic recessive fashion. Overall, these results indicate that the classical dichotomies between rare versus common diseases or recessive versus dominant inheritance modes do not reflect the reality where the same variant can generate a spectrum of phenotypic alterations. The severity of a CNV's expression is influenced by the ascertainment of the carriers, with the more severe expressions being clustered in clinical cohorts in which these CNVs have historically been studied. We show for 22q11.2 and 16p11.2 BP4-5 – two CNVs with notoriously high phenotypic heterogeneity – that findings from clinical and population cohorts converge on the same physiological systems, re-iterating the importance of studying these CNVs in diverse cohorts to capture their full phenotypic expression.

7.3 Perspectives

From a personalized medicine perspective, awareness of CNV pleiotropy and variable expressivity is key to improving diagnostic rates and anticipating comorbidities. While the review presented in Chapter 6 focuses on 16p11.2 BP4-5, many of the points raised relating to leveraging diversity to better understand phenotypic heterogeneity can be applied to other large, recurrent CNVs. In the following section, I explore more explicitly, and with the help of preliminary data, three key areas that I believe will answer open questions in the field and catalyze the translation of theoretical knowledge into clinical practice.

7.3.1 Mechanisms of CNV action

Protein-coding CNVs

I presented a few examples where dissecting the CNV-GWAS signal could pinpoint putative causal genes for a specific phenotype. Yet, doing so remains challenging and relies on the presence of multiple clusters of recurrent CNVs at the same loci and/or literature knowledge about the overlapping genes. If experimental approaches represent a great avenue

to decipher molecular mechanisms of CNV action, I will here focus on computational approaches that are more related to the work I conducted during my PhD. One strategy is to use MR to estimate the impact of changes in the expression of CNV-overlapping genes on the phenotypes linked to the region by CNV-GWAS, as done in Chapters 4 and 5. With the release of UKBB pQTLs (*4*), an exciting extension will be to determine how changes in transcript and protein levels of CNV-encompassed genes compare to results obtained through CNV-GWAS.

One limitation is that this approach determines the impact of individual genes. Yet, most trait-associated CNVs harbor multiple genes and there is increasing evidence for the presence of multiple driver and modifier genes per phenotype. As such, an extension of the single-gene approach would be to estimate the global causal effect of a CNV through changes in expression of encompassed genes. For deletions, this corresponds to the sum of the negative[2] gene-level MR effects weighted by the normalized expected change in expression upon deletion (Figure 7.1).

2: MR provides the phenotypic effect of one SD *increase* in expression. We assume that the deletion will *decrease* expression, hence the negative sign.

$$T = \sum_{i=1}^{n} -\alpha_{gene_i} \cdot \frac{1/2 \ \mu_{gene_i}}{SD_{gene_i}} \tag{7.1}$$

where

- n: Number of deleted genes.
- α_{gene_i}: MR effect of gene i.
- μ_{gene_i}: Average gene expression of gene i.
- SD_{gene_i}: SD of gene expression of gene i.

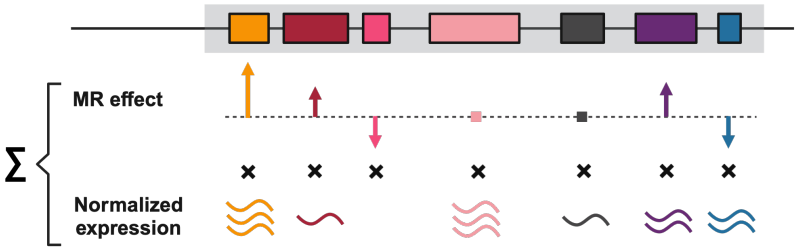


Figure 7.1: Deletion causal effect. Global causal effect of changes in gene expression induced by a multi-gene deletion on a trait, as described in Equation 7.1. For each gene (colored boxes) encompassed in the deletion region (gray box), the negative causal MR effect (arrows whose length and direction reflect the magnitude and sign of the causal effect) is multiplied by a normalized gene expression score. Individual gene-level estimates are summed up into a global causal effect.

These scores could be compared to CNV-GWAS estimates, with deviations indicating epistatic interactions or the presence of post-translational mechanisms aiming at buffering changes in gene expression. In practice, however, there are several limitations. Focusing on the MR effects, a first consideration is whether all deleted genes, or only the ones with a significant causal effect, should be considered to calculate T. While the former might add noise, the latter might neglect true effects that we were underpowered to detect. Related to that, some genes might lack sufficient IVs to be instrumented through MR. As genes will by definition be adjacent, they are also likely to share IVs. It is also unknown whether MR effects, that assess the impact of *increased* expression, can be extrapolated to the negative range. Concerning the weights, assuming halving of expression is likely an over-simplification (*227*), and the approach requires access to gene expression data (i.e., not only eQTLs), which might not always be available. In addition, knowledge about the relevant tissue is key, and calculating the score based on expression

data from an impertinent tissue might result in discrepancies with the CNV-GWAS effect, estimated at the organismal level. While larger and more diverse mRNA and protein expression studies are likely to solve some of these caveats, further theoretical work is needed to assess the implications of the abovementioned caveats.

Non-coding CNVs

Another exciting and underexplored area relates to non-coding CNVs. In the introduction, I briefly describe how SVs can act as eQTLs and have even more profound phenotypic consequences when disrupting TADs. Despite non-coding regions having lower probe coverage, a small proportion of our signals do not overlap any protein-coding region. For instance, a 260 kb 18q21.32 deletion downstream of *MC4R* was found to increase BMI by 4.2 kg/m^2 (*208*, *295*). As described in Figure 5.2, *MC4R* plays a key role in the leptin-melanocortin satiety pathway, providing a strong link to the obesity phenotype induced by the CNV. Similarly, we identify multiple independent association signals for height that map to the pseudo-autosomal region harboring *SHOX* (*208*), a gene whose haploinsufficiency is associated with penetrant forms of short stature (*316*, *317*). In the two described examples, the CNVs occur in **gene deserts**[3], and the adjacent genes are well-characterized. As the phenotypes caused by the CNV are reminiscent of the phenotypes caused by the genes' LoF, we can speculate that they exert their impact by disrupting gene regulatory regions. In other cases, the mechanism of action might not be as straightforward. The latter can be investigated by integrating data from **epigenomic assays**.

3: Large genomic regions (> 500 kb) devoided of protein-coding genes. They often occur near developmental transcription factors and tend to harbor regulatory elements for these genes.

Epigenomic assays

DNA is bound by a multitude of chemical groups (e.g., methyl groups) and proteins (e.g., histones), that together form **chromatin**. These modifications regulate the structure and packaging of DNA in a dynamic way, which in turn influences which parts of the genome are expressed at which time and in which cell. The study of these modifications is termed ***epi*genetics** (i.e., features *on top of* genetics). Several high-throughput assays have been developed to assess the epigenome:

- **ChIP-seq** (chromatin immunoprecipitation sequencing): Detects DNA binding sites for specific proteins, including transcription factors and histone modifications. The latter are useful to distinguish promoter from enhancer regions and determine whether they are in an active or repressed state.
- **ATAC-seq** (assay for transposase-accessible chromatin sequencing): Measures chromatin openness. This allows detection of accessible chromatin regions corresponding to enhancers or promoters.
- **Hi-C**: Measures chromatin conformation and interactions. Hi-C allows probing of the 3D genome structure, including the position of genomic compartments, TADs, and smaller-scale interactions.

Initiative such as the Encyclopedia of DNA Elements (ENCODE) (*767*), NIH Roadmap, (*768*), or BLUEPRINT (*769*), have generated epigenomic datasets that cover a larger number of cell types and states. These are made publicly, available, under the umbrella of the International Human Epigenome Consortium (IHEC).

Alternatively, if the region is conserved, model organisms harboring the mutation can be generated, allowing characterization of its functional consequences *in vivo*. While I did not have the time to pursue this line of investigation in depth, we started a collaboration with the laboratory of Guillaume Andrey at the University of Geneva to gain functional insights into the molecular mechanisms linking non-coding deletions in the gene desert surrounding *SHOX* to height. They will use the enSERT (enhancer insertion) protocol that uses the clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 technology to mediate site-directed insertion of an enhancer-reporter (*LacZ*) transgenic construct into the Hipp11 intergenic safe-harbor site (*770*). This approach has higher efficiency than random insertion transgenesis (~50% vs ~12%) and reduced ectopic expression as the transgenic cassettes are not subject to position effects, making the method more reproducible (*771*). Specifically, we will test whether two human candidate enhancers – selected based on our CNV-GWAS and epigenomic data (*772*, *773*) – are sufficient to drive expression in mouse through the *Shh* minimal promoter in mouse embryonic tissue at day 12.5-13.5. Of note, humans have two paralogs of the gene, *SHOX* on chromosome X, and *SHOX2* on chromosome 3, yet only the former has been associated with skeletal anomalies. Rodents have lost the copy mapping to chromosome X, and only harbor *Shox2*, suggesting that the latter has overtaken the function of both human genes. Correspondingly, this gene has often been studied in chicken or zebrafish where both copies are retained. Yet, *SHOX* and *Shox2* share an identical DNA-interacting homeodomain and conserved regulatory elements and it was demonstrated that human *SHOX* enhancers can be characterized in transgenic mice models (*774*–*777*).

7.3.2 Modulators of CNV impact

Genetic background
Following the model described in Figure 6.14 in Chapter 6, other genetic variants represent prime candidates to explain the variable expressivity and incomplete penetrance that we observed across CNVs associated with genomic disorders. While these interactions are highly complex, a strategy to study them is to decompose the genetic background into various components whose individual contributions and interactions with CNVs can be assessed. While there are many ways to do so, I here provide an analysis plan detailing how I would address the question.

Starting from a given CNV region, the simplest approach is to assess the presence of other variants on the remaining copies. This includes, for instance, the detection of compound heterozygotes generated through LoF of the single remaining copy in the case of a deletion, or the detection of more subtle effects triggered by haplotypes harboring non-coding variants whose combined effect leads to increased or decreased expression of encompassed genes. Phenotypes expressed only when compounded by another mutation have been described sporadically in clinical reports (e.g., Chapter 6), but no systematic study of CNV compounding has been performed in large population cohorts. The recent release of phased WGS and WES for UKBB (*778*), makes such analyses increasingly feasible.

Moving away from the CNV region itself, the next step involves assessing the contribution of other rare, pathogenic mutations, following the

logic of the previously described two-hit model (Figure 1.21F) (*285*, *286*). First, variants affecting genes linked to the same phenotype as the CNV should be identified. This can be done through literature review, leveraging resources such as OMIM or HPO that summarize findings from clinical studies. Alternatively, genes can be selected based on rare protein-coding burden tests from population cohorts. Once selected, variants are used to build predictors. In their simplest form, these describe the presence or absence of any of the selected variants. More complex predictors can be built as the sum of the number of present variants in a given individual, possibly weighted by pathogenicity, zygosity, and/or inheritance mode. The phenotype can then be modeled as a function of the CNV status, the rare variant predictor, and their interaction, allowing quantification of their respective contribution and possible interaction. An interesting extension would be to model these effects through time-to-event analysis.

Finally, the above-described framework can be extended by replacing the rare variant predictor with the PGS, which captures the contribution of common variants to the trait of interest. Using publicly available PGS weights calculated by LDpred2-auto (*779*), we explored how these scores relate to 108 CNV-trait associations described in Chapter 2 (Figure 7.3). We hypothesized that CNV carriers either have a similar PGS distribution to copy-neutral individuals (i.e., no interaction) or have a PGS that counteracts the effect of the CNV, based on the assumption that CNV carriers present in UKBB suffer from selection bias and have low CNV expressivity. It should be noted that differential PGS distribution between copy-neutral and CNV carriers could be due to either selection bias or PGS × CNV interactions, reiterating the importance of conducting studies in cohorts with various ascertainment biases to disentangle these mechanisms. While explicitly modeling the interaction did not yield any significant results, we were surprised to observe that for five associations, the PGS was synergetic to the CNV, pushing the phenotype toward the same direction. A similar trend was further observed for 13 additional associations. By partitioning the PGS (Figure 7.2), we demonstrate that three of these associations (i.e., hematological traits mapping to the *RHD* deletion described in Figure 2.15) can be explained by the local PGS. Tagging of a CNV by the PGS is only possible if the CNV appeared several generations ago (as opposed to *de novo*) and was inherited over multiple generations on the same haplotype. In line with this, the *RHD* deletion is the only common CNV (frequency 3.8%) among our set of studied CNV-trait pairs.

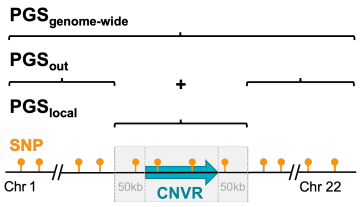


Figure 7.2: PGS partitioning.
Schematic representation of the genome-wide PGS (top) partitioned into an out (middle) and local (bottom) PGS. The out and local PGS are built by accounting for on all common variants (SNPs; orange) that *do not* or *do* overlap the CNV region (CNVR; blue) ± an adjacent region (here 50 kb, grey), respectively.

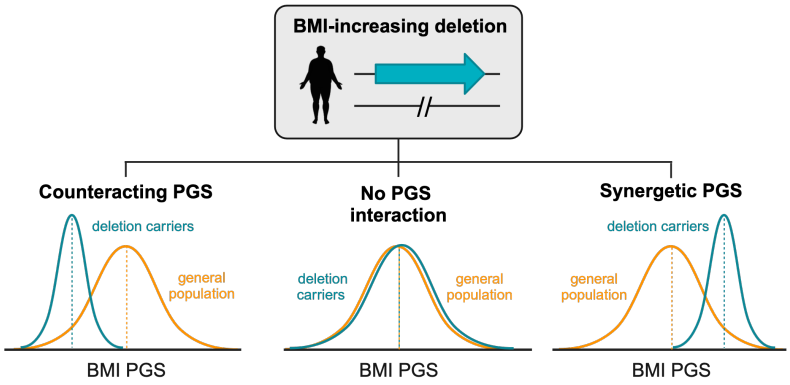


Figure 7.3: PGS of CNV carriers.
Possible scenarios of how the PGS of CNV carriers might compare to the PGS of non-carriers, illustrated here with a BMI-increasing deletion (top). Compared to the BMI PGS of the general population (orange distribution), the BMI PGS distribution of deletion carriers (blue distribution), might be: i) biased towards low PGS, thereby counteracting the deletion's effect (left), ii) similar to the general population (middle), iii) biased towards a high PGS, thereby exacerbating the deletion's effect (right).

While the other synergetic effects were weaker, they might be explained through assortative mating (Figure 7.4), a phenomenon that was recently proposed to contribute to the variable expressivity of rare variants linked to NDD (*727*). This, however, implies that a substantial fraction of CNVs is inherited. Yet, the average fraction of shared CNVs among siblings is about half of what is expected under Mendelian inheritance (see Chapter 2), implying a substantial *de novo* rate or selection bias. This analysis was not stratified by CNV pathogenicity and further research is needed to accurately determine CNV-specific rates of inheritance in the general population. Such studies will be limited by the low numbers of CNV carriers and relatives in UKBB, so that other cohorts with higher proportions of relatives might be more suitable for this type of analysis.

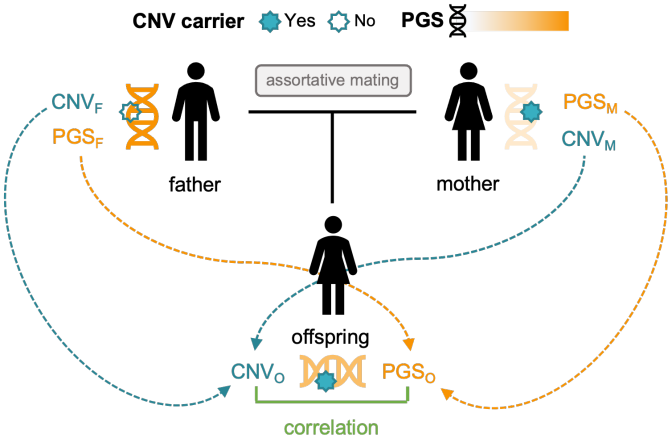


Figure 7.4: Assortative mating inducing CNV-PGS correlation.
In this scheme, assortative mating happens between the father (F), who has a high PGS for the trait on which assortment occurs but no CNV, and a mother (M), who has a low PGS for the trait on which assortment occurs but carries a CNV that associates with the trait on which assortment occurs. They produce an offspring (O) that inherits the CNV from the mother and a PGS of intermediate value that compounds to create a stronger phenotype than in either parent. This phenomenon can explain the positive correlation or synergetic effect of PGS and CNV.

Once the individual contribution of local versus distant, and rare versus common variants have been described in relation to CNVs, more complex models that provide a holistic description of how several mutational classes interact to produce the observed phenotype can be built. For instance, an interesting extension will be to determine whether CNV carriers are enriched among individuals whose observed phenotype strongly deviates from their PGS-predicted phenotype (*780*). Yet, even the most complex models are unlikely to perfectly explain phenotypic variance, due to the contribution of environmental factors. In the next section, I describe a research plan to study the role of one such factor, biological sex.

Biological sex

As the role of SV in the genetics of complex traits is emerging, so is our appreciation of biological sex as a modulator of genetic effects (*781*). Indeed, sex differences in phenotype are ubiquitous, and among the diseases studied in Chapter 3, most exhibit sex differences in prevalence. Interestingly, a substantial fraction of diseases had a significant sex-by-age interaction term, indicating that the risk of developing these diseases does not change at a constant rate across sexes, with life events such as menopause strongly impacting disease risk.

Genetics of sex differences

Recent studies have demonstrated that autosomal SNPs can exert small, albeit differential effects in males versus females (Figure 7.5A) (*782*–*787*). Explanations for sex differences in genetic architecture remain poorly understood and are likely multiple.

A prominent paradigm, known as the Carter effect, female/male protective effect, or **sex-dependent liability threshold**, stipulates that one of the sexes (often females) requires greater genetic liability to manifest the disease, a corollary of which being that, in theory, the protected sex presents higher heritability for the trait (Figure 7.5B) (*781*). Protective effects, where relatives of a proband of the protected sex are more likely to present the phenotype than relatives of a proband of the liable sex, are common (*788*–*792*). Yet, significant differences in heritability across sexes remain rare, even though they tend to corroborate the prediction that the protected sex has higher heritability (*793*–*795*). Intriguingly, a recent study in 1 million Swedish individuals found that females had an 11% *lower* heritability than males for ASD, despite females being less susceptible to the disease (*796*). **Greater genetic variability**, here in males, has been proposed as an explanation (Figure 7.5C), even though it should be noted that this model is not mutually exclusive with a sex-dependent liability threshold (*796*). Alternatively, sex modifies endogenous exposures by altering the hormonal milieu and determining life history events (e.g., pregnancies) while shaping environmental exposures through cultural and societal norms that influence lifestyle and behavior – including participation rates to genetic studies (*732*). In addition, chromosomal effects, such as dosage of genes on the X and Y chromosomes, might also contribute to sex differences. Interaction between these genetic and environmental factors and sex can modify total liability without affecting heritability (Figure 7.5D) (*781*). Recently, **amplification** was proposed as a mechanism to explain gene-by-sex interactions (*797*). Studying 27 quantitative traits in UKBB, gene-by-sex interactions were more often attributable to systematic sex differences in effect size magnitude across numerous variants (i.e., sex differential effects), rather than differences in the identity of causal variants (i.e., single-sex effects) or the direction of effects (i.e., dimorphic effects) (Figure 7.5A). The latter phenomenon could explain differences in genetic and phenotypic variance between sexes.

So far work investigating sex differences in CNV effects mainly focused on ASD, which has a much higher prevalence in males than in females (3:1 ratio) and for which CNVs represent an important risk factor (*728*, *729*). In line with the female-protective effect model, an excess of deleterious CNVs has been reported in female ASD cases (*313*–*315*, *326*), even though these results might be biased by differential clinical manifestation and societal gender biases leading to underdiagnosis in female individuals (*728*, *729*). Hence, while there seems to be a differential impact of CNVs between sexes, it remains unclear whether these observations are generalizable to a broader range of complex traits. Furthermore, sex differences of specific CNV regions (as opposed to the CNV burden) and how these relate to disparities in disease prevalence and CNV frequency across sexes remain poorly understood.

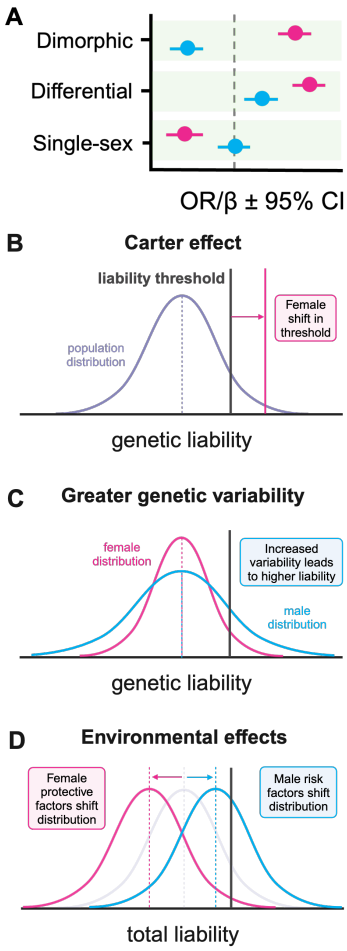


Figure 7.5: Genetics of sex differences. (**A**) Three types of sex differential effects (x-axis): opposite direction (top), same direction but of different strength (middle), or present only in one sex (bottom). (**B**) Illustration of the Carter effect or sex-dependent liability threshold model, where one sex requires a higher genetic liability to develop the disease. (**C**) Under the greater genetic variability, both sexes have the same mean genetic liability, but one sex has higher variance, so that more individuals pass the liability threshold.(**D**) The environmental effect model has a fixed liability threshold but protective and/or risk factors push the total liability distribution of females and/or males away or toward the threshold. (A-C) These are illustrations and the male (blue) and female (pink) effects might be inverted.

Conversely to what was observed in ASD cohorts (*313*–*315*, *326*), we did not identify a significantly higher CNV burden in female UKBB participants, even though it should be noted that we did not filter for

pathogenic CNVs specifically. Yet, studying the same quantitative traits as in Chapter 2, preliminary data identified 21 independent sex-differential CNV effects. Only three regions with sex-differential CNV effect showed a significant difference in CNV frequency across sexes, among which two exhibited a higher CNV frequency in the sex with the smaller size effect. This could indicate selection bias in the general population and/or (sex-differential) participation bias (*732*). Indeed, we show that UKBB is depleted for female 16p11.2 BP4-5 deletion carriers (Table 6.7) and a similar trend is observed for the BP2-3 region, with one explanation being sex-specific participation biases.

Examples of sex-specific CNV effects include a stronger increase in adiposity among female 16p11.2 BP2-3 deletion carriers, lower fluid intelligence among male 22q11.2 duplication carriers, or stronger decrease in grip strength among carriers of the Charcot-Marie-Tooth 17p12 duplication. Another interesting example is the association between *PDZK1*'s deletion and decreased serum urate levels, which was only significant in males. Several loci with sex differential impact on serum urate levels have been reported (*798*) and a variant upstream of *PDZK1* (rs1471633) was found to have a male-specific gout-increasing[4] effect (*799*). In line with this, a UKBB study found that a cluster of SNPs overlapping the region decreased risk for self-reported gout only in males (*782*). Experimental work showed that one of these SNPs modulates *PDZK1* expression by affecting binding of the transcription factor HNF4A, indicating that the region overlaps a *PDZK1* enhancer (*345*).

4: Gout is a disease characterized by inflammatory arthritis caused by hyperuricemia, i.e., elevated levels of uric acid. Among diseases studied in Chapter 3, it is the disease with the second strongest (after hernia) difference in prevalence across sexes ($OR_{males} = 6.9$, $p < 1 \times 10^{-300}$).

Overall, these preliminary results highlight that sex-specific effects might play an important role in the CNV architecture of complex traits, despite further research being required to assess the potential impact of sex-specific participation bias on the conclusions derived from such analyses. Importantly, other mutational classes and molecular data can be leveraged to elucidate the sex-specific underlying molecular mechanisms, even though such analyses are limited by the paucity of publicly available sex-stratified data. An interesting follow-up would be to extend these analyses to disease phenotypes, even though the already low power to detect CNV-disease effects in a sex-combined framework might prohibit such investigations at current sample sizes.

7.3.3 Clinical translation

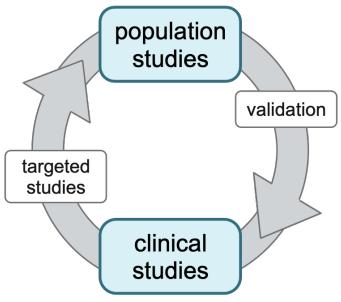


Figure 7.6: Clinical translation. Positive feedback loop that illustrates how knowledge generated by population and clinical studies reinforce each other.

During my thesis, I laid out a framework exemplifying how rare pathogenic CNVs can be studied in the general population to gain new clinical insights, with the ultimate goal that this knowledge will help diagnose and treat carriers. In the future, it will be key to consolidate findings by validating them in other biobanks (e.g., Table 1.3). This includes targeted PheWAS-based studies for clinically relevant CNVs, as presented in Chapters 4 to 6. By focusing on a single region, such studies can be tailored based on *a priori* knowledge about these regions. This allows, for instance, to refine phenotypic definitions or control for known confounders. In a second step, findings should be validated in clinical cohorts to assess their relevance. Future work should also aim at comparing frequency and effect size estimates resulting from both study settings, to quantify the impact of ascertainment and reveal the true spectrum of phenotypic consequences linked to clinically relevant CNVs. As such, I envision

that both types of studies can inform each other, generating a positive feedback loop (Figure 7.6). In parallel, more studies should assess the ethical considerations and impact of returning data to participants of biobanks, including the incidental finding of a CNV linked to a genomic disorder with variable expressivity. Current studies report a globally positive impact (*402*, *764*, *765*, *800*). Were these findings to be confirmed, it should promote the development of a biobank model wherein data sharing benefits both researchers and participants.

7.4 Conclusions

Because of challenges linked to their detection, CNVs remain an understudied mutational class. More specifically, at the start of my PhD, the bulk of knowledge about this mutational class stemmed from studies in clinical cohorts, where large, recurrent CNVs had been linked to genomic disorders. During my thesis, I adapted tools borrowed from the quantitative genetics field, which typically focuses on the study of common variants within population cohorts, to gain new insights into the phenotypic consequences of CNVs. This revealed that rare CNVs are present in a non-negligible fraction of the general population, where they act as pleiotropic phenotype modulators with variable expressivity. While their contribution at the population level remains marginal, they strongly modulate complex traits and common disease risk in carriers, making them highly valuable in the context of developing personalized medicine approaches.

Circling back to the title of my thesis, *where rare meets common* has a double meaning that is reflected both in the methodological approaches followed – at the intersection of statistical and medical genetics – as well as in the results of my research. Indeed, if I had to summarize the single most important conclusion of this body of work, it would be that the old paradigm linking rare variants to rare diseases represents an oversimplification of the complex biological reality. Instead, the same variant might produce a spectrum of phenotypic consequences, depending on the context in which it finds itself. While simplifications are sometimes necessary to grasp complexity, I believe that the next major advancements in human genetics will result from developing approaches that aim at embracing this diversity by providing a more nuanced and holistic understanding of the genetic architecture of complex human traits in the general population.

Bibliography

Here are the references in citation order.

1. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376,** 44–53 (2022) (cited on pages 3–5).
2. Codd, V. *et al.* Polygenic basis and biomedical consequences of telomere length variation. *Nature Genetics* **53,** 1425–1433 (2021) (cited on page 3).
3. UK Biobank Whole-Genome Sequencing Consortium. Whole-genome sequencing of half-a-million UK Biobank participants. *medRxiv,* 2023–12 (2023) (cited on pages 3, 6–8).
4. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622,** 329–338 (2023) (cited on pages 3, 26, 234).
5. Julkunen, H. *et al.* Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. *Nature Communications* **14,** 604 (2023) (cited on pages 3, 26).
6. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599,** 628–634 (2021) (cited on pages 3, 25, 26, 166, 192, 223).
7. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590,** 290–299 (2021) (cited on page 3).
8. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613,** 508–518 (2023) (cited on pages 3, 14, 123, 222).
9. Bick, A. G. *et al.* Genomic data in the All of Us Research Program. *Nature* (2024) (cited on pages 3, 6, 14, 222).
10. Beyter, D. *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nature Genetics* **53,** 779–786 (2021) (cited on pages 3, 8, 14, 36, 86).
11. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372,** eabf7117 (2021) (cited on pages 3, 7, 8, 36).
12. Gustafson, J. A. *et al.* Nanopore sequencing of 1000 Genomes Project samples to build a comprehensive catalog of human genetic variation. *medRxiv,* 2024–03 (2024) (cited on pages 3, 8).
13. Gong, J. *et al.* Long-read sequencing of 945 Han individuals identifies novel structural variants associated with phenotypic diversity and disease susceptibility. *medRxiv,* 2024–03 (2024) (cited on pages 3, 8).
14. Schloissnig, S. *et al.* Long-read sequencing and structural variant characterization in 1,019 samples from the 1000 Genomes Project. *bioRxiv,* 2024–04 (2024) (cited on pages 3, 8, 36).
15. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171,** 737–738 (1953) (cited on page 4).
16. Ganai, R. A. & Johansson, E. DNA replication—a matter of fidelity. *Molecular Cell* **62,** 745–755 (2016) (cited on page 6).
17. Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biology* **17,** 1–19 (2016) (cited on page 6).
18. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001) (cited on pages 6, 15).
19. Wang, T. *et al.* The Human Pangenome Project: A global resource to map genomic diversity. *Nature* **604,** 437–446 (2022) (cited on page 6).
20. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617,** 312–324 (2023) (cited on page 6).
21. Collins, R. L. *The Landscape and Consequences of Structural Variation in the Human Genome* PhD thesis (Harvard University, 2022) (cited on pages 7, 9, 37, 231).
22. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68 (2015) (cited on pages 7, 9).

23. Dhindsa, R. S. *et al.* A minimal role for synonymous variation in human disease. *The American Journal of Human Genetics* **109,** 2105–2109 (2022) (cited on page 7).

24. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Research* **11,** 863–874 (2001) (cited on page 8).

25. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics* **76,** 7–20 (2013) (cited on page 8).

26. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381,** eadg7492 (2023) (cited on page 8).

27. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* **17,** 405–423 (2015) (cited on page 8).

28. Blakes, A. J. *et al.* A systematic analysis of splicing variants identifies new diagnoses in the 100,000 Genomes Project. *Genome Medicine* **14,** 79 (2022) (cited on page 8).

29. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581,** 434–443 (2020) (cited on pages 8, 76, 166).

30. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nature Genetics* **46,** 944–950 (2014) (cited on page 8).

31. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46,** 310–315 (2014) (cited on page 8).

32. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625,** 92–100 (2024) (cited on page 8).

33. Schipper, M. & Posthuma, D. Demystifying non-coding GWAS variants: An overview of computational tools and methods. *Human Molecular Genetics* **31,** R73–R83 (2022) (cited on page 8).

34. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581,** 444–451 (2020) (cited on pages 8, 9, 36–38, 59, 86, 230, 231).

35. Abel, H. J. *et al.* Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583,** 83–89 (2020) (cited on pages 8, 9, 36–38, 59, 86, 231).

36. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015) (cited on pages 8, 36, 50, 81, 86).

37. Hassold, T. & Hunt, P. To err (meiotically) is human: The genesis of human aneuploidy. *Nature Reviews Genetics* **2,** 280–291 (2001) (cited on page 9).

38. Ohno, S. *Evolution by gene duplication* (Springer Berlin, Heidelberg, 1970) (cited on page 9).

39. Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nature Genetics* **47,** 296–303 (2015) (cited on pages 9, 55).

40. Giannuzzi, G. *et al.* The human-specific BOLA2 duplication modifies iron homeostasis and anemia predisposition in chromosome 16p11.2 autism individuals. *The American Journal of Human Genetics* **105,** 947–958 (2019) (cited on pages 9, 40, 86, 196, 216).

41. Nuttle, X. *et al.* Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* **536,** 205–209 (2016) (cited on pages 9, 194, 195, 216, 222).

42. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nature Genetics* **39,** 1256–1260 (2007) (cited on page 9).

43. Hoyt, S. J. *et al.* From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376,** eabk3112 (2022) (cited on page 9).

44. Tanudisastro, H. A., Deveson, I. W., Dashnow, H. & MacArthur, D. G. Sequencing and characterizing short tandem repeats in the human genome. *Nature Reviews Genetics,* 1–16 (2024) (cited on page 9).

45. Kloosterman, W. P. *et al.* Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Human Molecular Genetics* **20,** 1916–1924 (2011) (cited on page 9).

46. Schoeler, T., Pingault, J.-B. & Kutalik, Z. Self-report inaccuracy in the UK Biobank: Impact on inference and interplay with selective participation. *medRxiv,* 2023–10 (2023) (cited on page 10).

47. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: A resource for identification of autism genetic risk factors. *Neuron* **68,** 192–195 (2010) (cited on page 11).

48. Feliciano, P. *et al.* SPARK: A US cohort of 50,000 families to accelerate autism research. *Neuron* **97,** 488–493 (2018) (cited on page 11).

49. Pedersen, C. B. *et al.* The iPSYCH2012 case–cohort sample: New directions for unravelling genetic and environmental architectures of severe mental disorders. *Molecular Psychiatry* **23,** 6–14 (2018) (cited on page 11).

50. Fraser, A. *et al.* Cohort profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International Journal of Epidemiology* **42,** 97–110 (2013) (cited on page 12).

51. Rantakallio, P. The longitudinal study of the northern Finland birth cohort of 1966. *Paediatric and Perinatal Epidemiology* **2,** 59–88 (1988) (cited on page 12).

52. Olsen, J. *et al.* The Danish National Birth Cohort-its background, structure and aim. *Scandinavian Journal of Public Health* **29,** 300–307 (2001) (cited on page 12).

53. Magnus, P. *et al.* Cohort profile update: The Norwegian mother and child cohort study (MoBa). *International Journal of Epidemiology* **45,** 382–388 (2016) (cited on page 12).

54. Roden, D. M. *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clinical Pharmacology & Therapeutics* **84,** 362–369 (2008) (cited on page 12).

55. Carey, D. J. *et al.* The Geisinger MyCode community health initiative: An electronic health record–linked biobank for precision medicine research. *Genetics in Medicine* **18,** 906–913 (2016) (cited on page 12).

56. Staples, J. *et al.* Profiling and leveraging relatedness in a precision medicine cohort of 92,455 exomes. *The American Journal of Human Genetics* **102,** 874–889 (2018) (cited on page 12).

57. Boutin, N. T. *et al.* The evolution of a large biobank at Mass General Brigham. *Journal of Personalized Medicine* **12,** 1323 (2022) (cited on page 12).

58. Banerjee, D. & Girirajan, S. *Pathogenic variants and ascertainment: Neuropsychiatric disease risk in a health system cohort* 2023 (cited on page 12).

59. Fry, A. *et al.* Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American Journal of Epidemiology* **186,** 1026–1034 (2017) (cited on pages 12, 72, 81, 86, 123, 144, 165, 200).

60. Sudlow, C. *et al.* UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine* **12,** e1001779 (2015) (cited on pages 12, 14).

61. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209 (2018) (cited on pages 12, 14, 51, 52, 58, 86, 87, 89, 130, 151, 152, 157, 176, 199, 222).

62. Leitsalu, L. *et al.* Cohort profile: Estonian biobank of the Estonian genome center, university of Tartu. *International Journal of Epidemiology* **44,** 1137–1147 (2015) (cited on pages 12, 14, 51, 64, 87, 105, 152, 199, 222).

63. Zhou, W. *et al.* Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genomics* **2** (2022) (cited on page 13).

64. Schoeler, T. *et al.* Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nature Human Behaviour,* 1–12 (2023) (cited on page 13).

65. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *Journal of Epidemiology* **27,** S2–S8 (2017) (cited on pages 14, 222).

66. Jensson, B. O. *et al.* Actionable genotypes and their association with life span in Iceland. *New England Journal of Medicine* **389,** 1741–1752 (2023) (cited on page 14).

67. Feng, Y.-C. A. *et al.* Taiwan Biobank: A rich biomedical research database of the Taiwanese population. *Cell Genomics* **2** (2022) (cited on page 14).

68. Walters, R. G. *et al.* Genotyping and population characteristics of the China Kadoorie Biobank. *Cell Genomics* **3** (2023) (cited on page 14).

69. Brumpton, B. M. *et al.* The HUNT Study: A population-based cohort for genetic research. *Cell Genomics* **2** (2022) (cited on page 14).

70. Åsvold, B. O. *et al.* Cohort profile update: The HUNT study, Norway. *International Journal of Epidemiology* **53,** dyae013 (2023) (cited on page 14).

71. Fatumo, S. *et al.* Uganda Genome Resource: A rich research database for genomic studies of communicable and non-communicable diseases in Africa. *Cell Genomics* **2** (2022) (cited on page 14).

72. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: Past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics* **33,** 228–237 (2003) (cited on page 15).

73. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004) (cited on page 15).

74. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007) (cited on page 15).

75. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38,** 904–909 (2006) (cited on page 16).

76. Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L. & Kraft, P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *The American Journal of Human Genetics* **96,** 329–339 (2015) (cited on page 16).

77. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* **47,** D1005–D1012 (2019) (cited on pages 18, 54, 67).

78. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: Knowledgebase and deposition resource. *Nucleic Acids Research* **51,** D977–D985 (2023) (cited on pages 18, 156, 166, 223, 224).

79. Hayhurst, J. *et al.* A community driven GWAS summary statistics standard. *bioRxiv,* 2022–07 (2022) (cited on page 18).

80. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics* **19,** 807–812 (2011) (cited on pages 18, 91).

81. Visscher, P. M. *et al.* 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics* **101,** 5–22 (2017) (cited on pages 18, 26, 50, 77).

82. Auwerx, C. *et al.* Rare copy-number variants as modulators of common disease susceptibility. *Genome Medicine* **16,** 5 (2024) (cited on pages 18, 83, 131, 176, 177, 180, 200, 205–212, 214–217, 219).

83. Macé, A. *CNV Detection, Association and Interpretation* PhD thesis (Université de Lausanne, Faculté de biologie et médecine, 2017) (cited on page 18).

84. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* **11,** 499–511 (2010) (cited on page 19).

85. Gao, X., Starmer, J. & Martin, E. R. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology* **32,** 361–369 (2008) (cited on pages 19, 53, 93, 132, 155).

86. Benner, C. *et al.* FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32,** 1493–1501 (2016) (cited on page 20).

87. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **82,** 1273–1300 (2020) (cited on page 20).

88. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4,** s13742–015 (2015) (cited on pages 20, 23, 51, 87, 130, 176).

89. Hivert, V. *et al.* Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *The American Journal of Human Genetics* **108,** 786–798 (2021) (cited on pages 20, 21, 230).

90. Pazokitoroudi, A., Chiu, A. M., Burch, K. S., Pasaniuc, B. & Sankararaman, S. Quantifying the contribution of dominance deviation effects to complex trait variation in biobank-scale data. *The American Journal of Human Genetics* **108,** 799–808 (2021) (cited on pages 20, 21, 230).

91. Heng, T. H. *et al.* Widespread recessive effects on common diseases in a cohort of 44,000 British Pakistanis and Bangladeshis with high autozygosity. *medRxiv,* 2024–04 (2024) (cited on page 20).

92. Heyne, H. *et al.* Mono-and biallelic variant effects on disease at biobank scale. *Nature* **613,** 519–525 (2023) (cited on page 20).

93. Andersen, M. & Hansen, T. Genetics of metabolic traits in Greenlanders: Lessons from an isolated population. *Journal of Internal Medicine* **284,** 464–477 (2018) (cited on page 20).

94. Lim, E. T. *et al.* Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genetics* **10,** e1004494 (2014) (cited on page 20).

95. Xue, Y. *et al.* Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nature Communications* **8,** 15927 (2017) (cited on page 20).

96. Norio, R. The Finnish disease heritage III: The individual diseases. *Human Genetics* **112,** 470–526 (2003) (cited on page 20).

97. Gross, S. J., Pletcher, B. A. & Monaghan, K. G. Carrier screening in individuals of Ashkenazi Jewish descent. *Genetics in Medicine* **10,** 54–56 (2008) (cited on page 20).

98. Palmer, D. S. *et al.* Analysis of genetic dominance in the UK Biobank. *Science* **379,** 1341–1348 (2023) (cited on pages 20, 21).

99. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics* **9,** 255–266 (2008) (cited on page 21).

100. Mayhew, A. J. & Meyre, D. Assessing the heritability of complex traits in humans: Methodological challenges and opportunities. *Current Genomics* **18,** 332–340 (2017) (cited on page 21).

101. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42,** 565–569 (2010) (cited on pages 21, 22).

102. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88,** 76–82 (2011) (cited on page 21).

103. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47,** 291–295 (2015) (cited on page 21).

104. Patxot, M. *et al.* Probabilistic inference of the genetic architecture underlying functional enrichment of complex traits. *Nature Communications* **12,** 6972 (2021) (cited on page 21).

105. Burch, K. S. *et al.* Partitioning gene-level contributions to complex-trait heritability by allele frequency identifies disease-relevant genes. *The American Journal of Human Genetics* **109,** 692–709 (2022) (cited on page 21).

106. Weiner, D. J. *et al.* Polygenic architecture of rare coding variation across 394,783 exomes. *Nature* **614,** 492–499 (2023) (cited on pages 21, 223).

107. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature Genetics* **50,** 621–629 (2018) (cited on page 21).

108. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009) (cited on pages 21, 50).

109. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610,** 704–712 (2022) (cited on page 21).

110. Wainschtein, P. *et al.* Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics* **54,** 263–273 (2022) (cited on page 21).

111. Hawkes, G. *et al.* Whole genome association testing in 333,100 individuals across three biobanks identifies rare non-coding single variant and genomic aggregate associations with height. *bioRxiv,* 2023–11 (2023) (cited on page 21).

112. Schäffer, A. A. Digenic inheritance in medical genetics. *Journal of Medical Genetics* **50,** 641–652 (2013) (cited on page 21).

113. Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nature Genetics* **49,** 692–699 (2017) (cited on pages 21, 38).

114. Ferraro, N. M. *et al.* Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* **369,** eaaz5900 (2020) (cited on pages 21, 38).

115. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550,** 239–243 (2017) (cited on pages 21, 38).

116. Hernandez, R. D. *et al.* Ultrarare variants drive substantial cis heritability of human gene expression. *Nature Genetics* **51,** 1349–1355 (2019) (cited on pages 21, 38).

117. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature reviews genetics* **11,** 459–463 (2010) (cited on page 22).

118. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics* **46,** 100–106 (2014) (cited on page 22).

119. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nature Methods* **9,** 525–526 (2012) (cited on page 22).

120. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44,** 821–824 (2012) (cited on page 22).

121. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47,** 284–290 (2015) (cited on page 22).

122. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics* **51,** 1749–1755 (2019) (cited on page 22).

123. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics* **50,** 1335–1341 (2018) (cited on pages 22, 23).

124. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics* **53,** 1097–1103 (2021) (cited on pages 23, 25).

125. Ma, C., Blackwell, T., Boehnke, M., Scott, L. J. & Investigators, G. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic Epidemiology* **37,** 539–550 (2013) (cited on page 23).

126. Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika* **80,** 27–38 (1993) (cited on page 23).

127. Daniels, H. E. Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics,* 631–650 (1954) (cited on page 23).

128. Kuonen, D. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86,** 929–935 (1999) (cited on page 23).

129. Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *The American Journal of Human Genetics* **101,** 37–49 (2017) (cited on page 23).

130. Berman, J. J. *Rare diseases and orphan drugs: Keys to understanding and treating the common diseases* (Academic Press, London, 2014) (cited on pages 23, 39, 103).

131. Hughey, J. J. *et al.* Cox regression increases power to detect genotype-phenotype associations in genomic studies using the electronic health record. *BMC Genomics* **20,** 1–7 (2019) (cited on page 23).

132. Van Der Net, J. B. *et al.* Cox proportional hazards models have more statistical power than logistic regression models in cross-sectional genetic association studies. *European Journal of Human Genetics* **16,** 1111–1116 (2008) (cited on page 23).

133. Staley, J. R. *et al.* A comparison of Cox and logistic regression for use in genome-wide association studies of cohort and case-cohort design. *European Journal of Human Genetics* **25,** 854–862 (2017) (cited on page 23).

134. Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34,** 187–202 (1972) (cited on page 23).

135. Bi, W., Fritsche, L. G., Mukherjee, B., Kim, S. & Lee, S. A fast and accurate method for genome-wide time-to-event data analysis and its application to UK Biobank. *The American Journal of Human Genetics* **107,** 222–233 (2020) (cited on page 24).

136. He, L. & Kulminski, A. M. Fast algorithms for conducting large-scale GWAS of age-at-onset traits using cox mixed-effects models. *Genetics* **215,** 41–58 (2020) (cited on page 24).

137. Dey, R. *et al.* Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks. *Nature Communications* **13,** 5437 (2022) (cited on page 24).

138. Falconer, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics* **29,** 51–76 (1965) (cited on pages 25, 86).

139. Pedersen, E. M. *et al.* Accounting for age of onset and family history improves power in genome-wide association studies. *The American Journal of Human Genetics* **109,** 417–432 (2022) (cited on pages 25, 120, 121).

140. Xiao, R. & Boehnke, M. Quantifying and correcting for the winner's curse in genetic association studies. *Genetic Epidemiology* **33,** 453–462 (2009) (cited on page 25).

141. Nicolae, D. L. Association tests for rare variants. *Annual Review of Genomics and Human Genetics* **17,** 117–130 (2016) (cited on page 25).

142. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89,** 82–93 (2011) (cited on pages 25, 214, 223).

143. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics* **91,** 224–237 (2012) (cited on pages 25, 223).

144. Zhou, W. *et al.* SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. *Nature Genetics* **54,** 1466–1469 (2022) (cited on page 25).

145. Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genomics* **2** (2022) (cited on pages 25, 192, 223).

146. Bastarache, L., Denny, J. C. & Roden, D. M. Phenome-wide association studies. *JAMA* **327,** 75–76 (2022) (cited on page 26).

147. Katz, D. H. *et al.* Proteomic profiling platforms head to head: Leveraging genetics and clinical traits to compare aptamer-and antibody-based methods. *Science Advances* **8,** eabm5164 (2022) (cited on page 26).

148. Emwas, A.-H. M. The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. *Metabonomics: Methods and Protocols,* 161–193 (2015) (cited on page 26).

149. Oliva, M. *et al.* DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nature Genetics* **55,** 112–122 (2023) (cited on page 26).

150. Min, J. L. *et al.* Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nature Genetics* **53,** 1311–1321 (2021) (cited on page 26).

151. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369,** 1318–1330 (2020) (cited on pages 26, 30, 68, 157, 167).

152. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nature Genetics* **53,** 1290–1299 (2021) (cited on pages 26, 30).

153. De Klein, N. *et al.* Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. *Nature Genetics* **55,** 377–388 (2023) (cited on pages 26, 30).

154. Võsa, U. *et al.* Large-scale cis-and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics* **53,** 1300–1310 (2021) (cited on pages 26, 57, 133, 157, 166, 167).

155. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558,** 73–79 (2018) (cited on page 26).

156. Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nature Metabolism* **2,** 1135–1148 (2020) (cited on page 26).

157. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nature Genetics* **53,** 1712–1721 (2021) (cited on page 26).

158. Shin, S.-Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nature Genetics* **46,** 543–550 (2014) (cited on page 26).

159. Chen, Y. *et al.* Genomic atlas of the plasma metabolome prioritizes metabolites implicated in human diseases. *Nature Genetics* **55,** 44–53 (2023) (cited on page 26).

160. Lotta, L. A. *et al.* A cross-platform approach identifies genetic regulators of human metabolism and health. *Nature Genetics* **53,** 54–64 (2021) (cited on page 26).

161. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593,** 238–243 (2021) (cited on page 26).

162. Tambets, R., Kolde, A., Kolberg, P., Love, M. I. & Alasoo, K. Extensive co-regulation of neighbouring genes complicates the use of eQTLs in target gene prioritisation. *bioRxiv,* 2023–09 (2023) (cited on page 26).

163. Frayling, T. M. *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316,** 889–894 (2007) (cited on page 26).

164. Dina, C. *et al.* Variation in FTO contributes to childhood obesity and severe adult obesity. *Nature Genetics* **39,** 724–726 (2007) (cited on page 26).

165. Claussnitzer, M. *et al.* FTO obesity variant circuitry and adipocyte browning in humans. *New England Journal of Medicine* **373,** 895–907 (2015) (cited on page 26).

166. Wu, Y. *et al.* Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nature Communications* **9,** 918 (2018) (cited on page 27).

167. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics* **10,** e1004383 (2014) (cited on pages 27, 157, 167, 192).

168. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genetics* **17,** e1009440 (2021) (cited on page 27).

169. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics* **99,** 1245–1260 (2016) (cited on page 27).

170. Sanderson, E. *et al.* Mendelian randomization. *Nature Reviews Methods Primers* **2,** 6 (2022) (cited on pages 28, 29).

171. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics* **51,** 1339–1348 (2019) (cited on pages 28, 50, 77).

172. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44,** 512–525 (2015) (cited on page 29).

173. Porcu, E. *et al.* Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nature Communications* **10,** 3300 (2019) (cited on pages 29, 57, 68, 133, 157, 166, 192).

174. Porcu, E. *et al.* Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. *Nature Communications* **12,** 5647 (2021) (cited on pages 29, 30).

175. Sadler, M. C., Auwerx, C., Lepik, K., Porcu, E. & Kutalik, Z. Quantifying the role of transcript levels in mediating DNA methylation effects on complex traits and diseases. *Nature Communications* **13,** 7559 (2022) (cited on pages 29, 151, 157, 179).

176. Auwerx, C. *et al.* Exploiting the mediating role of the metabolome to unravel transcript-to-phenotype associations. *eLife* **12,** e81097 (2023) (cited on pages 29, 30).

177. Van der Graaf, A. *et al.* MR-link-2: Pleiotropy robust cis Mendelian randomization validated in four independent gold-standard datasets of causality. *medRxiv,* 2024–01 (2024) (cited on page 30).

178. Zuber, V. *et al.* Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *The American Journal of Human Genetics* **109,** 767–782 (2022) (cited on page 30).

179. Alasoo, K. *et al.* Genetic effects on promoter usage are highly context-specific and contribute to complex traits. *eLife* **8,** e41673 (2019) (cited on page 30).

180. Carvalho, C. M. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics* **17,** 224–238 (2016) (cited on pages 31, 32, 86, 194).

181. Weckselblatt, B. & Rudd, M. K. Human structural variation: Mechanisms of chromosome rearrangements. *Trends in Genetics* **31,** 587–599 (2015) (cited on page 31).

182. Gu, W., Zhang, F. & Lupski, J. R. Mechanisms for human genomic rearrangements. *Pathogenetics* **1,** 1–17 (2008) (cited on page 31).

183. Lehrman, M. A. *et al.* Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. *Science* **227,** 140–146 (1985) (cited on page 31).

184. Le Saux, O. *et al.* A spectrum of ABCC6 mutations is responsible for pseudoxanthoma elasticum. *The American Journal of Human Genetics* **69,** 749–764 (2001) (cited on pages 31, 40, 114).

185. Ringpfeil, F., Nakano, A., Uitto, J. & Pulkkinen, L. Compound heterozygosity for a recurrent 16.5-kb Alu-mediated deletion mutation and single-base-pair substitutions in the ABCC6 gene results in pseudoxanthoma elasticum. *The American Journal of Human Genetics* **68,** 642–652 (2001) (cited on pages 31, 40, 114).

186. Rossetti, L. C., Goodeve, A., Larripa, I. B. & De Brasi, C. D. Homeologous recombination between AluSx-sequences as a cause of hemophilia. *Human Mutation* **24,** 440–440 (2004) (cited on page 31).

187. Boone, P. M. *et al.* The Alu-rich genomic architecture of SPAST predisposes to diverse and functionally distinct disease-associated CNV alleles. *The American Journal of Human Genetics* **95,** 143–161 (2014) (cited on page 31).

188. Turner, D. J. *et al.* Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nature Genetics* **40,** 90–95 (2007) (cited on page 31).

189. Lopes, J. *et al.* Sex–dependent rearrangements resulting in CMT1A and HNPP. *Nature Genetics* **17,** 136–137 (1997) (cited on page 31).

190. Wirth, B. *et al.* De novo rearrangements found in 2% of index patients with spinal muscular atrophy: Mutational mechanisms, parental origin, mutation rate, and implications for genetic counseling. *The American Journal of Human Genetics* **61,** 1102–1111 (1997) (cited on page 31).

191. Hehir-Kwa, J. Y. *et al.* De novo copy number variants associated with intellectual disability have a paternal origin and age bias. *Journal of Medical Genetics* **48,** 776–778 (2011) (cited on page 31).

192. Lázaro, C. *et al.* Sex differences in mutational rate and mutational mechanism in the NF1 gene in neurofibromatosis type 1 patients. *Human Genetics* **98,** 696–699 (1996) (cited on page 31).

193. Duyzend, M. H. *et al.* Maternal modifiers and parent-of-origin bias of the autism-associated 16p11.2 CNV. *The American Journal of Human Genetics* **98,** 45–57 (2016) (cited on pages 31, 197, 220).

194. MacArthur, J. A. *et al.* The rate of nonallelic homologous recombination in males is highly variable, correlated between monozygotic twins and independent of age. *PLoS Genetics* **10,** e1004195 (2014) (cited on page 31).

195. Abyzov, A. *et al.* Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nature Communications* **6,** 7256 (2015) (cited on page 32).

196. Pannunzio, N. R., Li, S., Watanabe, G. & Lieber, M. R. Non-homologous end joining often uses microhomology: Implications for alternative end joining. *DNA Repair* **17,** 74–80 (2014) (cited on page 32).

197. Burssed, B., Zamariolli, M., Bellucco, F. T. & Melaragno, M. I. Mechanisms of structural chromosomal rearrangement formation. *Molecular Cytogenetics* **15,** 23 (2022) (cited on page 32).

198. Carvalho, C. M. *et al.* Replicative mechanisms for CNV formation are error prone. *Nature Genetics* **45,** 1319–1326 (2013) (cited on page 32).

199. Newman, S., Hermetz, K. E., Weckselblatt, B. & Rudd, M. K. Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. *The American Journal of Human Genetics* **96,** 208–220 (2015) (cited on page 33).

200. Winchester, L., Yau, C. & Ragoussis, J. Comparing CNV detection methods for SNP arrays. *Briefings in Functional Genomics and Proteomics* **8,** 353–366 (2009) (cited on page 34).

201. Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics* **40,** 1253–1260 (2008) (cited on page 34).

202. Colella, S. *et al.* QuantiSNP: An objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research* **35,** 2013–2025 (2007) (cited on pages 34, 35).

203. Wang, K. *et al.* PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17,** 1665–1674 (2007) (cited on pages 34, 51, 59, 87, 130, 151, 152, 158, 177).

204. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics* **40,** 1166–1174 (2008) (cited on page 35).

205. Valsesia, A., Macé, A., Jacquemont, S., Beckmann, J. S. & Kutalik, Z. The growing importance of CNVs: New insights for detection and clinical interpretation. *Frontiers in Genetics* **4,** 92 (2013) (cited on pages 35, 43, 50, 65, 123).

206. Mace, A. *et al.* New quality measure for SNP array based CNV detection. *Bioinformatics* **32,** 3298–3305 (2016) (cited on pages 35, 51, 52, 59, 81, 87, 88, 130, 131, 151, 152, 158, 177).

207. Lepamets, M. *et al.* Omics-informed CNV calls reduce false-positive rates and improve power for CNV-trait associations. *Human Genetics and Genomics Advances* **3** (2022) (cited on page 35).

208. Auwerx, C. *et al.* The individual and global impact of copy-number variants on complex human traits. *The American Journal of Human Genetics* **109,** 647–668 (2022) (cited on pages 35, 38, 47, 83, 86, 87, 90, 93–95, 99, 100, 102, 105–108, 110–112, 114, 116, 120, 122, 130–132, 141, 143, 150–153, 158, 169, 176, 177, 180, 181, 183, 200, 202, 203, 208–213, 215, 216, 235).

209. Fitzgerald, T. & Birney, E. CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank. *Cell Genomics* **2** (2022) (cited on pages 36, 68, 81, 86, 232).

210. Babadi, M. *et al.* GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. *Nature Genetics* **55,** 1589–1597 (2023) (cited on pages 36, 86, 232).

211. Hujoel, M. L. *et al.* Protein-altering variants at copy number-variable regions influence diverse human phenotypes. *Nature Genetics*, 1–10 (2024) (cited on pages 36, 86, 231, 232).

212. Danecek, P. *et al.* Detection and characterisation of copy number variants from exome sequencing in the DDD study. *Genetics in Medicine Open,* 101818 (2024) (cited on page 36).

213. Quinodoz, M. *et al.* Detection of elusive DNA copy-number variations in hereditary disease and cancer through the use of noncoding and off-target sequencing reads. *The American Journal of Human Genetics* **111,** 701–713 (2024) (cited on page 36).

214. Gabrielaite, M. *et al.* A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *Cancers* **13,** 6283 (2021) (cited on page 36).

215. Klambauer, G. *et al.* cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Research* **40,** e69–e69 (2012) (cited on page 36).

216. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology* **15,** 1–19 (2014) (cited on page 36).

217. Rausch, T. *et al.* DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28,** i333–i339 (2012) (cited on page 36).

218. Chen, X. *et al.* Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32,** 1220–1222 (2016) (cited on page 36).

219. Van Dijk, E. L. *et al.* Genomics in the long-read sequencing era. *Trends in Genetics* (2023) (cited on page 36).

220. Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nature Biotechnology,* 1–10 (2024) (cited on page 36).

221. Jiang, T. *et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology* **21,** 1–24 (2020) (cited on page 36).

222. Heller, D. & Vingron, M. SVIM: Structural variant identification using mapped long reads. *Bioinformatics* **35,** 2907–2915 (2019) (cited on page 36).

223. Marx, V. Method of the year: Long-read sequencing. *Nature Methods* **20,** 6–11 (2023) (cited on page 36).

224. Miller, D. E. *et al.* Targeted long-read sequencing identifies missing disease-causing variation. *The American Journal of Human Genetics* **108,** 1436–1449 (2021) (cited on page 37).

225. Van der Sanden, B. *et al.* Optical genome mapping enables accurate repeat expansion testing. *bioRxiv,* 2024–04 (2024) (cited on page 37).

226. Dremsek, P. *et al.* Optical genome mapping in routine human genetic diagnostics—its advantages and limitations. *Genes* **12,** 1958 (2021) (cited on page 37).

227. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nature Reviews Genetics* **14,** 125–138 (2013) (cited on pages 37, 86, 108, 234).

228. Harel, T. & Lupski, J. Genomic disorders 20 years on—mechanisms for clinical manifestations. *Clinical Genetics* **93,** 439–449 (2018) (cited on page 37).

229. Collins, R. L. *et al.* A cross-disorder dosage sensitivity map of the human genome. *Cell* **185,** 3041–3055 (2022) (cited on pages 38, 77, 86, 87, 93, 120, 159, 176, 198–200).

230. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nature Reviews Genetics* **19,** 453–467 (2018) (cited on page 38).

231. Han, L. *et al.* Functional annotation of rare structural variation in the human brain. *Nature Communications* **11,** 2990 (2020) (cited on page 38).

232. Benito-Sanz, S. *et al.* Identification of the first recurrent PAR1 deletion in Léri-Weill dyschondrosteosis and idiopathic short stature reveals the presence of a novel SHOX enhancer. *Journal of Medical Genetics* **49,** 442–450 (2012) (cited on page 38).

233. Jensen, T. D. *et al.* Integration of transcriptomics and long-read genomics prioritizes structural variants in rare disease. *medRxiv,* 2024–03 (2024) (cited on page 38).

234. Chen, R. *et al.* Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nature Biotechnology* **34,** 531–538 (2016) (cited on pages 39, 73, 79, 86, 121).

235. Wright, C. F. *et al.* Assessing the pathogenicity, penetrance, and expressivity of putative disease-causing variants in a population setting. *The American Journal of Human Genetics* **104,** 275–286 (2019) (cited on pages 39, 73, 79, 81, 86, 121, 130).

236. Goodrich, J. K. *et al.* Determinants of penetrance and variable expressivity in monogenic metabolic conditions across 77,184 exomes. *Nature Communications* **12,** 3505 (2021) (cited on pages 39, 73, 79, 81, 86, 121).

237. Kingdom, R. *et al.* Rare genetic variants in genes and loci linked to dominant monogenic developmental disorders cause milder related phenotypes in the general population. *The American Journal of Human Genetics* **109,** 1308–1316 (2022) (cited on pages 39, 86, 121, 122, 144).

238. Urpa, L. *et al.* Evidence for the additivity of rare and common variant burden throughout the spectrum of intellectual disability. *European Journal of Human Genetics*, 1–8 (2024) (cited on page 39).

239. Mars, N. *et al.* The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nature Communications* **11,** 6383 (2020) (cited on page 39).

240. Cirnigliaro, M. *et al.* The contributions of rare inherited and polygenic risk to ASD in multiplex families. *Proceedings of the National Academy of Sciences* **120,** e2215632120 (2023) (cited on page 39).

241. Kingdom, R. & Wright, C. F. Incomplete penetrance and variable expressivity: From clinical studies to population cohorts. *Frontiers in Genetics* **13,** 920390 (2022) (cited on pages 39, 130).

242. Phillips 3rd, J. & Cogan, J. Genetic basis of endocrine disease. 6. Molecular basis of familial human growth hormone deficiency. *The Journal of Clinical Endocrinology & Metabolism* **78,** 11–16 (1994) (cited on page 40).

243. Hobbs, H. H., Russell, D. W., Brown, M. S. & Goldstein, J. L. The LDL receptor locus in familial hypercholesterolemia: Mutational analysis of a membrane protein. *Annual Review of Genetics* **24,** 133–170 (1990) (cited on pages 40, 109).

244. Iacocca, M. A. & Hegele, R. A. Role of DNA copy number variation in dyslipidemias. *Current Opinion in Lipidology* **29,** 125–132 (2018) (cited on pages 40, 109).

245. Van de Steeg, E. *et al.* Complete OATP1B1 and OATP1B3 deficiency causes human Rotor syndrome by interrupting conjugated bilirubin reuptake into the liver. *The Journal of Clinical Investigation* **122,** 519–528 (2012) (cited on pages 40, 72).

246. Sleegers, K. *et al.* APP duplication is sufficient to cause early onset Alzheimer's dementia with cerebral amyloid angiopathy. *Brain* **129,** 2977–2983 (2006) (cited on page 40).

247. Rovelet-Lecrux, A. *et al.* APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nature Genetics* **38,** 24–26 (2006) (cited on page 40).

248. Duan, D., Goemans, N., Takeda, S., Mercuri, E. & Aartsma-Rus, A. Duchenne muscular dystrophy. *Nature Reviews Disease Primers* **7,** 13 (2021) (cited on page 40).

249. Campuzano, V. *et al.* Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* **271,** 1423–1427 (1996) (cited on page 40).

250. MacDonald, M. E. *et al.* A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72,** 971–983 (1993) (cited on page 40).

251. Lefebvre, S. *et al.* Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* **80,** 155–165 (1995) (cited on page 40).

252. Mailman, M. D. *et al.* Molecular analysis of spinal muscular atrophy and modification of the phenotype by SMN2. *Genetics in Medicine* **4,** 20–26 (2002) (cited on page 40).

253. Lakich, D., Kazazian Jr, H. H., Antonarakis, S. E. & Gitschier, J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nature Genetics* **5,** 236–241 (1993) (cited on page 40).

254. Becker, J. *et al.* Characterization of the factor VIII defect in 147 patients with sporadic hemophilia A: Family studies indicate a mutation type-dependent sex ratio of mutation frequencies. *The American Journal of Human Genetics* **58,** 657 (1996) (cited on page 40).

255. Tamary, H. & Dgany, O. *Alpha-thalassemia* (GeneReviews, Seattle (WA), 2020) (cited on page 40).

256. Avent, N. D. & Reid, M. E. The Rh blood group system: A review. *Blood* **95,** 375–387 (2000) (cited on pages 40, 68, 69).

257. Yang, Y. *et al.* Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): Low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *The American Journal of Human Genetics* **80,** 1037–1054 (2007) (cited on page 40).

258. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307,** 1434–1440 (2005) (cited on page 40).

259. Manickam, K. *et al.* Exome and genome sequencing for pediatric patients with congenital anomalies or intellectual disability: An evidence-based clinical guideline of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine* **23,** 2029–2037 (2021) (cited on page 40).

260. Leblond, C. S. *et al.* Operative list of genes associated with autism and neurodevelopmental disorders based on database review. *Molecular and Cellular Neuroscience* **113,** 103623 (2021) (cited on page 40).

261. Manning, M., Hudgins, L. & Professional Practice and Guidelines Committee. Array-based technology and recommendations for utilization in medical genetics practice for detection of chromosomal abnormalities. *Genetics in Medicine* **12,** 742–745 (2010) (cited on page 40).

262. Miller, D. T. *et al.* Consensus statement: Chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *The American Journal of Human Genetics* **86,** 749–764 (2010) (cited on page 40).

263. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annual Review of Medicine* **61,** 437–455 (2010) (cited on page 40).

264. Girirajan, S., Campbell, C. D. & Eichler, E. E. Human copy number variation and complex genetic disease. *Annual Review of Genetics* **45,** 203–226 (2011) (cited on page 40).

265. Firth, H. V. *et al.* DECIPHER: Database of chromosomal imbalance and phenotype in humans using ensembl resources. *The American Journal of Human Genetics* **84,** 524–533 (2009) (cited on pages 41, 50, 86, 153, 158, 195).

266. Golzio, C. *et al.* KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. *Nature* **485,** 363–367 (2012) (cited on pages 42, 196, 213, 221).

267. Lupski, J. R. *et al.* DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell* **66,** 219–232 (1991) (cited on page 42).

268. Patel, P. I. *et al.* The gene for the peripheral myelin protein PMP–22 is a candidate for Charcot–Marie–Tooth disease type 1A. *Nature Genetics* **1,** 159–165 (1992) (cited on page 42).

269. Ionasescu, V. V. *et al.* Severe Charcot-Marie-Tooth neuropathy type 1A with 1-base pair deletion and frameshift mutation in the peripheral myelin protein 22 gene. *Muscle & Nerve* **20,** 1308–1310 (1997) (cited on page 42).

270. Valentijn, L. J. *et al.* Identical point mutations of PMP–22 in Trembler–J mouse and Charcot–Marie–Tooth disease type 1A. *Nature Genetics* **2,** 288–291 (1992) (cited on page 42).

271. Suter, U. *et al.* Trembler mouse carries a point mutation in a myelin gene. *Nature* **356,** 241–244 (1992) (cited on page 42).

272. Qiu, Y. *et al.* Oligogenic effects of 16p11.2 copy-number variation on craniofacial development. *Cell Reports* **28,** 3320–3328 (2019) (cited on pages 42, 196, 213).

273. Arbogast, T. *et al.* Kctd13-deficient mice display short-term memory impairment and sex-dependent genetic interactions. *Human Molecular Genetics* **28,** 1474–1486 (2019) (cited on pages 42, 196, 213, 221, 224).

274. Kretz, P. F. *et al.* Dissecting the autism-associated 16p11.2 locus identifies multiple drivers in neuroanatomical phenotypes and unveils a male-specific role for the major vault protein. *Genome Biology* **24,** 261 (2023) (cited on pages 42, 196, 204, 221).

275. Li, G. *et al.* TBX6 as a cause of a combined skeletal-kidney dysplasia syndrome. *American Journal of Medical Genetics Part A* **188,** 3469–3481 (2022) (cited on pages 42, 214, 215).

276. Wu, N. *et al.* TBX6 null variants and a common hypomorphic allele in congenital scoliosis. *New England Journal of Medicine* **372,** 341–350 (2015) (cited on pages 42, 203, 214, 222).

277. Verbitsky, M. *et al.* The copy number variation landscape of congenital anomalies of the kidney and urinary tract. *Nature Genetics* **51,** 117–127 (2019) (cited on pages 42, 86, 110, 191, 198, 214, 215).

278. Nik-Zainal, S. *et al.* High incidence of recurrent copy number variants in patients with isolated and syndromic Müllerian aplasia. *Journal of Medical Genetics* **48,** 197–204 (2011) (cited on pages 42, 215).

279. Loviglio, M. N. *et al.* Chromosomal contacts connect loci associated with autism, BMI and head circumference phenotypes. *Molecular Psychiatry* **22,** 836–849 (2017) (cited on pages 42, 64, 150, 165, 207, 221, 224).

280. Loviglio, M. N. *et al.* The immune signaling adaptor LAT contributes to the neuroanatomical phenotype of 16p11.2 BP2-BP3 CNVs. *The American Journal of Human Genetics* **101,** 564–577 (2017) (cited on pages 42, 168, 221, 223, 224).

281. Pizzo, L. *et al.* Functional assessment of the "two-hit" model for neurodevelopmental defects in Drosophila and X. laevis. *PLoS Genetics* **17,** e1009112 (2021) (cited on page 42).

282. Weiner, D. J. *et al.* Statistical and functional convergence of common and rare genetic influences on autism at chromosome 16p. *Nature Genetics* **54,** 1630–1639 (2022) (cited on pages 42, 207, 221, 224).

283. Jensen, M. *et al.* Combinatorial patterns of gene expression changes contribute to variable expressivity of the developmental delay-associated 16p12.1 deletion. *Genome Medicine* **13,** 1–21 (2021) (cited on page 42).

284. Singh, M. D. *et al.* NCBP2 modulates neurodevelopmental defects of the 3q29 deletion in Drosophila and Xenopus laevis models. *PLoS Genetics* **16,** e1008590 (2020) (cited on page 42).

285. Girirajan, S. & Eichler, E. E. Phenotypic variability and genetic susceptibility to genomic disorders. *Human Molecular Genetics* **19,** R176–R187 (2010) (cited on pages 42, 220, 222, 237).

286. Girirajan, S. *et al.* Phenotypic heterogeneity of genomic disorders and rare copy-number variants. *New England Journal of Medicine* **367,** 1321–1331 (2012) (cited on pages 42, 197, 220, 237).

287. Pizzo, L. *et al.* Rare variants in the genetic background modulate cognitive and developmental phenotypes in individuals carrying disease-associated variants. *Genetics in Medicine* **21,** 816–825 (2019) (cited on pages 42, 218, 220).

288. Jensen, M. *et al.* A higher rare CNV burden in the genetic background potentially contributes to intellectual disability phenotypes in 22q11.2 deletion syndrome. *European Journal of Medical Genetics* **61,** 209–212 (2018) (cited on page 42).

289. Oetjens, M., Kelly, M., Sturm, A., Martin, C. & Ledbetter, D. Quantifying the polygenic contribution to variable expressivity in eleven rare genetic disorders. *Nature Communications* **10,** 4897 (2019) (cited on pages 42, 79, 221).

290. Bergen, S. E. *et al.* Joint contributions of rare copy number variants and common SNPs to risk for schizophrenia. *American Journal of Psychiatry* **176,** 29–35 (2019) (cited on pages 42, 221).

291. Davies, R. W. *et al.* Using common genetic variation to examine phenotypic expression and risk prediction in 22q11.2 deletion syndrome. *Nature Medicine* **26,** 1912–1918 (2020) (cited on pages 42, 130).

292. Aguirre, M., Rivas, M. A. & Priest, J. Phenome-wide burden of copy-number variation in the UK biobank. *The American Journal of Human Genetics* **105,** 373–383 (2019) (cited on pages 43, 51, 59, 61, 79, 86, 106, 141, 150, 158, 176, 183, 200, 209, 210, 212, 214, 215).

293. Crawford, K. *et al.* Medical consequences of pathogenic CNVs in adults: Analysis of the UK Biobank. *Journal of Medical Genetics* **56,** 131–138 (2019) (cited on pages 43, 51, 86, 95, 100, 104, 106, 116, 120, 141, 142, 150, 158, 176, 200, 209, 210, 212, 214–216).

294. Owen, D. *et al.* Effects of pathogenic CNVs on physical traits in participants of the UK Biobank. *BMC Genomics* **19,** 1–9 (2018) (cited on pages 43, 51, 79, 86, 106, 111, 116, 141, 143, 176, 200, 209, 212, 213).

295. Macé, A. *et al.* CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. *Nature Communications* **8,** 744 (2017) (cited on pages 43, 47, 51–53, 59, 61, 79, 81, 86, 93, 106, 122, 150, 158, 176, 181, 183, 200, 209, 235).

296. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nature Genetics* **50,** 1593–1599 (2018) (cited on pages 50, 77).

297. Shaikh, T. H. Copy number variation disorders. *Current Genetic Medicine Reports* **5,** 183–190 (2017) (cited on page 50).

298. Myocardial Infarction Genetics Consortium. *Nature Genetics* **41,** 334–341 (2009) (cited on page 50).

299. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464,** 713–720 (2010) (cited on page 50).

300. Kendall, K. M. *et al.* Cognitive performance among carriers of pathogenic copy number variants: Analysis of 152,000 UK Biobank subjects. *Biological Psychiatry* **82,** 103–110 (2017) (cited on pages 51, 79, 86, 122, 141, 200, 202).

301. Warland, A., Kendall, K. M., Rees, E., Kirov, G. & Caseras, X. Schizophrenia-associated genomic copy number variants and subcortical brain volumes in the UK Biobank. *Molecular Psychiatry* **25,** 854–862 (2020) (cited on page 51).

302. Kendall, K. M. *et al.* Association of rare copy number variants with risk of depression. *JAMA Psychiatry* **76,** 818–825 (2019) (cited on pages 51, 86, 141, 208).

303. Bracher-Smith, M. *et al.* Effects of pathogenic CNVs on biochemical markers: A study on the UK Biobank. *bioRxiv,* 723270 (2019) (cited on page 51).

304. Li, Y. R. *et al.* Rare copy number variants in over 100,000 European ancestry subjects reveal multiple disease associations. *Nature Communications* **11,** 255 (2020) (cited on pages 51, 86).

305. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nature Genetics* **53,** 185–194 (2021) (cited on pages 51, 67–69, 72, 79, 80, 86, 176, 200, 209, 210, 215).

306. Hujoel, M. L. *et al.* Influences of rare copy-number variation on human complex traits. *Cell* **185,** 4233–4248 (2022) (cited on pages 51, 61, 79, 81, 86, 95, 120, 122, 176, 209–212, 215, 216, 230, 231).

307. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38,** e164–e164 (2010) (cited on pages 51, 87).

308. Mägi, R. & Morris, A. P. GWAMA: Software for genome-wide association meta-analysis. *BMC Bioinformatics* **11,** 1–6 (2010) (cited on page 51).

309. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM codes to PheCodes: Workflow development and initial evaluation. *JMIR Medical Informatics* **7,** e14325 (2019) (cited on pages 52, 89, 131).

310. Dufour, D. R. *et al.* Diagnosis and monitoring of hepatic injury. II. Recommendations for use of laboratory tests in screening, diagnosis, and monitoring. *Clinical Chemistry* **46,** 2050–2068 (2000) (cited on pages 57, 68).

311. Van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* **45,** 1–67 (2011) (cited on page 58).

312. Jacquemont, S. *et al.* A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. *The American Journal of Human Genetics* **94,** 415–425 (2014) (cited on pages 59, 219).

313. Gilman, S. R. *et al.* Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* **70,** 898–907 (2011) (cited on pages 59, 239).

314. Levy, D. *et al.* Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70,** 886–897 (2011) (cited on pages 59, 239).

315. Sanders, S. J. *et al.* Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87,** 1215–1233 (2015) (cited on pages 59, 206, 239).

316. Rao, E. *et al.* Pseudoautosomal deletions encompassing a novel homeobox gene cause growth failure in idiopathic short stature and Turner syndrome. *Nature Genetics* **16,** 54–63 (1997) (cited on pages 62, 235).

317. Ellison, J. W. *et al.* PHOG, a candidate gene for involvement in the short stature of Turner syndrome. *Human Molecular Genetics* **6,** 1341–1347 (1997) (cited on pages 62, 235).

318. Fukami, M., Seki, A. & Ogata, T. SHOX haploinsufficiency as a cause of syndromic and nonsyndromic short stature. *Molecular Syndromology* **7,** 3–11 (2016) (cited on page 62).

319. Schiller, S. *et al.* Phenotypic variation and genetic heterogeneity in Leri-Weill syndrome. *European Journal of Human Genetics* **8,** 54–62 (2000) (cited on page 62).

320. Gaillard, F. *Madelung deformity | Radiology Reference Article | Radiopaedia.org — radiopaedia.org* https://radiopaedia.org/articles/7582 (cited on page 62).

321. Mefford, H. C. *et al.* Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *New England Journal of Medicine* **359,** 1685–1699 (2008) (cited on page 63).

322. Bernier, R. *et al.* Clinical phenotype of the recurrent 1q21.1 copy-number variant. *Genetics in Medicine* **18,** 341–349 (2016) (cited on page 63).

323. Brunetti-Pierri, N. *et al.* Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nature Genetics* **40,** 1466–1471 (2008) (cited on page 63).

324. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010) (cited on pages 64, 86).

325. Bochukova, E. G. *et al.* Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* **463,** 666–670 (2010) (cited on pages 64, 106, 150, 175, 176, 181, 197, 208).

326. Jacquemont, S. *et al.* Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* **478,** 97–102 (2011) (cited on pages 64, 107, 175, 176, 181, 197, 198, 209, 215, 239).

327. Bachmann-Gagescu, R. *et al.* Recurrent 200-kb deletions of 16p11.2 that include the SH2B1 gene are associated with developmental delay and obesity. *Genetics in Medicine* **12,** 641–647 (2010) (cited on pages 64, 150, 156).

328. Männik, K. *et al.* Leveraging biobank-scale rare and common variant analyses to identify ASPHD1 as the main driver of reproductive traits in the 16p11.2 locus. *BioRxiv,* 716415 (2019) (cited on pages 64, 196, 211, 215).

329. Kargi, A. Y. & Merriam, G. R. Diagnosis and treatment of growth hormone deficiency in adults. *Nature Reviews Endocrinology* **9,** 335–345 (2013) (cited on page 64).

330. Andrews, N. C. Genes determining blood cell traits. *Nature Genetics* **41,** 1161–1162 (2009) (cited on page 64).

331. Ganesh, S. K. *et al.* Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nature Genetics* **41,** 1191–1198 (2009) (cited on page 64).

332. Aguirre, G., De Ita, J. R., De La Garza, R. & Castilla-Cortazar, I. Insulin-like growth factor-1 deficiency and metabolic syndrome. *Journal of Translational Medicine* **14,** 1–23 (2016) (cited on page 67).

333. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nature Genetics* **53,** 1415–1424 (2021) (cited on page 67).

334. Rudd, M. K. *et al.* Segmental duplications mediate novel, clinically relevant chromosome rearrangements. *Human Molecular Genetics* **18,** 2957–2962 (2009) (cited on page 67).

335. Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nature Genetics* **43,** 838–846 (2011) (cited on pages 67, 86, 87, 122, 175, 202).

336. Yu, H. *et al.* A recurrent 1.71 Mb genomic imbalance at 2q13 increases the risk of developmental delay and dysmorphism. *Clinical Genetics* **81,** 257–264 (2012) (cited on page 67).

337. Riley, K. N. *et al.* Recurrent deletions and duplications of chromosome 2q11.2 and 2q13 are associated with variable outcomes. *American Journal of Medical Genetics Part A* **167,** 2664–2673 (2015) (cited on page 67).

338. Wolfe, K. *et al.* Delineating the psychiatric and behavioral phenotype of recurrent 2q13 deletions and duplications. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **177,** 397–405 (2018) (cited on page 67).

339. De Bruyne, E. *et al.* IGF-1 suppresses Bim expression in multiple myeloma via epigenetic and posttranslational mechanisms. *Blood* **115,** 2430–2440 (2010) (cited on page 67).

340. Anzai, N. *et al.* The multivalent PDZ domain-containing protein PDZK1 regulates transport activity of renal urate-anion exchanger URAT1 via its C terminus. *Journal of Biological Chemistry* **279,** 45942–45950 (2004) (cited on page 67).

341. Kolz, M. *et al.* Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. *PLoS Genetics* **5,** e1000504 (2009) (cited on pages 67, 69).

342. Köttgen, A. *et al.* Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nature Genetics* **45,** 145–154 (2013) (cited on pages 67, 69).

343. Yang, Q. *et al.* Multiple genetic loci influence serum urate levels and their relationship with gout and cardiovascular disease risk factors. *Circulation: Cardiovascular Genetics* **3,** 523–530 (2010) (cited on page 67).

344. Sulem, P. *et al.* Identification of low-frequency variants associated with gout and serum uric acid levels. *Nature Genetics* **43,** 1127–1130 (2011) (cited on page 67).

345. Ketharnathan, S. *et al.* A non-coding genetic variant maximally associated with serum urate levels is functionally linked to HNF4A-dependent PDZK1 expression. *Human Molecular Genetics* **27,** 3964–3973 (2018) (cited on pages 67, 69, 240).

346. Yuan, X. *et al.* Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *The American Journal of Human Genetics* **83,** 520–528 (2008) (cited on pages 68, 70).

347. Kamatani, Y. *et al.* Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nature Genetics* **42,** 210–215 (2010) (cited on pages 68, 70).

348. Chambers, J. C. *et al.* Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nature Genetics* **43,** 1131–1138 (2011) (cited on pages 68, 70).

349. Gurdasani, D. *et al.* Uganda genome resource enables insights into population history and genomic discovery in Africa. *Cell* **179,** 984–1002 (2019) (cited on page 68).

350. Seo, J. Y. *et al.* A genome-wide association study on liver enzymes in Korean population. *PLoS One* **15,** e0229374 (2020) (cited on page 68).

351. Pazoki, R. *et al.* Genetic analysis in European ancestry individuals identifies 517 loci associated with liver enzymes. *Nature Communications* **12,** 2579 (2021) (cited on page 68).

352. Vuckovic, D. *et al.* The polygenic and monogenic basis of blood traits and diseases. *Cell* **182,** 1214–1231 (2020) (cited on page 68).

353. Astle, W. J. *et al.* The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167,** 1415–1429 (2016) (cited on page 68).

354. Nash, R. & Shojania, A. Hematological aspect of Rh deficiency syndrome: A case report and a review of the literature. *American Journal of Hematology* **24,** 267–275 (1987) (cited on page 68).

355. Rai, D., Wilson, A. M. & Moosavi, L. *Histology, Reticulocytes* (StatPearls Publishing, Treasure Island (FL), 2019) (cited on page 68).

356. Goldstein, D. E. *et al.* Tests of glycemia in diabetes. *Diabetes Care* **27,** 1761–1773 (2004) (cited on page 69).

357. Akinlaja, O. Hematological changes in pregnancy-The preparation for intrapartum blood loss. *International Journal of Gynecology & Obstetrics* **4,** 00109 (2016) (cited on page 69).

358. Sanna, S. *et al.* Common variants in the SLCO1B3 locus are associated with bilirubin levels and unconjugated hyperbilirubinemia. *Human Molecular Genetics* **18,** 2711–2718 (2009) (cited on page 72).

359. Johnson, A. D. *et al.* Genome-wide association meta-analysis for total serum bilirubin levels. *Human Molecular Genetics* **18,** 2700–2710 (2009) (cited on page 72).

360. Kang, T.-W. *et al.* Genome-wide association of serum bilirubin levels in Korean population. *Human Molecular Genetics* **19,** 3672–3678 (2010) (cited on page 72).

361. Bielinski, S. J. *et al.* Mayo Genome Consortia: A genotype-phenotype resource for genome-wide association studies with an application to the analysis of circulating bilirubin levels. *Mayo Clinic Proceedings* **86,** 606–614 (2011) (cited on page 72).

362. Dai, X. *et al.* A genome-wide association study for serum bilirubin levels and gene-environment interaction in a Chinese population. *Genetic Epidemiology* **37,** 293–300 (2013) (cited on page 72).

363. Smith, N. F., Figg, W. D. & Sparreboom, A. Role of the liver-specific transporters OATP1B1 and OATP1B3 in governing drug elimination. *Expert Opinion on Drug Metabolism & Toxicology* **1,** 429–445 (2005) (cited on page 72).

364. Mitchel, M. W. *et al. 17q12 Recurrent Deletion Syndrome* (GeneReviews, Seattle (WA), 2020) (cited on page 73).

365. Mefford, H. *17q12 Recurrent Duplication* (GeneReviews, Seattle (WA), 2021) (cited on page 73).

366. Van Paassen, B. W. *et al.* PMP22 related neuropathies: Charcot-Marie-Tooth disease type 1A and hereditary neuropathy with liability to pressure palsies. *Orphanet Journal of Rare Diseases* **9,** 1–15 (2014) (cited on page 73).

367. Horowitz, G. L. & Staros, E. B. *Creatinine: Reference Range, Interpretation, Collection and Panels — emedicine.medscape.com* 2019. https://emedicine.medscape.com/article/2054342-overview#a2?form=fpf (cited on page 73).

368. Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. Where genotype is not predictive of phenotype: Towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human Genetics* **132,** 1077–1130 (2013) (cited on pages 73, 79).

369. Day, F. R. *et al.* Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nature Genetics* **49,** 834–841 (2017) (cited on pages 76, 81).

370. Gasner, A. & Rehman, A. *Primary Amenorrhea* (StatPearls Publishing, Treasure Island (FL), 2021) (cited on page 76).

371. Walker, M. H. & Tobler, K. J. *Female Infertility* (StatPearls Publishing, Treasure Island (FL), 2021) (cited on page 76).

372. Su, Y.-Q. *et al.* MARF1 regulates essential oogenic processes in mice. *Science* **335,** 1496–1499 (2012) (cited on pages 76, 77).

373. Kawaguchi, S., Ueki, M. & Kai, T. Drosophila MARF1 ensures proper oocyte maturation by regulating nanos expression. *PLoS One* **15,** e0231114 (2020) (cited on pages 76, 77).

374. Islam, R. *et al.* Genome-wide runs of homozygosity, effective population size, and detection of positive selection signatures in six Chinese goat breeds. *Genes* **10,** 938 (2019) (cited on pages 76, 77).

375. Katari, S. *et al.* Chromosomal instability in women with primary ovarian insufficiency. *Human Reproduction* **33,** 531–538 (2018) (cited on pages 76, 77).

376. Yang, X. *et al.* Gene variants identified by whole-exome sequencing in 33 French women with premature ovarian insufficiency. *Journal of Assisted Reproduction and Genetics* **36,** 39–45 (2019) (cited on pages 76, 77).

377. Su, Y.-Q., Sun, F., Handel, M. A., Schimenti, J. C. & Eppig, J. J. Meiosis arrest female 1 (MARF1) has nuage-like function in mammalian oocytes. *Proceedings of the National Academy of Sciences* **109,** 18653–18660 (2012) (cited on page 77).

378. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266,** 66–71 (1994) (cited on pages 77, 108).

379. McNally, E. J., Luncsford, P. J. & Armanios, M. Long telomeres and cancer risk: The price of cellular immortality. *The Journal of Clinical Investigation* **129,** 3474–3481 (2019) (cited on page 77).

380. Fisher, R. A. XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* **52,** 399–433 (1918) (cited on page 78).

381. Fahed, A. C. *et al.* Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nature Communications* **11,** 3635 (2020) (cited on page 79).

382. Dauber, A. *et al.* Genome-wide association of copy-number variation reveals an association between short stature and the presence of low-frequency genomic deletions. *The American Journal of Human Genetics* **89,** 751–759 (2011) (cited on pages 79, 122).

383. Wheeler, E. *et al.* Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. *Nature Genetics* **45,** 513–517 (2013) (cited on pages 79, 122).

384. Männik, K. *et al.* Copy number variations and cognitive phenotypes in unselected populations. *JAMA* **313,** 2044–2054 (2015) (cited on pages 79, 122, 165, 200, 202).

385. Saarentaus, E. C. *et al.* Polygenic burden has broader impact on health, cognition, and socioeconomic outcomes than most rare and high-risk copy number variants. *Molecular Psychiatry* **26,** 4884–4895 (2021) (cited on pages 79, 122, 199).

386. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316,** 445–449 (2007) (cited on pages 79, 86, 87, 122, 175, 197, 206).

387. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320,** 539–543 (2008) (cited on pages 79, 86, 87, 122, 176, 180, 197).

388. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455,** 237–241 (2008) (cited on page 79).

389. Mefford, H. C. *et al.* Genome-wide copy number variation in epilepsy: Novel susceptibility loci in idiopathic generalized and focal epilepsies. *PLoS Genetics* **6,** e1000962 (2010) (cited on pages 79, 86, 87, 112, 122, 176, 197).

390. Mefford, H. C. *et al.* Rare copy number variants are an important cause of epileptic encephalopathies. *Annals of Neurology* **70,** 974–985 (2011) (cited on pages 79, 205).

391. McDaid, A. F. *et al.* Bayesian association scan reveals loci associated with human lifespan and linked biomarkers. *Nature Communications* **8,** 15842 (2017) (cited on page 79).

392. Freathy, R. M. *et al.* Common variation in the FTO gene alters diabetes-related metabolic traits to the extent expected given its effect on BMI. *Diabetes* **57,** 1419–1426 (2008) (cited on page 80).

393. Würtz, P. *et al.* Metabolic signatures of adiposity in young adults: Mendelian randomization analysis and effects of weight change. *PLoS Medicine* **11,** e1001765 (2014) (cited on page 80).

394. Fall, T. *et al.* The role of adiposity in cardiometabolic traits: A Mendelian randomization analysis. *PLoS Medicine* **10,** e1001474 (2013) (cited on page 80).

395. Barker, D. J. *et al.* Type 2 (non-insulin-dependent) diabetes mellitus, hypertension and hyperlipidaemia (syndrome X): Relation to reduced fetal growth. *Diabetologia* **36,** 62–67 (1993) (cited on page 80).

396. Armengaud, J., Yzydorczyk, C., Siddeek, B., Peyter, A. & Simeoni, U. Intrauterine growth restriction: Clinical consequences on health and disease at adulthood. *Reproductive Toxicology* **99,** 168–176 (2021) (cited on page 80).

397. Halvorsen, M. *et al.* Increased burden of ultra-rare structural variants localizing to boundaries of topologically associated domains in schizophrenia. *Nature Communications* **11,** 1842 (2020) (cited on pages 81, 86).

398. Chen, L. *et al.* Association of structural variation with cardiometabolic traits in Finns. *The American Journal of Human Genetics* **108,** 583–596 (2021) (cited on pages 81, 86).

399. Li, J. *et al.* Whole genome distribution and ethnic differentiation of copy number variation in Caucasian and Asian populations. *PloS One* **4,** e7958 (2009) (cited on page 81).

400. Campbell, C. D. *et al.* Population-genetic properties of differentiated human copy-number polymorphisms. *The American Journal of Human Genetics* **88,** 317–332 (2011) (cited on page 81).

401. Chen, W. *et al.* Copy number variation across European populations. *PLoS One* **6,** e23087 (2011) (cited on page 81).

402. Martin, C. L. *et al.* Identification of neuropsychiatric copy number variants in a health care system population. *JAMA Psychiatry* **77,** 1276–1285 (2020) (cited on pages 81, 176, 199, 208, 217, 225, 241).

403. Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics* **10,** 451–481 (2009) (cited on page 86).

404. Kopal, J. *et al.* Rare CNVs and phenome-wide profiling highlight brain structural divergence and phenotypical convergence. *Nature Human Behaviour,* 1–17 (2023) (cited on pages 86, 204).

405. Senn, S. *Statistical issues in drug development* (Wiley, Chichester, 2021) (cited on page 86).

406. Mollon, J. *et al.* Impact of Copy Number Variants and Polygenic Risk Scores on Psychopathology in the UK Biobank. *Biological Psychiatry* **94,** 591–600 (2023) (cited on page 86).

407. Vaez, M. *et al.* Population-based Risk of Psychiatric Disorders Associated with Recurrent CNVs. *medRxiv* (2023) (cited on pages 86, 207, 208).

408. Sánchez, X. C. *et al.* Comparing copy number variations in a Danish case cohort of individuals with psychiatric disorders. *JAMA Psychiatry* **79,** 59–69 (2022) (cited on pages 86, 198, 199, 208).

409. Verbitsky, M. *et al.* Genomic disorders in CKD across the lifespan. *Journal of the American Society of Nephrology* **34,** 607–618 (2023) (cited on pages 86, 110, 215).

410. Montanucci, L. *et al.* Genome-wide identification and phenotypic characterization of seizure-associated copy number variations in 741,075 individuals. *Nature Communications* **14,** 4392 (2023) (cited on pages 86, 112, 198, 205).

411. Zamariolli, M. *et al.* The impact of 22q11.2 copy-number variants on human traits in the general population. *The American Journal of Human Genetics* **110,** 300–313 (2023) (cited on pages 86, 114, 122, 127).

412. Hanssen, R. *et al.* Chromosomal deletions on 16p11.2 encompassing SH2B1 are associated with accelerated metabolic disease. *Cell Reports Medicine* **4,** 101155 (2023) (cited on pages 86, 106, 122, 147, 210).

413. Vysotskiy, M. *et al.* Integration of genetic, transcriptomic, and clinical data provides insight into 16p11.2 and 22q11.2 CNV genes. *Genome Medicine* **13,** 1–26 (2021) (cited on pages 86, 120, 203, 206, 212–214).

414. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Research* **12,** 996–1006 (2002) (cited on pages 87, 151).

415. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010) (cited on page 89).

416. Therneau, T. M. *A Package for Survival Analysis in R* R package version 3.5-3 (2022) (cited on pages 95, 154).

417. Privé, F. Using the UK Biobank as a global reference of worldwide populations: Application to measuring ancestry diversity from GWAS summary statistics. *Bioinformatics* **38,** 3477–3480 (2022) (cited on page 95).

418. Levey, A. S. *et al.* A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine* **150,** 604–612 (2009) (cited on page 97).

419. Li, A. *et al.* Bietti crystalline corneoretinal dystrophy is caused by mutations in the novel gene CYP4V2. *The American Journal of Human Genetics* **74,** 817–826 (2004) (cited on page 105).

420. Zhou, W. *et al.* FAN1 mutations cause karyomegalic interstitial nephritis, linking chronic kidney failure to defective DNA damage repair. *Nature Genetics* **44,** 910–915 (2012) (cited on page 105).

421. Shinawi, M. *et al.* Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioral problems, dysmorphism, epilepsy, and abnormal head size. *Journal of Medical Genetics* (2010) (cited on pages 107, 150, 175, 176, 197, 202–204, 206, 207, 212–215, 217).

422. Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *New England Journal of Medicine* **358,** 667–675 (2008) (cited on pages 107, 150, 175, 197, 206, 208).

423. D'Angelo, D. *et al.* Defining the effect of the 16p11.2 duplication on cognition, behavior, and medical comorbidities. *JAMA Psychiatry* **73,** 20–30 (2016) (cited on pages 107, 176, 180, 183, 202–206, 208, 209, 211–214, 216, 217).

424. Reinthaler, E. M. *et al.* 16p11.2 600 kb Duplications confer risk for typical and atypical Rolandic epilepsy. *Human Molecular Genetics* **23,** 6069–6080 (2014) (cited on pages 107, 204, 205).

425. McCarthy, S. E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nature Genetics* **41,** 1223–1227 (2009) (cited on pages 107, 175, 176, 180, 197, 203, 207, 213).

426. Walters, R. G. *et al.* A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature* **463,** 671–675 (2010) (cited on pages 107, 150, 175, 176, 181, 191, 197, 208, 209).

427. Defesche, J. C. *et al.* Familial hypercholesterolaemia. *Nature Reviews Disease Primers* **3,** 1–20 (2017) (cited on page 109).

428. Iacocca, M. A. *et al.* Use of next-generation sequencing to detect LDLR gene copy number variation in familial hypercholesterolemia. *Journal of Lipid Research* **58,** 2202–2209 (2017) (cited on page 109).

429. Mach, F. *et al.* 2019 ESC/EAS Guidelines for the management of dyslipidaemias: Lipid modification to reduce cardiovascular risk: The Task Force for the management of dyslipidaemias of the European Society of Cardiology (ESC) and European Atherosclerosis Society (EAS). *European Heart Journal* **41,** 111–188 (2020) (cited on page 110).

430. Fajans, S. S., Bell, G. I. & Polonsky, K. S. Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young. *New England Journal of Medicine* **345,** 971–980 (2001) (cited on page 110).

431. Mefford, H. C. *et al.* Recurrent reciprocal genomic rearrangements of 17q12 are associated with renal disease, diabetes, and epilepsy. *The American Journal of Human Genetics* **81,** 1057–1069 (2007) (cited on page 110).

432. Wuttke, M. *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature Genetics* **51,** 957–972 (2019) (cited on page 111).

433. Stanzick, K. J. *et al.* Discovery and prioritization of variants and genes for kidney function in > 1.2 million individuals. *Nature Communications* **12,** 4350 (2021) (cited on page 111).

434. Girirajan, S. *et al.* A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nature Genetics* **42,** 203–209 (2010) (cited on page 111).

435. Stefansson, H. *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* **505,** 361–366 (2014) (cited on pages 111, 165, 200, 202, 211, 219).

436. Girirajan, S., Pizzo, L., Moeschler, J. & Rosenfeld, J. *16p12.2 Recurrent Deletion* (GeneReviews, Seattle (WA), 2018) (cited on page 111).

437. International League Against Epilepsy Consortium on Complex Epilepsies. GWAS meta-analysis of over 29,000 people with epilepsy identifies 26 risk loci and subtype-specific genetic architecture. *Nature Genetics* **55,** 1471–1482 (2023) (cited on page 113).

438. Howles, S. A. *et al.* Genetic variants of calcium and vitamin D metabolism in kidney stone disease. *Nature Communications* **10,** 5175 (2019) (cited on page 113).

439. Evangelou, E. *et al.* Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nature Genetics* **50,** 1412–1425 (2018) (cited on page 113).

440. De Kovel, C. G. *et al.* Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. *Brain* **133,** 23–32 (2010) (cited on page 112).

441. Heinzen, E. L. *et al.* Rare deletions at 16p13.11 predispose to a diverse spectrum of sporadic epilepsy syndromes. *The American Journal of Human Genetics* **86,** 707–718 (2010) (cited on page 112).

442. Alkuraya, F. S. *et al.* Human Mutations in NDE1 cause extreme microcephaly with lissencephaly. *The American Journal of Human Genetics* **88,** 536–547 (2011) (cited on page 113).

443. Bakircioglu, M. *et al.* The essential role of centrosomal NDE1 in human cerebral cortex neurogenesis. *The American Journal of Human Genetics* **88,** 523–535 (2011) (cited on page 113).

444. Ringpfeil, F., Lebwohl, M. G., Christiano, A. M. & Uitto, J. Pseudoxanthoma elasticum: Mutations in the MRP6 gene encoding a transmembrane ATP-binding cassette (ABC) transporter. *Proceedings of the National Academy of Sciences* **97,** 6001–6006 (2000) (cited on page 114).

445. Struk, B. *et al.* Mutations of the gene encoding the transmembrane transporter protein ABC-C6 cause pseudoxanthoma elasticum. *Journal of Molecular Medicine* **78,** 282–286 (2000) (cited on page 114).

446. Le Saux, O. *et al.* Mutations in a gene encoding an ABC transporter cause pseudoxanthoma elasticum. *Nature Genetics* **25,** 223–227 (2000) (cited on page 114).

447. Bergen, A. A. *et al.* Mutations in ABCC6 cause pseudoxanthoma elasticum. *Nature Genetics* **25,** 228–231 (2000) (cited on page 114).

448. Ralph, D. *et al.* Kidney stones are prevalent in individuals with pseudoxanthoma elasticum, a genetic ectopic mineralization disorder. *International Journal of Dermatology and Venereology* **3,** 198–204 (2020) (cited on pages 114, 122).

449. Legrand, A. *et al.* Mutation spectrum in the ABCC6 gene and genotype–phenotype correlations in a French cohort with pseudoxanthoma elasticum. *Genetics in Medicine* **19,** 909–917 (2017) (cited on pages 114, 122).

450. Letavernier, E. *et al.* ABCC6 deficiency promotes development of Randall plaque. *Journal of the American Society of Nephrology* **29,** 2337–2347 (2018) (cited on pages 114, 122).

451. Nitschke, Y. *et al.* Generalized arterial calcification of infancy and pseudoxanthoma elasticum can be caused by mutations in either ENPP1 or ABCC6. *The American Journal of Human Genetics* **90,** 25–39 (2012) (cited on page 114).

452. Le Boulanger, G. *et al.* An unusual severe vascular case of pseudoxanthoma elasticum presenting as generalized arterial calcification of infancy. *American Journal of Medical Genetics Part A* **152,** 118–123 (2010) (cited on page 114).

453. McDonald-McGinn, D. M. *et al.* 22q11.2 deletion syndrome. *Nature Reviews Disease Primers* **1,** 1–19 (2015) (cited on pages 114, 115, 129, 143).

454. Bartik, L. E. *et al.* 22q11.2 duplications: Expanding the clinical presentation. *American Journal of Medical Genetics Part A* **188,** 779–787 (2022) (cited on page 115).

455. Sharp, A. J. *et al.* A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nature Genetics* **40,** 322–328 (2008) (cited on page 115).

456. Lowther, C. *et al.* Delineating the 15q13.3 microdeletion phenotype: A case series and comprehensive review of the literature. *Genetics in Medicine* **17,** 149–157 (2015) (cited on page 115).

457. Gillentine, M. A. & Schaaf, C. P. The human clinical phenotypes of altered CHRNA7 copy number. *Biochemical Pharmacology* **97,** 352–362 (2015) (cited on page 115).

458. Golzio, C. & Katsanis, N. Genetic architecture of reciprocal CNVs. *Current Opinion in Genetics & Development* **23,** 240–248 (2013) (cited on pages 120, 224).

459. Kryger, M., Roth, T. & Goldstein, C. A. *Principles and Practice of Sleep Medicine* (Elsevier, Philadelphia (PA), 2021) (cited on page 121).

460. Of Us Research Program Investigators, A. The "All of Us" research program. *New England Journal of Medicine* **381,** 668–676 (2019) (cited on page 123).

461. Hunter-Zinck, H. *et al.* Genotyping array design and data quality control in the Million Veteran Program. *The American Journal of Human Genetics* **106,** 535–548 (2020) (cited on pages 123, 222).

462. Monteiro, F. P. *et al.* Defining new guidelines for screening the 22q11.2 deletion based on a clinical and dysmorphologic evaluation of 194 individuals and review of the literature. *European Journal of Pediatrics* **172,** 927–945 (2013) (cited on page 129).

463. Portnoï, M.-F. Microduplication 22q11.2: A new chromosomal syndrome. *European Journal of Medical Genetics* **52,** 88–93 (2009) (cited on page 130).

464. Verbesselt, J., Zink, I., Breckpot, J. & Swillen, A. Cross-sectional and longitudinal findings in patients with proximal 22q11.2 duplication: A retrospective chart study. *American Journal of Medical Genetics Part A* **188,** 46–57 (2022) (cited on pages 130, 143).

465. Yobb, T. M. *et al.* Microduplication and triplication of 22q11.2: A highly variable syndrome. *The American Journal of Human Genetics* **76,** 865–876 (2005) (cited on page 130).

466. Marshall, C. R. *et al.* Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature Genetics* **49,** 27–35 (2017) (cited on pages 130, 198, 207).

467. Lin, A. *et al.* Reciprocal copy number variations at 22q11.2 produce distinct and convergent neurobehavioral impairments relevant for schizophrenia and autism spectrum disorder. *Biological Psychiatry* **88,** 260–272 (2020) (cited on page 130).

468. Lin, A. *et al.* Mapping 22q11.2 gene dosage effects on brain morphometry. *Journal of Neuroscience* **37,** 6183–6199 (2017) (cited on page 130).

469. Savoia, A. *et al.* Spectrum of the Mutations in Bernard-Soulier Syndrome. *Human Mutation* **35,** 1033–1045 (2014) (cited on page 130).

470. Nunes, N. *et al.* CEDNIK syndrome in a Brazilian patient with compound heterozygous pathogenic variants. *European Journal of Medical Genetics* **65,** 104440 (2022) (cited on page 130).

471. Köhler, S. *et al.* The human phenotype ontology in 2021. *Nucleic Acids Research* **49,** D1207–D1217 (2021) (cited on pages 130, 142).

472. Bastarache, L. *et al.* Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* **359,** 1233–1239 (2018) (cited on page 131).

473. Voll, S. L. *et al.* Obesity in adults with 22q11.2 deletion syndrome. *Genetics in Medicine* **19,** 204–208 (2017) (cited on page 143).

474. Loid, P. *et al.* Targeted exome sequencing of genes involved in rare CNVs in early-onset severe obesity. *Frontiers in Genetics* **13,** 839349 (2022) (cited on page 143).

475. Campbell, I. M. *et al.* Platelet findings in 22q11.2 deletion syndrome correlate with disease manifestations but do not correlate with GPIb surface expression. *Clinical Genetics* **103,** 109–113 (2023) (cited on page 143).

476. Hierck, B. P. *et al.* A chicken model for DGCR6 as a modifier gene in the DiGeorge critical region. *Pediatric Research* **56,** 440–448 (2004) (cited on page 143).

477. Yu, A. *et al.* Genotypic and phenotypic variability of 22q11.2 microduplications: An institutional experience. *American Journal of Medical Genetics Part A* **179,** 2178–2189 (2019) (cited on page 143).

478. Courtens, W., Schramme, I. & Laridon, A. Microduplication 22q11.2: A benign polymorphism or a syndrome with a very large clinical variability and reduced penetrance?—Report of two families. *American Journal of Medical Genetics Part A* **146,** 758–763 (2008) (cited on page 143).

479. Zhang, X. *et al.* Local and global chromatin interactions are altered by large genomic deletions associated with human brain development. *Nature Communications* **9,** 5356 (2018) (cited on page 143).

480. Tucker, T. *et al.* Prevalence of selected genomic deletions and duplications in a French Canadian population-based sample of newborns. *Molecular Genetics & Genomic Medicine* **1,** 87–97 (2013) (cited on page 144).

481. Smajlagić, D. *et al.* Population prevalence and inheritance pattern of recurrent CNVs associated with neurodevelopmental disorders in 12,252 newborns and their parents. *European Journal of Human Genetics* **29,** 205–215 (2021) (cited on pages 144, 199).

482. Kirov, G. *et al.* The penetrance of copy number variations for schizophrenia and developmental delay. *Biological Psychiatry* **75,** 378–385 (2014) (cited on page 144).

483. Olsen, L. *et al.* Prevalence of rearrangements in the 22q11.2 region and population-based risk of neuropsychiatric and developmental disorders in a Danish population: A case-cohort study. *The Lancet Psychiatry* **5,** 573–580 (2018) (cited on page 144).

484. Baldini, G. & Phelan, K. D. The melanocortin pathway and control of appetite-progress and therapeutic implications. *Journal of Endocrinology* **241,** R1–R33 (2019) (cited on page 148).

485. Heymsfield, S. B. & Wadden, T. A. Mechanisms, pathophysiology, and management of obesity. *New England Journal of Medicine* **376,** 254–266 (2017) (cited on page 149).

486. Van der Klaauw, A. A. & Farooqi, I. S. The hunger genes: Pathways to obesity. *Cell* **161,** 119–132 (2015) (cited on page 150).

487. Clément, K. *et al.* MC4R agonism promotes durable weight loss in patients with leptin receptor deficiency. *Nature Medicine* **24,** 551–555 (2018) (cited on page 150).

488. Clément, K. *et al.* Efficacy and safety of setmelanotide, an MC4R agonist, in individuals with severe obesity due to LEPR or POMC deficiency: Single-arm, open-label, multicentre, phase 3 trials. *The Lancet Diabetes & Endocrinology* **8,** 960–970 (2020) (cited on pages 150, 170).

489. Kühnen, P. *et al.* Proopiomelanocortin deficiency treated with a melanocortin-4 receptor agonist. *New England Journal of Medicine* **375,** 240–246 (2016) (cited on page 150).

490. Maures, T. J., Kurzer, J. H. & Carter-Su, C. SH2B1 (SH2-B) and JAK2: A multifunctional adaptor protein and kinase made for each other. *Trends in Endocrinology & Metabolism* **18,** 38–45 (2007) (cited on page 150).

491. Ren, D., Li, M., Duan, C. & Rui, L. Identification of SH2-B as a key regulator of leptin sensitivity, energy balance, and body weight in mice. *Cell Metabolism* **2,** 95–104 (2005) (cited on pages 150, 168).

492. Duan, C., Yang, H., White, M. F. & Rui, L. Disruption of the SH2-B gene causes age-dependent insulin resistance and glucose intolerance. *Molecular and Cellular Biology* **24,** 7435–7443 (2004) (cited on pages 150, 168).

493. Li, M., Ren, D., Iseki, M., Takaki, S. & Rui, L. Differential role of SH2-B and APS in regulating energy and glucose homeostasis. *Endocrinology* **147,** 2163–2170 (2006) (cited on pages 150, 168).

494. Flores, A. *et al.* Crucial role of the SH2B1 PH domain for the control of energy balance. *Diabetes* **68,** 2049–2062 (2019) (cited on pages 150, 169).

495. Pearce, L. R. *et al.* Functional characterization of obesity-associated variants involving the α and β isoforms of human SH2B1. *Endocrinology* **155,** 3219–3226 (2014) (cited on page 150).

496. Doche, M. E. *et al.* Human SH2B1 mutations are associated with maladaptive behaviors and obesity. *The Journal of Clinical Investigation* **122,** 4732–4736 (2012) (cited on pages 150, 168).

497. Lupski, J. R. Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics* **14,** 417–422 (1998) (cited on page 150).

498. Steinman, K. J. *et al.* 16p11.2 deletion and duplication: Characterizing neurologic phenotypes in a large clinically ascertained cohort. *American Journal of Medical Genetics Part A* **170,** 2943–2955 (2016) (cited on pages 150, 202–206, 208, 213, 217).

499. Rosenfeld, J. A., Coe, B. P., Eichler, E. E., Cuckle, H. & Shaffer, L. G. Estimates of penetrance for recurrent pathogenic copy-number variations. *Genetics in Medicine* **15,** 478–481 (2013) (cited on pages 150, 197–199).

500. Zufferey, F. *et al.* A 600 kb deletion syndrome at 16p11.2 leads to energy imbalance and neuropsychiatric disorders. *Journal of Medical Genetics* **49,** 660 (2012) (cited on pages 150, 176, 181, 183, 191, 202–204, 206, 208, 209, 211–214, 217, 220).

501. Guha, S. *et al.* Implication of a rare deletion at distal 16p11.2 in schizophrenia. *JAMA Psychiatry* **70,** 253–260 (2013) (cited on page 150).

502. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Research* **11,** 1005–1017 (2001) (cited on pages 153, 194).

503. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297,** 1003–1007 (2002) (cited on page 153).

504. Field, Z., Miles, J. & Field, A. Discovering statistics using R. *Discovering Statistics Using R,* 1–992 (2012) (cited on page 156).

505. Lenth, R. V. Least-squares means: The R package lsmeans. *Journal of Statistical Software* **69,** 1–33 (2016) (cited on page 156).

506. Lenth, R. V. *emmeans: Estimated Marginal Means, aka Least-Squares Means* R package version 1.10.0 (2024) (cited on page 156).

507. Taylor, C. M. *et al. 16p11.2 Recurrent Deletion* (GeneReviews, Seattle (WA), 2021) (cited on pages 156, 225).

508. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851 (2007) (cited on page 157).

509. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics* **50,** 1505–1513 (2018) (cited on pages 157, 167).

510. Walters, R. G. *et al.* Rare genomic structural variants in complex disease: Lessons from the replication of associations with obesity. *PloS One* **8,** e58048 (2013) (cited on page 158).

511. Floegel, A. *et al.* Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. *Diabetes* **62,** 639–648 (2013) (cited on page 164).

512. Carlsson, E. R., Grundtvig, J. L. G., Madsbad, S. & Fenger, M. Changes in serum sphingomyelin after Roux-en-Y gastric bypass surgery are related to diabetes status. *Frontiers in Endocrinology* **9,** 172 (2018) (cited on page 164).

513. Roman-Urrestarazu, A. *et al.* Association of race/ethnicity and social disadvantage with autism prevalence in 7 million school children in England. *JAMA Pediatrics* **175,** e210054–e210054 (2021) (cited on page 165).

514. Styne, D. M. *et al.* Pediatric obesity—assessment, treatment, and prevention: An Endocrine Society clinical practice guideline. *The Journal of Clinical Endocrinology & Metabolism* **102,** 709–757 (2017) (cited on page 168).

515. Key, J. *et al.* Mid-gestation lethality of atxn2l-ablated mice. *International Journal of Molecular Sciences* **21,** 5124 (2020) (cited on page 168).

516. Groza, T. *et al.* The International Mouse Phenotyping Consortium: Comprehensive knockout phenotyping underpinning the study of human disease. *Nucleic Acids Research* **51,** D1038–D1045 (2023) (cited on page 168).

517. He, M. *et al.* Spns1 is a lysophospholipid transporter mediating lysosomal phospholipid salvage. *Proceedings of the National Academy of Sciences* **119,** e2210353119 (2022) (cited on page 168).

518. Simonds, S. E. *et al.* Leptin mediates the increase in blood pressure associated with obesity. *Cell* **159,** 1404–1416 (2014) (cited on page 169).

519. Metz, M. *et al.* Leptin increases hepatic triglyceride export via a vagal mechanism in humans. *Cell Metabolism* **34,** 1719–1731 (2022) (cited on page 169).

520. Jiang, L. *et al.* Neural deletion of Sh2b1 results in brain growth retardation and reactive aggression. *The FASEB Journal* **32,** 1830 (2018) (cited on page 169).

521. Sonoyama, T. *et al.* Human BDNF/TrkB variants impair hippocampal synaptogenesis and associate with neurobehavioural abnormalities. *Scientific Reports* **10,** 9028 (2020) (cited on page 169).

522. Javadi, M. *et al.* The SH2B1 adaptor protein associates with a proximal region of the erythropoietin receptor. *Journal of Biological Chemistry* **287,** 26223–26234 (2012) (cited on page 169).

523. Bijlsma, E. *et al.* Extending the phenotype of recurrent rearrangements of 16p11.2: Deletions in mentally retarded patients without autism and in normal individuals. *European Journal of Medical Genetics* **52,** 77–87 (2009) (cited on pages 175, 176, 183, 197, 198, 202, 204, 208, 212–215).

524. Rosenfeld, J. A. *et al.* Speech delays and behavioral problems are the predominant features in individuals with developmental delays and 16p11.2 microdeletions and microduplications. *Journal of Neurodevelopmental Disorders* **2,** 26–38 (2010) (cited on pages 175, 176, 183, 197, 202, 204, 207, 211–217).

525. Kumar, R. A. *et al.* Recurrent 16p11.2 microdeletions in autism. *Human Molecular Genetics* **17,** 628–638 (2008) (cited on pages 175, 197, 206).

526. Marshall, C. R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *The American Journal of Human Genetics* **82,** 477–488 (2008) (cited on pages 175, 197, 202, 206, 208, 209).

527. Hanson, E. *et al.* The cognitive and behavioral phenotype of the 16p11.2 deletion in a clinically ascertained population. *Biological Psychiatry* **77,** 785–793 (2015) (cited on pages 176, 202, 207, 208).

528. Green Snyder, L. *et al.* Autism spectrum disorder, developmental and psychiatric features in 16p11.2 duplication. *Journal of Autism and Developmental Disorders* **46,** 2734–2748 (2016) (cited on pages 176, 180, 206–208).

529. Niarchou, M. *et al.* Psychiatric disorders in children with 16p11.2 deletion and duplication. *Translational Psychiatry* **9,** 8 (2019) (cited on pages 176, 180, 202, 206–208).

530. Birnbaum, R., Mahjani, B., Loos, R. J. & Sharp, A. J. Clinical characterization of copy number variants associated with neurodevelopmental disorders in a large-scale multiancestry biobank. *JAMA Psychiatry* **79,** 250–259 (2022) (cited on pages 176, 199, 200, 222).

531. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7,** e34408 (2018) (cited on page 176).

532. Karczewski, K. J. *et al.* Pan-UK Biobank GWAS improves discovery, analysis of genetic architecture, and resolution into ancestry-enriched effects. *medRxiv,* 2024–03 (2024) (cited on page 178).

533. Okbay, A. *et al.* Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics* **54,** 437–449 (2022) (cited on page 178).

534. UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526,** 82–90 (2015) (cited on page 178).

535. Moix, S., Sadler, M. C., Kutalik, Z. & Auwerx, C. Breaking down causes, consequences, and mediating effects of age-related telomere shortening on human health. *medRxiv,* 2024–01 (2024) (cited on page 179).

536. Gale, C. R. *et al.* Pleiotropy between neuroticism and physical and mental health: Findings from 108,038 men and women in UK Biobank. *Translational Psychiatry* **6,** e791–e791 (2016) (cited on page 181).

537. Tyrrell, J. *et al.* Height, body mass index, and socioeconomic status: Mendelian randomisation study in UK Biobank. *BMJ* **352,** i582 (2016) (cited on page 183).

538. Yu, Y. *et al.* Age-and gender-dependent obesity in individuals with 16p11.2 deletion. *Journal of Genetics and Genomics* **38,** 403–409 (2011) (cited on pages 191, 208, 209).

539. Gill, R., Chen, Q., D'Angelo, D. & Chung, W. K. Eating in the absence of hunger but not loss of control behaviors are associated with 16p11.2 deletions. *Obesity* **22,** 2625–2631 (2014) (cited on pages 191, 209).

540. Abawi, O. *et al.* Genetic Obesity Disorders: Body Mass Index Trajectories and Age of Onset of Obesity Compared with Children with Obesity from the General Population. *The Journal of Pediatrics* **262,** 113619 (2023) (cited on pages 191, 209).

541. Verbitsky, M. *et al.* Copy number variant analysis and genome-wide association study identify loci with large effect for vesicoureteral reflux. *Journal of the American Society of Nephrology* **32,** 805–820 (2021) (cited on pages 191, 215).

542. Yang, N. *et al.* Human and mouse studies establish TBX6 in Mendelian CAKUT and as a potential driver of kidney defects associated with the 16p11.2 microdeletion syndrome. *Kidney International* **98,** 1020–1030 (2020) (cited on pages 191, 214, 215).

543. Shiow, L. R. *et al.* Severe combined immunodeficiency (SCID) and attention deficit hyperactivity disorder (ADHD) associated with a Coronin-1A mutation and a chromosome 16p11.2 deletion. *Clinical Immunology* **131,** 24–30 (2009) (cited on pages 191, 216, 220).

544. Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M. & He, X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics* **52,** 740–747 (2020) (cited on page 192).

545. Darrous, L., Mounier, N. & Kutalik, Z. Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics. *Nature Communications* **12,** 7274 (2021) (cited on page 192).

546. Vollger, M. R. *et al.* Segmental duplications and their variation in a complete human genome. *Science* **376,** eabj6965 (2022) (cited on page 194).

547. Migliavacca, E. *et al.* A potential contributory role for ciliary dysfunction in the 16p11.2 600 kb BP4-BP5 pathology. *The American Journal of Human Genetics* **96,** 784–796 (2015) (cited on pages 194, 221, 224).

548. Tai, D. J. *et al.* Tissue-and cell-type-specific molecular and functional signatures of 16p11.2 reciprocal genomic disorder across mouse brain and human neuronal models. *The American Journal of Human Genetics* **109,** 1789–1813 (2022) (cited on pages 194, 225).

549. Wallace, A. S. *et al.* Longitudinal report of child with de novo 16p11.2 triplication. *Clinical Case Reports* **6,** 147–154 (2018) (cited on page 195).

550. Badar, S. A., Breman, A. M., Christensen, C. K., Graham, B. H. & Golomb, M. R. Girl-Boy Twins with Developmental Delay from 16p11.2 Triplication due to Biparental Inheritance from Two Parents with 16p11.2 Duplication. *Cytogenetic and Genome Research* **162,** 40–45 (2022) (cited on pages 195, 220).

551. Pohovski, L. M., Sansović, I., Vulin, K. & Odak, L. The first case report of distal 16p12.1p11.2 trisomy and proximal 16p11.2 tetrasomy inherited from both parents. *Croatian Medical Journal* **64,** 339–343 (2023) (cited on pages 195, 220).

552. Horev, G. *et al.* Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. *Proceedings of the National Academy of Sciences* **108,** 17076–17081 (2011) (cited on pages 195, 210).

553. Portmann, T. *et al.* Behavioral abnormalities and circuit defects in the basal ganglia of a mouse model of 16p11.2 deletion syndrome. *Cell Reports* **7,** 1077–1092 (2014) (cited on pages 195, 210).

554. Arbogast, T. *et al.* Reciprocal effects on neurocognitive and metabolic phenotypes in mouse models of 16p11.2 deletion and duplication syndromes. *PLoS Genetics* **12,** e1005709 (2016) (cited on pages 195, 210).

555. Gundersen, B. B. *et al.* Towards Preclinical Validation of Arbaclofen (R-baclofen) Treatment for 16p11.2 Deletion Syndrome. *bioRxiv,* 2023–09 (2023) (cited on pages 196, 225).

556. Martin Lorenzo, S. *et al.* Changes in social behavior with MAPK2 and KCTD13/CUL3 pathways alterations in two new outbred rat models for the 16p11.2 syndromes with autism spectrum disorders. *Frontiers in Neuroscience* **17,** 1148683 (2023) (cited on pages 196, 207, 219).

557. Skarnes, W. C. *et al.* A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* **474,** 337–342 (2011) (cited on page 196).

558. Blaker-Lee, A., Gupta, S., McCammon, J. M., De Rienzo, G. & Sive, H. Zebrafish homologs of genes within 16p11.2, a genomic region associated with brain disorders, are active during brain development, and include two deletion dosage sensor genes. *Disease Models & Mechanisms* **5,** 834–851 (2012) (cited on page 196).

559. Iyer, J. *et al.* Pervasive genetic interactions modulate neurodevelopmental defects of the autism-associated 16p11.2 deletion in Drosophila melanogaster. *Nature Communications* **9,** 2548 (2018) (cited on pages 196, 221, 224).

560. McCammon, J. M., Blaker-Lee, A., Chen, X. & Sive, H. The 16p11.2 homologs fam57ba and doc2a generate certain brain and body phenotypes. *Human Molecular Genetics* **26,** 3699–3712 (2017) (cited on pages 196, 205).

561. Kim, J. *et al.* Dissecting 16p11.2 hemi-deletion to study sex-specific striatal phenotypes of neurodevelopmental disorders. *Molecular Psychiatry,* 1–12 (2024) (cited on pages 196, 221).

562. Ghebranious, N., Giampietro, P. F., Wesbrook, F. P. & Rezkalla, S. H. A novel microdeletion at 16p11.2 harbors candidate genes for aortic valve development, seizure disorder, and mild mental retardation. *American Journal of Medical Genetics Part A* **143,** 1462–1471 (2007) (cited on pages 197, 202, 204, 208, 211, 213, 214, 216, 217).

563. Auwerx, C., Moix, S., Kutalik, Z. & Reymond, A. Disentangling mechanisms behind the pleiotropic effects of proximal 16p11.2 BP4-5 CNVs. *medRxiv,* 2024–03 (2024) (cited on pages 197, 200–202, 205, 206, 209–212, 216, 220).

564. Gillentine, M. A., Lupo, P. J., Stankiewicz, P. & Schaaf, C. P. An estimation of the prevalence of genomic disorders using chromosomal microarray data. *Journal of Human Genetics* **63,** 795–801 (2018) (cited on pages 198, 200).

565. Zhou, W. *et al.* Study of the association between schizophrenia and microduplication at the 16p11.2 locus in the Han Chinese population. *Psychiatry Research* **265,** 198–199 (2018) (cited on pages 198, 207).

566. Gudmundsson, O. O. *et al.* Attention-deficit hyperactivity disorder shares copy number variant risk with schizophrenia and autism spectrum disorder. *Translational Psychiatry* **9,** 258 (2019) (cited on pages 198, 199, 208).

567. Rees, E. *et al.* Analysis of intellectual disability copy number variants for association with schizophrenia. *JAMA Psychiatry* **73,** 963–969 (2016) (cited on pages 198, 207).

568. Hanson, E. *et al.* Cognitive and behavioral characterization of 16p11.2 deletion syndrome. *Journal of Developmental & Behavioral Pediatrics* **31,** 649–657 (2010) (cited on page 198).

569. Green, E. *et al.* Copy number variation in bipolar disorder. *Molecular Psychiatry* **21,** 89–93 (2016) (cited on pages 198, 207).

570. Glessner, J. T. *et al.* Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* **459,** 569–573 (2009) (cited on page 198).

571. Kushima, I. *et al.* Comparative analyses of copy-number variation in autism spectrum disorder and schizophrenia reveal etiological overlap and biological insights. *Cell Reports* **24,** 2838–2856 (2018) (cited on page 198).

572. Sanders, S. J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70,** 863–885 (2011) (cited on page 198).

573. Liu, N., Li, H., Li, M., Gao, Y. & Yan, H. Prenatally diagnosed 16p11.2 copy number variations by SNP Array: A retrospective case series. *Clinica Chimica Acta* **538,** 15–21 (2023) (cited on page 200).

574. Kang, H. *et al.* Pathogenic recurrent copy number variants in 7,078 pregnancies via chromosomal microarray analysis. *Journal of Perinatal Medicine* **52,** 171–180 (2024) (cited on page 200).

575. Moreno-De-Luca, A. *et al.* The role of parental cognitive, behavioral, and motor profiles in clinical variability in individuals with chromosome 16p11.2 deletions. *JAMA Psychiatry* **72,** 119–126 (2015) (cited on pages 202, 220).

576. Taylor, C. M. *et al.* Phenotypic shift in copy number variants: Evidence in 16p11.2 duplication syndrome. *Genetics in Medicine* **25,** 151–154 (2023) (cited on pages 202, 220).

577. Hippolyte, L. *et al.* The number of genomic copies at the 16p11.2 locus modulates language, verbal memory, and inhibition. *Biological Psychiatry* **80,** 129–139 (2016) (cited on page 202).

578. Kim, S. H. *et al.* Language characterization in 16p11.2 deletion and duplication syndromes. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **183,** 380–391 (2020) (cited on page 202).

579. Demopoulos, C. *et al.* Abnormal speech motor control in individuals with 16p11.2 deletions. *Scientific Reports* **8,** 1274 (2018) (cited on page 202).

580. Raca, G. *et al.* Childhood Apraxia of Speech (CAS) in two patients with 16p11.2 microdeletion syndrome. *European Journal of Human Genetics* **21,** 455–459 (2013) (cited on page 202).

581. Fedorenko, E. *et al.* A highly penetrant form of childhood apraxia of speech due to deletion of 16p11.2. *European Journal of Human Genetics* **24,** 302–306 (2016) (cited on page 202).

582. Mei, C. *et al.* Deep phenotyping of speech and language skills in individuals with 16p11.2 deletion. *European Journal of Human Genetics* **26,** 676–686 (2018) (cited on page 202).

583. Bernier, R. *et al.* Developmental trajectories for young children with 16p11.2 copy number variation. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **174,** 367–380 (2017) (cited on pages 202, 207).

584. Goldman, S. *et al.* Quantitative gait assessment in children with 16p11.2 syndrome. *Journal of Neurodevelopmental Disorders* **11,** 1–5 (2019) (cited on page 203).

585. Gur, R. *et al.* Neurocognitive Profiles of 22q11.2 and 16p11.2 Deletions and Duplications. *Research Square,* 2023–12 (2023) (cited on page 203).

586. Qureshi, A. Y. *et al.* Opposing brain differences in 16p11.2 deletion and duplication carriers. *Journal of Neuroscience* **34,** 11199–11211 (2014) (cited on pages 203, 204, 213).

587. Martin-Brevet, S. *et al.* Quantifying the effects of 16p11.2 copy number variants on brain structure: A multisite genetic-first study. *Biological Psychiatry* **84,** 253–264 (2018) (cited on pages 203, 213).

588. Maillard, A. *et al.* The 16p11.2 locus modulates brain structures common to autism, schizophrenia and obesity. *Molecular Psychiatry* **20,** 140–147 (2015) (cited on pages 203, 204, 213).

589. Cárdenas-De-La-Parra, A. *et al.* Developmental trajectories of neuroanatomical alterations associated with the 16p11.2 copy number variations. *NeuroImage* **203,** 116155 (2019) (cited on pages 203, 213).

590. Sundberg, M. *et al.* 16p11.2 deletion is associated with hyperactivation of human iPSC-derived dopaminergic neuron networks and is rescued by RHOA inhibition in vitro. *Nature Communications* **12,** 2897 (2021) (cited on pages 203, 225).

591. Urresti, J. *et al.* Cortical organoids model early brain development disrupted by 16p11.2 copy number variants in autism. *Molecular Psychiatry* **26,** 7560–7580 (2021) (cited on pages 203, 225).

592. Deshpande, A. *et al.* Cellular phenotypes in human iPSC-derived neurons from a genetic model of autism spectrum disorder. *Cell Reports* **21,** 2678–2687 (2017) (cited on page 203).

593. Blackmon, K. *et al.* Focal cortical anomalies and language impairment in 16p11.2 deletion and duplication syndrome. *Cerebral Cortex* **28,** 2422–2430 (2018) (cited on page 203).

594. Owen, J. P. *et al.* Brain MR imaging findings and associated outcomes in carriers of the reciprocal copy number variation at 16p11.2. *Radiology* **286,** 217–226 (2018) (cited on pages 203, 213).

595. Schaaf, C. P. *et al.* Expanding the clinical spectrum of the 16p11.2 chromosomal rearrangements: Three patients with syringomyelia. *European Journal of Human Genetics* **19,** 152–156 (2011) (cited on pages 203, 213).

596. Stingl, C. S., Jackson-Cook, C. & Couser, N. L. Ocular findings in the 16p11.2 microdeletion syndrome: A case report and literature review. *Case Reports in Pediatrics* **2020** (2020) (cited on pages 203, 213, 216).

597. Buchan, J. G. *et al.* Are copy number variants associated with adolescent idiopathic scoliosis? *Clinical Orthopaedics and Related Research®* **472,** 3216–3225 (2014) (cited on page 203).

598. Takeda, K. *et al.* Compound heterozygosity for null mutations and a common hypomorphic risk haplotype in TBX6 causes congenital scoliosis. *Human Mutation* **38,** 317–323 (2017) (cited on pages 203, 214).

599. Rodà, D., Gabau, E., Baena, N. & Guitart, M. *Phenotype variability in thirteen 16p11.2 deletion patients* in *Anales de Pediatria* **89** (2017), 62–63 (cited on pages 203, 206).

600. Yoon, J. G. *et al.* Molecular diagnosis of craniosynostosis using targeted next-generation sequencing. *Neurosurgery* **87,** 294–302 (2020) (cited on pages 203, 213).

601. Chang, Y. S. *et al.* Reciprocal white matter alterations due to 16p11.2 chromosomal deletions versus duplications. *Human Brain Mapping* **37,** 2833–2848 (2016) (cited on page 203).

602. Owen, J. P. *et al.* Aberrant white matter microstructure in children with 16p11.2 deletions. *Journal of Neuroscience* **34,** 6214–6223 (2014) (cited on page 203).

603. Berman, J. I. *et al.* Abnormal auditory and language pathways in children with 16p11.2 deletion. *NeuroImage: Clinical* **9,** 50–57 (2015) (cited on page 203).

604. Ahtam, B., Link, N., Hoff, E., Ellen Grant, P. & Im, K. Altered structural brain connectivity involving the dorsal and ventral language pathways in 16p11.2 deletion syndrome. *Brain Imaging and Behavior* **13,** 430–445 (2019) (cited on page 203).

605. Mazzucchelli, C. *et al.* Knockout of ERK1 MAP kinase enhances synaptic plasticity in the striatum and facilitates striatal-mediated learning and memory. *Neuron* **34,** 807–820 (2002) (cited on pages 204, 223).

606. Richter, M. *et al.* Altered TAOK2 activity causes autism-related neurodevelopmental and cognitive abnormalities through RhoA signaling. *Molecular Psychiatry* **24,** 1329–1350 (2019) (cited on page 204).

607. Wang, Q.-W. *et al.* 16p11.2 CNV gene Doc2α functions in neurodevelopment and social behaviors through interaction with Secretagogin. *Cell Reports* **42,** 112691 (2023) (cited on page 204).

608. Bertero, A. *et al.* Autism-associated 16p11.2 microdeletion impairs prefrontal functional connectivity in mouse and human. *Brain* **141,** 2055–2065 (2018) (cited on page 204).

609. Moreau, C. A. *et al.* Mutations associated with neuropsychiatric conditions delineate functional brain connectivity dimensions contributing to autism and schizophrenia. *Nature Communications* **11,** 5272 (2020) (cited on page 204).

610. Maillard, A. M. *et al.* Pervasive alterations of intra-axonal volume and network organization in young children with a 16p11.2 deletion. *Translational Psychiatry* **14,** 95 (2024) (cited on page 204).

611. Berman, J. I. *et al.* Relationship between M100 auditory evoked response and auditory radiation microstructure in 16p11.2 deletion and duplication carriers. *American Journal of Neuroradiology* **37,** 1178–1184 (2016) (cited on page 204).

612. Jenkins 3rd, J. *et al.* Auditory evoked M100 response latency is delayed in children with 16p11.2 deletion but not 16p11.2 duplication. *Cerebral Cortex* **26,** 1957–1964 (2016) (cited on page 204).

613. Matsuzaki, J. *et al.* Abnormal auditory mismatch fields in children and adolescents with 16p11.2 deletion and 16p11.2 duplication. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* **5,** 942–950 (2020) (cited on page 204).

614. LeBlanc, J. J. & Nelson, C. A. Deletion and duplication of 16p11.2 are associated with opposing effects on visual evoked potential amplitude. *Molecular Autism* **7,** 1–7 (2016) (cited on page 204).

615. Al-Jawahiri, R., Jones, M. & Milne, E. Atypical neural variability in carriers of 16p11.2 copy number variants. *Autism Research* **12,** 1322–1333 (2019) (cited on page 204).

616. Hudac, C. M. *et al.* Modulation of mu attenuation to social stimuli in children and adults with 16p11.2 deletions and duplications. *Journal of Neurodevelopmental Disorders* **7,** 1–13 (2015) (cited on page 204).

617. Hinkley, L. B. *et al.* Sensorimotor cortical oscillations during movement preparation in 16p11.2 deletion carriers. *Journal of Neuroscience* **39,** 7321–7331 (2019) (cited on page 204).

618. Yin, X. *et al.* Delayed motor learning in a 16p11.2 deletion mouse model of autism is rescued by locus coeruleus activation. *Nature Neuroscience* **24,** 646–657 (2021) (cited on page 204).

619. Kumar, K. *et al.* Subcortical brain alterations in carriers of genomic copy number variants. *American Journal of Psychiatry* **180,** 685–698 (2023) (cited on page 204).

620. Sønderby, I. E. *et al.* Effects of copy number variations on brain structure and risk for psychiatric illness: Large-scale studies from the ENIGMA working groups on CNVs. *Human Brain Mapping* **43,** 300–328 (2022) (cited on page 204).

621. Moufawad El Achkar, C. *et al.* Clinical characteristics of seizures and epilepsy in individuals with recurrent deletions and duplications in the 16p11.2 region. *Neurology. Genetics* **8,** e200018 (2022) (cited on pages 204, 205).

622. Bedoyan, J. K. *et al.* Duplication 16p11.2 in a child with infantile seizure disorder. *American Journal of Medical Genetics Part A* **152,** 1567–1574 (2010) (cited on page 205).

623. Hino-Fukuyo, N. *et al.* Genomic analysis identifies candidate pathogenic variants in 9 of 18 patients with unexplained West syndrome. *Human Genetics* **134,** 649–658 (2015) (cited on page 205).

624. Dimassi, S. *et al.* A subset of genomic alterations detected in rolandic epilepsies contains candidate or known epilepsy genes including GRIN2A and PRRT2. *Epilepsia* **55,** 370–378 (2014) (cited on page 205).

625. Vlaskamp, D. R. *et al.* PRRT2-related phenotypes in patients with a 16p11.2 deletion. *European Journal of Medical Genetics* **62,** 265–269 (2019) (cited on page 205).

626. Forrest, M. P. *et al.* Rescue of neuropsychiatric phenotypes in a mouse model of 16p11.2 duplication syndrome by genetic correction of an epilepsy network hub. *Nature Communications* **14,** 825 (2023) (cited on page 205).

627. Lipton, J. & Rivkin, M. J. 16p11.2-related paroxysmal kinesigenic dyskinesia and dopa-responsive parkinsonism in a child. *Neurology* **73,** 479–480 (2009) (cited on pages 205, 206).

628. Dale, R. C., Grattan-Smith, P., Nicholson, M. & Peters, G. B. Microdeletions detected using chromosome microarray in children with suspected genetic movement disorders: A single-centre study. *Developmental Medicine & Child Neurology* **54,** 618–623 (2012) (cited on page 205).

629. Silveira-Moriyama, L. *et al.* Clinical features of childhood-onset paroxysmal kinesigenic dyskinesia with PRRT2 gene mutations. *Developmental Medicine & Child Neurology* **55,** 327–334 (2013) (cited on page 205).

630. Weber, A., Köhler, A., Hahn, A., Neubauer, B. & Müller, U. Benign infantile convulsions (IC) and subsequent paroxysmal kinesigenic dyskinesia (PKD) in a patient with 16p11.2 microdeletion syndrome. *Neurogenetics* **14,** 251–253 (2013) (cited on page 205).

631. Termsarasab, P. *et al.* Paroxysmal kinesigenic dyskinesia caused by 16p11.2 microdeletion. *Tremor and Other Hyperkinetic Movements* **4,** 274 (2014) (cited on page 205).

632. Heron, S. E. & Dibbens, L. M. Role of PRRT2 in common paroxysmal neurological disorders: A gene with remarkable pleiotropy. *Journal of Medical Genetics* **50,** 133–139 (2013) (cited on pages 205, 223).

633. Ebrahimi-Fakhari, D., Saffari, A., Westenberger, A. & Klein, C. The evolving spectrum of PRRT2-associated paroxysmal diseases. *Brain* **138,** 3476–3495 (2015) (cited on page 205).

634. Brueckner, F. *et al.* Unusual variability of PRRT2 linked phenotypes within a family. *European Journal of Paediatric Neurology* **18,** 540–542 (2014) (cited on page 205).

635. Maas, R. P. *et al.* Benign nocturnal alternating hemiplegia of childhood: A clinical and nomenclatural reappraisal. *European Journal of Paediatric Neurology* **22,** 1110–1117 (2018) (cited on page 205).

636. Sen, K., Genser, I., DiFazio, M. & DiSabella, M. Haploinsufficiency of PRRT2 Leading to Familial Hemiplegic Migraine in Chromosome 16p11.2 Deletion Syndrome. *Neuropediatrics* **53,** 279–282 (2022) (cited on page 205).

637. Zhang, L. *et al.* A Girl with PRRT2 Mutation Presenting with Benign Familial Infantile Seizures Followed by Autistic Regression. *Case Reports in Pediatrics* **2024,** 5539799 (2024) (cited on page 206).

638. Roeben, B., Blum, D., Gabriel, H. & Synofzik, M. Atypical parkinsonism with severely reduced striatal dopamine uptake associated with a 16p11.2 duplication syndrome. *Journal of Neurology* **266,** 775–776 (2019) (cited on page 206).

639. Kamara, D., De Boeck, P., Lecavalier, L., Neuhaus, E. & Beauchaine, T. P. Characterizing sleep problems in 16p11.2 deletion and duplication. *Journal of Autism and Developmental Disorders,* 1–14 (2023) (cited on page 206).

640. Bamonte, L. Developmental presentation, medical complexities, and service delivery for a child with 16p11.2 deletion syndrome. *Pediatric Physical Therapy* **27,** 90–99 (2015) (cited on pages 206, 209, 212, 216).

641. Choi, A. *et al.* Circuit mechanism underlying fragmented sleep and memory deficits in 16p11.2 deletion mouse model of autism. *bioRxiv,* 2023–12 (2023) (cited on pages 206, 216).

642. Lu, H.-C., Pollack, H., Lefante, J. J., Mills, A. A. & Tian, D. Altered sleep architecture, rapid eye movement sleep, and neural oscillation in a mouse model of human chromosome 16p11.2 microdeletion. *Sleep* **42,** zsy253 (2019) (cited on page 206).

643. Angelakos, C. C. *et al.* Hyperactivity and male-specific sleep deficits in the 16p11.2 deletion mouse model of autism. *Autism Research* **10,** 572–584 (2017) (cited on pages 206, 219).

644. Simons VIP Consortium. Simons Variation in Individuals Project (Simons VIP): A genetics-first approach to studying autism spectrum and related neurodevelopmental disorders. *Neuron* **73,** 1063–1067 (2012) (cited on pages 206, 222).

645. Rusu, A. *et al.* Day-to-day spontaneous social behaviours is quantitatively and qualitatively affected in a 16p11.2 deletion mouse model. *Frontiers in Behavioral Neuroscience* **17,** 1294558 (2023) (cited on page 207).

646. Yang, M. *et al.* 16p11.2 deletion syndrome mice display sensory and ultrasonic vocalization deficits during social interactions. *Autism Research* **8,** 507–521 (2015) (cited on page 207).

647. Bristow, G. C. *et al.* 16p11 duplication disrupts hippocampal-orbitofrontal-amygdala connectivity, revealing a neural circuit endophenotype for schizophrenia. *Cell Reports* **31** (2020) (cited on page 207).

648. Rein, B. *et al.* Reversal of synaptic and behavioral deficits in a 16p11.2 duplication mouse model via restoration of the GABA synapse regulator Npas4. *Molecular Psychiatry* **26,** 1967–1979 (2021) (cited on page 207).

649. Levinson, D. F. *et al.* Copy number variants in schizophrenia: Confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. *American Journal of Psychiatry* **168,** 302–316 (2011) (cited on page 207).

650. Vassos, E. *et al.* Penetrance for copy number variants associated with schizophrenia. *Human Molecular Genetics* **19,** 3477–3481 (2010) (cited on page 207).

651. Jutla, A., Turner, J. B., Green Snyder, L., Chung, W. K. & Veenstra-VanderWeele, J. Psychotic symptoms in 16p11.2 copy-number variant carriers. *Autism Research* **13,** 187–198 (2020) (cited on pages 207, 208).

652. Ismail, Z. *et al.* Psychosis in Alzheimer disease—mechanisms, genetics and therapeutic opportunities. *Nature Reviews Neurology* **18,** 131–144 (2022) (cited on page 207).

653. Zheng, X. *et al.* A rare duplication on chromosome 16p11.2 is identified in patients with psychosis in Alzheimer's disease. *PLoS One* **9,** e111462 (2014) (cited on page 207).

654. Malhotra, D. *et al.* High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron* **72,** 951–963 (2011) (cited on page 207).

655. Charney, A. W. *et al.* Contribution of rare copy number variants to bipolar disorder risk is limited to schizoaffective cases. *Biological Psychiatry* **86,** 110–119 (2019) (cited on page 207).

656. Grozeva, D. *et al.* Rare copy number variants: A point of rarity in genetic risk for bipolar disorder and schizophrenia. *Archives of General Psychiatry* **67,** 318–327 (2010) (cited on page 207).

657. Williams, N. M. *et al.* Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: A genome-wide analysis. *The Lancet* **376,** 1401–1408 (2010) (cited on page 207).

658. O'Dushlaine, C. *et al.* Rare copy number variation in treatment-resistant major depressive disorder. *Biological Psychiatry* **76,** 536–541 (2014) (cited on page 208).

659. Tansey, K. *et al.* Copy number variants and therapeutic response to antidepressant medication in major depressive disorder. *The Pharmacogenomics Journal* **14,** 395–399 (2014) (cited on page 208).

660. Rucker, J. J. *et al.* Genome-wide association analysis of copy number variation in recurrent depressive disorder. *Molecular Psychiatry* **18,** 183–189 (2013) (cited on page 208).

661. Degenhardt, F. *et al.* Association between copy number variants in 16p11.2 and major depressive disorder in a German case–control sample. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **159,** 263–273 (2012) (cited on page 208).

662. McGrath, L. M. *et al.* Copy number variation in obsessive-compulsive disorder and Tourette syndrome: A cross-disorder study. *Journal of the American Academy of Child & Adolescent Psychiatry* **53,** 910–919 (2014) (cited on page 208).

663. Kotov, R., Gamez, W., Schmidt, F. & Watson, D. Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin* **136,** 768–821 (2010) (cited on page 208).

664. Fernandez, B. A. *et al.* Phenotypic spectrum associated with de novo and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. *Journal of Medical Genetics* **47,** 195–203 (2010) (cited on pages 208, 214, 215).

665. Maillard, A. *et al.* 16p11.2 Locus modulates response to satiety before the onset of obesity. *International Journal of Obesity* **40,** 870–876 (2016) (cited on page 209).

666. Narayan, K. V., Boyle, J. P., Thompson, T. J., Gregg, E. W. & Williamson, D. F. Effect of BMI on lifetime risk for diabetes in the US. *Diabetes Care* **30,** 1562–1566 (2007) (cited on page 209).

667. Shah, A. S. & Nadeau, K. J. The changing face of paediatric diabetes. *Diabetologia* **63,** 683–691 (2020) (cited on page 209).

668. Kostopoulou, E. *et al.* Hyperinsulinaemic hypoglycaemia: A new presentation of 16p11.2 deletion syndrome. *Clinical Endocrinology,* 766–769 (2019) (cited on page 209).

669. Hoytema van Konijnenburg, E. M. *et al.* Hyperinsulinism in a patient with a Zellweger Spectrum Disorder and a 16p11.2 deletion syndrome. *Molecular Genetics and Metabolism Reports* **23,** 100590 (2020) (cited on pages 209, 220).

670. Chan, W.-K. *et al.* Metabolic dysfunction-associated steatotic liver disease (MASLD): A state-of-the-art review. *Journal of Obesity & Metabolic Syndrome* **32,** 197–213 (2023) (cited on page 210).

671. Yki-Järvinen, H. Non-alcoholic fatty liver disease as a cause and a consequence of metabolic syndrome. *The Lancet Diabetes & Endocrinology* **2,** 901–910 (2014) (cited on page 210).

672. Cusi, K. *et al.* American Association of Clinical Endocrinology clinical practice guideline for the diagnosis and management of nonalcoholic fatty liver disease in primary care and endocrinology clinical settings: Co-sponsored by the American Association for the Study of Liver Diseases (AASLD). *Endocrine Practice* **28,** 528–562 (2022) (cited on page 210).

673. Menzies, C. *et al.* Distinct basal metabolism in three mouse models of neurodevelopmental disorders. *Eneuro* **8** (2021) (cited on page 210).

674. Wang, D. *et al.* Microduplication of 16p11.2 locus Potentiates Hypertrophic Obesity in Association with Imbalanced Triglyceride Metabolism in White Adipose Tissue. *Molecular Nutrition & Food Research* **66,** 2100241 (2022) (cited on page 210).

675. Béland-Millar, A. *et al.* 16p11.2 haploinsufficiency reduces mitochondrial biogenesis in brain endothelial cells and alters brain metabolism in adult mice. *Cell Reports* **42,** 112485 (2023) (cited on page 210).

676. Tomasello, D. L. *et al.* 16pdel lipid changes in iPSC-derived neurons and function of FAM57B in lipid metabolism and synaptogenesis. *iScience* **25,** 103551 (2022) (cited on page 211).

677. Welling, M. S. *et al.* Effects of glucagon-like peptide-1 analogue treatment in genetic obesity: A case series. *Clinical Obesity* **11,** e12481 (2021) (cited on page 211).

678. Richardson, T. G., Sanderson, E., Elsworth, B., Tilling, K. & Smith, G. D. Use of genetic variation to separate the effects of early and later life adiposity on disease risk: Mendelian randomisation study. *BMJ* **369,** m1203 (2020) (cited on page 211).

679. Gardner, E. J. *et al.* Reduced reproductive success is associated with selective constraint on human genes. *Nature* **603,** 858–863 (2022) (cited on page 211).

680. Puvabanditsin, S. *et al.* Microdeletion of 16p11.2 associated with endocardial fibroelastosis. *American Journal of Medical Genetics Part A* **152,** 2383–2386 (2010) (cited on page 211).

681. Shen, Y. *et al.* Intra-family phenotypic heterogeneity of 16p11.2 deletion carriers in a three-generation Chinese family. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **156,** 225–232 (2011) (cited on pages 211, 214, 216).

682. Zhu, X. *et al.* Identification of copy number variations associated with congenital heart disease by chromosomal microarray analysis and next-generation sequencing. *Prenatal Diagnosis* **36,** 321–327 (2016) (cited on page 211).

683. Lefebvre, M. *et al.* Autosomal recessive variations of TBX6, from congenital scoliosis to spondylocostal dysostosis. *Clinical Genetics* **91,** 908–912 (2017) (cited on pages 211, 214).

684. Karunanithi, Z., Vestergaard, E. M. & Lauridsen, M. H. Transposition of the great arteries-a phenotype associated with 16p11.2 duplications? *World Journal of Cardiology* **9,** 848–852 (2017) (cited on page 211).

685. Maya, I. *et al.* Prenatal microarray analysis in right aortic arch—a retrospective cohort study and review of the literature. *Journal of Perinatology* **38,** 468–473 (2018) (cited on page 211).

686. Xie, H. & Hong, N. Identification of rare copy number variants associated with pulmonary atresia with ventricular septal defect. *Frontiers in Genetics* **10,** 436730 (2019) (cited on page 211).

687. Szelest, M., Stefaniak, M., Ręka, G., Jaszczuk, I. & Lejman, M. Three case reports of patients indicating the diversity of molecular and clinical features of 16p11.2 microdeletion anomaly. *BMC Medical Genomics* **14,** 1–11 (2021) (cited on pages 211–214, 216).

688. Cai, M. *et al.* Prenatal Diagnosis of genetic aberrations in fetuses with pulmonary stenosis in southern China: A retrospective analysis. *BMC Medical Genomics* **16,** 119 (2023) (cited on page 211).

689. Li, R. *et al.* Isolated aberrant right subclavian artery: An underlying clue for genetic anomalies. *The Journal of Maternal-Fetal & Neonatal Medicine* **36,** 2183762 (2023) (cited on page 211).

690. Lin, S. *et al.* Contribution of genetic variants to congenital heart defects in both singleton and twin fetuses: A Chinese cohort study. *Molecular Cytogenetics* **17,** 2 (2024) (cited on page 211).

691. Ehrlich, L. & Prakash, S. K. Copy-number variation in congenital heart disease. *Current Opinion in Genetics & Development* **77,** 101986 (2022) (cited on page 211).

692. Assimopoulos, S. *et al.* Genetic mouse models of autism spectrum disorder present subtle heterogenous cardiac abnormalities. *Autism Research* **15,** 1189–1208 (2022) (cited on page 211).

693. Turcotte, A.-F. *et al.* Association between obesity and risk of fracture, bone mineral density and bone quality in adults: A systematic review and meta-analysis. *PloS One* **16,** e0252487 (2021) (cited on page 212).

694. Min, B.-J. *et al.* Whole-exome sequencing identifies mutations of KIF22 in spondyloepimetaphyseal dysplasia with joint laxity, leptodactylic type. *The American Journal of Human Genetics* **89,** 760–766 (2011) (cited on pages 213, 214).

695. Escamilla, C. O. *et al.* Kctd13 deletion reduces synaptic transmission via increased RhoA. *Nature* **551,** 227–231 (2017) (cited on page 213).

696. Lin, G. N. *et al.* Spatiotemporal 16p11.2 protein network implicates cortical late mid-fetal brain development and KCTD13-Cul3-RhoA pathway in psychiatric diseases. *Neuron* **85,** 742–754 (2015) (cited on page 213).

697. Tian, D. *et al.* Contribution of mGluR5 to pathophysiology in a mouse model of human chromosome 16p11.2 microdeletion. *Nature Neuroscience* **18,** 182–184 (2015) (cited on page 213).

698. Pucilowska, J. *et al.* The 16p11.2 deletion mouse model of autism exhibits altered cortical progenitor proliferation and brain cytoarchitecture linked to the ERK MAPK pathway. *Journal of Neuroscience* **35,** 3190–3200 (2015) (cited on page 213).

699. Blizinsky, K. D. *et al.* Reversal of dendritic phenotypes in 16p11.2 microduplication mouse model neurons by pharmacological targeting of a network hub. *Proceedings of the National Academy of Sciences* **113,** 8520–8525 (2016) (cited on pages 213, 225).

700. Nascimento, L. P. C. *et al.* 16p11.2 Microduplication Syndrome with Increased Fluid in the Cisterna: Coincidence or Phenotype Extension? *Genes* **14,** 1583 (2023) (cited on page 213).

701. Démurger, F. *et al.* Array-CGH analysis suggests genetic heterogeneity in rhombencephalosynapsis. *Molecular Syndromology* **4,** 267–272 (2013) (cited on page 213).

702. Bardakjian, T. M., Kwok, S., Slavotinek, A. M. & Schneider, A. S. Clinical report of microphthalmia and optic nerve coloboma associated with a de novo microdeletion of chromosome 16p11.2. *American Journal of Medical Genetics Part A* **152,** 3120–3123 (2010) (cited on pages 214, 217).

703. Al-Kateb, H. *et al.* Scoliosis and vertebral anomalies: Additional abnormal phenotypes associated with chromosome 16p11.2 rearrangement. *American Journal of Medical Genetics Part A* **164,** 1118–1126 (2014) (cited on page 214).

704. Shimojima, K., Inoue, T., Fujii, Y., Ohno, K. & Yamamoto, T. A familial 593-kb microdeletion of 16p11.2 associated with mental retardation and hemivertebrae. *European Journal of Medical Genetics* **52,** 433–435 (2009) (cited on page 214).

705. Yang, N. *et al.* TBX6 compound inheritance leads to congenital vertebral malformations in humans and mice. *Human Molecular Genetics* **28,** 539–547 (2019) (cited on page 214).

706. Liu, J. *et al.* TBX6-associated congenital scoliosis (TACS) as a clinically distinguishable subtype of congenital scoliosis: Further evidence supporting the compound inheritance and TBX6 gene dosage model. *Genetics in Medicine* **21,** 1548–1558 (2019) (cited on page 214).

707. Ren, X. *et al.* Increased TBX6 gene dosages induce congenital cervical vertebral malformations in humans and mice. *Journal of Medical Genetics* **57,** 371–379 (2020) (cited on page 214).

708. Dabbas, N., Adams, K., Pearson, K. & Royle, G. Frequency of abdominal wall hernias: Is classical teaching out of date? *JRSM Short Reports* **2,** 1–6 (2011) (cited on page 214).

709. Wat, M. J. *et al.* Genomic alterations that contribute to the development of isolated and non-isolated congenital diaphragmatic hernia. *Journal of Medical Genetics* **48,** 299–307 (2011) (cited on page 214).

710. Brady, P. *et al.* Identification of dosage-sensitive genes in fetuses referred with severe isolated congenital diaphragmatic hernia. *Prenatal Diagnosis* **33,** 1283–1292 (2013) (cited on page 214).

711. Sandbacka, M. *et al.* TBX6, LHX1 and copy number variations in the complex genetics of Müllerian aplasia. *Orphanet Journal of Rare Diseases* **8,** 1–13 (2013) (cited on page 215).

712. Chu, C. *et al.* Whole-exome sequencing identified a TBX6 loss of function mutation in a patient with distal vaginal atresia. *Journal of Pediatric and Adolescent Gynecology* **32,** 550–554 (2019) (cited on page 215).

713. Chen, N. *et al.* Perturbations of genes essential for Müllerian duct and Wölffian duct development in Mayer-Rokitansky-Küster-Hauser syndrome. *The American Journal of Human Genetics* **108,** 337–345 (2021) (cited on page 215).

714. Su, K. *et al.* Recurrent human 16p11.2 microdeletions in type I Mayer–Rokitansky–Küster–Hauser (MRKH) syndrome patients in Chinese Han population. *Molecular Genetics & Genomic Medicine* **12,** e2280 (2024) (cited on page 215).

715. Seth, A. *et al.* Gene dosage changes in KCTD13 result in penile and testicular anomalies via diminished androgen receptor function. *The FASEB Journal* **36,** e22567 (2022) (cited on page 215).

716. Haller, M., Au, J., O'Neill, M. & Lamb, D. J. 16p11.2 transcription factor MAZ is a dosage-sensitive regulator of genitourinary development. *Proceedings of the National Academy of Sciences* **115,** E1849–E1858 (2018) (cited on page 215).

717. Su, J. *et al.* Association of prenatal renal ultrasound abnormalities with pathogenic copy number variants in a large Chinese cohort. *Ultrasound in Obstetrics & Gynecology* **59,** 226–233 (2022) (cited on page 215).

718. Khoreva, A. *et al.* Novel hemizygous CORO1A variant leads to combined immunodeficiency with defective platelet calcium signaling and cell mobility. *Journal of Allergy and Clinical Immunology: Global* **3,** 100172 (2024) (cited on pages 216, 220).

719. Wang, L. A., Larson, A. & Abbott, J. K. The Immune Status of Patients with 16p11.2 Deletion Syndrome. *Journal of Clinical Immunology* **43,** 1792–1795 (2023) (cited on page 216).

720. Giannuzzi, G. *et al.* Possible association of 16p11.2 copy number variation with altered lymphocyte and neutrophil counts. *NPJ Genomic Medicine* **7,** 38 (2022) (cited on pages 216, 224).

721. Stocker, T. J. *et al.* The actin regulator coronin-1A modulates platelet shape change and consolidates arterial thrombosis. *Thrombosis and Haemostasis* **118,** 2098–2111 (2018) (cited on page 216).

722. Wang, B. *et al.* A foundational atlas of autism protein interactions reveals molecular convergence. *bioRxiv,* 2023–12 (2023) (cited on page 216).

723. Smith, H., Lane, C., Al-Jawahiri, R. & Freeth, M. Sensory processing in 16p11.2 deletion and 16p11.2 duplication. *Autism Research* **15,** 2081–2098 (2022) (cited on page 216).

724. Osório, J. M. A. *et al.* Touch and olfaction/taste differentiate children carrying a 16p11.2 deletion from children with ASD. *Molecular Autism* **12,** 1–14 (2021) (cited on page 216).

725. Ventura, M. *et al.* Bifocal germinoma in a patient with 16p11.2 microdeletion syndrome. *Endocrinology, Diabetes & Metabolism Case Reports* **2019** (2019) (cited on page 217).

726. Egolf, L. E. *et al.* Germline 16p11.2 microdeletion predisposes to neuroblastoma. *The American Journal of Human Genetics* **105,** 658–668 (2019) (cited on page 217).

727. Smolen, C. *et al.* Assortative mating and parental genetic relatedness contribute to the pathogenicity of variably expressive variants. *The American Journal of Human Genetics* **110,** 2015–2028 (2023) (cited on pages 218, 220, 238).

728. Loomes, R., Hull, L. & Mandy, W. P. L. What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry* **56,** 466–474 (2017) (cited on pages 219, 239).

729. Santos, S., Ferreira, H., Martins, J., Goncalves, J. & Castelo-Branco, M. Male sex bias in early and late onset neurodevelopmental disorders: Shared aspects and differences in Autism Spectrum Disorder, Attention Deficit/hyperactivity Disorder, and Schizophrenia. *Neuroscience & Biobehavioral Reviews* **135,** 104577 (2022) (cited on pages 219, 239).

730. Ratto, A. B. *et al.* What about the girls? Sex-based differences in autistic traits and adaptive skills. *Journal of Autism and Developmental Disorders* **48,** 1698–1711 (2018) (cited on page 219).

731. Polyak, A., Rosenfeld, J. A. & Girirajan, S. An assessment of sex bias in neurodevelopmental disorders. *Genome Medicine* **7,** 1–11 (2015) (cited on pages 219, 220).

732. Pirastu, N. *et al.* Genetic analyses identify widespread sex-differential participation bias. *Nature Genetics* **53,** 663–671 (2021) (cited on pages 219, 239, 240).

733. Lynch III, J. F. *et al.* Comprehensive behavioral phenotyping of a 16p11.2 Del mouse model for neurodevelopmental disorders. *Autism Research* **13,** 1670–1684 (2020) (cited on page 219).

734. Grissom, N. *et al.* Male-specific deficits in natural reward learning in a mouse model of neurodevelopmental disorders. *Molecular Psychiatry* **23,** 544–555 (2018) (cited on page 219).

735. Ouellette, J. *et al.* Vascular contributions to 16p11.2 deletion autism syndrome modeled in mice. *Nature Neuroscience* **23,** 1090–1101 (2020) (cited on page 219).

736. Agarwalla, S. *et al.* Male-specific alterations in structure of isolation call sequences of mouse pups with 16p11.2 deletion. *Genes, Brain and Behavior* **19,** e12681 (2020) (cited on page 219).

737. Giovanniello, J., Ahrens, S., Yu, K. & Li, B. Sex-specific stress-related behavioral phenotypes and central amygdala dysfunction in a mouse model of 16p11.2 microdeletion. *Biological Psychiatry Global Open Science* **1,** 59–69 (2021) (cited on page 219).

738. Dastan, J. *et al.* Exome sequencing identifies pathogenic variants of VPS13B in a patient with familial 16p11.2 duplication. *BMC Medical Genetics* **17,** 1–6 (2016) (cited on page 220).

739. Amor, D. J. & Bijlsma, E. K. Letter regarding the article" Extending the phenotype of recurrent rearrangements of 16p11.2: Deletions in mentally retarded patients without autism and in normal individuals ()" and the diagnosis of coexisting Mowat-Wilson syndrome in a patient with 16p11.2 deletion. *European Journal of Medical Genetics* **61,** 48–49 (2018) (cited on page 220).

740. Pelliccia, V., Ferranti, S., Mostardini, R. & Grosso, S. A case of Friedreich ataxia in an adolescent with 16p11.2 microdeletion syndrome. *Neurological Sciences* **41,** 721–722 (2020) (cited on page 220).

741. Posey, J. E. *et al.* Resolution of disease phenotypes resulting from multilocus genomic variation. *New England Journal of Medicine* **376,** 21–31 (2017) (cited on page 220).

742. Blumenthal, I. *et al.* Transcriptional consequences of 16p11.2 deletion and duplication in mouse cortex and multiplex autism families. *The American Journal of Human Genetics* **94,** 870–883 (2014) (cited on page 221).

743. Kostic, M. *et al.* Patient brain organoids identify a link between the 16p11.2 copy number variant and the RBFOX1 gene. *ACS Chemical Neuroscience* **14,** 3993–4012 (2023) (cited on page 221).

744. Schultz, L. M. *et al.* Copy number variants differ in frequency across genetic ancestry groups. *medRxiv,* 2024–03 (2024) (cited on page 222).

745. Hsieh, P. *et al.* Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science* **366,** eaax2083 (2019) (cited on page 222).

746. Liu, F. *et al.* Haplotype-specific MAPK3 expression in 16p11.2 deletion contributes to variable neurodevelopment. *Brain* **146,** 3347–3363 (2023) (cited on page 222).

747. Hudac, C. M. *et al.* Evaluating heterogeneity in ASD symptomatology, cognitive ability, and adaptive functioning among 16p11.2 CNV carriers. *Autism Research* **13,** 1300–1310 (2020) (cited on page 223).

748. Hasegawa, Y. *et al.* Microglial cannabinoid receptor type 1 mediates social memory deficits in mice produced by adolescent THC exposure and 16p11.2 duplication. *Nature Communications* **14,** 6559 (2023) (cited on page 223).

749. Hou, K. *et al.* Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nature Genetics* **55,** 549–558 (2023) (cited on page 223).

750. Crepel, A. *et al.* Narrowing the critical deletion region for autism spectrum disorders on 16p11.2. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **156,** 243–245 (2011) (cited on page 223).

751. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics* **48,** 1443–1448 (2016) (cited on page 224).

752. Steinberg, S. *et al.* Common variant at 16p11.2 conferring risk of psychosis. *Molecular Psychiatry* **19,** 108–114 (2014) (cited on page 223).

753. Stoppel, L. J. *et al.* R-baclofen reverses cognitive deficits and improves social interactions in two lines of 16p11.2 deletion mice. *Neuropsychopharmacology* **43,** 513–524 (2018) (cited on page 225).

754. Rein, B., Conrow-Graham, M., Frazier, A., Cao, Q. & Yan, Z. Inhibition of histone deacetylase 5 ameliorates abnormalities in 16p11.2 duplication mouse model. *Neuropharmacology* **204,** 108893 (2022) (cited on page 225).

755. Walsh, J. J. *et al.* 5-HT release in nucleus accumbens rescues social deficits in mouse autism model. *Nature* **560,** 589–594 (2018) (cited on page 225).

756. Walsh, J. J. *et al.* Systemic enhancement of serotonin signaling reverses social deficits in multiple mouse models for ASD. *Neuropsychopharmacology* **46,** 2000–2010 (2021) (cited on page 225).

757. Panzini, C. M., Ehlinger, D. G., Alchahin, A. M., Guo, Y. & Commons, K. G. 16p11.2 deletion syndrome mice perseverate with active coping response to acute stress–rescue by blocking 5-HT 2A receptors. *Journal of Neurochemistry* **143,** 708–721 (2017) (cited on page 225).

758. Mitchell, E. J. *et al.* Drug-responsive autism phenotypes in the 16p11.2 deletion mouse model: A central role for gene-environment interactions. *Scientific Reports* **10,** 12303 (2020) (cited on page 225).

759. Martin Lorenzo, S., Nalesso, V., Chevalier, C., Birling, M.-C. & Herault, Y. Targeting the RHOA pathway improves learning and memory in adult Kctd13 and 16p11.2 deletion mouse models. *Molecular Autism* **12,** 1–13 (2021) (cited on page 225).

760. Pucilowska, J. *et al.* Pharmacological inhibition of ERK signaling rescues pathophysiology and behavioral phenotype associated with 16p11.2 chromosomal deletion in mice. *Journal of Neuroscience* **38,** 6640–6652 (2018) (cited on page 225).

761. Nadeau, J. H. & Auwerx, J. The virtuous cycle of human genetics and mouse models in drug discovery. *Nature Reviews Drug discovery* **18,** 255–272 (2019) (cited on page 225).

762. Butter, C. E. *et al.* Experiences and concerns of parents of children with a 16p11.2 deletion or duplication diagnosis: A reflexive thematic analysis. *BMC Psychology* **12,** 137 (2024) (cited on page 225).

763. Chung, W. K., Herrera, F. F. & Simon's Searchlight Foundation. Health supervision for children and adolescents with 16p11.2 deletion syndrome. *Molecular Case Studies* **9,** a006316 (2023) (cited on page 225).

764. Leitsalu, L. *et al.* Reporting incidental findings of genomic disorder-associated copy number variants to unselected biobank participants. *Personalized Medicine* **13,** 303–314 (2016) (cited on pages 225, 241).

765. Wilkins, E. J., Archibald, A. D., Sahhar, M. A. & White, S. M. "It wasn't a disaster or anything": Parents' experiences of their child's uncertain chromosomal microarray result. *American Journal of Medical Genetics Part A* **170,** 2895–2904 (2016) (cited on pages 225, 241).

766. Rees, E. & Kirov, G. Copy number variation and neuropsychiatric illness. *Current Opinion in Genetics & Development* **68,** 57–63 (2021) (cited on page 231).

767. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583,** 699–710 (2020) (cited on page 235).

768. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015) (cited on page 235).

769. Adams, D. *et al.* BLUEPRINT to decode the epigenetic signature written in blood. *Nature Biotechnology* **30,** 224–226 (2012) (cited on page 235).

770. Osterwalder, M. *et al.* in *Craniofacial Development: Methods and Protocols* 147–186 (Springer, 2021) (cited on page 236).

771. Kvon, E. Z. *et al.* Comprehensive in vivo interrogation reveals phenotypic impact of human enhancer variants. *Cell* **180,** 1262–1271 (2020) (cited on page 236).

772. Gerrard, D. T. *et al.* An integrative transcriptomic atlas of organogenesis in human embryos. *eLife* **5,** e15657 (2016) (cited on page 236).

773. Gerrard, D. T. *et al.* Dynamic changes in the epigenomic landscape regulate human organogenesis and link to developmental disorders. *Nature Communications* **11,** 3920 (2020) (cited on page 236).

774. Rosin, J. M., Abassah-Oppong, S. & Cobb, J. Comparative transgenic analysis of enhancers from the human SHOX and mouse Shox2 genomic regions. *Human Molecular Genetics* **22,** 3063–3076 (2013) (cited on page 236).

775. Liu, H. *et al.* Functional redundancy between human SHOX and mouse Shox2 genes in the regulation of sinoatrial node formation and pacemaking function. *Journal of Biological Chemistry* **286,** 17029–17038 (2011) (cited on page 236).

776. Cobb, J., Dierich, A., Huss-Garcia, Y. & Duboule, D. A mouse model for human short-stature syndromes identifies Shox2 as an upstream regulator of Runx2 during long-bone development. *Proceedings of the National Academy of Sciences* **103,** 4511–4515 (2006) (cited on page 236).

777. Abassah-Oppong, S. *et al.* A gene desert required for regulatory control of pleiotropic Shox2 expression and embryonic survival. *bioRxiv,* 2020–11 (2020) (cited on page 236).

778. Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nature Genetics* **55,** 1243–1249 (2023) (cited on page 236).

779. Privé, F. *et al.* Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics* **109,** 12–23 (2022) (cited on page 237).

780. Hawkes, G. *et al.* Identification and analysis of individuals who deviate from their genetically-predicted phenotype. *PLoS Genetics* **19,** e1010934 (2023) (cited on page 238).

781. Khramtsova, E. A., Davis, L. K. & Stranger, B. E. The role of sex in the genomics of human complex traits. *Nature Reviews Genetics* **20,** 173–190 (2019) (cited on pages 238, 239).

782. Bernabeu, E. *et al.* Sex differences in genetic architecture in the UK Biobank. *Nature Genetics* **53,** 1283–1289 (2021) (cited on pages 239, 240).

783. Randall, J. C. *et al.* Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genetics* **9,** e1003500 (2013) (cited on page 239).

784. Döring, A. *et al.* SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. *Nature Genetics* **40,** 430–436 (2008) (cited on page 239).

785. Dumitrescu, L. *et al.* Sex differences in the genetic predictors of Alzheimer's pathology. *Brain* **142,** 2581–2589 (2019) (cited on page 239).

786. Hartiala, J. A. *et al.* Genome-wide association study and targeted metabolomics identifies sex-specific association of CPS1 with coronary artery disease. *Nature Communications* **7,** 10558 (2016) (cited on page 239).

787. Johnston, K. J. *et al.* Sex-stratified genome-wide association study of multisite chronic pain in UK Biobank. *PLoS Genetics* **17,** e1009428 (2021) (cited on page 239).

788. Carter, C. & Evans, K. Inheritance of congenital pyloric stenosis. *Journal of Medical Genetics* **6,** 233 (1969) (cited on page 239).

789. Robinson, E. B., Lichtenstein, P., Anckarsäter, H., Happé, F. & Ronald, A. Examining and interpreting the female protective effect against autistic behavior. *Proceedings of the National Academy of Sciences* **110,** 5258–5262 (2013) (cited on page 239).

790. Taylor, M. J. *et al.* Is there a female protective effect against attention-deficit/hyperactivity disorder? Evidence from two representative twin samples. *Journal of the American Academy of Child & Adolescent Psychiatry* **55,** 504–512 (2016) (cited on page 239).

791. Kruse, L. M., Buchan, J. G., Gurnett, C. A. & Dobbs, M. B. Polygenic threshold model with sex dimorphism in adolescent idiopathic scoliosis: the Carter effect. *The Journal of Bone and Joint Surgery* **94,** 1485–1491 (2012) (cited on page 239).

792. Kantarci, O. *et al.* Men transmit MS more often to their children vs women: the Carter effect. *Neurology* **67,** 305–310 (2006) (cited on page 239).

793. Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R. & Smoller, J. W. Phenome-wide heritability analysis of the UK Biobank. *PLoS Genetics* **13,** e1006711 (2017) (cited on page 239).

794. Stringer, S., Polderman, T. J. & Posthuma, D. Majority of human traits do not show evidence for sex-specific genetic and environmental effects. *Scientific Reports* **7,** 8688 (2017) (cited on page 239).

795. Traglia, M. *et al.* Genetic mechanisms leading to sex differences across common diseases and anthropometric traits. *Genetics* **205,** 979–992 (2017) (cited on page 239).

796. Sandin, S. *et al.* Examining sex differences in autism heritability. *JAMA Psychiatry* (2024) (cited on page 239).

797. Zhu, C. *et al.* Amplification is the primary mode of gene-by-sex interaction in complex human traits. *Cell Genomics* **3,** 100297 (2023) (cited on page 239).

798. Halperin Kuhns, V. L. & Woodward, O. M. Sex differences in urate handling. *International Journal of Molecular Sciences* **21,** 4269 (2020) (cited on page 240).

799. Narang, R. K. *et al.* Interactions between serum urate-associated genetic variants and sex on gout risk: Analysis of the UK Biobank. *Arthritis Research & Therapy* **21,** 1–9 (2019) (cited on page 240).

800. Vears, D. F. *et al.* Return of individual research results from genomic research: A systematic review of stakeholder perspectives. *PloS One* **16,** e0258646 (2021) (cited on page 241).