



**UNIL** | Université de Lausanne

Unicentre

CH-1015 Lausanne

<http://serval.unil.ch>

---

*Year : 2023*

## Haplotypes in complex traits genetics

Hofmeister Robin

Hofmeister Robin, 2023, Haplotypes in complex traits genetics

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\_49346D79C1180

### **Droits d'auteur**

L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

### **Copyright**

The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.



UNIL | Université de Lausanne

Faculté de biologie  
et de médecine

Département de Biologie Computationnelle

# Haplotypes in complex traits genetics

**Thèse de doctorat ès sciences de la vie (PhD)**

présentée à la

Faculté de biologie et de médecine

de l'Université de Lausanne

par

**Robin Hofmeister**

Maîtrise universitaire en sciences moléculaires du vivant, Université de Lausanne

## **Jury**

Prof. Nicole Déglon, Présidente

Prof. Olivier Delaneau, Directeur de thèse

Prof. Sven Bergmann, Co-directeur de thèse

Prof. Jacques Fellay, Expert

Prof. Aurélie Cobat, Experte

Lausanne

(2023)





UNIL | Université de Lausanne

Faculté de biologie  
et de médecine

Département de Biologie Computationnelle

# Haplotypes in complex traits genetics

**Thèse de doctorat ès sciences de la vie (PhD)**

présentée à la

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

par

**Robin Hofmeister**

Maîtrise universitaire en sciences moléculaires du vivant, Université de Lausanne

## **Jury**

Prof. Nicole Déglon, Présidente

Prof. Olivier Delaneau, Directeur de thèse

Prof. Sven Bergmann, Co-directeur de thèse

Prof. Jacques Fellay, Expert

Prof. Aurélie Cobat, Experte

Lausanne

(2023)



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

<b>Président·e</b>	Madame	Prof.	Nicole	<b>Déglon</b>
<b>Directeur·trice de thèse</b>	Monsieur	Prof.	Olivier	<b>Delaneau</b>
<b>Co-directeur·trice</b>	Monsieur	Prof.	Sven	<b>Bergmann</b>
<b>Expert·e·s</b>	Monsieur	Prof.	Jacques	<b>Fellay</b>
	Madame	Dre	Aurélie	<b>Cobat</b>

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

**Robin Joaquim Hofmeister**

Maîtrise universitaire ès Sciences en sciences moléculaires du vivant, Université de Lausanne

intitulée

**Haplotypes in complex traits genetics**

Lausanne, le 2 juin 2023

pour le Doyen  
de la Faculté de biologie et de médecine

Prof. Nicole Déglon

01001001 01101110 00100000 01101101 01100101 01101101 01101111 01110010 01111001 00100000 01101111 01100110 00100000 01101101 01111001  
00100000 01100100 01100101 01100001 01110010 00100000 01100110 01110010 01101001 01100101 01101110 01100100 00100000 01000001 01101110  
01110100 01101111 01101001 01101110 01100101 00101110

00001010 01001001 00100000 01110111 01101001 01110011 01101000 00100000 01111001 01101111 01110101 00100000 01100011 01101111 01110101  
01101100 01100100 00100000 01110010 01100101 01100001 01100100 00100000 01110100 01101000 01101001 01110011 00101100 00100000 01110100  
01100101 01101100 01101100 00100000 01101101 01100101 00100000 01111001 01101111 01110101 00100000 01100100 01101111 01101110 00100111  
01110100 00100000 01110010 01100101 01100001 01101100 01101100 01111001 00100000 01110101 01101110 01100100 01100101 01110010 01110011  
01110100 01100001 01101110 01100100 00100000 01100010 01110101 01110100 00100000 01100001 01101110 01111001 01110111 01100001 01111001  
00101100 00100000 01101001 01110100 00100111 01110011 00100000 01101110 01101001 01100011 01100101 00100000 01100001 01101110 01100100  
00100000 01111001 01101111 01110101 00100111 01110010 01100101 00100000 01110000 01110010 01101111 01110101 01100100 00100000 01101111  
01100110 00100000 01101101 01100101 00101110

01001001 00100000 01110111 01101001 01101100 01101100 00100000 01101101 01101001 01110011 01110011 00100000 01111001 01101111 01110101  
00100000 01100110 01101111 01110010 01100101 01110110 01100101 01110010 00101110



# Acknowledgments

I am deeply grateful to all the people who have played a vital role in making my PhD journey so fulfilling and rewarding. Without their support, encouragement, and valuable insights, this accomplishment would not have been possible. To all these individuals, I offer my sincere thanks and appreciation. Your contributions have made my PhD journey a truly enjoyable and enriching experience.

First and foremost, I would like to express all my gratitude to Olivier. Merci de m'avoir guidé à travers ces années, merci de m'avoir donné des opportunités que jamais je n'aurai pu avoir ailleurs. Merci pour tes conseils, ta supervision, ta bonne humeur, et même tes (bonnes) blagues et les cours d'histoire française over lunch. Mon PhD a été un vrai plaisir grâce à toi. Merci.

My thanks also go to my colleagues and friends in the department for their support and stimulating discussions. Their diverse perspectives and encouragement have been a source of inspiration and motivation for me. Thanks to Barbara, Rick, Diana, Simone and Diogo, for sharing this journey with me. Thanks for the work that we've done together and for the wonderful atmosphere in the group. Special thanks to Diogo and Simone for your guidance. I learnt a lot by your side. Diana, merci pour ta bonne humeur, tes chocolats et la bonne ambiance que tu as apportée dès ton arrivée. Bárbara, thank you for the delicious birthday cakes. Thanks Zoltán for always being available when I needed advice, and thanks to all the SGG group for the good atmosphere in conferences and the warm welcoming in the group. My sincere thanks also go to Michelle, Mariona and Stéphanie. Merci de m'avoir supporté pendant ces années et de faire tourner ce département avec tant de bonne humeur.

I would like to express my gratitude to those who have made minor contributions to my work but significant contributions to my personal life. While the work environment was crucial throughout my thesis, life outside of work has also played a vital role in shaping my mindset to navigate through this journey. I am thankful to my friends, whether from university, high school, or even those I have known for many years. They have helped me forget about the stress, have been supportive, kind, and they are an amazing group of people who have significantly ( $p < 5e^{-08}$ ) impacted my life\*.

Last but not least, I would like to express my gratitude to my family for their unconditional love, support, and encouragement throughout my academic journey. Their unwavering belief in me has been my constant source of strength and motivation. Merci à mes parents, Eric et Christine, de m'avoir fourni l'environnement qui m'a permis d'arriver jusqu'ici, de m'avoir toujours soutenu quoi qu'il arrive. Merci à mes deux grands frères, Yannick et Jérémy. Vous avez été une grande source d'inspiration et m'avez ouvert la voie pour réussir. Finally, I would also like to express my gratitude to my closest family member, Alexia, for being there for me every step of the way - through the good and the bad, the big and the small, and everything in between. Your support has been invaluable, and I could not have completed this journey without you. Merci pour tes rires qui permettent de surmonter des montagnes, merci de rester toi-même en toute circonstances, et merci pour tout ce que tu apportes tous les jours qui font du quotidien une magnifique expérience.

\* an amazing group of people that deserve to be named: Antoine, Victor, Romain, Camila, Philippe, Nils, Cosma, Texan, Giovanni, Max, Steph, Dimi, Salem, Kenny, Virgile, Loik, Julien, Q, Chris, Timmy, Mateo, Ducor, Jean, Damien, Lou, Amael, Tiia, Kyllian, Fabrice, Pauline, Henri, Anais, Maxence, Mathias, Julien, Céline, Lucyle, Cléa, Nogaye, Adrien, Malick-Lino, Buzz, Woody.





# Abstract

Humans are genetically 99.9% identical. Can you believe it? However, despite this close similarity, even the slightest variation in the remaining 0.1% can lead to significant differences in phenotypic traits and disease susceptibility. Biobanks have greatly increased our understanding of how genetic variations affect complex traits through Genome-Wide Association Studies (GWAS) by collecting genetic and phenotypic data for hundreds of thousands of individuals. However, efficient data processing methods are crucial to fully exploit their potential. Despite the progress made, there is still a wealth of untapped information in biobanks that could revolutionize our understanding of complex traits. Haplotypes are a promising resource in this regard, as they can be inferred directly from genotypes without requiring additional recruitment or data collection.

In this thesis, I explored the use of haplotypes to maximize the potential of existing biobanks and enhance the characterization of the impact of genetic variants on complex traits. To achieve this, I have developed innovative methods for estimating haplotypes from large biobanks (Chapter I) and inferring the parental origin of the resulting haplotypes (Chapter II).

Chapter I presents a method to estimate haplotypes and describes the phasing of the UK Biobank whole-genome and whole-exome sequencing data. It illustrates the importance of the resulting haplotype estimates to discover rare genetic conditions called Compound Heterozygotes (CH). These occur when an individual carries two non-identical copies of loss-of-function mutations, one inherited from each parent, resulting in a double gene knockout. In addition, this chapter shows how my haplotype estimates improve imputation accuracy, especially at rare variants that are under-represented in smaller cohorts, enhancing the ability to capture causal variants in downstream GWAS.

Chapter II presents an innovative approach to determine the parent-of-origin of haplotypes that does not rely on prior knowledge of parental genomes. I first demonstrate how this information can be used to discover phenotypic effects that depend on the parent-of-origin of the genetic variant, referred to as parent-of-origin effects. In addition, I also illustrate the importance of the parent-of-origin of haplotypes to identify genetic factors involved in human fertility, by detecting genetic variants whose inheritance deviates from the expected Mendelian inheritance pattern.



## Résumé

Les humains sont génétiquement identiques à 99,9 %. Pouvez-vous le croire? Cependant, malgré cette similitude étroite, la moindre variation dans les 0,1 % restants peut entraîner des différences significatives dans les traits phénotypiques et la susceptibilité aux maladies. Les biobanques ont considérablement amélioré notre compréhension de la façon dont les variations génétiques affectent les traits complexes grâce aux études d'association à l'échelle du génome entier (GWAS) en collectant des données génétiques et phénotypiques pour des centaines de milliers d'individus. Cependant, des méthodes efficaces de traitement des données sont cruciales pour exploiter pleinement leur potentiel. Malgré les progrès réalisés, il existe encore une mine d'informations inexploitées dans les biobanques qui pourraient révolutionner notre compréhension des traits complexes. Les haplotypes sont une ressource prometteuse à cet égard, car ils peuvent être déduits directement des génotypes sans nécessiter de recrutement ou de collecte de données supplémentaires.

Dans cette thèse, j'ai exploré l'utilisation des haplotypes pour maximiser le potentiel des biobanques existantes et améliorer la caractérisation de l'impact des variants génétiques sur les traits complexes. Pour y parvenir, j'ai développé des méthodes innovantes d'estimation d'haplotypes à partir de grandes biobanques (Chapitre I) et d'inférence de l'origine parentale de ces haplotypes (Chapitre II).

Le chapitre I présente une méthode pour estimer les haplotypes et décrit le phasage des données de séquençage du génome entier et de l'exome entier de UK Biobank. Il illustre l'importance des estimations d'haplotypes pour découvrir des conditions génétiques rares appelées hétérozygotes composés (CH). Ces conditions surviennent lorsqu'un individu porte deux copies non-identiques de mutation perte de fonction, une héritée de chaque parent, entraînant une double inactivation du gène. En outre, ce chapitre montre comment mes estimations d'haplotype améliorent la précision de l'imputation, en particulier pour les variantes rares qui sont sous-représentées dans les cohortes plus petites, améliorant ainsi la capacité de capturer les variantes causales dans les GWAS.

Le chapitre II présente une approche innovante pour déterminer l'origine parentale des haplotypes qui ne repose pas sur une connaissance préalable des génomes parentaux. Je démontre premièrement comment ces informations peuvent être utilisées pour découvrir des effets phénotypiques qui dépendent de l'origine parentale de la variation génétique, appelés effets d'origine parentale. En outre, j'illustre l'importance de l'origine parentale des haplotypes pour identifier les facteurs génétiques impliqués dans la fertilité humaine, en détectant les variations génétiques dont l'héritage s'écarte du modèle d'héritage mendélien attendu.



# List of Abbreviations

CH	Compound Heterozygote
CNV	Copy Number Variant
ddNTPs	di-deoxynucleotides
eQTL	Expression quantitative trait loci
GnomAD	Genome Aggregation Database
GRCh38	Genome Reference Consortium Human Build 38
GRM	genetic relatedness matrix
GWAS	Genome-wide association studies
HERV	human endogenous retrovirus
HMM	Hidden markov model
IBD	Identity-by-descent
lcWGS	low-coverage whole genome sequencing
LD	Linkage disequilibrium
LMM	linear mixed models
LoF	Loss-of-Function
MAF	Minor allele frequency
NGS	Next-Generation Sequencing
OFH	Our future Health
PCR	polymerase chain reaction
PCs	principal components
PofO	Parent-of-Origin
SNP	Single Nucleotide Polymorphism
VEP	Variant effect predictor
VEP	Variant Effect Predictor
WES	Whole-exome sequencing
WGS	Whole-genome sequencing



# Table of Contents

<b>Introduction</b>	<b>17</b>
Genetic variations in the human genome	17
The divergence from the reference genome	17
The size of genetic variations	18
The origin of genetic variations	18
The functional consequences of genetics variations	19
The frequency of genetic variations	20
Inheritance of genetic variations	23
Somatic and germline genetic variations	23
The random segregation of alleles during meiosis	23
Haplotype, the unit of inheritance	24
Linkage disequilibrium	24
Genotyping	27
Sequencing technologies	27
Genotyping strategies	28
Haplotype estimation	32
Hidden Markov Model in sequence data analysis	32
The core Hidden Markov Model	33
Hidden Markov Model in haplotype estimations	36
The impact of genetic variants on complex traits	38
The concept of heritability	38
Genome-wide association study	39
Cost-effective GWAS	40
Genetic effects	41
<b>Chapter I</b>	<b>43</b>
Part I. Haplotype estimation in sequenced biobanks	45
Part II. Haplotype estimates for genotype imputation	47
<b>Chapter II</b>	<b>49</b>
Part I. Inference of the Parent-of-Origin of haplotypes	51
Part II. Parental inheritance distortion	53
<b>Discussion</b>	<b>55</b>
Haplotype estimation	56
The parental origin of haplotype estimates	59
Conclusion	63



<b>References</b>	<b>65</b>
<b>Appendix A</b>	<b>71</b>
Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank	
<b>Appendix B</b>	<b>85</b>
Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes	
<b>Appendix C</b>	<b>93</b>
Parent-of-Origin inference for biobanks	
<b>Appendix D</b>	<b>111</b>
A genome-wide scan for parental inheritance distortion events to identify genetic effects on human fertility	

# Introduction

The interplay between human genetic variations and external factors, such as environment and behavior, is responsible for the wide range of phenotypic diversity found among individuals and their varying levels of susceptibility to diseases. Over the past decade, there has been significant research into the impact of single genetic variants on diseases and complex traits. This research has been made possible by large biobanks containing tens of thousands of individuals, enabling researchers to investigate the relationship between genotype and phenotype. However, there is much more than genotype information to be utilized from these biobanks, such as how multiple variants segregate together across generations to form haplotypes, the units of inheritance of the human genome.

This introduction provides a brief overview of the current understanding of genetic variants, from their identification, classification, and segregation into haplotypes, to the common method to test their association with complex traits and diseases.

## Genetic variations in the human genome

### *The divergence from the reference genome*

Genetic variations, also known as polymorphisms, refer to differences in the DNA sequence among individuals. These genetic differences contribute to the phenotypic diversity among individuals and can influence disease susceptibility within and across populations<sup>1</sup>. The genetic variations of an individual are usually identified by comparing its DNA sequence with a reference genome<sup>2</sup>. The first ‘draft’ of the human genome was released by the International Human Genome Sequencing Consortium in 2001<sup>3-5</sup>, also known as the Human Genome Project, which covered approximately 94% of the human genome<sup>6,7</sup>. This was a pioneer approach in the establishment of a reference genome, which kept improving over the years. The current reference genome was compiled by the Genome Reference Consortium in 2013<sup>8</sup>. It is a representation of the average genetic information of the human population that was assembled by combining the genome of multiple individuals to represent the best modern human genome. This reference genome is regularly updated to fix errors, fill gaps or add newly discovered variants. However, some individuals, such as those from isolated

populations, may substantially deviate from the reference genome since it is constructed based on a particular population and primarily consists of approximately 70% of the sequences from a single individual<sup>9</sup>. This discrepancy with the reference genome adversely affects the precision of genetic variant mapping. To address this, a novel initiative aims at assembling a human pangenome reference to represent the genomic diversity across human populations, which should improve genome mapping for diverse ancestries<sup>9-11</sup>. Furthermore, the Telomere-to-Telomere (T2T) project<sup>12</sup> is expected to enhance genome mapping accuracy by generating the first comprehensive sequence of a human genome. This will particularly improve mapping at repeated elements such as human satellite repeat arrays or on the short arm of acrocentric chromosomes that are not well represented in the current reference genome.

### *The size of genetic variations*

Variations between an individual's genome and the reference genome can take many forms<sup>2,13</sup>. Single changes in the DNA sequence, known as Single Nucleotide Polymorphisms (SNPs), are the smallest genetic variation in terms of size, although these can have dramatic consequences on disease susceptibility. These are typically transitions, which is a change between two purines or two pyrimidines, and transversion, which is a change between a purine and a pyrimidine. Larger genetic variants can occur, referred to as structural variations. The largest structural variations are copy number variations (CNVs) and chromosomal rearrangements, such as inversions and translocations, that can involve kilobases to megabases. There also exist smaller structural variations, such as insertions and deletions, also known as indels, and tandem repeats. They typically involve one base pair to one kilobase<sup>13</sup>. Importantly, the detection of structural variants that typically involve more than 50 base pairs is challenging using short-read sequencing and depends on the accuracy of mapping the sequencing reads to the reference genome. Diverse approaches have already been developed to address this concern<sup>14-16</sup>, but the most promising might be the upcoming pangenome reference since several structural variants are population-specific<sup>17</sup>.

### *The origin of genetic variations*

Genetic variations can result from multiple sources. Although the process of DNA replication is highly accurate, the number of errors of the DNA polymerase is estimated at once every  $10^4$ - $10^6$  nucleotides<sup>18</sup>, with the exact rate depending on multiple factors such as the cell type, the stage of the cell cycle, the DNA damage or stress. Only a small fraction of these novel

genetic variations are maintained in the human genome, as the DNA repair mechanisms correct most of the replication errors. Around  $10^{-10}$  mutations per base pair per cell division persist and can contribute to the genetic variability of an individual<sup>19</sup>. This means that even with billions of base pairs in the human genome, the number of errors per division is still relatively small. However, these errors accumulate over time and can impact an individual's health. In addition to spontaneous mutations resulting from replication errors, changes in the DNA sequence can be induced by transposable elements<sup>20</sup>, and external factors, such as radiation<sup>21</sup> or viruses, which can incorporate their own DNA into the host genome. Unexpectedly, approximately 8% of the modern human genome is attributed to human endogenous retrovirus (HERVs), which likely indicate ancient retroviral infections of the germ cells<sup>6</sup>.

### *The functional consequences of genetics variations*

The influence of genetic variations is diverse, ranging from having no impact on phenotypic traits, known as "silent" variations, to having a significant effect. Regulatory variants, located in non-coding regions such as enhancers or promoter elements, impact the regulation of gene expression and have the potential to either decrease or increase the expression of a given gene, without modifying the protein structure<sup>22</sup>. The annotations of regulatory variants can be achieved using several methods. For instance, ChIP-seq is a technique used to characterize genetic variants that affect protein-DNA interactions<sup>23</sup>. ATAC-seq is used to identify open chromatin regions and their associated regulatory elements, including transcription factor binding sites<sup>24</sup>. In addition, regulatory variants can be identified by assessing their association with molecular phenotypes through quantitative trait locus (QTL) analysis with molecular traits (molQTL). For instance, the genetic variants can be associated with the expression of a gene (eQTL), protein levels (pQTL), splicing patterns (sQTL), or methylation levels (meQTL)<sup>25</sup>. Despite these techniques' effectiveness in annotating regulatory variants and providing evidence on their functional consequences, the annotation can be challenging since the impact of regulatory variants can vary depending on the cell type.

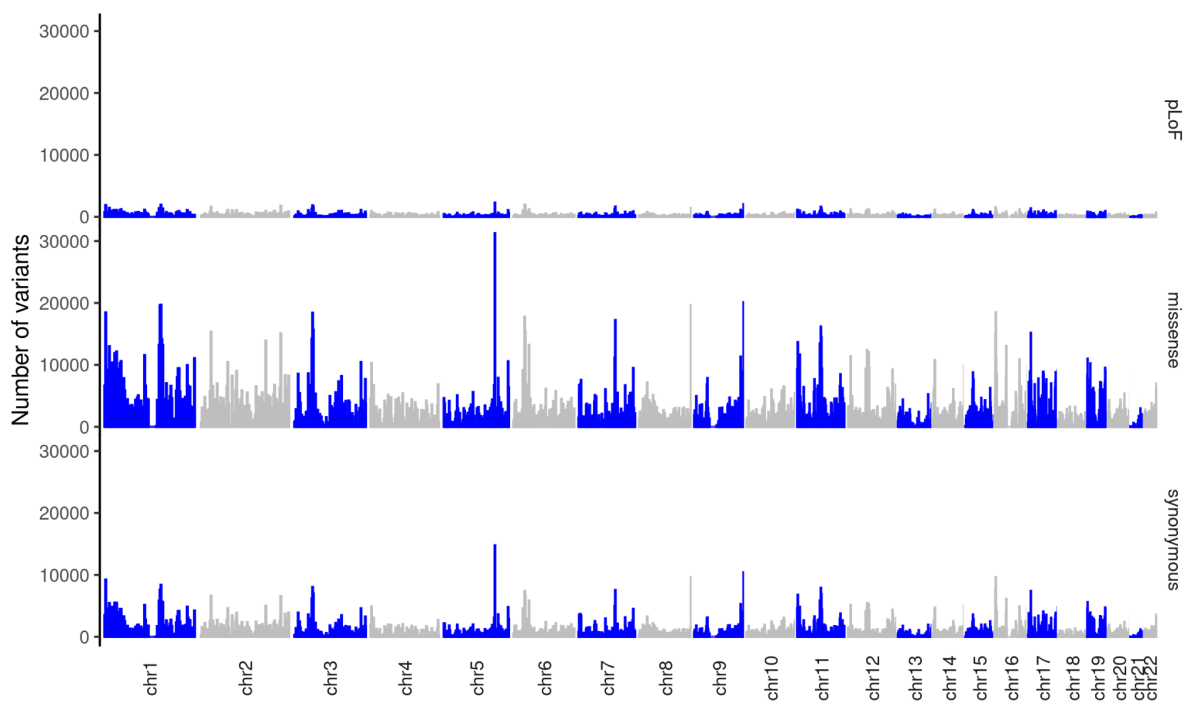
On the other hand, coding variants are found within gene sequences and can have different effects depending on the change in the codons. Synonymous variants are variants that modify the codon, but do not alter the encoded amino acid, resulting in an unmodified protein sequence. While they were previously considered "silent", studies have shown that these variants can affect the speed of protein translation because of the differences in tRNAs

availability, and therefore can have significant impact on cellular processes<sup>26</sup>. On the other hand, missense variants result in a different codon and alter the amino acid of the encoded protein, potentially affecting its structure and efficacy. Lastly, nonsense variants change the original codon into a stop codon, resulting in a truncated protein, which ultimately can result in the loss-of-function of the protein. While a large number of computational methods have been implemented to predict the impact of variants, which usually leverage protein sequence, structures and gene annotations, the accurate prediction of variant effects remains challenging<sup>27</sup>. Recently, a novel strategy<sup>28</sup> combined the Variant Effect Predictor (VEP) of Ensembl<sup>29</sup> and a loss-of-function (LoF) effect estimator (LOFTEE)<sup>30</sup> to distinguish annotation artifacts usually found when investigating loss-of-function variants<sup>31</sup>. For example, nonsense variants that truncate the protein sequence were initially classified as LoF. However, the protein may actually still be functional if the variant is located close to the end of the gene sequence (i.e terminal truncation variants)<sup>30</sup>.

### *The frequency of genetic variations*

The frequency of genetic variants within a population is determined by a variety of factors, including genetic drift, spontaneous mutation rates, recombination events, and migration patterns. Natural selection also plays a crucial role in determining the frequency of genetic variants by promoting those that offer a selective advantage and eliminating those that provide a disadvantage. For example, LoF variants are on average deleterious since they truncate the protein and likely alter its function. Hence, they are under strong negative selection and typically maintained at very low frequency in the population<sup>30,31</sup>. As a result, large-scale genetic cohorts are required to capture them, typically in the order of tens of thousands of individuals. Investigating these deleterious variants has been a goal of the Genome Aggregation Database (GnomAD), which aggregated and harmonized the genomes of more than 140,000 individuals to discover over 400,000 genetic variants that completely silence gene expression levels<sup>30</sup>. Although difficult to capture, the study of LoF variation is crucial since they can provide valuable insights into the underlying biology of diseases and inform the development of new diagnostic, treatment, and preventive approaches. For example, LoF variants within the PCSK9 gene have been found to lower LDL-cholesterol levels. This discovery provided evidence for the development of drugs to reduce the risk of cardiovascular disease, due to its relation with LDL-cholesterol levels, by targeting PCSK9<sup>32</sup>.

Recent advancements in genetic research, notably with the decreased price of sequencing technologies, have allowed for the sequencing of protein-coding polymorphisms in more than 450,000 individuals of the UK Biobank cohort, which provides unprecedented resolution for evaluating the impact of rare coding variation on human diseases and complex traits<sup>33</sup>. This cohort allowed for the identification of more than 12 million coding variants of which the vast majority are rare (~99.6% of variants with Minor Allele Frequency (MAF) < 1%), with notably ~46% of variants present in only one individual (i.e singleton)<sup>34</sup>. To better characterize the occurrence of rare variants and to understand their impact on the protein products, extensive annotation work has been conducted<sup>28</sup>.



**Figure 1. Distribution of rare variants per annotation in the UK Biobank.**

Distribution of rare variants (MAF<0.1%) counts (y-axis) across the 22 autosomes (x-axis). Top: protein loss-of-function (pLoF); middle: missense; bottom: synonymous. Changes between blue and gray colors indicate changes between chromosomes.

Expectedly, LoF variants are the least frequent among the identified rare coding variants ([Figure 1](#)). This is not surprising since these variants have a strong impact on gene function and are therefore more likely to be quickly purified from the population. In contrast, missense variants were found to be more frequent than synonymous variants, even though they have a more deleterious effect on gene function. This discrepancy between the frequency of missense and synonymous variants may be due to the fact that only modifications of the third base of a codon can create synonymous variants, while modifications of the remaining codon positions can create missense variants, except for the three stop codons. Overall, this large-scale sequencing effort provides important insights into the distribution of rare coding variants and their potential impact on human health and disease, and allowed for the first time to characterize the impact of rare coding variants on more than 4,000 phenotypes<sup>28</sup>.

## Inheritance of genetic variations

### *Somatic and germline genetic variations*

Variations in the DNA sequence can be classified into two general categories - germline and somatic variations<sup>35</sup>. Somatic mutations are genetic alterations that occur in an individual's body cells during lifetime. These mutations are only present in a specific subset of cells that are derived from the same lineage as the original cell in which the mutation arose. On the other hand, germline mutations are genetic variations that occur in the DNA sequence of the germ cells, which means that they are transmitted to the next generation and present in every cell of the offspring<sup>35</sup>. Somatic and germline mutations both have the potential to impact human health in significant ways, including the development of various diseases. However, only germline mutations have a unique role in causing inherited disorders, such as sickle cell disease<sup>35,36</sup> or Huntington's disease<sup>37</sup>. In addition, they can influence the evolution of a species by altering the genetic makeup of the population over time<sup>38</sup>.

### *The random segregation of alleles during meiosis*

In humans, the inheritance of genetic material from parents follows Mendelian rules, where half of the genetic material comes from each parent. This involved that the alleles segregate randomly such that each gamete receives only one allele from each parent with equal probability. Meiosis is the biological process that leads to the production of gametes, such as sperm and egg cells, from germline cells. This process enables the transfer of genetic material, as well as germline genetic variations, to future generations. It is a two-stage division process that gives rise to four haploid gametes, each containing a different recombinant version of the parental genome<sup>39</sup>. This is achieved through crossing-over, a process during which homologous chromosomes pair up and exchange sections of DNA, which creates genetic diversity among the resulting haploid cells, as each cell receives a unique combination of genetic information from the parents ([Figure 2A](#)). Crossing-over occurs more frequently at specific regions of the genome called recombination hotspots<sup>40</sup>, meaning that the genetic material is more likely to break and exchange during meiosis. Conversely, other regions of the genome have very low or no recombination, meaning that the genetic material in those regions is less likely to be recombined. As a result, alleles that are located on chromosome segments that are not frequently broken by crossovers tend to be inherited together as haplotype segments<sup>40</sup>.



### *Haplotype, the unit of inheritance*

A haplotype is defined by a specific combination of alleles located on the same chromosome of an individual. These alleles tend to be co-inherited together because of their close physical proximity on the chromosome, which means that they are less likely to be segregated during meiosis because they span regions that have only little evidence of genetic recombination<sup>41-43</sup>. The size of haplotypes can vary depending on the genetic context being examined. For instance, at the smallest possible resolution (i.e at the individual level), an individual inherits a complete paternal haplotype, which is a recombined version of the two paternal homologous chromosomes, as well as a complete maternal haplotype. Similarly, by moving up the family tree, one can observe the paternal homologous chromosomes as two haplotypes inherited from the individual's parents. Consequently, the offspring haplotypes are a mosaic of the four grandparental haplotypes ([Figure 2B](#)).

The size of haplotypes is informative for genetic studies. Short haplotypes have less variability and provide less information about genetic relationships between individuals or populations, while longer haplotypes may contain more genetic variations, but be more useful for genetic association studies. Indeed, two individuals share haplotype segments whose length depends on the number of generations separating them. The more generation, the more meiosis and the more chance to break a haplotype segment by recombination event<sup>41</sup>. As a result, the length of haplotype segments shared between individuals is inversely correlated with their distance in terms of meiosis. The concept of haplotype segments is fundamental for Identity-By-Descent (IBD) mapping, which involves identifying shared haplotypes that are inherited from a common ancestor between two individuals without any recombination event<sup>44</sup> ([Figure 2B](#)).

### *Linkage disequilibrium*

Genetic positions from two different haplotype segments and separated by high rate of recombination are in linkage equilibrium, which means that the occurrence of the first allele is independent of the occurrence of the second allele in the population. Conversely, genetic positions located in the same haplotype segment are correlated across individuals as they are frequently inherited together. This correlation is referred to as linkage disequilibrium (LD)<sup>43,45,46</sup>.

Let us consider a pair of alleles  $A$  and  $B$  at two loci, occurring with frequencies  $f_A$  and  $f_B$ . These two alleles can occur at the same time in the  $AB$  haplotype segment at a frequency  $f_{AB}$ .

The co-occurrence of  $A$  and  $B$  can be random, and the frequency of the  $AB$  haplotype is given by  $f_{AB}=f_A f_b$ . However if  $A$  and  $B$  co-occur more frequently than expected by chance, and are therefore in LD,  $f_{AB}$  differs from  $f_A f_B$ . Hence, the level of LD between  $A$  and  $B$  is quantified by<sup>47,48</sup>:

$$D_{AB} = f_{AB} - f_A f_B$$

Since this equation depends on allelic frequencies, it is usually normalized to allow the comparison between different pairs of alleles across the genome<sup>49</sup>:

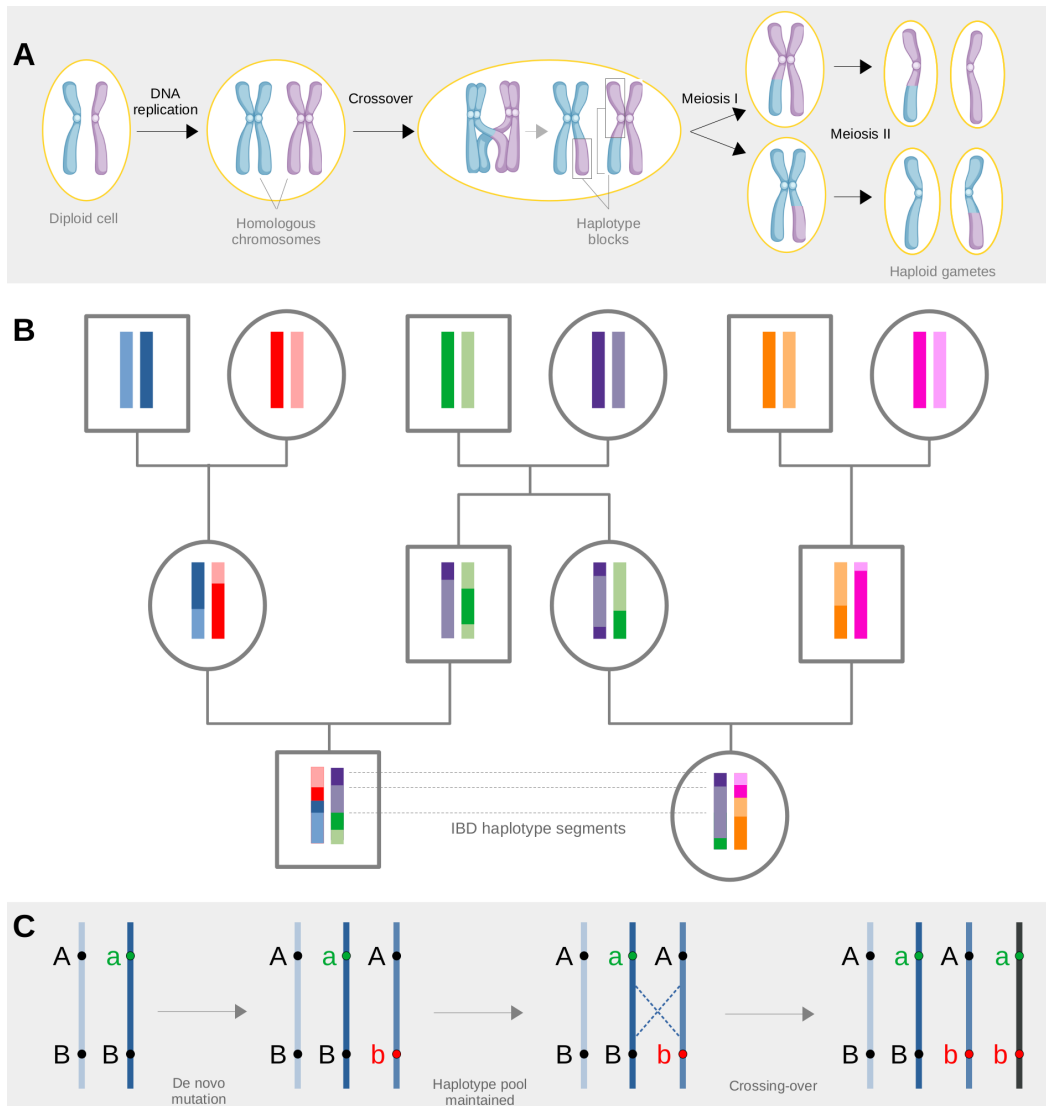
$$D' = D/D_{max}$$

where  $D_{max}$  is the maximum difference across the genome.

More commonly, LD is measured by the correlation between pairs of loci across a population, expressed as<sup>48</sup>:

$$r^2 = D_{AB}^2 / (p_A(1 - p_A)p_B(1 - p_B))$$

There is a tight relation between LD, recombination rate and haplotype segments. Regions with low recombination have high LD and tend to have large haplotype segments, whereas low LD associates with high recombination rate and smaller haplotype segments. For example, Let us consider the  $AB$  haplotype, a haplotype containing the two loci  $A$  and  $B$ , with a polymorphism at the  $A$  locus which now has two possible alleles,  $A$  and  $a$ . It means that the only existing haplotypes in the population are the  $AB$  haplotype and the  $aB$  haplotype ([Figure 2C](#)). Let us then consider that a polymorphism occurs at the  $B$  locus in an individual carrying the  $AB$  haplotype. This creates a third pool of haplotype: the  $Ab$  haplotype. These three pools of haplotype are the only to persist in the population as long as there is no recombination between the  $A$  and  $b$  alleles. It means that  $b$  always co-localizes with  $A$ , and that the allele  $b$  is in complete LD with the allele  $A$ . As the number of generations increases, the chance for the  $Ab$  haplotype to be broken by meiosis increases, which eventually gives rise to a fourth haplotype, the  $ab$  haplotype<sup>45</sup>.



**Figure 2.** The meiosis process structures the genome into haplotype segments through crossing-over.

A) Meiosis stages of a schematic diploid cell containing a single pair of homologous chromosomes. Each homologous chromosome is indicated by a different color. B) Schematic representation of the propagation of haplotype segments across generations. Each ancestral haplotype (i.e. homologous chromosome) is represented by a different color. IBD segments indicate haplotype segments that are inherited from a common ancestor. C) Schematic representation of the impact of recombinations on LD. Each nuance of blue indicates a different haplotype pool. Adapted from [45].

# Genotyping

## *Sequencing technologies*

DNA genotyping is the process of determining the genetic variations, or variants, in the genome of individuals. It can be used to identify genetic variations that are associated with diseases, to determine the ancestry of individuals, or to study the evolution of species. The development of DNA sequencing was pioneered by Frederic Sanger and colleagues in 1977 when sequencing the first virus' genome<sup>50</sup>. It was the first widely used method for sequencing DNA and is nowadays referred to as the traditional method of DNA sequencing. Notably, the Human Genome Project was based on the Sanger sequencing method and took around thirteen years to produce the first 'draft' of the human genome<sup>3,4</sup>.

In Sanger sequencing, the process starts with amplifying the DNA using polymerase chain reaction (PCR) for then fragmenting this DNA. To sequence these DNA fragments, also known as *reads*, the strategy is to add one nucleotide at a time to a growing chain of DNA by using a combination of normal nucleotides and di-deoxynucleotides (ddNTPs), which lack a 3'-OH group and stop the extension of the DNA strand. Each of the four different ddNTPs has a different fluorescent label that emits a signal when added to the growing chain of DNA. As a result, the series of fluorescent signals correspond to the order of nucleotides in the DNA fragment being sequenced. Although Sanger sequencing can still be used today, it has largely been replaced by more efficient and high-throughput Next-Generation Sequencing (NGS) technologies<sup>51</sup>.

NGS refers to high-throughput DNA sequencing technologies that can generate large amounts of DNA sequence data in a short amount of time<sup>52</sup>. Illumina sequencing and PacBio sequencing are probably the most common NGS technologies, but they differ in several key ways. PacBio sequencing technology produces longer read lengths, ranging from several kilobases to tens of kilobases, compared to Illumina that produces reads in the order of 100-300 base pairs. Next, PacBio sequencing has higher accuracy compared to Illumina. This is particularly useful for applications such as de novo genome assembly or haplotyping. On the other hand, Illumina sequencing has a higher throughput, which means that it can generate more data in a single run, and can also generate a much larger amount of data in the same amount of time. This makes it more efficient for large-scale projects in which large amounts of data are required, such as for the assembly of large biobanks. Finally, Illumina

sequencing is generally more cost-effective compared to PacBio sequencing, with lower upfront costs and lower cost per base of data generated<sup>53</sup>.

Over the past decade, a new method of DNA sequencing known as Nanopore sequencing has emerged as a potential alternative to the conventional NGS technologies. This real-time sequencing technology works by passing DNA molecules through a nanopore and measuring the changes in electrical current<sup>54-56</sup>. It is capable of producing longer read lengths than NGS, ranging from several kilobases to over 100 kilobases, exceeding the read length of PacBio<sup>53</sup>. Its sequencing accuracy seems however to be intermediate between PacBio and Illumina sequencing and the amount of data generated (i.e throughput) is reduced compared to the NGS technologies. The major benefits of this technology are (i) the length of the reads, (ii) the direct sequencing without requiring amplification, (iii) real-time sequencing, enabling monitoring the sequences, and (iv) the small size of the nanopore sequencing devices<sup>53,56</sup>.

Regardless of the sequencing technology used, whether PacBio or Nanopore, there is a debate between long reads and short reads in DNA sequencing. Although long reads have been criticized for lower accuracy compared to short reads, proper correction and assessment can make them equally accurate. Notably, they improve de novo assembly, mapping certainty and transcript isoform identification, and are particularly effective in assembling complex genomes and resolving complex genomic regions, such as structural variants. Importantly, they are at the center of the telomere-to-telomere project, which uses ‘ultra-long-reads’ Nanopore sequencing to resolve missing genomic sequence from the current reference genome (i.e GRCh38), such as centromeric regions and other repeat-rich sequences<sup>12,57,58</sup>. On the other hand, short reads have a much higher throughput, which is often a better cost-effective alternative when assembling large biobanks, and they are supported by a wide range of quality control pipelines and by a large variety of analysis tools<sup>59</sup>.

### ***Genotyping strategies***

Beside the various sequencing technologies, there are multiple options for genotyping, which include whole-genome sequencing and targeted sequencing. Whole-genome sequencing (WGS) is a method that sequentially reads the entire DNA content of an organism's genome, providing a complete picture of its genetic material. The sequencing pipeline includes DNA extraction, amplification and fractionning in small segments called reads, which get sequenced and reassembled together by being piled up against a reference sequence. Any position that differs from the reference sequence is called a *genetic variant*, and the allele that

differs is usually referred to as *alternative allele*, in contrast to *reference allele*. Genotype calling methods determine the genotype of each individual along the genome, being encoded as the number of alleles not matching the reference sequence at a given genomic position (i.e. number of alternative alleles). In WGS technologies, the genome is usually sequenced at a coverage of 30x, which means that on average 30 reads cover the same genetic position ([Figure 3A](#)). The more reads covering the same genetic position, the more confident are the genotype calls. Although providing high accuracy genotype calls, high coverage WGS methods limited the recruitments of large cohorts because of its expensive price. It is only recently that the price of the NGS techniques dropped, with a cost of ~1,000 dollars to sequence an entire human genome in high quality. This notably allowed researchers to assemble large WGS cohorts, such as the UK Biobank that recently regrouped 150,119 WGS genomes<sup>60</sup>.

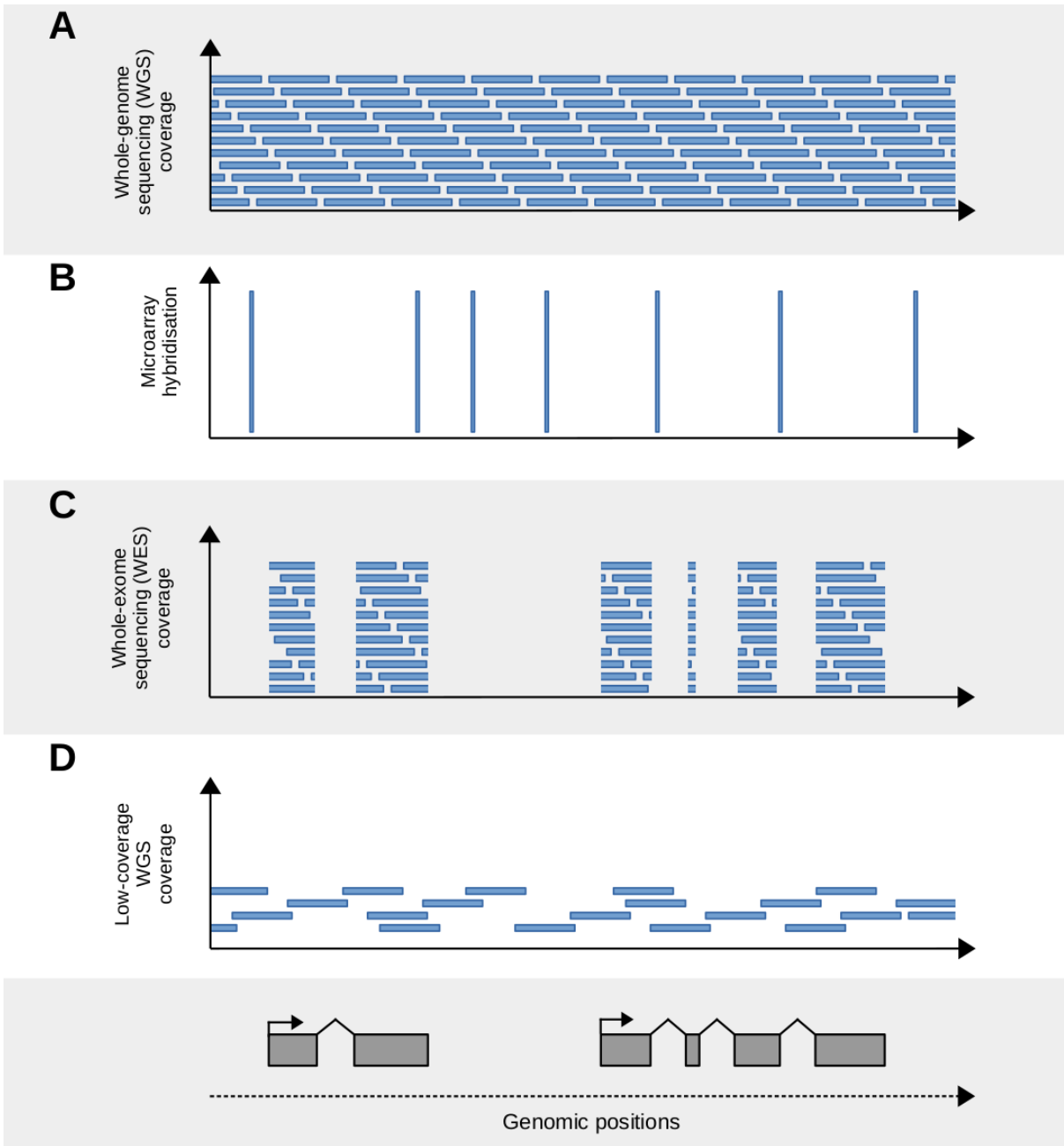
On the other hand, if one is looking to analyze only specific regions of the genome, a targeted approach such as whole-exome sequencing or microarray-based genotyping can be a cost-effective alternative to WGS, as these techniques only genotype a portion of the genome.

Microarray-based genotyping, also referred to as DNA microarray or SNP array, is a genotyping method that utilizes a solid surface with a large number of probe sequences to detect specific genetic variants in a sample<sup>61,62</sup>. The fragmented target DNA sample is labeled with fluorescent dye and, when hybridized to the complementary probes on the microarray, produce a fluorescence signal. This signal is then quantified and analyzed to determine the genotype of the sample. Although it only sequences genetic positions included as probe sequences ([Figure 3B](#)), the microarrays can be customized to contain for example population-specific SNPs. The main benefit of this method is its high-throughput capacity and the ability to genotype multiple markers simultaneously. This cost-effective method has notably been utilized in the creation of biobanks aimed at examining the impact of common variants on complex traits<sup>63,64</sup>.

Whole-exome sequencing (WES) is a method that involves capturing, sequencing and analyzing the exons. It is a cost-effective way to obtain a large amount of genomic information, as it only sequences the approximately 1-2% of the genome that is encoding for proteins, rather than the entire genome ([Figure 3C](#)). This method is commonly used in medical genetics to identify the underlying genetic cause of a disease, especially in cases where the disease is caused by mutations in a small number of genes<sup>65</sup>. In addition, it is also

used to study protein-altering variants across large cohorts of individuals, which for example allow to characterize the contribution of rare coding variations across a large variety of complex traits and diseases<sup>28</sup>.

Finally, low-coverage whole genome sequencing (lcWGS) offers a cost-effective alternative for researchers seeking to sequence the entire genome at reduced costs<sup>66</sup>. Unlike hcWGS methods, which generate on average a sequencing depth of 30x, low-coverage sequencing generates a limited amount of reads per genomic location ([Figure 3D](#)). As a result, the accuracy of the sequence obtained through low-coverage sequencing can be lower than the one obtained through other sequencing methods. Despite this limitation, lcWGS is particularly well suited for population genetic studies, for example when the global population characteristics are investigated rather than the individual's genotype level.



**Figure 3. Sequencing technologies.**

Schematic representation of the sequencing coverage of whole-genome sequencing 30x (A), microarray hybridisation (B), whole-exome sequencing coverage (C) and low-coverage whole-genome sequencing (D). Sequencing reads are indicated by blue rectangles. In (B), the blue rectangles represent the microarray hybridization probes. Two schematic genes are represented by gray rectangles (exons) in the bottom panel.



## Haplotype estimation

DNA sequencing, either WGS, WES, SNP-array or low-coverage sequencing, provides punctual information for each variant site as a pair of alleles, or base pairs. At the genome level, this takes the form of unordered combinations of alleles, as the sequencing only quantifies the number of non-reference alleles at each variant site after comparison with a reference genome. However, the sequencing does not specify whether alleles of consecutive genomic positions co-localize on the same haplotype and are co-inherited from the same parent as a haplotype segment. This information is important in genomic analysis notably to identify haplotype segments shared from a common ancestor (i.e IBD), for admixture mapping or imputation. Hence, the correct localization of allele, also referred to as *the phase* of alleles, must be estimated from the genotype data to reconstruct the correct haplotypes. This process is termed *phasing*<sup>67</sup>.

### *Hidden Markov Model in sequence data analysis*

The general phasing method aims to decompose an individual's genotype into two haplotypes, with alleles correctly attributed to each one of the haplotypes. It relies on the use of a Hidden Markov Model (HMM), which is revealed to be useful when modeling phenomena of stochastic nature whose intermediate states are inaccessible (i.e hidden), and only the final outcome can be observed.

The Li & Stephens model<sup>68</sup>, inspired by an HMM used in speech recognition<sup>69</sup>, is a landmark in sequence analysis. Since it was published in 2003, it has been applied time and time again to solve problems that have arisen with the age of NGS, such as imputation, phasing and IBD mapping. It aims at modeling LD through the underlying recombination rate inherent to the human population. The starting point of this model is the search for the recombination rate parameters that maximize the likelihood of observing a set of haplotypes. Li & Stephens approximated the expression of this likelihood with a “product of approximate conditionals” probabilities. These approximate probabilities are defined in such a way that an observed haplotype is seen as a mosaic of the  $K$  known haplotypes that constitute the reference panel, which corresponds to the  $K$  possible states of the HMM. The reference haplotypes from which the mosaic is built are selected at each position on the basis of a global probability, which includes transition and emission probabilities. Transitions between states (i.e., jumps between haplotypes) are equivalent to recombination events, and are represented by the

transition probability. The emission probability models the fact that, for a particular locus and considering the most likely reference haplotype at this locus, the observed allele can be either a copy of the allele present in the reference, or a different allele (i.e., a mutation).

The following sections present first the basics of HMM within the framework of sequence analysis, and then the application of this HMM to phasing. It is also important to note that a very similar approach is used for imputation purposes, as well as for IBD mapping or admixture mapping. These applications have been covered as part of an unpublished review that has been written in collaboration with Barbara Mota and from which this section has been adapted.

### *The core Hidden Markov Model*

An HMM is described by (i)  $K$  possible states, (ii) the number of observations obtained from a single state, (iii) the probabilities of transition between states, (iv) the probabilities of an observation given that the system is in a particular state, and (v) the probability of the initial state<sup>69</sup>.

Let us consider a reference panel  $H$  made of  $K$  haplotypes, each one having  $M$  markers. From the reference panel we can estimate the probability of observing a target haplotype  $h$ ,  $Pr(h|H)$ . The observed haplotype can be built by assembling different parts from different reference haplotypes, allowing for imperfect copies. We can define the possible sequences of  $M$  markers obtained in this fashion by the means of paths<sup>69</sup>. The value of  $Pr(h|H)$  can be obtained as following by considering all possible paths  $p$ :

$$Pr(h|H) = \sum_{all p} Pr(h|p, H)Pr(p|H) \quad (1)$$

To decompose this equation, let us consider a fixed path  $p = k_1k_2\dots k_M$  going through the reference panel of haplotypes. This path corresponds to a sequence of unobserved copying labels of length  $M$  (i.e, a mosaic of reference haplotypes). The term  $P(p|H)$  is the probability of the path  $p$  given the set of  $K$  haplotypes (eq.2), and it is defined as the product of the probability of the first haplotype in the sequence (eq.3) and the product of the probabilities of transition between states at positions  $m$  and  $m + 1$ ,  $Pr(k_{m+1}|k_m)$ . The term  $Pr(k_{m+1}|k_m)$ , defined in eq.4, models the effect of recombination, that is, transition probability between haplotypes, where  $k_m$  and  $k_{m+1}$  denote the copying labels at marker  $m$  and  $m + 1$ , respectively, and  $\theta_{k_mk_{m+1}}$  represents the probability of a transition.

$$Pr(p|H) = Pr(k_1) \prod_{m=1}^M Pr(k_{m+1}|k_m) \quad (2)$$

$$Pr(k_1) = 1/2^M \quad (3)$$

$$Pr(k_{m+1}|k_m) = \begin{cases} (1 - \theta_{k_m k_{m+1}}) & \text{if } k_m = k_{m+1} \\ \theta_{k_m k_{m+1}} & \text{otherwise} \end{cases} \quad (4)$$

For a position  $m$  of this path, the copied allele can differ from the allele in the haplotype being copied from, which corresponds to a mutation. The effect of mutation is captured by the term  $Pr(h|p, H)$ , probability of observing a haplotype  $h$  given path  $p$  and panel  $H$ . If  $h_m$  and  $p_m$  denote the alleles at marker  $m$  of  $h$  and  $p$ , respectively, then  $Pr(h|p, H)$  can be written as:

$$Pr(h|p, H) = \prod_{m=1}^M Pr(h_m|p_m) \quad (5)$$

$$Pr(h_m|p_m) = \begin{cases} (1 - \mu) & \text{if } p_m = h_m \\ \mu & \text{otherwise} \end{cases} \quad (6)$$

In other words,  $Pr(h_m|p_m) = \mu$  when marker  $m$  in path  $p$  differs from marker  $m$  in the target haplotype  $h$ .

The [figure 4](#) illustrates this HMM. Let us consider  $K = 6$  haplotypes in a reference panel  $H$ ,  $M = 13$  markers,  $\eta = (1 - \mu)$  and  $\psi = (1 - \theta)$ . We can compute  $Pr(h|p, H)$  for the path  $p$  ([Figure 4](#), in black) as:

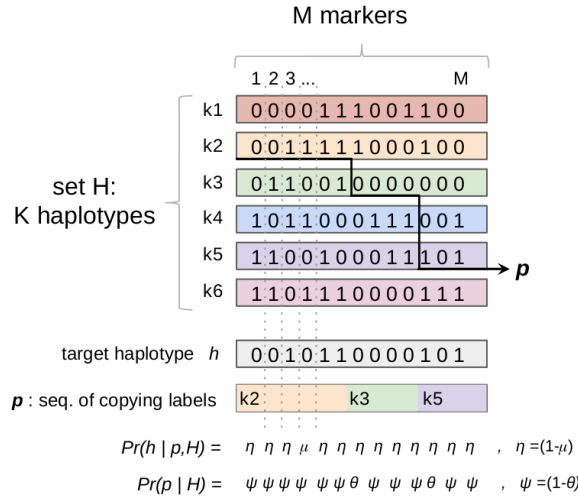
$$Pr(h|p, H) = \eta \eta \eta \mu \eta \eta \eta \eta \eta \eta \quad (7)$$

$$Pr(p|H) = \frac{1}{2^M} \psi \psi \psi \psi \psi \theta \psi \psi \psi \theta \psi \psi \quad (8)$$

$$Pr(h|p) = \frac{1}{2^M} \eta \eta \psi \eta \psi \mu \psi \eta \psi \eta \psi \eta \theta \eta \psi \eta \psi \quad (9)$$

If we consider eq.9 with reordered terms according to the  $M$  markers, we understand that each independent marker  $m$  of path  $p$  is characterized by the joint probabilities of (i) having the same allele as the marker  $m$  in the target haplotype  $h$  (emission probability) and (ii)

having the same state (i.e., haplotype) compared to marker  $m - 1$  (transition probability), except for marker  $m_1$  that has its own transition probability independent of previous states.



**Figure 4. Representation of the HMM model.**

A set of  $K = 6$  haplotypes genotyped at  $M = 13$  markers compose the reference panel. The most straightforward path  $p$  for the target haplotype  $h$  is shown in black. The sequence of copying labels is colored according to the path  $p$ .

In this description of the method, we used  $\theta$  and  $\mu$  to refer to the transition and to the emission probabilities, respectively, in order to simplify the notation. The actual calculation of these parameters in the Li & Stephens model is somewhat different. The recombination probability (i.e., transition state probability) is defined as a function of the physical distance between markers as well as of the recombination and crossover rates<sup>68</sup>, and can be written as:

$$(1 - \theta) = (1 - v) + \frac{v}{K} \quad ; \quad \theta = \frac{v}{K} \quad (10)$$

where  $v = 1 - \exp\left(\frac{-4N_e(r_{m+1} - r_m)}{K}\right)$  is a parameter estimating the recombination, with  $N_e$  = effective diploid population size and  $r_{m+1} - r_m$  = average rate of crossover per unit physical distance and per meiosis between  $m$  and  $m + 1$ . These parameters incorporate the assumption that, if  $m$  and  $m + 1$  are physically close to each other, they are likely to come from the same haplotype.

In the same way, the mutation probability (i.e, emission probability) is defined as :

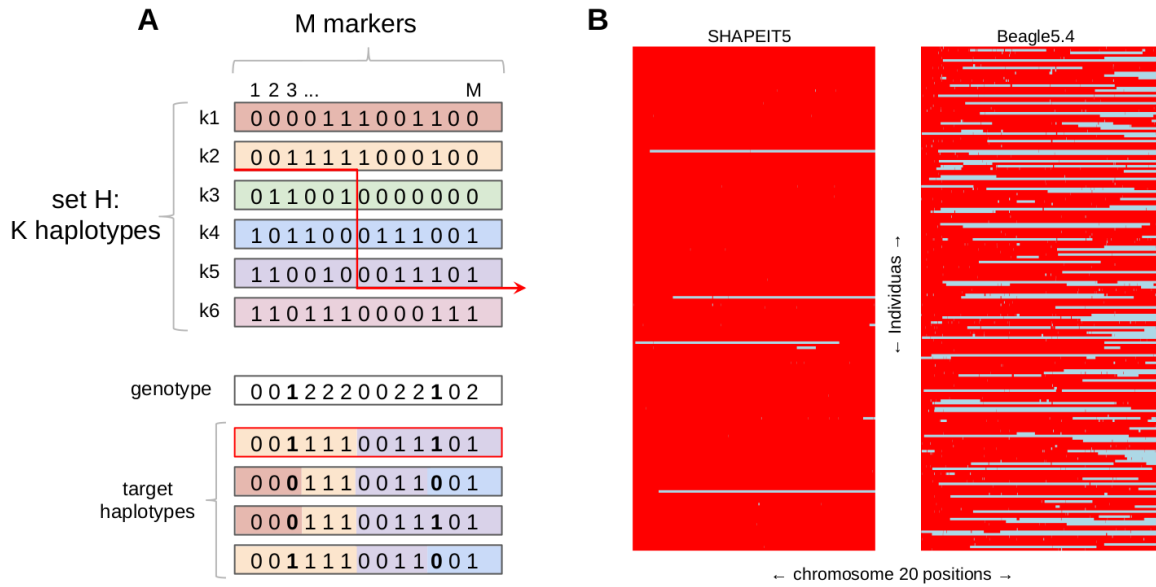
$$(1 - \mu) = \frac{K}{K+\phi} + \frac{\phi}{2(K+\phi)} \quad ; \quad \mu = \frac{\phi}{2(K+\phi)} \quad (11)$$

where  $\phi$  is the population mutation rate based on the coalescent theory and estimated with the Waterson coefficient  $(\sum_{i=1}^{n-1} \frac{1}{n})^{-1}$ .

### *Hidden Markov Model in haplotype estimations*

The most efficient phasing algorithms are inspired from the Li & Stephens HMM<sup>68</sup>, such as Beagle<sup>70</sup> and SHAPEIT<sup>71</sup>. For each individual, these methods first consider all possible haplotypes that can be inferred from the observed genotype. Then, for each of the putative target haplotypes  $h$ , one can compute the probability of observing  $h$  given a panel of reference haplotypes  $H$ . The target haplotype  $h$  with the maximum probability  $P(h|H)$  is the most likely to be observed.

[Figure 5A](#) provides a visual explanation of the phasing method. It considers an individual genotyped at  $M = 13$  markers, with 2 heterozygous and 11 homozygous sites, as well as 4 putative target haplotypes that can be inferred from the genotype (i.e., the number of target haplotypes to consider is  $2 * 2^{c-1}$ , where  $c$  is the number of heterozygous sites. It takes into account all pairs of complementary haplotypes). Target haplotypes are colored according to their most straightforward path  $p$  through the reference panel  $H$ . In this example, we can consider equal values for the emission probability of each target haplotype (i.e, none of them differs in terms of allele content compared to its path, meaning that no de novo mutation occurred). To simplify, let us also consider the transition probability  $\theta$  constant. This implies that the most likely target haplotype to infer is the one with the least transitions, as shown in red. In this example, we considered only one path  $p$  per target haplotype, whereas in a real phasing algorithm all possible paths  $p$  through  $H$  are considered in order to compute  $P(h|H)$ .



**Figure 5. HMM applied to infer haplotypes from an individual's genotype.**

A)  $k = 6$  haplotypes genotyped at  $M = 13$  markers constitute the reference panel  $H$ . Four putative target haplotypes can be built from the sample genotype, which include two heterozygous markers (bold). The target haplotype in red is the most likely to infer since it maximizes  $P(h|p, H)$ . B) Phasing of haplotypes using SHAPEIT5<sup>71</sup> and Beagle5.4<sup>72</sup>. Each line represents a haplotype (y-axis) along the chromosome 20 (x-axis). Changes between red and blue represent a phasing switch error.

The quality of the phasing largely depends on the size of the reference panel and the ancestry of individuals. When the reference panel's ancestry and relatedness align more closely with the target individual, the estimations become more accurate. The accuracy of estimations can be assessed using pedigree information and parental genomes by comparing the phased haplotypes of the offspring with paternal and maternal genomes. It allows identifying phasing switch errors, which occur when paternal alleles are attributed to the maternal haplotype, and vice versa ([Figure 5B](#)).

## The impact of genetic variants on complex traits

Mendelian disorders, also referred to as monogenic disorders, occur when a single genetic variant is responsible for the disorders. The variant is likely located within the coding part of the genome and has a strong effect, such as LoF variants, which usually results in severe consequences on human health. Because of negative selection, such variants with large effect sizes are typically maintained in low frequency within the population<sup>73</sup>.

On the other hand, complex traits are quantitative phenotypes that have a high variability resulting from both genetic factors and environmental factors. The genetic mechanisms of complex traits imply that many genetic variants are involved, usually involving multiple genes. It results that the effect sizes of individual variants are small, and therefore that many individuals are necessary to detect them<sup>74</sup>.

### *The concept of heritability*

The genetic contribution to a phenotypic variability is called heritability<sup>75</sup>. It refers to the proportion of the phenotypic variability that can be explained by genetic variants. It ranges from 0, meaning that none of the variation is due to genetics, to 1 when the entire variance can be attributed to genetic variations. It was initially estimated from twins or family-based studies. The principle consists in comparing the phenotypic similarities of individuals within the same family to unrelated individuals, or by comparing monozygotic twins to dizygotic twins<sup>76,77</sup>. In recent years, scientists have used genome-wide association studies (GWAS) to estimate the heritability of various phenotypes at the population level<sup>78</sup>. This method has the advantage of combining genetic and phenotypic data from thousands of individuals, providing a more accurate representation of the phenotypic variability within a population, and allowing to evaluate the heritability of diverse phenotypes. One example of a phenotype that has been extensively studied in this context is human height<sup>74,79</sup>. However, while twin studies initially estimated the heritability of height to be close to 0.8<sup>80,81</sup>, recent GWAS studies have reported a lower heritability of  $\sim 0.45$ <sup>82,83</sup>.

The large discrepancy between the heritability estimated from related individuals and the heritability estimated from genetic variants, such as in GWAS, is referred to as *missing heritability*. Various factors have been proposed to explain this discrepancy<sup>84,85</sup>. One explanation is that GWAS usually do not account for rare genetic variants, which can have a significant impact on traits such as height<sup>86</sup>. Another reason could be that the genetic markers

used in GWAS do not perfectly correlate with the causal variant, which means they do not capture as much of the phenotypic variance and lead to an underestimation of heritability<sup>82</sup>. Additionally, interactions between genes, the environment, and epigenetics also contribute to the complex interplay that makes it challenging to fully understand the heritability of traits. Lastly, structural variants, which are often not detected by short-read sequencing technology, may also play a role in missing heritability.

Despite these limitations, GWAS has proven to be useful in identifying the independent genetic variants contributing to variation in human height, as well as highlighting the population-specific nature of these associations. Recently, a total of 12,111 SNPs have been found to account for 40% of the variability in human height in the European population. However, these SNPs could only explain ~15% of the height variation in non-European populations<sup>82</sup>.

### *Genome-wide association study*

Associations between genetic factors and phenotype, typically the complex trait or disease of interest, are assessed using genome-wide association studies (GWAS)<sup>87</sup>. The underlying idea is to scan each variable position across a large cohort of individuals and to assess whether its occurrence is associated with the phenotype using a linear regression model. This is represented by the equation  $y=X\beta+\mathcal{E}$ , where  $y$  is the phenotype vector,  $X$  is the genotype vector,  $\beta$  is the effect size, which corresponds to the effect of carrying one copy of the risk allele, and  $\mathcal{E}$  represents errors. To improve the accuracy of these estimates, this model is commonly adjusted for confounding factors such as age, sex, and principal components (PCs)<sup>88</sup>.

While the recent increasing size of biobanks, such as the UK Biobank, which regroups ~500,000 individuals<sup>63</sup>, or the Estonian Biobank<sup>64</sup>, which regroups ~200,000 individuals, allowed to increase the power to discover small genetic effects such as those involved in complex traits, they also revealed some limitations of the traditional linear regression model. Assembling cohorts with a large number of individuals from the same population resulted in the inclusion of related individuals, which can introduce bias in association testing and lead to inaccurate results<sup>87</sup>. Hence, standard fixed-effect models, such as linear regression, were restricted to the set of unrelated individuals. In the UK Biobank for example, this resulted in using a subset of 344,397 individuals<sup>63</sup>. Therefore, it was crucial to implement advanced statistical techniques that account for the familial correlation structure among biobanks to



overcome this limitation and ensure accurate results. This has been addressed by the use of linear mixed models (LMM)<sup>89</sup>, which explicitly account for relatedness by conditioning on a genetic relatedness matrix (GRM). In the UK Biobank, this substantially increased the sample size to 456,422 individuals<sup>90</sup>, providing a ~30% increase in sample size compared to when using fixed-effect models. This approach is nowadays being implemented in the most efficient GWAS softwares that are capable of handling association testing for hundreds of thousands of individuals<sup>91,92</sup>.

Although modern biobanks collect phenotype and genotype data for hundreds of thousands of individuals, recent research suggests that millions of individuals are necessary to saturate the genome in association signals<sup>82</sup>. Therefore, it is crucial to continually increase the sample size of cohorts to improve the precision and accuracy of GWAS findings. A larger sample size permits more precise effect size estimation, increases statistical power, and enables the detection of smaller genetic effects. Currently, the most significant sample sizes for GWAS result from collaborative efforts among researchers who share data across multiple studies. By conducting meta-analyses, these efforts can surmount these challenges and augment the GWAS's effective sample size<sup>82</sup>.

### *Cost-effective GWAS*

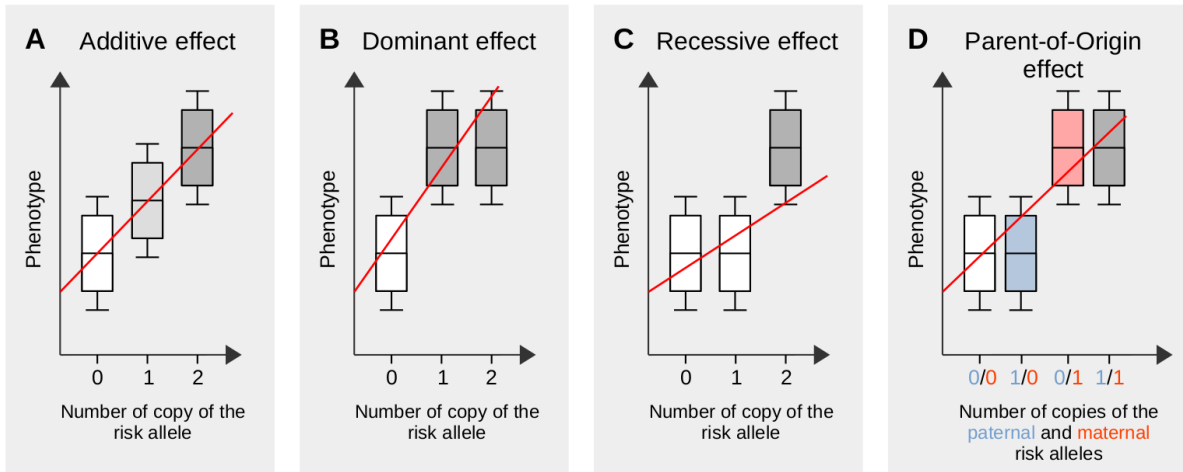
Increasing the sample size of GWAS studies can be challenging due to logistical and financial constraints. In the case of large biobanks, financial constraints are often the primary limiting factor. As a result, cost-effective alternatives to WGS have been employed, such as microarray genotyping. However, since it genotypes only a subset of genomic positions, the missing information must be predicted from a reference panel of haplotypes, a process referred to as imputation<sup>93</sup>. While this improves the chances of pinpointing causal variants, the accuracy of genotype imputation largely depends on a reference panel of phased haplotype that is used to perform the genotype predictions. The more accurate the phasing, the more informative the reference panel will be, and the better the imputation accuracy will be. In addition to the accuracy of the phasing, the size of the reference panel is also an important factor that affects the accuracy of imputation. A larger reference panel provides more genetic diversity, which increases the chances of finding haplotypes that match the missing genotypes in the study sample. This leads to better imputation accuracy, particularly for rare variants that may not be well-represented in smaller reference panels. Therefore, it is

essential to use a high-quality phasing algorithm and a reference panel that is large enough to capture the diversity of genetic variation present in the target population.

### *Genetic effects*

While efficient methods have been developed to associate genetic variants with complex traits and diseases, they are mostly designed to account for additive effects, which means that the effect of having two copies of the risk allele is assumed to be twice the effect of having a single copy of the risk allele. In other terms, an additive effect measures the independent contribution of each allele at the same locus. However, it is well known that there exists a wide range of genetic effects that can not be fully captured by linear models. The most common are probably the dominant and recessive effects, which measure the interaction of the two alleles at the same locus. In the dominant effect, one risk allele is sufficient to cause the phenotype, and there is no phenotypic difference between having one or two copies. In the recessive effect, two risk alleles are necessary to provoke the phenotype, and there is no phenotypic difference between having no or one copy ([Figure 6](#)).

Less common but not least, genetic effects can also be specific to haplotypes and depends on the epigenetic genetic background. Parent-of-Origin effects are phenotypic effects that depend on the parental origin of the risk allele. Therefore, the risk allele has an effect only when it locates on a specific parental haplotype, for example on the maternally inherited haplotype ([Figure 6D](#)). Such effects likely result from genomic imprinting, a mechanism of DNA methylations (i.e imprints) that silence genes. Imprints are sex-specific and established during the gametogenesis, meaning that the offspring inherit the paternal haplotype together with male-specific imprints, as well as the maternal haplotype with female-specific imprints. It results that some genes are always expressed from the same parental haplotype across individuals, and that risk alleles have an effect only when locating on this particular haplotype.



**Figure 6. Genetic effects on complex traits and diseases.**

Schematic representations of Additive (A), Dominant (B), Recessive (C) and Parent-of-Origin (D) effect. The phenotypic effects (y-axis) is stratified by the copy number of the risk allele (x-axis). Red lines indicate simulated linear estimates (i.e what can be captured using linear models). In (D), risk alleles are stratified by copy number and parental origin.

# Chapter I

## Haplotype estimation

---

The recent release of the whole-genome sequencing data for 150,199 individuals by the UK Biobank represents a significant milestone in the field of genomics, as it provides an unprecedented opportunity to assemble a large and diverse reference panel for genetic analyses. However, despite the potential benefits of such a resource, there are also significant challenges associated with processing and analyzing such a large amount of data. Current phasing methods are not well-suited to handle such a large amount of sequencing data, particularly when it comes to phase rare variants that are present in less than 1/1000 individuals. Phasing rare variants is particularly challenging and can lead to inaccurate haplotype estimates which have major impacts on downstream analysis<sup>94</sup>.



## Part I. Haplotype estimation in sequenced biobanks

The first part of this chapter outlines my contribution and involvement in the phasing software SHAPEIT5, which builds on previous versions of the software<sup>95</sup>. The manuscript is presented in [Appendix A](#). In this work, we introduce a new version of the SHAPEIT software, which is specifically designed to effectively and accurately phase rare genetic variants in large sequenced biobanks. The manuscript describes the phasing of the WGS and WES data of the UK Biobank cohort, showcases the accuracy of this phasing compared to concurrent methods and provides evidence for the utility of the phased haplotypes in detecting compound heterozygotes events.

I am the **co-first author** of this manuscript. It consists in collaboration within the research group, with three first authors that contributed equally to the work. My contributions were focused on **phasing the WGS and WES data**, as well as writing the corresponding manuscript sections. Furthermore, we created a dedicated website containing software and documentation, in which **I authored the phasing tutorial for the UK Biobank WGS, WES, and SNP-array data**.

This manuscript is currently in review in Nature Genetics. Alongside with the manuscript, **I am responsible for generating the haplotypes for the upcoming release of the UK Biobank data** in July 2023. The dataset comprises over 700 million variants across 200,031 individuals, with most being rare (~97% having MAF<0.1%). This call set will be the most efficient reference panel for imputing individuals of European ancestry. Consequently, it will likely be used in hundreds of GWAS.



## Part II. Haplotype estimates for genotype imputation

Phased haplotypes are commonly utilized as a reference panel for genotype imputation. In the part I of this chapter, the construction of a phased reference panel from UK Biobank WGS data is described, alongside with a brief summary of its effectiveness in CH event discovery and genotype imputation. However, a more detailed explanation of the utility of this reference panel for genotype imputation is presented in a separate manuscript.

The second part of this chapter outlines my involvement in a manuscript presenting a novel implementation of the low-coverage imputation software Glimpse<sup>96</sup>, named Glimpse2. The manuscript is presented in [Appendix B](#). The purpose of this software is to handle the recent improvement of reference panels for genotype imputation, since existing softwares do not scale efficiently with hundreds of thousands of reference haplotypes. The manuscript explains the method and demonstrates the effectiveness of using the UK Biobank phased haplotype as a reference panel for imputation, in comparison to alternative reference panels and across various populations. In addition, it also showcases the increased power of downstream GWAS using sequencing coverages as low as 0.5x compared to SNP array. I am the **second author** of this manuscript. My contribution to this manuscript includes **conducting the GWAS experiments**, assessing the **impact of sequencing coverages on GWAS accuracy**, writing the relevant section of the manuscript, and discussing the design of the experiments and the rationale of the project. The manuscript is currently in review in Nature Genetics.





# Chapter II

## The parental origin of haplotype estimates

---

Although phasing algorithms can be used to reconstruct haplotypes from genotype data, they cannot determine whether a haplotype was inherited from the mother or the father. The typical method to determine the parent-of-origin (PofO) of haplotypes compares the offspring haplotypes with parental genomes. However, due to the limited availability of parental genomes in modern biobanks, it can be challenging to assign the origin of haplotypes for a large number of individuals. What sparked my interest is that, while increasing the sample size in standard GWAS necessitates genotyping more individuals, there are numerous existing haplotypes for which the parent-of-origin information is not yet available. In the UK Biobank for example, the PofO can be inferred from parental genomes for ~5,000 individuals, representing only 1% of the available haplotypes. Therefore, increasing the number of individuals with parent-of-origin assigned can be done by developing innovative methods to analyze existing data.



## **Part I. Inference of the Parent-of-Origin of haplotypes**

The first part of this chapter describes the implementation of an approach to infer the parent-of-origin of alleles using close relatives instead of parental genomes. The article is presented in [Appendix C](#). Compared to the traditional approach using parental genomes, this allowed us to increase by 5 times the number of individuals with PofO assigned in the UK Biobank. Briefly, this approach combines (i) kinship estimates to identify close relatives and to group them into parental groups, (ii) IBD sharing and phasing to assign parental origin to haplotypes, and (iii) haploid imputation to increase the SNPs density. Finally we tested the parental origin of alleles for association with phenotypes to characterize parent-of-origin effects in the human genome.

**I am the main author of this article.** I worked on the study design, the implementation of the method, performed the GWAS experiments, wrote the manuscript and created an online database to host the summary statistics<sup>97</sup>.



## **Part II. Parental inheritance distortion**

The second part of chapter II illustrates an alternative use of the parent-of-origin of alleles, which consists in investigating genetic factors contributing to human fertility. The fundamental concept behind this approach is that alleles inherited less frequently from one parent may have a significant impact on reproductive functions or gametic competition. The advantage of such an approach is that, while GWAS studying genetic effect on human fertility usually use proxy phenotypes, such as the number of children ever born or the age at first birth, our approach does not require any phenotype. It only assesses distortion from the expected Mendelian inheritance pattern.

This is an ongoing project. **I am the main researcher on this project.** The [Appendix D](#) presents the preliminary results under the form of a draft manuscript of the current state of the project and is formatted into Abstract, Introduction, Results, Future analysis, and Methods section.



## Discussion

The central theme of this thesis is the importance of haplotypes in genomic analysis. Haplotypes are derived from genotype data and have various applications, including detecting compound heterozygote events, studying parental origin and their effects on complex traits, and investigating Mendelian inheritance patterns. The primary motivation for focusing on haplotypes is their underutilization in biobanks, despite being obtainable at no extra cost, beyond computations, from existing data. As a result, this thesis aims to showcase how efficient method developments can leverage haplotypes from existing genotype data to maximize the potential of current biobanks.

Two main approaches have been developed to achieve this goal. The first involves estimating haplotypes from genotype data, which is notably essential for assembling large reference panels of haplotypes used for genotype imputation. The second approach involves inferring the parental origin of haplotypes using available close relatives, which is a significant breakthrough in parent-of-origin effect mapping since it largely increases the sample size compared to the traditional inference that uses parental genomes.

Although the two chapters in this thesis are distinct, they are closely linked. Accurately estimating haplotypes is crucial for performing PofO inference of resulting haplotypes. The final section of this thesis outlines potential future applications and improvements of both chapters, culminating in a novel perspective on evaluating the phenotypic impact of rare variants.



## Haplotype estimation

Chapter I of this thesis presents a novel implementation of the SHAPEIT phasing software that has been specifically designed to cope with the large number of individuals and variants contained in modern sequenced biobanks, with a particular focus on the phasing of rare variants. In addition, the phasing of the UK Biobank's WGS and WES data is described, along with the application of the resulting haplotypes to compound heterozygous calling and array imputation, showcasing the utility of such phasing for genetic analysis.

This research article focuses on the phasing of the initial UK Biobank WGS release, encompassing 150,119 individuals. However, subsequent releases have expanded the dataset to 200,031 individuals, with plans to include approximately 500,000 individuals by November 2023<sup>98</sup>. As sample sizes increase, novel haplotype estimation becomes necessary to improve phasing accuracy. This is particularly important for rare variants, where phasing accuracy improves with a larger minor allele count<sup>71</sup>. Larger sample sizes provide more accurate phasing for variants, increasing their value for downstream analysis. Therefore, it is crucial to update the haplotype estimation method for each release to ensure the best possible accuracy for genetic analysis. The method and pipeline developed in this study will be used to process the upcoming release of the UK Biobank and provide the research community with the most accurate haplotype estimates possible.

The limited knowledge about the phase of rare alleles in large cohorts of unrelated individuals previously limited several research areas. However, accurate phasing of rare variants in this study allows for their inclusion in downstream CH event detection, which is crucial since rare LoF variants are often the primary contributors to disease<sup>99,100</sup> and are potential therapeutic targets<sup>32</sup>. Previously, CH investigations were limited to families, where parental genomes were used to determine independent inheritance of two mutations within the same gene. This approach helped to assess the contribution of rare and severe CH events to diseases but did not provide insight into the prevalence of CH events in healthy populations. By expanding the sample size used to detect CH events, a better understanding of the genetic basis of diseases, especially regarding gene essentiality, can be achieved.

The contribution of phasing to CH event detection will become particularly important with two key aspects. First, the upcoming release of the UK Biobank WGS data, scaling up to ~500,000 individuals, will increase the number of observed gene double knockout and

contribute to a better characterization of CH events. Secondly, haplotype estimates serve as a reference panel for imputation. Larger reference panel sizes result in more accurate imputation, with a significant increase in accuracy for rare variants. This accurate imputation of rare variants enables their use in the CH event detection process.

A second manuscript developed by the research group and introduced in Chapter II demonstrated the use of a reference panel derived from UK Biobank WGS data for genotype imputation, specifically for large biobanks that employ microarray genotyping technology to reduce costs<sup>101</sup>. For instance, the OFH project aims to recruit 5 million UK participants and will likely use microarray genotyping technology<sup>102</sup>. Since the ancestry of these individuals will be similar to the UK Biobank cohort, **the reference panel I generated will likely be used to impute those 5 millions individuals**. In addition this reference panel is the most effective to impute any cohort of European ancestry and **will likely be used in hundreds of GWAS studies**. The current reference panel constructed using 150,000 individuals from the UK Biobank WGS data provides high accuracy imputation of variants found in 1/1,000 individuals. However, the upcoming release of the UK Biobank WGS is expected to significantly improve imputation accuracy, enabling the imputation of variants present in 1/10,000 individuals with sufficient accuracy for downstream analyses. This improved reference panel will be particularly advantageous for enhancing the imputation of rare variants, including protein-modifying variants that are only present in a few copies in the UK Biobank WES. Consequently, the ability to map CH events using imputed variants will be strengthened, as LoF variants will be more common in the population with the larger sample size.

Although the phasing performed on the UK Biobank WGS data set is highly accurate, there is still room for improvement. The current call set provides a phasing probability per variant per individual, enabling easy identification and exclusion of badly phased variants for downstream analysis. However, given the frequency of singletons in the dataset ( $\sim 46\%$ )<sup>60</sup>, and the low phase confidence reported for singletons (mean accuracy =  $\sim 65\%$ ), losing this amount of information is undesirable. Thus, improving the phase at singleton and any other low confidence phasing sites is crucial for efficient downstream analysis, such as detecting CH events. To address this issue, our group is currently working on a follow-up which briefly consists in identifying variants with low phasing probability for each individual, and searching for nearby common variants with high phasing probability that co-localize on the

same read as the low probability rare variant. This approach allows to deduce the phase of the rare variants as being the same as the phase of the common variants. While computationally demanding, this approach will allow to re-localize poorly phased alleles onto the correct haplotype and to considerably increase the phasing accuracy, in particular at singletons.

Finally, the accurate phase of rare variants provided by my work opens novel perspectives in large-scale association analysis of rare variants. Current methods for assessing the impact of rare variants on complex traits involve burden tests, which aggregate deleterious variants within a gene and test the resulting gene burden for association with a trait<sup>28</sup>. While these tests typically focus on protein-modifying variants, alternative approaches are emerging, such as testing rare intergenic variants within a gene cis-window that are likely located in regulatory elements. However, no previous study has integrated haplotype information into these analyses due to the lack of accurate phase at rare variants. Using our haplotype estimation method in the UK Biobank WES and WGS, researchers can test the gene burden at the haplotype level. In particular, this approach is interesting for investigating the burden of rare variants at known imprinted genes, for which only the paternal or maternal copy of the gene is expressed. Indeed, [Appendix C](#) shows that testing paternal and maternal alleles separately leads to stronger significance compared to normal additive tests in case of parent-of-origin effects. Therefore, it is reasonable to expect that burden testing at imprinted genes will be more efficient when considering the parental haplotypes separately. Considering that burden testing usually involves a small power due to the limited number of individuals carrying rare variants, this approach has the potential to increase the characterization of the effect of rare variants at imprinted loci.

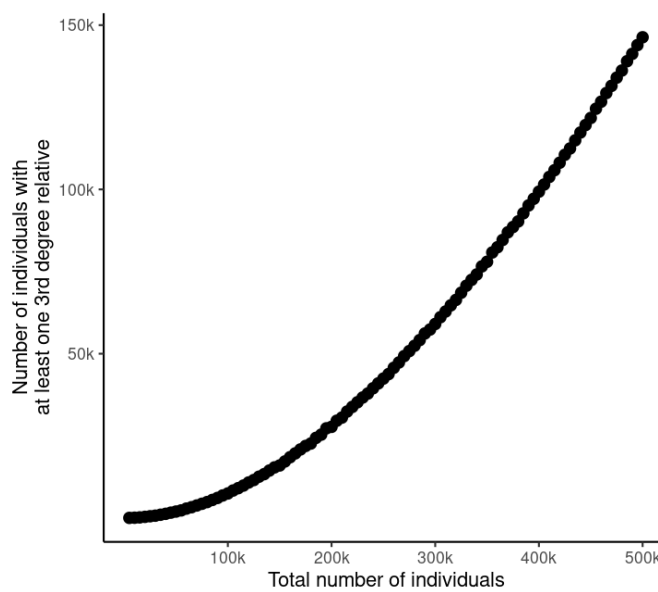
## The parental origin of haplotype estimates

Chapter II presents an approach to infer the PofO of individuals without any prior knowledge of genealogy or available parental genomes. This approach leads to a considerable expansion in the number of individuals with PofO assignments, which enhances the statistical power to discover PofO effects. Furthermore, it showcases the importance of the parent-of-origin information in the identification of genetic factors associated with human fertility. Despite caveats of this approach as discussed in the published article, it constitutes a promising alternative to family-based studies since it benefits from the increasing sample size of biobanks.

Indeed, as the number of individuals in a cohort increases, the number of individuals with PofO inferred can be proportionally increased. Moreover, the number of individuals with at least a third-degree relative in the cohort increases quadratically with the total number of individuals ([Figure 7](#)). As a result, since the approach presented relies on the availability of close relatives, it suggests that it has an exponential potential on very large datasets. The upcoming Our Future Health (OFH) project<sup>102</sup>, which aims to recruit 5 million UK participants, will be enriched in close relatives since individuals will be recruited from the same population. This is anticipated to increase the PofO sample size to approximately 20% of the total number of individuals, representing **1 million individuals with PofO inference**. First, this cohort will constitute the largest available with PofO inference and significantly strengthen the benefit of this approach that exploits the inherent degree of relatedness of modern biobanks. Second, considering that 5.4 million individuals have allowed to saturate the association signal for standing height, I anticipate that one million individuals will provide **a saturated map of PofO effects across the human genome**.

The PofO approach can identify genetic loci with PofO effects on phenotypes, but it only provides candidate genes and does not specify the parental-specific expression nature of these candidates. Although PofO loci discovered can be associated with imprinted genes, the PofO associations can also underlie a more complex mechanism in which non-imprinted genes interact with imprinted genes to generate PofO effects<sup>103</sup>, which requires further investigation. To advance our understanding of PofO effects on phenotypes, omics data, specifically RNA sequencing, needs to be integrated. First, by combining GWAS signals with RNA sequencing data, novel candidate imprinted genes can be identified, and RNA sequencing can confirm

the PofO specific expression patterns, improving classification accuracy. Second, by analyzing gene co-expression and co-regulation via joint expression quantitative trait loci (eQTL) analysis, this approach can identify networks of gene interactions consisting of imprinted and non-imprinted genes that have the potential to cause PofO effects and contribute to complex traits. This would be particularly interesting since current catalogs of imprinting genes are thought to be incomplete because they imperfectly capture imprinting in adults<sup>104</sup>.



**Figure 7. Relatedness in the UK Biobank cohort.**

Number of individuals with at least one third degree relative (y-axis) among an increasing number of individuals randomly sampled from the UK Biobank cohort (x-axis).

The majority of knowledge about imprinted genes comes from animal breeding, where initially, genomic regions likely to contain imprinted genes have been identified by phenotypic screening of uniparental disomy mice<sup>105</sup>. In contrast, human studies have focused on investigating the parental-specific allelic expression of candidate imprinted genes using family data, which has confirmed some of the imprinted genes identified in animal studies<sup>106,107</sup>. In addition, human imprinted genes involved in severe disorders have been characterized, such as in the Prader-Willi and Angelman syndromes<sup>108,109</sup>. However, the

current classification of imprinted genes in humans is considered incomplete and mainly includes imprinted genes with complete imprinting patterns. Recent studies suggest that subtle imprinting patterns may exist in humans, but these are challenging to detect due to the small differences in parental allelic expression and require large-scale family transcriptome data<sup>104,110</sup>. However, most existing large-scale expression data contain unrelated individuals, making it difficult to study such patterns. In addition, the degree of parental-specific monoallelic expression varies depending on the tissue and developmental stage<sup>111</sup>, which adds another layer of complexity to the study of imprinted genes. To address this challenge, a recent study aimed to identify novel candidate imprinted genes using allele-specific expression data from different tissues<sup>104</sup>. Although unrelated individuals were used to identify the candidate genes, they were validated notably using family transcriptome data to distinguish the paternally and maternally inherited alleles, emphasizing the current need of including family data in the study of imprinted genes.

In this context, the PofO inference method developed here can provide a significant advance in the identification of candidate imprinted genes from gene expression data, since it allows to determine the PofO of alleles across a set of unrelated individuals, and therefore does not require family data to validate the findings. Notably, it would allow the study of imprinting at two different layers. First, for examining the parent-of-origin specific allelic expression of genes. Second, by scanning for PofO specific association between genetic variants and gene expression level, namely PofO eQTLs. However, the caveat of such approaches is to require a transcriptome cohort large enough to contain close relatives, typically in the order of tens of thousands of individuals.

One potential solution to this challenge is to completely eliminate the need for family data and to develop a method to infer the PofO of alleles at the gene expression level, by taking advantage of the current classification of genes exhibiting a complete imprinting expression pattern. This can be achieved by mapping RNA reads to haplotypes and utilizing genes with complete imprinting as a reference for haplotype labeling: RNA reads corresponding to maternally expressed genes will map to the maternally inherited haplotype, while RNA reads corresponding to paternally expressed genes will map to the paternally inherited haplotype. The utilization of this approach is expected to increase the number of individuals for which the PofO can be inferred as it does not depend on the availability of close relatives, and would be a promising approach if large biobanks start RNA sequencing. However, this

approach is limited by the extent at which phasing can be achieved. Specifically, current phasing methods can efficiently resolve the co-inheritance of alleles located on the same chromosome (i.e. intra-chromosomal phasing)<sup>71</sup>. Regrettably, these methods are incapable of resolving inter-chromosomal phasing, meaning that the first haplotype of a given chromosome may not necessarily be co-inherited with the first haplotype of the next chromosome. As a result, the proposed solution necessitates the presence of at least one imprinted gene per chromosome to label the haplotypes. In addition, it is also limited by the tissue-specific nature of imprinting<sup>104,111</sup>, which might limit the use of the current set of known imprinted genes. Therefore, the approach would be more effective if applied to multiple tissues, which would enable for a better understanding of the tissue-specificity of imprinting and improve the accuracy of haplotype labeling.

## Conclusion

Despite more than 15 years of GWAS research, the scientific community's interest in association analysis has not waned. While GWAS has been successful in identifying genetic variants associated with complex traits, there is still room for improvement in utilizing the vast amount of existing data. To fully leverage the potential of existing biobanks data, there is a need to enhance data processing, inference, and testing methodologies. In this thesis, I developed efficient methods for inferring haplotypes and their parental origin from existing biobanks, and I demonstrated the practical applications of my inferences. I am confident that my work will have a significant impact on the community for various reasons.

The phased haplotypes generated as part of this thesis, which will be continuously updated with the upcoming data releases, constitute a resource that will be employed in various analyses. Firstly, these haplotypes enable the detection of compound heterozygote events in large-scale population cohorts. Secondly, they enable integrating the phase of rare variants in gene burden analysis and allow for these analyses to be conducted at the haplotype level. Thirdly, they represent the best available reference panel for the European population. Consequently, the reference panel that I generated will be utilized in numerous GWAS studies.

The PofO inference method I developed is a significant advance that enables the study of PofO effects in large-scale biobanks. This methodology is expected to be employed in numerous large-scale cohorts, uncovering numerous novel signals that will improve our current comprehension of PofO effects on complex traits and the underlying biology of the imprinting mechanism. Within a few years, this methodology will likely enable the study of PofO effects in one million individuals, providing a saturated map of PofO effects across the entire human genome.

Although my work demonstrates two methods for maximizing the potential of current biobanks, additional innovative strategies are necessary. To uncover novel associations that traditional methods may have missed, future efforts should concentrate on creating and combining diverse methodologies.





## References

1. Kwok, A. J., Mentzer, A. & Knight, J. C. Host genetics and infectious disease: new tools, insights and translational opportunities. *Nat. Rev. Genet.* **22**, 137–153 (2020).
2. Lappalainen, T., Scott, A. J., Brandt, M. & Hall, I. M. Genomic analysis in the age of human genome sequencing. *Cell* **177**, 70–84 (2019).
3. Collins, F. S. & Fink, L. The Human Genome Project. *Alcohol Health Res. World* **19**, 190–195 (1995).
4. Hood, L. & Rowen, L. The Human Genome Project: big science transforms biology and medicine. *Genome Med.* **5**, 1–8 (2013).
5. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
6. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
7. Gibbs, R. A. The Human Genome Project changed everything. *Nat. Rev. Genet.* **21**, 575–576 (2020).
8. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
9. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* vol. 604 437–446 Preprint at <https://doi.org/10.1038/s41586-022-04601-8> (2022).
10. Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nat. Rev. Genet.* **21**, 243–254 (2020).
11. Kaye, A. M. & Wasserman, W. W. The genome atlas: navigating a new era of reference genomes. *Trends Genet.* **37**, 807–818 (2021).
12. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
13. Ku, C. S., Loy, E. Y., Salim, A., Pawitan, Y. & Chia, K. S. The discovery of human genetic variations and their use as disease markers: past, present and future. *J. Hum. Genet.* **55**, 403–415 (2010).
14. Fiegler, H. *et al.* Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Res.* **16**, 1566–1574 (2006).
15. Liu, Z. *et al.* Towards accurate and reliable resolution of structural variants for clinical diagnosis. *Genome Biol.* **23**, 68 (2022).
16. Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).
17. Conrad, D. F. & Hurles, M. E. The population genetics of structural variation. *Nat. Genet.* **39**, S30–6 (2007).
18. Kunkel, T. A. Evolving views of DNA replication (in) fidelity. *Cold Spring Harb. Symp. Quant. Biol.* **74**, 91–101 (2009).
19. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. Rates of spontaneous mutation. *Genetics* **148**, 1667–1686 (1998).
20. Senft, A. D. & Macfarlan, T. S. Transposable elements shape the evolution of mammalian development. *Nat. Rev. Genet.* **22**, 691–711 (2021).

21. Dubrova, Y. E., Plumb, M., Gutierrez, B., Boulton, E. & Jeffreys, A. J. Transgenerational mutation by radiation. *Nature* vol. 405 37–37 Preprint at <https://doi.org/10.1038/35011135> (2000).
22. Rojano, E., Seoane, P., Ranea, J. A. G. & Perkins, J. R. Regulatory variants: from detection to predicting impact. *Brief. Bioinform.* **20**, 1639–1654 (2019).
23. Furey, T. S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.* **13**, 840–852 (2012).
24. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).
25. Molecular quantitative trait loci. *Nat. Rev. Methods Primers* **3**, (2023).
26. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* **12**, 32–42 (2010).
27. Livesey, B. J. & Marsh, J. A. Interpreting protein variant effects with computational predictors and deep mutational scanning. *Dis. Model. Mech.* **15**, (2022).
28. Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genom* **2**, 100168 (2022).
29. Ensembl Variant Effect Predictor (VEP). <https://www.ensembl.org/info/docs/tools/vep/index.html>.
30. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
31. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
32. Cohen, J. C., Boerwinkle, E., Mosley, T. H., Jr & Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).
33. Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat. Genet.* **53**, 942–948 (2021).
34. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
35. Moore, L. *et al.* The mutational landscape of human somatic and germline cells. *Nature* **597**, 381–386 (2021).
36. Kavanagh, P. L., Fasipe, T. A. & Wun, T. Sickle Cell Disease: A Review. *JAMA* **328**, 57–68 (2022).
37. McColgan, P. & Tabrizi, S. J. Huntington’s disease: a clinical review. *Eur. J. Neurol.* **25**, 24–34 (2018).
38. Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.* **1**, 40–47 (2000).
39. Marston, A. L. & Amon, A. Meiosis: cell-cycle controls shuffle and deal. *Nat. Rev. Mol. Cell Biol.* **5**, 983–997 (2004).
40. Choi, K. & Henderson, I. R. Meiotic recombination hotspots - a comparative view. *Plant J.* **83**, 52–61 (2015).
41. Greenwood, T. A., Rana, B. K. & Schork, N. J. Human haplotype block sizes are negatively correlated with recombination rates. *Genome Res.* **14**, 1358–1361 (2004).

42. Crawford, D. C. & Nickerson, D. A. Definition and clinical importance of haplotypes. *Annu. Rev. Med.* **56**, 303–320 (2005).
43. Wall, J. D. & Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics* vol. 4 587–597 Preprint at <https://doi.org/10.1038/nrg1123> (2003).
44. Thompson, E. A. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics* **194**, 301–326 (2013).
45. Ardlie, K. G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **3**, 299–309 (2002).
46. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
47. Jennings, H. S. The Numerical Results of Diverse Systems of Breeding, with Respect to Two Pairs of Characters, Linked or Independent, with Special Relation to the Effects of Linkage. *Genetics* **2**, 97–154 (1917).
48. Slatkin, M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**, 477–485 (2008).
49. Lewontin, R. C. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* **49**, 49–67 (1964).
50. Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687–695 (1977).
51. Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).
52. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
53. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
54. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26**, 1146–1153 (2008).
55. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34**, 518–524 (2016).
56. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348–1365 (2021).
57. Marx, V. Method of the year: long-read sequencing. *Nat. Methods* **20**, 6–11 (2023).
58. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
59. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).
60. Halldorsson, B. V. *et al.* The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
61. Bumgarner, R. Overview of DNA microarrays: types, applications, and their future. *Curr. Protoc. Mol. Biol.* **Chapter 22**, Unit 22.1. (2013).
62. Bier, F. F. *et al.* DNA microarrays. *Adv. Biochem. Eng. Biotechnol.* **109**, 433–453 (2008).
63. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.

- Nature* **562**, 203–209 (2018).
64. Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).
  65. Ross, J. P., Dion, P. A. & Rouleau, G. A. Exome sequencing in genetic disease: recent advances and considerations. *F1000Res.* **9**, (2020).
  66. Lou, R. N., Jacobs, A., Wilder, A. P. & Therkildsen, N. O. A beginner’s guide to low-coverage whole genome sequencing for population genomics. *Mol. Ecol.* **30**, 5966–5993 (2021).
  67. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).
  68. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
  69. Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE Inst. Electr. Electron. Eng.* **77**, 257–286 (1989).
  70. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).
  71. 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. *bioRxiv* 2022.10.19.512867 (2022) doi:10.1101/2022.10.19.512867.
  72. Browning, B. L. & Browning, S. R. Statistical phasing of 150,119 sequenced genomes in the UK Biobank. *Am. J. Hum. Genet.* **110**, 161–165 (2023).
  73. Antonarakis, S. E. & Beckmann, J. S. Mendelian disorders deserve more attention. *Nat. Rev. Genet.* **7**, 277–282 (2006).
  74. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
  75. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
  76. Mayhew, A. J. & Meyre, D. Assessing the Heritability of Complex Traits in Humans: Methodological Challenges and Opportunities. *Curr. Genomics* **18**, 332–340 (2017).
  77. Barry, C.-J. S. *et al.* How to estimate heritability: a guide for genetic epidemiologists. *Int. J. Epidemiol.* (2022) doi:10.1093/ije/dyac224.
  78. Zaitlen, N. & Kraft, P. Heritability in the genome-wide association era. *Hum. Genet.* **131**, 1655–1664 (2012).
  79. Guo, M. H., Hirschhorn, J. N. & Dauber, A. Insights and Implications of Genome-Wide Association Studies of Height. *J. Clin. Endocrinol. Metab.* **103**, 3155–3168 (2018).
  80. Silventoinen, K. *et al.* Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res.* **6**, 399–408 (2003).
  81. Macgregor, S., Cornes, B. K., Martin, N. G. & Visscher, P. M. Bias, precision and heritability of self-reported and clinically measured height in Australian twins. *Hum. Genet.* **120**, 571–580 (2006).
  82. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).
  83. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

84. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450 (2010).
85. Young, A. I. Solving the missing heritability problem. *PLoS Genet.* **15**, e1008222 (2019).
86. Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).
87. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 1–21 (2021).
88. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
89. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
90. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* **51**, 1749–1755 (2019).
91. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
92. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
93. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
94. Appadurai, V. *et al.* Accuracy of haplotype estimation and whole genome imputation affects complex trait analyses in complex biobanks. *Commun Biol* **6**, 101 (2023).
95. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 1–10 (2019).
96. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 120–126 (2021).
97. The Parent-of-Origin Effects Catalog. <https://poedb.dcsr.unil.ch/>.
98. Genetic data. <https://www.ukbiobank.ac.uk/enable-your-research/about-our-data/genetic-data>.
99. Niu, Y. *et al.* Loss-of-Function Genetic Screening Identifies Aldolase A as an Essential Driver for Liver Cancer Cell Growth Under Hypoxia. *Hepatology* **74**, 1461–1479 (2021).
100. Bayona, A. *et al.* Loss-of-function mutation of PCSK9 as a protective factor in the clinical expression of familial hypercholesterolemia: A case report. *Medicine* **99**, e21754 (2020).
101. Rubinacci, S., Hofmeister, R., da Mota, B. S. & Delaneau, O. Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. *bioRxiv* 2022.11.28.518213 (2022) doi:10.1101/2022.11.28.518213.
102. Our Future Health. *Our Future Health* <https://ourfuturehealth.org.uk/>.
103. Macias-Velasco, J. F. *et al.* Parent-of-origin effects propagate through networks to shape metabolic traits. *Elife* **11**, (2022).
104. Baran, Y. *et al.* The landscape of genomic imprinting across diverse adult human tissues. *Genome Res.* **25**, 927–936 (2015).

105. Cattanach, B. M. & Kirk, M. Differential activity of maternally and paternally derived chromosome regions in mice. *Nature* **315**, 496–498 (1985).
106. Ishida, M. & Moore, G. E. The role of imprinted genes in humans. *Mol. Aspects Med.* **34**, 826–840 (2013).
107. Pilvar, D., Reiman, M., Pilvar, A. & Laan, M. Parent-of-origin-specific allelic expression in the human placenta is limited to established imprinted loci and it is stably maintained across pregnancy. *Clin. Epigenetics* **11**, 1–14 (2019).
108. Morison, I. M., Ramsay, J. P. & Spencer, H. G. A census of mammalian imprinting. *Trends Genet.* **21**, 457–465 (2005).
109. Morison, I. M., Paton, C. J. & Cleverley, S. D. The imprinted gene and parent-of-origin effect database. *Nucleic Acids Res.* **29**, 275–276 (2001).
110. Henckel, A. & Arnaud, P. Genome-wide identification of new imprinted genes. *Brief. Funct. Genomics* **9**, 304–314 (2010).
111. Prickett, A. R. & Oakey, R. J. A survey of tissue-specific genomic imprinting in mammals. *Mol. Genet. Genomics* **287**, 621–630 (2012).

## Appendix A

# Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank

This manuscript is presented in [Chapter I](#).

The online version and the supplementary material can be downloaded from <https://www.nature.com/articles/s41588-023-01415-w>.







# Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank

Received: 27 October 2022

Accepted: 4 May 2023

Check for updates

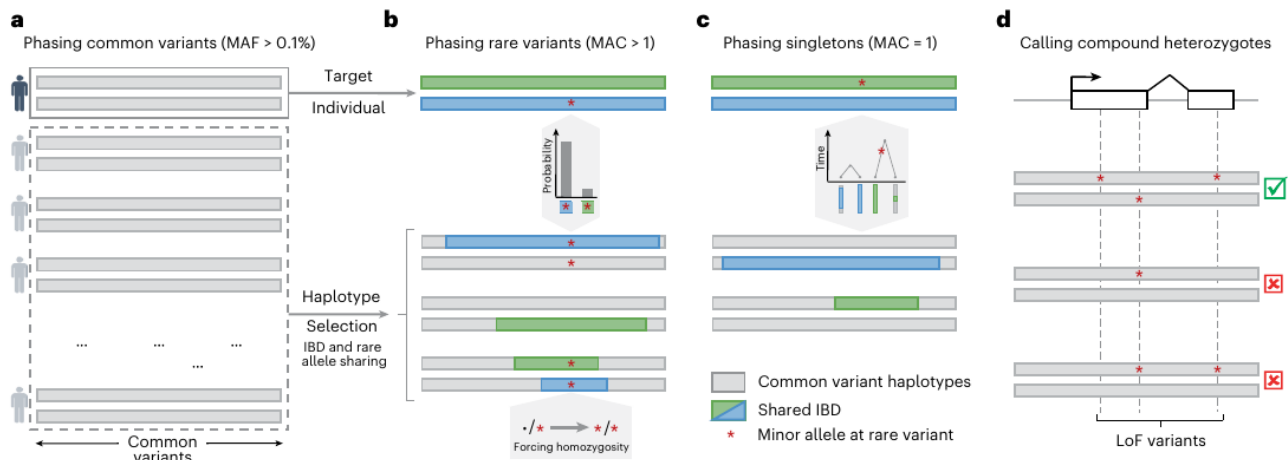
Robin J. Hofmeister<sup>1,2</sup>, Diogo M. Ribeiro<sup>1,2</sup>, Simone Rubinacci<sup>1,2</sup>  
& Olivier Delaneau<sup>1</sup>✉

Phasing involves distinguishing the two parentally inherited copies of each chromosome into haplotypes. Here, we introduce SHAPEIT5, a new phasing method that quickly and accurately processes large sequencing datasets and applied it to UK Biobank (UKB) whole-genome and whole-exome sequencing data. We demonstrate that SHAPEIT5 phases rare variants with low switch error rates of below 5% for variants present in just 1 sample out of 100,000. Furthermore, we outline a method for phasing singletons, which, although less precise, constitutes an important step towards future developments. We then demonstrate that the use of UKB as a reference panel improves the accuracy of genotype imputation, which is even more pronounced when phased with SHAPEIT5 compared with other methods. Finally, we screen the UKB data for loss-of-function compound heterozygous events and identify 549 genes where both gene copies are knocked out. These genes complement current knowledge of gene essentiality in the human genome.

Modern genetic association studies are increasingly based on whole-genome or whole-exome sequencing (WGS/WES) for hundreds of thousands of samples collected as part of nationwide biobanking initiatives<sup>1,2</sup>. Compared with previous studies based on single nucleotide polymorphism (SNP) arrays, WGS and WES data can identify rare variants (e.g., minor allele frequency below 1%), allowing a systematic characterization of their contribution to trait heritability<sup>3</sup>, functional relevance<sup>4</sup> and effects on various traits and diseases<sup>5,6</sup>. In this context, haplotype phasing of rare variants, which involves distinguishing the two parentally inherited copies of each chromosome into haplotypes, adds a layer of biologically relevant information and unlocks new analyses. For instance, phasing is crucial to identify compound heterozygous events, which occur when both copies of a gene contain nonidentical, heterozygous mutations. In the case of Mendelian disorders, compound heterozygosity is one of the most common inheritance models for rare recessive diseases in nonconsanguineous individuals<sup>7,8</sup>. Previous efforts to identify compound heterozygous events in large cohorts provided valuable insights, yet these either relied on imputed data<sup>9</sup> or ignored

phasing information<sup>6</sup>. Compound heterozygous event identification requires high-confidence phase information to be considered when rare variants are analyzed, such as in gene-based burden test analysis<sup>10</sup>. The most common approach to phase rare variants without parental genomes or long-reads in large cohorts of individuals is statistical phasing, which leverages information across individuals to make estimation of haplotypes<sup>11</sup>. This technique is well established for common variants typed on SNP arrays, where phase information is used, for instance, to perform genotype imputation<sup>12</sup>, admixture analysis<sup>13</sup> and genealogy estimation<sup>14</sup>. Phasing methods have been optimized to scale to the thousands of samples in modern SNP array datasets, and the time is ripe to do the same for the millions of rare variant sites present in WGS/WES datasets. As an example, the WGS data for 150,119 UKB samples comprise three orders of magnitude more variants than the Axiom array data, around 96% of them having a minor allele frequency (MAF) below 0.1%. Phasing large scale WGS/WES datasets is challenging and new methods able to handle large amounts of rare variants are now emerging<sup>15</sup>. Recently, a computationally efficient solution for rare

<sup>1</sup>Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. <sup>2</sup>These authors contributed equally: Robin J. Hofmeister, Diogo M. Ribeiro, Simone Rubinacci. ✉e-mail: [olivier.delaneau@unil.ch](mailto:olivier.delaneau@unil.ch)



**Fig. 1 | Rationale of SHAPEIT5.** **a**, All samples are phased at common variants (MAF  $\geq$  0.1%). **b**, Phasing of a given rare variant onto the haplotypes at common variants. Conditioning haplotypes used in the estimation share long matches with the target (green and blue) and are not monomorphic at the rare variant. Since heterozygous genotypes for the rare variant are unphased,

the minor alleles at those are assumed to be on both haplotypes (i.e., forcing homozygosity). **c**, Singleton phasing by assigning the new allele on the target haplotype with the shortest match. **d**, Compound heterozygous event mapping based on the rare variant phasing (a–c).

variant phasing has been implemented in Beagle v.5.4 (refs. 16,17), in which common and rare variants are phased separately: in a first step, a standard phasing method is used to obtain haplotypes at common variants, and in a second step rare heterozygous sites are phased onto the resulting haplotypes using genotype imputation technique. This type of strategy, based on haplotype scaffolds, has been used in other contexts, such as in genotype imputation<sup>18</sup>, integration of family data<sup>19</sup> and external phasing information<sup>20</sup>.

In this work, we describe SHAPEIT5, a method designed to accurately phase rare variants in large WGS/WES datasets, including singletons, with moderate accuracy, while attributing phasing confidence scores. We applied it to estimate haplotypes for 150,119 and 452,644 UKB samples with WGS and WES data, respectively. We demonstrate the benefit of using these two haplotype collections as reference panels for SNP array imputation and finally show that the phase inferred at rare variants in the WES dataset can be screened to reliably identify compound heterozygous loss-of-function (LoF) mutations, probably leading to complete gene knockouts.

## Results

### Overview of the SHAPEIT5 phasing method

SHAPEIT5 performs haplotype phasing of WGS or WES data using three different phasing models, each focusing on a specific type of variants: (1) common variants are phased using the SHAPEIT4 model<sup>20</sup>, (2) rare variants are phased onto the resulting haplotypes using an imputation model and (3) singletons are phased using a coalescent-inspired model. See Fig. 1 for an illustration of the phasing scheme. Common variants are defined as having a MAF above 0.1% and are phased using an optimized version of the SHAPEIT4 algorithm, known to perform well on large sample sizes (Fig. 1a).

The resulting haplotypes are used in a second stage as a scaffold onto which rare variants (MAF < 0.1%) are phased one after another, following a methodology similar to that of Beagle v.5.4 (refs. 16,17). To cope with the large numbers of rare variants, SHAPEIT5 uses a sparse data representation for rare variants: only genotypes carrying at least one copy of the minor allele are stored in memory and considered for computation, thereby discarding all genotypes being homozygous for the major allele<sup>21,22</sup>. SHAPEIT5 phases each rare heterozygous genotype conditioning on a small number of informative haplotypes (Fig. 1b).

For a specific rare variant, these conditioning haplotypes are chosen so that (1) they belong to samples being locally identical-by-descent (IBD) with the target sample and (2) they are polymorphic at the rare variant (that is, at least a few carry a copy of the minor allele). To comply with the first requirement, SHAPEIT5 uses a positional Burrows-Wheeler transform (PBWT) data structure<sup>23</sup> built on all the scaffold haplotypes at common variants. This allows rapid identification of shared segments between haplotypes. To ensure representation of the minor allele in the conditioning set (second requirement), the method performs a second PBWT pass restricted to the subset of samples carrying a copy of the minor allele. This second pass is performed efficiently by leveraging the sparse representation of the genotypes. We then determine the alleles carried by the conditioning haplotypes at the rare variant of interest, which is straightforward when homozygous. However, when a conditioning sample is heterozygous, the allele carried by each of its two haplotypes is unknown. In this case, our model assumes that both haplotypes carry the minor allele as done in Beagle v.5.4 (refs. 16,17). Once the conditioning set of haplotypes is assembled, SHAPEIT5 uses the Li and Stephens model<sup>24</sup> to get the most likely phase configuration of the rare allele by imputation (that is, either on its first or second target haplotype; Supplementary Fig. 1). The strength of our model resides in the guarantee that each rare heterozygous genotype is phased from a conditioning set containing long haplotype matches and carrying copies of the two possible alleles.

For singleton variants (minor allele count (MAC) of 1), SHAPEIT5 uses another phasing model that (1) assumes singletons to be recent mutation events and (2) leverages IBD sharing patterns between haplotypes to make inference (Fig. 1c). Specifically, our model identifies the longest possible match in the dataset for each target haplotype. By definition, these matches point to haplotypes sharing recent common ancestors with the target and their lengths indicate the number of generations separating them: the shorter the match, the older the common ancestor. Our model assumes that an older common ancestor means more time for a mutation to occur on that lineage and therefore assigns the minor alleles of singletons to the target haplotype with the shortest match<sup>25</sup>.

### Phasing UKB exomes and genomes

We used SHAPEIT5 to phase haplotypes for three different UKB sequencing datasets: (1) WGS data on chromosome 20 for 147,754 samples and

around 13.8 million SNPs and indels after quality control, (2) WES data for 452,644 samples and around 26 million variants and (3) WGS data for the full set of 150,119 samples and around 603 million variants. For (1) and (2), we included only samples for which Axiom array data are available and excluded parental genomes for duos (parent–offspring pairs) and trios (parent–offspring triplets) to measure phasing accuracy in the offspring. Numbers of samples, trios, duos and variants after quality control are given in Supplementary Table 1. Phasing of the WES dataset was performed for each chromosome independently and phasing of the WGS was done in overlapping chunks of around 4.5 Mb on average to leverage parallelization on the UKB Research Analysis Platform (RAP). We compare the performance of our method with Beagle v.5.4 (refs. 16,17) (default parameters) on the WES and WGS datasets on chromosome 20.

### Phasing performance in the UKB data

To assess phasing performance, we used the available white British trios (719 for WES, 31 for WGS) and duos (432 for WGS). Using these, we (1) derived a true set of haplotypes for the offspring using inheritance logic, (2) performed statistical phasing of the WES and WGS datasets after having excluded parental genomes and (3) compared the offspring haplotypes obtained by statistical phasing with the true set obtained in (1). We assessed how close the two sets of haplotypes are by measuring the switch error rate (SER), which is the fraction of successive heterozygous genotypes phased differently. When looking at overall SER using different validation sets (duos, trios), different sets of variants (all variants or common variants only) and different sample sizes, we found minor differences between SHAPEIT5 and Beagle v.5.4 on the WGS data (Supplementary Fig. 2a–c). However, when considering only Axiom array positions, lower SER is observed with SHAPEIT5 (Supplementary Fig. 2d). We did not find the same pattern when phasing the Axiom array data only ( $n = 5,000$  to  $n = 480,000$ ): the two methods exhibit similar accuracy regardless of sample size (Extended Data Fig. 1). We obtained low SER ( $<0.2\%$ ) on the largest sample sizes for both methods, to the point that switch errors and genotyping errors cannot be distinguished (Extended Data Fig. 2).

A key feature of the WES and WGS datasets is the large number of rare variants they contain. The number of heterozygous genotypes is low at these variants and they have a small contribution in global SER measurements. We therefore stratified the SER within bins of MACs to focus on rare variants. We assigned heterozygous genotypes to different MAC bins depending on the variant frequency and computed in each MAC bin the fraction of them being correctly phased (relative to the previous heterozygous genotype, regardless of its MAC). When doing so, we found that SHAPEIT5 phases rare variants with higher accuracy than Beagle v.5.4 in both the WGS and WES datasets (Fig. 2a,b). For instance, SHAPEIT5 and Beagle v.5.4 phase rare variants in the WGS data (MAC between 11 and 20) with SER of 4.36% and 8.76%, respectively, which is a 50.2% drop. In the WES dataset, the same variant category is phased by SHAPEIT5 with a switch error rate of 2.93% compared with 5.18% with Beagle v.5.4 (42.67% reduction). Overall, SHAPEIT5 phases rare variants in the WES and WGS with 20% to 50% fewer switch errors compared with Beagle v.5.4, depending on MAC. This improvement in accuracy is also observed when only using trios for validation (Supplementary Fig. 3) and depends on sample size (Supplementary Fig. 4). Significant differences between the two methods are observed in datasets comprising at least 50,000 samples and increase with sample size.

In a large sequencing dataset, a singleton can be the product of several causes, including recent mutation, de novo mutation, somatic mutation or genotyping error. SHAPEIT5 aims to resolve the phase of recent mutations. We estimated the fraction of singletons falling in this category using duos and trios in the WGS data. We measured the fraction of singletons in offspring that is not supported by the genotype data available for the parents. In duos, we found that 47.36% of the singletons are supported by the genotyped parent, whereas

52.64% are not (Extended Data Fig. 3a), deviating from the expected 50% and suggesting that 5.26% of the singletons are not inherited from parents (assuming no inheritance bias). Consistently, in trios we found that 4.52% of the singletons in the offspring are not inherited from the parents (none of the parents carry the minor allele; Mendel inconsistency; Extended Data Fig. 3b). Together, this shows that most singletons (~95%) are inherited and can therefore be phased using both inheritance logic in trios and duos and our model. In the WGS dataset, we obtained SER of 35.1% and 36.6%, respectively (Extended Data Fig. 3c,d). In the WES dataset, we obtained an SER of 35.2% (Fig. 2b). While relatively high, this is a significant deviation from the expected 50% from previous models (binomial test  $P$  values  $<3.7 \times 10^{-15}$ ; Extended Data Fig. 3c,d).

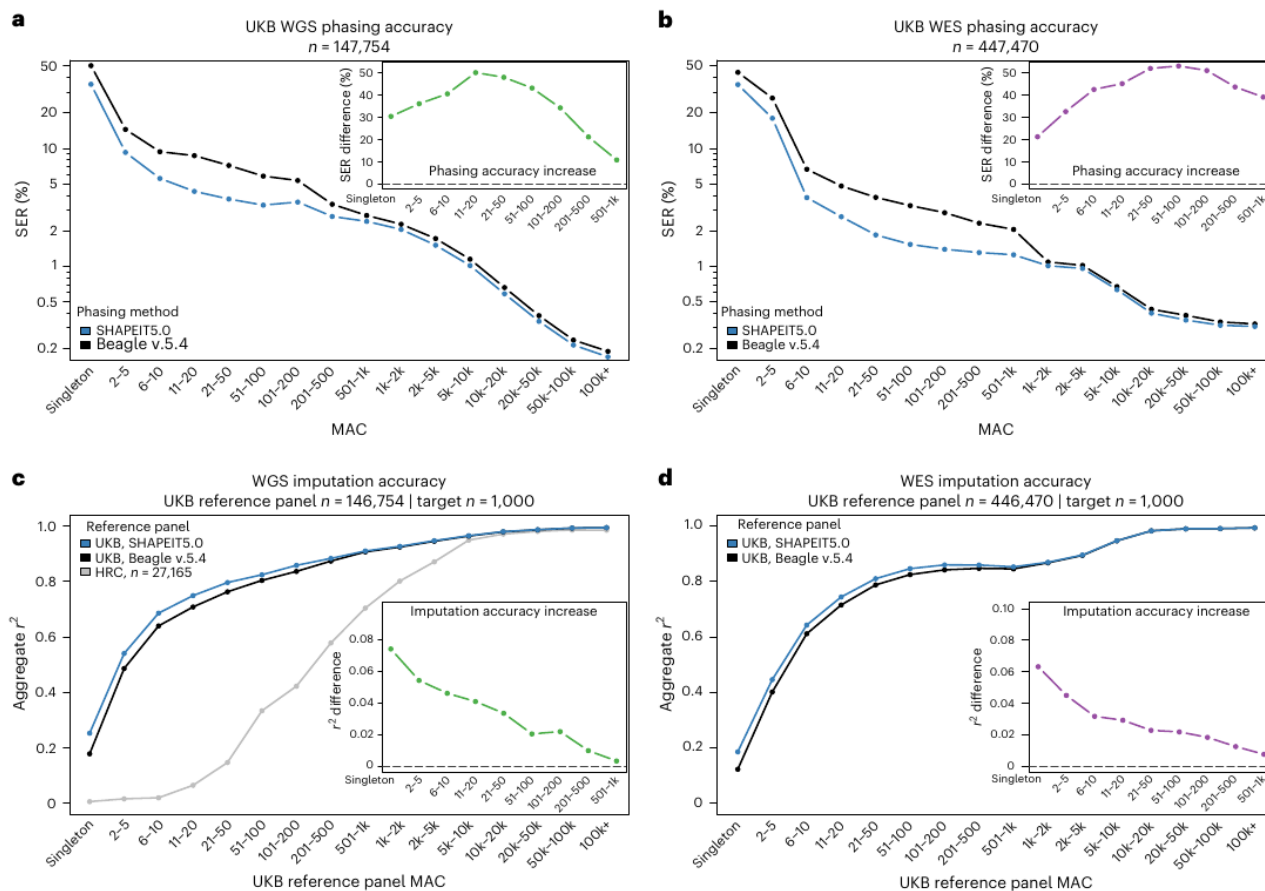
All computations were performed on the UKB RAP. The RAP offers a choice of two priority levels for computations: 'spot' (lower cost) and 'on demand' (higher cost). Assuming that all computing is performed on demand, Beagle v.5.4 and SHAPEIT5 require £57.80 and £65.20 of computing costs (as of October 2022) to phase chromosome 20 WGS data ( $n = 147,754$ ), which correspond to approximately £2,890 and £3,258 for the entire genome (Supplementary Table 2). However, these are conservative estimates, as SHAPEIT5 allows phasing of the data in chromosomal chunks (in parallel), therefore greatly reducing the need for using 'on demand' priority.

### SHAPEIT5 phasing improves genotype imputation accuracy

Several downstream analyses in disease and population genetics require haplotype-level data. One example is genotype imputation<sup>26</sup>, which uses WGS data as a reference panel to predict missing genotypes in SNP array data. As the accuracy of genotype imputation depends on the reference panel, we quantified phasing errors using genotype imputation, which has two main advantages. First, it provides a validation alternative to SER that is easy to partition by minor allele frequency. Second, it assesses the phasing quality across all samples, and not only on a small subset with parental genomes available. We imputed a subset of 1,000 UKB British samples with SNP array data available, together with WGS and WES as validation.

First, we show that genotype imputation using the UKB WGS reference panel greatly outperforms the previous generation of reference panels, such as the Haplotype Reference Consortium (HRC)<sup>27</sup> (Fig. 2c), in line with previous findings showing that large WGS panels enhance imputation<sup>2</sup>. For both UKB WGS and WES, we find that the reference panels phased with SHAPEIT5 outperform those phased with Beagle v.5.4 at rare variants (MAC  $< 500$ ; Fig. 2c,d and Extended Data Fig. 4), consistent with the SER estimates reported in Fig. 2a,b. As an example, imputation using the WGS or WES reference panel phased with SHAPEIT5 provides an increase of squared Pearson coefficient of around 0.05 for variants with a MAC between 2 and 5. In an association study, this corresponds to an increase of 5% in effective sample size when testing these variants for association, due only to better reference panel phasing<sup>28</sup>. Even singletons are better imputed using the SHAPEIT5 panel. Despite the low overall accuracy at these variants, which restricts their utility in downstream analyses, this confirms on a larger scale the validity of our singleton phasing.

SHAPEIT5 introduces a metric of phasing confidence at rare heterozygous genotypes (MAF  $< 0.1\%$ ), which corresponds to the probability of the reported phase. This allows controlling for phasing errors and utilizing phasing certainty in downstream analyses. Phasing confidence lies between 0.5 and 1, where 1 indicates no uncertainty in the phase and 0.5 means that the two phasing possibilities are equally likely. Singletons are attributed a phasing confidence of 0.5 as phasing confidence cannot be computed for them. We assessed the phasing accuracy at different confidence scores (Extended Data Fig. 5) and show that filtering variants with a threshold of 0.99 controls the SER to a maximum of around 2% for WGS data and around 1% for WES data while keeping most variants (for instance,  $>75\%$  and  $>40\%$  variants with



**Fig. 2 | Phasing performance.** **a, b.** SER (y axis, log scale) of SHAPEIT5 (blue) compared with Beagle v.5.4 (black) stratified by MAC (x axis) for the UKB WGS (**a**) and WES (**b**). The zoomed-in views show the relative reduction of SER using SHAPEIT5 compared with Beagle v.5.4 at rare variants. **c, d.** Imputation accuracy (Aggregate  $r^2$ , y axis) for 1,000 white British samples genotyped with the Axiom

array when using reference panels phased with either SHAPEIT5 (blue) or Beagle v.5.4 (black) WGS (**c**) or WES (**d**). In (**c**) the data were also imputed using the HRC reference panel (gray). The zoomed-in views show the increase of imputation accuracy at rare variants using the UKB dataset phased with SHAPEIT5 compared with Beagle v.5.4 as a reference panel.

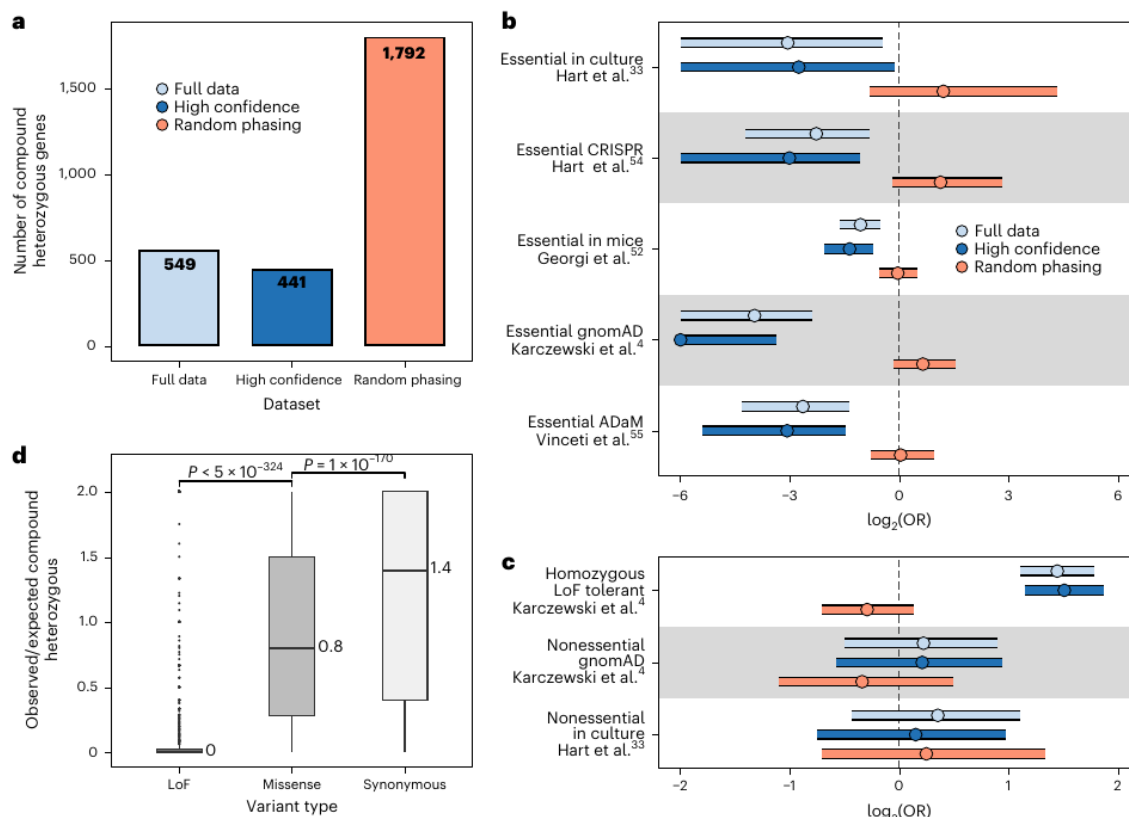
MAC 2–5 are retained). This allows researchers to confidently use rare heterozygous genotypes in their analyses.

### Identification of LoF compound heterozygotes

Compound heterozygous events occur in an individual when both copies of a gene contain at least one heterozygous variant. Compound heterozygosity is often studied in the context of LoF variants, which are expected to have highly deleterious effects on genes—equivalent to a homozygous gene knockout. Indeed, compound heterozygous events have been linked to several diseases including cancer, birth defects and Alzheimer’s disease<sup>8,29–32</sup>. The accurate haplotype phasing across the UKB performed in this study, including extremely rare variants, allows the identification of individuals and genes with compound heterozygous events. For this, we gathered 383,637 high-confidence LoF variants (stop-gain, frameshift or essential splice variants) phased across 374,826 white British individuals and 17,689 protein-coding genes (Methods). We found that a gene has, on average, 22.3 LoF variants across the cohort and an individual has, on average, 7.8 LoF variants (Extended Data Fig. 6). To determine compound heterozygous events, we identify individuals with LoF mutations in both copies of a gene. Owing to their higher error rates and the risk of introducing false positives, we opted to exclude singletons from these analyses. A total of 2,150 (12%) out of 17,689 protein-coding genes tested had at least one individual with

two or more LoF variants, and thus liable for compound heterozygous identification. From those 2,150 genes, we found 549 (26%) genes with one or more individuals with compound heterozygous LoF variants (Fig. 3a), for a total of 779 gene-individual events (766 distinct individuals; Extended Data Fig. 7 and Supplementary Data 1). When considering only high-confidence haplotype calls (phasing confidence score >0.99), we still identify 80% (441) genes and 79% (614) of the compound heterozygous events identified in the full dataset, indicating that these mostly rely on high-confidence haplotype calls (Fig. 3a and Extended Data Fig. 7). We found that the 549 compound heterozygous genes are highly depleted in several lists of known essential genes, compared with the 2,150 genes with two or more LoF variants (odds ratio (OR) 0.1–0.48 across essential gene lists,  $P < 9.7 \times 10^{-3}$ ; Fig. 3b). Conversely, compound heterozygous genes are enriched in lists of nonessential and homozygous LoF tolerant genes (OR 1.2–2.7 across nonessential gene lists; Fig. 3c). The comparison with genes with two or more LoF variants in the same individual ensures that the signal observed is not due to the mere presence or absence of LoF variants in those genes, but rather the avoidance of them occurring in both gene copies. As the UKB is composed largely of healthy individuals, a depletion of compound heterozygous events in essential genes is expected.

When comparing with phasing performed with Beagle v.5.4, we found 673 compound heterozygous genes (962 events) that are



**Fig. 3 | Compound heterozygous identification in the UKB WES data phased with SHAPEIT5.** **a**, Number of genes with at least one individual with compound heterozygous LoF variants across several categories: Full data, all LoF variants in the study, except singletons; High confidence, LoF variants excluding calls with phasing confidence score <0.99; and Random phasing, shuffling phasing of all LoF variants (once). **b**, Two-way Fisher’s exact test odds ratios  $\pm$  95% confidence interval ( $\log_2$ -scaled) of compound heterozygous genes versus noncompound heterozygous genes presence in several lists of essential genes (Methods). Background is composed of 3,018 genes with  $\geq 2$  LoF mutations; x axis is capped at  $-6$ . **c**, Same as **b** but across lists of nonessential or LoF tolerant genes. **d**, Ratio

between the number of individuals with compound heterozygous events and the expected number of individuals given the number of variants, per gene. Missense ( $n = 14,336$  genes) and synonymous ( $n = 9,816$  genes) events are shown in addition to LoF events ( $n = 2,150$  genes) as a comparison. The length of the box corresponds to the interquartile range (IQR) with the center line and values corresponding to the median, and the upper and lower whiskers represent the largest or lowest value no further than  $1.5 \times$  IQR from the third and first quartile, respectively.  $P$  values between categories correspond to two-sided Wilcoxon test  $P$  values.

significantly depleted in essential genes but at reduced levels compared with SHAPEIT5 phasing (Extended Data Fig. 8). Finally, as a control, we attributed the phase of variants randomly, which led to 1,792 compound heterozygous genes and 17,241 events (Fig. 3a), which did not display depletion in essential genes, as expected (Fig. 3b). Together, these results indicate that accurate haplotype phasing is crucial for the identification of bona fide compound heterozygous events.

The finding that compound heterozygous genes are depleted in essential genes indicates that such events are avoided, at least in a subset of the genes. To explore this further, we compared the number of expected and observed compound heterozygous events per gene, based on the variant distribution in the UKB cohort, assuming that each variant phase is independent (Methods). For LoF variants, we observed a marked decrease in observed compound heterozygous events compared with expected, confirming evidence for negative selection (Fig. 3d). Conversely, when considering variants with synonymous effect (Extended Data Fig. 9 and Supplementary Data 1), the number of observed compound heterozygous events is not depleted (median ratio = 1.4; Fig. 3d), indicating no or low selective pressure to reduce synonymous variant compound heterozygous events for most genes. When considering missense or low-confidence LoF variants (referred to as missense for simplicity), we observed a mild decrease

in observed events compared with expected (mean ratio = 0.8; Fig. 3d and Supplementary Data 1), consistent with the possible deleterious effect of some missense variants. In addition, we found that missense compound heterozygous genes had only mild or no depletion for essential genes, whereas synonymous compound heterozygous genes either had no significant depletions or were even enriched in some essential gene sets (Extended Data Fig. 9). Overall, our results demonstrate that the accurate phasing at rare variants with SHAPEIT5 allows us to screen for compound heterozygous events across the UKB cohort with high confidence, revealing that LoF compound heterozygous events are under strong selective pressure in essential genes, as expected by their high negative impact.

### Discussion

We present SHAPEIT5, a tool for phasing rare variants in large sequencing datasets. SHAPEIT5 phases common variants first to create a haplotype scaffold. Subsequently, rare variants are phased one at a time on this scaffold. A key difference from Beagle v.5.4 is the use of individualized panels of haplotypes for rare variant phasing. SHAPEIT5 ensures representation of the minor alleles at rare variants, which leads to accuracy improvements that are more pronounced in larger sample sizes. We produced phased genomes for the UKB WGS and

WES data for a compute cost below £4,000. The haplotype estimates have low SERs, with rare variants down to doubletons being phased with high confidence. This accurate phasing enables highly accurate genotype imputation when used as a reference panel. Beyond measuring error rates, we also validated phased haplotypes biologically by identifying compound heterozygous events, which we found highly depleted in essential genes, as expected. In addition, we achieved singleton phasing, albeit with higher error rates and therefore with limited downstream utility. However, we view this as an advance in phasing models as previous approaches were unable to phase singletons.

Although of substantial interest, previous knowledge of compound heterozygous cases comes mostly from case studies in families<sup>28</sup> and there is currently no method to identify these events in large biobanks systematically. Here, we show that high-quality phasing of rare variants with SHAPEIT5 allows compound heterozygosity to be studied at the biobank-scale level, which can greatly increase the number of events characterized compared with the use of family data, in addition to exploring their association with new phenotypes. As a proof-of-principle, we screened all protein-coding genes for compound heterozygous events with high-confidence LoF variants and found 549 genes predicted to be fully knocked out across 816 UKB individuals out of the 374,826 individuals considered in this study. This complements other lists of nonessential genes<sup>33</sup>, with the main difference that these knockouts are found in vivo in humans. Approximately 0.22% of the UKB cohort had at least one gene knockout by compound LoF heterozygous events. This observed frequency of events matches previous estimates in outbred healthy cohorts<sup>34</sup>. UKB participants are not expected to have any rare and/or severe genetic diseases as their average age is 56 years, which is after the age of onset for most rare diseases. This partially explains why the gene knockouts observed are strongly depleted in several lists of essential genes. However, we still found 52 genes deemed as essential in at least one of the essential gene lists we analyzed. We can conceive three possible scenarios to explain these specific cases. First, the mutations had a moderate impact on the individual and did not result in severe disease. As an example, we found one individual with pulmonary embolism while having a knockout of the essential gene *ADAM19*—a gene reported for its involvement in pulmonary disease<sup>35,36</sup>. Second, compensatory mutations can rescue the deleterious effect of the knockout. For instance, we observed one individual with a knockout of *CFTR*—an essential gene found to be rescued by several gain-of-function mutations across the genome<sup>37–39</sup>. Finally, some of the compound heterozygous events discovered may be false positives driven by incorrect phasing or erroneous LoF annotations.

We foresee that rare variant phasing in large sequencing studies such as the UKB has the potential to unlock many applications and analyses. First, other types of functional variants can be screened for compound heterozygous effects, for instance, combining LoF and missense or regulatory variants<sup>40</sup>. Second, phase information can be included in rare variant burden testing approaches, which usually consider only a mixture of the two haplotypes. Third, using accurately phased reference panels allows phasing of extremely rare variants with high accuracy, even singletons to some extent, for any new sequenced genome from the same population. This is beneficial for diagnosis of rare and severe diseases caused by compound heterozygous effects, such as in the Genomics England dataset<sup>41</sup>, in which diagnosis yield could be increased by incorporating phase information.

### Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-023-01415-w>.

### References

1. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
2. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
3. Wainschein, P. et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat. Genet.* **54**, 263–273 (2022).
4. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
5. Backman, J. D. et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
6. Wang, Q. et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527–532 (2021).
7. Miller, D. B. & Piccolo, S. R. Compound heterozygous variants in pediatric cancers: a systematic review. *Front. Genet.* **11**, 493 (2020).
8. Miller, D. B. & Piccolo, S. R. A survey of compound heterozygous variants in pediatric cancers and structural birth defects. *Front. Genet.* **12**, 640242 (2021).
9. Sulem, P. et al. Identification of a large set of rare complete human knockouts. *Nat. Genet.* **47**, 448–452 (2015).
10. Povysil, G. et al. Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat. Rev. Genet.* **20**, 747–759 (2019).
11. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).
12. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
13. Hellenthal, G. et al. A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).
14. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).
15. Halldorsson, B. V. et al. The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
16. Browning, B. L. & Browning, S. R. Statistical phasing of 150,119 sequenced genomes in the UK Biobank. *Am. J. Hum. Genet.* **110**, 161–165 (2023).
17. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).
18. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
19. Delaneau, O., Marchini, J. & The 1000 Genomes Project Consortium. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat. Commun.* **5**, 3934 (2014).
20. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
21. Rubinacci, S., Delaneau, O. & Marchini, J. Genotype imputation using the positional Burrows Wheeler transform. *PLoS Genet.* **16**, e1009049 (2020).
22. Wertenbroek, R., Rubinacci, S., Xenarios, I., Thoma, Y. & Delaneau, O. XSI—a genotype compression tool for compressive genomics in large biobanks. *Bioinformatics* **38**, 3778–3784 (2022).
23. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).
24. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).

25. Platt, A., Pivrotto, A., Knoblauch, J. & Hey, J. An estimator of first coalescent time reveals selection on young variants and large heterogeneity in rare allele ages among human populations. *PLoS Genet.* **15**, e1008340 (2019).
26. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
27. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
28. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
29. Allen, K. J. et al. Iron-overload-related disease in HFE hereditary hemochromatosis. *N. Engl. J. Med.* **358**, 221–230 (2008).
30. Hoogmartens, J. et al. Contribution of homozygous and compound heterozygous missense mutations in VWA2 to Alzheimer's disease. *Neurobiol. Aging* **99**, 100.e17–100.e23 (2021).
31. Mendonça, L. O. et al. A case report of a novel compound heterozygous mutation in a Brazilian patient with deficiency of Interleukin-1 receptor antagonist (DIRA). *Pediatr. Rheumatol. Online J.* **18**, 67 (2020).
32. Wang, R.-R. et al. Novel compound heterozygous mutations T2C and 1149insT in the KCNQ1 gene cause Jervell and Lange-Nielsen syndrome. *Int. J. Mol. Med.* **28**, 41–46 (2011).
33. Hart, T., Brown, K. R., Sircoulomb, F., Rottapel, R. & Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol. Syst. Biol.* **10**, 733 (2014).
34. Minikel, E. V. et al. Evaluating drug targets through human loss-of-function genetic variation. *Nature* **581**, 459–464 (2020).
35. London, S. J. et al. ADAM19 and HTR4 variants and pulmonary function: Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium targeted sequencing study. *Circ. Cardiovasc. Genet.* **7**, 350–358 (2014).
36. Sakornsakolpat, P. et al. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat. Genet.* **51**, 494–505 (2019).
37. Corvol, H. et al. Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nat. Commun.* **6**, 8382 (2015).
38. Trzcinska-Daneluti, A. M. et al. High-content functional screen to identify proteins that correct F508del-CFTR function. *Mol. Cell. Proteom.* **8**, 780–790 (2009).
39. Wang, X. et al. Hsp90 cochaperone Aha1 downregulation rescues misfolding of CFTR in cystic fibrosis. *Cell* **127**, 803–815 (2006).
40. Castel, S. E. et al. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat. Genet.* **50**, 1327–1334 (2018).
41. Investigators, G. P. P. et al. 100,000 genomes pilot on rare-disease diagnosis in health care—preliminary report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023



## Methods

### Ethics statement

This study relied on analyses of genetic data from the UKB cohort, which was collected with informed consent obtained from all participants. Data for this study were obtained under the UKB applications licence number 66995. All data used in this research are publicly available to registered researchers through the UKB data-access protocol.

### Common variant phasing

For common variant phasing ( $MAF \geq 0.1\%$ ), SHAPEIT5 is based largely on the previous SHAPEIT version (v.4). Briefly, it updates the phase of each sample in turn within a Gibbs sampler iteration scheme: each sample is phased by conditioning on other samples' haplotypes using the Li and Stephens model<sup>24</sup>. Two main features, already part of the SHAPEIT4 model, allow fast phasing at common variants: (1) first, the haplotype sampling step has linear complexity in the number of conditioning states<sup>32</sup> and is multithreaded so that several samples are phased in parallel; (2) second, the sampling is based on a parsimonious and highly informative set of haplotypes, identified in constant time using the PBWT data structure.

However, one computational limitation of SHAPEIT4 resides in its inability to parallelize the construction of the PBWT, which can become relatively long in very large datasets. In SHAPEIT5, the main improvement we introduced for common variant phasing is a parallelization scheme for the PBWT construction: several PBWT passes are run in parallel on several central processing unit cores, each one running for a different chunk of 4 cM by default, achieving a notable reduction of the wallclock running time of the method.

### Rare variant phasing

To accurately phase rare variants ( $MAF < 0.1\%$ ), SHAPEIT5 uses the haplotypes derived at common variants as haplotype scaffolds onto which heterozygous genotypes are phased one rare variant at a time. For a single heterozygous genotype, we aim to determine which of the two target chromosomes carries the minor allele (as opposed to the major allele). To do so, our method uses the Li and Stephens model to compute the probabilities of the two possible phases. The probabilistic inference is based on a set of haplotypes carried by other samples in the dataset, that we call conditioning haplotypes. Similarly, we call a conditioning sample, any sample carrying at least one conditioning haplotype and conditioning set, the collection of conditioning haplotypes used for inference. The outcome of the estimation is a posterior probability of the most likely phase for each of the rare heterozygotes. Specifically, our model comprises five main features:

**Sparse representation.** We use a sparse matrix representation of the genotypes at rare variants to efficiently store large amounts of genotype data in memory and speed up computations. Only genotypes carrying at least one copy of the rare allele are stored in memory together with the necessary indexes to determine the sample and variant to which the genotype corresponds. As most of the rare variants are homozygous for the major allele, this representation allows for a large reduction in memory usage and a fast identification of heterozygous genotypes at a given rare variant. To quickly retrieve rare genotypes at both the sample and variant levels, we store this sparse genotype matrix in memory together with its transpose.

**Haplotype selection.** To get the most informative haplotypes in the conditioning set, we require that they (1) share long haplotype matches with the target and (2) are not monomorphic at the rare variant of interest. The first condition ensures that the haplotypes in the conditioning set are informative for the copying model. The second condition ensures that the conditioning set contains carriers of the two possible alleles at the rare variant of interest. The latter is required to accurately contrast the two possible phasing possibilities of the rare heterozygous

variant. To efficiently retrieve haplotypes complying with these properties, we use the PBWT data structure of the haplotype data derived at common variants. We perform both forward and backward PBWT sweeps so that we can identify long matches between haplotypes centered in the position of the rare variant by interrogating the flanking prefix arrays. This gives a first set of haplotypes that complies with condition (1), but not necessarily with condition (2). Therefore, we do a second identification of matches in the PBWT, this time restricting the search to the subset of samples carrying the minor allele. We achieve this second pass efficiently by taking advantage of the sparse genotype representation: we interrogate only the PBWT prefix arrays at the sparse indexes.

**Forcing homozygosity.** The conditioning set defined before contains a set of haplotypes that share large segments with the target haplotype at common variants, but they have not been phased yet at the rare variant of interest. When the conditioning sample (that is, the sample carrying the haplotype) is homozygous, this is not an issue as its two haplotypes carry the same allele. However, when the conditioning sample is heterozygous, we do not know the allele carried by each one of its two haplotypes. We solve this by simply assigning the minor allele to both haplotypes<sup>17</sup>. As a consequence of the two previous steps, the conditioning set of haplotypes is guaranteed to contain carriers of the two possible alleles at the rare variant of interest.

**Copying model.** We can now perform phasing of rare heterozygous genotypes based on the conditioning set of haplotypes that have been constructed as part of all the previous steps. SHAPEIT5 computes the probability that each target haplotype carries the minor allele by using a haploid version of the Li and Stephens model<sup>24</sup> as implemented in Impute5 (ref. 21) (for a definition of the HMM parameters and a formal description of the imputation model used, see Rubinacci et al.<sup>21</sup> and Howie et al.<sup>43</sup>). Specifically, it runs a forward-backward pass as done in the context of genotype imputation (see Marchini<sup>44</sup> for details) to get the probabilities that each target haplotype carries the minor allele at the rare variant. In practice, the vector of copying probabilities is obtained at each rare variant by averaging the copying probabilities computed at the two closest flanking common variants. Here, the conditioning set of haplotypes serves as a local reference panel for imputing the alleles at the rare variant in the target sample. Of note, accurate inference is made possible since the conditioning set we chose is guaranteed to comprise carriers of both the major and minor alleles at the rare variant of interest. Having only carriers of a single allele would not be informative for making inference here. Finally, we use these imputation probabilities to derive phasing probabilities (Supplementary Fig. 1), which we can use to get the most probable phase or as phasing confidence scores to propagate phasing uncertainty in downstream analyses.

**Singleton phasing.** In the case of singletons, only the target sample carries a copy of the minor allele at the rare variant. Therefore, none of the conditioning haplotypes carries the minor allele and the whole copying model described above is unable to make inference. This is a well-known limitation of all statistical phasing methods. SHAPEIT5 can provide inference at these sites by using the Viterbi algorithm for the Li and Stephens model<sup>24</sup>, to obtain the longest shared IBD segment between each one of the two target haplotypes and the conditioning haplotypes. The minor allele of singletons is then assigned to the target haplotype with the shortest shared segment. The idea behind this model presumes that the shorter the IBD sharing between two haplotypes, the older their most recent common ancestor is, and therefore, the chance for new mutations to occur in that lineage is increased.

### Validation of haplotype estimates

To validate haplotype estimates, we use trios (two parents, one offspring) for WES data and both duos (parent-offspring pairs) and trios

for WGS data. To identify parent–offspring relationships, we use the kinship estimate and the IBSO as provided as part of the UKB SNP array release. We select parent–offspring relationships as having a kinship coefficient lower than 0.3553 and greater than 0.1767 and an IBSO lower than 0.0012 (refs. 1,45). In addition, we require that the difference in age between parents and offspring is greater than 15 years and that the two parents have different sex for trios. We finally keep only self-declared white British individuals for which ancestry was confirmed by principal component analysis (PCA, UKB field 22006). The number of trios used in the validation for all three datasets (Array, WES or WGS) is shown in Supplementary Table 1. Validation of haplotypes is a two-step procedure. First, we statistically phase a given dataset including only the offspring samples. Second, we use the parents to measure the SER—a metric commonly used to assess how close estimated and true haplotypes are. The SER is defined as the fraction of successive pairs of heterozygous genotypes being correctly phased. In the context of this work, we measured SER stratified by bins of MAC. We assigned each heterozygous genotype to a given MAC bin and counted the fraction of heterozygous genotypes being correctly phased per MAC bin, relative to the previous heterozygous genotypes (this one can belong to any MAC bin). This definition of SER has the advantage of showing how well statistical phasing performs depending on the frequency of the variants it phases (either common or rare).

#### UKB SNP array dataset

We used the UKB Axiom array in PLINK format and converted it into VCF format using plink2 (v.2.00a3.1LM). This resulted in 784,256 variant sites across autosomes for 488,377 individuals. We then applied quality control on the data using the UKB SNPs and samples QC file (UKB Resource 531) to only retain SNPs and individuals that have been used for the official phasing of the Axiom array data<sup>1</sup>, resulting in 670,741 variant sites across 486,442 individuals. This includes 897 white British parent–offspring trios and 4,373 white British parent–offspring duos (Supplementary Table 1).

#### UKB WGS dataset

We use the whole-genome GraphTyper joint call pVCFs from the UKB RAP. We first decomposed multiallelic variants into biallelic variants using bcftools (v.1.15.1) norm -m<sup>46</sup>. We then performed quality control of the variant sites and filtered out SNPs and indels for (1) Hardy–Weinberg  $P$  value  $< 10^{-30}$ , (2) more than 10% of the individuals having no data (GQ score = 0; missing data), (3) heterozygous excess less than 0.5 or greater than 1.5 and (4) alternative alleles with AAscore  $< 0.5$ . Additionally, we kept only variant sites with the tag 'FILTER = PASS', as suggested by the data providers<sup>15</sup>. This resulted in a total of 603,925,301 variant sites, including 20,662,402 common variant sites (MAF  $\geq 0.1\%$ ) and 583,262,899 rare variant sites (MAF  $< 0.1\%$ ), across a total of 150,119 individuals. This WGS dataset includes 31 trios and 432 duos (Supplementary Table 1). To assess the accuracy of the phasing, we use chromosome 20 only. For this analysis, we used only samples being also genotyped with the UKB Axiom array, resulting in 147,754 individuals (Supplementary Table 1). We phased chromosome 20 using chunks of, on average, 4.5 Mb with overlapping buffers of 250 kb. We used Beagle v.5.4 (refs.16,17) with default parameters on the entire chromosome 20.

#### UKB WES dataset

We used the WES files in pVCF format as released on UKB RAP. The quality control pipeline has been described in Szustakowski et al.<sup>47</sup>. To phase WES data, we first merged it with the unphased SNP array data. The aim of this was to increase the number of common variants that are phased in the first step of SHAPEIT5 (that is, common variants phasing), which improves the quality of the haplotype scaffold onto which rare variants are phased, in particular at intergenic regions. We kept only individuals with both the SNP array and the WES data, resulting in 452,644 total individuals, including 719 white British parent–offspring trios and

3,014 white British parent–offspring duos. When a variant is listed in both the WES and the SNP array, we keep the SNP array copy as the SNP array is expected to be more robust to SNP calling errors<sup>48</sup>. This resulted in retaining a total of 26,199,614 variants, including 977,517 common variants (MAF  $\geq 0.1\%$ ) and 25,222,097 rare variants (MAF  $< 0.1\%$ ) (Supplementary Table 1). Phasing the 452,644 individuals with both WES and Axiom array available data is performed for each chromosome independently in a single chunk. We also used Beagle v.5.4 (refs. 16,17) with default parameters.

#### Genotype imputation

To perform genotype imputation from the phased WGS and WES datasets, we extracted 1,000 samples with British ancestry that are unrelated to any other sample in the dataset, and for which we had Axiom SNP array data available. We therefore used a reference panel composed of the remaining 146,754 WGS samples and 446,470 WES samples for both SHAPEIT5 and Beagle v.5.4. For the HRC reference panel, we used the PICARD toolkit (<http://broadinstitute.github.io/picard/>) to lift over the data to the Human genome assembly GRCh38, retaining 99.8% of the original variants.

We used Beagle v.5.4 for genotype imputation of SNP array data, allowing prephasing from the reference panel. We accessed imputation accuracy by measuring the squared Pearson correlation between imputed and high-coverage genotypes using the GLIMPSE\_concordance tool<sup>49</sup> (-gt-val option) at custom allele count bins (-ac-bins 15 10 20 50 100 200 500 1000 2000 5000 10000 20000 50000 100000 146754 for WGS, --ac-bins 15 10 20 50 100 200 500 1000 2000 5000 10000 20000 50000 100000 446470 for WES). A drop of correlation quantifies the reduction in effective sample size in association testing due to imperfect imputation. For instance, a difference of 0.05 involves a power loss equivalent to losing 5% of the data.

We also evaluated the nonreference discordance rate using the GLIMPSE\_concordance<sup>49</sup> tool. The nonreference discordance<sup>46</sup> is calculated as  $NRD = (e_{rr} + e_{ra} + e_{aa}) / (e_{rr} + e_{ra} + e_{aa} + m_{ra} + m_{aa})$ , where  $e_{rr}$ ,  $e_{ra}$  and  $e_{aa}$  are the counts of the mismatches for the homozygous reference, heterozygous and homozygous alternative genotypes, respectively, and  $m_{ra}$  and  $m_{aa}$  are the counts of the matches at the heterozygous and homozygous alternative genotypes. NRD is an error rate that excludes the homozygous reference matches, which are the most frequent at rare variants, giving more weight to the other matches. We computed the nonreference discordance rate within frequency bins in the reference panel.

#### Compound heterozygosity detection

We restricted the analysis to the cohort of self-declared white British individuals for which the ancestry is confirmed by PCA (UKB field 22006) with both SNP array and exome-seq data, excluding parental individuals ( $n = 374,826$ ). Only WES variants with MAF  $< 0.1\%$  before sample filtering were considered. Variant annotations (LoF, Synonymous and Missense|LC) were obtained from the Genebase database<sup>50</sup> through Hail (gene-level results, results.mt). Briefly, these variants had been annotated by Ensembl VEP v.95 (ref. 51) and LoF variants (stop-gain, frameshift and splice donor/acceptor sites) were further processed by LOFTEE<sup>4</sup>, separating high-confidence (used as 'LoF') from low-confidence (used in the 'Missense|LC' category). Only unique canonical transcripts for protein-coding genes were considered. LoF, synonymous and missense variants were gathered in the UKB cohort using bcftools (v.1.15.1) isec function, with the '-c none' parameter to match variants by chromosome, position, reference and alternative alleles. Singleton variants were excluded from this analysis.

Identification of compound heterozygous events was performed with custom Python (v.3.7) scripts. Briefly, for each variant type (LoF, synonymous, missense) and for each gene, individuals with at least two mutations were assessed for compound heterozygosity by having at least one variant in each of the two haplotypes. In addition, for each

gene, we calculated the expected number of individuals with compound heterozygosity as  $\sum_{i=1}^n 1 - \frac{1}{2^{(\nu-1)}}$ , where  $\nu$  indicates the number of variants in individual  $i$  in the gene. To compare the number of LoF compound heterozygous genes and events without phasing, we randomized phasing at all variants by attributing 0.5 probability for each variant to fall in either of the two haplotypes, independently for each variant.

### Essential and nonessential gene lists

We obtained lists of essential and nonessential genes from several sources (described below). For each of these gene lists, we performed Fisher's exact tests (two-sided) for several categories of compound heterozygous genes versus noncompound heterozygous genes, considering a background of 2,150 genes with at least one individual with two LoF mutations. For synonymous and missense variants, the background included 10,119 and 14,914 genes, respectively. The following lists of genes were obtained: (1) essential in mice ( $n = 2,454$ ) from Georgi et al.<sup>52</sup> includes genes where homozygous knockout in mice results in pre-, peri- or postnatal lethality and was extracted with ortholog human gene symbols from McArthur's laboratory<sup>53</sup>; (2) essential in culture ( $n = 360$ ) core essential genes from genomic perturbation screens were obtained from Hart et al.<sup>33</sup>; (3) nonessential in culture ( $n = 927$ ) putatively nonessential genes (shRNA screening) were obtained from Hart et al.<sup>33</sup>; (4) essential CRISPR ( $n = 684$ ) genes essential in culture from CRISPR screening were obtained from Hart et al.<sup>54</sup>; (5) essential ADaM ( $n = 1,075$ ) genes annotated by the ADaM analysis of a large collection of gene dependency profiles (CRISPR-Cas9 screens) across 855 human cancer cell lines (Project Score and Project Achilles 20Q2) were obtained from Vinceti et al.<sup>55</sup>; (6) essential gnomAD ( $n = 1,920$ ) genes at the bottom LOEUF decile from gnomAD v.2.1.1 (that is, most constrained genes) were obtained from <https://gnomad.broadinstitute.org/> (ref. 4); (7) nonessential gnomAD ( $n = 1,919$ ) genes at the top LOEUF decile from gnomAD v.2.1.1 (that is, least constrained genes) were obtained from <https://gnomad.broadinstitute.org/> (ref. 4); and (8) homozygous LoF tolerant ( $n = 1,815$ ) genes with homozygous LoF variants observed in the gnomAD cohort were obtained from Karczewski et al.<sup>4</sup> (Supplementary Data 7).

### Statistics and reproducibility

This study was based on the UKB SNP array, WES and WGS datasets. Variants and samples were selected based on quality controls and ancestry as described in the SNP array, WES and WGS data processing methods. In certain analyses, only individuals including both WGS/WES and SNP array data were included. Statistical analyses, including Fisher's exact tests, binomial and Wilcoxon tests were performed with R v.4.2. All code to reproduce analyses is publicly available.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

The lists of compound heterozygous events and genes are available in Supplementary Data 1. The phased WGS reference panel can be accessed via the UKB RAP: <https://ukbiobank.dnanexus.com/landing>. RAP is open to researchers who are listed as collaborators on UKB-approved access applications. Lifter was performed using a chain file provided by UCSC (<https://hgdownload.cse.ucsc.edu/goldenpath/hg19/liftOver/>). The publicly available subset of the Haplotype Reference Consortium dataset is available from the European Genome-Phenome Archive at the European Bioinformatics Institute, accession EGAS00001001710. Source data are provided with this paper.

### Code availability

SHAPEIT5 is available under MIT license at <https://github.com/odelaneau/shapeit5>. This includes code to the phase\_common, phase\_rare,

ligate and switch tools and the scripts used to phase WES and WGS data on the UKB RAP. The documentation is available at <https://odelaneau.github.io/shapeit5>. Code and source data to reproduce analysis and plots have been deposited in the linked Zenodo repository: <https://doi.org/10.5281/zenodo.7828479> (ref. 56).

### References

- Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
- Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
- Marchini, J. in *Handbook of Statistical Genomics* Vol. 4 (ed. Balding, D. J.) 87–114 (Wiley, 2019).
- Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
- Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
- Szostakowski, J. D. et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat. Genet.* **53**, 942–948 (2021).
- Yi, M. et al. Performance comparison of SNP detection tools with illumina exome sequencing data—an assessment using both family pedigree information and sample-matched SNP array data. *Nucleic Acids Res.* **42**, e101 (2014).
- Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 120–126 (2021).
- 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**, 100168 (2022).
- McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
- Georgi, B., Voight, B. F. & Bućan, M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet.* **9**, e1003484 (2013).
- Minikel, E. et al. macarthur-lab/gene\_lists: stable release. *Zenodo* <https://doi.org/10.5281/zenodo.6724346> (2022).
- Hart, T. et al. Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. *G3 (Bethesda)* **7**, 2719–2727 (2017).
- Vinceti, A. et al. CoRe: a robustly benchmarked R package for identifying core-fitness genes in genome-wide pooled CRISPR-Cas9 screens. *BMC Genomics* **22**, 828 (2021).
- 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. *Zenodo* <https://doi.org/10.5281/zenodo.7828479> (2023).

### Acknowledgements

This work was funded by a Swiss National Science Foundation (SNSF) project grant (PP00P3\_176977). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

### Author contributions

O.D. and S.R. developed the method. R.J.H. performed the phasing experiments. S.R. performed the imputation experiments. D.M.R. performed compound heterozygous analyses. All authors wrote and reviewed the manuscript. O.D. designed and supervised the study. These authors contributed equally and are listed alphabetically: R.J.H., D.M.R. and S.R.

### Competing interests

The authors declare no competing interests.

### Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41588-023-01415-w>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41588-023-01415-w>.

**Correspondence and requests for materials** should be addressed to Olivier Delaneau.

**Peer review information** *Nature Genetics* thanks Arnaldur Gylfason, Tobias Marschall and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).



## Appendix B

# Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes

This manuscript is presented in [Chapter I](#).

The online version and the supplementary material can be downloaded from <https://www.nature.com/articles/s41588-023-01438-3> .



# Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes

Received: 28 November 2022

Accepted: 31 May 2023

Published online: 29 June 2023

 Check for updatesSimone Rubinacci<sup>1,2</sup>, Robin J. Hofmeister<sup>1,2</sup>, Bárbara Sousa da Mota<sup>1,2</sup>  
& Olivier Delaneau<sup>1,2</sup>✉

The release of 150,119 UK Biobank sequences represents an unprecedented opportunity as a reference panel to impute low-coverage whole-genome sequencing data with high accuracy but current methods cannot cope with the size of the data. Here we introduce GLIMPSE2, a low-coverage whole-genome sequencing imputation method that scales sublinearly in both the number of samples and markers, achieving efficient whole-genome imputation from the UK Biobank reference panel while retaining high accuracy for ancient and modern genomes, particularly at rare variants and for very low-coverage samples.

Recent work and method advances<sup>1–4</sup> highlight the advantages of low-coverage whole-genome sequencing (lcWGS), followed by genotype imputation from a large reference panel, as a cost-effective genotyping technology for statistical and population genetics. Large-scale whole-genome sequencing projects, such as the recent release of 150,119 samples from the UK Biobank<sup>5</sup> (UKB), offer new opportunities to improve lcWGS imputation, potentially improving accuracy at rare variants (minor allele frequency (MAF) < 0.1%). However, current methods struggle to scale to the size of this new generation of reference panels resulting in prohibitive computational costs. To address this issue, we propose GLIMPSE v.2 (GLIMPSE2), a major improvement of GLIMPSE<sup>1</sup>, that scales to a reference panel containing millions of reference haplotypes, with high imputation accuracy at rare variants (MAF < 0.1%) and for very low-coverage samples (0.1× to 0.5×).

To demonstrate the benefits of using sequenced biobanks for lcWGS imputation, we phased the recent release of the UKB WGS data<sup>5,6</sup> using SHAPEIT5 (ref. 7) and created a UKB reference panel of 280,238 haplotypes and 582,534,516 markers (Supplementary Note 1). We used the UKB panel to impute lcWGS samples with GLIMPSE2 and other recently released imputation methods: GLIMPSE1 (ref. 1) and QUILT v1.0.4 (ref. 2). Compared to other reference panels, the UKB leads to considerable accuracy improvements for British samples across all tested depths of coverage. Furthermore, GLIMPSE2 outperforms GLIMPSE1, particularly at rare variants (MAF < 0.1%) and for very low-coverage (for 0.1× and 1.0× data at 0.01% MAF, GLIMPSE1 and GLIMPSE2 obtain an  $r^2$  of 0.561 and 0.892 compared to 0.725 and 0.927, respectively) and matches QUILT v1.0.4 accuracy, designed to condition on the full set of reference haplotypes

(for 0.1× and 1.0× data at 0.01% MAF, QUILT v1.0.4 obtained an  $r^2$  of 0.728 and 0.925, respectively; Fig. 1a, Supplementary Note 2, Supplementary Figs. 1–3 and Supplementary Tables 2–4). We also find that the accuracy of GLIMPSE2 and QUILT v1.0.4 methods is similar when imputing 42 non-European samples from 1,000 Genomes Project using the UKB reference panel (Supplementary Note 2, Supplementary Fig. 4 and Supplementary Table 5).

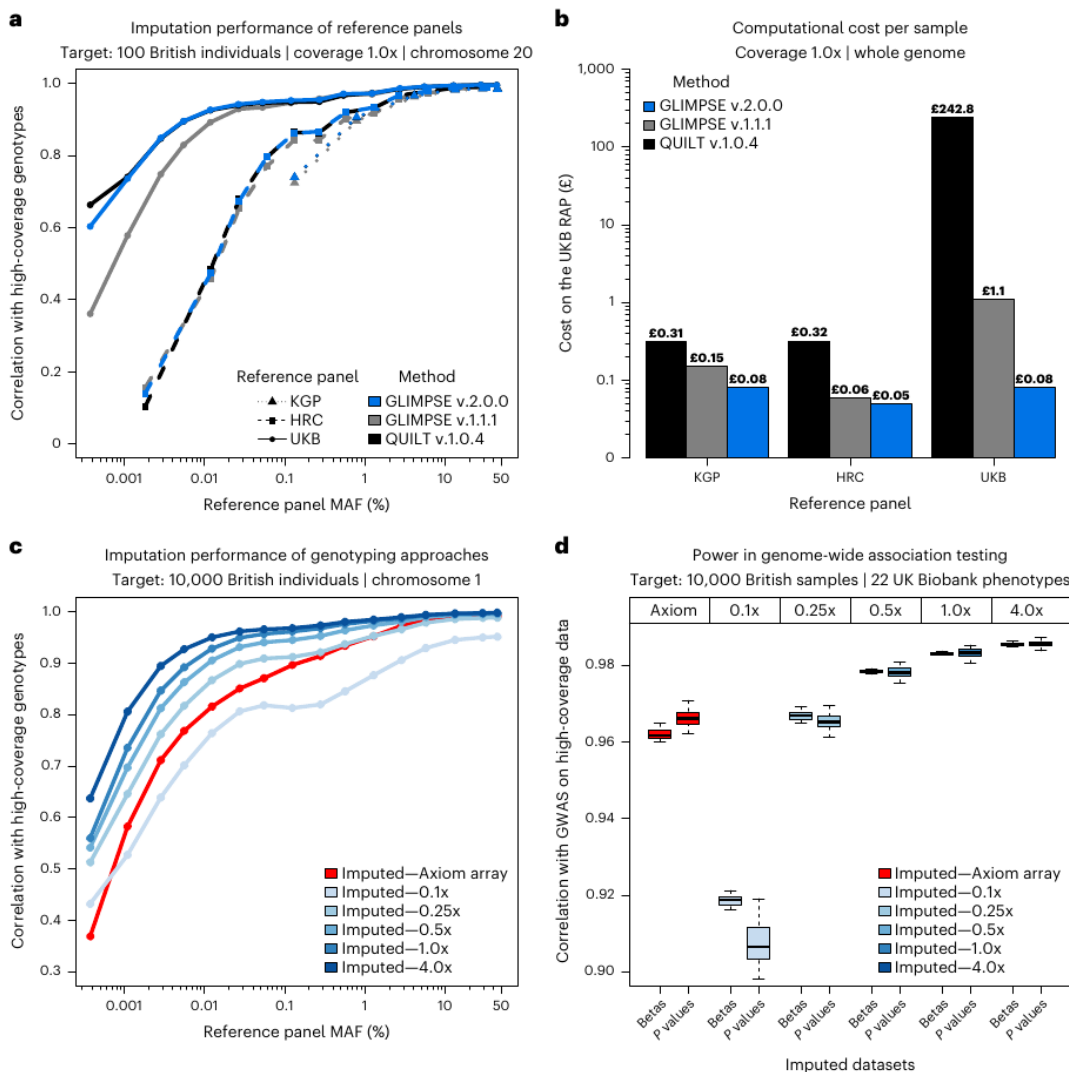
We further investigate the effect of the reference panel by imputing individuals of 129 human populations from the Simons Genome Diversity Project and we show that the UKB panel drastically improves imputation accuracy of European samples compared to the 1,000 Genomes Project reference panel, in particular of Northern Europe origin, for which the UKB reference panel obtains a reduction of non-reference discordance rate >67% (Supplementary Note 3, Extended Data Fig. 2 and Supplementary Fig. 8). Additionally, we imputed three ancient Europeans and a Yamnaya sample for which high-coverage data (>18×) are available and find similar improvements (Supplementary Note 4 and Supplementary Fig. 9), showing that some ancient populations, such as Viking, Western Hunter-Gatherer and Yamnaya could be well imputed from the UKB reference panel.

The imputation of a single lcWGS genome using the UKB reference panel is expensive or prohibitive using existing methods. On the UKB research analysis platform (RAP), the cost is £1.11 and £242.80 for GLIMPSE1 and QUILT v1.0.4, respectively. In contrast, the same task performed with GLIMPSE2 only costs £0.08, due to major algorithmic improvements that drastically reduce the imputation time for rare variants (Fig. 1b, Supplementary Note 2 and Supplementary Figs. 5 and 6). We confirm this trend for up to 2 million reference haplotypes,

<sup>1</sup>Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. <sup>2</sup>Swiss Institute of Bioinformatics, Lausanne, Switzerland.

✉ e-mail: [olivier.delaneau@unil.ch](mailto:olivier.delaneau@unil.ch)





**Fig. 1 | Accuracy, running time and power of low-coverage imputation using the UKB WGS data. a, b,** Imputation performance of different imputation methods: QUILT v.1.0.4 (black), GLIMPSE1 (gray) and GLIMPSE2 (blue); across the 1,000 Genomes Project (KGP), HRC and UKB reference panels, for 100 UKB British samples at 1.0× coverage. **a,** Accuracy on chromosome 20 (Pearson  $r^2$ , y axis), of imputation methods and reference panels: KGP (dotted line), HRC (dashed line) and UKB (full line). Accuracy is plotted against MAF of the appropriate reference panel (x axis, log scale). **b,** Cost per sample on the RAP for whole-genome imputation (y axis, log scale) across different reference panels (x axis). **c, d,** Performance of imputed data using the UKB reference panel across

coverages (0.1–4.0×, different shades of blue, GLIMPSE2 imputation) and Axiom array data (red). **c,** Accuracy on chromosome 1 of 10,000 UKB British samples (Pearson  $r^2$ , y axis) against MAF of the appropriate reference panel (x axis, log scale). **d,** Power in association testing of 10,000 UKB British samples compared to high-coverage data. Correlation of betas and  $P$  values (Pearson  $r^2$ , y axis) of different imputed datasets (x axis) across 22 UKB phenotypes. Lower and upper limits of the box plots represent the first and third quartiles (Q1 and Q3); the median is marked at the center of the box. Lower and upper whiskers are defined as  $Q1 - 1.5(Q3 - Q1)$  and  $Q3 + 1.5(Q3 - Q1)$ , respectively.

using simulated data (Supplementary Note 2 and Supplementary Fig. 7). These improvements in imputation running time and memory requirements are crucial to keep lcWGS close to single nucleotide polymorphism (SNP) arrays in terms of computational costs<sup>8,9</sup> (Supplementary Note 5) while maintaining the major advantage of providing better genotype calls. Indeed, we find that imputation of 0.5× data yields similar or more accurate results compared to the UKB Axiom array, with a notable difference at rare variants (for 0.5× coverage, accuracy improvement of  $r^2 > 0.1$  for variants with a MAF < 0.01%, Fig. 1c). Using simulated SNP arrays, we further confirm that 0.5× yields at least the same imputation accuracy as the densest SNP array model tested (Omni 2.5 array; Extended Data Fig. 3).

To assess the impact of these improvements on genome-wide association studies (GWAS), we imputed 10,000 UKB samples that we used to test 22 quantitative traits for association, comparing the respective abilities of lcWGS and SNP array data to recover the signals found with high-coverage sequencing data (Supplementary Note 6). We find that 0.5× leads to  $P$  values and effect size estimates as accurate as those obtained from Axiom array data (Fig. 1d and Supplementary Figs. 10–12) while delimiting regions of association with matching sensitivity and specificity (Supplementary Note 6 and Extended Data Fig. 4). We also look at rare loss-of-function, missense and synonymous variants<sup>10</sup> and show that 1.0× outperforms the Axiom array for all categories of variants, an improvement that will be reflected in downstream

burden-test analysis (Supplementary Note 7 and Extended Data Fig. 5). Altogether, this shows that lcWGS constitutes a powerful alternative to SNP array for downstream GWAS and rare-variant analysis.

In this work, we introduce several improvements to the GLIMPSE method that solve the computational problem of imputing lcWGS data from the 150,119 WGS samples in the UKB. We demonstrate that this reference panel leads to striking accuracy improvements across several sample ancestries, allele frequencies and depths of coverages. Our study further confirms the advantage of lcWGS over SNP arrays for GWAS, by showing that using imputed data with coverage as low as 0.5× are enough to outperform SNP array data, particularly at rare variants. Our work can be applied to other sequenced and diverse biobanks, such as Trans-Omics for Precision Medicine<sup>11</sup>, gnomAD<sup>12</sup> or AllofUs<sup>13</sup>, thereby facilitating lcWGS imputation of non-European individuals. We believe that the difference between low-coverage and high-coverage WGS will become increasingly smaller as large reference panels will keep collecting more human haplotype diversity.

### Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-023-01438-3>.

### References

1. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 120–126 (2021).
2. Davies, R. W. et al. Rapid genotype imputation from sequence with reference panels. *Nat. Genet.* **53**, 1104–1111 (2021).
3. Martin, A. R. et al. Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *Am. J. Hum. Genet.* **108**, 656–668 (2021).
4. Li, J. H., Mazur, C. A., Berisa, T. & Pickrell, J. K. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Genome Res.* **31**, 529–537 (2021).
5. Halldorsson, B. V. et al. The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
6. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
7. 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. *Nat. Genet.* <https://doi.org/10.1038/s41588-023-01415-w> (2023).
8. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
9. Rubinacci, S., Delaneau, O. & Marchini, J. Genotype imputation using the positional Burrows Wheeler transform. *PLoS Genet.* **16**, e1009049 (2020).
10. 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**, 100168 (2022).
11. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
12. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
13. The All of Us Research Program Investigators. The ‘All of Us’ research program. *N. Engl. J. Med.* **381**, 668–676 (2019).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

## Methods

This study relies on analyses of genetic data from the UKB cohort, which was collected with informed consent obtained from all participants. Data for this study were obtained under the UKB applications licence number 66995 and are available to registered researchers through the UKB data-access protocol. Additional data used in this study are all publicly available.

### GLIMPSE2

To perform imputation of low-coverage WGS data, GLIMPSE2 uses a Gibbs sampler algorithm that alternates between haploid imputation and phasing, using a modified version of the Li and Stephens hidden Markov model (HMM)<sup>14</sup>. The method necessitates a genotype likelihoods matrix for the target samples and a reference panel of haplotypes as input. The initialization step begins with the selection of a set of haplotypes from the reference panel via rare-variant calls derived from the low-coverage genotype likelihoods. Following that, two consecutive steps of haploid imputation are executed, one for each of the two target haplotypes. At the end of the initialization step, a diploidy is assigned to each target sample. GLIMPSE2 subsequently runs a series of burn-in and main Gibbs iterations to refine the genotype calls and phasing of each target sample. The algorithm determines haploid likelihoods for one of the two target haplotypes, based on the original genotype likelihoods and conditional on the current estimate of the other haplotype. To integrate over phasing uncertainty, the approach averages imputation posteriors across all main iterations.

Conversely from the GLIMPSE1 method, GLIMPSE2 approach is primarily focused on imputation only from the reference panel and it optimizes this task by incorporating new features. First, the reference panel is represented sparsely in memory, allowing for efficient storage of dense cohorts. The sparse representation of the reference panel facilitates the introduction of a new data structure to hasten haplotype matching and an efficient implementation of the HMM, which calculates posterior probabilities by leveraging the sparsity of the panel. Additional features of GLIMPSE2 include a genotype caller that integrates genotype likelihood computations directly into the GLIMPSE software and imputation of small insertions and deletions and low-quality variants separately from SNPs, by performing imputation into a haplotype scaffold obtained from high-quality SNPs.

The subsequent sections will provide a more comprehensive explanation of three of the previously referenced features, which are critical for the ability of the model to scale when applied to deeply sequenced reference panels. Further details regarding the method can be found in Supplementary Note 1.2.2.

### Sparse reference panel representation

GLIMPSE2 represents the reference panel as a sparse matrix, encoding haplotypes with one bit per allele if the variant is defined as common ( $MAF \geq 0.001$  by default) and storing the indices of the haplotypes that carry the minor allele, otherwise. This data representation allows for small memory usage but also for a fast identification of the haplotypes carrying a rare variant. Additionally, the transpose of the data structures gives efficient access to the rare variants of each haplotype. More details can be found in Supplementary Note 1.2.2.1.

We encoded the sparse reference panel representation in a binary file format to be efficiently stored on the disk. The file format translates directly into the memory data structures used by GLIMPSE2 and does not require any general-purpose compression algorithm. Together with the reference file format, we store the run-length encoded sparse positional Burrows–Wheeler transform (PBWT) data structure in the same file file, together with the recombination map. As a result, all the data related to the reference panel can be quickly loaded in memory, in much faster running times than standard file formats, such as VCF and BCF.

### Sparse positional Burrows–Wheeler transform matching

One of the key components of the GLIMPSE1 model is to reduce the state space using PBWT<sup>15</sup>, a data structure that allows efficient query searches in haplotype cohorts, linear in the number of samples and markers. Similarly, GLIMPSE2 extends the PBWT and proposes an algorithm designed for large sequencing cohorts, here called sparse PBWT.

By using the sparse representation of the reference panel, rare variants are treated differently than common variants, allowing the computation of smaller PBWTs which speeds up the algorithm. This is based on the idea that between two adjacent common variants most of the haplotypes do not contain the minor allele in the region and therefore most of the haplotypes would form a single invariable block of major alleles that preserves their relative haplotype order. Therefore, a smaller PBWT is constructed only on haplotypes that have at least one minor allele between two adjacent common variants. The positional prefix array of the small PBWT at the end of the rare-variant interval is simply concatenated with the positional prefix array of other haplotypes that are not changing in the interval. A schematic illustration of the sparse PBWT is shown in Extended Data Fig. 1 and more details are provided in Supplementary Note 1.2.2.2.

Haplotype selection is performed by querying target samples in the sparse PBWT, looking at neighboring haplotypes at common variants (at 0.1 cM intervals by default). The selection is complemented with variant sharing at rare variants, as rare-variant sharing is likely to arise from a recent common ancestor.

### Sparse HMM computations

Imputation and phasing are performed using the forward–backward algorithm on the Li and Stephens HMM<sup>14</sup>, where reference haplotypes represent the states of the HMM. The computation of posterior probabilities is a computationally intensive task, linear in the number of haplotypes and markers.

The sparse matrix representation of the reference haplotypes in GLIMPSE2 implementation allows to remove the linear component at the marker level during the HMM calculations. GLIMPSE2 selects only  $K$  (default  $K = 2,000$ ) haplotypes with the sparse PBWT selection to assemble a custom reference panel in which most of the rare variants present in the original reference panel are monomorphic. In the forward–backward algorithm these monomorphic variants do not contribute to the overall state probability. Therefore, in GLIMPSE2 the forward–backward probabilities are computed only at sites that are polymorphic in the custom reference panel, adjusting the transition probability to consider the physical distance between two consecutive polymorphic sites. Posterior probabilities of variants that are monomorphic in the custom reference panel can be quickly computed using the appropriate emission probability.

Our method takes advantage of low-level programming language (AVX2 intrinsics) to optimize the HMM forward–backward computations at the hardware level, working on blocks of eight floats. This allows the method to be efficient in the core part of the algorithm and therefore use twice the number of states and larger imputation windows compared to the previous version of GLIMPSE. More details are provided in Supplementary Note 1.2.2.3.

### Evaluation of imputation accuracy

We measured imputation performance as the squared Pearson correlation between high-coverage genomes and imputed dosages. We pooled all validation and imputed dosages belonging to the same frequency bin and computed a single squared Pearson correlation value per bin. Statistics summarizing the number of variants falling in each allele count bin are provided in Supplementary Tables 2–4. We used the GLIMPSE2 concordance tool to measure the squared Pearson correlation by streaming the imputed and validation data to maintain low memory requirements.

We also evaluated the non-reference discordance rate (NRD), defined as the rate between mismatches at the three possible genotypes, divided by the same mismatches plus heterozygous and homozygous alternative matches. We define the non-reference concordance rate as  $NRC = 1 - NRD$ . We provide more information about the benchmark and measurement of imputation accuracy in Supplementary Notes 1.3 and 1.3.1, respectively.

### Evaluation of association tests

We used chromosome 1 data for a subset of 10,000 unrelated UKB individuals of white British ancestry randomly sampled and a total of 99 phenotypes, selected as phenotypes with <10% of missing data in our call set across anthropomorphic traits and blood measurements. We performed association tests using plink2 (ref. 16) with default parameters and the first ten principal components plus sex and age as covariates to test phenotypes for associations with the seven call sets we generated: high-coverage WGS, five low-coverage WGS (0.1×, 0.25×, 0.5×, 1.0× and 4.0×) and the UKB Axiom array. We selected associations that are genome-wide significant ( $P < 5 \times 10^{-8}$ ) and independent (being at least 500 kilobases apart). Out of the phenotypes analyzed, a total of 22 showed significant associations on chromosome 1 in the high-coverage dataset. These 22 phenotypes were chosen for comparison across the six imputed call sets.

To assess the accuracy of GWAS performed using imputed call sets, we compared association strength and effect sizes by computing the Pearson correlation between imputed and high-coverage GWAS experiments. We additionally assess the ability of GWAS experiments to distinguish significant from non-significant signals, considering the high-coverage GWAS to be the ground truth. For this, we computed the sensitivity, the proportion of genome-wide significant associations that can be retrieved, and the specificity, the proportion of genome-wide non-significant associations that can be retrieved using imputed call sets.

### Statistics and reproducibility

This study was based on the UKB SNP array and WGS datasets, Simons Genome Diversity Project, 1,000 Genomes Project and the Haplotype Reference Consortium (HRC). Variants and samples selected are based on quality controls and ancestry as described by the respective dataset. For certain analysis samples were extracted randomly from the UKB cohort, according to their ancestry. Statistical analyses, including Wilcoxon tests were performed with R v.4.0. All code to reproduce analyses is publicly available (Code availability section).

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

The 1,000 Genomes Project phase 3 dataset sequenced at high coverage by the New York Genome Center is available on the European Nucleotide Archive under accession no. PRJEB31736, the International Genome Sample Resource (IGSR) data portal and the University of Michigan school of public health ftp site (<ftp://share.sph.umich.edu/1000g-high-coverage/freeze9/phased/>). The publicly available subset of the HRC dataset is available from the European Genome-phenome Archive at the European Bioinformatics Institute under accession no. EGAS00001001710. The publicly available Simons Genome Diversity project is available on the IGSR data portal and Cancer Genomics Cloud, powered by Seven Bridges. The UKB WGS data and phenotypes can be accessed via RAP: <https://ukbiobank.dnanexus.com/landing>. The phased WGS reference panel can be accessed via RAP: <https://ukbiobank.dnanexus.com/landing>. Source data are provided with this paper.

### Code availability

GLIMPSE2 source code is available with MIT licence from <https://github.com/odelaneau/GLIMPSE> and <https://odelaneau.github.io/GLIMPSE/>. This includes code to the chunk, split\_reference, phase, ligate and concordance. The documentation is available at <https://odelaneau.github.io/GLIMPSE/>. Code and source data to reproduce analysis and figures have been deposited in a Zenodo repository<sup>17</sup>.

### References

- Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
- Durbin, R. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).
- Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
- Rubinacci, S., Hofmeister, R. J., Sousa da Mota, B. & Delaneau, O. Source data, scripts and code for the manuscript ‘Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes’. *Zenodo* <https://doi.org/10.5281/ZENODO.7860468> (2023).

### Acknowledgements

This work was funded by a Swiss National Science Foundation project grant 373 (PPOOP3\_176977) and conducted under UKB project 66995. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank the participants of the UKB. The sequencing of 150,119 UKB samples used in this study has been funded by the UKB WGS consortium. DNA sequencing was performed at the Wellcome Trust Sanger Institute and deCODE genetics. The New York Genome Center 1000 Genomes data were generated at the New York Genome Center with funds provided by a National Human Genome Research Institute grant no. 3UM1HG008901–03S1.

### Author contributions

S.R. and O.D. designed the study. S.R. and O.D. developed the algorithms and wrote the software. R.J.H. performed the GWAS experiments. S.R. and B.S.M. performed imputation of ancient samples. B.S.M. provided interpretation regarding imputed ancient samples. S.R. performed the remaining experiments. O.D. supervised the project. All authors reviewed the final paper.

### Competing interests

The authors declare no competing interests.

### Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41588-023-01438-3>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41588-023-01438-3>.

**Correspondence and requests for materials** should be addressed to Olivier Delaneau.

**Peer review information** *Nature Genetics* thanks Arnaldur Gylfason, Tobias Marschall and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).



## Appendix C

### Parent-of-Origin inference for biobanks

This article is presented in [Chapter II](#).

The online version and the supplementary material can be downloaded from <https://www.nature.com/articles/s41467-022-34383-6> .




# Parent-of-Origin inference for biobanks

Received: 18 March 2022

Accepted: 24 October 2022

Published online: 05 November 2022

 Check for updatesRobin J. Hofmeister<sup>1,2</sup>, Simone Rubinacci<sup>1,2</sup>, Diogo M. Ribeiro<sup>1,2</sup>,  
Alfonso Buil<sup>3,4</sup>, Zoltán Kutalik<sup>1,2,5</sup> & Olivier Delaneau<sup>1,2</sup> ✉

Identical genetic variations can have different phenotypic effects depending on their parent of origin. Yet, studies focusing on parent-of-origin effects have been limited in terms of sample size due to the lack of parental genomes or known genealogies. We propose a probabilistic approach to infer the parent-of-origin of individual alleles that does not require parental genomes nor prior knowledge of genealogy. Our model uses Identity-By-Descent sharing with second- and third-degree relatives to assign alleles to parental groups and leverages chromosome X data in males to distinguish maternal from paternal groups. We combine this with robust haplotype inference and haploid imputation to infer the parent-of-origin for 26,393 UK Biobank individuals. We screen 99 phenotypes for parent-of-origin effects and replicate the discoveries of 6 GWAS studies, confirming signals on body mass index, type 2 diabetes, standing height and multiple blood biomarkers, including the known maternal effect at the *MEG3/DLKI* locus on platelet phenotypes. We also report a novel maternal effect at the *TERT* gene on telomere length, thereby providing new insights on the heritability of this phenotype. All our summary statistics are publicly available to help the community to better characterize the molecular mechanisms leading to parent-of-origin effects and their implications for human health.

Parent-of-Origin (PoFO) effects refer to genetic variations having an effect on a phenotype that depends on the parent from which alleles are inherited<sup>1,2</sup>. PoFO effects are thought to mainly result from genomic imprinting, a mechanism relying on parent-specific DNA methylation, named imprints, that silence one of the parental copies of a gene. Such parent-specific imprints are established during spermatogenesis and oogenesis and are maintained in all somatic cells of the offspring<sup>3</sup>. This leads to some genes, called imprinted genes, to exhibit an allele-specific expression pattern that depends on the PoFO of the underlying genetic sequence. This allele-specific expression can be maintained throughout life or specific to some development states<sup>4</sup>. One of most studied imprinted loci in the human genome is probably the HI9 loci at 11p15.5 that is involved in growth and development disorders such as the Beckwith–Wiedemann or Silver–Russel syndromes<sup>5</sup>. Multiple studies have investigated PoFO effects on complex traits, notably for the

*KCNQ1* and *KLF14* genes whose associations with type 2 diabetes risk depends only on the maternal copies<sup>6</sup>, as well as for the *MEG3/DLKI* imprinted locus associated with age at menarche<sup>7</sup> and platelet count<sup>8</sup>.

Searching for PoFO effects on a genome-wide scale requires knowing the PoFO of each individual allele. The most direct approach to obtain this information relies on the availability of parental genomes, which allows using the Mendelian principles of inheritance to determine the parent from which a specific allele is inherited. Study cohorts usually include a small number of genotyped parent-offspring duos and trios, resulting in a low discovery power and a challenging detection of PoFO effects. To alleviate this problem, multiple approaches have been explored so far. First, by deploying large efforts in data collection, such as the study performed on the DiscovEHR cohort<sup>9</sup>, representing one of the largest PoFO study done to date, with hundreds of phenotypes assessed for more than 22,000 samples with at

<sup>1</sup>Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. <sup>2</sup>Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland.

<sup>3</sup>Institute of Biological Psychiatry, Mental Health Services, Copenhagen University Hospital, Copenhagen, Denmark. <sup>4</sup>Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark. <sup>5</sup>University Center for Primary Care and Public Health, University of Lausanne, Lausanne, Switzerland. ✉e-mail: [olivier.delaneau@unil.ch](mailto:olivier.delaneau@unil.ch)



least one genotyped parent. Alternatively, this can also be achieved by meta-analysis across multiple cohorts regrouping duos and trios, with the caveat of restricting the analysis to the subset of phenotypes in common across datasets<sup>7,10</sup>. Second, statistical approaches have been proposed to test for PoFO effects in large collections of unrelated samples by exploiting the differences in phenotypic variance between heterozygous and homozygous individuals, with the caveat of also detecting effects unrelated to PoFO such as gene-environment interactions<sup>11</sup>. Third, it has been shown that the PoFO of an individual's alleles can also be determined by the use of cousins as *surrogate parents* when parental genomes are not available<sup>6</sup>. This latter approach is particularly well suited for datasets comprising many samples from the same generation but also requires the genealogy of most individuals in the study cohort, which is not the case in large datasets such as the UK Biobank<sup>12</sup>.

In this work, we present a probabilistic method to infer the PoFO of alleles in biobank scale datasets from second- and third-degree relatives without requiring any parental genomes nor explicit genealogy to be known. To do so, our approach combines multiple estimation steps, involving surrogate parent groups formation, parental status assignment based on chromosome X, haplotype inference, Identity-By-Descent (IBD) detection, and haploid imputation. When applied to the UK Biobank dataset, this allows us to infer the PoFO for 21,484 samples with high confidence in addition to the 4909 duos/trios for which we perform direct inference from parental genomes, resulting in a dataset comprising a total of 26,393 samples and 7.6 million variants. Considering duos/trios as the ground truth, we show that our PoFO estimations from second- and third-degree relatives have a high call rate (~75%) and low error rate (<1%) at heterozygous genotypes. Taking advantage of the vast phenotypic diversity of the UK Biobank, we carried out genome-wide association scans for PoFO effects for a total of 99 phenotypes, replicating well-known imprinted loci as well as discovering novel putative PoFO associations, thereby demonstrating that our method has the potential to further reveal the contribution of PoFO effects to complex traits. All the summary statistics for the conducted association scans are publicly available (<http://poedb.dcsr.unil.ch/>) and allow the exploration of the PoFO effects for variants of interest across phenotypes.

## Results

### PoFO inference from genotype data

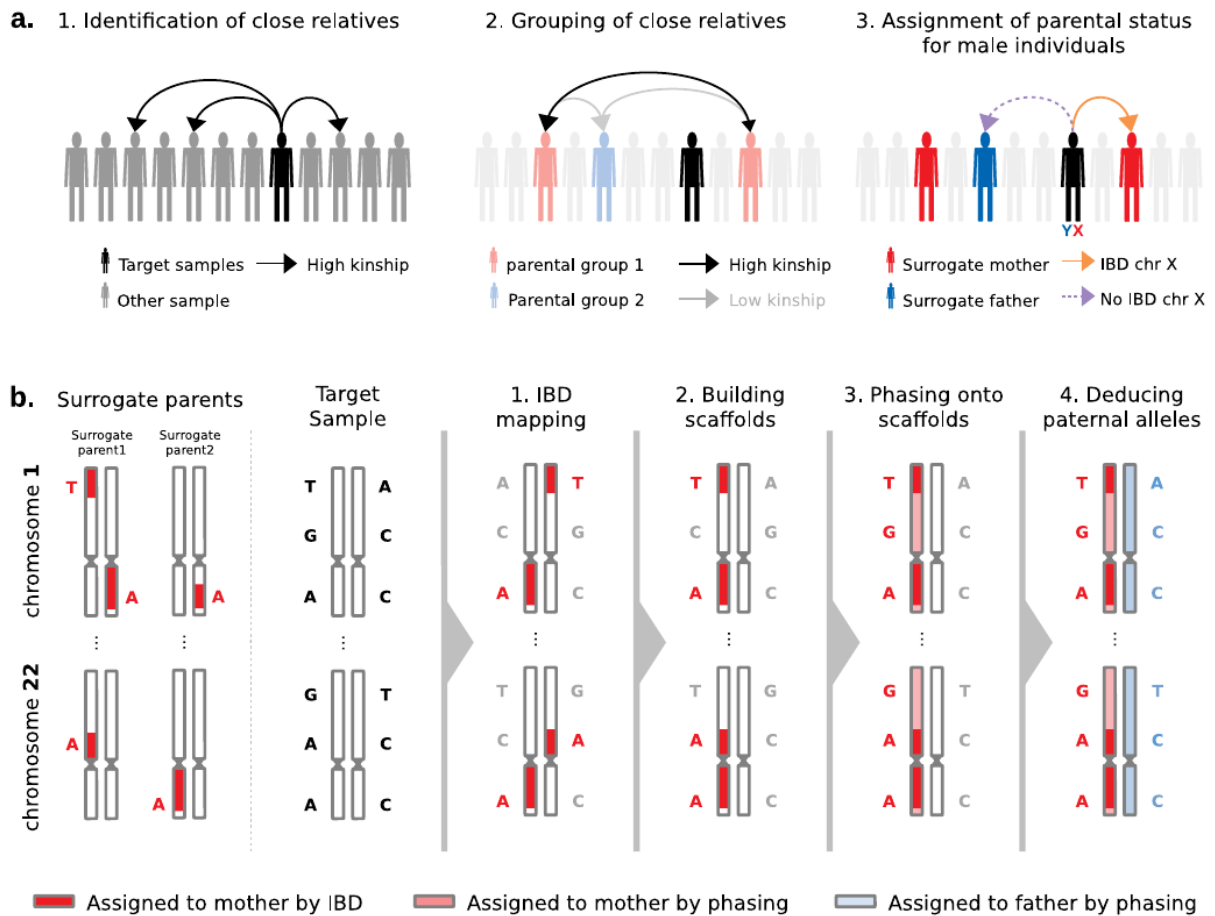
To infer the PoFO of all alleles carried by a given target sample, we proceed in two consecutive steps as detailed below:

1. *Identification of surrogate parents* (Fig. 1a). For each target sample (white British individual of the UK Biobank), we identify close relatives, and we determine which of the two parents (mother or father) conveys the relatedness. For this, we first look at pairwise kinship estimates given by KING<sup>13</sup> to identify second- or third-degree relatives and group them into the two parental groups based on their relatedness: they cluster in the same group if they are related and in different groups otherwise. Then, we assign parental status (maternal or paternal) to parental groups for male targets only by exploiting the fact that their single chromosome X copy is maternally inherited. Therefore, we search for relatives sharing portions of their chromosome X Identical-By-Descent (IBD) with the target sample and we label them as surrogate mothers. We also propagate the information to other relatives: those from the same parental group are also labeled as surrogate mothers and those from the other parental group as surrogate fathers. In case no IBD is found, we cannot annotate parental groups as maternal or paternal and we exclude the target sample from the dataset. Hereafter, we call *surrogate parents* the close relatives we identified using this approach.
2. *Assignment of PoFO to alleles* (Fig. 1b). After the identification of surrogate parents, we assign PoFO to the target's alleles. First, we

search for autosomal shared IBD segments between the target and the surrogate parents using IBD mapping robust to both phasing and genotyping errors (see **Methods**, Supplementary Fig. 1). Then, we classify the resulting IBD segments as being maternally or paternally inherited depending on the surrogate parent they map to. This delimits a subset of alleles that are co-inherited from the same parent within and across chromosomes (i.e., that co-localize on the transmitted set of homologous chromosomes). This leaves another subset of alleles for which we do not know the PoFO (i.e., those not shared IBD with any of the surrogate parents). For those, we extrapolate the PoFO using statistical phasing: we model alleles for which we know the PoFO status as a haplotype scaffold<sup>14</sup> onto which all remaining alleles are probabilistically phased using SHAPEIT4<sup>15</sup> (Supplementary Fig. 1). The PoFO assignment of these remaining alleles is then given by their frequency of co-localization onto each haplotype scaffold, which also reflects how reliable the phasing is (i.e., phasing certainty, Supplementary Fig. 1). Finally, we extrapolate the PoFO for untyped variants by performing haploid imputation of each parental haplotype in turn using IMPUTE5<sup>16</sup> and the HRC as reference panel<sup>17</sup>.

### Validation of the PoFO inference on duos and trios

To assess the accuracy of our approach, we used 443,993 genotyped UK Biobank samples of British or Irish ancestry, together with their pairwise kinship estimates, to identify a subset of samples with parents and second-to-third degree relatives. For these samples, we inferred the PoFO using two approaches: directly from the parents or using second-to-third degree relatives as surrogate parents. We compared the quality of the PoFO inference given by surrogate parents to the direct approach based on parental genomes, considered to be the ground truth. We found a total of 3872 parent-offspring duos and 1037 trios, of which 1090 duos and 309 trios also have groups of surrogate parents. We used this subset of 1399 samples to assess optimal parameters and the accuracy of the method. We focused on two metrics: (i) the error rate, which is the percentage of heterozygous genotypes with incorrect PoFO assignment and (ii) the call rate, which is the percentage of heterozygous genotypes for which a PoFO call could be made (see **Methods**). We explored a range of different parameter settings for the IBD detection and PoFO confidence score (i.e., phasing certainty onto the haplotype scaffold) and found that using haplotype segments longer than 3 cM as scaffold and a phasing certainty above 0.7 lead to a good trade-off between call rate and error rate (Fig. 2a). This resulted in a whole genome error rate of 0.51% and a call rate of 74.5%. As expected, the error and call rate depend on the number of available surrogate parents per target, with the call rate increasing and the error rate decreasing as the number of surrogate parents increases (Fig. 2b). The majority of our targets only have a single surrogate parent (75.95% of the target samples, Fig. 2c) and even in this case, a call rate of 70.9% and an error rate of 0.6% is achieved (Fig. 2b). We then considered the genomic localization of variants: we found a lower call rate and a slightly higher error rate as we approach telomeres, which results from phasing edge effects (Fig. 2d). Overall, we found small error rates for the majority of the variants: 79% have an error rate <1% and 56% inferred perfectly (Fig. 2e). This low error rate mostly results from the high phasing accuracy that can be achieved in the UK Biobank using SHAPEIT4<sup>15</sup>. Overall, we obtained a whole genome switch error rate of 0.0845% between consecutive heterozygous genotypes when comparing to parental genomes, with only small variations across chromosomes (Supplementary Fig. 2A). When looking at the distribution of these switch errors along the genome, we found that they mostly occur within small segments and that long range errors are almost entirely corrected by the use of haplotype scaffolds (Supplementary Fig. 2B). As a result, we obtained haplotypes that are resolved across entire chromosomes with only a few sporadic errors that, given their



**Fig. 1 | Rationale of PofO inference.** **a** Identification of surrogate parents in 3 steps: (1) identification of close relatives for a target sample of interest using the pairwise kinship estimates, (2) clustering of close relatives by maximizing and minimizing the inter- and intra-groups relatedness, respectively, (3) assignment of parental status to close relatives' groups (i.e., surrogate parents) using IBD sharing on chromosome X for male targets. **b** Parent-of-origin inference in 4 steps: (1)

identification of autosomal IBD segments shared between the target and the surrogate parents, (2) scaffold construction with co-inherited alleles localized on the same homologous chromosome across all autosomes, (3) statistical phasing of all remaining alleles against the scaffold and (4) whole genome deduction of the maternal and paternal origins of alleles from phasing probabilities.

frequency (<0.1% error rate), we believe to result mostly from genotyping errors.

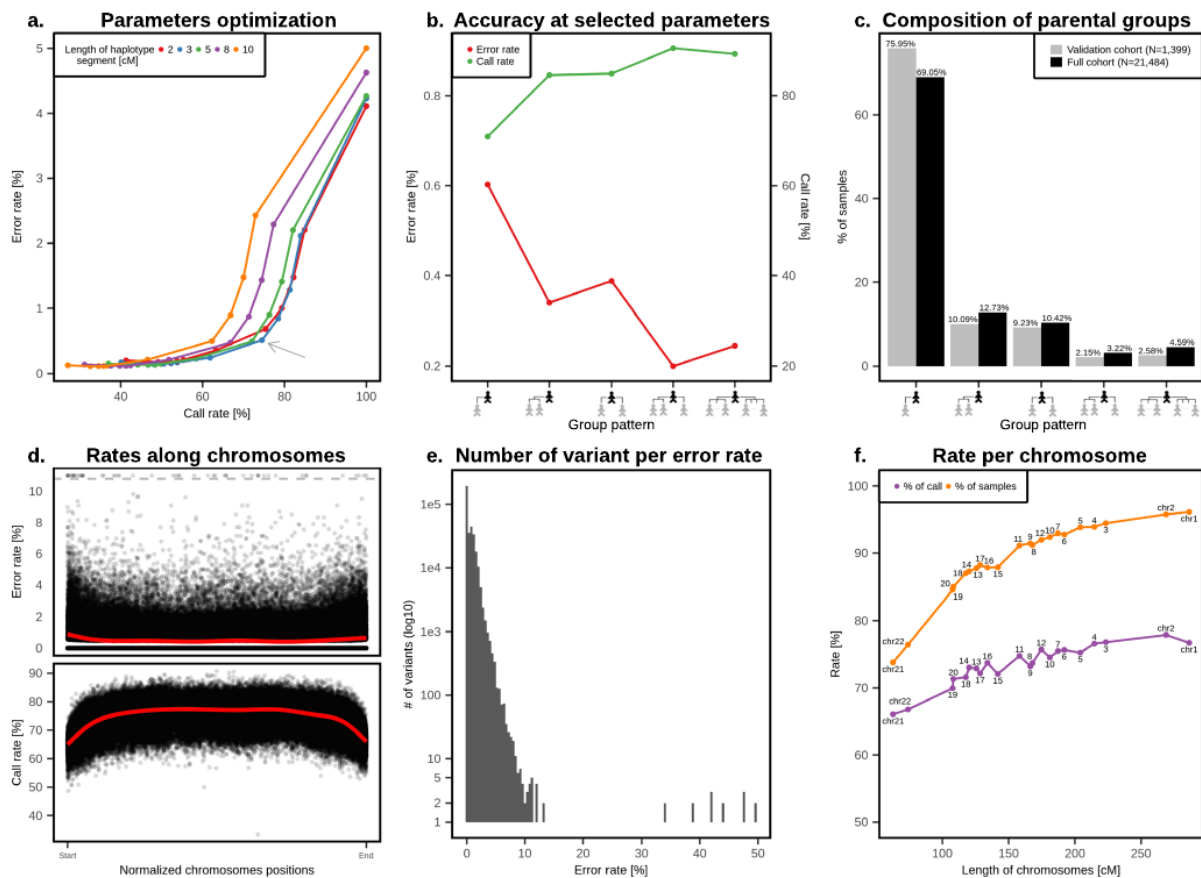
### PofO inference in 26,393 individuals

For all genotyped British and Irish individuals in the UK Biobank without any genotyped parent ( $N=438,993$ ), we inferred the PofO using the method based on surrogate parents, as described above. In total, we found 105,826 samples with second-to-third degree relatives forming groups of surrogate parents. Amongst those, we could assign parental status to surrogate parent groups to a subset of 21,484 samples using IBD matching on chromosome X. Comparing the distribution of surrogate parents per target sample, we found a remarkable match between the full ( $N=21,484$ ) and the validation ( $N=1399$ ) datasets (Fig. 2c), suggesting that we can expect similar error rates between datasets. As our method requires IBD sharing between the targets and the surrogate parents, no inference can be made for chromosomes where no IBD sharing is found. The number of samples with PofO inference thus varies across chromosomes depending on their length, ranging from 15,645 samples (72.8%) for chromosome 21 to 20,381 samples (94.8%) for chromosome 1 (Fig. 2f). It follows that the call rate also varies across chromosomes, ranging from 66% for chromosome 21 to 77.9% for chromosome 2 (Fig. 2f). From the sample

point-of-view, we found that 31.3% of the samples have inference for the 22 autosomes and 96.1% have inference for more than 15 chromosomes (Supplementary Fig. 3). Finally, we merged the 21,484 samples with PofO inferred from surrogate parents together with the 4909 samples with PofO inferred from genotyped parents (3872 duos and 1037 trios) to get a final set comprising a total of 26,393 individuals with PofO inference (22,652 males and 3741 females) across 7.6 million variants genome-wide (Supplementary Table 1). Together with deep phenotyping provided by the UK Biobank, this represents a unique dataset to study PofO effects on complex traits.

### Discovery of PofO associations

Distinguishing paternally from maternally inherited alleles allows us to design different association scans to test the PofO specificity of associations: (i) maternal, to test only the maternally inherited alleles, (ii) paternal, to test only the paternally inherited alleles, (iii) differential, to compare paternally and maternally inherited alleles at heterozygous genotypes only and (iv) additive, as a control to test minor alleles regardless of PofO. Using these models we scanned for association 99 quantitative phenotypes of the UK biobank using BOLT-LMM<sup>18</sup> (Supplementary Data 1), for which we provide all summary statistics online (<http://poedb.dcsr.unil.ch/>).



**Fig. 2 | Validation of the PoFo inference.** **a** Call rate (x-axis) and error rate (y-axis) as a function of (i) the minimal length of IBD tracks for scaffold construction and (ii) the minimal phasing probability used to call a heterozygote as phased. Each point corresponds to a given phasing probability threshold going from 0.5 (right most point) to 1.0 (left most point) with steps of 0.05. The grey arrow indicates the parameters we used in our analysis (3 cM long IBD tracks and 0.7 minimal phasing probability). **b** Call rate (left y-axis) and error rate (right y-axis) as a function of the composition of the parental groups (x-axis). The latter ranges from one parental group with one surrogate parent (left) to two parental groups comprising multiple surrogate parents (right). **c** Fraction of targets as a function of the composition of

the parental groups (x-axis): in the validation data ( $N = 1399$ ) in gray and in the call set ( $N = 21,484$ ) in black. **d** Error rate (top panel; y-axis) and call rate (bottom panel; y-axis) per variant site as a function of their normalized positions relative to each telomere (x-axis). Red lines are fitted density curves. Error rates greater than 10% are capped to 11% as indicated by the dashed gray line. **e** Distribution of error rates per number of variant sites (y-axis, log scale). **f** Fraction of samples (purple) and heterozygotes (i.e., call rate; orange) in the call set for which PoFo is inferred, as a function of chromosome length (cM, x-axis). Chromosome numbers are shown next to the points in black. Source data are provided as a Source Data file.

In a first pass, we focused on variants being Bonferroni significant in both the differential and additive scans ( $p < 5 \times 10^{-6}$ ) and used the paternal and maternal scans to determine the parental origin and the direction of the effect. We found two signals fulfilling these criteria. The first is a PoFo association with platelet phenotypes at the *MEG3/DLKI* imprinted locus<sup>19</sup> (Table 1 and Fig. 3a, b). The lead SNP rs59228823 is an eQTL for *MEG3* in blood samples<sup>20</sup> and is associated with platelet count and platelet crit under the additive, maternal and differential scans but not under the paternal scan (Table 1). The minor allele C at this SNP significantly decreases the platelet count and crit when maternally inherited (Table 1 and Fig. 3c). A similar maternal effect has been previously reported on platelet count for another SNP in the same locus<sup>8,9</sup>: rs1555405, which is in linkage disequilibrium with rs59228823 ( $R^2 = 0.59$ ). We also replicated the association at rs1555405 (maternal, paternal, differential  $p$ -values =  $2 \times 10^{-16}$ , 0.13,  $1.4 \times 10^{-6}$ , Supplementary Fig. 4A), suggesting that these two associations capture the same effect. The second PoFo association we found is at SNP rs2735940 for the leukocyte Telomere Length (TL) phenotype, with the minor allele G decreasing TL only when maternally inherited (Table 1 and Fig. 4a–c). This SNP is located -1.5 kb upstream of the

promoter of *TERT* (Fig. 4b), a gene encoding for the catalytic subunit of the telomerase, an enzyme involved in TL maintenance<sup>21</sup>. This SNP is in high linkage disequilibrium with the SNP rs2853677 ( $r^2 = 0.6$ ), previously reported in different GWAS for multiple cancers<sup>22,23</sup>, blood cell counts<sup>24</sup>, aging<sup>25</sup> and telomere length<sup>26</sup>. When directly testing rs2853677 in our data, we find a strong maternal effect similarly to the lead SNP (maternal, paternal, differential  $p$ -values =  $4.6 \times 10^{-17}$ , 0.8,  $7.9 \times 10^{-11}$ , Supplementary Fig. 4B). This suggests that a parent-of-origin effect underlies this pleiotropic locus.

In a second pass, we focused on associations that are Bonferroni significant in the differential scans but not supported by the additive scans. In total, we found 14 of these associations that we classified as putative PoFo effects (Supplementary Table 2). This includes three maternal associations, four paternal associations, and seven associations with opposite effect between the paternally and maternally inherited alleles which are consistent with a pattern of bipolar dominance<sup>2</sup>. To confirm these results, we used a method developed by Hoggart et al.<sup>11</sup> designed to capture PoFo effects by detecting an increased variance across heterozygous compared to homozygous. Using this method on the full set of British samples ( $N = 443,993$ ), all

**Table 1 | Discovery of PoFO associations**

Phenotypes	SNP	Chr	Position (hg19)	Risk allele	MAF	Add.P	Add.B	Pat.P	Pat.B	Mat.P	Mat.B	Diff.P	PoFO effect	Mapped gene	UKB phenotype code	Hoggart et al. P.
Adjusted T/S ratio, Telomeres length	rs2735940	5	1296486	G	0.49	3.7e-08	-0.049	0.46	0.008	2.1e-19	-0.121	4.3e-13	Maternal	TERT	22191	0.00183
Platelet crit	rs59228823	14	101185187	C	0.24	1.6e-10	-0.065	0.66	-0.007	6.6e-17	-0.124	2.9e-09	Maternal	MEG3, DLK1	30090	0.00180
Platelet count	rs59228823	14	101185187	C	0.24	8.6e-10	-0.064	0.58	-0.011	6.8e-16	-0.123	2.3e-08	Maternal		30080	0.00180

PoFO effects were identified based on a Bonferroni correction ( $5 \times 10^{-6}$ ) on both the differential and the additive scans. Risk allele represents the minor allele tested in each of the four scans. Additive betas represent the phenotypic effects of minor alleles; Paternal betas represent the phenotypic effects of paternally inherited minor alleles. Maternal betas represent the phenotypic effects of maternally inherited minor alleles. Differential betas are not shown since they depend only on the parental alleles taken as reference. Genes were mapped using either eQTLs or ensemble Variant Effect Predictor (VEP). P-values are computed using BOLT-LMM<sup>40</sup>, expect Hoggart et al.<sup>11</sup>. P-values that are computed using the increased variance method (see Methods). Add Additive, Pat. Paternal, Mat. Maternal, Diff. Differential, P = p-values; B = betas; Chr=Chromosome.

associations have  $p$ -values  $<0.007$  (Supplementary Table 2). The strongest opposite PoFO effect involves the variant rs77403171 at 2q22.3, intronic to *ARHGAP15* and decreasing the eosinophil percentage when maternally inherited while increasing the trait when paternally inherited. *ARHGAP15* is a Rho GTPase-activating protein that has already been associated with multiple blood cell phenotypes, notably neutrophils, leukocytes, and eosinophils<sup>27-29</sup>. These are examples of genetic effects missed by the additive model as paternal and maternal contribution at heterozygous sites cancel out when considered together.

Finally, we used the PoFO associations at the *MEG3/DLK1* locus and at the *TERT* locus to illustrate the benefit of using our PoFO inference on the discovery power of PoFO effects. To do so, we used 4909 UK Biobank duos/trios as baseline and gradually added random subsets of 5000 samples for which PoFO inference could be made from surrogate parents, ending up with the full set of 26,393 samples. Doing so led to a clear boost in association strength for the additive, maternal and differential signals, with maternal scans ranging from non-significant on the duos/trios for both platelet crit and TL ( $n=4909$ ;  $p$ -value =  $6.28 \times 10^{-04}$  and  $8.36 \times 10^{-05}$ , respectively) to strongly significant on the full sample size ( $n=26,393$ ;  $p$ -value =  $6.6 \times 10^{-17}$  and  $2.1 \times 10^{-19}$ , respectively; Fig. 5a, b), while the paternal signal remained non-significant. Similarly, we also looked at the effects of errors in the PoFO inference on the discovery power by randomizing the PoFO assignment for an increasing number of samples. This progressively diluted the maternal signal onto the two parental origins while leaving the additive signal unchanged (Fig. 5c, d). Interestingly, the association with TL remains significant with up to 10% of errors in the PoFO inference, suggesting that PoFO testing could tolerate relatively high error rates with our sample size.

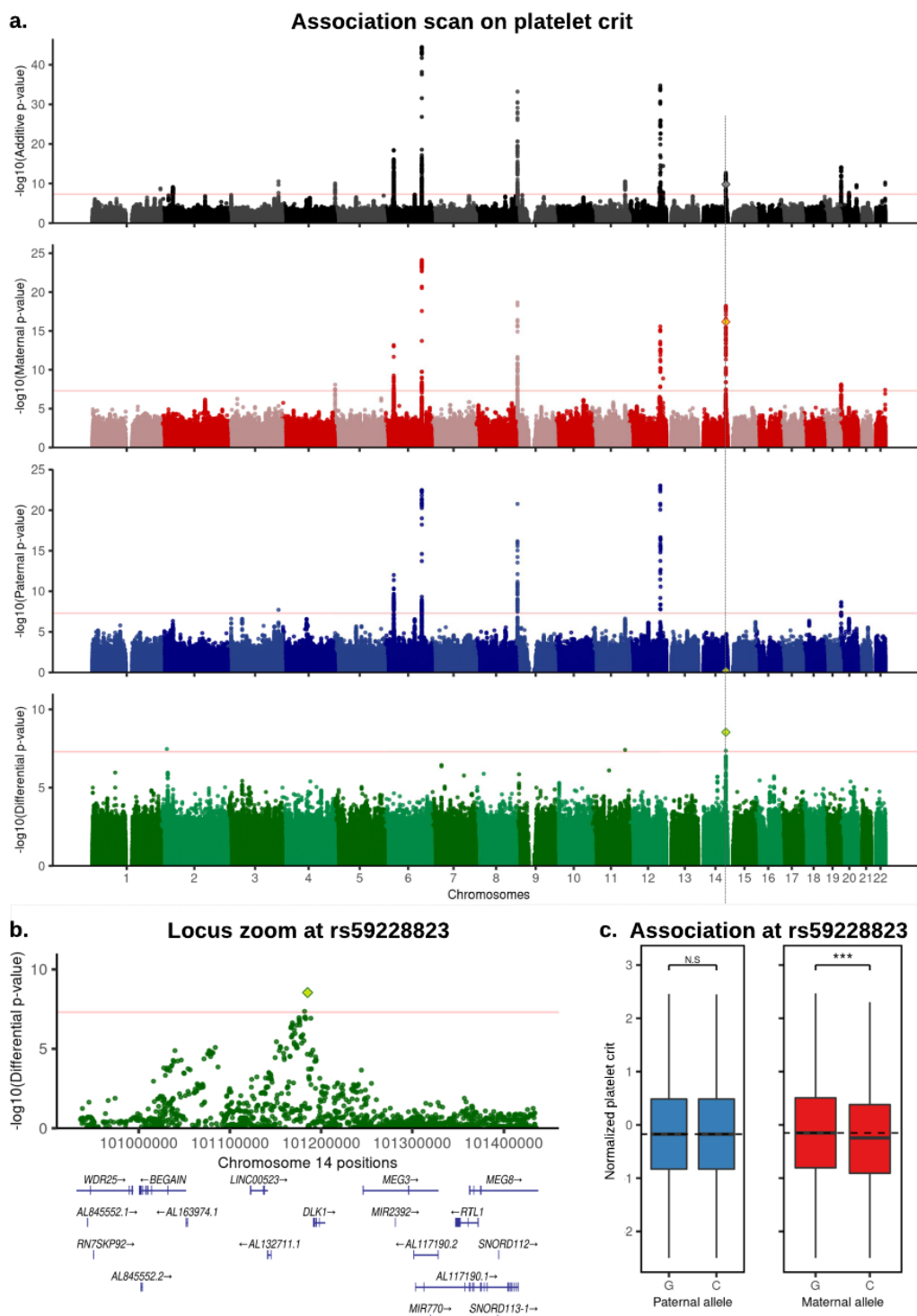
### Replication of PoFO associations

The PoFO callset for the UK Biobank generated with our method provides a powerful resource to replicate independent PoFO associations or to annotate other types of associations as PoFO effects. To show this, we assessed the ability of our method to replicate the results of seven GWAS studies across multiple phenotypes often studied in the context of PoFO effects. These studies belong to three different categories: (i) PoFO studies using trios or known genealogies, (ii) PoFO studies across unrelated individuals using an increased-variance method, and (iii) studies investigating genotype-environment (GxE) interactions.

**Standing height.** We focused on the 11 PoFO associations reported in three studies making use of genealogy-based PoFO inference<sup>30-32</sup>, 9 of which could be assessed in our data (identical SNP-phenotype pair in the UK Biobank). Seven of these associations are located in two well-known imprinted regions, 11p15.5 and 14q32, and the remaining two are located in the HLA region. Only one association has been replicated across the two of the three studies at rs143840904. In contrast, we replicated 8 associations out of the 9 we could test, with the same parent and direction of effects (Table 2A-C), thereby reinforcing the role of these two well-known imprinted regions on height and providing further evidence on the PoFO effect at the HLA region.

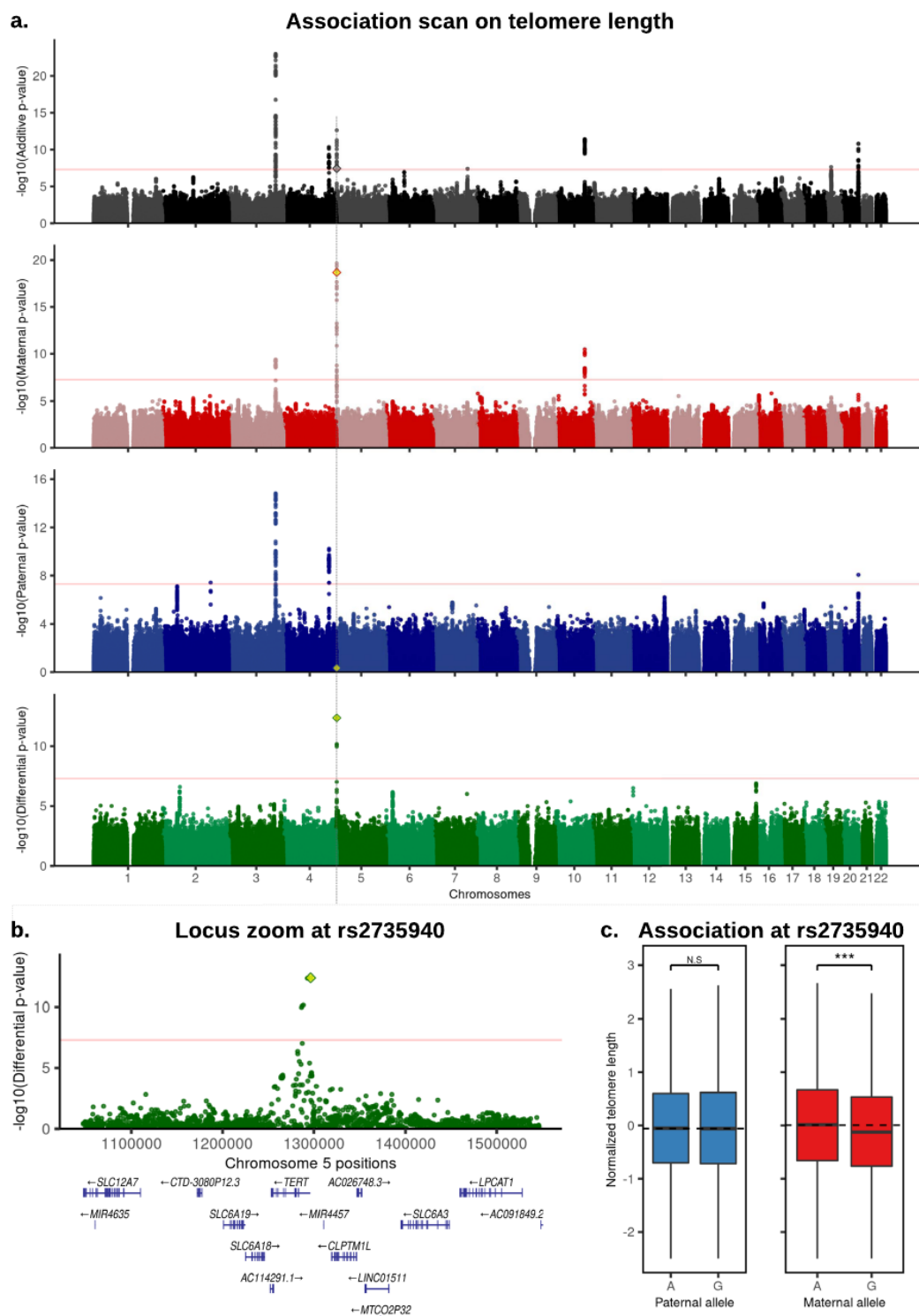
**Blood biomarkers.** A recent study<sup>9</sup> examined multiple blood biomarkers and reported a total of 10 PoFO associations within imprinted loci using trios-based PoFO inference. In our dataset we were able to assess 7 of these associations and replicated 5 of them, with the same parent and direction of effects (Table 2D). This included the PoFO effect on platelet phenotypes at the *MEG3/DLK1* locus we reported earlier.

**Type 2 diabetes.** Kong et al.<sup>6</sup> reported a total of 4 PoFO associations on type 2 diabetes (T2D) using genealogy-based PoFO inference. They fall within two distinct regions that harbor well-documented imprinted



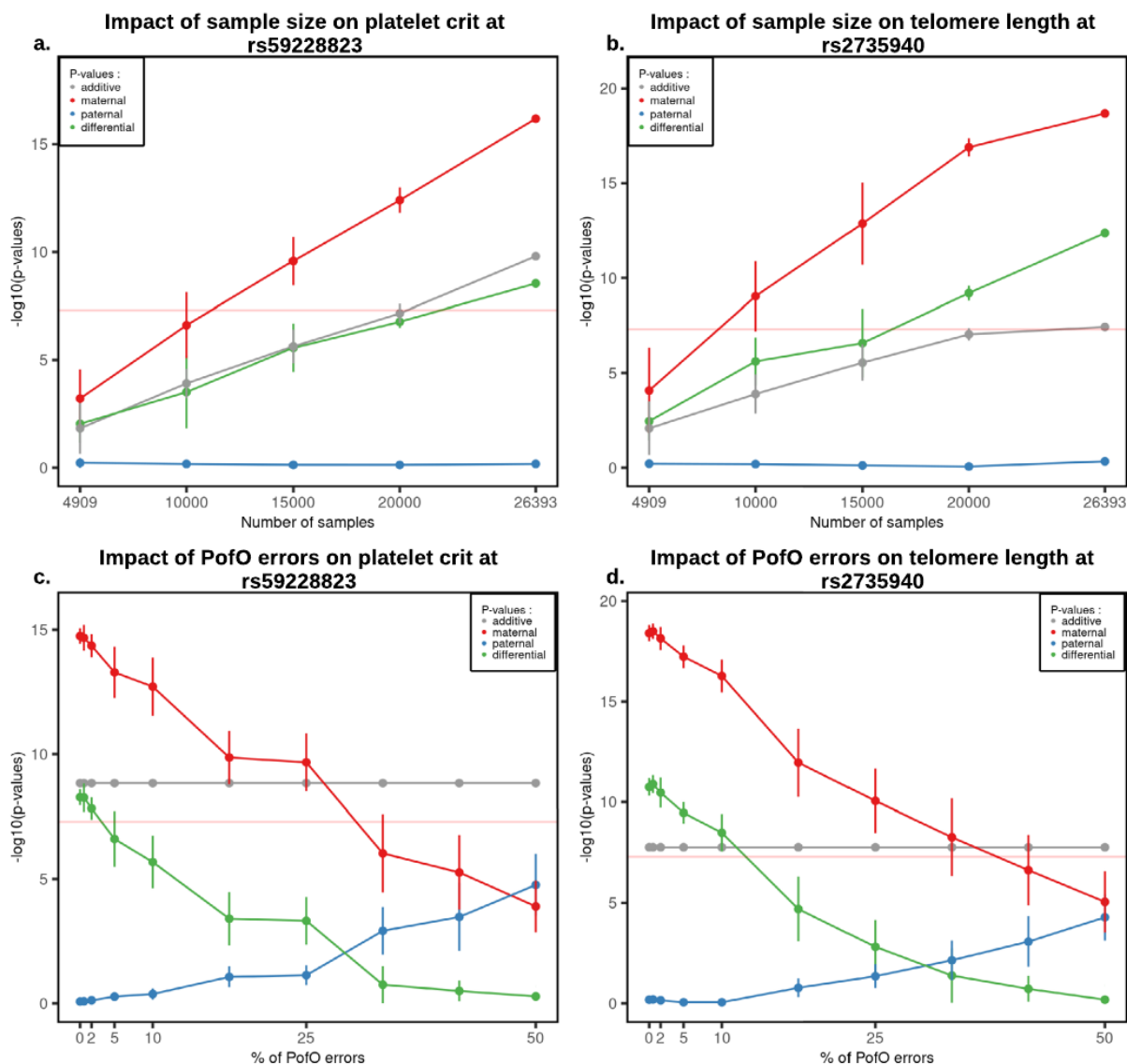
**Fig. 3 | Association scans for Pofo effects on platelet crit.** **a** Manhattan plots of four association scans with platelet crit. From top to bottom plots are shown results for additive (black), maternal (red), paternal (blue) and differential (green) scans. The lead variant mentioned in this study (rs59228823) is shown with a diamond. Red horizontal lines indicate genome-wide significance threshold at  $-\log_{10}(5 \times 10^{-8})$ . **b** Locus zoom at rs59228823 on the differential scan. **c** Box plot of the normalized platelet crit (y-axis) stratified by risk alleles and origin at SNP rs59228823; paternal in

blue and maternal in red (x-axis). The horizontal dotted lines represent the phenotypic median of the major allele G. Boxes bound the 25th, 50th (median), and the 75th quantiles. Whiskers range from minima (lower) to maxima (upper). Sample sizes are  $n_{\text{paternal}}(G/C) = 16,285/4,769$  and  $n_{\text{maternal}}(G/C) = 16,368/4686$  individuals. N.S non-significant ( $p$ -value = 0.66); \*\*\*=significant ( $p$ -value =  $6.6 \times 10^{-17}$ ) (computed with BOLT-LMM<sup>15</sup>). Source data for (a) and (b) are provided as a Source Data file.



**Fig. 4 | Association scans for PoFo effects on telomere length.** **a** Manhattan plots of four association scans with telomere length. From top to bottom plots are shown results for additive (black), maternal (red), paternal (blue) and differential (green) models. The lead variant mentioned in this study (rs2735940) is shown with a diamond. Red horizontal lines indicate genome-wide significance threshold at  $-\log_{10}(5 \times 10^{-8})$ . **b** Locus zoom at rs2735940 on the differential scan. **c** Box plot of the normalized telomere length (y-axis) stratified by risk alleles and origin at SNP

rs2735940; paternal in blue and maternal in red (x-axis). The horizontal dotted lines represent the phenotypic median of the major allele A. Boxes bound the 25th, 50th (median), and the 75th quantile. Whiskers range from minima (lower) to maxima (upper). Sample sizes are  $n_{\text{paternal}}(A/G) = 10,627/10,337$  and  $n_{\text{maternal}}(A/G) = 10,635/10,329$  individuals. N.S non-significant ( $p$ -value = 0.46); \*\*\* = significant ( $p$ -value =  $2.1 \times 10^{-19}$ ) (computed with BOLT-LMM<sup>16</sup>). Source data for (a) and (b) are provided as a Source Data file.



**Fig. 5 | Robustness of the PofO testing.** **a, b** Association strength as  $-\log_{10}(p\text{-value})$  for rs59228823 and rs2735940 (y-axis) on platelet crit and TL, respectively, as a function of the number of randomly chosen samples included in the analysis under the additive (black), paternal (blue), maternal (red) and differential (green) scans. Each point for  $N = [10,000; 15,000; 20,000]$  represents the median  $p$ -value obtained after 10 randomizations with vertical bars representing the standard error. Points for  $N = 4909$  and  $N = 26,393$  represent the  $p$ -values obtained using only the samples with genotyped parents and using our full sample size,

respectively. **c, d** Association strength as  $-\log_{10}(p\text{-value})$  for rs59228823 and rs2735940 (y-axis) on platelet crit and TL, respectively, as a function of the fraction of samples for which PofO has been randomly drawn (x-axis, 100% = 26,393). Samples included are those for which the PofO has been inferred from the surrogate parents. Each point represents the median  $p$ -value obtained after 10 randomizations with vertical bars representing the standard errors.  $P$ -values are computed with BOLT-LMM<sup>48</sup>. Source data are provided as a Source Data file.

gene clusters, 11p15.5<sup>33</sup> and 7q32<sup>34,35</sup>. As we could not directly test T2D status due to the small number of cases in our dataset, we tested the biomarker most correlated with T2D: glycated hemoglobin (HbA1c, <https://ukbb-rg.hail.is/>). By doing so, we replicated the three strongest associations with the same parental effect (Table 2E). In addition, we phenome-wide analyzed these four variants in our dataset and found 22 associations with differential  $p$ -value  $< 0.01$  (Supplementary Data 2) for 20 distinct phenotypes, many of them closely related to T2D. This illustrates how the deep phenotyping of UK Biobank can help to provide new mechanistic insights for these four T2D risk alleles at the biomarker level.

**BMI by increased variance.** Hoggart et al.<sup>11</sup> reported a total of 6 PofO associations with BMI using an increased-variance method designed to capture PofO effects, two of which were replicated using independent family datasets. These include variants associated with known imprinted genes, *SLC2A10* at 20q13.12 and *KCNK9* at 8q24.3. We could replicate the strongest association in our dataset at the *KCNK9* locus, with the T allele of rs2471083 increasing BMI when maternally inherited (Table 2F). Here, our replication offers additional support for the *KCNK9* locus and confirmation of the maternal origin of this effect, an information that the increased-variance approach can not provide.

**Table 2 | Replication of PoFo associations**

SNP	External studies					Our study					Locus				
	Add.P	Add.B	Pat.P	Mat.P	Diff.P	UKB pheno-type code	Add.P	Add.B	Pat.P	Mat.P		Diff.P			
A. Benonisclottir et al. <sup>31</sup>	rs147239461	2.8e-03	-0.043	5.9e-13	9.4e-04	0.056	1.2e-13	0.5	-0.0095	<b>0.017</b>	-0.05	0.076	0.044	3.5e-03	IGF2, H19
	rs7482510	4.5e-04	-0.03	5.1e-11	-0.065	0.018	4.7e-09	0.011	-0.019	0.055	-0.02	0.18	-0.016	0.7	
	rs143840904	1.3e-05	-0.11	0.042	0.057	<b>2.0e-17</b>	1.6e-14	2.2e-06	-0.1	0.03	-0.06	<b>6.2e-07</b>	-0.16	0.041	KCNQ1
	rs41286560	0.078	<b>-0.031</b>	<b>2.2e-08</b>	-0.12	1.7e-03	7.4e-10	0.47	-0.013	<b>0.02</b>	-0.06	0.32	0.025	0.018	R7L1
B. Zoledziwska et al. <sup>30</sup>	rs143840904	4.58e-05	-0.152	0.9653	-0.0021	<b>3.92e-08</b>	7.55e-05	2.2e-06	-0.10	0.03	-0.069	<b>6.2e-07</b>	-0.16	0.041	KCNQ1
	rs2075870	2.65e-05	-0.158	0.793	-0.0172	<b>6.97e-08</b>	2.0e-04	1.3e-04	-0.07	0.1	-0.05	<b>9.6e-07</b>	-0.15	0.03	
	rs149658560	1.01e-04	-0.161	0.8183	-0.0121	<b>2.93e-07</b>	3.0e-04	0.11	-0.02	0.85	0.009	<b>1.0e-04</b>	-0.108	4.6e-03	
	rs67004488	1.2e-06	-0.157	0.3875	-0.40	<b>5.21e-07</b>	2.4e-03	7.2e-04	-0.04	0.024	-0.055	<b>4.7e-04</b>	-0.074	0.33	
C. Granot-Hershkovitz et al. <sup>32</sup>	rs1042136	1.82e-02	-0.006	<b>1.55e-08</b>	-0.023	4.21e-01	1.39e-04	0.3	-0.0072	<b>0.036</b>	-0.025	0.47	0.010	0.064	HLA
	rs1431403	8.54e-05	-0.004	<b>5.41e-06</b>	-0.011	5.88e-03	6.72e-03	0.097	-0.010	<b>0.014</b>	-0.023	0.85	0.003	0.052	
	rs9332053	4.85e-01	0.153	2.28e-01	1.286	9.59e-06	3.17e-03	0.38	0.025	0.35	0.039	0.87	-0.001	0.45	RC8TB2
	rs117989553	9.6e-27	-0.09	1.9e-05	-0.13	0.15	-	3.9e-10	-0.113	1.6e-04	-0.096	5.6e-06	-0.126	0.59	GRB10
	rs117421106	3.2e-19	-0.12	<b>7.9e-05</b>	-0.19	0.084	-	4.7e-04	-0.089	<b>1.7e-04</b>	-0.138	0.27	-0.05	0.087	
	rs117515500	1.1e-07	0.07	0.14	0.07	1.5e-05	0.2	0.66	-0.011	0.74	-0.009	0.67	-0.026	0.82	IGF2R
	rs12154627	1.2e-23	0.04	0.064	0.03	<b>2.1e-16</b>	0.13	0.0018	0.031	0.6	0.001	<b>2.7e-07</b>	<b>0.066</b>	3.4e-04	CPA4, MEST, KLF14
	rs12154627	5.1e-17	-0.04	0.56	-0.01	<b>1.4e-09</b>	-0.1	5.7e-04	-0.035	0.42	-0.015	<b>1.6e-05</b>	-0.06	0.055	
	rs4758459	8.2e-09	0.03	0.073	0.03	<b>5.3e-05</b>	0.07	0.061	0.014	0.89	0.001	<b>3.5e-04</b>	<b>0.044</b>	0.0097	H19, IGF2
	rs10146962	3.2e-20	-0.04	0.12	0.02	<b>2.3e-11</b>	-0.1	1.5e-11	-0.062	0.14	-0.021	<b>7.0e-17</b>	-0.112	6.5e-07	DLK1, MEG3
E. Kong et al. <sup>6</sup>	rs2237892	0.043	1.15 (OR)	0.24	1.03 (OR)	<b>0.0084</b>	1.30 (OR)	0.34	-0.021	0.14	0.037	<b>0.047</b>	-0.061	0.014	KCNQ1
	rs231362	0.013	1.10 (OR)	0.73	0.98 (OR)	<b>6.2e-05</b>	1.23 (OR)	0.0044	-0.029	0.45	-0.015	<b>4.7e-04</b>	-0.050	0.039	KCNQ1
	rs2334499	0.034	1.08 (OR)	<b>4.7e-10</b>	1.35 (OR)	0.002	0.86 (OR)	0.13	0.011	<b>4.3e-04</b>	0.043	0.17	-0.0184	2.6e-04	HCCA2
	rs4731702	0.039	1.08 (OR)	0.79	0.99 (OR)	0.001	1.17 (OR)	0.011	-0.023	0.061	-0.021	0.063	-0.028	0.99	KLF14
F. Hoggart et al. <sup>7</sup>	rs2471083	<b>9.34E-07</b>	(p-value PoFo effect)					0.039	0.019	0.45	-0.009	<b>1.2e-04</b>	0.055	5.5e-04	KCNK9
	rs3091869	4.7E-06	(p-value PoFo effect)					0.81	-0.002	0.76	-0.003	0.92	-0.0009	0.84	SLC2A10
G. Kerin et al. <sup>36</sup>	rs539515	p-value GxE = 6.5e-12	beta GxE = -0.0117					1.3e-07	0.052	<b>7.3e-09</b>	0.083	0.047	0.028	0.0069	SEC16B
	rs2153960	p-value GxE = 6.5e-9	beta GxE = -0.0098					0.31	-0.009	0.33	-0.014	0.62	-0.006	0.72	FOXO3

For trio-based and genealogy-based PoFo associations (A-E), we considered as replicated associations for which the same parental effect and direction of the effect could be retrieved. For the increased variance method (F) and the GxE interactions (G), we considered associations as being PoFo effect when only one parental scan was significant ( $p < 0.05$ ). Additive betas represent the phenotypic effects of maternally inherited minor alleles. Paternal betas represent the phenotypic effects of paternally inherited minor alleles. Maternal betas represent the phenotypic effects of maternally inherited minor alleles. Differential betas are not shown since they depend only on the parental alleles taken as reference. Genes were mapped by the respective studies. P-values (our study) are computed using BOLT-LMM<sup>35</sup>. Add Additive, Pat Paternal, Mat Maternal, Diff. Differential, P = p-values; B = betas. Bold indicate replicated associations.



**BMI by GxE.** We hypothesized that some GxE signals detected for BMI could be due to PoFO effects. Using the UK Biobank, Kerin et al.<sup>36</sup> reported two GxE associations at rs2153960 and rs539515, mapping to *FOXO3* and *SEC16B*, respectively, with the latter replicated by another study<sup>37</sup>. In our study, we found a paternal effect of rs539515 on BMI (Table 2G, maternal, paternal, differential  $p$ -values = 0.047,  $7.3 \times 10^{-09}$ , 0.0069). Interestingly, when performing a phenome-wide scan of the Iq25.2 locus harboring rs539515, we found paternal associations between four SNPs in high LD (rs527065, rs539515, rs8030 and rs531385;  $r^2 > 0.5$ ) with weight, waist circumference, hip circumference, basal metabolic rate and arm/leg mass (maternal  $p$ -values  $> 0.05$ , paternal  $p$ -values  $< 5 \times 10^{-8}$ , differential  $p$ -values  $< 0.005$ ; Supplementary Table 3). All these SNPs also map to *SEC16B* (either intronic or splicing QTLs) and have already been associated with weight- or obesity-related phenotypes under the additive model<sup>38–40</sup>. Altogether, this suggests that the GxE effect of *SEC16B* on BMI is likely due to a paternal effect.

**Birth weight.** We investigated the 22 PoFO associations reported with birth weight phenotype<sup>41</sup> and were not able to replicate any of these associations in our data, nor any of the additive ones (Supplementary Data 3). This is most likely due to some phenotype misspecification of birth weight in the UK Biobank as this phenotype is self-reported by individuals between 39 and 73 years old, which is likely less reliable than newborn birth weight reported by the mother. Additionally, the individuals with available birth weight specification represent only half of our samples (Supplementary Data 1) which considerably decreases the discovery power.

## Discussion

Studying PoFO effects requires parental genomes or genealogies to determine the set of alleles transmitted to the offspring by each of the two parents. As a consequence, this prevents the study of PoFO effects in biobanks, usually comprising a large and diverse panel of phenotypes. In this work, we propose an approach that leverages the high degree of relatedness between individuals inherent to biobank-scale datasets in order to infer the PoFO of alleles for many individuals and variant sites without any parental genomes or genealogy being available. When applied on the UK Biobank, this approach could predict the PoFO of alleles for around 5% of the total number of samples, resulting in a dataset comprising the PoFO inference for more than 26,000 samples at 7.6 million variants. Together with deep phenotyping, this dataset allows studying PoFO effects on a large scale with improved discovery power, as demonstrated by our ability to replicate many known PoFO associations as well as discover new ones.

We looked at PoFO associations at three different levels. First, we reported two clear PoFO associations supported by additive signals: a maternal effect on platelet phenotypes located in the *MEG3/DLKI* imprinted locus that has already been described<sup>8,9</sup> and another maternal effect on TL at the *TERT* locus, a gene repeatedly associated with TL under an additive model. This new PoFO signal at the *TERT* locus is particularly interesting, not only for its implication in cancer<sup>42</sup>, but also because TL has been found to be highly heritable and proposed to be under imprinting mechanisms<sup>43–47</sup>, which has not yet been confirmed. In this work, we highlight a strong maternal genetic effect at the *TERT* locus, thereby providing additional evidence of the parent-of-origin component in TL heritability and hypothesis on the imprinting status of *TERT*. In addition to this, we also reported 14 new putative PoFO associations across multiple complex traits and confirmed them by replicating the signals in a larger UK Biobank sample set using an increased variance method. These new associations represent interesting candidates of PoFO effects in the human genome and would deserve further investigation and replication in independent datasets. Interestingly, none of

them fall in imprinting regions, suggesting that the current annotation of imprinted genes is still incomplete or that the molecular mechanisms underlying PoFO effects are not necessarily directly linked to genomic imprinting<sup>48</sup>. Finally, we replicated the results of 6 GWAS on PoFO out of the 7 we investigated, confirming PoFO effects on BMI, T2D, standing height and multiple blood biomarkers. We also showed that the summary statistics we provide can be used to annotate additive signals (e.g., *TERT*) or variance QTL (GxE, e.g., *SEC16B*) as PoFO. We believe that an increase of power is still necessary to detect additional PoFO effects with strong confidence but that the current approach already provides a useful resource that can reveal many other associations by meta-analysis. Besides, we also believe that our dataset can be used for more targeted GWAS scans and reveal new putative PoFO effects by focusing only on known imprinted loci, only on additive associations or on both criteria together<sup>6,9</sup>, thereby decreasing the cost of multiple testing corrections.

One of the strengths of our PoFO inference method resides in its ability to make PoFO calls with a low error rate. Regardless, the presence of errors in the inference is unlikely to produce false positive PoFO associations, but only decrease the statistical power of the study, since inference errors are expected to be drawn independently from the phenotypes. Instead, errors are expected to lead to false negatives as PoFO signals get diluted onto the two parental origins and thus decrease association power. In this work, we controlled for this by focusing exclusively on high-confidence PoFO calls, which corresponds to a call for 74.5% of heterozygous genotypes with an estimated error rate below 1%. The overall high accuracy in our estimates could be achieved thanks to recent progress in the statistical estimation of haplotypes for very large sample sizes<sup>15,49</sup> so that the PoFO status inferred within IBD tracks could be confidently propagated to entire chromosomes. Further improvements in phasing algorithms could be made by explicitly modeling IBD sharing between close relatives, eventually through inter- and intra-chromosomal scaffolding as we performed in this work.

Our ability to infer PoFO depends on the availability of close relatives. Surprisingly, even when only a single third-degree relative is available for IBD mapping, we achieve a high call rate and a low error rate. We believe this could be further improved by using more distant relatives, even if they will contribute less to the inference than second- and third-degree relatives. In addition, our PoFO inference depends on the ability to assign parental status to relatives based on IBD sharing on chromosome X, which comes with some flaws. First, our current inference is only possible for males as it leverages chromosome X haploidy, which means that only non-sex specific and male specific PoFO effects can be investigated. As a result, female specific PoFO effects, which could be of great interest given the recent findings on sexual dimorphisms<sup>50</sup>, notably for anthropomorphic traits, are likely missed by this approach. Potential improvements should come with whole genome sequencing (WGS) data: parental status assignment based on rare variant matching on chromosome Y and mitochondrial DNA would likely become possible. In the UK Biobank, this has the potential to substantially increase the sample size above the ~26,000 samples we have so far to a theoretical upper bound of 105,826 samples, which corresponds to the number of samples for which we found groups of close relatives in the dataset. This could further boost the discovery power of downstream PoFO association scans. Second, this approach can be confounded by high levels of inbreeding which could lead a sample to share portions of the chromosome X IBD with close relatives on both sides of the family, therefore greatly complexifying sex assignment. However, we consider this issue to be almost negligible in this study as the UK biobank mostly comprises outbred individuals. Conversely, admixture affects kinship estimation and therefore our ability to find surrogate parents, although this can be compensated

by using a robust method for kinship estimation in admixed populations<sup>51</sup>.

Overall, this study is a valuable resource to further characterize PofO effects and investigate the impact of imprinting genes on complex traits. Although the multiple successive steps of this approach (IBD mapping, phasing, imputation) are difficult to fully automatize, we expect it to be applicable to other biobanks, such as those collected by the FinnGen research project (<https://finngen.gitbook.io/documentation/>), the Million Veteran Program<sup>52</sup> or The Estonian Biobank<sup>53</sup>. Collective efforts would allow the detection of PofO effects with an unprecedented sample size by meta-analyzing PofO effects across multiple biobanks and therefore greatly help future research on the molecular mechanisms leading to PofO effects and their implication for human health.

## Methods

### Duos/Trios identification

To identify trios and duos we used pairwise kinship and IBSO estimates up to third degree relative computed using KING<sup>13</sup> and provided as part of the UK biobank study. Following Manichaikul et al.<sup>13</sup> and Bycroft et al.<sup>12</sup>, we defined offspring-parent pairs as having a kinship coefficient between 0.1767 and 0.3535 and an IBSO below 0.0012 (Supplementary Fig. 5). We also added the condition of age difference greater than 15 years between parent-offspring pairs. We used the age and sex of the individuals to distinguish parents and offspring. For the trios, we also ensured that the two parents have different sex. Starting from 147,731 UKB individuals with at least one third degree relative, we found a total of (i) 1064 samples with both mother and father (i.e., trios) and (ii) 4123 samples with mother or father (i.e., duos). We used the reported ancestry of individuals to keep only genotyped individuals of British and Irish ancestry ( $N=443,993$ ), which resulted in 1037 trios and 3872 duos.

### IBD based group inference

We used pairwise kinship and IBSO estimates up to the third degree relative to identify sibling pairs (kinship between 0.1767 and 0.3535 and IBSO above 0.0012), and second- and third-degree relatives' pairs (kinship below 0.1767) for all genotyped individuals of British and Irish ancestry ( $N=443,993$ ) (Supplementary Fig. 5). For the following steps, we used only second- and third-degree relatives to form surrogate parent groups. We excluded siblings as they share the same two parental genomes and therefore are not informative to distinguish the paternal from the maternal genome. We found 106,414 individuals with at least one second or third degree relative and 21,255 sibling pairs. For individuals with two or more second- and third-degree relatives, we separated those relatives into groups, representing the groups of relatives on each side of the family (i.e., mother-side relatives and father-side relatives). To do so, we used the relatedness in-between these relatives: those related to each other are expected to be on the same side of the family, while those unrelated to each other are expected to be on different sides of the family. We built for each individual a kinship symmetric matrix of size  $N \times N$ , where  $N$  is the number of second-to-third relatives of the target individual considered, filled with the kinship values in-between each relative. We then used the 'igraph' R package to cluster these relatives into groups based on their relatedness similarly to what has been done by Bycroft et al.<sup>12</sup>. As we wanted a maximum of two distinct groups (i.e., one paternal and one maternal), we excluded samples with more than two clusters of relatives from the analysis as it indicates ambiguous cases. Similarly, if a second-to-third degree relative is related to the two clusters, we also excluded the sample to avoid ambiguous cases. Importantly, this is often a symmetric assignment: when A is part of the group of relatives of B, this usually involves B is part of the relatives of A. We identified a total of 105,826 individuals with groups of relatives, ranging from one group of one relative to two groups of

more than two relatives. This includes 309 individuals having also both parents genotyped (i.e., trios) and 1090 having a single parent genotyped in the data (i.e., duos). These 1399 individuals with at least one genotyped parent and groups of close relatives constitute our validation data set on which we applied our PofO inference method using the close relatives as surrogate parents, ignoring the parental genomes. We then used parental genomes to compute the accuracy of our inference.

### Group assignment

We assigned parental status (i.e., mother or father) to groups of close relatives by examining shared IBD segments on chromosome X using XIBD<sup>54</sup>, a software specifically designed to map IBD on chromosome X (Fig. 1c). This assignment was only possible for males as they inherit their only chromosome X copy from their mother: a close relative, male or female, sharing IBD on chromosome X with the target is expected to be from the maternal side of the family. To empirically determine the IBD threshold above which only mother-side relatives are found, we used the 1399 samples of our validation set (i.e., with close relatives' groups and genotyped parents). We computed the IBD sharing on chromosome X for each target-relative pair, knowing the correct parental side of the relatives from the kinship in between the relatives and the available parents. We found that only mother-side relatives share more than 0.1 of IBD1 on chromosome X (Supplementary Fig. 6), a value that we used as a threshold to assign maternal status. Across the 107,038 individuals having groups of close relatives, 48,814 individuals are males, and we assigned the group of close relatives to the maternal side of the family for 20,620 of them. By extension, we propagated the maternal status to the relatives from the same parental group, and we labeled as paternal the relatives from the other group. We then used the underlying idea that siblings share the same set of cousins, uncle, and aunt to enrich our set of samples. We searched for siblings of these 20,620 individuals having the exact same close relatives' groups. We found 864 such siblings, resulting in a total of 21,484 individuals with close relatives' groups assigned to parental status (i.e., surrogate parents). Notably, this strategy allowed us to assign parental status for a small additional subset of female individuals ( $N=775$ , Supplementary Table 1).

### Genotype processing

We used the UK biobank SNP array data in PLINK format. We converted the UK biobank PLINK files into VCF format using PLINK v1.90b5<sup>55</sup>, which resulted in 784,256 variant sites across the autosomes for 488,377 individuals. We then used the UK biobank SNPs QC file (UK biobank resource 1955) to keep only variants used for the phasing of the original UK biobank release, resulting in 670,741 variant sites.

### Validation and production datasets

We assembled two distinct datasets comprising different collections of samples of British or Irish ancestries by subsampling the original dataset with BCFtools v1.8. The first one includes all UK Biobank samples excluding the parental genomes for the  $N=1399$  validation samples for which we have both parental genomes and surrogate parents. We ran our inference on  $N=1399$  validation samples and we assessed its performance by comparing our estimates to the truth given by parental genomes. It is important to note that parental genomes have been used only at the validation stage and not during any phasing runs nor PofO inference. The second dataset includes this time all available UK Biobank samples and has been used to produce the final set of individuals with PofO inference that has been used for association testing. This includes  $N=21,484$  samples for which PofO could be inferred from surrogate parents and  $N=4909$  samples for which PofO could be directly inferred from the

trios/duos. The final dataset includes 22,652 males (85.8%) and 3741 females (14.2%).

### PoFO inference step1: IBD mapping

In this first stage, we inferred PoFO for alleles shared IBD with surrogate parents. To do so, we started by an initial phasing run of the data using SHAPEIT v4.2.1<sup>15</sup> with default parameters so that all data consists of haplotypes. Then, we designed a Hidden Markov Model (HMM)<sup>56</sup> to identify IBD sharing between the target haplotypes and a reference panel mixing haplotypes from two different sources: from the surrogate parents of the target (labeled as mother or father) and from unrelated samples. We aimed for such a probabilistic model for its robustness to phasing and genotyping errors compared to approaches based on exact matching such as the positional Burrows–Wheeler transform (PBWT). The model then uses a forward-backward procedure to compute, for each allele of a target haplotype, the probability of copying the allele from (i) the surrogate mother haplotypes, (ii) the surrogate father haplotypes or (iii) unrelated haplotypes. Here, we used 100 unrelated haplotypes as decoys so that the model is not forced to systematically copy from surrogate parents. When the model copies the target haplotype from a specific surrogate parent at a given locus with high probability, we can therefore infer the PoFO at this locus from the parental group the surrogate parent belongs to. When the model copies from unrelated haplotypes, no inference can be made at the locus (Supplementary Figs. 1, 7 panels 1, 2). We implemented this approach in an open-source tool available on GitHub<sup>57</sup>. As a result of this procedure, we obtained PoFO calls within haplotypes segments shared IBD with surrogate parents.

### PoFO inference step2: extrapolation by phasing

In this second stage, we inferred PoFO for all remaining genotyped alleles. First, we built a haplotype scaffold comprising all alleles for which we know PoFO from IBD sharing with surrogate parents<sup>54</sup>. In other words, we forced all alleles that we knew to be co-inherited from the same ancestor to locate on the same homologous chromosome (Supplementary Figs. 1, 2B). In the scaffolds, we only included IBD tracks longer than 3 cM. We empirically determined this length on the validation set of samples by maximizing and minimizing the call rate and the error rate, respectively (see “Methods” section ‘Accuracy and parameters optimization’). In addition, we considered in the haplotype scaffold only alleles having a PoFO probability greater or equal to 95%. As a result, we could build paternal and maternal haplotype scaffolds that we used in a second step to rephase the entire dataset using SHAPEIT4 v4.2.1<sup>15</sup>. The goal of this second round of phasing is three-folds: (i) to ensure that the pool of alleles coming from the same parent land onto the same haplotype, (ii) to propagate the PoFO assignment from IBD tracks to all alleles along the chromosomes and (ii) to correct long range switch errors. Point (ii) is made possible as all alleles with PoFO unknown (i.e., not in IBD tracks) are phased relatively to the haplotype scaffold so that we can extrapolate their PoFO from the scaffold they co-localize with (paternal/maternal). In practice, we ran SHAPEIT4 with two main options: *-scaffold* to specify the scaffolds of haplotypes to be used in the estimation and *-bingraph* to output the haplotype reconstructions together with phasing uncertainties. The latter provides the haplotype reconstructions as parsimonious graphs encapsulating phasing uncertainty so that likely haplotype pairs can be rapidly sampled without being forced to rerun the complete phasing run. As a consequence, we sampled for each target sample a 1000 haplotype pairs using different seeds and computed the probability for a given allele to be paternal or maternal from its frequency of co-localization across the 1000 pairs onto the paternal and maternal haplotype scaffolds, respectively (Supplementary Figs. 1 and 7A–H panels 3). This frequency indicates the certainty we have in phasing

and therefore is a probabilistic measurement of the confidence in the PoFO assignment. For instance, a specific allele being phased with a certainty of 0.8 onto the paternal haplotype scaffold has an 80% chance to be of paternal origin. In all downstream analysis, we considered only heterozygous genotypes with a phasing probability above 0.7; a threshold that we empirically determined from the validation set of samples by maximizing and minimizing the call rate and the error rate (see “Methods” section on ‘Accuracy and parameters optimization’).

### PoFO inference step3: extrapolation by imputation

In this third stage, we inferred PoFO for untyped alleles, i.e., not included on the SNP array. To do so, we imputed the data using IMPUTE5 v1.1.4<sup>16</sup> with the Haplotype Reference Consortium<sup>17</sup> as a reference panel. As our data is phased with each haplotype assigned to a specific parent, we used the parameter *-out-ap-field* to run a haploid imputation of the data and separately imputed the paternal haplotype and the maternal haplotype. Of note, we filtered out all heterozygous genotypes with a phasing certainty below 0.7 prior to imputation (see previous section). As a result of haploid imputation, the PoFO of imputed alleles can be probabilistically deduced from the imputation dosages: an allele imputed with a dosage of 0.85 on the paternal haplotype has 85% probability of being inherited from the father (i.e., PoFO probability = 85%). Finally, we filtered out variants with an INFO score below 0.8 and obtained a dataset comprising 22,156,064 variants.

### Accuracy and parameters optimization

We used samples with both genotyped parents and groups of surrogate parents (i.e., validation set of samples  $N=1399$ ) to compute the errors in the PoFO inference and to optimize the parameters of our inference method. For the trios ( $N=309$ ) and the duos ( $N=1090$ ), we determined the correct parental origin of offspring heterozygous genotypes at sites where a parent is homozygous, excluding sites with Mendel inconsistencies. We assessed the impact of two parameters on the call rate (percentage of heterozygous genotypes with PoFO assignment) and the error rate (percentage of heterozygous genotypes with incorrect PoFO assignment) of the PoFO inference: (i) the length in centimorgan (cM) of the haplotype segments that we included in the scaffold for the second phasing run and (ii) the phasing certainty threshold we used to assume PoFO to be known at heterozygous genotypes. To do so, we compute the call rate and the error rate for all combinations of the following parameters (Fig. 2a): haplotype segments of 2, 3, 5, 8, and 10 cM and threshold on the phasing certainty between 0.5 and 1.0 by steps of 0.05. Overall, we found that a phasing certainty above 0.7 and haplotype segments above 3 cM to be a good trade-off between call rate and error rate and used these values in all downstream analyses.

### Association testing for PoFO

We tested 99 quantitative phenotypes of the UK biobank data set (Supplementary Data 1) from 4 phenotypic categories to allow genome-wide association analysis of variants of interest: body size measurements, body composition by impedance, blood biochemistry and blood count. We additionally tested telomere length and birth weight which are not included in one of these categories. For telomere length, we removed individuals with reported blood cancer or malignancies. We considered only phenotypes with less than 50% of missing data. We rank-transformed each phenotype using the ‘*rnttransform*’ function from the GenABEL v1.8-OR package<sup>58</sup>. We used the sex, age and the method used to infer the PoFO of alleles as covariates (i.e., surrogate parents or direct parents). We used BOLT-LMM v2.3.4<sup>18</sup> to run all association tests. As recommended by the authors, we performed the model fitting only on the genotyped variants. For the additive GWAS scans, we used the *-dosagefile* parameter to test

imputed alleles dosages, as recommended in the documentation. For the PofO GWAS scans (i.e., maternal scan and paternal scan), we used the `-dosageFile` parameter to test the PofO dosages of alleles. In practice, we only used imputed allele dosages (i) of the paternal haplotype for the paternal-specific GWAS and (ii) of the maternal haplotype for the maternal-specific GWAS, so that PofO assignment uncertainty is propagated to association testing. We conducted a third PofO GWAS scan that compares the effect of maternally and paternally inherited minor alleles at heterozygous genotypes (i.e., differential scan). For this, we used only heterozygous genotypes with imputed minor allele dosages greater or equal to 0.95 to keep only genotypes with high confidence in the PofO. We encoded such alleles as 0 when inherited from the father and 1 when inherited from the mother. We again used the `-dosageFile` parameter to test whether the paternal and maternal alleles have differential effect at heterozygous sites with all homozygous genotypes set to missing. Prior to running association testing, we coded all variants so that we systematically tested the effects of minor alleles. We filtered out all variants with a minor allele frequency (MAF) below 1% which resulted in 7,645,537 variants for association testing.

### GWAS hits identification

We identify independent hits as having Linkage Disequilibrium (LD,  $R^2$ , computed with PLINK v1.90b5<sup>55</sup>) < 0.05 and being located at least 500 kb apart. If two hits are not independent, we select the one with the lowest  $p$ -value. We identify PofO associations as being Bonferroni significant ( $p < 5 \times 10^{-08}$ ) in the differential scan and in the additive scan. We inferred the parent and direction of the effect using the paternal and maternal scans.

### Replication of PofO hits using the increased-variance method

We used the QUICKTEST software<sup>11</sup>, designed to capture PofO effects as described by Hoggart et al. Software and documentation were accessed on 12.22.2021 (<https://wp.unil.ch/sgg/program/quicktest/>). We restricted the analysis to the subset of 443,993 genotyped samples of British or Irish ancestry. We used as covariates age, sex and the first ten PCs.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

### Data availability

The summary statistics for the four GWAS models across the 99 phenotypes are available here for download: <http://poedb.dcsr.unil.ch/>. The UK Biobank genetic data are available under restricted access for privacy policy reason, access can be obtained by application via the UK Biobank Access Management System (<https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access>). Source data are provided with this paper.

### Code availability

Repository [https://github.com/RJHFMSTR/PofO\\_inference](https://github.com/RJHFMSTR/PofO_inference) hosts the source code of the IBD mapper used as part of this study, a full documentation of the pipeline<sup>37</sup>, as well as the custom code used for the analysis and for the data visualization.

### References

- Tucci, V., Isles, A. R., Kelsey, G., Ferguson-Smith, A. C. & Erice Imprinting, G. Genomic imprinting and physiological processes in mammals. *Cell* **176**, 952–965 (2019).
- Lawson, H. A., Cheverud, J. M. & Wolf, J. B. Genomic imprinting and parent-of-origin effects on complex traits. *Nat. Rev. Genet.* **14**, 609–617 (2013).
- Kacem, S. & Feil, R. Chromatin mechanisms in genomic imprinting. *Mamm. Genome* **20**, 544–556 (2009).
- Barlow, D. P. Competition—a common motif for the imprinting mechanism? *EMBO J.* **16**, 6899–6905 (1997).
- Poole, R. L. et al. Beckwith-Wiedemann syndrome caused by maternally inherited mutation of an OCT-binding motif in the IGF2/H19-imprinting control region, ICR1. *Eur. J. Hum. Genet.* **20**, 240–243 (2012).
- Kong, A. et al. Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–874 (2009).
- Perry, J. R. et al. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92–97 (2014).
- Zink, F. et al. Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. *Nat. Genet.* **50**, 1542–1552 (2018).
- Kim, H. I. et al. Genome-wide survey of parent-of-origin-specific associations across clinical traits derived from electronic health records. *HGG Adv.* **2**, 100039 (2021).
- Horikoshi, M. et al. Genome-wide associations for birth weight and correlations with adult disease. *Nature* **538**, 248–252 (2016).
- Hoggart, C. J. et al. Novel approach identifies SNPs in SLC2A10 and KCNK9 with evidence for parent-of-origin effect on body mass index. *PLoS Genet.* **10**, e1004508 (2014).
- Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
- Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
- Delaneau, O. & Marchini, J., Genomes Project, C. & Genomes Project, C. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat. Commun.* **5**, 3934 (2014).
- Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
- Rubinacci, S., Delaneau, O. & Marchini, J. Genotype imputation using the positional Burrows Wheeler transform. *PLoS Genet.* **16**, e1009049 (2020).
- McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
- Loh, P. R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
- Sanli, I. et al. Meg3 non-coding RNA expression controls imprinting by preventing transcriptional upregulation in cis. *Cell Rep.* **23**, 337–348 (2018).
- Vosa, U. et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
- Leao, R. et al. Mechanisms of human telomerase reverse transcriptase (hTERT) regulation: clinical impacts in cancer. *J. Biomed. Sci.* **25**, 22 (2018).
- Bhat, G. R. et al. Association of newly identified genetic variant rs2853677 of TERT with non-small cell lung cancer and leukemia in population of Jammu and Kashmir, India. *BMC Cancer* **19**, 493 (2019).
- Brandes, N., Linal, N. & Linal, M. Genetic association studies of alterations in protein function expose recessive effects on cancer predisposition. *Sci. Rep.* **11**, 14901 (2021).
- Vuckovic, D. et al. The polygenic and monogenic basis of blood traits and diseases. *Cell* **182**, 1214–1231 e11 (2020).
- McCartney, D. L. et al. Genome-wide association studies identify 137 genetic loci for DNA methylation biomarkers of aging. *Genome Biol.* **22**, 194 (2021).
- Codd, V. et al. Polygenic basis and biomedical consequences of telomere length variation. *Nat. Genet.* **53**, 1425–1433 (2021).

27. Margraf, A. et al. ArhGAP15, a RacGAP, acts as a temporal signaling regulator of Mac-1 affinity in sterile inflammation. *J. Immunol.* **205**, 1365–1375 (2020).
28. Persson, H. et al. Transcriptome analysis of controlled and therapy-resistant childhood asthma reveals distinct gene expression profiles. *J. Allergy Clin. Immunol.* **136**, 638–648 (2015).
29. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays, and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
30. Zoledziewska, M. et al. Height-reducing variants and selection for short stature in Sardinia. *Nat. Genet.* **47**, 1352+ (2015).
31. Benonisdotir, S. et al. Epigenetic and genetic components of height regulation. *Nat. Commun.* **7**, 13490 (2016).
32. Granot-Hershkovitz, E. et al. Searching for parent-of-origin effects on cardiometabolic traits in imprinted genomic regions. *Eur. J. Hum. Genet.* **28**, 646–655 (2020).
33. Smith, A. C., Choufani, S., Ferreira, J. C. & Weksberg, R. Growth regulation, imprinted genes, and chromosome 11p15.5. *Pediatr. Res.* **61**, 43R–47R (2007).
34. Bentley, L. et al. The imprinted region on human chromosome 7q32 extends to the carboxypeptidase A gene cluster: An imprinted candidate for Silver-Russell syndrome. *J. Med. Genet.* **40**, 249–256 (2003).
35. Carrera, I. A. et al. Microdeletions of the 7q32.2 imprinted region are associated with Silver-Russell syndrome features. *Am. J. Med. Genet. A* **170**, 743–749 (2016).
36. Kerin, M. & Marchini, J. Inferring gene-by-environment interactions with a Bayesian whole-genome regression model. *Am. J. Hum. Genet.* **107**, 698–713 (2020).
37. Marderstein, A. R. et al. Leveraging phenotypic variability to identify genetic interactions in human phenotypes. *Am. J. Hum. Genet.* **108**, 49–67 (2021).
38. Hotta, K. et al. Association between obesity and polymorphisms in SEC16B, TMEM18, GNPDA2, BDNF, FAIM2 and MC4R in a Japanese population. *J. Hum. Genet.* **54**, 727–731 (2009).
39. Graff, M. et al. Genome-wide analysis of BMI in adolescents and young adults reveals additional insight into the effects of genetic loci over the life course. *Hum. Mol. Genet.* **22**, 3597–3607 (2013).
40. Sahibdeen, V. et al. Genetic variants in SEC16B are associated with body composition in black South Africans. *Nutr. Diabetes* **8**, 43 (2018).
41. Juliusdotir, T. et al. Distinction between the effects of parental and fetal genomes on fetal growth. *Nat. Genet.* **53**, 1135–1142 (2021).
42. Yuan, X., Larsson, C. & Xu, D. Mechanisms underlying the activation of TERT transcription and telomerase activity in human cancer: Old actors and new players. *Oncogene* **38**, 6172–6183 (2019).
43. Buxton, J. L. et al. Human leukocyte telomere length is associated with DNA methylation levels in multiple subtelomeric and imprinted loci. *Sci. Rep.* **4**, 4954 (2014).
44. Nordfjall, K., Larefalk, A., Lindgren, P., Holmberg, D. & Roos, G. Telomere length and heredity: Indications of paternal inheritance. *Proc. Natl Acad. Sci. USA* **102**, 16374–16378 (2005).
45. Weng, Q. et al. The known genetic loci for telomere length may be involved in the modification of telomeres length after birth. *Sci. Rep.* **6**, 38729 (2016).
46. Prescott, J. et al. Genome-wide association study of relative telomere length. *PLoS One* **6**, e19635 (2011).
47. Barrett, E. L. & Richardson, D. S. Sex differences in telomeres and lifespan. *Aging Cell* **10**, 913–921 (2011).
48. Guilmatre, A. & Sharp, A. J. Parent of origin effects. *Clin. Genet.* **81**, 201–209 (2012).
49. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).
50. Pulit, S. L. et al. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum. Mol. Genet.* **28**, 166–174 (2019).
51. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* **98**, 127–148 (2016).
52. Gaziano, J. M. et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
53. Leitsalu, L. et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).
54. Henden, L., Wakeham, D. & Bahlo, M. XIBD: Software for inferring pairwise identity by descent on the X chromosome. *Bioinformatics* **32**, 2389–2391 (2016).
55. Purcell, S. et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
56. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
57. Hofmeister, R. J. et al. Parent-of-Origin Inference for Biobanks. GitHub, <https://doi.org/10.5281/zenodo.7085471> (2022).
58. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: An R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).

## Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 66995 and funded by the Swiss National Science Foundation (SNSF) project grant 373 (PPOOP3\_176977). We thank Jonathan Marchini for highlighting the GxE interaction of the SEC16B locus and Chiara Auwerx for the discussion on the telomere length phenotype.

## Author contributions

R.J.H. and O.D. designed the study and wrote the paper. R.J.H. performed experiments. R.J.H. and O.D. developed the IBD mapping algorithm. R.J.H. and S.R. performed the imputation. R.J.H. and D.M.R. interpreted the biological relevance of the results. R.J.H., O.D., and Z.K. designed the GWAS models. This study was initiated after discussions between A.B. and O.D. The project has been supervised by O.D. All authors reviewed the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-34383-6>.

**Correspondence** and requests for materials should be addressed to Olivier Delaneau.

**Peer review information** *Nature Communications* thanks Wei-Min Chen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022



## **Appendix D**

### **A genome-wide scan for parental inheritance distortion events to identify genetic effects on human fertility**

This manuscript is presented in [Chapter II](#).

It contains unpublished preliminary results.





# A genome-wide scan for parental inheritance distortion events to identify genetic effects on human fertility

Robin J. Hofmeister, Olivier Delaneau

## Abstract

Fertility measurement is challenging due to its complex nature, which is influenced by a range of biological processes, such as hormonal regulation, gametogenesis, fertilization, and implantation. Consequently, researchers often use proxies to assess genetic factors involved in human fertility, such as the number of children a person has or their age at first birth, to measure fertility indirectly. However, these proxies can be influenced by factors like ancestry, socioeconomic status, or lifestyle choices, which can confound the results. Furthermore, these methods may not be suitable for detecting genetic factors associated with fertility in diverse populations. Recently, transmission distortion tests (TDTs) have been proposed as an alternative to study genetic factors linked to human fertility, by assessing the frequency of allele transmission from parents to offsprings. However, this method is limited to family studies as it requires prior knowledge of parental genotypes, and the sample size is therefore insufficient to detect moderate distortions of transmission, that are those to likely persist in the population. Here, we propose an innovative approach that investigates genetic contribution to human fertility by detecting variants whose parental inheritance deviates from the expected Mendelian ratio that does not rely on the availability of parental genomes. We identified a strong paternal distortion signal at 22q13.33 whose associated genes *RABL2* and *ACR* impact sperm function.

## Introduction

Mendel's Law of segregation implies that the offspring of an heterozygous parent has an equal probability of inheriting either allele. Deviations from the expected Mendelian inheritance pattern, which occurs when one allele is preferentially transmitted, is termed Transmission Ratio Distortion (TRD)<sup>1,2</sup>. TRD can arise due to various selective stages that occur during different biological processes or developmental stages. Meiotic drive, for example, refers to the phenomenon in which "driving alleles" influence the meiotic process to increase their transmission, leading to a deviation from the expected inheritance pattern. In the asymmetric female meiosis, only one of the four haploid products becomes an oocyte, which leaves room for gamete competition and selection before the fertilization. On the other hand, male gametogenesis produces many small gametes that have to compete for fertilization and during which selection can occur. Finally, the viability of individuals can also play a role in shaping TRD in the human genome. An allele that confers a survival advantage, for example at the zygote stage, will be more frequently represented in the population, leading to deviations from the typical Mendelian inheritance pattern<sup>1,2</sup>.

The most documented events of TRD are 'gamete killers'. In mice, the t-haplotype confers a fertilization advantage which results in increasing its transmission and hence its frequency in the population<sup>3</sup>. In drosophila, segregation distorter (SD) locus prevents wild-type gametes (i.e not carrying the distorter form) from developing normally, resulting in skewed transmission in favor of the SD form<sup>4,5</sup>. Despite the presence of several documented instances of TRD in other organisms, the extent and impact of TRD in humans remains largely unknown. Yet, it is likely that distorter variants exist across the human genome. According to speculative reports, between 50-75% of all human conceptions are lost before the first missed menstrual period, and infertility affects one in every couple trying to conceive<sup>6</sup>. These factors suggest that the influence of distorter variants on human reproduction and the deviation from expected Mendelian inheritance patterns may be more prevalent than previously thought.

Attempts have been made to understand the extent of TRD in human populations using different strategies. First, studies investigated the excess of allele sharing across siblings and

twins. In one study using 143 nuclear families of Hutterite ancestry, a genome-wide excess of allele sharing among siblings was found, which could indicate a departure from the expected Mendelian inheritance pattern at many loci<sup>7</sup>. However, this evidence was contradicted by a study on dizygotic twins from Australia and the Netherlands that found no excess of allele sharing, either across the entire genome or at the HLA locus, the human ortholog of the mouse t-haplotype<sup>8</sup>. This is also in contrast to previous findings that showed a higher degree of HLA haplotype sharing in dizygotic twins<sup>9</sup>. These conflicting results highlight the need for further research and alternative strategies to better comprehend the occurrence and impact of TRD in different populations.

Second, deviations from Mendel's pattern of inheritance have been identified using the Transmission Disequilibrium Test (TDT), which measures the non-random transmission of an allele from heterozygous parents to their affected offspring. It is a widely used approach in family-based studies to examine the possible connection between a genetic marker and a particular illness. This notably allows the discovery of distortion events associated with diseases such as Crohn's disease<sup>9,10</sup> or the long-QT syndrome<sup>11</sup>. Finally, the TDT has been expanded at the population level by considering each offspring as "affected"<sup>12-14</sup>. This allowed the study of TRD without restricting analyses to disease cases and considerably increased the study sample size by including any available family in the analysis. It is important to note that by evaluating TRD without considering a specific phenotype, it allows to identify genetic loci that affect the likelihood of survival. This is similar to selecting "alive" as the phenotype, as it highlights genetic factors that impact the chances of survival.

Several studies have used this generalized TDT approach to investigate loci affecting reproduction and survival. Hanchard *et al.* analyzed MHC regions across 380 newborns and found modest evidence for TRD in the *CLIC1* gene ( $p=0.025$ ) after restricting the test to 13 SNPs based on LD<sup>6</sup>. Santos *et al.* focused on the human region syntenic with the mouse t-haplotype, containing notably the human MHC region<sup>13</sup>. They adjusted for multiple testing by using tag SNPs and permutations. They observed a significant deviation ( $p=2e^{-04}$ ) in the allelic transmission among 30 CEU male parents with a strong ratio: 17 of the 18 transmission were for the same allele. Patterson *et al.* assessed TRD genome-wide by combining both parents from the Framingham Heart Study (FHS) and found eight candidate

distortion events, one of which reach genome-wide significance ( $p=7.4e^{-10}$ ). Meyer *et al.* used human pedigree datasets and revealed two significant loci among 90 individuals of the Autism Genetic Resource Exchange (AGRE) dataset: the first at 10q26.12 when combining both paternal and maternal transmission ( $p=4.55e^{-08}$ ); the second at 6p21.1 that only showed a significant signal for paternal transmission ( $p=1.77e^{-05}$ )<sup>14</sup>. Although the signal at 6p21.1 confirms the previous observation in 30 CEU males<sup>13</sup>, these signals could not be replicated in the FHS cohort, which likely suffers from genotyping error noise<sup>14,15</sup>. Finally, Liu *et al.* searched for paternal- or maternal-specific TRD events in the FHS cohort, correcting for genotyping errors<sup>15</sup>. They found two maternal-specific loci, at the *LRP2* and *ZNF133* genes ( $p=4.2e^{-08}$  and  $2.6e^{-08}$ , respectively). Despite setting the basis on the extent of TRD genome-wide, only one of these signals seemed to be replicated across two independent cohorts (i.e at 6p21.1), the remaining still requiring further validation as suggested<sup>15</sup>.

Although the generalized TDT appears to be the most promising method for detecting TDR in healthy populations, it also has several limitations and challenges that need to be considered. These include issues with sample size, population stratification, and technical factors such as genotyping errors. Indeed, the TDT method is limited to family studies as it relies on prior knowledge of the parental genotypes. This restriction limits the sample size and reduces the potential for discovery, resulting in an ability to detect only strong cases of TDR. However, such strong TDR distortions are not stable in a population, as the dominant allele becomes fixed more quickly as the distortion becomes stronger<sup>14,15</sup>. As a result, it is unlikely that strong TRD will be observed in multiple populations simultaneously and the replication of TRD signals becomes challenging<sup>15</sup>. On the other hand, moderate distortions are more challenging to detect since they require extremely large sample sizes. For example, considering an heterozygote frequency of 10%, approximately 20,000 parent-offspring trios are necessary to achieve 80% power to detect distortion of 5% at  $\alpha=10^{-7}$  according to simulations<sup>16</sup>. Finally, previous studies have highlighted the possibility of genotyping error to introduce false positive TRD detection<sup>14,15,17</sup>. This is often indicated by an excess of genome-wide low p-values and lack of consistency between variants in close proximity, while a true TRD signal is expected to spread to neighboring variants due to linkage disequilibrium<sup>14</sup>. Due to these limitations of the TDT, the detection of transmission distortion events in healthy populations remains a significant challenge.

To address this, we present an innovative method for analyzing distortion events that modifies the conventional use of TDT. Our approach consists of two key components. First, instead of examining the transmission of minor and major alleles from parents to offspring, we propose investigating the frequency at which minor alleles are inherited by the offspring from each parent. This automatically takes into account the differential impact of the variant on the paternal and maternal reproductive processes or gametic competition and enables the detection of parent-specific variants. Second, in order to enhance the sample size of our study, we have utilized a recently developed method that infers the parent-of-origin of alleles from close relatives, eliminating the need for parental genotypes<sup>18</sup>. This notably allows us to increase the sample size compared to family-based studies, resulting in a more robust and effective analysis of transmission distortion events in healthy populations.

In the UK Biobank whole-genome sequencing (WGS) dataset, we inferred the PofO for 10,150 samples and tested the resulting call set for Parental Inheritance Distortion (PID) event. We identified a strong paternal distortion signal at 22q13.33 that could be technically validated in the UK Biobank whole-exome sequencing (WES) data and whose associated genes impact sperm function. Beside improving upon the traditional TDT method, our results demonstrate a more reliable way to test for genetic factors involved in human fertility compared to GWAS studies that use proxy phenotypes<sup>19</sup>.

## Results

### Genome-wide Parental Inheritance Distortion scan

We identified 10,150 individuals from the UK Biobank WGS data who had surrogate parents labeled as either paternal or maternal, representing 38% of the original sample size reported using the UK Biobank axion array data<sup>18</sup>. We inferred the parent-of-origin at common variants (MAF>0.1%, see methods) for these individuals, resulting in a call set of 16,449,701 variant sites across the 22 autosomes.

To detect variants that deviated from the expected Mendelian inheritance pattern of 50%, we performed a Parental Inheritance Distortion (PID) test (see methods) on the resulting call set. Using a Bonferroni genome-wide significance threshold ( $p < 5e^{-08}$ ), we identified 9 genome-wide significant loci (**Figure 1A**). Among them, 6 showed signals extending to nearby variants due to linkage disequilibrium (chromosomes 3,11,12,14,19,22), similar to what is observed in GWAS signals which suggests that they are true positives<sup>14</sup>, while the other 3 were isolated variants that are likely to be false positives (chromosomes 4,5,6).

We conducted a comparison of allelic frequency between males and females in the UK Biobank WGS cohort to eliminate the possibility of sex-specific variants causing the distortions of inheritance. Indeed, genetic variants over-represented in one gender may lead to unequal transmission between paternal and maternal alleles. The UK Biobank WGS cohort consists of 67,290 and 82,651 genetically confirmed males and females, respectively. We did not observe any significant discrepancies between the allelic frequencies of males and females at loci exhibiting genome-wide significant distortions (**Supplemental Figure 1**). This implies that the observed distortion signals are not driven by imbalance allele frequencies between male and females. [However, no statistical measure of the difference between men and women has yet been calculated.]\*(*remains to be investigated*)

To validate the signals, we employed two different strategies. Firstly, since the UK Biobank project includes multiple releases (WGS, WES, and SNP array) for the same set of individuals, we tested the distortion events across different genotyping technologies for the same cohort. This serves can be considered as technical replicates, and can provide insights into the true nature of the signals, helping to distinguish between genuine distortion events and

errors arising from sequencing, mapping, or genotype calling<sup>14,15,17</sup>. Secondly, we aim at validating these signals in other cohorts, which can be considered as biological replicates. This approach helps to confirm the robustness and generalizability of the findings beyond the initial cohort. [However, no biological validation was yet successful]\*(*remains to be investigated*)

### Technical replication

Using the UK Biobank WES data, we inferred the parent-of-origin (PofO) for 26,393 individuals for which we performed genotype imputation using the UK Biobank WGS as a reference panel (see methods). We selected all variants exhibiting genome-wide significant distortion ( $p < 5e^{-08}$ ) in the WGS call set, which represents 144 non-independent variants. Out of these, only 92 could be tested in the WES call set due to the filtering out of poorly imputed variants. Interestingly, these 92 variants are located on chromosomes 3, 11, 14, and 22, which are the signals more likely to show true positive signals due to the spreading of significant distortions to nearby variants (**Figure 1A**). This could already indicate that the remaining poorly imputed variants are false positive signals. We tested these 92 candidate variants for PID and found that only the locus on chromosome 22 also shows genome-wide significance ( $p < 5e^{-08}$ ) in the WES call set (**Figure 1B**).

The SNP leading the signal on chromosome 22 (rs2747986,  $p_{\text{wgs}} = 4.1e^{-47}$ ) is  $\sim 1.5$  times less inherited from fathers (ratio=0.39, count<sub>pat</sub>=1829, count<sub>mat</sub>=2808). This variant is located in an intron of *RABL2* (**Figure 1C**), a gene that plays a role in sperm tail structure and has been implicated in male fertility<sup>20,21</sup>. In addition, among the signal confirmed in the WES call set ( $p_{\text{wgs}} < 5e^{-08}$  and  $p_{\text{wes}} < 5e^{-08}$ ), the lead SNP is rs199928666 ( $p_{\text{wgs}} = 1.6e^{-39}$ ,  $p_{\text{wes}} = 3.3e^{-10}$ ), a variant in strong LD with rs2747986 ( $r^2 = 0.42$ )<sup>22</sup> and that exhibits a similar distortion ratio (ratio=0.43). This variant is a splice-eQTL in testis for *ACR*<sup>23</sup>, a gene that encodes the acrosin protein, the main protease of the acrosome, which plays a role in penetrating the zona pellucida. A decrease in acrosin protein levels has been linked to delayed fertilization<sup>24,25</sup>. These findings suggest that sperm cells carrying the risk allele are less efficient in the fertilization process than wild-type sperm cells, making them less likely to be inherited. However, since these



alterations are unlikely to completely impair sperm cells, fathers who are homozygous for the risk allele likely exhibit reduced fertility rather than complete infertility.

The reason why we only observed the signal on chromosome 22 in the WES call set could be attributed to the presence of genotyped variants that exhibit strong distortion at this locus, leading to improved imputation accuracy of neighboring variants. As the PID test is sensitive to genotyping error, it may also be sensitive to imputation error, which makes it challenging to detect signals in imputed call sets. [However, this needs to be validated for other significant regions identified in the WGS dataset.]\*(*remains to be investigated*). Additionally, it is possible that all the signals, except for the one on chromosome 22, are a result of sequencing or genotype calling artifacts.

We finally evaluated the efficacy of traditional family-based studies in detecting signals at this locus using a subset of individuals (N=518) with parental genome data available for PofO inference in the UK Biobank WGS data. We did not observe any distortion for the locus on chromosome 22, highlighting the advantage of our call set over the conventional TDT approach that uses family data (**Figure 2A**).

### Biological replication

Next, we aimed to validate the signals by replicating them in independent datasets. To achieve this, we utilized the Estonian Biobank (EBB)<sup>26</sup>, in which we identified 29,650 parent-offspring duos and 10,502 parent-offspring trios. For these, we inferred the PofO using available parental genomes. We tested the resulting call set for PID, and we could not replicate the signal on chromosome 22 (**Figure 2F**) nor any of the other genome-wide significant signals detected on the UK Biobank WGS call set.

We discovered that the genotyped variants in the EBB were not dense enough to impute the locus accurately on chromosome 22 (**Figure 2F**). As a comparison, we were unable to recover the signal on chromosome 22 using the UK Biobank SNP array data (**Figure 2G**), which contains a similar density of variants as the EBB SNP array data. To explore further, we evaluated whether the recovery of the signal on chromosome 22 was due to the genotyped data or the reference panel. For this, we used different reference panels for both the UK Biobank SNP array imputation (**Figure 2G-I**) and for the UK Biobank WES imputation (**Figure**

**2C-E).** None of the SNP array imputed call sets allowed us to retrieve the signal. On the other hand, all the UK Biobank WES imputed call sets retrieved the signal. This confirms the limitation of SNP array based imputed call sets and highlights the benefit of using WGS and WES data in this context.

## Future analysis

Since SNP array data seems to not allow recovering signals located close to telomeres, we aimed to validate our findings using WGS cohorts. For this, we initiated a collaboration with the Qatar Genome Project<sup>27</sup>. A first overview of the data allowed us to identify a total of 2359 parent-offspring duos and 691 parent-offspring trios across a total of 13,896 individuals. In addition, we also aim to use the publicly available 1000 Genome Project<sup>28</sup>, which includes a total of 602 parent-offspring trios.

Furthermore, given that the signals identified are predominantly situated near telomeres, we do not exclude the possibility that they result from phasing edge effect or parent-of-origin inference edge effects. Nevertheless, we do not perceive any scenario in which phasing errors are associated with the parental origin of variants, resulting in the identification of these signals. In addition, we will also investigate the mappability of the regions. Low mappability can result in poor phasing accuracy.

## Method

### Parent-of-origin inference from close relatives

To infer the parent-of-origin from close relatives in the UK Biobank data set, we used a method previously developed as part of our research group<sup>18</sup>. Briefly, it consists in (i) the identification of close relatives using the kinship estimate, (ii) the grouping of close relatives into parental groups, (iii) the labeling of parental groups as paternal or maternal using the IBD sharing on the chromosome X for male individuals, and (iv) phasing to assign parental origin to haplotypes.

### Parent-of-origin inference from parental genomes

To infer the parent-of-origin from available parental genomes, we used the phasing software SHAPEIT5<sup>29</sup>, which includes an option `--pedigree` implementing a ‘Mendelian’ phasing. This option uses parental genomes to solve the phase for heterozygous offspring. When parental genomes can not be used, such as in the case where both parents are heterozygous, it solves the phase from the reference panel using the typical phasing model.

### Parental inheritance distortion test

To test the deviation from Mendelian inheritance pattern, we used the function `binom.test(P, P+M, 0.5)` implemented in R, where:

P=number of heterozygous individuals with paternally inherited minor allele,

M=number of heterozygous individuals with maternally inherited minor allele,

0.5 = expected ratio based on Mendelian rules.

Additionally, we computed the ratio of paternally versus maternally inherited alleles computed as  $r = P/(P+M)$ . This ratio equals 0.5 when there is no distortion.

### UK Biobank array data processing

In the UK Biobank SNP-array data, we inferred the parent-of-origin from close relatives as described in *Hofmeister et al*<sup>18</sup>. We followed the same quality control procedure. It resulted in 26,393 individuals with PofO inferred. For the SNP-array data, this method is followed by haploid imputation to increase the variant density on each of the two parental haplotype separately.

### UK Biobank sequencing data processing

We use the 150,119 individuals available WGS data and the 452,644 individuals with both WES and SNP array data. As described in the SHAPEIT5 documentation<sup>29</sup>, we merged the WES with the SNP array data to increase the variant density and improve the accuracy of the phasing procedure. For both the WGS and the WES data, we followed the same quality control procedure as in the SHAPEIT5 manuscript.

To infer the parent-of-origin in the WGS and WES data, we proceed in a multi-step process. First we inferred the parent-of-origin from close relatives in the SNP-array data (see above). Then, we used the resulting haplotypes as a scaffold to phase the sequencing data using the option *--scaffold* of the SHAPEIT5\_phase\_common tool.

For the WES data, we additionally increased the variant density using haploid imputation.

### Estonian Biobank data processing

We used the SNP array data of the Estonian Biobank (EBB)<sup>26</sup> pre-QCed as provided by the official release. We use the software KING<sup>30</sup> to compute the relatedness among individuals. We identified parent-offspring duos and trios as relationships as having a kinship coefficient lower than 0.3553 and greater than 0.1767 and an IBS0 lower than 0.0012<sup>30,31</sup>. In addition, we require that the difference in age between parents and offspring is greater than 15 years and that the two parents have different sex for trios. This resulted in the identification of 10,502 trios and 29,650 duos. We inferred the PofO from parental genomes using SHAPEIT5 and the *--pedigree* option<sup>29</sup>. In addition, we performed haploid imputation using HRC as a reference panel<sup>32</sup>.

### 1000GP data processing

We used the publicly available 1000 Genome Project 30x GRCh38 data<sup>28</sup>. We identified trios and duos using the provided pedigree file. It consists of 602 trios and 6 duos. For these, we inferred the PofO from parental genomes using SHAPEIT5 and the *--pedigree* option<sup>29</sup>.

### Haploid Imputation

Haploid imputation has been performed using the option *-out-ap-field* of IMPUTE5<sup>33</sup> such as described in *Hofmeister et al.* For this procedure, we used different reference panels in order to assess their impact on the signal of imputed call sets:

- Haplotype Reference Consortium (HRC)<sup>32</sup> GRCh37
- 1000 Genome Project 30x GRCh38, publicly available
- UK Biobank WGS GRCh38, produced as part of the SHAPEIT5 manuscript<sup>29</sup>.

Each imputed call set has been filtered to remove variant sites with INFOscore below 0.8.

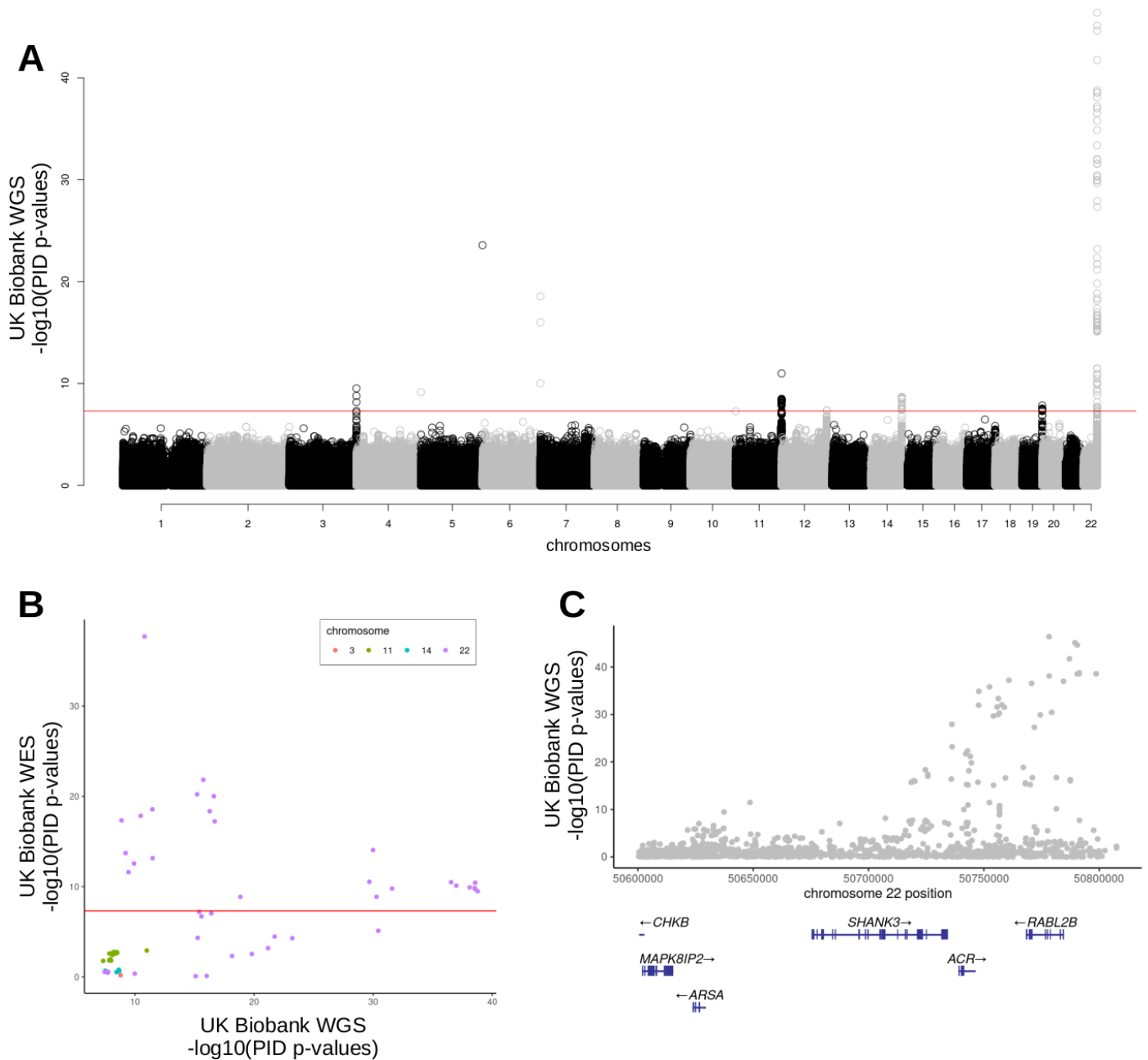
### LiftOver

For imputation purposes, we lifted over the UK Biobank SNP array data and the HRC data from GRCh37 to GRCh38 using a vcf liftover tool available as part of the SHAPEIT5 release<sup>29</sup>.

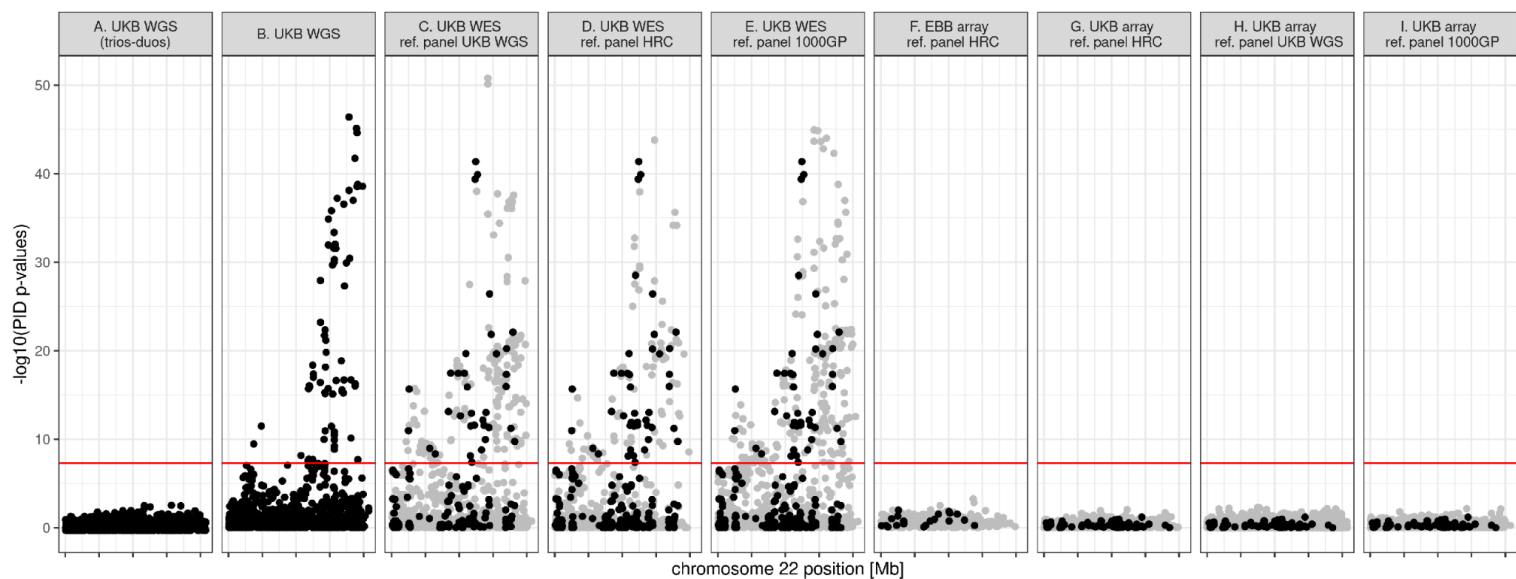
### Qatar Genome Project data processing

We use the software KING<sup>30</sup> to compute the relatedness among individuals. We identified parent-offspring duos and trios as relationships as having a kinship coefficient lower than 0.3553 and greater than 0.1767 and an IBS0 lower than 0.0012<sup>30,31</sup>. In addition, we require that the difference in age between parents and offspring is greater than 15 years and that the two parents have different sex for trios. It resulted in 691 trios and 2359 duos.

## Figures



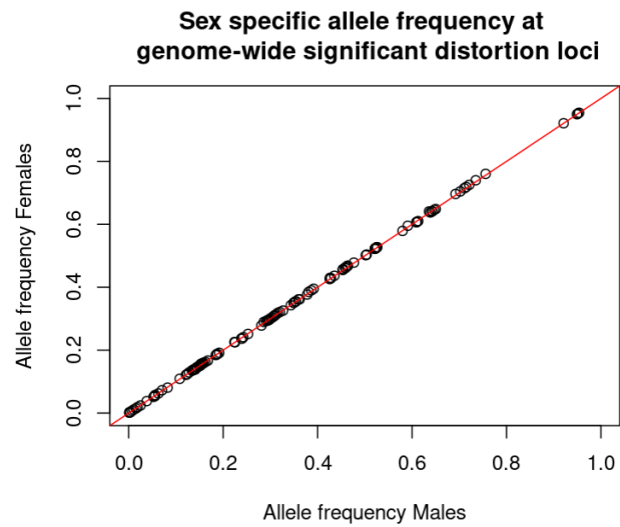
**Figure 1. Parental Inheritance Distortion scan. A)** PID test significance (y-axis,  $-\log_{10}(\text{p-value})$ ) across the 22 autosomes (x-axis). Even chromosomes are shown in black; odd chromosomes are shown in gray. The red line indicates the Bonferroni genome-wide significance threshold ( $5e^{-08}$ ). **B)** PID significance in the UK Biobank WES data (y-axis) versus the UK Biobank WGS data for variants exhibiting PID genome-wide significance in the WGS data. **C)** Locus zoom at 22q13.33 on the UK Biobank WGS PID scan.



**Figure 2. Locus zoom at 22q13.33. A-I)** PID significance (y-axis) along chromosome 22 positions (x-axis) across different call sets (A-I). **A)** UK Biobank WGS using only duos and trios for which the PofO was inferred from parental genomes (N=518). **B)** UK Biobank WGS using N=10,150 individuals for which the PofO was inferred from close relatives. **C-E)** UK Biobank WES using N=26,393 individuals for which the PofO was inferred from close relatives across different reference panels for genotype imputation. **F)** Estonian Biobank SNP array imputed with HRC as a reference panel, using only duos and trios for which the PofO was inferred from parental genomes (N=40,152). **G-I)** UK Biobank SNP array using N=26,393 individuals for which the PofO was inferred from close relatives across different reference panels for genotype imputation. **Black** dots indicate genotyped variants. **Gray** dots indicate imputed variants.



## Supplementary figures



**Supplementary figure 1.** Allelic frequency at PID genome-wide significant loci. Males allele frequencies (x-axis) versus female allele frequencies (y-axis) in the UK Biobank WGS cohort.

## References

1. Friocourt, G. *et al.* Bypassing Mendel's First Law: Transmission Ratio Distortion in Mammals. *Int. J. Mol. Sci.* **24**, (2023).
2. Huang, L. O., Labbe, A. & Infante-Rivard, C. Transmission ratio distortion: review of concept and implications for genetic association studies. *Hum. Genet.* **132**, 245–263 (2013).
3. Olds-Clarke, P. Models for male infertility: the t haplotypes. *Rev. Reprod.* **2**, 157–164 (1997).
4. Larracuente, A. M. & Presgraves, D. C. The selfish Segregation Distorter gene complex of *Drosophila melanogaster*. *Genetics* **192**, 33–53 (2012).
5. Kozielska, M., Weissing, F. J., Beukeboom, L. W. & Pen, I. Segregation distortion and the evolution of sex-determining mechanisms. *Heredity* **104**, 100–112 (2009).
6. Hanchard, N. *et al.* An investigation of transmission ratio distortion in the central region of the human MHC. *Genes Immun.* **7**, 51–58 (2006).
7. Zöllner, S. *et al.* Evidence for extensive transmission distortion in the human genome. *Am. J. Hum. Genet.* **74**, 62–72 (2004).
8. Montgomery, G. W. *et al.* HLA and genomewide allele sharing in dizygotic twins. *Am. J. Hum. Genet.* **79**, 1052–1058 (2006).
9. Jawaheer, D., MacGregor, A. J., Gregersen, P. K., Silman, A. J. & Ollier, W. E. Unexpected HLA haplotype sharing in dizygotic twin pairs discordant for rheumatoid arthritis. *J. Med. Genet.* **33**, 1015–1018 (1996).
10. Liu, L. Y., Schaub, M. A., Sirota, M. & Butte, A. J. Transmission distortion in Crohn's disease risk gene ATG16L1 leads to sex difference in disease association. *Inflamm. Bowel Dis.* **18**, 312–322 (2012).
11. Imboden, M. *et al.* Female predominance and transmission distortion in the long-QT syndrome. *N. Engl. J. Med.* **355**, 2744–2751 (2006).
12. Naumova, A. K., Greenwood, C. M. & Morgan, K. Imprinting and deviation from Mendelian transmission ratios. *Genome* **44**, 311–320 (2001).
13. Santos, P. S. C. *et al.* Assessment of transmission distortion on chromosome 6p in healthy individuals using tagSNPs. *Eur. J. Hum. Genet.* **17**, 1182–1189 (2009).
14. Meyer, W. K. *et al.* Evaluating the evidence for transmission distortion in human pedigrees. *Genetics* **191**, 215–232 (2012).
15. Liu, Y. *et al.* Identification of two maternal transmission ratio distortion loci in pedigrees of the Framingham heart study. *Sci. Rep.* **3**, 2147 (2013).
16. Evans, D. M., Morris, A. P., Cardon, L. R. & Sham, P. C. A note on the power to detect transmission distortion in parent-child trios via the transmission disequilibrium test. *Behav. Genet.* **36**, 947–950 (2006).
17. Mitchell, A. A., Cutler, D. J. & Chakravarti, A. Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *Am. J. Hum. Genet.* **72**, 598–610 (2003).
18. Hofmeister, R. J. *et al.* Parent-of-Origin inference for biobanks. *Nat. Commun.* **13**, 1–15 (2022).
19. Mathieson, I. *et al.* Genome-wide analysis identifies genetic effects on reproductive success and ongoing natural selection at the FADS locus. *Nature Human Behaviour* 1–12 (2023) doi:10.1038/s41562-023-01528-6.
20. Hosseini, S. H., Sadighi Gilani, M. A., Meybodi, A. M. & Sabbaghian, M. The impact of

- RABL2B gene (rs144944885) on human male infertility in patients with oligoasthenoteratozoospermia and immotile short tail sperm defects. *J. Assist. Reprod. Genet.* **34**, 505–510 (2017).
21. Lo, J. C. Y. *et al.* RAB-Like 2 Has an Essential Role in Male Fertility, Sperm Intra-Flagellar Transport, and Tail Assembly. *PLoS Genet.* **8**, e1002969 (2012).
  22. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).
  23. GTEx Portal. <https://gtexportal.org/home/>.
  24. Adham, I. M., Nayernia, K. & Engel, W. Spermatozoa lacking acrosin protein show delayed fertilization. *Mol. Reprod. Dev.* **46**, 370–376 (1997).
  25. Yamagata, K. *et al.* Acrosin accelerates the dispersal of sperm acrosomal proteins during acrosome reaction. *J. Biol. Chem.* **273**, 10470–10474 (1998).
  26. Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).
  27. Al Thani, A. *et al.* Qatar Biobank Cohort Study: Study Design and First Results. *Am. J. Epidemiol.* **188**, 1420–1433 (2019).
  28. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
  29. 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. *bioRxiv* 2022.10.19.512867 (2022) doi:10.1101/2022.10.19.512867.
  30. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
  31. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
  32. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
  33. Rubinacci, S., Delaneau, O. & Marchini, J. Genotype imputation using the Positional Burrows Wheeler Transform. *PLoS Genet.* **16**, e1009049 (2020).