



UNIL | Université de Lausanne

Unicentre

CH-1015 Lausanne

<http://serval.unil.ch>

Year : 2018

Integrative statistical analysis of-omics and GWAS data

Rüegger Sina

Rüegger Sina, 2018, Integrative statistical analysis of-omics and GWAS data

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_A922203F8CB34

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 universitaire de médecine et santé communautaires

INTEGRATIVE STATISTICAL ANALYSIS OF -OMICS AND GWAS DATA

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

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

par

Sina RÜEGER

Master de Zurich University of Applied Sciences

Jury

Prof. Bogdan Draganski, Président
Prof. Zoltán Kutalik, Directeur de thèse
Prof. Valentin Rousson, Co-directeur
Prof. Pierre-Yves Bochud, Co-directeur
Prof. Daniel Wegmann, expert
Prof. Bart Deplancke, expert
Prof. Iris Heid, expert

Lausanne 2018



UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Département universitaire de médecine et santé communautaires

INTEGRATIVE STATISTICAL ANALYSIS OF -OMICS AND GWAS DATA

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

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

par

Sina RÜEGER

Master de Zurich University of Applied Sciences

Jury

Prof. Bogdan Draganski, Président
Prof. Zoltán Kutalik, Directeur de thèse
Prof. Valentin Rousson, Co-directeur
Prof. Pierre-Yves Bochud, Co-directeur
Prof. Daniel Wegmann, expert
Prof. Bart Deplancke, expert
Prof. Iris Heid, expert

Lausanne 2018

Imprimatur

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

Président·e

Monsieur Prof. Bogdan **Draganski**

Directeur·rice de thèse

Monsieur Prof. Zoltan **Kutalik**

Co-directeurs·rices

Monsieur Prof. Valentin **Rousson**

Monsieur Prof. Pierre-Yves **Bochud**

Experts·es

Monsieur Prof. Daniel **Wegmann**

Monsieur Prof. Bart **Deplancke**

Madame Prof. Iris **Heid**

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

Madame Sina Rüeger

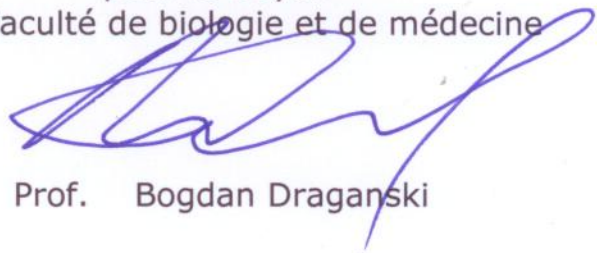
Master of Science in Engineering Zürcher Hochschule für Angewandte Wissenschaften,
Winterthur

intitulée

Integrative statistical analysis of -omics and GWAS data

Lausanne, le 21 septembre 2018

pour le Doyen
de la Faculté de biologie et de médecine



Prof. Bogdan Draganski

Integrative statistical analysis of -omics and GWAS data

PhD thesis of Sina Rüeger

Institute of Social and Preventive Medicine, Lausanne

Supervision by Prof. Zoltán Kutalik, Prof. Pierre-Yves Bochud and Prof. Valentin Rousson

10 July 2018

Acknowledgements

Throughout my PhD, I was incredibly lucky to receive support from colleagues at work, friends and family. I would like to thank:

- My **supervisors**: Most and foremost, I would like to thank **Zoltán Kutalik**. As my main supervisor, he shared his knowledge in statistical genetics, guided me into the right direction whenever necessary and helped me to overcome fears and struggle. I appreciate his loyalty and that he was always there for me when I needed his help. **Pierre-Yves Bochud**: for providing me with my first genomics project to work on when I started my internship in Lausanne and supporting me in the initial phase of my PhD. **Valentin Rousson**: for his external view on my projects and the helpful discussions in these years.
- My former and current **colleagues at work**: for their IT support, asking critical questions about my projects, coffee breaks, running to seminars and much more. I always enjoyed the collegial spirit in our group — helping each other out and growing as a group through different skill sets. **Ninon Mounier** and I share enthusiasm for crafting goods, but also similar standards when it comes to project management and data science. **Jonathan Sulc** was always willing to have a look at my writing (followed by rigorous editing) and always up for lunch. **David Lamparter** was eager to explain all sorts of complex biological and statistical topics to me. **Eleonora Porcu** has been an excellent role model and always happy to help and encourage, while treating me as an equal.
- **Angeline Chatelan**: for the discussions on topics around public health, nutrition and the impact of environment and genetics.
- The **R community**: for great support and tooling. Being part of the *R-Ladies* community has given me a confidence boost that I was missing at an earlier stage of my PhD. On a related note, the R-package bookdown to create and format this thesis, helped me focus purely on its content.
- **Beate Sick**: for introducing me to SNPs and eventually pointing me to Sven Bergmann, which redirected me to Zoltán.
- **Wei Wei**: for being a constant companion and giving solid advice on succeeding in my PhD.
- My **partner**: for accompanying me through the up and downs of my final PhD phase and helping me improve my work-life balance.
- My extended **familiy**: My **godparents** Vreni + Fritz and Thomas + Linda took their role very seriously and have always backed whatever I did in life. My **parents** not only provided me with genetic material, but also with an environment that allowed me to express my curiosity. With my father being a mechanic and my mum a web programmer, I believe, much of my enthusiasm for science and technology comes from them. While my father has sadly passed away, it was my mum who was there for me during the last few years.
- Finally, doctoral studies are not an easy ride for anyone. When I had difficulties that eventually lead to mental distress, the Consultation psychothérapeutique at UNIL helped out quickly and referred me to a CBT specialist who was of great help. I think, learning this new way of thinking was key for continuing my PhD (and probably for my future life in general).

Abstract

Complex traits such as human height or cardiovascular disease are highly polygenic, influenced by environmental factors and common in the population. By studying complex traits, we might be able to answer questions regarding the genetic contribution to a complex trait, gain insight into their genetic architecture and narrow down the responsible genetic variants. Such findings can ultimately lead to better treatment, prevention, diagnosis or prognosis of diseases.

Cost-effective DNA microarrays have made it possible to perform genetic studies at a large scale. A genome-wide association study (GWAS) aims to quantify the statistical association of each available genetic variant across the whole genome with a trait of interest in a group of individuals.

To eventually gain insight into the biological pathways underpinning traits, GWAS results (association summary statistics) can be used for follow-up studies by integrating summary statistics with external *-omics* data and applying additional statistical methods. For example, the heritability explained by typed genetic variants can be estimated from GWAS association summary statistics. Another example is a Mendelian randomisation, a method that is able to estimate the causal effect of one trait on another, and vice versa.

These statistical follow-up methods often use either individual-level genotype data or summary statistics combined external sequencing data as input. However, because effect sizes of genetic variants involved in complex traits are typically small, studies with larger sample size have more statistical power, which creates the need for combining public summary statistics, because access to individual-level data is often limited. What is more, summary statistics-based methods require information for the same set of SNPs for each study. To impute summary statistics of untyped variants, summary statistics imputation is used.

Summary statistic imputation follows the intuition that parts of the genome tend to be inherited together, which creates sets of correlated SNPs in close proximity ("in linkage disequilibrium (LD)"). Having information about a subset of SNPs and knowing the local LD structure from external reference panels, we can infer the summary statistics of untyped SNPs.

During my PhD, I investigated the limitations and potential of *summary statistic imputation*. First, I first improved the measure of imputation quality. Second, I extended the method, to have higher accuracy for imputation in cosmopolitan population cohorts. Third, I compared summary statistic imputation to genotype imputation and identified groups of genetic variants that are hard to impute. Fourth, I applied summary statistic imputation in a case study and discovered 34 additional height associated variants (19 of which replicated).

Résumé

Les traits complexes tels que la taille humaine, les maladies cardiovasculaires ou d'autres maladies souvent fréquentes dans la population, sont hautement polygéniques mais aussi influencés par des facteurs environnementaux. L'étude de ces traits complexes pourrait nous permettre de quantifier la contribution des facteurs génétique impliqués, de mieux comprendre leur architecture génétique et d'affiner l'identification des variants génétiques responsables. Ces résultats peuvent finalement conduire à améliorer à la fois le traitement, le diagnostic et le pronostic des maladies, mais également les stratégies de prévention mises en place.

L'arrivée sur le marché de puces à ADN à des prix accessibles a permis d'effectuer des études génétiques à grande échelle. Les études d'association pangénomique (GWAS) visent à mettre en évidence et à quantifier l'association statistique de chaque variant génétique ("Single Nucleotide Polymorphism" ou SNP) avec un trait d'intérêt dans un groupe d'individus (cohorte).

Pour obtenir un aperçu des mécanismes biologiques sous-jacents, les résultats de GWAS (statistiques synthétiques d'association) peuvent être utilisés pour des études additionnelles. Il est possible d'utiliser ces statistiques synthétiques pour appliquer des méthodes analytiques supplémentaires ou bien de les combiner avec des données -omiques externes. Par exemple, l'héritabilité expliquée par les variants génotypés peut être estimée à partir des statistiques synthétiques d'un GWAS. Un autre exemple d'analyse, appelé randomisation Mendélienne, permet d'estimer l'effet de causalité d'un trait sur un autre.

Ces méthodes d'analyses complémentaires nécessitent souvent des données génétiques au niveau individuel ou bien des statistiques synthétiques combinées avec des données de corrélation entre les SNPs. Cependant, les effets génétiques observés sont généralement modestes, et il est intéressant de combiner plusieurs cohortes pour augmenter la taille d'échantillon et ainsi obtenir une puissance statistique plus importante. C'est pourquoi les méthodes basées sur les statistiques synthétiques sont souvent préférées. En effet, l'accès aux données individuelles est limité, tandis que les statistiques synthétiques sont usuellement partagées publiquement. Néanmoins, pour pouvoir être combinées, ces statistiques synthétiques doivent être disponibles pour un même ensemble de variants génétiques. Afin d'imputer les statistiques synthétiques des variants non génotypés, et donc non disponibles dans certaines cohortes, l'imputation à partir de statistiques synthétiques est utilisée.

L'imputation à partir de statistiques synthétiques repose sur le fait que certaines parties du génome tendent à être héritées ensemble, ce qui crée des ensembles de SNPs, corrélés (en déséquilibre de liaison, ou LD). A partir des statistiques synthétiques d'association pour un sous-ensemble de SNP et d'informations sur la structure LD locale obtenue grâce à un panel de référence externe, il est possible d'inférer les statistiques synthétiques des SNPs non génotypés.

Pendant mon doctorat, j'ai étudié les limites et le potentiel de l'imputation à partir de statistiques synthétiques. Premièrement, j'ai amélioré la mesure de la qualité d'imputation de la méthode. Dans un second temps, j'ai également amélioré la méthode elle-même, de manière à obtenir une meilleure précision lors de l'imputation de cohortes multi-ethniques. Troisièmement, j'ai comparé l'imputation statistique à partir de statistiques synthétiques à l'imputation basée sur les données génomiques au niveau individuel et identifié des groupes de variants difficiles à imputer. Enfin, j'ai appliqué l'imputation à partir de statistiques synthétiques à une étude de cas sur la taille humaine, ce qui a permis d'identifier 34 nouveaux marqueurs génétiques associés avec les variations de taille humaine observées.

Contents

1	<i>Introduction</i>	7
1.1	<i>Recent development in genetics of complex traits</i>	7
1.2	<i>Complex traits</i>	9
1.3	<i>Genome-wide association studies (GWAS)</i>	10
1.4	<i>Bridging the knowledge gap</i>	13
1.5	<i>Beyond GWASs</i>	16
1.6	<i>Imputation methods</i>	24
2	<i>Results</i>	29
2.1	<i>Evaluation and application of summary statistic imputation</i>	29
2.2	<i>Improving summary statistic imputation for mixed populations</i>	30
2.3	<i>Applications of summary statistic imputation</i>	30
2.4	<i>Contributions to research in infectious diseases</i>	31
2.5	<i>Minor contributions to other publications</i>	32
3	<i>Discussion</i>	33
3.1	<i>Summary statistic imputation: limitations and future work</i>	34
3.2	<i>Key resources for future GWASs</i>	37
3.3	<i>Future work on translational GWASs</i>	38
3.4	<i>Conclusion</i>	42
	<i>References</i>	45

List of Figures

- 1.1 **Genetic variation:** Although ~ 3.2 bio bp long, the genome contains much less genetic variation. Sequencing efforts such as HapMap and the 1000 Genomes Project, give evidence to more than 15 mio variants. GWAS data can be imputed by inferring LD structure from the most recent reference panel, see Figure 1.8 8
- 1.2 **Allele frequency versus penetrance** 11
- 1.3 **How clinical findings of GWASs play into the four goals of health-care** 12
- 1.4 **Central dogma and biological cascade** 14
- 1.5 **Mendelian randomisation:** This illustration shows the underlying model used for Mendelian randomisation. An instrumental variable, a genetic marker G , affects an outcome Y through a risk factor X . The dashed line illustrates potentially violated assumptions. In Figure 1.4, the risk factor would be anywhere after the DNA, e.g. a molecular phenotype or a modifiable exposure. α and γ represent the association summary statistic GWAS with the outcome and the risk factor. β is the MR effect size. 19
- 1.6 **Different types of pleiotropy.** When association signals of multiple phenotypes colocate, this can have many reasons: **(A) Spurious pleiotropy:** A GWAS variant (black dot) tags two other variants: one (orange) that is linked to trait I and one (red), that is linked to trait II. **(B) Mediated pleiotropy:** The causal variant affects trait II through trait I. This type of pleiotropy can be detected through MR and genetic correlation. **(C) Biological pleiotropy:** the causal variant affects two traits through independent pathways. **(D) Pleiotropy through confounding:** genetic correlation is introduced by common confounder. (B) and (D): will display genetic correlation on a genome-wide level, (A) and (D): will display pleiotropy at a locus level. 20

- 1.7 **Individual-level genotype data to summary statistics: (A)** Genotype data for K individuals and M markers. Note that for SNV k there is no data. **(B)** LD structure between SNVs estimated through the squared correlation. **(C¹)** Association summary statistics: estimating per allele effect sizes and the standard error. Using these two summary statistics, along with sample size (N) and MAF , we can derive other association summary statistics, such as Z-statistics (Z), the explained phenotypic variance (r^2) or the standardised effect size (r). **(C²)** Aggregated data form, also called *summary statistics*, listing sample size, effect allele frequency, effect size and its standard error. Summary statistics that quantify the association are called *association summary statistics*. **(D)** Imputation of the untyped SNV k as a function of genotype data and haplotype information ($D(A, B)$), or as a function of summary statistics and LD structure ($D(C^2, B)$). For both options, haplotype information and LD structure is estimated from external reference panels. See also Figure 1.8. 23
- 1.8 **Imputation with reference panels** 25
- 2.1 **Download** *Evaluation and application of summary statistic imputation to discover new height-associated loci* **here.** 29
- 2.2 **Download** *Improved imputation of summary statistics for mixed populations* **here.** 30
- 2.3 **Download** *Rare and low-frequency coding variants alter human adult height* **here.** 31
- 2.4 **Download** *Bayesian association scan reveals loci associated with human lifespan and linked biomarkers* **here.** 31
- 2.5 **Download** *Impact of common risk factors of fibrosis progression in chronic hepatitis C* **here.** 31
- 3.1 **Combining MR and drugbank info into a score matrix** 39
- 3.2 **Approach A - comparing drugs** 40
- 3.3 **Approach B - comparing traits** 41
- 3.4 **Approach A in Total Cholesterol:** Y-axis shows the scores, each point is a drug, with a boxplot overlayed. LHS displays TC-unspecific-drugs, the RHS are the TC-specific drugs. One-sided Wilcoxon rank sum test ($P = 5 \times 10^{-5}$) 41
- 3.5 **Approach B in Total Cholesterol:** The x-axis shows scores, the y-axis the TC-specific drugs. Each dot is one trait, with TC as a solid dot. Drugs are ordered according to the ranking of the black dot (score for TC). Binomial test rank 1st or 2nd (#successes=8, #trials=17): $P = 7.6 \times 10^{-7}$. 42

1

Introduction

My PhD thesis is titled *Integrative statistical analysis of -omics¹ and GWAS² data*. To summarise, my thesis is about how statistical methods applied to -omics and GWAS data can lead to additional insights into the genetic basis of diseases, and also demonstrates the need for such methodology as summary level data can be accessed freely, while individual-level data is scarce.

This document is a collection of my PhD output, accomplished with the help of others in the domain of statistical genetics and under the primary supervision of Zoltán Kutalik, as well as Valentin Rousson and Pierre-Yves Bochud. I started my PhD with a manuscript on estimating the attributable fraction of genetic and environmental/lifestyle factors in patients with Fibrosis after a Hepatitis C infection. The majority of the remaining time was spent around understanding and improving summary statistic imputation, while also applying this method to various real data.

This chapter is an introduction to the analysis of complex traits³, the utility of GWASs and the inference of summary statistics of untyped markers. The next chapter provides a summary of my work. A discussion will set my output into context, discuss the current challenges and give an outlook to future work. Lastly, my two first author papers are attached.

First, let us define the **overarching goal of genetic epidemiology**. Most diseases (or health, if viewed the other way around) are partially driven by genetic risk factors. Statistical methods enable us to quantify the genetic contribution to a disease, help to gain insight into genetic architecture⁴ and narrow down the responsible genetic variants⁵, ultimately leading to better treatment, prevention, diagnosis and prognosis.

1.1 Recent development in genetics of complex traits

After the millennium, the technical progress made it possible to measure biomedical data at low cost. For measuring DNA, having low-cost DNA microarrays meant to move from expensive and hypothesis driven candidate-gene studies and linkage analyses to a hypothesis-free genome-wide approach.

In 2003, the **first human genome sequence** was completed

¹ -omics: Data ending in -omics, such as genomics, transcriptomics or metabolomics.

² *Genome-wide association study (GWAS)*: Aims to identify genetic variants associated with a trait of interest. This is done by testing the association of each available genetic variant across the genome with a trait in a group of individuals.

³ In this document, I use the terms *disease*, *trait* and *phenotype* interchangeably.

⁴ *Genetic architecture* refers to the landscape of genetic contributions to a given phenotype. It comprises the number of genetic variants that influence a phenotype, the size of their effects on the phenotype, the frequency of those variants in the population and their interactions with each other and the environment (Timpson et al. 2018).

⁵ *Genetic variant* or *genetic marker* are terms that include SNVs, copy number variation, methylation and other epigenetic variation. *SNP*: Single nucleotide polymorphism. Strictly speaking only common variants (minor allele frequency > 0.05). *SNVs*: Single nucleotide variant. A general term that includes all variants irrespective of minor allele frequency.

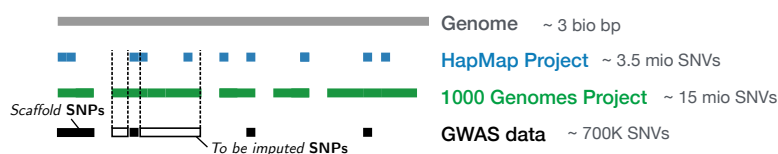
(Collins et al. 2004). This milestone allowed other **worldwide genotyping & sequencing projects** to take their course. The International HapMap Consortium (2003), completed in 2005, and later the 1000 Genomes Project Consortium (2010), completed in 2012, explored the genetic variation in a multitude of different populations. One outcome of this investigation lead to insight into the **linkage disequilibrium (LD)**⁶ structure across the genome and ultimately to a list of tag SNPs⁷ that are able to capture the majority of common genetic variation (Visscher et al. 2012).

Cost-effective DNA microarrays (SNP arrays) make use of LD structure, by containing a small fraction of all genome-wide SNPs, each of them tagging multiple other variants (1.1).

Having an inexpensive way to extract DNA information, the genetic variation affecting a complex trait can be explored in a **hypothesis-free, genome-wide manner**. This is done by regressing the trait in question onto each genotype in an univariate fashion, whereby the resulting association summary statistics then helps to decide, whether this variant is associated with the trait or not. This is the basis of every **genome-wide association study (GWAS)**. It is an experimental design that scans systematically over the genome in a set of individuals (Visscher et al. 2012).

The first GWAS, published in 2005 (Klein, Zeiss, and Chew 2005), identified gene *CFH* to be associated with age-related macular degeneration. The authors analysed 96 cases and 50 controls in age-related macular degeneration, by screening 116'204 SNPs throughout the genome.

Since then, a vast number⁸ of GWASs and meta-analysed GWASs have been performed.



The initial hope was, to identify the genetic mutations involved by simply extrapolating the analysis approaches done in Mendelian disorders to complex traits (Visscher et al. 2012). This hope was quickly diminished as it became clear, that the genetic architecture underlying complex traits is far more complicated than in Mendelian disorders.

Correctly applied, GWAS can **give insight into biology** from different angles. Firstly, GWAS results can propose candidate genes, which are later verified in laboratory experiments. Secondly, GWAS results can be used for follow-up studies **integrating GWAS association results with external data** and applying additional analytical methods, can allow insight into the biological pathways underpinning traits.

⁶ *Linkage disequilibrium (LD)*: Non-random association of SNPs. In humans, *meiosis* reduces the number of chromosomes to 23 in the maternal and paternal cell. *Genetic recombination* then combines maternal and paternal chromosome pairs by splitting the chromosomes into pieces, shuffling it and rearranging them. Because of this mechanism, an individuals' chromosome will consist of a unique combination of maternal and paternal DNA. The chromosome is broken down at similar positions, therefore certain blocks of DNA tend to be inherited together. Hence the variants within these blocks are in *linkage disequilibrium*. Other influences of LD structure include selection, drift, mutation rate.

⁷ *Tag SNPs*: SNPs that are correlated with neighbouring SNPs. When tag SNPs are typed, they can serve as a surrogate for untyped SNPs.

⁸ The GWAS Catalog (Welter et al. 2014), a curated repository for GWAS results, contains SNP associations from 3'395 publications (May 21 2018).

Figure 1.1: **Genetic variation**: Although ~ 3.2 bio bp long, the genome contains much less genetic variation. Sequencing efforts such as HapMap and the 1000 Genomes Project, give evidence to more than 15 mio variants. GWAS data can be imputed by inferring LD structure from the most recent reference panel, see Figure 1.8

This latter option is what my PhD thesis aims to tackle.

1.2 Complex traits

In genetics, traits are classified as *monogenic* (or Mendelian) or *complex*. In contrast to rare *monogenic traits*, *complex traits*⁹ involve more than one genetic variant that (often mildly) alters the predisposition to a trait or disease. *Complex traits* are often highly polygenic and thus common.

Traditionally, genetics has been done in **monogenic diseases** that follow Mendelian inheritance. If large pedigrees are available, such Mendelian diseases are easier to study, because the genetic variant(s) responsible for the disease susceptibility are rare and have maximal penetrance. Finding the gene(s) involved in Mendelian diseases involve sequencing and studying pedigrees.

Complex traits¹⁰ can be grouped into various (partially overlapping) subcategories of classical medicine:

- Anthropometric traits (e.g. height or BMI)
- Neurological traits (e.g. Alzheimer disease or schizophrenia)
- Immune-related traits (e.g. asthma or Crohn's disease)
- Haematological traits (e.g. haemoglobin)
- Metabolic traits (e.g. type 2 diabetes)
- Reproductive traits (e.g. age at menarche or number of offspring)
- Social traits (e.g. educational attainment)
- Cardiovascular traits (e.g. coronary artery disease)
- Cancer (e.g. prostate cancer)

Like Mendelian traits, **complex traits** have been studied since early 1900 in families to investigate heritability. For example, the heritability of **human height** was studied in 1918 already (Fisher 1918).

Human height is an ideal model trait to study. With an estimated broad-sense heritability of around 80% it is highly heritable and an easy-to-measure model trait, and only moderately influenced by environmental factors.

Studying complex traits involve much **larger sample sizes** than linkage studies because of **polygenicity** with **low penetrance** variants. GWASs precisely aim to do that by estimating the association summary statistics¹¹ between each genetic variant and a complex trait for a set of individuals. This way, thousands genetic variants have been identified to be associated with complex traits (Welter et al. 2014). For example, for human height the genetic variants found, explain each up to 0.43% of the phenotypic variance (Wood and others 2014; Marouli and others 2017)¹².

Although GWASs are thought to be *hypothesis-free* (in the sense of a non-targeted genome-wide scan), there are some assumption regarding the **underlying model**, e.g. SNPs affect a trait in an additive way.

⁹ For a review on complex traits, see McCarthy et al. (2008).

¹⁰ Complex traits are either quantitative traits (for example height) or a disease (for example depression). However, to simplify terminology, I will refer to *traits*.

¹¹ *Summary statistics* are aggregated forms of individual-level data. In the context of GWASs these are for example sample size or minor allele frequency for each SNP. *Association summary statistic* is more specific in that it quantifies the *association* with a trait of interest, for example effect size, Z-statistic or the corresponding P-value for each SNP.

¹² For comparison, the explained phenotypic variance of the FTO variant (rs1558902) is 0.22% (Locke and others 2015).

McCarthy et al. (2008) illustrated (see Fig 1.2), how the **penetrance** of a variant is **related** to its **MAF**¹³. Common variants tend to have smaller effect sizes (except for traits not under natural selection) (Guo et al. 2018), while it is assumed that rare variants can cover the full range of effect sizes. Although, so far, the rare variants discovered mainly display large effects. In order to be able to detect small effects of rare variants, sample size must be large (to have enough power and to observe the variant) and the data must be high quality.

Besides effect sizes, the **explained phenotypic variance** is tied to allele frequency too. For each variant, the explained phenotypic variance can be estimated as $2\beta^2 f(1-f)$, with f being the MAF, β the effect size on the phenotype. Because of this relation, Robinson, Wray, and Visscher (2014) make the point, that rare variants with high penetrance will still only explain a small fraction of the phenotypic variance.

The **underlying biological model** of a complex trait is thought to involve large gene regulatory networks, affecting for example gene expression level of hundreds of acting genes, protein translation or protein folding, in combination with environmental influences (that potentially interact with the gene regulatory pathways). Currently, there are two models proposed for complex traits. The *polygenic* model is, that the causal genetic variants (e.g. SNPs, rare variants, gene-gene interactions, copy number variants, DNA methylation or histone modification) cluster into key pathways that are relevant for a disease. The *omigenic* model (Boyle, Li, and Pritchard 2017) is, that all gene regulatory networks are interconnected; all genetic variation is associated to a complex trait, but apart from the core genes that represent real biology, the peripheral genes show association only through indirect, non-trait specific effects. These two views — polygenic and omnigenic — are part of an on-going discussion (Wray et al. 2018).

To this end, GWAS results have been guiding several proposed underlying biological models of complex traits.

1.3 Genome-wide association studies (GWAS)

In a GWAS¹⁴ a genome-wide scan of statistical genotype-trait associations in a set of individuals is done, to tests whether DNA variation is related to an alteration in the trait.

To identify the genetic markers that act on complex traits, the framework of GWASs is used to **estimate univariate genotype-trait associations**. Typically, the association between each genetic marker and the trait is estimated using linear **regression** (or another statistical model depending on the outcome), with a trait as the outcome, the genetic marker (genotype dosage) as a predictor, and covariates to increase estimation accuracy and correct for confounders¹⁵. Such lean models are computationally inexpensive to run and make follow-up analyses, such as meta-analyses, easier to

¹³ *Minor allele frequency (MAF)*: The allele frequency of the less common allele. Minor alleles can change across populations, which is why large-scale meta-GWASs often report EAF instead. *Effect allele frequency (EAF)*: The allele frequency of the effect allele (and not the reference allele).

¹⁴ For a review on GWASs see McCarthy et al. (2008), Visscher et al. (2012) or Visscher et al. (2017).

¹⁵ *Confounder*: A variable that is associated with a predictor and the outcome. For example, GWASs often account for population structure as a confounding factor.

Adapted from McCarthy et al. (2008)

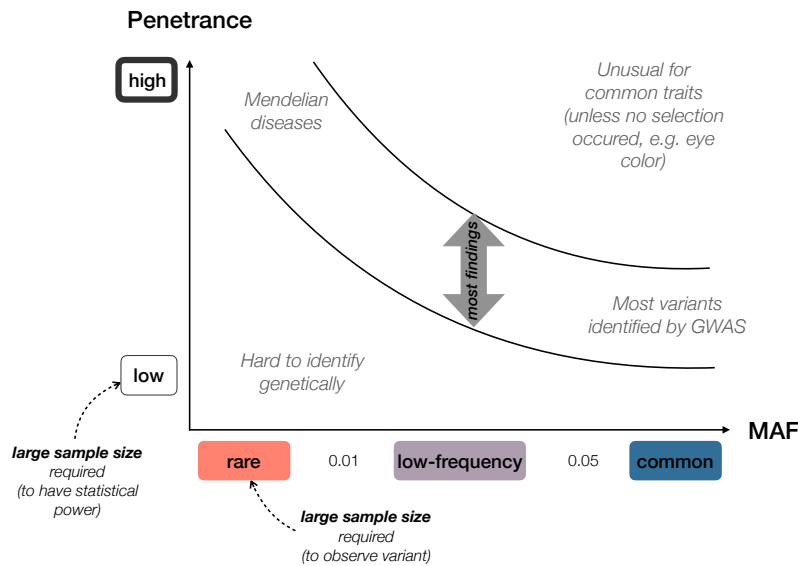


Figure 1.2: **Allele frequency versus penetrance:** This figure is a schematic representation of the (proposed) relationship between allele frequency and penetrance (or effect size). Most GWAS findings so far have been within the black arch ranging from low-MAF-high-effect-size to high-MAF-low-effect-size. In order to detect genetic variants residing in the lower left triangle, sample size must be large for two reasons. Firstly, to observe rare variants in a set of individuals, and secondly, to have large enough power to estimate low effect sizes.

apply, as described later.

Although the central **goal of GWASs** is to gain insight into the genetic architecture of complex traits, the ultimate aim of doing so is manifold (Fig 1.3) and can be broadly grouped into the **four cornerstones of healthcare**: *therapy, prevention, prognosis and diagnosis* (McCarthy et al. 2008). In the context of genetic epidemiology, these four goals can be achieved with different means (Figure 1.3). **Biomarkers**, such as RNA levels, qualify to support all four goals; **polygenic risk scores** help to predict disease progression, improve diagnosis and prevent diseases; **single SNPs** have (so far) only been useful in very special cases for diagnosis of diseases and stratifying for treatment options. The case of gene *IL28B* in patients infected with hepatitis C virus (Ge et al. 2009; Rauch et al. 2010) showed that SNPs located in this specific gene can be used to predict treatment response and spontaneous clearance, therefore allowing to personalise treatment.

To tackle these four goals, we need (among others) to quantify heritability, understand the underlying biology, identify (all) disease associated markers and build strong predictors.

Because of the genetic architecture underlying complex traits, **large sample sizes** are required to study them to have enough power to detect genetic variants with small effect sizes. For example, if we assume that a variant has an explained phenotypic variance of 0.3%¹⁶, then the sample size sufficient to detect such a variant with a statistical test, needs to be larger than 7062 (given 80% power). As the explained phenotypic variance decreases, sample size needs to increase.

To accumulate such large sample sizes, population cohorts with identical phenotypes and genetic data at hand collaborate in **consortia**.

¹⁶ This is the explained phenotypic variance of the FTO variant (Locke and others 2015), the genetic variant explaining most of the phenotypic variance of BMI to date.

Clinical findings from GWASs			
	Single SNP	Polygenic risk score	Biomaker (e.g. RNA)
Therapy	(✓)		✓
Prevention		✓	✓
Prognosis		✓	(✓)
Diagnosis	(✓)	✓	✓

Figure 1.3: How clinical findings of GWASs play into the four goals of healthcare: The four main cornerstones of healthcare are: therapy (including personalised treatment options), prevention, prognosis and diagnosis. Currently, a single SNP can be used to personalise therapy and to improve diagnosis. Polygenic risk scores are used for prevention plans, prognosis of disease progression and diagnosis. Biomarker, for example RNA levels, provide the most complete capacity, ranging from therapy to diagnosis.

However, privacy restrictions limit sharing individual-level data (and analysing individual data from hundreds of thousands of individuals would demand computing power and storage). To circumvent these constraints, each cohort executes a GWAS according to a study plan and then deposits the association summary statistics on a server. Once every cohort has submitted its results, an inverse variance weighted **meta-analysis** is run. The resulting meta-analysed summary statistics are then published along with a publication, ready for other researchers to be used. As an example, GIANT, a consortium focusing on anthropometric traits, has released complete summary statistics for 15 publications.

The challenges that such consortia face are rooted in the design of the meta-analysis and quality control. The contributing cohort might be heterogeneous in terms of design (for example a mix of birth, prospective, cross-sectional or longitudinal cohorts) and ancestry (mixed or admixed populations). Furthermore, there is little control over the analysis performed by an individual cohort, beyond using quality control tooling (Winkler et al. 2014).

Although these meta-analysed GWAS deserve a special term (e.g. mGWAS¹⁷), I will for simplicity continue to call them GWAS.

The **caveats** of GWASs can be broadly assigned into three groups: experimental design, model formulation and interpretation. The first group concerns for example how many individuals from which population are selected and if there is a replication cohort (to date, there are only a few study cohorts with more than 100K participants). Caveats concerning statistical modelling include outcome and predictor transformation, accounting for measured environmental correlates and confounders (such as population structure),

¹⁷ Another caveat is, that the *m* in mGWAS could stand for *meta*, *methylation*, *microbiom* or *metabolom*.

multiple testing correction and power issues. When interpreting or utilising these results as input for other methods, a phenomenon such as winner's curse should be accounted for. Another interpretation fallacy is to interpret the identified lead SNPs as the causal ones. Instead, lead SNPs derived through pruning¹⁸ are often tagging one or more causal SNPs through LD. Complex LD structure and sampling error make it hard to identify the causal variant(s) with such simple strategies.

Most of the lead SNP arising from GWAS results found to date are residing in **non-coding regions**, indicating regulatory mechanisms (Schork et al. 2013; Astle et al. 2017) or tagging of coding variants. Taking LD structure between SNPs and information about molecular function of SNPs (such as eQTL, methQTL or chromatin-QTLs) into account can help narrowing down causal variants.

In a next step, summary statistics can be used for **follow-up analyses**, for which traditionally individual-level data was used. For example, Mendelian randomisation, a causal inference method, uses GWAS summary statistics to identify how phenotypes are related and is potentially able to assign cause and consequence.

1.4 Bridging the knowledge gap

Univariate GWAS summary statistics provide us with information about the statistical association of genetic variants and a trait, while controlling for some environmental covariates and genetic confounding.

Combining¹⁹ such GWAS summary statistics with additional, external data (such as the LD structure, curated pathways, other GWAS studies, or simply algorithms) can leverage the GWAS results to answer questions regarding heritability, narrow down regions that harbour causal variants, quantify how traits relate to each other and determine the actors in the genetic cascade of a certain gene (shown in Figure 1.4).

Before going into the details of analytic approaches, let us first define what we want to explore: (1) the **"black box"** view, where we care mostly about **prediction and heritability** (because effect sizes can be translated into heritability or phenotypic variation), or (2) **understanding the biological system**, where we are interested the molecular pathways underlying traits, and how they work in concert across traits. In the following I will describe both views in detail.

1.4.1 The gap between heritability and GWAS findings

The heritability of a trait quantifies, how much of the variation of that trait can be attributed to genetic variation (including additive and non-additive effects, such as dominant, epistatic effects). This is called **broad-sense heritability** (H^2) and can be estimated from family studies (H_{Ped}^2).

¹⁸ *Pruning* or *clumping* is a selection strategy based on LD, MAF and/or association summary statistic to define top SNPs in a genomic loci.

¹⁹ For an overview on GWAS and how it can be used to discover the biology of diseases, translate into new therapeutics, and understand the underlying genetic architecture see Visscher et al. (2017) and Timpson et al. (2018).

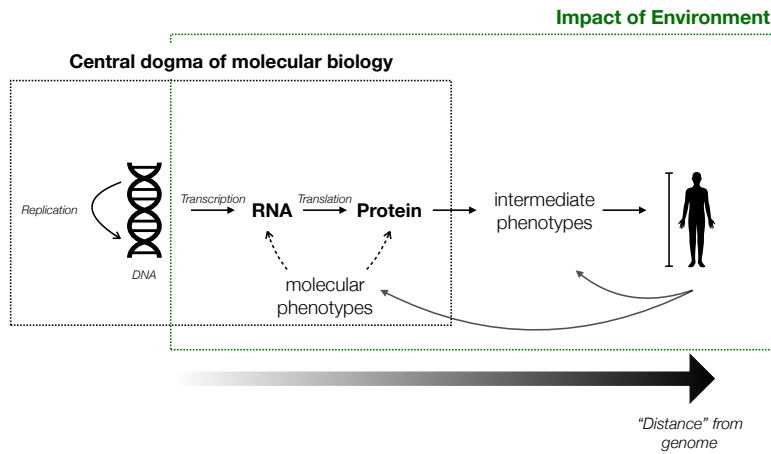


Figure 1.4: **Central dogma and biological cascade:** This illustration shows how the central dogma of molecular biology fits into a the (hypothetical) biological cascade. DNA information flows within DNA, or from DNA to proteins through RNA. The downstream effect of genetic variation therefore affects everything from RNA onwards. Environmental changes (everything that is not genetic) do not impact genetic variation (unless radiation occurs), but affect molecular, intermediate and more distant phenotypes. Any phenotype to the right can also loop back into molecular or intermediate phenotypes. Height, illustrated with a human body, is assumed to be very distal from the genome, involving many biological processes to act on, and therefore highly polygenic.

Narrow-sense heritability (h^2) is simply the variance explained²⁰ by additive genetic effects. This is sometimes called *SNP-heritability* h^2_{SNP} , the proportion of phenotypic variance explained by SNPs. h^2_{GWAS} is the total explained variance of genome-wide significant loci. By definition, $h^2_{GWAS} < h^2_{SNP} < H^2_{Ped}$.

The term *missing heritability* is the discrepancy between the H^2_{Ped} and h^2_{GWAS} for most traits.

For example, **height** has an estimated broad-sense heritability of 80%. Over the years, new findings have decreased the missing heritability. In 2008, 40 loci associated with human height only explained 5% of phenotypic variance (Visscher 2008). In 2014, an increase in sample size to 250K discovered 697 additional height-associated SNPs, that explain 19% of the phenotypic variance (Wood and others 2014). In 2017, a focus on coding variants discovered 120 new height associated loci that raised the explained phenotypic variance to 22% (Marouli and others 2017). Finally, in 2018, a meta-analysis of $\sim 700K$ individuals explained 34.7% of the phenotypic variance (using $\sim 15'000$ SNPs with $P < 0.001$) (Yengo et al. 2018).

The discrepancy between SNP-heritability (h^2_{SNP}) and the heritability of associated SNPs only (h^2_{GWAS}) is due to the fact, that many genetic variants that might be associated with height do not reach genome-wide significance because of power issues due to insufficient sample size or are not genotyped/not imputable.

The discrepancy between the broad-sense heritability observed in family studies (H^2_{Ped}) and the narrow-sense heritability (h^2_{SNP}) (80% versus 55% in human height) is due to neglecting G-by-G²¹ interaction, G-by-E²² interaction, other genetic variation such as copy number variants or rare variants.

In principle, there are three ways to tackle missing heritability:

²⁰ *Variance explained:* Proportion of variance in an outcome explained by a statistical model.

²¹ *G-by-G:* Gene-gene interaction (epistasis)

²² *G-by-E:* Gene-environment interaction

1. **Increase sample size** in order to increase power and precision.
2. **Genotype or sequence variants** that are poorly imputable.
3. **Exploit genetic correlation or causation** between traits (McDaid et al. 2017; Turley et al. 2018; Maier et al. 2018).

Broad-sense heritability is a measure that informs us about the maximum potential of GWASs. Narrow-sense heritability guides the discovery of new variants and tells us if the search has been exhausted.

Details about the estimation of heritability are below.

1.4.2 *Understanding biology*

A second type of methods tries to understand the biological system in its details. In Figure 1.4 this would be to **discover** all the **molecular processes** that lead from **genetic variation to a trait** of interest. Dermitzakis (2008) showed how genetic variation is linked to gene expression that can translate into disease risk. Most importantly, the article points out, how gene expression is a cellular phenotype, that varies from tissue to tissue. The GTEx Consortium (2013) addressed this hypothesis of genotype and tissue-specific gene expression levels being correlated.

There are many **global projects** that seek to understand the layers in between genetic variation and diseases:

- ENCODE is a consortium that aims to describe the functional elements in the human genome (The ENCODE Project Consortium 2012) .
- The Roadmap epigenomics project focuses on epigenetic variation (DNA methylation, histone modification, chromatin accessibility, small RNA transcripts) in stem cells and ex vivo tissues (Roadmap Epigenomics Consortium 2015).
- GTEx provides univariate genotype-gene expression analysis results for multiple human (post-mortem) tissues (The GTEx Consortium 2013).

These datasets can be used to find **eQTLs**²³ (Westra et al. 2013; Zhernakova et al. 2017) or mQTLs (metabolomic QTLs) (Rueedi et al. 2017).

Finding **causal DNA-to-trait pathways** can be approached from different angles.

Tools such as PASCAL (Lamparter et al. 2016) and DEPICT (Pers et al. 2015) combine GWAS summary statistic with tissue specific pathway information and report back relevant gene sets for the trait of interest.

A first step to understand the mechanism of action is to start from GWAS association results and to narrow down causal variants with **fine-mapping** methods.

Yet another way of unravelling SNP-trait pathways, is to look at all intermediate layers between genetic variation and the trait

²³ *Expression quantitative trait loci (eQTL)*: A QTL study is essentially a GWAS with a molecular phenotype as an outcome.

(e.g. gene expression or related traits), and to integrate this information with the GWAS association results. Such methods are called **causal inference** or **multi-trait analyses**.

1.5 Beyond GWASs

This section is dedicated to introduce methods that make use of GWAS association summary statistics²⁴.

In the following I am listing for each class of methods at least one approach that uses summary statistics, as well as one using individual-level data. There is a **trade-off** between easy-to-use *summary statistics*, and using *individual-level genotype data*. Because, sample size is crucial, and summary statistics can be easier shared and the necessary LD can be estimated from external reference panels, methods using summary statistics are preferred.

1.5.1 Heritability

As introduced above, the heritability²⁵ of a trait quantifies, how much of the variation of that trait can be attributed to genetic variation. This “general” heritability is called *broad-sense* heritability. *Narrow-sense* heritability is a subset of *broad-sense* heritability and more straightforward to estimate.

Broad-sense heritability (H^2) can be estimated with family-based studies, such as parent-offspring regression or twin studies (H^2_{Ped}). For example; Visscher et al. (2006) studied full-sib pairs and estimated H^2_{Ped} for height to be 0.8, although this figure will vary depending on the variation of the environment (Visscher, Hill, and Wray 2008).

Cost-effective genotyping and GWAS data make it possible to estimate narrow-sense heritability (the proportion of phenotypic variance explained by SNPs) or *SNP heritability* (h^2_{SNP}) from genotype data of unrelated individuals (Yang et al. 2017). There are two principle methods that estimate h^2_{SNP} from *individual-level genotype data*:

- **GCTA** (Yang et al. 2011)
- and **LDAK** (Speed et al. 2016).

Both methods fit a linear mixed effects model, treating all SNPs as random effects and calculating the genetic relationship matrix between individuals. The variance explained by all SNPs can then be estimated using restricted maximum likelihood (REML) approach. While GCTA expects that each SNP contributes equally to heritability, LDAK uses a more generalisable approach, by allowing the expected heritability of each SNP to vary with LD levels, genotyping quality, and estimating relationship between heritability and MAF.

To estimate narrow-sense heritability from *summary statistics*, two methods can be used:

²⁴ For a review on statistical methods that leverage summary association data see Pasaniuc and Price (2017).

²⁵ An introduction to missing heritability is provided by Visscher, Hill, and Wray (2008), further reading by Yang et al. (2017).

- LD-score regression (**LDSC**) (B. K. Bulik-Sullivan et al. 2015) uses the same model as GCTA and displays therefore the same weaknesses.
- **SUMHER** (Speed and Balding 2018) is analogous to LDAK and models the heritability as a function of MAF, LD levels and genotyping uncertainty.

Additionally, LDSC and LDAK can be used to quantify confounding in a GWAS and estimate genetic correlation between traits.

All four presented methods are based on genome-wide data. There are also attempts to calculate heritability locus-wise, for example with BOLT-LMM (Loh et al. 2015) or partitioning the heritability by functional annotation using LDSC (Finucane et al. 2015).

Finally, one approach is to calculate the heritability from genome-wide significant SNPs only. However, this heavily underestimates the narrow-sense heritability and is bound to winner's curse (Wood and others 2013).

1.5.2 Fine-mapping

The interpretation of GWAS results is often difficult because the **statistically significant variants** detected by GWASs are typically **not the causal ones**, but may rank high in terms of statistical significance because of complex LD structure. Simply prioritising variants by *P*-values is especially sub-optimal when more than one causal variant is present at a locus (Pasaniuc and Price 2017). Moreover, the associated variants might fall into non-coding regions.

Linking the causal variant with a particular gene is complicated through allelic heterogeneity, the presence of multiple causal variants of a trait at the same locus (Hormozdiari et al. 2017).

Fine-mapping²⁶ is a technique that aims to refine the genomic localisation of causal variants at a given locus that are most likely to be functional (which is different from variants that are in LD with the causal variant).

²⁶ Review on fine-mapping: Schaid, Chen, and Larson (2018)

Conditional analysis is one way to refine the association signals. *Approximate conditional analysis* is the umbrella term for identifying multiple signals at a locus using *summary statistics*:

- **COJO** (Yang et al. 2010) applies a step-wise **conditional analysis** and assumes multivariate normal distribution.
- **SOJO** (Ning et al. 2017) uses the Lasso framework to select variants. The equivalent of these methods for **individual-level genotype** data are Lasso and conditional analysis to jointly fit multiple SNPs.

Another approach is, to compute the **posterior probabilities** of causality for every SNP using *summary statistics*, for example CAVIAR (Hormozdiari et al. 2017), CAVIARBF (Chen et al. 2015) (CAVIAR using Bayesian factors), FINEMAP (Benner et al. 2018) or

fastPAINTOR (Kichaev et al. 2017). This type of analysis is computationally intensive; hence the maximum number of causal variants is often restricted. Except for FINEMAP, all methods need the number of causal variants at a given locus as input. fastPAINTOR also calculates the posterior probability using Bayesian statistics, but integrates functional annotation as well. There are other Bayesian fine-mapping methods that use *individual-level data* as input, such as BIMBAM (Servin and Stephens 2007).

Note that, after applying fine-mapping techniques, further follow-up investigations are essential, either through replication in different studies or laboratory functional studies (Schaid, Chen, and Larson 2018).

1.5.3 Causal inference

Causal inference²⁷ methods help to determine the strength and direction between connected phenotypes.

²⁷ For an overview on Mendelian randomisation see Pingault et al. (2018).

The fundamental aim of epidemiology is to determine the causes of diseases. Many epidemiological analyses focus on **whether an exposure modifies** the severity or the risk of the disease.

These risk factors can be modifiable exposures (e.g. smoking or nutrition), molecular phenotypes (proteins, gene expression, metabolom), or related traits (e.g. cardiovascular disease, socioeconomic status).

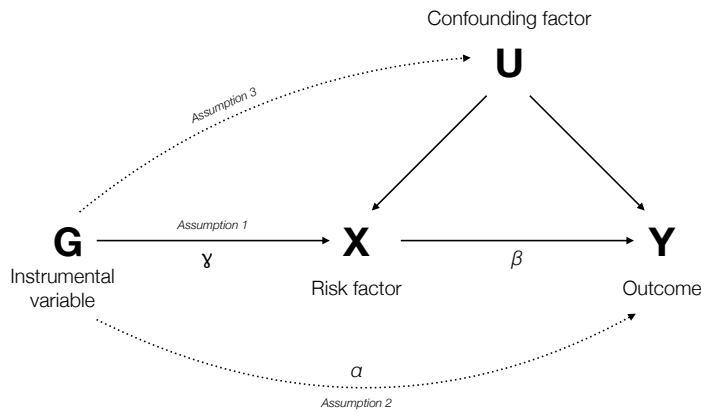
However, causal inference of such complex networks with **traditional techniques** such as randomised control trials are **often impossible** due to ethical issues, low sample size, limited time and limited funding. Furthermore, randomised control trials require careful randomisation and monitoring of participants is challenging.

Mendelian randomisation (MR) (Davey Smith and Ebrahim 2003; Burgess, Butterworth, and Thompson 2013; Bowden et al. 2015) exploits a natural randomising scheme of genetic variants through meiosis, where genetic variants can be used as instrumental variables (Didelez and Sheehan 2007).

In order to apply MR, three key **assumptions** have to hold. They are all centred around the genetic variants, the exposure/risk factor and the outcome, illustrated in Figure 1.5. The instrumental variable G :

1. is associated with the risk factor X ,
2. is not associated with any confounder C of the risk factor–outcome association,
3. is conditionally independent of the outcome Y given the risk factor X and confounders C .

MR had a boost over the recent years. The reason is, that GWAS results enable instrumental variables that satisfy assumption #1 much better.



Adapted from Bowden et al. 2015

Figure 1.5: **Mendelian randomisation:** This illustration shows the underlying model used for Mendelian randomisation. An instrumental variable, a genetic marker **G**, affects an outcome **Y** through a risk factor **X**. The dashed line illustrates potentially violated assumptions. In Figure 1.4, the risk factor would be anywhere after the DNA, e.g. a molecular phenotype or a modifiable exposure. α and γ represent the association summary statistic GWAS with the outcome and the risk factor. β is the MR effect size.

To perform MR with *individual-level data*, two-stage least squares (TSLS) MR (Baum, Schaffer, and Stillman 2003) can be used. Unless large sample sizes are used (such as UK Biobank), this is not feasible anymore.

Analogous to TSLS, there is **two sample MR** (Burgess, Butterworth, and Thompson 2013) (summarized inverse-variance weighted estimation) for *summary statistics*. Other pleiotropy-robust methods include the ratio method, MR-Egger estimation (Bowden et al. 2015) and the weighted median estimator (Bowden et al. 2015).

Recent advances include **SMR** (summary-data-based MR) and **GSMR** (generalised summary-data-based MR). These methods can use eQTLs as instrumental variables, gene expression summary statistics as risk exposure, and common complex traits as outcomes (Zhu et al. 2016; Verbanck et al. 2018; Porcu and others 2018). This setting allows identifying genes whose expression levels are associated with the trait.

MR is closely linked to **multi-trait analysis**. MR explores how traits are causally connected, while a multi-trait analysis exploits the more general concept of (genetic) correlation between traits (*correlation \neq causation*).

1.5.4 Multi-trait analysis

Multi-trait analyses can help to understand how traits are related to each other.

From a genetic perspective, there are two ways how traits can be related. Either on a locus-level through shared genetic variants or loci (**pleiotropy**) or via a systematic **correlation between SNP** effects (Pasaniuc and Price 2017). Figure 1.6 illustrates these two

classes row-wise. The top row shows locus-level pleiotropy (*spurious pleiotropy* and *biological pleiotropy*), while the bottom row shows systematic pleiotropy/correlation (*mediated pleiotropy* and *pleiotropy through confounding*).

Quantifying the associations between traits, requires distinguishing these different types of pleiotropy - *mediated pleiotropy* and *true pleiotropy* - using colocalisation methods, MR or genetic correlation. Seemingly related traits might fall under *spurious pleiotropy* or *pleiotropy through confounding*.

Adapted from Hackinger & Zeggini (2017)

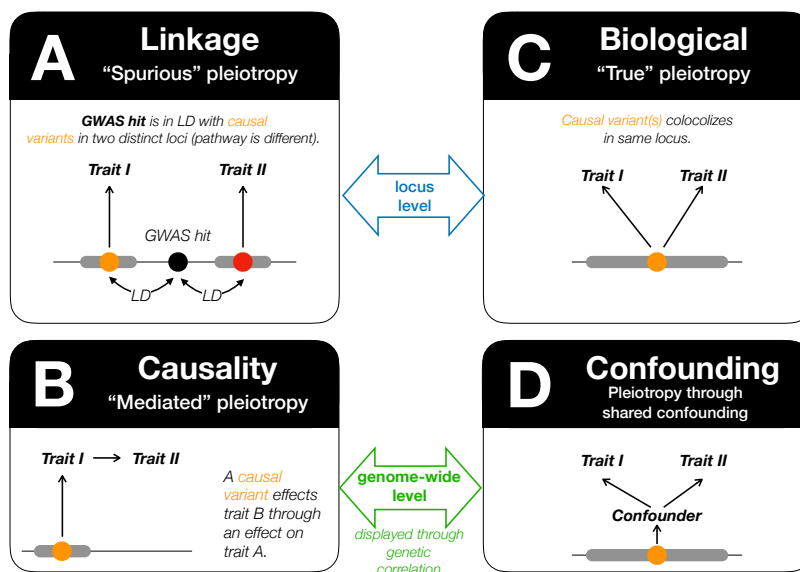


Figure 1.6: Different types of pleiotropy. When association signals of multiple phenotypes colocalise, this can have many reasons: **(A)** *Spurious pleiotropy*: A GWAS variant (black dot) tags two other variants: one (orange) that is linked to trait I and one (red), that is linked to trait II. **(B)** *Mediated pleiotropy*: The causal variant affects trait II through trait I. This type of pleiotropy can be detected through MR and genetic correlation. **(C)** *Biological pleiotropy*: the causal variant affects two traits through independent pathways. **(D)** *Pleiotropy through confounding*: genetic correlation is introduced by common confounder. (B) and (D): will display genetic correlation on a genome-wide level, (A) and (D): will display pleiotropy at a locus level.

Polygenic risk scores (PRS) can be utilised to detect forms of pleiotropy. PRS are an important concept in GWAS methodology, originally developed to predict an individual's disease risk. To build a PRS, SNP effects are estimated from a discovery sample, a score is build according to certain criteria and then applied in an independent sample. Eventually, such PRS could be used in a clinical setting to make genetic prediction of a disease for a single individual. However, currently the prediction accuracy is for many complex traits not high enough, so rather than making predictions, PRSs are used to indicate which individuals locate at the lower or the higher tail of the disease prediction distribution. Another application is relevant for pleiotropy. If an out-of-sample PRS is applied to individuals, and the genetic prediction is correlated with other phenotypes, then there might be some type of pleiotropy.

Pleiotropy

True biological pleiotropy²⁸ focuses on the influence that a *single* genetic marker has on multiple traits.

When a **single genetic variant** is causally implicated in more than one trait, this is called *pleiotropy* (or *true biological pleiotropy*

²⁸ See Hackinger and Zeggini (2017) for a comprehensive review of methods detecting different kinds of pleiotropy.

as there are other types, see Figure 1.6, C). The problem is, that true biological pleiotropy is hard to study. In reality, other types of pleiotropy are more common (Hackinger and Zeggini 2017), such as mediated and spurious pleiotropy, and pleiotropy through confounding (A, B & D in Figure 1.6).

Colocalisation with multiple phenotypes to dissect association signals can be done with a number of methods:

- **SCOPA** (Mägi et al. 2017) uses reverse regression on *individual-level data* (outcome is the genotype of an individual SNP and phenotypes are predictors). **META-SCOPA** is the analogous tool for *summary statistics*.
- **MultiPhen** (O'Reilly et al. 2012) models multiple phenotypes simultaneously using ordinal regression.
- Aschard et al. (2014) applies **principle component analysis (PCA)** on multiple traits.

However, using these tools, it is almost impossible to separate any scenarios from each other (because different scenarios can give rise to identical GWAS data).

Genetic correlation

Genetic correlation quantifies the extent to which genetic effects are shared between two traits. Contrary to local pleiotropy, this is a more systematic view of genome-wide “pleiotropy”.

The primary reason for genetic correlation is a common (heritable) *confounding* factor that can create a seemingly strong genetic correlation between the two traits (as in D in Figure 1.6). Another reason is *mediated pleiotropy*, where genetic variants are acting on a trait through another, intermediate trait, in a causal manner (as in B in Figure 1.6).

Methods that explore genetic correlation are closely linked to causal inference methods.

The growing number of consortia, the presence of semi-public large-scale genetic data, and the increase in publicly available GWAS results has made a variety of multi-trait analyses possible:

- Using **LDSC**, B. Bulik-Sullivan et al. (2015) estimated 276 genetic correlations among 24 traits.
- O'Connor and Price (2017) proposed a model in which a latent causal variable (**LCV**) mediates the genetic correlation between two traits.
- Lu et al. (2016) established a better understanding of the genetic basis and linking between anthropometric and cardiovascular traits by analysing *cross-phenotype associations*.
- Pickrell (2014) used a *hierarchical modelling approach* to analyse 18 human traits jointly (GWAS summary statistics and genetic annotation as input).

- Pickrell et al. (2015) performed a systematic search for genetic variants that *influence pairs of traits* (40 in total). Additionally, they inferred causal relationships between traits.
- **GSMR** and **SMR** (Zhu et al. 2016; Zhu et al. 2018; Porcu and others 2018) are multi-instrument MR methods based on summary statistic that aim to establish causal links from exposure to phenotype level.
- Beyond establishing the relationship between traits, analysing multiple traits at the same time can also **boost power** to detect relevant loci (e.g. Turley et al. (2018) or McDaid et al. (2017)). **MTAG** (Turley et al. 2018), is a method to analyse different study results with overlapping samples (therefore boosting power).

Having access to *individual-level data* has the additional advantage of estimating the trait-trait correlation directly, that otherwise needs to be estimated from other sources (Cichonska et al. 2016).

1.5.5 Data

The statistical methods mentioned in the previous section use either *individual data* or an aggregated form called *summary statistics*. A-C in Figure 1.7 illustrate the difference between individual and aggregated data in terms of information loss. Most importantly, because of privacy restrictions, the LD structure between markers cannot be retrieved anymore, once the data is aggregated and published, and thus has to be estimated from external data.

Below is a list of resources of genetic data, in summarised or individual form.

Individual-level data

This section refers to part (A) in Figure 1.7.

- **HapMap Project** phase III includes sequencing data from 1'397 individuals, 11 ancestry groups (The International HapMap Consortium 2003) → this database is not maintained anymore.
- **1000 Genomes Project** (1000 Genomes Project Consortium 2010): phase III contains sequencing data from 2'504 individuals, from 26 populations, for 84.4 million variants markers → public data.
- **UK10K** (Moayyeri et al. 2013; Boyd et al. 2013): sequencing data from 4000 individuals of European/British ancestry → restricted access.
- **UK Biobank** (Sudlow et al. 2015; UK Biobank Phasing and Imputation Documentation 2015): genotype and genotype imputed data for 500K individuals → restricted access.

Currently, 1000 Genomes Project and UK10K data are often used as **reference panels** for imputation (part (B) in Figure 1.7).

GWAS summary statistic results

This section refers to part (C) in Figure 1.7.

Adapted from Pasaniuc & Price (2017)

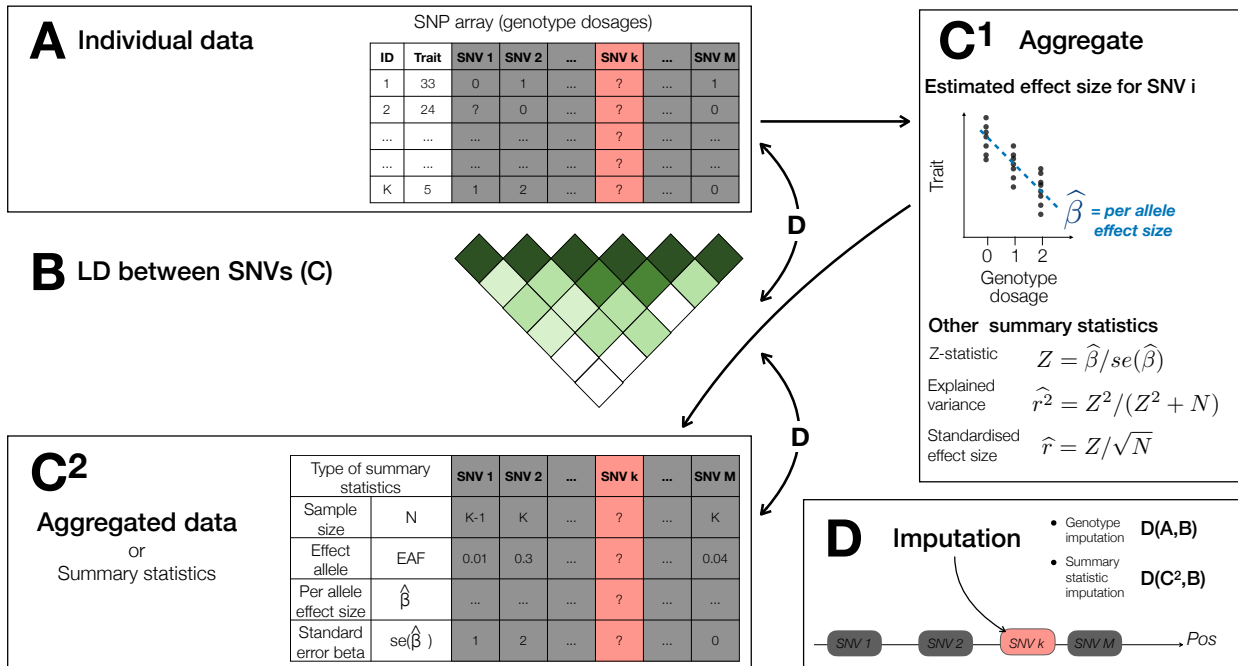


Figure 1.7: **Individual-level genotype data to summary statistics:** (A) Genotype data for K individuals and M markers. Note that for SNV k there is no data. (B) LD structure between SNVs estimated through the squared correlation. (C¹) Association summary statistics: estimating per allele effect sizes and the standard error. Using these two summary statistics, along with sample size (N) and MAF, we can derive other association summary statistics, such as Z-statistics (Z), the explained phenotypic variance (r^2) or the standardised effect size (r). (C²) Aggregated data form, also called *summary statistics*, listing sample size, effect allele frequency, effect size and its standard error. Summary statistics that quantify the association are called *association summary statistics*. (D) Imputation of the untyped SNV k as a function of genotype data and haplotype information ($D(A, B)$), or as a function of summary statistics and LD structure ($D(C^2, B)$). For both options, haplotype information and LD structure is estimated from external reference panels. See also Figure 1.8.

- **GIANT**, a consortium focusing on anthropometric traits has released all summary statistics of 15 large-scale meta-GWASs.
- The **Catalog of published GWAS studies** (Welter et al. 2014) provides a curated list of all GWAS results.
- **GRASP** (Leslie, O'Donnell, and Johnson 2014) provides a genome-wide repository of associations between SNPs and phenotypes.
- **UK Biobank results** (Abbott et al. 2017): GWAS summary statistics from genotype imputed UK Biobank data, including 337'000 individuals and 2'419 phenotypes. Based on this, **UKB phewas** takes a variant identifier as input and returns the association results for the most relevant phenotypes.
- **LD-Hub** (Zheng et al. 2016): a centralised database of summary-level GWAS results.
- **eQTL** (Westra et al. 2013): Cis- and trans-eQTLs results in whole blood samples, limited to FDR=0.5.
- **GTEx** (The GTEx Consortium 2013): eQTL summary statistics from over 20 different tissues.
- **PhenoScanner**: lookup of curated large-scale GWAS results, takes a variant identifier as input.
- **ExAC** (Lek et al. 2016): Summary statistics (allele frequencies) available through exome sequencing of 60'706 individuals (various disease-specific and population genetic studies, large-scale sequencing projects).
- Table 1 in Pasaniuc and Price (2017) lists resources.

Follow-up summary statistics

Performing methods such as the ones described in ‘Beyond GWASs’, also output summary statistics that can be accessed.

- **MR-base** (Hemani et al. 2016): a web application that displays the result of a systematically performed MR analyses on a number of traits, using over 1000 GWAS summary statistic results. **MR-base PheWas** takes a variant identifier as input, and returns the traits with relevant MR results as output.
- **LD-Hub** (Zheng et al. 2016): a web application to look up pre-run LDSC results.

1.6 Imputation methods

The methods described above (heritability estimation, fine-mapping, causal inference and multi-trait analysis) require (in parts) 1) large sample sizes, 2) the results to be harmonised in terms of markers, and some 3) full genome summary statistics. Both, 2) and 3) can be achieved by imputation (which will ultimately increase sample size too).

The intuition of all imputation methods is, that parts of the genome tend to be inherited together, which creates sets of correlated SNPs in close proximity (‘in linkage disequilibrium (LD)’), are also called *haplotypes*. Therefore, having information about a subset of SNPs and knowing the local LD structure, we can reconstruct the remaining SNPs with a certain confidence (see part (B) and (D) in Figure 1.7, and Figure 1.8).

There are two commonly used methods to infer the summary statistic of an unobserved genotype. The method of choice is **genotype imputation**. Imputing missing genotypes per individual involves a two-step procedure, after which common association tests can be performed at both, genotyped and imputed SNPs. An alternative solution is **summary statistic imputation**, that uses summary statistics resulting from a genotype-based GWAS and imputes the missing SNP estimates directly. Both imputation methods rely on external **reference panels** to facilitate **haplotype** or **LD estimation**.

In my thesis I focus on *summary statistic imputation*. Although *genotype imputation* is more accurate than *summary statistic imputation*, it is in many cases the only option to impute summary statistics. **Privacy, logistic or computational constraints make individual-level genotype data less likely to be shared**, therefore *genotype imputation* is oftentimes not feasible. For example, large consortia do meta-analyses on HapMap or 1000 Genomes imputed markers. But the reference panels are changing and growing rapidly. Therefore, in order to have the newest results, every cohort that is part of a consortium would need to re-impute with the newest reference panel, re-submit and re-analyse the meta-analysis. Another example where *summary statistic imputation*

comes in handy, are independently conducted downstream analyses on GWAS data, such as the methods mentioned above, where researcher need genome-wide association results at highly overlapping genomic resolution.

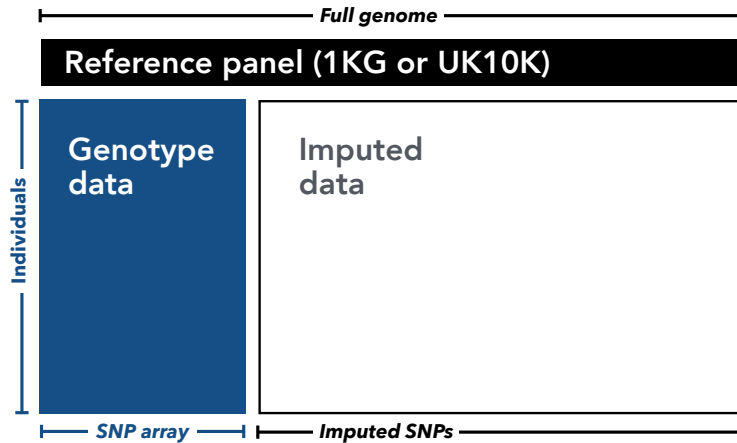


Figure 1.8: **Imputation with reference panels:** SNP arrays cover a relatively small fraction of genetic variation in the human genome, but are cost-effective, hence more individuals can be genotyped (blue). The power of using SNP arrays is, to integrate them with information from a relatively low number of densely sequenced individuals (black), leading to the genotype imputed data (black rectangle).

Imputation of genetic variants can be done for completely missing or only partially missing variants.

1.6.1 Genotype imputation

Genotype imputation models the missing genotypes for each individual. In brief, genotype imputation looks at each individual and assigns - with the help of similar (distantly related) individuals - a genotype dosage for a variant that was not genotyped. Having a complete data set available one can then run a GWAS and obtain summary statistics.

Genotype imputation²⁹ (Marchini and Howie 2010, Howie et al. (2012)) comes in different flavours and involves two steps:

1. *Pre-phasing*: estimating haplotypes for each individual within the GWAS sample in an iterative process (for example MaCH (Li et al. 2010), IMPUTE (Howie et al. 2012) or SHAPEIT (Delaneau et al. 2013)).
2. *Imputation of missing genotypes* into phased reference haplotypes using hidden Markov models (minimac (Fuchsberger, Abecasis, and Hinds 2015) or IMPUTE (Howie et al. 2012)).

IMPUTE is a software that is able to perform both steps.

Phased reference haplotypes (so called *template haplotypes*) have to be estimated each time a new reference panel is published.

Genotype imputation has a **high accuracy** for a **allele frequencies** down to 0.5%. Because it uses haplotypes, the accuracy depends on the population diversity (and the size) of the reference

²⁹ Introduction to genotype imputation: Marchini and Howie (2010).

panel (it is sufficient that the relevant haplotypes are present in the reference panel, but the overall allele frequency does not need to match the GWAS allele frequency).

Two downsides of genotype imputation are the **long computation time** and the **storage space** required. For example, for UK Biobank, genotype imputation would take 4200 CPU days (compared to 8.3 CPU days with summary statistic imputation), and storing imputed and compressed UK Biobank data for 500K individuals, requires 5 TB of space.

Michigan Imputation Server offers automated genotype imputation using the Haplotype Reference Consortium (2016).

1.6.2 Imputation using summary statistics

Summary statistic imputation (Pasaniuc et al. 2014) has been proposed as an **efficient solution** that only requires **summary statistics** and the **LD information** estimated from the latest reference panel to directly impute up-to-date meta-analysis summary statistics (Pasaniuc and Price 2017). Because *summary statistic imputation* uses summarised data as input, it is **not bounded to privacy restrictions** related to the use of individual-level data. Another advantage is its substantially lower computation time and storage space compared to genotype imputation.

Summary statistic imputation in the context of genomic data was **first described** by Wen and Stephens (2010), where they inferred allele frequencies for an untyped SNV, by a linear combination of observed allele frequencies. Lee et al. (2013) and Lee et al. (2014) then further extended the method to the application of linear regression estimates and a covariance matrix shrinkage depending on the reference panel size. Later, Pasaniuc et al. (2014) included a sliding window, which allowed partitioning of the genome into smaller pieces to facilitate imputation on a larger scale, and introduced a different shrinkage approach. Since then a few extensions have been published, that mainly concentrate on summary statistic imputation for admixed populations (Donghyung Lee, Bigdeli, et al. 2015; Donghyung Lee, Williamson, et al. 2015; Park et al. 2015), or including covariates (Xu et al. 2015).

Summary statistic imputation works through providing 1) summary statistics for a set of genotyped marks (called *tag SNVs*), and 2) the LD structure, described in part (B) and (C) in Figure 1.7.

For accurate inference, the current *summary statistic imputation* method makes a few **assumptions** that I partially addressed in my work.

- The **LD structure** reflects the correlation between Z-statistics (meaning, the LD structure was estimated from the same set of individuals).
 - LD structure is typically estimated from an external reference panel that might **misrepresent** the LD structure. *Approach:*

incorporate population (ad)mixture using the weighted LD structure of subpopulations in the reference panel (see summary).

- The reference panel size might be small, which (1) requires to **tune the shrinkage** to the correlation matrix (called λ), and (2) leads to imprecise estimation of low frequency variants ($\text{MAF} < 1\%$). *Approach: investigate shrinkage parameter λ (see discussion).*
- An unbiased estimation incorporates the **imputation quality**.
 - Without accounting for imputation quality, the summary statistic estimation is underestimated for variants with imputation quality < 1 . However, because of issues related to LD estimation (item before) imputation quality is imprecisely estimated.
 - *Approach: Improve the imputation quality towards a more accurate, yet fast to compute measure (see summary in the next chapter).*
- Sample size is **constant** over all tag SNVs. *Approach (by Aaron McDaid): account for varying sample size.*
- Imputed variants are a **linear combination** of tag variant. The imputation of summary statistics of an untyped SNV is essentially the linear combination of the summary statistics of the tag SNVs. Such a model cannot capture non-linear dependence between tag- and target SNVs, which is often the case for rare variants. *Future research direction: estimation of imputed SNVs as a non-linear function of tag SNPs.*

Other contribution to *summary statistic imputation* was to

- compare *summary statistic imputation* to *genotype imputation*,
- and test the utility of *summary statistic imputation* on a real case study on human height.

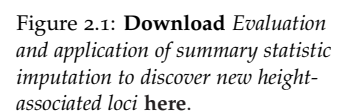
The following chapter lists my contribution to *summary statistic imputation* and other publications.

Results

2.1 Evaluation and application of summary statistic imputation

The manuscript entailed four parts:

- This manuscript is appended at the end of this thesis or can be



downloaded here.

2.2 Improving summary statistic imputation for mixed populations

Parts of the computation of *summary statistic imputation* (and its imputation quality) involve estimating the **LD structure** among SNPs in the GWAS population¹. However, by choosing an ad-hoc reference panel, the estimated LD structure often **misrepresents** the correlation structure in the GWAS, either caused by a mismatching or a too small reference panel. In fact, for *summary statistic imputation* to effectively work, the reference panel has to be large and matching in terms of population structure. With limited reference panel size and choice, and GWASs being highly heterogeneous in terms of ancestry, this is currently not possible.

In *Improved imputation of summary statistics for mixed populations* (Rüeger, McDaid, and Kutalik 2017) I addressed this problem with the standard reference panels at hand and approximating the diagonal of the LD matrix with minor allele frequencies reported in the GWAS. The method combines the **LD matrices of sub-populations** in the reference panel with weights w , so that the overall error (MSE) in the LD matrix is minimised. The weights w are determined for each genomic region separately.

For performance comparison, I used UK10K data that was up-sampled to 25'000 individuals and simulated phenotype, and compared the results our approach to existing methods. I observed a **variance-bias trade-off**, with too small reference panels having the MSE dominating by the variance, ultimately implying that optimisation of admixture is less relevant for small reference panels. One drawback of the method is, that allele frequencies are often not reported in a GWAS.

This manuscript² is appended at the end of this thesis or can be downloaded here.

2.3 Applications of summary statistic imputation

The publication *Rare and low-frequency coding variants alter human adult height* (Marouli and others 2017) by the GIANT consortium identified 120 new height associated loci (122 variants) (hereafter called *exome study*). Previously, there had been 697 variants discovered using HapMap imputed data of 253'288 individuals (hereafter called *HapMap study*) (Wood and others 2014). The *exome study* initially proposed 606 variants that had reached the exome-wide significance threshold. Some of these variants were located nearby *HapMap study* hits. Conditional analysis was used to **fine-map** the proposed 606 *exome study* and 697 *HapMap study* hits. To do so, the 606 exome variants were conditioned onto 697 previously discovered variants from the HapMap study using UK Biobank data. As



Figure 2.2: Download Improved imputation of summary statistics for mixed populations here.

¹ More precisely, it is the correlation among the Z-statistics that is approximated by the correlation among genetic variants.

² This manuscript is currently hosted on bioRxiv and has been submitted to Bioinformatics, where it is under revision.

For the article *Bayesian association scan reveals loci associated with human lifespan and linked biomarkers* (McDaid et al. 2017) I **provided** the first author with the **summary statistic imputation algorithm**, in order to enable fast imputation of all GWASs to the same set of SNVs.

Knowing that genetics plays a role in disease progression and treatment of HCV, I wanted to investigate what the **impact of genetic and non-genetic risk factors** was on FPR. As genetic risk factors, I used 10 SNPs known to affect HCV progression. I quantified the impact of a risk factor by estimating the **attributable fraction (AF)**. The AF is the fraction of cases that would be prevented if the risk factor could be eliminated. In the context of FPR, this is, the

Coding variants in new and known height loci
Many of the height-associated variants discovered in this study are located near common variants previously associated with height.

The enormous variation in human lifespan is in part due to a myriad of sequence variants, only a few of which have been revealed to date. Since many life-shortening events are related to diseases, we developed a Mendelian randomization-based method combining 52 disease-related GWAS studies to derive longevity priors for all HapMap SNPs. A Bayesian association scan revealed these priors, for parental age at death in the UK Biobank study ($n=379$) revealed 10 independent SNPs with significant associations. The average life expectancy discovery rate (FDR), eleven of them reached 1% (FDR) in five independent longevity studies combined; all but three are depleted of the life-shortening alleles in older Biobank participants. Further analyses revealed that brain expression levels of nearby genes (*RSMS*, *SLC7A1* and *CHRNA5*) might be causally implicated in longevity. Gene expression and cellular restriction experiments in model organisms confirm the conserved role for *RSMS* and *SLC7A1* in modulating lifespan.

Figure 2.5: **Download** Impact of common risk factors of fibrosis progression in chronic hepatitis C **here**.

fraction of fast progressors that could be turned into slow progressors. To estimate the AF, I used data from Swiss Hepatitis C Cohort Study, along with three additional cohorts to replicate the results.

The conclusion was, that most factors accelerating liver fibrosis progression in chronic hepatitis C are **unmodifiable** (age at infection, sex, route of infection, HCV genotype, *rs738409*, *rs4374383* and *rs910049*). Furthermore, differently from the Scottish cohort results, we did not observe a significant effect of alcohol consumption. One major difficulty was to avoid pitfalls in estimating the attributable fraction, as well as the lack of complete replication cohort data.

This manuscript can be downloaded [here](#).

In Gauthiez et al. (2017), I supported the first author with conducting a **meta-analysis** of polymorphisms influencing hepatitis C virus clearance.

Lastly, I calculated **power** for various Cox proportional hazard regression models for two publications of Swiss Transplant Cohort Study data (Wójtowicz, Lecompte, et al. 2015; Wójtowicz, Gresnigt, et al. 2015).

2.5 *Minor contributions to other publications*

Winkler et al. (2015) looked at the influence of age and sex stratification on heritability of body size and shape. Using GCTA (Yang et al. 2011), I calculated **heritability** of anthropometric traits in various age- and sex-groups in the CoLaus study (Firmann et al. 2008).

3

Discussion

In the introduction I presented **various statistical methods** that can be applied to genetic data — individual-level data or in a summarised form — to answer questions regarding heritability and predictability of complex traits, or to explore the underlying disease aetiology.

- Narrow-sense *heritability* — the proportion of phenotypic variance explained by SNPs — can be estimated in various ways from individual-level genotype data or from summarised data. Updated heritability estimations from GWAS results are key measures that guide and assess novel findings.
- *Fine-mapping* is used to identify causal variants in a genomic region associated with a complex trait. Fine-mapping methods are mostly based on summary statistics, Bayesian modelling, and functional annotation.
- Most *causal inference* methods are centred around Mendelian randomisation (MR) and extensions of it. Because two sample MR has been developed and more and larger GWASs have become publicly available, it has become easier to use reliable instrumental variables for MR.
- Finally, *multi-trait analyses* give insight into genetic correlation between traits and potential pleiotropy.

Note, that because effect sizes of genetic variants involved in complex traits are typically small, sample size is key; studies with **larger sample size** will have more statistical power.

These statistical methods often work with individual-level genotype data or with summary statistics combined with LD data¹ as input. However, methods relying on access to genetic data are limited by the analyst's access to cohort data, hence sample size is limited. The use of **summary statistics**, combined with information from **external reference panels** is a compromise to make up for the lack of individual level data.

Most publications that are centred around GWAS data use such methods. For example, Porcu and others (2018) used summary statistics-based MR to estimate the **causal impact of gene expression** on 43 complex traits. Doing so, Porcu *et al.* uncovered 2'277 putative genes causally associated with at least one complex trait,

¹ *Linkage disequilibrium (LD)* structure, the non-random association of SNPs, structure is estimated from an external reference panel, such as the 1000 Genomes Project Consortium (2010).

while also evaluating shared causal effects of gene expression on pairs of traits. In the case of Porcu *et al.*, SNP summary statistics (instrumental variables) need to be available for both sets of summary statistics: the eQTL study and the GWAS of the complex trait.

In another example Winkler *et al.* (2015) screened for **age- and sex-specific effects in BMI and WHRadjBMI** (weight-hip-ratio adjusted for BMI) via a genome-wide interaction meta-analysis. Formally, this is a G-by-E² design, where E is dichotomous, leading to four strata (men $\leq 50y$, men $> 50y$, women $\leq 50y$, women $> 50y$). The meta-analysis included $> 320'000$ from 114 studies with differing SNP panels (HapMap imputed, Illumina MetaboChip), yielding a common set of 2.8 million SNPs. SNPs were only included in the analysis if being available in at least half of the maximum sample size in all four strata, leading to SNPs not being tested or having lower sample size (thus lower power).

In fact, oftentimes, summary statistics-based methods require results for the **same set of SNPs** to have full power (e.g. a meta-analysis). To harmonise results, imputation methods are used.

Therefore, in order to answer questions regarding heritability and disease aetiology, not only is it important to have statistical methods available that use summary statistics as input, but also a methodology that **imputes summary statistics** for unmeasured SNPs.

My PhD evolved around methods that integrate GWAS and other -omics data with statistical methods. The results are described in the previous chapter. My main occupation was the **improvement of summary statistic imputation**. In parallel, I also ran GWASs, other statistical analyses and contributed to collaborations. In this discussion, I will therefore focus on *summary statistic imputation*, the future of GWAS, and give an outlook into *future work* that involves integrating GWAS results and *drugbank* data.

3.1 Summary statistic imputation: limitations and future work

In my two publications (Rüeger, McDaid, and Kutalik 2018; Rüeger, McDaid, and Kutalik 2017) I proposed an improved *summary statistic imputation*³ method (improved imputation quality and optimised assembly of the LD matrix). I also compared *summary statistic imputation* to *genotype imputation*, identified groups of genetic variants that are hard to impute, and demonstrated in a case study the utility of *summary statistic imputation*.

3.1.1 Estimation error of LD structure

Estimating the correlation matrix between the Z-statistics is one of the major challenges in *summary statistic imputation*⁴. In reality, the correlation structure between SNPs in the GWAS is approximated through the correlation between the SNPs from external reference panels. The accuracy of imputation is reduced when the **LD struc-**

² G-by-E: Gene-environment interaction

³ *Summary statistic imputation* is a statistical method that is used to impute the summary statistic (often the Z-statistic) of an untyped SNP by combining summary statistics of typed SNPs with LD information. More specifically, Z-statistics are modelled to follow a multivariate normal distribution. The correlation between typed and untyped Z-statistics is estimated through the correlation between SNPs estimated from an external reference panel (LD).

⁴ Estimating the LD structure with the correlation matrix is a general problem to any method that uses GWAS summary data combined with reference panel data. Deng and Pan (2018) pointed out an increased type 1 error rate in the context of approximate conditional and joint analysis.

ture between typed and imputed SNPs is **misspecified**.

Ideally, the reference panel population needs to match the GWAS population. Even if this is the case, the reference panel needs to be sufficiently large in order not to run into sparsity problems when inverting the LD matrix (which is part of the summary imputation algorithm). The current situation of reference panels requires careful considerations of the choice of sub-populations, and shrinkage of the LD matrix in order to reduce MSE at the cost of introducing some bias.

Ultimately, imprecise LD estimations lead to more false positives and false negatives, and also to imprecise imputation quality estimation.

I addressed the issue of estimating the LD matrix in two ways:

First, I searched for an **adaptive shrinkage** method of the **LD matrix**. Shrinkage methods help to process correlation matrices that describe a large number of variables from only a few samples ($n \gg p$ problem). As a simple example consider a correlation matrix of 2000 neighbouring SNPs, estimated from 100 individuals. SNPs at the very start and at the very end of the range are likely in low LD with each other, but still display some non-zero correlation due to estimation error, which can be inflated when inverting the matrix.

By multiplying the off-diagonal values in the LD matrix with a scalar ($0 \leq \lambda \leq 1$), the matrix becomes invertible. Any $\lambda < 1$ will make the correlation matrix invertible, but a stronger shrinkage can reduce estimation error. Choosing the optimal λ is key to keep the estimation error low. For example, λ can be applied as a function of the reference panel size n : $\lambda = 2/\sqrt{n}$. I also worked on an alternative approach where the shrinkage parameter would change according to the underlying local genetic architecture of each region, however, was not successful. I hypothesised that optimal shrinkage depends on local LD structure, and sought to optimise λ for each genomic region using the effect sizes of tag SNVs as training data set in a leave-one-out fashion. When looking at null variants, however, maximum shrinkage ($\lambda = 1$) usually leads to the smallest MSE. Therefore, when looking at a specific genomic region with a mixture of null and associated SNVs, the selected λ will be shifted towards 1 and shrink the estimation of associated SNVs towards 0, which is not ideal.

Second, in the article Rüeger, McDaid, and Kutalik (2017) I showed how **imputation accuracy changes** according to **reference panel composition** and **reference panel size**. My results imply that simply enforcing allele frequencies between the GWAS and reference panel to match might decrease the bias of imputing the summary statistic but increase the variance to a much larger extent. This phenomenon is known as the **variance-bias trade-off**. The MSE of an estimate (in this case the imputed summary statistic) is composed of the squared bias and the variance. The goal is a minimal MSE. Parameter configurations (such as the reference panel

composition) can increase or decrease the contribution of the bias and the variance term, and as such the MSE.

To improve imputation of admixed or mixed GWASs, access to larger and more diverse reference panels are needed.

3.1.2 *Improve the estimation of imputation quality*

Imputation quality is defined as the **squared correlation** r^2 between the imputed and true genotypes. In reality, this is estimated through the variance of the imputed version of the SNP divided by the variance of the true underlying variance (which can be estimated from the reference panel) \hat{r}^2 of the tag SNPs. Therefore, the imputation quality can vary from 0 to 1, with $\hat{r}^2 = 1$ indicating perfect imputation.

Because \hat{r}^2 is **estimated from an external reference panel**, it suffers from similar problems to the ones mentioned above (reference panel size and composition). I tackled this problem by looking at variations of the classical \hat{r}^2 (e.g. using ridge regression for tag SNP selection or by accounting for sample size and the effective number of variants).

Additionally, \hat{r}^2 — being **simply the tag-ability of a SNP** — does not incorporate any of the factors that might decrease confidence in an imputed summary statistic (such as mismatching reference panel, the impact of the shrinkage parameter λ , number of tag SNPs or variable missingness among tag SNPs). This last topic has not been explored yet.

3.1.3 *Summary statistic imputation before meta-analysis*

To date, most published GWASs from consortia are meta-analyses based on HapMap imputed genotype data. **Updating genotype imputed data to newer reference panels** (such as 1000 Genomes Project Consortium (2010) or Haplotype Reference Consortium (2016)) is cumbersome for contributing cohorts, often leading to **partial** contribution, hence only a few consortia have done this (e.g. Early Growth Genetics Consortium (Horikoshi et al. 2016), CKDGEN Consortium (Gorski et al. 2017)).

This problem could be avoided (and sample size increased) by using *summary statistic imputation* instead. For example, cohorts could provide summary statistics on a varying genomic resolution. The consortium could then first **impute each single cohort summary statistics** to a common set of SNPs (and using an appropriate reference panel), before meta-analysing all cohort summary statistics. Imputing each single cohort *summary statistics* before meta-analysing them is important for imputation accuracy, as different cohorts require different settings (e.g. choice of reference panel). This is a topic that could be explored in the future.

3.1.4 Software implementation

I believe that in order for *summary statistic imputation* to be used widely as an alternative to genotype imputation, an **easy-to-use software implementation** must be available. People developing statistical methods are often capable of programming and can easily implement methods from other researchers. But others, with an expertise in a different domain than programming and statistical genetics, will likely struggle with an implementation, therefore not using the algorithm and turning to an easier solution instead.

Currently, there are various different implementations of summary statistic imputation (e.g. ImpG-Summary). These methods are **not straightforward to use** and **computationally very slow**.

This is why, I and Aaron McDaid, a former colleague, aimed for a more flexible *summary statistic imputation* implementation. Although I already had an implementation in R, Aaron opted for a C++ implementation, allowing **fast computation** and **command line operation**. Meanwhile, I was responsible for the conceptual planning and design of the tool.

More specifically, my aims in terms of **functionality** were:

- to use commonly used file formats as input (GWAS output such as that from PLINK, QuickTest, METAL or SNPTEST),
- to provide a simple output as a text format,
- to include sanity checks (such as re-imputing tag SNPs),
- to allow flexibility in terms of SNP identifiers,
- to have the user semi-guided in terms of parameter choice (e.g. shrinkage parameter λ is set to $2/\sqrt{n}$, but can be adapted to any value between 0 and 1),
- to compute the correlation matrices on the fly (after the reference panel has been downloaded),
- to implement a my new version of imputation quality and account for variable missingness,
- to allow quick imputation of single SNPs.

Having the software on **Github**, I hope that future improvements of *summary statistic imputation* can be directly implemented by others.

3.2 Key resources for future GWASs

In this section I list elements that, if improved, would best facilitate advancement of GWAS and could therefore ensure a thriving future of methods using association summary statistics⁵.

Large & diverse reference panels: To improve imputation of admixed or mixed GWASs, access to larger reference panels would be needed. For example, Haplotype Reference Consortium (2016) could release LD matrices, which could then be used as input for *summary statistic imputation* and similar methods.

Mutliethnic GWASs: Conducting GWASs in non-European individuals. So far, GWASs have predominantly focussed on samples

⁵ The concept of GWASs has its own challenges (Marigorta et al. 2018).

with European ancestry, and have therefore missed a substantial portion of the genetic variation that is present in the human population. According to Popejoy and Fullerton (2016), in 2016 only 19% of the GWAS in the GWAS Catalog (Welter et al. 2014) were non-European and 14% were of Asian descent. Meaning, the remaining 5% were shared among GWASs of individuals in African, Hispanic & Latin American, Pacific Islander, Arab & Middle Eastern, and indigenous descent. This has implications on (non-)transferability of GWAS results, for example in the context of polygenic risk scores (Martin et al. 2017; Kim et al. 2017).

Public data: To boost power with *summary statistic imputation* and other *summary statistic* based methods, GWAS results must be publicly easily accessible, and include information for all tested SNPs regarding: effect size, standard error of the effect size, sample size, effect allele, and effect allele frequency.

More -omics data: Most GWASs focus on SNP data. But exploring other -omics data as outcome and predictors (e.g. transcriptome, proteome, metabolome) will be needed for new discoveries.

Rare variant detection: Genotype imputation and summary statistic imputation both have limitations to impute rare variants. Therefore, specially designated sequencing studies are the only option to explore rare variants. Other genetic variation than SNVs (e.g. copy number variation, sex chromosomes) will also be needed. In principle, as long as a correlation structure is provided, *summary statistic imputation* can be performed for other genetic variation.

3.3 Future work on translational GWASs

To conclude this discussion I would like to repeat the overarching goal that was stated in the beginning of the introduction: applying statistical methods to genetic data to **identify the genetic risk factors underlying complex diseases**, which then could lead to improved treatment, prevention, diagnosis and prognosis.

So far, I have discussed statistical methods that analyse genetic data to identify key genes, SNPs or other genetic actors. However, I have not elaborated on how to translate these findings into better therapies for patients. In this section I will present a method that connects GWAS and MR findings with a pharmacological database to **enable repositioning existing drugs**.

A few GWAS results have been validated using knockout models in animals or in-vitro experiments. These approaches are expensive and not applicable in a high-throughput manner. In the past there have been various attempts to define the druggable part of the genome by linking GWAS results to drug targets in a systematic way (Finan et al. 2017; Gaspar and Breen 2017). However, using GWAS results only, these studies were underpowered and the signal seems to be weak.

Recently, **two sample MR methods** have been developed that combine GWAS results with eQTL data, leading to a **list of putative**

causal genes for a number of complex traits (Zhu et al. 2016). Such a well-powered study was performed by my colleague Eleonora Porcu (Porcu and others 2018), which estimated gene expression-trait associations in 43 traits using an extended MR technique. At the same time, a web database called **drugbank.ca** (Wishart et al. 2018) contains gene target information for over 7'000 drugs. I combined the MR results with information of drug-gene pairs from the drugbank.

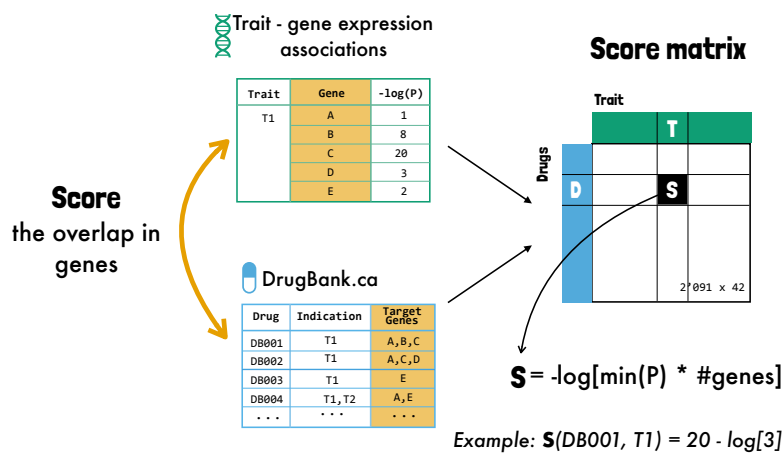
My goal was:

- To validate the MR study results. Do drugs for specific diseases tend to target genes whose expression is causally linked to those diseases?
- To understand why some of the identified genes overlap with the trait-specific treatment of the complex traits but others do not.
- To evaluate repositioning of drugs.

3.3.1 Method

From my colleague I received a dataset with three columns (gene, trait, and the association p -value) and 603'404 rows.⁶

I then combined all 603'404 causal effect estimates with known drug target genes using the drugbank database. For this, I developed a **score** to quantify for each drug-trait pair how well the drug targets correspond to the set of genes causally implicated for the given trait. Figure 3.1 illustrates this with a fictional example. This step provided me with a matrix of 89'913 trait-drug scores (2'091 drugs, 43 traits).



⁶ The association was estimated with an extended MR approach to whole blood expression of 15'985 genes (exposure) and 43 traits/diseases. This yielded 603'404 causal effect estimates (of which 5'009 are significantly non-zero).

Figure 3.1: **Combining MR and drugbank info into a score matrix:** These two datasets are a schematic illustration of the MR results (LSH-top) and the drugbank database (LHS-bottom). For any trait-drug combination, I calculated a score (RHS). As an example, we can calculate the overlap of the gene targets of drug DB001 and the gene-trait association of trait T1: Drug DB001 has gene A, B, C as gene targets. These genes have $-\log_{10}(P)$ -values of 1, 8 and 20. We can distil this info into a score (e.g. sum of the $-\log_{10}(P)$ -values).

Next, I defined a set of drugs for each complex trait. I call these drugs *trait-specific drugs* (and drugs not used to treat the trait *trait-*

unspecific drugs). Because I manually curated these looking up indications, I did this only for a limited number of diseases (below, I will list results for *total cholesterol*).

Finally, I analysed the score matrix in two ways. **Approach (A)** compares the scores for a given trait between **trait-specific drugs** and **trait-unspecific drugs**. If scores of trait-specific drugs have on average higher scores, this means that genes detected through MR are specifically related (target/transporter) to the trait-specific drugs. Furthermore, trait-unspecific drugs that have high scores can potentially be repositioned for this trait. Figure 3.2 shows a schematic view.

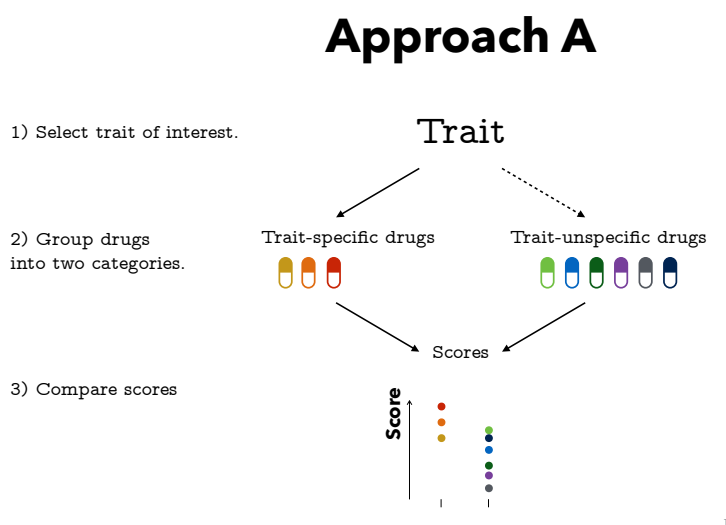


Figure 3.2: **Approach A - comparing drugs:** For a given trait I compared the scores of trait-specific drugs (LHS) and trait-unspecific drugs (RHS).

Approach (B) focuses on **trait-specific drugs** and compares the scores to other traits, see Figure 3.3 for an illustration. If scores of trait-specific drugs rank high, this means that the MR results are highly trait specific (in terms of drugs).

3.3.2 Results

As a showcase, I will present the results for *total cholesterol* (TC).

In approach (A) I analysed whether the scores of TC-specific drugs are enriched in MR results compared to TC-unspecific drugs scores. Figure 3.4 shows that the scores of TC drugs are indeed enriched (one-sided Wilcoxon rank sum test, $P = 5 \times 10^{-5}$), pointing to an agreement between the trait-specific drugs and the MR results. There are a number of TC-unspecific drugs that rank higher than the maximum score in TC-specific drugs. For example *Olsalazine* (an anti-inflammatory drug used in the treatment of the intestines), *Bendroflumethiazide* (a high-blood pressure treatment) or *Cetuximab* (cancer treatment). By understanding why these TC-unspecific drugs are in such high agreement with the MR results in TC, and

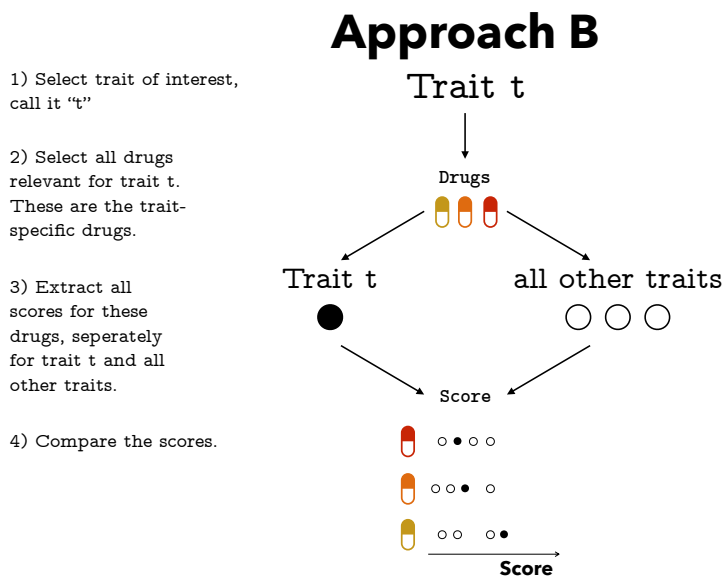


Figure 3.3: **Approach B - comparing traits:** For a specific trait "t" I extracted all trait-specific drugs, and then compared their scores for that specific trait "t" (solid dots) to all other traits (circles).

investigating their mechanism of action (including side effects), some TC-unspecific drugs could potentially be candidates for high cholesterol treatment.

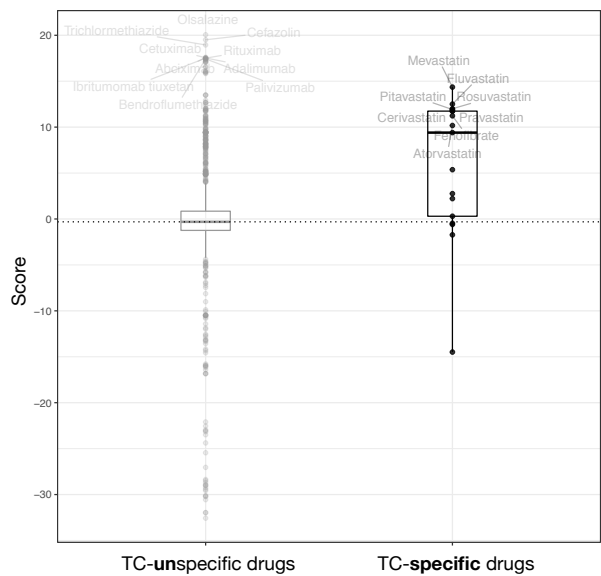


Figure 3.4: **Approach A in Total Cholesterol:** Y-axis shows the scores, each point is a drug, with a boxplot overlayed. LHS displays TC-unspecific-drugs, the RHS are the TC-specific drugs. One-sided Wilcoxon rank sum test ($P = 5 \times 10^{-5}$)

In approach (B) I analysed whether TC-specific drugs are enriched in MR results in TC, compared to other traits. Figure 3.5 shows that for 8 out of 17 drugs, scores for TC ranks first or second ($P = 7.6 \times 10^{-7}$).

3.3.3 Limitations

This work-in-progress method has a number of limitations.

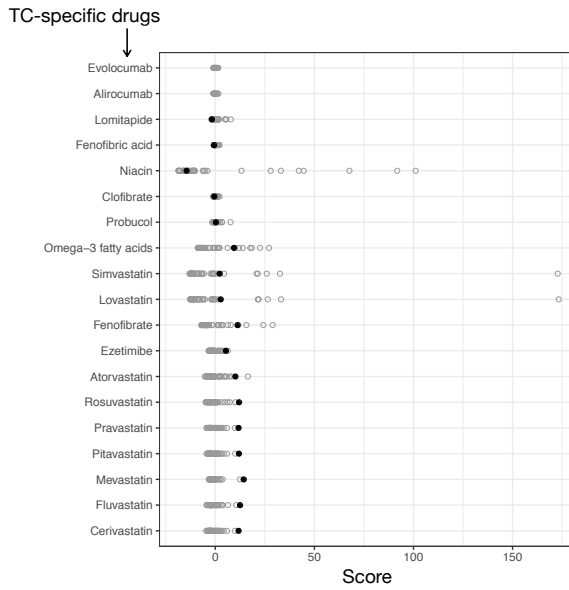


Figure 3.5: **Approach B in Total Cholesterol:** The x-axis shows scores, the y-axis the TC-specific drugs. Each dot is one trait, with TC as a solid dot. Drugs are ordered according to the ranking of the black dot (score for TC). Binomial test rank 1st or 2nd (#successes=8, #trials=17): $P = 7.6 \times 10^{-7}$.

Approach (A) works best with a **large number of drugs**. For example, for amyotrophic lateral sclerosis (ALS) there are currently only two drugs listed in the drugbank. However, diseases such as ALS precisely need drug discoveries.

Some diseases display **symptoms** that are **highly general** and for which a vast number of drugs is available. For example, rheumatoid arthritis (RA) has 164 drugs enlisted. This makes approach (B) difficult to work, as many diseases will have the same drugs as treatment. Excluding diseases with similar treatments, or accounting for such diseases might be a solution to that.

There are two factors that increase uncertainty of the results of this method. First, the **gene targets listed** in the drugbank were established with various methods of differing reliability. Second, the MR results are **limited to peripheral blood**, thus not generalisable to other (important) tissues. Some known drug targets are not part of the results. Other tissue was analysed, but sample size restricts statistical power.

Most importantly, for this method to work, establishing a general and widely applicable **disease-indication-drug network** is key. Such a network could be established through using ATC codes.

3.4 Conclusion

With my work on *summary statistic imputation* I have helped to improve linking incomplete GWAS summary statistics with follow-up methods. I have emphasised the need of publicly available summary statistics for a range of traits, as well as large and diverse reference panels. In combination with statistical methods, these two data sources could help to unravel the biology of diseases currently relevant in public health.

For example, besides better treatment, a key factor in improving life quality for sick individuals are **early stage diagnoses** of diseases that require well-timed therapeutic interventions (e.g. multiple sclerosis). Future efforts in solving such riddles in genetic epidemiology involve the usual suspects, such as the exploration of rare variation with more sequencing studies, deep phenotyping and cohorts from diverse ethnicity.

What I think will be highly relevant in the future are **public health intervention** that tackle highly complex and prevalent diseases such as obesity. Understanding how genetic and environmental risk factors related to obesity are causally linked to each other can for example be done with deep molecular phenotyping that can help to bridge the connection between genetic variants and obesity. Identifying the responsible (actionable) risk factors for obesity is currently far from being solved; although statistical methods are being developed that can handle more complex models across many different populations and environments.

Another important factor for future genetic epidemiology is how data is collected. Currently, genetic and phenotypic data is in the hands of researcher (e.g. cohort studies) or companies (e.g. genetic testing companies). This setup has one principle limitation: the data “owners” are in charge of giving researchers access to the data - and not the individuals, that provided the data. New data collection initiatives let an **individual own and control its data** (e.g. Nebula Genomics). This is done with new technologies that enhance data privacy, and as such potentially increase the size of genomic data.

Bringing the attention back to the individuals that provided the data, is also part of the last point that I believe will become more relevant in the upcoming years: **returning genetic research results back to individuals**. Individuals that become part of a cohort or data collection must sign a consent form (see an example [here](#)). Part of such a consent form is whether the individual wishes to have the genetic risk reported back to them. An individual has also the right “not to know”. So far, mostly rare diseases were reported back, ideally through a genetic counselling. However, for complex diseases reporting back the disease risk proved more difficult, as these diseases needed more investigation to derive reliable risk predictors. For well-studied complex diseases, we can now report back an individual’s disease risk through polygenic risk scores derived from large-scale GWASs. The Estonian Biobank (EGCUT) is making a step into this direction. A new pilot programme returns an individual risk information to EGCUT participants for certain diseases, for example type 2 diabetes.

References

- 1000 Genomes Project Consortium. 2010. "A map of human genome variation from population-scale sequencing." *Nature* 467 (7319): 1061–73. doi:10.1038/nature09534.
- Abbott, Liam, Verner Anttila, Krishna Aragam, Jon Bloom, Sam Bryant, Claire Churchhouse, Joanne Cole, et al. 2017. "Rapid GWAS of thousands of phenotypes for 337'000 samples in the UK Biobank." <http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank>.
- Aschard, Hugues, Bjarni J. Vilhjálmsson, Nicolas Greliche, Pierre Emmanuel Morange, David Alexandre Trégouët, and Peter Kraft. 2014. "Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies." *American Journal of Human Genetics* 94 (5): 662–76. doi:10.1016/j.ajhg.2014.03.016.
- Astle, William J, Heather Elding, Tao Jiang, Willem H Ouwehand, Adam S Butterworth, Nicole Soranzo, William J Astle, et al. 2017. "The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease Resource The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease." *Cell*, 1415–29. doi:10.1016/j.cell.2016.10.042.
- Baum, Christopher F, Mark E. Schaffer, and Steven Stillman. 2003. "Instrumental variables and GMM: Estimation and testing." *Stata Journal* 3 (1): 1–31. doi:The Stata Journal.
- Benner, Christian, Chris C A Spencer, Aki S Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. 2018. "Genetics and population analysis FINEMAP : efficient variable selection using summary data from genome-wide association studies." *Bioinformatics* 32 (May): 1493–1501. doi:10.1093/bioinformatics/btw018.
- Bowden, Jack, George Davey Smith, Philip C Haycock, and Stephen Burgess. 2015. "Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator." *Genetic Epidemiology*. doi:10.1002/gepi.21965.
- Boyd, Andy, Jean Golding, John Macleod, Debbie A. Lawlor, Abigail Fraser, John Henderson, Lynn Molloy, Andy Ness, Susan Ring, and George Davey Smith. 2013. "Cohort profile: The 'Children of the 90s'-The index offspring of the avon longitudinal study of parents and children." *International Journal of Epidemiology* 42 (1): 111–27. doi:10.1093/ije/dyr207.
- Boyle, Evan A, Yang I Li, and Jonathan K Pritchard. 2017. "Perspective An Expanded View of Complex Traits : From Polygenic to

- Omnigenic." *Cell* 169 (7): 1177–86. doi:10.1016/j.cell.2017.05.038.
- Bulik-Sullivan, B, H Finucane, V Anttila, A Gusev, F Day, P Loh, ReproGen Consortium, et al. 2015. "An atlas of genetic correlations across human diseases and traits." *Nature Genetics* 47 (11): 1236–41. doi:10.1038/ng.3406.
- Bulik-Sullivan, Brendan K, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. 2015. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies." *Nature Genetics* 47 (3): 291–95. doi:10.1038/ng.3211.
- Burgess, Stephen, Adam Butterworth, and Simon G. Thompson. 2013. "Mendelian randomization analysis with multiple genetic variants using summarized data." *Genetic Epidemiology* 37 (7): 658–65. doi:10.1002/gepi.21758.
- Chen, Wenan, Beth R. Larrabee, Inna G. Ovsyannikova, Richard B. Kennedy, Iana H. Haralambieva, Gregory A. Poland, and Daniel J. Schaid. 2015. "Fine mapping causal variants with an approximate bayesian method using marginal test statistics." *Genetics* 200 (3): 719–36. doi:10.1534/genetics.115.176107.
- Cichonska, Anna, Juho Rousu, Pekka Marttinen, Antti J. Kangas, Pasi Soininen, Terho Lehtimäki, Olli T. Raitakari, et al. 2016. "MetaCCA: Summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis." *Bioinformatics* 32 (13): 1981–9. doi:10.1093/bioinformatics/btw052.
- Collins, F. S., E. S. Lander, J. Rogers, and R. H. Waterson. 2004. "Finishing the euchromatic sequence of the human genome." *Nature* 431 (7011): 931–45. doi:10.1038/nature03001.
- Davey Smith, George, and Shah Ebrahim. 2003. "'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease?" *International Journal of Epidemiology* 32 (1): 1–22. doi:10.1093/ije/dy070.
- Delaneau, Olivier, Bryan Howie, Anthony J. Cox, Jean François Zagury, and Jonathan Marchini. 2013. "Haplotype estimation using sequencing reads." *American Journal of Human Genetics* 93 (4): 687–96. doi:10.1016/j.ajhg.2013.09.002.
- Deng, Yangqing, and Wei Pan. 2018. "Improved Use of Small Reference Panels for Conditional and Joint Analysis with GWAS Summary Statistics." *Genetics*. doi:10.1534/genetics.118.300813.
- Dermitzakis, Emmanouil T. 2008. "From gene expression to disease risk." *Nature Genetics* 40 (5): 492–93. doi:10.1038/ng0508-492.
- Didelez, Vanessa, and Nuala a. Sheehan. 2007. "Mendelian randomization as an instrumental variable approach to causal inference." *Statistical Methods in Medical Research* 16 (4): 309–30.

doi:10.1177/0962280206077743.

- Finan, Chris, Anna Gaulton, Felix A. Kruger, R. Thomas Lumbers, Tina Shah, Jorgen Engmann, Luana Galver, et al. 2017. "The druggable genome and support for target identification and validation in drug development." *Science Translational Medicine* 9 (383): 1–16. doi:10.1126/scitranslmed.aag1166.
- Finucane, Hilary K, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verner Anttila, et al. 2015. "Partitioning heritability by functional annotation using genome-wide association summary statistics." *Nature Genetics* 47 (11): 1228–35. doi:10.1038/ng.3404.
- Firmann, Mathieu, Vladimir Mayor, Pedro Marques Vidal, Murielle Bochud, Alain Pécoud, Daniel Hayoz, Fred Paccaud, et al. 2008. "The CoLaus study : a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome." *BMC Cardiovascular Disorders* 11: 1–11. doi:10.1186/1471-2261-8-6.
- Fisher, Ronald A. 1918. "XV.—The Correlation Between Relatives on the Supposition of Mendelian Inheritance." *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* 52 (2). Royal Society of Edinburgh Scotland Foundation: 399–433.
- Fuchsberger, Christian, Gonçalo R. Abecasis, and Davis A. Hinds. 2015. "Minimac2: Faster genotype imputation." *Bioinformatics* 31 (5): 782–84. doi:10.1093/bioinformatics/btu704.
- Gaspar, H. A., and G. Breen. 2017. "Drug enrichment and discovery from schizophrenia genome-wide association results: An analysis and visualisation approach." *Scientific Reports* 7 (1): 1–9. doi:10.1038/s41598-017-12325-3.
- Gauthiez, Emeline, Ines Habfast-Robertson, Sina Rüeger, Zoltan Kutalik, Vincent Aubert, Thomas Berg, Andreas Cerny, et al. 2017. "A systematic review and meta-analysis of HCV clearance." *Liver International* 37 (10): 1431–45. doi:10.1111/liv.13401.
- Ge, Dongliang, Jacques Fellay, Alexander J. Thompson, Jason S. Simon, Kevin V. Shianna, Thomas J. Urban, Erin L. Heinzen, et al. 2009. "Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance." *Nature* 461 (7262): 399–401. doi:10.1038/nature08309.
- Gorski, Mathias, Peter J. Van Der Most, Alexander Teumer, Audrey Y. Chu, Man Li, Vladan Mijatovic, Ilja M. Nolte, et al. 2017. "1000 Genomes-based metaanalysis identifies 10 novel loci for kidney function." *Scientific Reports* 7 (February): 1–11. doi:10.1038/srep45040.
- Guo, Jing, Yang Wu, Zhihong Zhu, Zhili Zheng, Maciej Trza-

- skowski, Jian Zeng, Matthew R Robinson, Peter M Visscher, and Jian Yang. 2018. "Shaped By Natural Selection in Humans." *Nature Communications*, 1–9. doi:10.1038/s41467-018-04191-y.
- Hackinger, Sophie, and Eleftheria Zeggini. 2017. "Statistical methods to detect pleiotropy in human complex traits → **Review**." doi:10.1098/rsob.170125.
- Haplotype Reference Consortium. 2016. "A reference panel of 64,976 haplotypes for genotype imputation." *Nature Genetics* 48 (10). doi:10.1038/ng.3643.
- Hemani, Gibran, Jie Zheng, Kaitlin H Wade, Charles Laurin, Benjamin Elsworth, Stephen Burgess, Jack Bowden, et al. 2016. "MR-Base: A Platform for Systematic Causal Inference Across the Phenome Using Billions of Genetic Associations." *bioRxiv*. doi:10.1101/078972.
- Horikoshi, Momoko, Robin N. Beaumont, Felix R. Day, Nicole M. Warrington, Marjolein N. Kooijman, Juan Fernandez-Tajes, Bjarke Feenstra, et al. 2016. "Genome-wide associations for birth weight and correlations with adult disease." *Nature* 538 (7624): 248–52. doi:10.1038/nature19806.
- Hormozdiari, Farhad, Anthony Zhu, Gleb Kichaev, Chelsea J-T Ju, Ayellet V Segrè, Jong Wha J Joo, Hyejung Won, et al. 2017. "Widespread Allelic Heterogeneity in Complex Traits." *American Journal of Human Genetics* 100 (5): 789–802. doi:10.1016/j.ajhg.2017.04.005.
- Howie, Bryan, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R. Abecasis. 2012. "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing." *Nature Genetics* 44 (8): 955–59. doi:10.1038/ng.2354.
- Innes, Hamish A, Sharon J Hutchinson, Stephen Barclay, Elaine Cadzow, John F Dillon, Andrew Fraser, David J Goldberg, Peter R Mills, Scott A McDonald, and Judith Morris. 2010. "Quantifying the Fraction of Cirrhosis Attributable to Alcohol Among Chronic Hepatitis C Virus Patients: Implications for Treatment Cost-Effectiveness." *Hepatology*, no. 1: 451–60. doi:10.1002/hep.26051.
- Kichaev, Gleb, Megan Roytman, Ruth Johnson, Eleazar Eskin, Sara Lindström, Peter Kraft, and Bogdan Pasaniuc. 2017. "Improved Methods for Multi-Trait Fine Mapping of Pleiotropic Risk Loci." *Bioinformatics* 33 (2): 248–55. doi:10.1093/bioinformatics/btw615.
- Kim, Michelle S, Kane P Patel, Andrew K Teng, Ali J Berens, and Joseph Lachance. 2017. "Ascertainment bias can create the illusion of genetic health disparities." *bioRxiv*, 195768. doi:10.1101/195768.
- Klein, Robert J, Caroline Zeiss, and Emily Y Chew. 2005. "Complement Factor H Polymorphism in Age-Related Macular Degener-

- ation." *Science*, no. April: 385–89.
- Lamparter, David, Daniel Marbach, Rico Rueedi, Zoltán Kutalik, and Sven Bergmann. 2016. "Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics." *PLoS Computational Biology* 12 (1): 1–20. doi:10.1371/journal.pcbi.1004714.
- Lee, D., V. S. Williamson, T. B. Bigdeli, B. P. Riley, a. H. Fanous, V. I. Vladimirov, and S.-a. Bacanu. 2014. "JPEG: a summary statistics based tool for gene-level joint testing of functional variants." *Bioinformatics* 31 (8). doi:10.1093/bioinformatics/btu816.
- Lee, Donghyung, T Bernard Bigdeli, Vernell S Williamson, Vladimir I Vladimirov, P Riley, Ayman H Fanous, and Silviu-alin Bacanu. 2015. "Genome Analysis Dismix : Direct Imputation of Summary Statistics for Unmeasured Snps from Mixed Ethnicity Cohorts." *Bioinformatics*. doi:10.1093/bioinformatics/btv348.
- Lee, Donghyung, T. Bernard Bigdeli, Brien P. Riley, Ayman H. Fanous, and Silviu Alin Bacanu. 2013. "DIST: Direct imputation of summary statistics for unmeasured SNPs." *Bioinformatics* 29 (22). doi:10.1093/bioinformatics/btt500.
- Lee, Donghyung, Vernell S. Williamson, T. Bernard Bigdeli, Brien P. Riley, Bradley T. Webb, Ayman H. Fanous, Kenneth S. Kendler, Vladimir I. Vladimirov, and Silviu-Alin Bacanu. 2015. "JPEG-MIX: gene-level joint analysis of functional SNPs in cosmopolitan cohorts." *Bioinformatics* 32. doi:10.1093/bioinformatics/btv567.
- Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, et al. 2016. "Analysis of protein-coding genetic variation in 60,706 humans." *Nature* 536 (7616): 285–91. doi:10.1038/nature19057.
- Leslie, Richard, Christopher J. O'Donnell, and Andrew D. Johnson. 2014. "GRASP: Analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database." *Bioinformatics* 30 (12): 185–94. doi:10.1093/bioinformatics/btu273.
- Li, Yun, Cristen J. Willer, Jun Ding, Paul Scheet, and Gonçalo R. Abecasis. 2010. "MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes." *Genetic Epidemiology* 34 (8): 816–34. doi:10.1002/gepi.20533.
- Locke, Adam E, and others. 2015. "Genetic studies of body mass index yield new insights for obesity biology." *Nature* 518 (7538): 197–206. <http://dx.doi.org/10.1038/nature14177> <http://10.0.4.14/nature14177>.
- Loh, Po-Ru, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, et al. 2015. "Efficient Bayesian mixed model analysis increases association power in large cohorts." *Nature Genetics* 47

- (3): 284–90. doi:10.1038/ng.3190.
- Lu, Yingchang, Felix R. Day, Stefan Gustafsson, Martin L. Buchkovich, Jianbo Na, Veronique Bataille, Diana L. Cousminer, et al. 2016. “New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk.” *Nature Communications* 7. doi:10.1038/ncomms10495.
- Maier, Robert M., Zhihong Zhu, Sang Hong Lee, Maciej Trzaskowski, Douglas M. Ruderfer, Eli A. Stahl, Stephan Ripke, et al. 2018. “Improving genetic prediction by leveraging genetic correlations among human diseases and traits.” *Nature Communications* 9 (1): 989. doi:10.1038/s41467-017-02769-6.
- Marchini, Jonathan, and Bryan Howie. 2010. “Genotype imputation for genome-wide association studies → **Review**.” *Nature Reviews Genetics* 11 (7): 499–511. doi:10.1038/nrg2796.
- Marigorta, Urko M, Juan Antonio Rodríguez, Greg Gibson, and Arcadi Navarro. 2018. “Replicability and Prediction: Lessons and Challenges from GWAS.” *Trends in Genetics* xx: 1–14. doi:10.1016/j.tig.2018.03.005.
- Marouli, Eirini, and others. 2017. “Rare and low-frequency coding variants alter human adult height.” *Nature*. doi:10.1038/nature21039.
- Martin, Alicia R., Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, and Eimear E. Kenny. 2017. “Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations.” *American Journal of Human Genetics* 100 (4): 635–49. doi:10.1016/j.ajhg.2017.03.004.
- Mägi, Reedik, Yury V Suleimanov, Geraldine M Clarke, Marika Kaakinen, Krista Fischer, Inga Prokopenko, and Andrew P Morris. 2017. “SCOPA and META-SCOPA : software for the analysis and aggregation of genome-wide association studies of multiple correlated phenotypes.” *BMC Bioinformatics*, 4–11. doi:10.1186/s12859-016-1437-3.
- McCarthy, Mark I, Gonçalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P.A. Ioannidis, and Joel N. Hirschhorn. 2008. “Genome-wide association studies for complex traits: Consensus, uncertainty and challenges → **Review**.” *Nature Reviews Genetics* 9 (5): 356–69. doi:10.1038/nrg2344.
- McDaid, Aaron F., Peter K. Joshi, Eleonora Porcu, Andrea Komljenovic, Hao Li, Vincenzo Sorrentino, Maria Litovchenko, et al. 2017. “Bayesian association scan reveals loci associated with human lifespan and linked biomarkers.” *Nature Communications* 8 (May). doi:10.1038/ncomms15842.
- Moayyeri, Alireza, Christopher J. Hammond, Ana M. Valdes, and Timothy D. Spector. 2013. “Cohort profile: TwinsUK and healthy

- ageing twin study." *International Journal of Epidemiology* 42 (1): 76–85. doi:10.1093/ije/dyr207.
- Ning, Zheng, Youngjo Lee, Peter K Joshi, James F Wilson, Yudi Pawitan, and Xia Shen. 2017. "A Selection Operator for Summary Association Statistics Reveals Allelic Heterogeneity of Complex Traits." *The American Journal of Human Genetics* 101 (6): 903–12. doi:10.1016/j.ajhg.2017.09.027.
- O'Connor, Luke, and Alkes L. Price. 2017. "Distinguishing genetic correlation from causation across 52 diseases and complex traits." *Doi.Org*, 205435. doi:10.1101/205435.
- O'Reilly, Paul F., Clive J. Hoggart, Yotsawat Pomyen, Federico C.F. Calboli, Paul Elliott, Marjo Riitta Jarvelin, and Lachlan J.M. Coin. 2012. "MultiPhen: Joint model of multiple phenotypes can increase discovery in GWAS." *PLoS ONE* 7 (5). doi:10.1371/journal.pone.0034861.
- Park, D. S., B. Brown, C. Eng, S. Huntsman, D. Hu, D. G. Torgerson, E. G. Burchard, and N. Zaitlen. 2015. "Adapt-Mix: learning local genetic correlation structure improves summary statistics-based analyses." *Bioinformatics* 31 (12). doi:10.1093/bioinformatics/btv230.
- Pasaniuc, Bogdan, and Alkes L Price. 2017. "Dissecting the genetics of complex traits using summary association statistics." *Nature Reviews Genetics* 18 (2): 117–27. doi:10.0.4.14/nrg.2016.142.
- Pasaniuc, Bogdan, Noah Zaitlen, Huwenbo Shi, Gaurav Bhatia, Alexander Gusev, Joseph Pickrell, Joel Hirschhorn, David P Strachan, Nick Patterson, and Alkes L Price. 2014. "Fast and accurate imputation of summary statistics enhances evidence of functional enrichment." *Bioinformatics* 30 (20). doi:10.1093/bioinformatics/btu416.
- Pers, Tune H., Juha M. Karjalainen, Yingleong Chan, Harm Jan Westra, Andrew R. Wood, Jian Yang, Julian C. Lui, et al. 2015. "Biological interpretation of genome-wide association studies using predicted gene functions." *Nature Communications* 6. doi:10.1038/ncomms6890.
- Pickrell, Joseph K. 2014. "Joint analysis of functional genomic data and genome-wide association studies of 18 human traits." *American Journal of Human Genetics* 94 (4): 559–73. doi:10.1016/j.ajhg.2014.03.004.
- Pickrell, Joseph, Tomaz Berisa, Laure Segurel, Joyce Y Tung, and David Hinds. 2015. "Detection and interpretation of shared genetic influences on 40 human traits." *bioRxiv*, no. April: 019885. doi:10.1101/019885.
- Pingault, Jean-Baptiste, Paul F O'Reilly, Tabea Schoeler, George B Ploubidis, Frühling Rijdsdijk, and Frank Dudbridge. 2018. "Using genetic data to strengthen causal inference in observational research → **Review**." *Nature Reviews Genetics*. doi:10.1038/s41576-

018-0020-3.

Popejoy, Alice B., and Stephanie M. Fullerton. 2016. "Genomics is failing on diversity." *Nature* 538 (7624): 161–64. doi:10.1038/538161a.

Porcu, Eleonora, and others. 2018. "Mendelian randomization integrating GWAS and eQTL reveals genetic determinants of complex and clinical traits." *Manuscript in Preparation*.

Rauch, Andri, Zoltán Kutalik, Patrick Descombes, Tao Cai, Julia Di Iulio, Tobias Mueller, Murielle Bochud, et al. 2010. "Genetic Variation in IL28B Is Associated With Chronic Hepatitis C and Treatment Failure: A Genome-Wide Association Study." *Gastroenterology* 138 (4): 1338–1345.e7. doi:10.1053/j.gastro.2009.12.056.

Roadmap Epigenomics Consortium. 2015. "Integrative analysis of 111 reference human epigenomes." *Nature* 518 (February): 317. doi:10.0.4.14/nature14248.

Robinson, Matthew R., Naomi R. Wray, and Peter M. Visscher. 2014. "Explaining additional genetic variation in complex traits." *Trends in Genetics* 30 (4): 124–32. doi:10.1016/j.tig.2014.02.003.

Rueedi, Rico, Roger Mallol, Johannes Raffler, David Lamparter, Nele Friedrich, Peter Vollenweider, Gérard Waeber, Gabi Kas-tenmüller, Zoltán Kutalik, and Sven Bergmann. 2017. "Metabo-matching: Using genetic association to identify metabolites in proton NMR spectroscopy." *PLoS Computational Biology* 13 (12): 1–17. doi:10.1371/journal.pcbi.1005839.

Rüeger, Sina, and others. 2015. "Impact of Common Risk Factors of Fibrosis Progression in Chronic Hepatitis C." *Gut* 64 (10): 1605–15.

Rüeger, Sina, Aaron McDaid, and Zoltán Kutalik. 2017. "Improved Imputation of Summary Statistics for Realistic Settings." *bioRxiv*, 203927.

Rüeger, Sina, Aaron McDaid, and Zoltán Kutalik. 2018. "Evaluation and application of summary statistic imputation to discover new height-associated loci." *PLoS Genetics* 14 (5): 1–32. doi:10.1371/journal.pgen.1007371.

Schaid, Daniel J., Wenan Chen, and Nicholas B. Larson. 2018. "From genome-wide associations to candidate causal variants by statistical fine-mapping → **Review**." *Nature Reviews Genetics*, 1. doi:10.1038/s41576-018-0016-z.

Schork, Andrew J, Wesley K Thompson, Phillip Pham, Ali Torkamani, J Cooper Roddey, Patrick F Sullivan, John R Kelsoe, Michael C O Donovan, and Helena Furberg. 2013. "All SNPs Are Not Created Equal : Genome-Wide Association Studies Reveal a Consistent Pattern of Enrichment among Functionally

- Annotated SNPs" 9 (4). doi:10.1371/journal.pgen.1003449.
- Servin, Bertrand, and Matthew Stephens. 2007. "Imputation-based analysis of association studies: Candidate regions and quantitative traits." *PLoS Genetics* 3 (7): 1296–1308. doi:10.1371/journal.pgen.0030114.
- Speed, Doug, and David J Balding. 2018. "Better estimation of SNP heritability from summary statistics provides a new understanding of the genetic architecture of complex traits." *bioRxiv*. doi:10.1101/284976.
- Speed, Doug, Na Cai, Michael Johnson, Sergey Nejentsev, and David Balding. 2016. "Re-evaluation of SNP heritability in complex human traits." *bioRxiv*, no. April: 074310. doi:10.1101/074310.
- Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, et al. 2015. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age." *PLoS Medicine* 12 (3): 1–10. doi:10.1371/journal.pmed.1001779.
- The ENCODE Project Consortium. 2012. "An integrated encyclopedia of DNA elements in the human genome." *Nature* 489 (September): 57. <http://dx.doi.org/10.1038/nature11247> <http://10.0.4.14/nature11247>.
- The GTEx Consortium. 2013. "The Genotype-Tissue Expression (GTEx) project." *Nature Genetics* 45 (6): 580–85. doi:10.1038/ng.2653.
- The International HapMap Consortium. 2003. "The International HapMap Project." *Nature* 426: 789–96. doi:10.1038/nature02168.
- Timpson, Nicholas J., Celia M.T. Greenwood, Nicole Soranzo, Daniel J. Lawson, and J. Brent Richards. 2018. "Genetic architecture: The shape of the genetic contribution to human traits and disease → **Review**." *Nature Reviews Genetics* 19 (2): 110–24. doi:10.1038/nrg.2017.101.
- Turley, Patrick, Raymond K. Walters, Omeed Maghzian, Aysu Okbay, James J. Lee, Mark Alan Fontana, Tuan Anh Nguyen-Viet, et al. 2018. "Multi-trait analysis of genome-wide association summary statistics using MTAG." *Nature Genetics* 50 (2): 229–37. doi:10.1038/s41588-017-0009-4.
- UK Biobank Phasing and Imputation Documentation. 2015. https://biobank.ctsu.ox.ac.uk/crystal/docs/impute_ukb_v1.pdf.
- Verbanck, Marie, Chia-yen Chen, Benjamin Neale, and Ron Do. 2018. "Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases." *Nature Genetics* 50 (May). doi:10.1038/s41588-018-0099-7.
- Visscher, Peter M, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark

- I McCarthy, Matthew A Brown, and Jian Yang. 2017. "10 Years of GWAS Discovery : Biology, Function, and Translation → **Review.**" *The American Journal of Human Genetics* 101 (1): 5–22. doi:10.1016/j.ajhg.2017.06.005.
- Visscher, Peter M. 2008. "Sizing up human height variation." *Nature Genetics* 40 (5): 489–90. doi:10.1038/ng0508-489.
- Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang. 2012. "Five Years of GWAS Discovery → **Review.**" *The American Journal of Human Genetics* 90 (1): 7–24. doi:10.1016/j.ajhg.2011.11.029.
- Visscher, Peter M., William G. Hill, and Naomi R. Wray. 2008. "Heritability in the genomics era - Concepts and misconceptions → **Review.**" *Nature Reviews Genetics* 9 (4): 255–66. doi:10.1038/nrg2322.
- Visscher, Peter M., Sarah E. Medland, Manuel A.R. Ferreira, Katherine I. Morley, Gu Zhu, Belinda K. Cornes, Grant W. Montgomery, and Nicholas G. Martin. 2006. "Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings." *PLoS Genetics* 2 (3): 0316–25. doi:10.1371/journal.pgen.0020041.
- Welter, Danielle, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, et al. 2014. "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations." *Nucleic Acids Research* 42 (D1): 1001–6. doi:10.1093/nar/gkt1229.
- Wen, Xiaoquan, and Matthew Stephens. 2010. "Using linear predictors to impute allele frequencies from summary or pooled genotype data." *Annals of Applied Statistics* 4 (3): 1158–82.
- Westra, Harm-Jan, Marjolein J Peters, Tonu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W Christiansen, et al. 2013. "Systematic identification of trans eQTLs as putative drivers of known disease associations." *Nature Genetics* 45 (10): 1238–43. doi:10.0.4.14/ng.2756.
- Winkler, Thomas W., Felix R. Day, Damien C. Croteau-Chonka, Andrew R. Wood, Adam E. Locke, Reedik Mägi, Teresa Ferreira, et al. 2014. *Nature Protocols* 9 (5): 1192–1212. doi:10.1038/nprot.2014.071.
- Winkler, Thomas W., Anne E. Justice, Mariaelisa Graff, Llida Barata, Mary F. Feitosa, Su Chu, Jacek Czajkowski, et al. 2015. "The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study" 11 (10): e1005378. doi:10.1371/journal.pgen.1005378.
- Wishart, David S., Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, et al. 2018. "DrugBank 5.0: A major update to the DrugBank database for 2018." *Nucleic Acids Research* 46 (D1): D1074–D1082. doi:10.1093/nar/gkx1037.
- Wood, Andrew R, and others. 2014. "Defining the role of common

- variation in the genomic and biological architecture of adult human height." *Nature Genetics* 46 (11). doi:10.1038/ng.3097.
- Wood, Andrew R., and others. 2013. "Imputation of Variants from the 1000 Genomes Project Modestly Improves Known Associations and Can Identify Low-Frequency Variant - Phenotype Associations Undetected by Hapmap Based Imputation." *PLoS ONE* 8 (5): 1–13. doi:10.1371/journal.pone.0064343.
- Wójtowicz, Agnieszka, Mark S. Gresnigt, Thanh Lecompte, Stephanie Bibert, Oriol Manuel, Leo A.B. Joosten, Sina Rüeger, et al. 2015. "IL1B and DEFB1 polymorphisms increase susceptibility to invasive mold infection after solid-organ transplantation." *Journal of Infectious Diseases* 211 (10): 1646–57. doi:10.1093/infdis/jiu636.
- Wójtowicz, Agnieszka, T. Doco Lecompte, Stephanie Bibert, Oriol Manuel, Sina Rüeger, Christoph Berger, Katia Boggian, et al. 2015. "PTX3 Polymorphisms and Invasive Mold Infections after Solid Organ Transplant." *Clinical Infectious Diseases* 61 (4): 619–22. doi:10.1093/cid/civ386.
- Wray, Naomi R., Cisca Wijmenga, Patrick F. Sullivan, Jian Yang, and Peter M. Visscher. 2018. "Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model." *Cell* 173 (7): 1573–80. doi:10.1016/j.cell.2018.05.051.
- Xu, Zheng, Qing Duan, Song Yan, Wei Chen, Mingyao Li, Ethan Lange, and Li Yun. 2015. "Genome Analysis DISSCO : Direct Imputation of Summary Statistics allowing COvariates." *Bioinformatics*. doi:10.1093/bioinformatics/btv168.
- Yang, Jian, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, and others. 2010. "Common SNPs explain a large proportion of the heritability for human height." *Nature Genetics* 42 (7): 565–69. doi:10.1371/journal.pgen.1003355.
- Yang, Jian, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. 2011. "GCTA: A tool for genome-wide complex trait analysis." *American Journal of Human Genetics* 88 (1): 76–82. doi:10.1016/j.ajhg.2010.11.011.
- Yang, Jian, Jian Zeng, Michael E. Goddard, Naomi R. Wray, and Peter M. Visscher. 2017. "Concepts, estimation and interpretation of SNP-based heritability." *Nature Genetics* 49 (9): 1304–10. doi:10.1038/ng.3941.
- Yengo, Loic, Julia Sidorenko, Kathryn E Kemper, Zhili Zheng, Andrew R Wood, Michael N Weedon, Timothy M Frayling, Joel Hirschhorn, and Jian Yang. 2018. "Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry," 1–25. doi:10.1101/274654.
- Zheng, Jie, Mesut Erzurumluoglu, Benjamin Elsworth, Laurence

Howe, Philip Haycock, Gibran Hemani, Katherine Tansey, et al. 2016. "LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis." *bioRxiv*, no. September: 051094. doi:10.1101/051094.

Zhernakova, Daria V, Patrick Deelen, Martijn Vermaat, Maarten Van Iterson, Michiel Van Galen, Wibowo Arindrarto, Peter Van Hof, et al. 2017. "Identification of context-dependent expression quantitative trait loci in whole blood" 49 (1). doi:10.1038/ng.3737.

Zhu, Zhihong, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, et al. 2016. "Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets." *Nature Genetics* 48 (5): 481–87. doi:10.1038/ng.3538.

Zhu, Zhihong, Zhili Zheng, Futao Zhang, Yang Wu, Maciej Trzaskowski, Robert Maier, Matthew R. Robinson, et al. 2018. "Causal associations between risk factors and common diseases inferred from GWAS summary data." *Nature Communications* 9 (1). doi:10.1038/s41467-017-02317-2.

RESEARCH ARTICLE

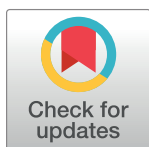
Evaluation and application of summary statistic imputation to discover new height-associated loci

Sina Rüeger^{1,2}✉, Aaron McDaid^{1,2}✉, Zoltán Kutalik^{1,2*}

1 Institute of Social and Preventive Medicine, Lausanne University Hospital, Lausanne, 1010, Switzerland, **2** Swiss Institute of Bioinformatics, Lausanne, 1015, Switzerland

✉ These authors contributed equally to this work.

* zoltan.kutalik@unil.ch



OPEN ACCESS

Citation: Rüeger S, McDaid A, Kutalik Z (2018) Evaluation and application of summary statistic imputation to discover new height-associated loci. *PLoS Genet* 14(5): e1007371. <https://doi.org/10.1371/journal.pgen.1007371>

Editor: Michael P. Epstein, Emory University, UNITED STATES

Received: December 1, 2017

Accepted: April 18, 2018

Published: May 21, 2018

Copyright: © 2018 Rüeger et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the Leenaards Foundation (<http://www.leenaards.ch>), the Swiss National Science Foundation [31003A-143914, 31003A-169929]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

As most of the heritability of complex traits is attributed to common and low frequency genetic variants, imputing them by combining genotyping chips and large sequenced reference panels is the most cost-effective approach to discover the genetic basis of these traits. Association summary statistics from genome-wide meta-analyses are available for hundreds of traits. Updating these to ever-increasing reference panels is very cumbersome as it requires reimputation of the genetic data, rerunning the association scan, and meta-analysing the results. A much more efficient method is to directly impute the summary statistics, termed as *summary statistics imputation*, which we improved to accommodate variable sample size across SNVs. Its performance relative to *genotype imputation* and practical utility has not yet been fully investigated. To this end, we compared the two approaches on real (genotyped and imputed) data from 120K samples from the UK Biobank and show that, *genotype imputation* boasts a 3- to 5-fold lower root-mean-square error, and better distinguishes true associations from null ones: We observed the largest differences in power for variants with low minor allele frequency and low imputation quality. For fixed false positive rates of 0.001, 0.01, 0.05, using *summary statistics imputation* yielded a decrease in statistical power by 9, 43 and 35%, respectively. To test its capacity to discover novel associations, we applied *summary statistics imputation* to the GIANT height meta-analysis summary statistics covering HapMap variants, and identified 34 novel loci, 19 of which replicated using data in the UK Biobank. Additionally, we successfully replicated 55 out of the 111 variants published in an exome chip study. Our study demonstrates that *summary statistics imputation* is a very efficient and cost-effective way to identify and fine-map trait-associated loci. Moreover, the ability to impute summary statistics is important for follow-up analyses, such as Mendelian randomisation or LD-score regression.

Author summary

Genome-wide association studies (GWASs) quantify the effect of genetic variants and traits, such as height. Such estimates are called *association summary statistics* and are typically publicly shared through publication. Typically, GWASs are carried out by genotyping $\sim 500'000$ SNVs for each individual which are then combined with sequenced reference panels to infer untyped SNVs in each individual's genome. This process of *genotype imputation* is resource intensive and can therefore be a limitation when combining many GWASs. An alternative approach is to bypass the use of individual data and directly impute summary statistics. In our work we compare the performance of *summary statistics imputation* to *genotype imputation*. We observe that *genotype imputation* shows a 3- to 5-fold lower RMSE compared to *summary statistics imputation*, as well as a better capability to distinguish true associations from null results. Furthermore, we demonstrate the potential of *summary statistics imputation* by presenting 34 novel height-associated loci, 19 of which were confirmed in UK Biobank. Our study demonstrates that given current reference panels, *summary statistics imputation* is a very efficient and cost-effective way to identify common or low-frequency trait-associated loci.

Introduction

Genome-wide association studies (GWASs) have been successfully applied to reveal genetic markers associated with hundreds of traits and diseases. The genotyping arrays used in these studies only interrogate a small proportion of the genome and are therefore typically unable to pinpoint the causal variant. Such arrays have been designed to be cost-effective and include only a set of tag single nucleotide variants (SNVs) that allow the inference of many other unmeasured markers. To date, thousands of individuals have been sequenced [1, 2] to provide high resolution haplotypes for *genotype imputation* tools such as IMPUTE and minimac [3, 4], which are able to infer sequence variants with ever-increasing accuracy as the reference haplotype set grows.

Downstream analyses such as Mendelian randomisation [5], approximate conditional analysis [6], heritability estimation [7], and enrichment analysis using high resolution annotation (such as DHS) [8] often require genome-wide association results at the highest possible genomic resolution. *Summary statistics imputation* [9] has been proposed as a solution that only requires summary statistics and the linkage disequilibrium (LD) information estimated from the latest sequencing panel to directly impute up-to-date meta-analysis summary statistics [10]. Because *summary statistics imputation* uses summarised data as input, it is not bounded to privacy restrictions related to the use of individual data. Another advantage is its substantially lower computation time compared to genotype imputation. For example, for imputation of the UK Biobank data, it is about 500 times faster (4200 vs 8.3 CPU days comparing Minimac [4] to our SSIMP software [11]).

This study compares *summary statistics imputation* directly to genotype imputation and focuses on its practical advantages using real data. In particular, we evaluated two experiments: 1) we ran a GWAS on both simulated traits and human height using data from 120'086 individuals from the UK Biobank and compared the performances of *summary statistics imputation* and *genotype imputation*, using direct genotyping/sequencing as gold standard; 2) we imputed association summary statistics from a HapMap-based GWAS study [12] using the UK10K reference panel to explore new potential height-associated variants which we validated using results from Marouli *et al.* [13] and the UK Biobank height GWAS ($n = 336'474$). We

extended *summary statistics imputation* [9, 14] which yields increased imputation accuracy by accounting for variable sample sizes. For all applications presented in this manuscript we are using this improved version of *summary statistics imputation*.

Materials and methods

Summary statistics imputation (SSimp)

By combining summary statistics for a set of variants and the fine-scale LD structure in the same region, we can estimate summary statistics of new, untyped variants at the same locus.

We assume a set of univariate effect size estimates a_i are available for SNVs $i = 1, \dots, I$ from a linear regression between a continuous phenotype y and the corresponding genotype g^i measured in N individuals. Without loss of generality we assume that both vectors are normalised to have zero mean and unit variance. Thus $a_i = \frac{(g^i)'y}{N}$ and $\mathbf{a} = (a_1, a_2, \dots, a_I)'$ $\sim \mathcal{N}(\alpha, \Sigma)$. Σ represents the pairwise covariance matrix of effect sizes of all $i = 1, \dots, I$ SNVs.

To estimate the univariate effect size α_u of an untyped SNV u in the same sample, one can use the conditional expectation of a multivariate normal distribution. The conditional mean of the effect of SNV u can be expressed using the effect size estimates of the tag SNVs [9, 15]:

$$\hat{a}_u = a_{u|\mathcal{M}} = \alpha_u + \Sigma_{u\mathcal{M}}\Sigma_{\mathcal{M}\mathcal{M}}^{-1}(\mathbf{a} - \alpha), \quad (1)$$

where \mathcal{M} is a vector of so-called *tag* SNVs, $\Sigma_{u\mathcal{M}}$ represents the covariance between SNV u and all \mathcal{M} markers and $\Sigma_{\mathcal{M}\mathcal{M}}$ represents the covariance between all \mathcal{M} markers.

We assume that estimates for the two covariances are available from an external reference panel with n individuals and denote them $\mathbf{s} = \hat{\Sigma}_{u\mathcal{M}}$, $\mathbf{S} = \hat{\Sigma}_{\mathcal{M}\mathcal{M}}$. The corresponding correlation matrices are γ and Γ , with $\mathbf{c} = N \cdot \mathbf{s}$ and $\mathbf{C} = N \cdot \mathbf{S}$ being the estimates for the correlation matrices. Further, by assuming that SNV u and the trait are independent conditioned on the \mathcal{M} markers, i.e. $\alpha_u - \Sigma_{u\mathcal{M}}\Sigma_{\mathcal{M}\mathcal{M}}^{-1}\alpha = 0$, Eq (1) becomes

$$\hat{a}_u = a_{u|\mathcal{M}} = \mathbf{s}'\mathbf{S}^{-1}\mathbf{a} = \mathbf{c}'\mathbf{C}^{-1}\mathbf{a} \quad (2)$$

One can also choose to impute the Z-statistic instead, as derived by Pasaniuc *et al.* [9]:

$$\hat{z}_{u|\mathcal{M}} = \mathbf{c}'\mathbf{C}^{-1}\mathbf{z} \quad (3)$$

with $\mathbf{z} = \mathbf{a}\sqrt{N}$, when the effect size is small (as is the case in typical GWAS).

Similar to Pasaniuc *et al.* [9], we chose \mathcal{M} to include all measured variants within at least 250 Kb of SNV u . To speed up the computation when imputing SNVs genome-wide, we apply a windowing strategy, where SNVs within a 1 Mb window are imputed simultaneously using the same set of \mathcal{M} tag SNVs the 1 Mb window plus 250 Kb flanking regions on each side.

Shrinkage of SNV correlation matrix. To estimate \mathbf{C} (and \mathbf{c}) we use an external reference panel of n individuals. Since the size of \mathbf{C} often exceeds the number of individuals ($q \gg n$), shrinkage of matrix \mathbf{C} is needed to guarantee that it is invertible.

Off-diagonal values of \mathbf{C} are shrunk towards zero and the extent of which is characterised by a shrinkage parameter λ . As a consequence, it also lowers the RMSE in *summary statistics imputation* [16], as values in \mathbf{C} close to zero, may represent pure noise (and zero LD), which can be inflated when inverting the matrix.

By applying shrinking, the modified matrix becomes

$$\mathbf{C}_\lambda = (1 - \lambda)\mathbf{C} + \lambda\mathbf{I} \quad (4)$$

Even though \mathbf{c} is not inverted, we still shrink it to curb random fluctuations in the LD estimation in case of no LD.

$$\mathbf{c}_\lambda = (1 - \lambda)\mathbf{c} \quad (5)$$

Inserting \mathbf{c}_λ and \mathbf{C}_λ , Eq (2) then becomes

$$\hat{\mathbf{a}}_u = \mathbf{a}_{u|\mathcal{M}} = \mathbf{c}'_\lambda \mathbf{C}_\lambda^{-1} \mathbf{a} \quad (6)$$

Note that λ can vary between 0 and 1, with $\lambda = 1$ turning \mathbf{C} to the identity matrix, while $\lambda = 0$ leaves \mathbf{C} unchanged. Schäfer & Strimmer [16] find an optimal λ by minimising the variance of matrix \mathbf{C} . Wen & Stephens [17] propose to adjust matrix \mathbf{C} in a way that they represent recombination hotspots correctly. A similar idea is to set small absolute correlation values to 0. Here, we mainly focus on two commonly used λ values: λ fixed at 0.1 [9], and λ changing with the reference panel size n : $\lambda = 2/\sqrt{n}$ [18].

Imputation quality. Imputation quality, r^2 , is defined as the squared correlation between the imputed and true genotypes. An r^2 value of 1 means perfect imputation, whereas r^2 of 0 indicates poor imputation [19]. In *summary statistics imputation* this quantity is the total variance explained by a linear model where the imputed genotype is regressed onto all measured markers. It was proposed by Pasanuic *et al.* [9] to be estimated as

$$\hat{r}_{\text{pred}}^2 = \mathbf{c}'_\lambda \mathbf{C}_\lambda^{-1} \mathbf{c}_\lambda \quad (7)$$

Furthermore, we introduce an adjusted form to account for the ratio between the number of parameters (q) and sample size (n) [20]. Due to the fact that many measured SNVs are correlated, we further modify the formula by adjusting the number of parameters in the formula to the effective number of variants q_{eff} [21]:

$$\hat{r}_{\text{pred,adj}}^2 = 1 - (1 - \hat{r}_{\text{pred}}^2) \frac{n - 1}{n - q_{\text{eff}} - 1} \quad (8)$$

Negative values in Eq (8) are set to zero.

Summary statistics imputation accounting for varying sample size and missingness. All previously published *summary statistics imputation* methods assume that all effect estimates are based on the same set of N individuals. This assumption does not hold most of the time since meta-analysis studies use different genotyping chips or different imputation reference panels. As a result, the covariance between effect estimates will change. In the extreme case when effect estimates are computed in two non-overlapping samples, the correlation will be zero even if there is very high LD between the two SNVs.

To perform imputation, we require the correlation between any target complete Z-statistic, z_w , and any observed partial Z-statistic, z_k° , (with $k \in \mathcal{M}$),

$$\mathbf{d}_k := \text{Cor}[z_u, z_k^\circ] = c_{uk} \sqrt{\frac{N_k}{N_{\max}}}$$

We define N_k as the sample size of SNV k , \mathbf{N} as a vector recording the sample size of each tag SNV, N_{\max} as the maximum in \mathbf{N} , and assume that every tag SNV k the sample of individuals is a subset of a complete sample of N_{\max} individuals.

By defining $\delta_{kl} := \frac{N_{k \cap l}}{\sqrt{N_k N_l}}$, we can calculate the adjusted (estimated) correlation matrix \mathbf{D} , where each element is calculated as follows:

$$\mathbf{D}_{kl} = c_{kl} \delta_{kl}. \quad (9)$$

We present two estimators of δ_{kl} . Typically, we do not know the details of the exact sample overlap for every pair of SNVs, $N_{k \cap l}$, and instead simply know N_{\max} and the vector \mathbf{N} . Therefore, we must derive the sample overlap $N_{k \cap l}$ based on assumptions about the dependence structure of missingness.

The most conservative assumption is maximum possible overlap, resulting in maximum dependence, as this minimises the imputed Z-statistic. If each SNV has a corresponding binary missingness vector, the correlation between these missingness vectors will be maximised when the sample overlap is at its maximum, $N_{k \cap l} = \min(N_k, N_l)$. To enable the *dependent* approach, we construct a \mathbf{D} matrix by replacing $N_{k \cap l}$ with $\min(N_k, N_l)$,

$$\mathbf{D}_{kl}^{(dep)} = \mathbf{C}_{kl} \hat{\delta}_{kl}^{(dep)} = \mathbf{C}_{kl} \min \left(\frac{\sqrt{N_k}}{\sqrt{N_l}}, \frac{\sqrt{N_l}}{\sqrt{N_k}} \right). \quad (10)$$

If the missingness vectors are *independent* of each other, the expected overlap can be estimated as

$$\mathbf{D}_{kl}^{(ind)} = \mathbf{C}_{kl} \hat{\delta}_{kl}^{(ind)} = \mathbf{C}_{kl} \frac{\sqrt{N_k N_l}}{N_{\max}}. \quad (11)$$

Finally, we impute $z_u | \mathbf{z}_{\mathcal{M}}^{\circ}$ as

$$\hat{z}_u = \mathbf{d}' \mathbf{D}^{-1} \mathbf{z}_{\mathcal{M}}^{\circ}. \quad (12)$$

by using \mathbf{d} from Eq (9) and \mathbf{D} from either, Eq (10) or Eq (11).

In order to convert \hat{z}_u into the corresponding estimate of the standardised effect, we consider

$$\hat{a}_u = \frac{\hat{z}_u}{\sqrt{N_{\max} \mathbf{d}' \mathbf{D}^{-1} \mathbf{d}}}. \quad (13)$$

Note that $\mathbf{d}' \mathbf{D}^{-1} \mathbf{d}$ is the corresponding imputation quality. Details to the estimation of δ can be found in [S3 Appendix](#).

Comparison of summary statistics imputation versus genotype imputation

UK Biobank data. The UK Biobank [22] comprises health related information about 500'000 individuals based in the United Kingdom and aged between 40-69 years in 2006-2010. For our analysis we used Caucasians individuals (amongst people who self-identified as British) from the first release of the genetic data ($n = 120'086$). For SNVs, the number of individuals range between $n = 3'431$ and $n = 120'082$. Additionally to custom SNP array data, UK Biobank contains imputed genotypes [23]. A subset of 820'967 variants were genotyped and imputed, and 72M variants were imputed by UK Biobank, using SHAPEIT2 and IMPUTE2 [23].

Imputation of height GWAS summary statistics conducted in UK Biobank. We imputed GWAS Z-statistics (ran on directly genotyped data) using *summary statistics imputation* within 1 Mb-wide regions, by blinding one at the time and therefore allowing the remaining SNVs to be used for tagging. As tag SNVs we used all SNVs except the focal SNV within a 1.5 Mb window.

Selection of regions and SNVs. We selected 706 regions in total, consisting of 535 loci containing height-associated SNVs [12, 13] and 171 regions not containing any height-associated (all $P \geq 10^{-5}$) SNV. More specifically, within each height-associated region we only imputed SNVs that have $LD_{\max} > 0.2$. LD_{\max} was defined as the largest squared correlation between a SNV and all height-associated SNVs on the same chromosome. In the 171 null regions we chose only those variants with $LD_{\max} \leq 0.05$ with any associated marker on the same chromosome. These selection criteria lead to 44'992 variants being imputed. We did not analyse palindromic SNVs (A/T and C/G) (3'306 variants), SNVs with missing genotypes for more than 36'024 (30%) individuals (2'317 variants), SNVs with $MAF < 1\%$ (3'010 variants). These restrictions left us with 37'467 of the 44'992 imputed SNVs.

Comparison of summary statistics imputation and genotype imputation. To compare the performance between *summary statistics imputation* and *genotype imputation* followed by association we compared each method to the directly genotyped data association as gold standard. Fig 1 gives an overview of how these three types of summary statistics are related and compared. We used RMSE, bias, correlation, and the regression slope (no intercept) to evaluate these approaches against the truth.

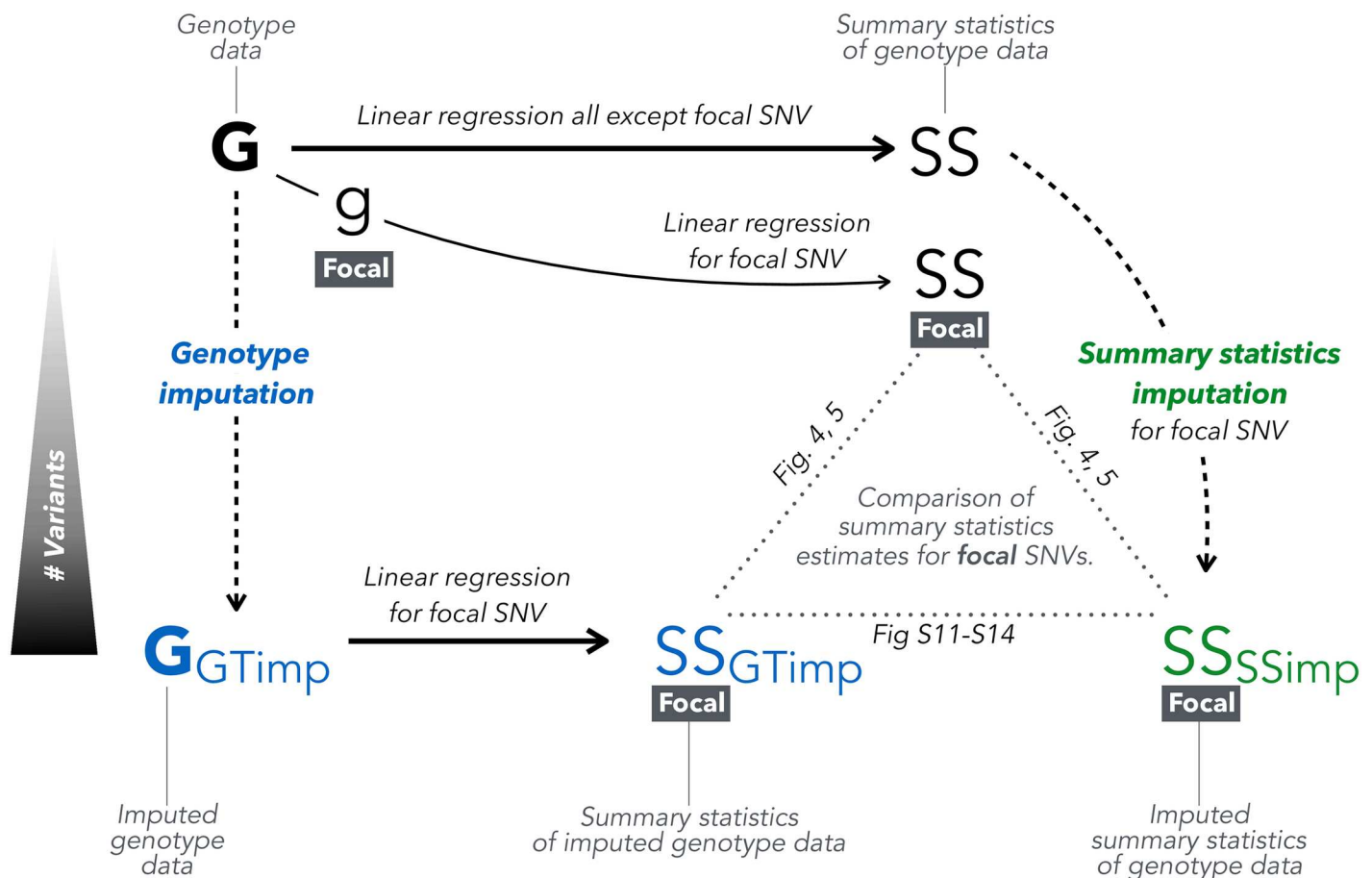


Fig 1. Overview of genotype vs. summary statistics imputation. From genotype data (top-left, G) we can calculate summary statistics (top-right, SS). Summary statistics for an unmeasured/masked SNV can be obtained via two ways: we can impute genotype data (bottom-left, G-Timp) using *genotype imputation* and then calculate summary statistics via linear regression (bottom-middle, SS-G-Timp), or by applying *summary statistics imputation* on the summary statistics calculated from genotype data (bottom-right, SS-SSimp). For the purpose of our analysis, we are only looking at genotyped (and genotype imputed) SNVs, thus masking one focal SNV in Figs 4, 5 and S11–S14.

<https://doi.org/10.1371/journal.pgen.1007371.g001>

More precisely, the RMSE and the Bias for a set of $k = 1 \dots K$ SNVs is:

$$\begin{aligned} d_k &= Z_k^{SSimp} - Z_k \\ \text{RMSE} &= \sqrt{\frac{1}{K} \sum_{k=1}^K d_k^2} \\ \text{Bias} &= \frac{1}{K} \sum_{k=1}^K d_k \end{aligned}$$

with Z_k^{SSimp} being the Z-statistic resulting from *summary statistics imputation* for SNV k and Z_k the Z-statistic resulting from genotype data for SNV k (our gold standard). Likewise, we replaced Z_k^{SSimp} with Z_k^{GTimp} , to calculate RMSE and bias for *genotype imputation*.

Note that for height-associated SNPs with missing genetic data we rescaled the association Z-statistic Z_u as follows $Z_u^* = Z_u \cdot \sqrt{\frac{N_{\max}}{N_u}}$ in order to make it comparable with its imputed version (Z^{GTimp} , Z^{SSimp}), derived from the full sample.

Additionally, we calculated power and false positive rate (FPR) for each method. For this, we randomly selected 3'390 SNVs and used each once as null and once as associated SNV. For the null scenario, we simulated a random, standard normal phenotype. For the alternative scenario, we simulated a phenotype such that the SNV explained 0.01% of the simulated phenotype variance (corresponding to typical a GWAS effect size). For both scenarios we calculated the summary statistics via *genotype imputation* and *summary statistics imputation*. For *summary statistics imputation*, we first ran a GWAS within ± 0.75 Mb of the focal SNV, and subsequently used the estimated summary statistics to perform *summary statistics imputation*. For SNVs with a real association we calculated the power as the fraction q_A of SNVs with a $P < \alpha$ ($q_A = f_A/m_A$, with m_A being the number of associated SNVs and f_A among them those with $P < \alpha$), whereas for SNVs with no association we calculated FPR as the fraction q_N of SNVs with $P < \alpha$ ($q_N = f_N/m_N$, with m_N being the number of null SNVs and f_N among them those with $P < \alpha$). We varied α between 0 and 1 and visualised FPR versus power for each method. The standard deviation was calculated based on the assumption of a binomial distribution for f_A and f_N : $f_i \sim B(m_i, q_i)$. The respective variance estimation for q_i is then: $\text{Var}(q_i) = q_i(1 - q_i)/m_i$.

Stratifying results. The obtained (summary statistics) imputation results were grouped based on the imputed SNVs (i) being correlated ($\text{LD} > 0.3$) to any height-associated SNV on the same chromosome or being a null SNV ($\text{LD} < 0.05$); (ii) low-frequency ($1\% < \text{MAF} \leq 5\%$) or common SNV ($\text{MAF} > 5\%$); (iii) being badly- ($\hat{r}_{\text{pred,adj}}^2 \leq 0.3$), medium- ($0.3 < \hat{r}_{\text{pred,adj}}^2 \leq 0.7$) or well-imputed ($0.7 < \hat{r}_{\text{pred,adj}}^2 \leq 1$). Height-associated SNVs are exclusively from 535 regions and termed *associated* SNVs, while SNVs not associated with height stem from 171 regions and are termed *null* SNVs. Throughout the manuscript, LD is estimated as the squared correlation [24].

Summary statistics imputation of the height GWAS of the GIANT consortium

GIANT consortium summary statistics. In 2014 the GIANT consortium published meta-analysed height summary statistics involving 79 cohorts, 253'288 individuals of European ancestry, and 2'550'858 autosomal HapMap SNVs [12], leading to the discovery of 423 height-associated loci (697 variants). Later, Marouli *et al.* [13] published summary statistics of the exome array meta-analysis (241'419 SNVs in up to 381'625 individuals), finding 122 novel

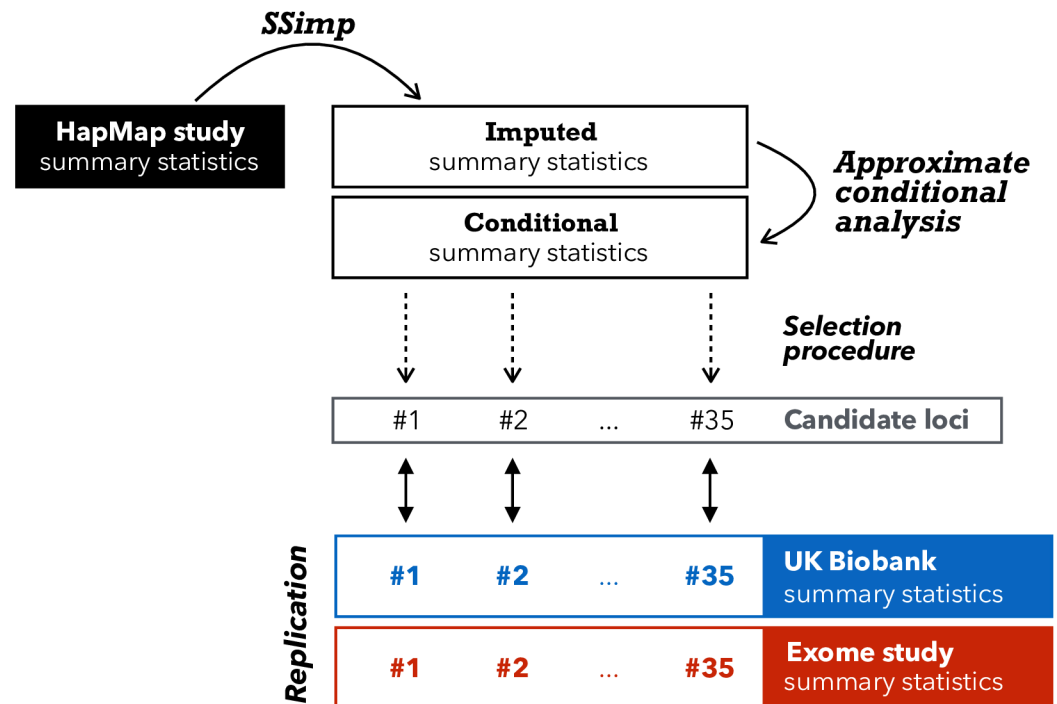


Fig 2. Overview of imputation and replication scheme. This illustration gives an overview how we used > 2M GIANT HapMap summary statistics (black rectangle) as tag SNVs to impute > 10M variants with $MAF \geq 0.1\%$ in UK10K. After adjusting the summary statistics for conditional analysis we applied a selection process that resulted in 35 candidate loci. To confirm these 35 loci we used summary statistics from UK Biobank (blue) as replication as well as summary statistics from the exome chip study, if available [13] (red). Loci that had not been discovered by the exome chip study, were termed *novel*.

<https://doi.org/10.1371/journal.pgen.1007371.g002>

variants (located in 120 loci) associated with height. Of the 122 exome variants, four were not available in UK10K and seven were on chromosome X, and could therefore not be imputed (because Wood *et al.* [12] did not include chromosome X), leaving 111 variants. We refer to the summary statistics by Wood *et al.* [12] as HapMap study, and to Marouli *et al.* [13] as exome chip study.

Summary statistics imputation of Wood *et al.* We imputed all non-HapMap variants that were available in UK10K, using the summary statistics in Wood *et al.* [12] as tag SNVs. In general, we only imputed variants with $MAF_{UK10K} \geq 0.1\%$ (this allows a minimal allele count of $8 \simeq 0.001 \cdot 3781 \cdot 2$), except for the 111 exome variants reported in Marouli *et al.* [13], which we imputed regardless of their MAF. We divided the genome into 2'789 core windows of 1 Mb. We imputed the summary statistics of each variant using the tag SNVs within its respective window and 250 Kb on each side. Fig 2 gives an overview of the datasets and methods involved.

Definition of a candidate locus. After applying *summary statistics imputation* we screened for SNVs with $\hat{r}_{pred,adj}^2 \geq 0.3$ and an (imputed) P -value $\leq 10^{-8}$ and applied conditional analysis, aiming to limit the results to SNVs acting independently from known HapMap findings. The significance threshold of 10^{-8} was chosen based on the effective number of SNVs evaluated ($< 9'276'018$). For each imputed 1 Mb window, we started the conditional analysis by defining two sets of SNVs. The first set contained all imputed SNVs that had an imputed P -value $\leq 10^{-8}$, ranging from position $bp^{(1)}$ to $bp^{(2)}$. The second SNV set contained all reported HapMap SNVs (697 in total) within a range of $bp^{(1)} - 1$ Mb and $bp^{(2)} + 1$ Mb. Having two SNV sets—the first set with newly detected variants, the second set with reported HapMap

variants—we could then condition each SNV in the first set on all SNVs in the second set, using approximate conditional analysis [25] and UK10K as the reference panel. Next, we declared a region as a candidate locus if at least one imputed variant in that locus had a conditional P -value $\leq 10^{-8}$. Additionally, for each (35) lead variant in the candidate regions we performed conditional analysis using each HapMap SNV (in turn) within 1 Mb vicinity. Finally, we performed a conditional analysis for nearby candidate loci (neighbouring windows), to avoid double counting. In each candidate locus we report the imputed variant with the smallest conditional P -value as the top variant.

Replication of candidate loci emerging from summary statistics imputation. We replicate our findings using our UK Biobank height GWAS results and for SNVs present on the exome chip we also use the recent height GWAS [13]. For both attempts to replicate our findings, UK Biobank and the exome chip study, the significance threshold for replication is $\alpha = 0.05/k$, with k as the number of candidate loci.

For replication using UK Biobank we used summary statistics based on the latest release of genetic data with $n = 336,474$ individuals, provided by the Neale lab [26]. For SNVs that were not present in the latest release we used summary statistics from the first release of genetic data ($n = 120,086$).

Annotation of candidate loci. We use two databases to annotate newly discovered SNVs. First, we use GTEx [27], an eQTL database with SNV-gene expression association summary statistics for 53 tissues. Second, we conduct a search in Phenoscanner [28], to identify previous studies (GWAS and metabolites) where the newly discovered SNVs had already appeared. For these two databases we report the respective summary statistics that pass the significance threshold of $\alpha = 10^{-6}$. We only extract the information for variants that were defined as novel discoveries.

Simulation

We simulated genetic data on 25,000 individuals was used. In brief, we used data from the five European subpopulations CEU, GBR, FIN, TSI and IBR of the 1000 Genomes reference panel [1]. We chose to up-sample chromosome 15 using HAPGEN2 [29] to 5,000 individuals for each subpopulation, yielding a total of 25,000 individuals. Of these, half of the data was used to estimate the LD structure C and the other half to simulate the association study with an *in silico* phenotype. The simulation procedure is described in more detail in [S1 Appendix](#). Forty regions were selected with one non-HapMap causal variant in each and all HapMap SNVs were used as tag SNVs. Sample size distributions were drawn from two published GWAS studies (on HDL [30] and T2D [31]). Missingness was assigned at random positions while respecting the missingness correlation parameter θ_{miss} , with zero value reflecting missingness at random and one corresponding to the maximum possible sample overlap between SNVs.

Reference panels

To estimate LD structure in C and c (Eq (2)) we used 3,781 individuals from UK10K data [32, 33], a reference panel of British ancestry that combines the TWINSUK and ALSPAC cohorts.

Software

All analysis was performed with R-3.2.5 [34] programming language, except GWAS summary statistics computation for UK Biobank genotype and genotype imputed data, for which SNPTTEST-5.2 [35] was used. For summary statistics imputation we used SSIMP [11].

Results

To assess the performance of *summary statistics imputation* in realistic scenarios we used two different datasets. In Section “Comparison of *summary statistics imputation* versus *genotype imputation*” we compare the performance of *summary statistics imputation* to *genotype imputation*, using measured and imputed genotype data from 120'086 individuals in the UK Biobank. In Section “*Summary statistics imputation* of the height GWAS of the GIANT consortium”, we use published association summary statistics from 253'288 individuals to show that *summary statistics imputation* can be used to identify novel associations. For all analyses we used an improved estimation of the standardised effect sizes that is robust to variable sample missingness. We validate this method in the next Section “Varying sample size and missingness”. Both analyses are centered around the genetics of human height. In the following we will often refer to two GIANT (Genetic Investigation of ANthropometric Traits) publications: Wood *et al.* [12], an analysis of HapMap variants that revealed 423 loci, and Marouli *et al.* [13], an exome chip based analysis that revealed 120 new height-associated loci. Together, these two studies—the HapMap and the exome chip study—constitute the most complete collection of genetic associations with height.

Varying sample size and missingness

The conventional estimate of the standardised effect of a SNV u , $\hat{a}_u^{(conv)}$, (Eq (2)) is unbiased, under certain assumptions, but can have large variance when there is variation in the sample sizes recorded in N_M . In this section, we used upsampled 1000 Genomes data [1] and simulated phenotype with known standardised effect α and various missingness design. We compare the MSE of the conventional estimation to the MSE of two other estimators, Eq (13) using $D^{(dep)}$ and $D^{(ind)}$, derived in the method section.

In general, the size of the overlap is unknown and we recommend using the assumption of maximum dependence ($D^{(dep)}$) as it is the most conservative assumption. An alternative is to assume randomly distributed missingness ($D^{(ind)}$). Most pairs of SNVs in GIANT attain close to the maximum possible missingness-overlap (S10 Fig) and therefore this assumption is not overly-conservative.

The results in Fig 3 demonstrate that the conventional method has the largest MSE across all the simulation parameters tested. Where the variance in sample size is very large (top row of Fig 3), the true correlation is often very close to zero. Both of our methods effectively make this same (correct) assumption of low correlation and therefore they both perform equally well.

Where the variation in sample size is less extreme, as in the simulations on the bottom row of Fig 3, there is less shrinkage of correlation and the simulated missingness correlation becomes more relevant. Where the simulated data has the maximum possible missingness correlation (on the right hand side of the subplots in Fig 3), i.e. the sample overlap between each pair of SNVs is as large as possible given their two sample sizes, $D^{(dep)}$ performs better (as expected). With lower overlap (first column) $D^{(ind)}$ performs better.

Comparison of *summary statistics imputation* versus *genotype imputation*

By having two types of genetic data at hand, genotype and imputed genotype data, we were able to compare summary statistics of 37'467 typed SNVs resulting from (1) associations calculated from original genotype data (ground truth); (2) associations calculated from imputed genotype data (*genotype imputation*) and (3) associations imputed from summary statistics calculated using genotype data (Fig 1). For our analysis, we defined 706 genomic regions in total,

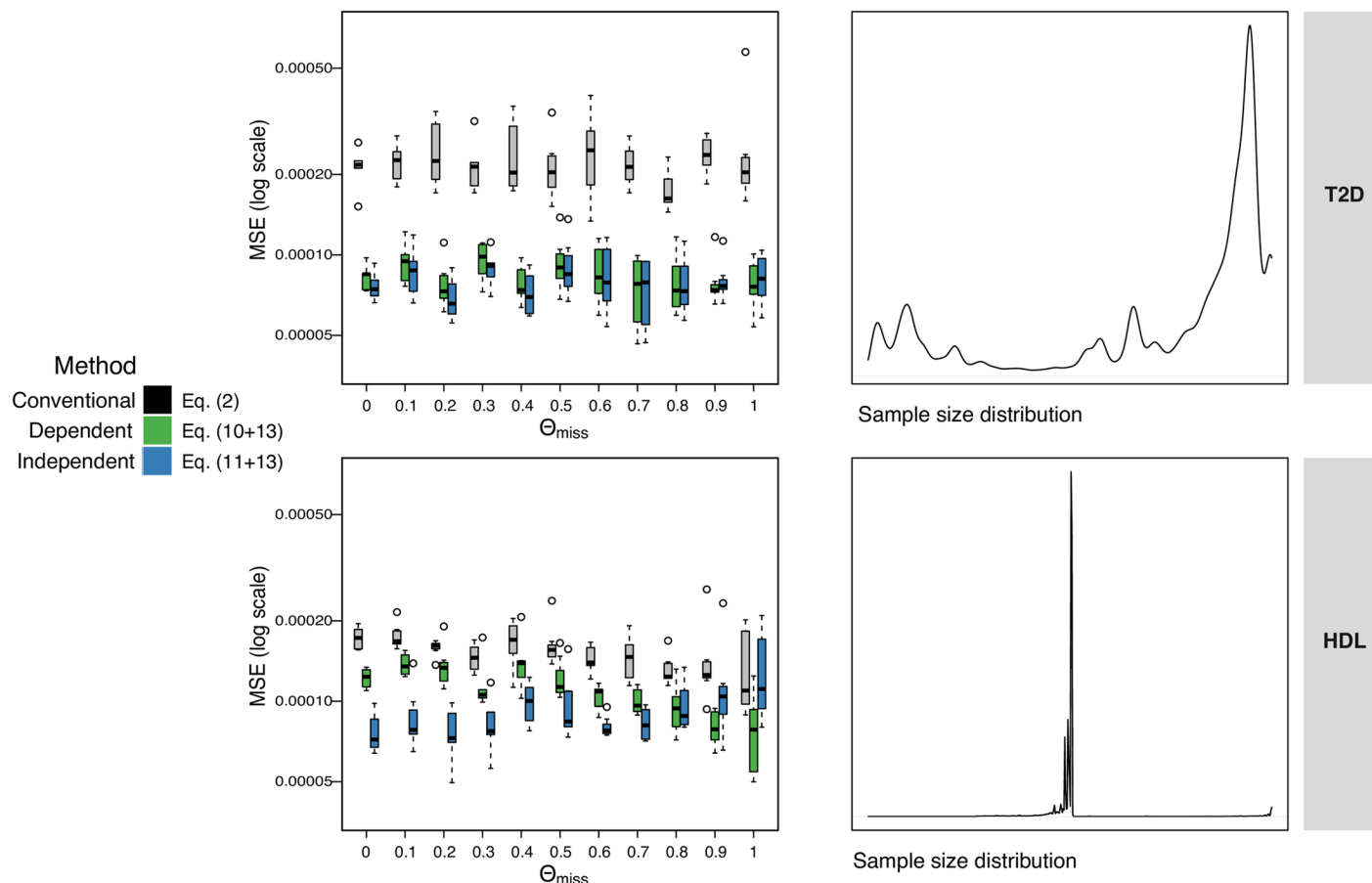


Fig 3. Accounting for variable sample size. Effect of missingness on accuracy of imputation of standardised effects, evaluated via simulations where true effect is known. The y-axis is the MSE (on log-scale) between the true standardised effect and the conventional estimate which ignores missingness (Eq (1), grey), our estimate $D^{(dep)}$ (Eq (10), green), and our estimate $D^{(ind)}$ (Eq (11), blue). The x-axis is the ‘missingness-correlation’ (θ_{miss}), where a value of 1 means the number of individuals in the samples had maximum overlap with each other, and 0 means they were simulated independently leading to smaller overlap. Each boxplot shows the MSEs across the 40 regions simulated. Top row is where the N ’s (simulated sample sizes) are selected randomly from a study of T2D [31], with sample sizes varying between 13 and 110’219 individuals. Bottom row is based on HDL [30], with sample sizes ranging between 50’000 and 187’167 individuals. All sample sizes are scaled to 0-to-12500 as this is the size of the simulated GWAS.

<https://doi.org/10.1371/journal.pgen.1007371.g003>

among which 535 contain SNVs associated with height [12, 13], while the remaining 171 regions were selected to be free of any known height associated SNVs.

We examined imputation results for different SNV categories. These were grouped based on (i) their association status (being correlated with the causal SNV vs. null SNVs) with the lead SNV of each of the 535 height-associated regions (6’080 variants were correlated, 31’567 were not); (ii) frequency (MAF: 1% < low-frequency \leq 5% < common; 13’857 and 23’790 variants, respectively); and (iii) imputation quality based on *summary statistics imputation* ($\hat{r}_{pred,adj}^2$: low \leq 0.3 < medium \leq 0.7 < high; 724, 9’792, and 27’131 variants, respectively). S1 and S2 Figs show the distribution of SNV counts in each of these twelve subgroups. We term the 6’080 SNVs correlated with a height-associated lead SNV as *associated* SNVs. Conversely, we refer to the 31’567 SNVs that are not correlated with any height-associated lead SNV as *null* SNVs. For both, null and associated SNV groups, the largest group of analysed variants were common and well-imputed (S1 Fig). The fraction of SNVs with low quality imputation increases with lower minor allele frequency (S2 Fig). However, the number of rare variants (MAF < 1%) were too small (2’411 variants, among these only 13 associated variants), similar

to the number of badly-imputed SNVs (724 variants, among these only one associated variant) to draw meaningful conclusions and hence we limited our analysis to common and low-frequency, and medium- and well-imputed variants.

We focused on two aspects of the imputation results. First, we compared how *summary statistics imputation* and *genotype imputation* perform relative to the ground truth (direct genotyping). For this we used four measures: the root mean squared error (RMSE), bias, the linear regression slope, and the correlation. Second, we calculated power and false positive rate for *genotype imputation* and *summary statistics imputation* directly.

Genotype imputation outperforms summary statistics imputation for low allele frequency. Fig 4 shows in green the comparison between summary statistics resulting from measured genotype data (ground truth) and imputed summary statistics for 6'080 height-associated variants. As expected, the performance drops as the imputation quality and as the MAF decrease. For well-imputed common SNVs (the largest subgroup with 5'714 variants), *summary statistics imputation* performs on average well with a correlation and a slope close to 1 ($\text{cor} = 0.998$ and $\text{slope} = 0.98$), but it drops to $\text{cor} = 0.928$ and a $\text{slope} = 0.83$ for low imputation quality, low-frequency variants. On the other hand, for *genotype imputation* (Fig 4, blue dots) all subgroups of SNVs show near perfect slope and correlation. Note that imputation quality for *summary statistics imputation* and *genotype imputation* differ in definition and we find that the latter was consistently higher (S3 and S4 Figs) and showed little variation across SNVs. To be able to compare the performance between *genotype imputation* and *summary statistics imputation* for the same subgroups of SNVs we used the imputation quality defined by *summary statistics imputation* to classify SNVs.

For the 31'567 null SNVs we present the same metrics as for associated SNVs. We analysed 13'556 low-frequency and 18'011 common variants. First, the green dots in Fig 5 show summary statistics from genotype data and *summary statistics imputation*. We find that both the correlation and slope gradually decrease with dropping imputation quality and MAF. For example, the correlation is 0.91–0.94 for well-imputed, 0.73–0.76 for medium and 0.42–0.66 for badly-imputed SNVs. The blue dots in Fig 5 show the respective results for *genotype imputation*, which exhibits an almost perfect (> 0.98) slope and correlation.

Effect estimate accuracy and precision. We then compared *summary statistics imputation* and *genotype imputation* in terms of RMSE among associated variants (for the same six SNV categories), shown in the upper part of Table 1. For all six subgroups, *genotype imputation* had a smaller RMSE than *summary statistics imputation*. The difference between the two methods in terms of RMSE increases as imputation quality decreases. For the largest SNV subgroup—well-imputed and common SNVs—*summary statistics imputation* had a RMSE of 0.33 versus 0.093 for *genotype imputation*. In case of *summary statistics imputation*, the RMSE is more influenced by a decrease in imputation quality than by a reduction of MAF. For example, the RMSE for common variants with medium-quality imputation is 1.02 (a 3.1-fold increase), while the RMSE for low-frequency variants with high-quality imputation is 0.48 (a 1.4-fold increase). However, for *genotype imputation* a decrease in MAF or imputation quality seems to have a similar effect. For example, the RMSE for well-imputed, low-frequency variants is 0.14 for *genotype imputation* (a 1.5-increase), and the RMSE for medium-imputed, common variants is 0.19 for *genotype imputation* (a 2.1-increase) (Fig 6). For null SNVs we observe for *summary statistics imputation* a RMSE of 0.38 for well-imputed and common SNVs up to 0.95 for badly-imputed and low-frequency SNVs (lower part in Table 1). For *genotype imputation* the RMSE ranges are much lower, between 0.09 for badly-imputed and common SNVs and 0.19 for badly-imputed and low-frequency SNVs. The bias is very close to zero for both approaches and for null and associated SNVs, and does not significantly vary with MAF or imputation quality.

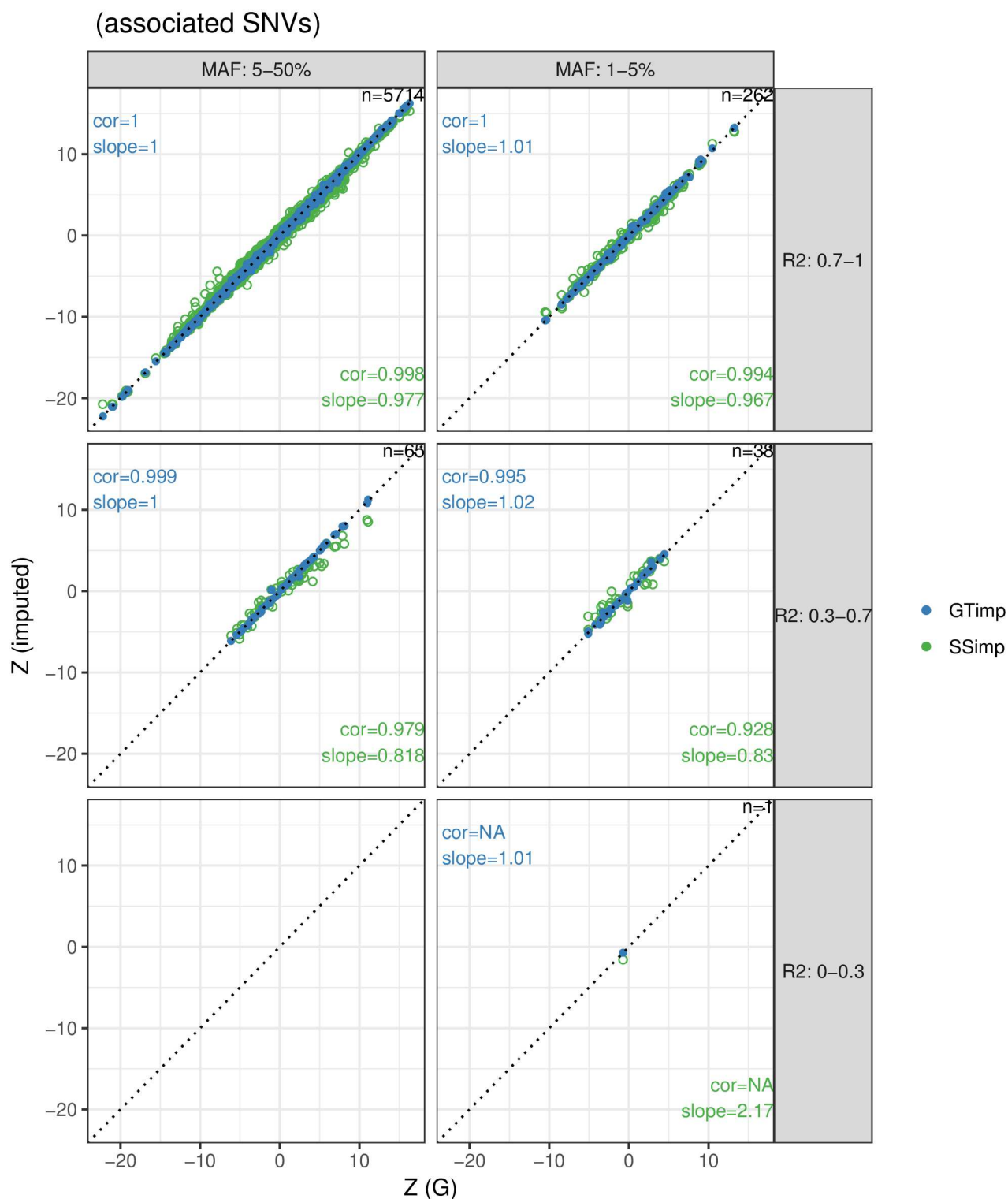


Fig 4. Summary statistics imputation versus genotype imputation for associated variants. The x-axis shows the Z-statistics of the genotype data (ground truth), while the y-axis shows the Z-statistics from *summary statistics imputation* (green) or *genotype imputation* (blue). Results are grouped according to MAF (columns) and imputation quality (rows) categories and the numbers top-right in each window refers to the number of SNVs represented. The identity line is indicated with a dotted line. The estimation for correlation and slope are noted in the bottom-right corner for *summary statistics imputation* and in the top-left corner for *genotype imputation*. Blue dots are plotted over the green ones. [S11](#) and [S13](#) Figs provide scatterplots with the imputation quality of *summary statistics imputation* and *genotype imputation* as colors.

<https://doi.org/10.1371/journal.pgen.1007371.g004>

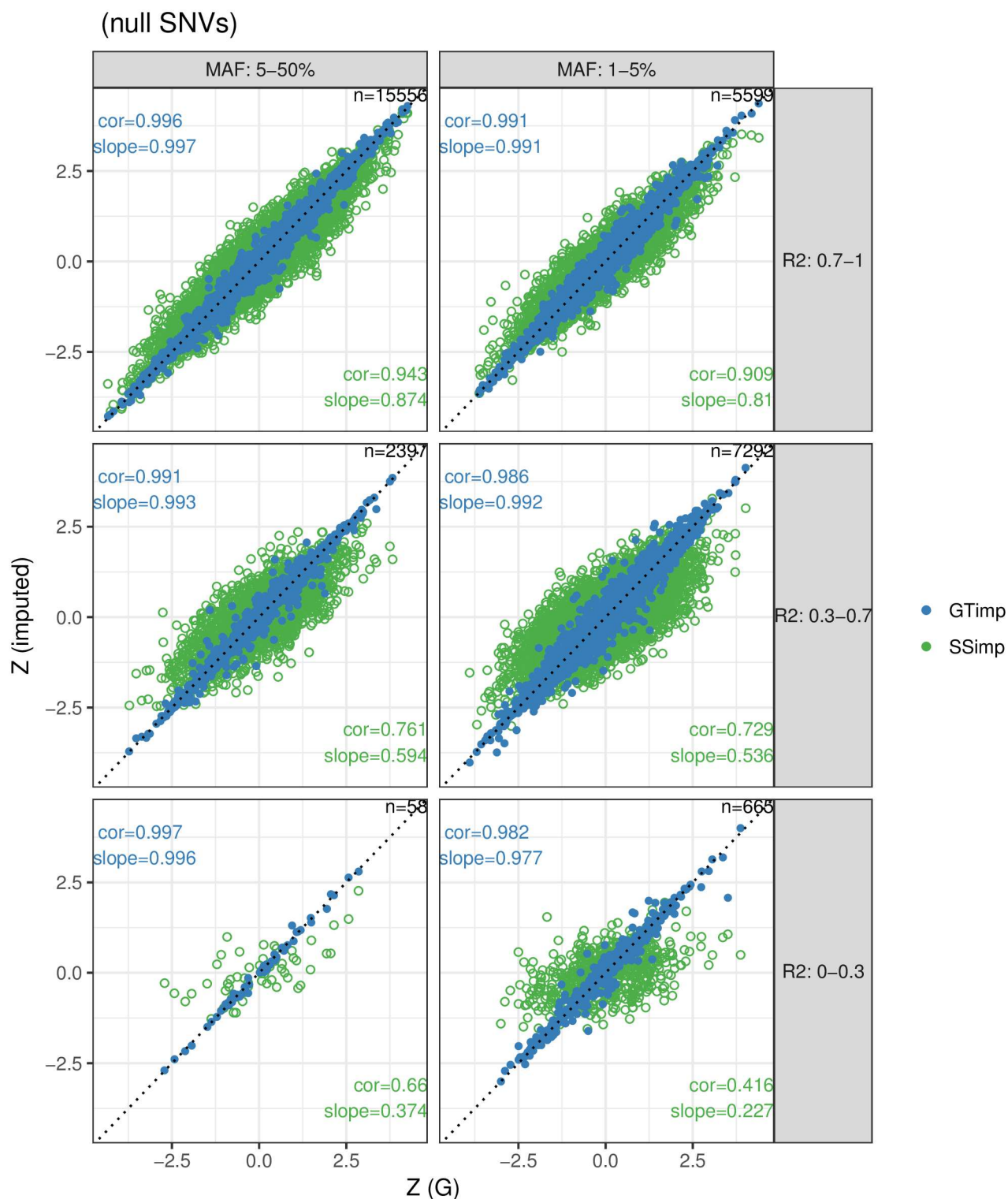


Fig 5. Summary statistics imputation versus genotype imputation for null variants. The x-axis shows the Z-statistics of the genotype data (ground truth), while the y-axis shows the Z-statistics from *summary statistics imputation* (green) or *genotype imputation* (blue). Results are grouped according to MAF (columns) and imputation quality (rows) categories and the numbers top-right in each window refers to the number of SNVs represented. The identity line is indicated with a dotted line. The estimation for correlation and slope are noted in the bottom-right corner for *summary statistics imputation* and in the top-left corner for *genotype imputation*. Blue dots are plotted over the green ones. [S12](#) and [S14](#) Figs provide scatterplots with the imputation quality of *summary statistics imputation* and *genotype imputation* as colors.

<https://doi.org/10.1371/journal.pgen.1007371.g005>

Table 1. RMSE for *summary statistics imputation* and *genotype imputation*.

	MAF	$\hat{r}_{\text{pred,adj}}^2$	SSimp		GTimp		# SNVs
			RMSE	Bias	RMSE	Bias	
Associated	1-5%	0-0.3	0.8484	-0.8484	0.0059	-0.0059	1
	1-5%	0.3-0.7	1.0120	0.1960	0.2729	0.0170	38
	1-5%	0.7-1	0.4785	-0.0137	0.1407	0.0073	262
	5-50%	0.3-0.7	1.0266	-0.3455	0.1916	-0.0041	65
	5-50%	0.7-1	0.3333	0.0011	0.0929	-0.0023	5714
Null	1-5%	0-0.3	0.9479	-0.0267	0.1944	0.0083	665
	1-5%	0.3-0.7	0.7262	0.0006	0.1765	0.0006	7292
	1-5%	0.7-1	0.4549	-0.0002	0.1491	0.0022	5599
	5-50%	0-0.3	0.8780	0.0057	0.0926	-0.0077	58
	5-50%	0.3-0.7	0.6906	-0.0115	0.1445	-0.0013	2397
	5-50%	0.7-1	0.3816	-0.0010	0.1022	-0.0004	15556

This table shows RMSE and bias for *summary statistics imputation* (SSimp) and *genotype imputation* (GTimp) in each variant subgroup (based on MAF and imputation quality) for associated SNVs (upper rectangle) and null SNVs (lower rectangle). The rightmost column reports the number of variants in each SNV subgroup. For MAF and $\hat{r}_{\text{pred,adj}}^2$ notation, the lower bound is excluded while the upper bound is included. For example, 1 – 5% is equivalent to $1 < \text{MAF} \leq 5$. RMSE differences are also displayed in Fig 6.

<https://doi.org/10.1371/journal.pgen.1007371.t001>

Summary statistics imputation displays lower false positive rate. Analogous to a ROC curve Fig 7 presents simultaneously power and false positive rate (FPR) with varying significance threshold (α from 0 to 1) for simulated phenotypes. As before, we stratified the results by MAF and imputation quality categories. We observe that for common SNVs with $\hat{r}_{\text{pred,adj}}^2 > 0.7$ the results for *genotype imputation* and *summary statistics imputation* are almost identical in terms of FPR and power. For low-frequency and well-imputed variants, *genotype imputation* offers some power advantage compared to *summary statistics imputation*, in particular for intermediate FPRs. As we approach lower imputation quality and MAF, *genotype imputation* advantage becomes more and more apparent for all range of FPR values. Averaged over all SNV categories, for false positive rates of 0.001, 0.01, 0.05, *summary statistics imputation* yielded a decrease in statistical power by 9, 43 and 35%, respectively.

Summary statistics imputation of the height GWAS of the GIANT consortium

While previous studies have examined the role of (common) HapMap variants for height [12, 36], the impact of rare coding variants could not be investigated until bespoke genotyping chips (interrogating low-frequency and rare coding variants) were designed to address this question in a cost-effective manner. Such an exome chip based study was conducted by the GIANT consortium in 381'000 individuals and revealed 120 height-associated loci, of which 83 loci were rare or low-frequency [13]. These association results enabled us to compare the usefulness of imputation-based inference with direct genotyping done in Wood *et al.* [12], since the two studies are highly comparable in terms of ancestry composition and statistical analysis, evidenced by S6 Fig confirming very high concordance between summary statistics for the subset of 2'601 SNVs correlated to a height-associated variant which were available in both studies.

Discovery and replication of 19 new loci. By imputing > 6M additional SNVs summary statistics using HapMap variants [12] as tag SNPs we were interested in two aspects: (1) discovering new height-associated candidate loci, and (2) replicating these candidate

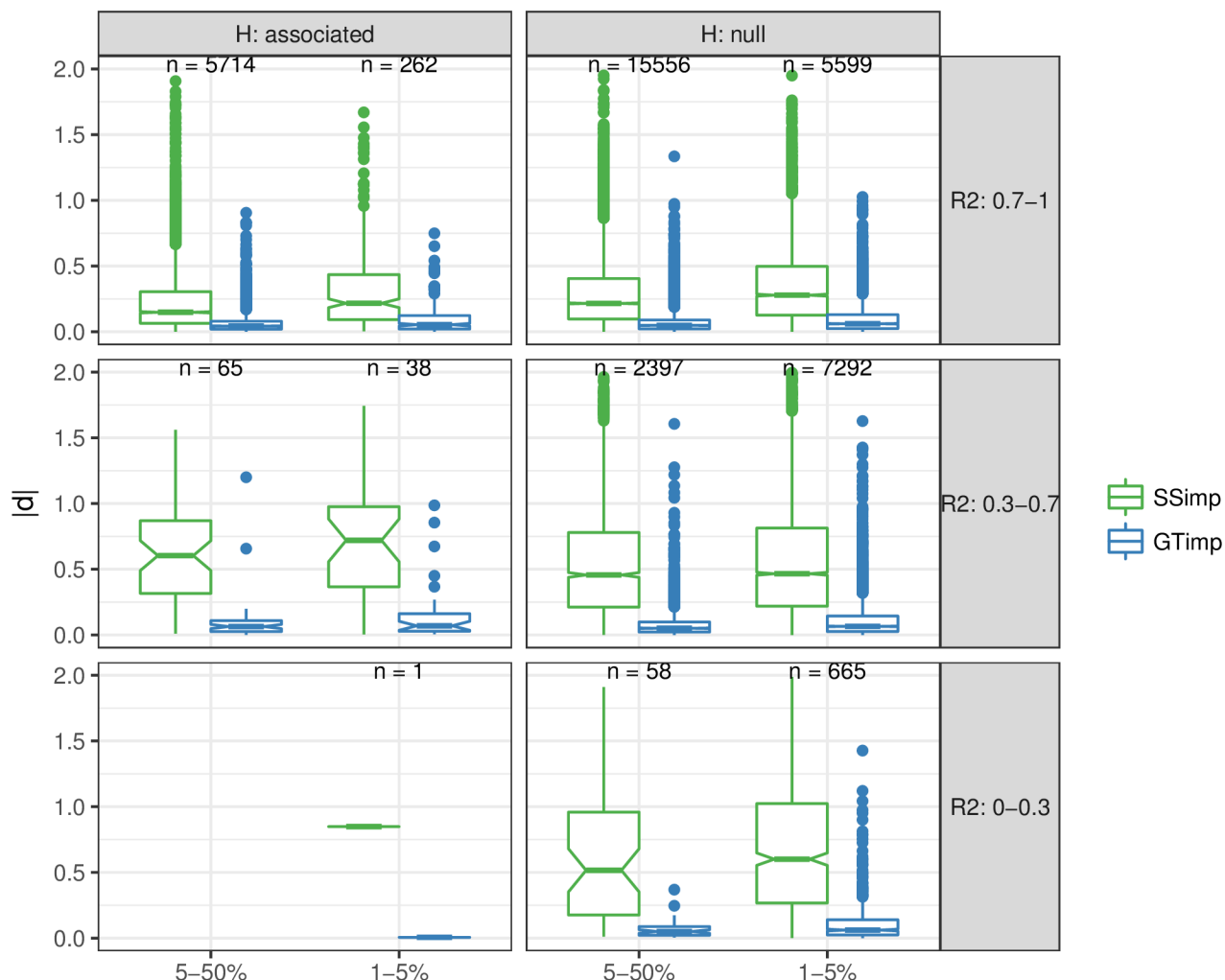


Fig 6. Visualising RMSE of summary statistics imputation and genotype imputation. This figure uses boxplots to compare the absolute difference $|d|$ (used for calculation of RMSE) for each variant between Z-statistics of *summary statistics imputation* (SSimp, green) and *genotype imputation* (GTimp, blue) of associated SNVs (left column) and null SNVs (right column). Results are grouped according to MAF (x-axis) and imputation quality (rows) categories. The numbers printed above the boxplot represents the number of SNVs used for the $|d|$ calculation in that MAF and imputation quality subgroup. The corresponding $RMSE = \sqrt{\frac{1}{n} \sum_i d_i^2}$ is shown in Table 1.

<https://doi.org/10.1371/journal.pgen.1007371.g006>

loci in the UK Biobank and the GIANT exome chip look-up (Fig 2). We used the HapMap-based height study and the UK10K reference panel as inputs for *summary statistics imputation* and used all HapMap SNVs as tag SNVs. We imputed variants that were available in UK10K with a $MAF_{UK10K} \geq 0.1\%$, as well as all reported exome variants in Marouli *et al.* [13]. In total we imputed 10'966'111 variants, of which 9'276'018 (84%) had an imputation quality ≥ 0.3 .

We subjected all 9'276'018 variants with an imputation quality ≥ 0.3 to a scan for novel candidate loci. A region was defined as a candidate locus if at least one imputed variant was independent from any reported HapMap variant nearby (conditional P -value $\leq 10^{-8}$). We identified 35 such candidate loci. Within each locus we defined the imputed variant with the lowest conditional P -value as the top variant. All 35 variants are listed in S1 Table and locus-zoom plots are provided in S7 Fig.

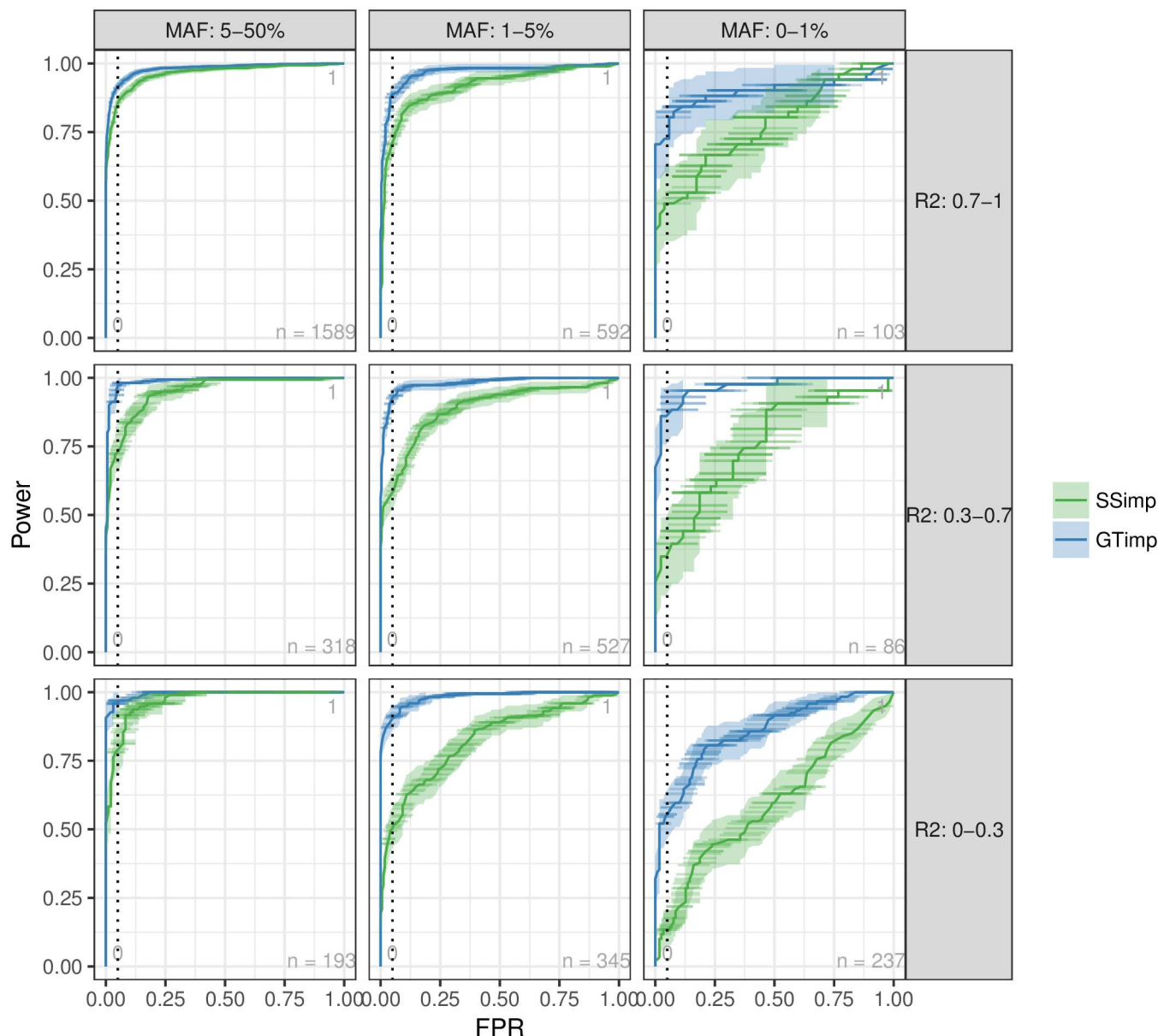


Fig 7. FPR versus power. This figure compares the false positive rate (FPR) (x-axis) versus the power (y-axis) for *genotype imputation* (blue) and *summary statistics imputation* (green) for different significance thresholds (α), including a 95%-confidence interval in both directions (vertically as a ribbon and horizontally as lines). The vertical, dashed line represents FPR = 0.05. Results are grouped according to MAF (columns) and imputation quality (rows) categories. A zoom into the area of FPR between 0 and 0.1 can be found in S5 Fig.

<https://doi.org/10.1371/journal.pgen.1007371.g007>

Next, we used the UK Biobank to replicate the associations with height of these 35 candidate variants and subsequently grouped them into replicating (20 variants) and not replicating (15 variants) (at $\alpha = 0.05/35$ level).

An overview of the 20 replicating variants is given in Table 2. One region had already been discovered in the GIANT exome chip study: rs28929474, located in gene *SERPINA1*. Fig 8 shows this region as locus-zoom plot with summary statistics from the HapMap study, *summary statistics imputation*, and the exome chip study. To annotate these 20 novel candidate variants further, we investigated whether they are eQTLs or associated with other traits. We

Table 2. Twenty replicating candidate loci for height.

#	SNV	Chr	Pos	Allele	Gene ⁽¹⁾	MAF ⁽²⁾	SSimp		UK Biobank		Group
				R/E			P	N	P	N	
1	rs112635299(*)	14	94838142	G/T	-	2.33%	4.21E-14	234380	5.16E-77	336474	(i)
2	rs76306191	1	155006451	C/G	<i>DCST1</i> [E]	20.30%	6.51E-10	245908	2.74E-16	336474	(ii)
3	rs73029259	6	164111348	T/A	-	12.77%	7.61E-09	251161	1.02E-15	336474	(ii)
4	rs67807996	1	149995265	G/A	-	40.16%	1.48E-43	219605	2.75E-102	336474	(iii)
5	rs12795957	11	67242216	G/A	-	5.46%	1.52E-24	193457	1.75E-76	336474	(iii)
6	rs503035	5	134353734	A/G	-	30.39%	6.34E-24	248110	5.46E-39	336474	(iii)
7	rs568777	6	81809121	C/G	-	26.61%	7.08E-24	252456	3.11E-35	336474	(iii)
8	rs75975831	19	17264961	G/C	<i>MYO9B</i> [I]	22.52%	3.59E-10	233765	9.19E-22	336474	(iii)
9	rs56006730	12	103132740	G/A	-	10.41%	1.80E-09	250070	1.05E-19	336474	(iii)
10	rs35374532	6	26163345	A/AT	<i>HIST1H2BD</i> [I]	38.85%	2.97E-27	252327	8.64E-18	120086	(iii)
11	rs80171383	11	46084677	G/A	<i>PHF21A</i> [I]	14.72%	3.53E-16	247885	2.05E-16	336474	(iii)
12	rs13108218	4	3443931	A/G	<i>HGFAC</i> [I]	39.72%	2.15E-10	222502	5.05E-15	336474	(iii)
13	rs428925	5	173022921	G/A	-	27.59%	1.34E-16	206987	4.31E-13	336474	(iii)
14	rs6085649	20	6665532	A/G	-	45.61%	1.24E-09	251393	1.65E-12	336474	(iii)
15	rs78566116	6	32396146	G/T	-	7.67%	2.74E-19	248592	4.18E-12	336474	(iii)
16	rs350889	19	4118481	A/G	<i>MAP2K2</i> [I]	24.28%	8.17E-10	207571	7.11E-12	336474	(iii)
17	rs7955819	12	20677958	T/C	<i>PDE3A</i> [I]	23.23%	6.13E-10	250048	3.25E-08	336474	(iii)
18	rs7971674	12	1513526	A/T	<i>ERC1</i> [I]	14.12%	8.10E-09	240270	2.19E-07	336474	(iii)
19	rs12939056	17	7754993	G/A	<i>KDM6B</i> [E]	43.26%	1.09E-12	245015	7.64E-07	336474	(iii)
20	rs58402222	1	46059835	T/TA	<i>NASP</i> [I]	45.72%	7.50E-13	252901	1.79E-04	120086	(iii)

This table presents 20 regions that contain at least one imputed variant that is independent from top HapMap variants nearby and that replicated in the UK Biobank (at $\alpha = 0.05/35$ level). Each row represents one region (#), indicating the SNV with the lowest conditional *P*-value. The first seven columns provide general information for each variant, followed by the *P*-value and sample size from *summary statistics imputation*, *P*-value and sample size from the UK Biobank. The second last column assigns each of the 35 candidate loci to one of three groups: candidate loci (i) that were reported by [13] already, (ii) that had no reported HapMap variant nearby and (iii) that had reported HapMap variants nearby. $\hat{r}^2_{\text{pred,adj}}$ of all variants listed was greater than or equal to 0.3. We provide a more detailed table for all 35 variants (both replicating and not replicating) in S1 Table.

(*) rs28929474, exome chip study results: $P = 1.39 \times 10^{-45}$, $N = 365'451$.

⁽¹⁾ [I] intronic, [E] exonic, - intergenic.

⁽²⁾ MAF was computed in UK10K.

<https://doi.org/10.1371/journal.pgen.1007371.t002>

report this in Table 3 where we list eQTLs detected by GTEx [27] and Table 4 that presents a curated association-trait list by Phenoscanner [28]. In the following we describe variants that replicated in UK Biobank which are either eQTLs or have previously been associated with another trait.

We can classify the 35 candidate loci into three categories that reflect the type of conditional analysis performed. Group (i) includes SNVs replicating already published exome chip associations (one locus), group (ii) includes SNVs that contain no reported HapMap variant nearby (three loci), and group (iii) includes SNVs that contain one or more reported independent HapMap variants nearby (31 loci). Replication success with UK Biobank is 1/1 in group (i), 2/3 in group (ii), 17/31 in group (iii). We only term categories (ii) and (iii) as *novel* candidate loci, therefore limiting the number of novel candidate loci to 34, with 19 replicating in UK Biobank.

Although group (ii) only contains loci that had no reported HapMap variants nearby, three candidate loci (#2, #3, #21 in S1 Table) contain borderline significant HapMap signals (*P*-value between 10^{-6} and 10^{-8} in [12]).

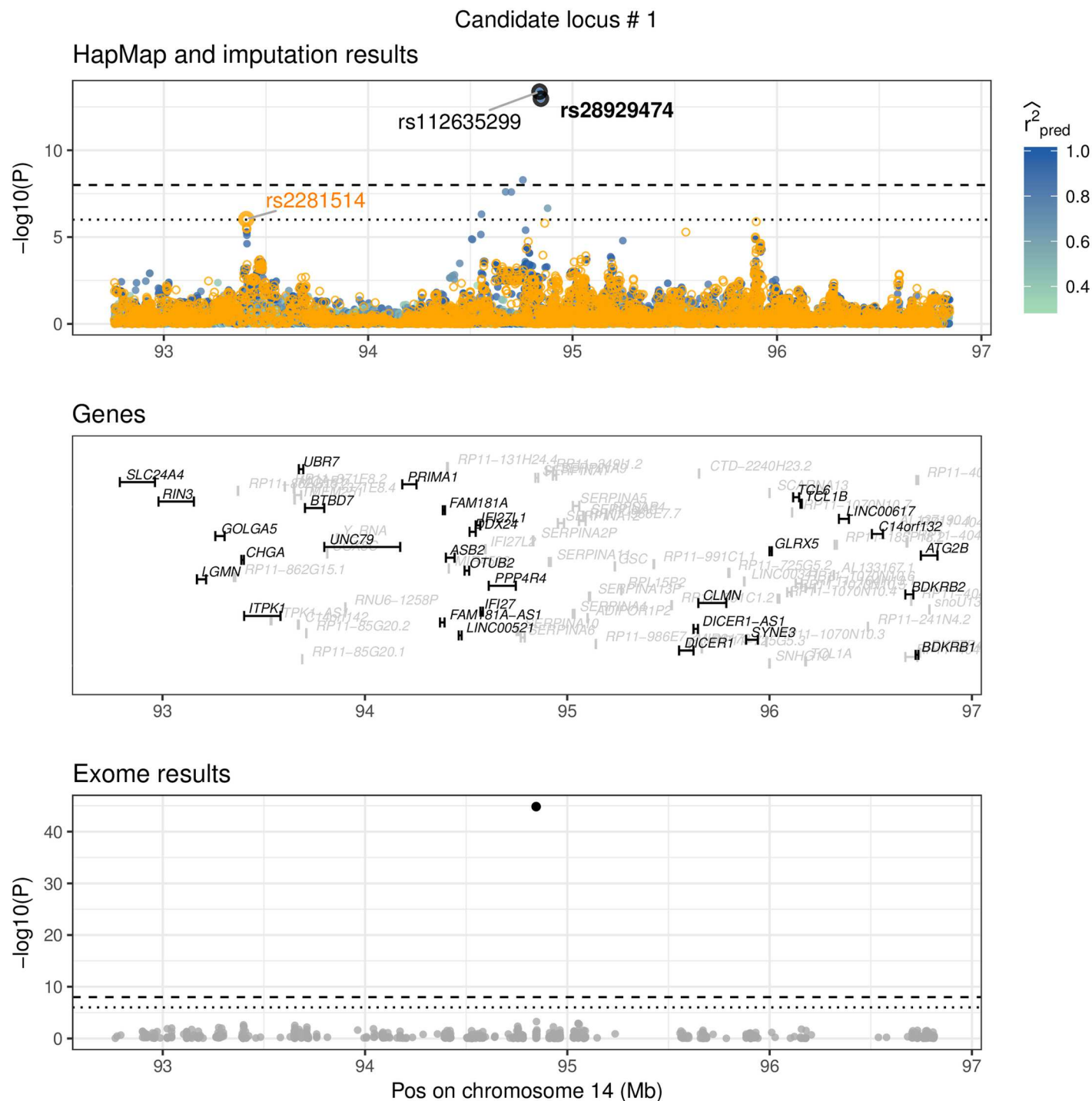


Fig 8. Replication of exome variant. rs28929474 is a missense variant on chromosome 14 in gene *SERPINA1*, low-frequency (MAF = 2.3%), imputed summary statistics ($P_{\text{Simp}} = 1.06 \times 10^{-13}$), replication in the UK Biobank ($P_{\text{UKBB}} = 6.49 \times 10^{-78}$). rs112635299 has the strongest signal in this region ($P = 4.21 \times 10^{-14}$), but is highly correlated to rs28929474 ($LD = 0.95$). This figure shows three datasets: Results from the HapMap and the exome chip study, and imputed summary statistics. The top window shows HapMap P -values as orange circles and the imputed P -values (using *summary statistics imputation*) as solid circles, with the colour representing the imputation quality (only $\hat{r}_{\text{pred,adj}}^2 \geq 0.3$ shown). The bottom window shows exome chip study results as solid, grey dots. Each dot represents the summary statistics of one variant. The x-axis shows the position (in Mb) on a ≥ 2 Mb range and the y-axis the $-\log_{10}(P)$ -value. The horizontal line shows the P -value threshold of 10^{-6} (dotted) and 10^{-8} (dashed). Top and bottom window have annotated summary statistics: In the bottom window we mark dots as black if it is part of the 122 reported hits of [13]. In the top window we mark the rs-id of variants that are part of the 122 reported variants of [13] in bold black, and if they are part of the 697 variants of [12] in bold orange font. Variants that are black (plain) are imputed variants (that had the lowest conditional P -value). Variants in orange (plain) are

HapMap variants, but were not among the 697 reported hits. Each of the annotated variants is marked for clarity with a bold circle in the respective colour. The genes annotated in the middle window are printed in grey if the gene has a length < 5'000 bp or is an unrecognised gene (RP-).

<https://doi.org/10.1371/journal.pgen.1007371.g008>

We observed that variants with higher MAF have higher chance to replicate. Among the 20 candidate variants that did replicate in UK Biobank, 19 were common and one a low-frequency variant (*rs112635299*, MAF = 2.32%). Conversely, among the 15 candidate variants that did not replicate in the UK Biobank, 10 are rare, three are low-frequency variants, and only two are common.

Locus #1: *rs112635299* (imputed P -value 4.21×10^{-14}), is a proxy of *rs28929474* (LD = 0.88), has been associated with alpha-1 globulin [37] and is associated with multiple lipid metabolites [38]. *rs28929474* was identified in the GIANT exome chip study to be height-associated ($P = 1.39 \times 10^{-45}$) [13]. The P -value calculated with *summary statistics imputation* was $P = 1.06 \times 10^{-13}$. *rs28929474* is a low-frequency variant (MAF = 2.3%) and replicates in the UK Biobank with $P = 1.66 \times 10^{-25}$.

Locus #2: *rs76306191* is a common variant on chromosome 1, located in gene *DCST1*. There was no reported HapMap variant nearby to condition on. However, the absolute correlation to the HapMap variant with the lowest P -value ($> 10^{-8}$) in the same region was 0.8. One of the 122 variants reported by the exome chip study, *rs141845046*, was in this region, but had an imputed P -value $> 10^{-3}$. *rs76306191* replicated in the UK Biobank with $P = 1.09 \times 10^{-7}$. *rs76306191* is an eQTL in artery (tibial) for gene *ZBTB7B* and in thyroid gland for gene *DCST2*.

Locus #5: *rs12795957* is a variant on chromosome 11 and an eQTL for gene *RAD9A* in artery (tibial).

Locus #6: *rs503035* is a variant on chromosome 5. It is an eQTL for gene *PITX1* in testis tissue. *rs62623707*, one of the 122 reported exome variants, was in this region, but had an imputed P -value $> 10^{-3}$.

Locus #15: *rs78566116* is a variant on chromosome 6. *rs78566116* has been associated with HPV8 seropositivity in cancer [39], rheumatoid arthritis [40] and ulcerative colitis [41].

Table 3. GTEx annotation results for variants in eQTLs.

#	SNV	P_{SSimp}	P_{UKBB}	GTEx tissue	Gene	P
2	<i>rs76306191</i>	6.51E-10	1.09E-07	Artery_Tibial	<i>ZBTB7B</i>	3.97E-09
				Thyroid	<i>DCST2</i>	2.41E-08
5	<i>rs12795957</i>	1.52E-24	6.17E-41	Artery_Tibial	<i>RAD9A</i>	6.48E-10
6	<i>rs503035</i>	6.34E-24	8.06E-12	Testis	<i>PITX1</i>	2.91E-07
20	<i>rs58402222</i>	7.50E-13	1.79E-04	Cells_Transformed_fibroblasts	<i>MAST2</i>	8.84E-23
				Cells_Transformed_fibroblasts	<i>CCDC163P</i>	1.11E-19
				Cells_Transformed_fibroblasts	<i>TMEM69</i>	2.16E-08
				Thyroid	<i>GPBP1L1</i>	3.26E-11

This table shows SNVs which are significant eQTLs in GTEx [27]. We only report SNV-gene expression associations where the summary statistics pass the significance threshold of $\alpha = 10^{-6}$. The first four columns represent the region number, SNV, P -value from *summary statistics imputation* and the P -value in the UK Biobank. The four remaining columns are information extracted from GTEx, with the tissue name, gene name, the P -value of the association between the SNV and the gene expression, and the gene type. For each region, we only include the tissue with the lowest P -value per SNV-gene associations. The full version of this table is available in S2 Table. # refers to the region number.

<https://doi.org/10.1371/journal.pgen.1007371.t003>

Table 4. Known trait association results for variants in Table 2.

#	SNV	P_{SSimp}	P_{UKBB}	Study	PMID	Ancestry	Trait	P	N
1	rs112635299	4.21E-14	3.52E-25	Wood	23696881	Mixed	Alpha 1 globulin	2.51E-12	5278
				Kettunen J	27005778	European	Glycoprotein acetyls	1.27E-10	17772
							mainly a1Lacid glycoprotein		
				Kettunen J	27005778	European	Total cholesterol in small LDL	6.59E-10	20057
				Kettunen J	27005778	European	M.LDL.C	4.03E-09	20060
				Kettunen J	27005778	European	Cholesterol esters in medium LDL	6.19E-09	17774
				Kettunen J	27005778	European	Total lipids in medium LDL	7.26E-09	17774
				Kettunen J	27005778	European	Total cholesterol in LDL	8.66E-09	20060
				Kettunen J	27005778	European	Total lipids in small LDL	1.56E-08	17774
				Kettunen J	27005778	European	Conc. of medium LDL particles	1.67E-08	17774
				Kettunen J	27005778	European	Conc. of small LDL particles	2.77E-07	17774
				Kettunen J	27005778	European	Cholesterol esters in large LDL	4.72E-07	17774
				Kettunen J	27005778	European	Total cholesterol in large LDL	7.36E-07	20053
				Kettunen J	27005778	European	Total lipids in large LDL	9.86E-07	17774
15	rs78566116	2.74E-19	9.80E-04	Chen D	21896673	Mixed	HPV8 seropositivity in cancer	3.30E-16	6885
				Okada Y	24390342	European	Rheumatoid arthritis	3.80E-94	58284
				Okada Y	24390342	Mixed	Rheumatoid arthritis	2.30E-90	80799
				IBDGC	26192919	European	Ulcerative colitis	4.06E-08	27432

This table describes SNVs previously associated with other traits. The search was conducted with Phenoscanner [28]. We only list SNVs for which Phenoscanner had information available regarding GWAS traits or metabolites. The first four columns specify region, SNV-id, followed by the P -value from *summary statistics imputation* and the P -value from the UK Biobank. Column five to ten contain information extracted from Phenoscanner. We report the respective summary statistics that pass the significance threshold of $\alpha = 10^{-6}$. # refers to the region number, conc. to concentration.

<https://doi.org/10.1371/journal.pgen.1007371.t004>

Locus #20: rs58402222 is an intronic variant on chromosome 1, located in gene *NASP*. It is an eQTL for genes *CCDC163P*, *MAST2* and *TMEM69* in cells (transformed fibroblasts); and for *GPBP1L1* in thyroid tissue.

Replication of 55/111 reported GIANT exome chip variants. Next, we focussed on 122 novel variants of Marouli *et al.* [13]. For this analysis we did not apply any MAF restrictions. Of these 122 variants, 11 variants were either not referenced in UK10K or on chromosome X, and were therefore not imputed, limiting the number of loci and variants to 111—78 common, 25 low-frequency, eight variants rare (S3 Table). By grouping results below or above the P -value threshold of $\alpha = 0.05/111$ we could classify variants into the ones that replicated and those that failed replication. This is summarised in Table 5 and S8 Fig, which shows that 55 of

Table 5. 111 variants: Fraction of top variants in exome chip study retrieved with imputation of HapMap study.

$\hat{r}_{\text{pred.adj}}^2$	MAF		
	5 – 50%	1 – 5%	0 – 1%
0.7-1	65% (49/75)	50% (4/8)	-
0.3-0.7	67% (2/3)	0% (0/17)	0% (0/3)
0-0.3	-	-	0% (0/5)

This table presents *summary statistics imputation* results, limited to 111 variants identified as “novel” by [13]. We summarised the results according to their allele frequency and imputation quality category. For each subgroup we calculated the fraction of top exome variants that had a P -value $\leq 0.05/111$ with *summary statistics imputation*.

<https://doi.org/10.1371/journal.pgen.1007371.t005>

the 111 variants could be retrieved, four of them with $MAF \leq 5\%$. When looking at imputation quality, of the 111 top variants 83 variants were imputed with high confidence ($\hat{r}_{pred,adj}^2 \geq 0.7$). Of these, 53 were retrieved when using the typical candidate SNV threshold (0.05/111). Details to the imputation of all 111 variants are listed in [S3 Table](#).

Discussion

In this article, we focussed on the comparison between *genotype* and *summary statistics imputation*. In contrast to previous work by others [9, 14], here we systematically assessed the performance and limitations of *summary statistics imputation* through real data applications for different SNV subgroups characterised by allele frequency, imputation quality and association status (null/associated).

First, we adapted the published *summary statistics imputation* method [9], by allowing the LD structure to be adaptive according to varying sample size in summary statistics of tag SNVs. Our simulation study has shown that this version of *summary statistics imputation* has a lower MSE in all scenarios. We then evaluated the performance of our improved *summary statistics imputation* method in terms of different measures and showed that *summary statistics imputation* is a very efficient and fast method to separate null from associated SNVs. However, *genotype imputation* outperforms *summary statistics imputation* by a clear margin in terms of accuracy of effect size estimation. By imputing GIANT HapMap-based summary statistics we have demonstrated that *summary statistics imputation* is a rapid and cost-effective way to discover novel trait associated loci. We also highlight that the principal limitations of *summary statistics imputation* are rooted in the LD estimation and in imputing very rare variants with sufficient confidence. Finally, we implemented *summary statistics imputation* that accounts for varying sample size as a command-line tool [11].

Accounting for varying sample size

Imputation accuracy is affected by the varying sample size across tag SNVs. If two SNVs were observed in two different samples, the correlation between the summary statistics will decrease with the number of individuals in common between the two samples. Our approach addresses this problem by shrinking the correlation matrix according to sample size overlap. We present two ways of estimating this overlap: $D^{(ind)}$ for *independent* missingness, which is randomly distributed; and $D^{(dep)}$ for *dependent* missingness, which is highly correlated.

To evaluate the performance of these two methods we simulated data with two different distributions of missingness (narrow or wide range of sample sizes) and varying correlation in missingness between variants (from completely random to maximal overlap, [Fig 3](#)). We then compared the performances of conventional *summary statistics imputation* and our proposed dependent ($D^{(dep)}$) and independent ($D^{(ind)}$) approaches. Overall, replacing C and c with D and d yields a lower RMSE. Furthermore, we note that the dependent approach has lower RMSE when the sample size variance is low and the missingness correlation approaches one. [S15 Fig](#) shows the comparison between the conventional estimation and using $D^{(dep)}$ for imputing GIANT height association summary statistics.

Ideally, for any pair of SNVs that are in LD with each other, we would know the exact number of individuals that are in the overlap, i.e. the number of individuals for which both SNVs were genotyped. Using the individual study missingness and sample sizes from the Genetic Investigation of ANthropometric Traits (GIANT) consortium, we demonstrate in [Fig. S10 Fig](#) that the size of the overlap is generally larger than would be the case under a strict ‘missing independently at random’ assumption. Furthermore, the correlation of missingness is typically positive ($N_{k \cap l} > \frac{N_k N_l}{N_{max}}$) and often approaches the maximum possible overlap ($N_{k \cap l} = \min(N_k, N_l)$).

The reason for this is that SNPs are either entirely missing from a study or being available for all study participants depending on its genotyping chip or imputation panel, which induces positive missingness correlation between markers.

Comparison of *summary statistics imputation* versus *genotype imputation*

We compared *summary statistics imputation* and *genotype imputation* by using individual-level data from the UK Biobank.

In general, imputation using *summary statistics imputation* leads to a larger RMSE than *genotype imputation* in all twelve SNV subgroups investigated (Fig 6). Among associated SNVs, *summary statistics imputation* performs similar to *genotype imputation* for well-imputed SNVs, but shows a trend for underestimation of the Z-statistics and lower correlation with the true effect size for medium-imputed SNVs (Fig 4). Conversely, *genotype imputation* has more consistent results for most of the twelve SNV subgroups (Figs 4 and 5), that is reflected in a correlation close to one between Z-statistics from genotype data and *genotype imputation* data.

When investigating power and FPR for both methods (Fig 7) we observe that for a given significance threshold, *summary statistics imputation* has lower power compared to *genotype imputation*, an effect that is amplified for SNVs with lower imputation quality ($\hat{r}_{\text{pred,adj}}^2 \leq 0.7$) and lower MAF ($\text{MAF} \leq 5\%$).

Underestimation for null and associated SNVs

Ultimately, the underestimation of imputed Z-statistics with *summary statistics imputation* leads to a lower type I error. This effect is amplified for SNV groups with lower imputation quality ($\hat{r}_{\text{pred,adj}}^2 < 1$). For associated SNVs with $\hat{r}_{\text{pred,adj}}^2 < 1$ we expect an underestimation for associated SNVs due to the fact that we are imputing summary statistics under the null model, whereas for null SNVs with $\hat{r}_{\text{pred,adj}}^2 < 1$ we expect an underestimation due to decreased variance of the *summary statistics imputation* estimation.

Ideally, for an unbiased estimation of causal and null SNVs, the imputed Z-statistics (Eq (2)) should be divided by \hat{r}^2 . However, as the imputation quality $\hat{r}_{\text{pred,adj}}^2$ is noisily estimated from small reference panels (discussed below) and it is not guaranteed that the SNV we impute is causal, we risk to overestimate the summary statistics of associated SNVs. This is the reason why refrain from doing so.

S9 Fig shows the *P*-value distribution of *summary statistics imputation* for null SNVs with an accumulation of low *P*-values for well-imputed SNVs and an accumulation of high *P*-values for badly-imputed SNVs. We think that two factors are in play here. First, mostly due to polygenicity, the genomic lambda for height is $\lambda_{GC} = 1.94$, therefore we expect even seemingly null variants to show inflation. Second, for null SNVs, the sample variance of the imputed Z-statistics should be proportional to the average imputation quality. We calculated for each of the null SNV subgroups the ratio between the sample variance for Z-statistics from *summary statistics imputation* and the sample variance for Z-statistics from genotype data. For common null SNVs we observe a ratio that gradually decreases with imputation quality (0.86 for perfectly-, 0.61 for medium- and 0.32 for badly imputed SNVs). For low-frequency null variants the ratio is up to 0.6 lower (0.80 for perfectly-, 0.54 for medium- and 0.30 for badly imputed SNVs). The inflation for well-imputed SNVs can be explained by the genomic lambda, while for badly-imputed SNVs it is aggravated by the underestimated standard error.

Atypical allele frequency distribution and rare variants exclusion

Because the number of associated SNVs with $MAF < 1\%$ was too low (13 variants) to draw any meaningful conclusions, we refrained from analysing this MAF group. One other reason to exclude rare variants from this analysis is, that the reference panel used (UK10K) contains 3'871 individuals and therefore estimations for LD of rare variants are unreliable and rare variants can (in theory) only be covered down to $MAF = 1/(2 \cdot 3'871)$. We believe improving *summary statistics imputation* for rare variants will require not only larger reference panels to allow estimation of LD of rare variants, but also methods which would allow non-linear tagging of variants. It should be kept in mind that, just like for *genotype imputation*, even with very large reference panels, one will not be able to impute variants with extremely rare allele counts. To investigate these SNVs full genome sequencing is indispensable [42].

Imputation quality metric discrepancies

We find that our imputation quality measure $\hat{r}_{pred,adj}^2$ is conservative and probably underestimates the true imputation quality (S4 Fig). To calculate the imputation quality $\hat{r}_{pred,adj}^2$, we need—similar to imputing summary statistics in Eq (2)—to compute correlation matrices c and C estimated from a reference panel (Eq (8)) and therefore encounter similar challenges as summary statistic imputation itself due to difficulties of reliable LD estimation.

The discrepancy in imputation quality metric between *summary statistics imputation* and *genotype imputation* (S4 Fig) can be explained by the fact that: (1) genotyped variants that were imputed too, were also used for phasing, (2) it is indeed more difficult to impute summary statistics using *summary statistics imputation*, and therefore the imputation quality is shifted towards zero, and (3) $\hat{r}_{pred,adj}^2$ is an estimation that can either be erroneous due to choosing the wrong reference panel (and therefore $\hat{r}_{pred,adj}^2$ does not represent the true imputation quality) or it can be imprecise due to small sample size of the reference panel. For example, UK10K contains 3'871 individuals and is too small to precisely estimate these matrices (the standard error for a correlation estimated from $n = 3'871$ is 0.016), which becomes problematic in cases of low correlation.

Summary statistics imputation of the height GWAS of the GIANT consortium

As a showcase of the utility of *summary statistics imputation* we imputed Wood *et al.* [12] to higher genomic resolution (limited to variants with $MAF \geq 0.1\%$ as well as 111 previously reported exome variants) [13], then selected imputed variants that act independently from all variants reported in Wood *et al.* and from each HapMap SNP, we then replicated these using (independent) UK Biobank data.

While Wood *et al.* [12] is the largest height study to date in terms of number of markers (covering HapMap variants in 253'288 individuals), Marouli *et al.* [13] exceeds their sample size by more than 100'000 individuals, but is limited to 241'419 exome variants. The similarity between both GIANT studies made the exome chip study ideal for replication. We chose the UK Biobank as a second replication dataset, despite its limitation to individuals of British ancestry, as it covers more variants than the exome chip study.

We identify 35 regions, of which one had already been identified in the recent GIANT height exome chip study ($rs28929474$) and 19 replicated in UK Biobank (at $\alpha = 0.05/35$ level). Two candidate loci (#2, #3 in Table 2) that replicate in UK Biobank have borderline significant HapMap signals in close proximity (P -value between 10^{-6} and 10^{-8} in [12]) and were therefore not reported in the study in 2014.

The 15 non-replicating candidate loci were on average on a lower allele frequency spectrum (ten are rare, three are low-frequency variants, and two are common). Allele frequency was higher among the 20 replicating candidate variants (19 were common and one a low-frequency variant).

We also ran an additional approximate conditional analysis, where we conditioned each of the 35 variants onto their neighbouring HapMap SNP (one-by-one). The resulting maximum conditional *P*-value per locus, is provided as an additional column [S1 Table](#). Correcting for the testing of 529 windows ($\alpha = 0.05/529$) we find evidence that 18 of the 35 variants are not only independent from all [12] reported SNPs, but also of each HapMap variant too.

Replicating GIANT exome chip imputation results. We then focussed on the *summary statistics imputation* of the the 111 reported exome chip variants [13]. Knowing from our previous findings that rare variants are challenging to impute due to reference panel size, we expected to retrieve a larger fraction of common and low-frequency than rare variants. Among variants with lower imputation quality only two common and medium-imputed variants could be retrieved. As shown in [Figs 4 and 7](#), the power of *summary statistics imputation* decreases with lower MAF and imputation quality.

Limitations

For replication of summary statistics from European individuals we use the UK Biobank, which represents only a subset of all European ancestries and is genotype-imputed (instead of sequenced), but on the other hand provides a reliable resource due to its sample size.

Furthermore, in UK Biobank, *genotype imputation* done for genotyped variants can only partially be compared to *genotype imputation* for untyped variants, as genotyped variants were used for phasing (therefore *genotype imputation* of genotyped variants is easier and leads imputation qualities close to one, [S4 Fig](#)). Due to the small number of height-associated rare variants (13) we can not draw meaningful conclusions for this group and hence avoided their analysis.

The choice of the reference panel to conduct summary statistics imputation depends on the fine balance between maximising the sample size of the reference panel (which determines the error in estimated LD) and matching the population diversity of the conducted GWAS. At the first glance, 1000 Genomes reference panel could have been used to appropriately match GIANT allele frequencies, however, the 8-fold higher sample size of UK10K panel offers a larger benefit, ultimately reducing the RMSE [43].

For the simulation study comparing standard *summary statistics imputation* to our method taking into account variable missingness, we used an upsampling technique called HAPGEN2 [29], which limits the lower bound of the global allele frequency to $1/(2 \cdot 503)$. Furthermore, the outcome used for the simulated GWAS is based on one causal variant with an explained variance of 0.02, therefore it might not be fully representative for a polygenic phenotype with more than one causal variant.

The *summary statistics imputation* method itself has several limitations too.

Due to the size of publicly available sequenced reference panels we can not explore the performance of rare variants ($MAF < 1\%$).

The imputation of summary statistics of an untyped SNV is essentially the linear combination of the summary statistics of the tag SNVs ([Eq \(2\)](#)). Such a model cannot capture non-linear dependence between tag- and target SNVs [10], which is often the case for rare variants [44, 45]. In contrast, *genotype imputation* is able to capture such non-linear relationships by estimating the underlying haplotypes (a non-linear combination of tagging alleles). Furthermore, in case of *genotype imputation* it is sufficient that the relevant haplotypes are present in

the reference panel, but the overall allele frequency does not need to match the GWAS allele frequency.

Summary statistics imputation relies on fine tuning of parameters, such as shrinkage of the correlation matrix. Any $\lambda > 0$ will make the correlation matrix invertible, but a stronger shrinkage can compensate for estimation error. We hypothesised that optimal shrinkage depends on local LD structure, and sought to optimise λ for each genomic region using the effect sizes of tag SNVs as training data set in a leave-one-out fashion. When looking at null variants, however, maximum shrinkage ($\lambda = 1$) usually leads to the smallest RMSE. Therefore, when looking at a region with a mixture of null and associated SNVs, the selected λ will be shifted towards 1 and shrink the estimation of associated SNVs towards 0, which is not ideal.

The imputation quality metric $\hat{r}_{\text{pred,adj}}^2$ tends to be inaccurate in case of small reference panels. The metric is commonly estimated as the total explained variance of a linear model given the reference panel, where the unmeasured SNV is regressed onto all measured markers in the reference panel (Eq (7)). We noticed that for reference panel sizes smaller than 1000 individuals, the conventional estimation of imputation quality in Eq (7) is biased towards overestimation. We extend the existing imputation quality (Eq (7)) by accounting for sample size and the effective number of variants (Eq (8)). The most accurate imputation quality estimations are obtained using an out-of-sample prediction after model selection by fitting a ridge regression model for each unmeasured SNV (\hat{r}_{ridge}^2). However, due to the computational complexity, the calculation takes longer than the actual imputation. We provide a more detailed analysis in [S2 Appendix](#).

Supporting information

S1 Fig. UK Biobank: Absolute frequencies of allele frequency and imputation quality of imputed SNVs. This figure shows how many of the null and associated SNVs were categorised into common, low-frequency and rare MAF subgroups, and into well-imputed, medium imputed and badly imputed imputation subgroups. Associated SNVs are presented in the left window, and null SNVs are presented in the right window. MAF category (x-axis), # of SNVs on the y-axis, colour refers to imputation quality category.
(PDF)

S2 Fig. UK Biobank: Relative frequencies of imputation quality within each allele frequency group. This figure shows the fraction of badly-, medium- and well-imputed SNVs within each MAF subgroup. Null and associated SNVs were categorised into common, low-frequency and rare MAF subgroup, and into well-imputed, medium imputed and badly imputed imputation subgroup. Associated SNVs are presented in the left window, and null SNVs are presented in the right window. MAF category (x-axis), fraction of SNVs on the y-axis, colour refers to imputation quality category. Numbers within the stacked barplot refer to the number of SNVs imputed in each subgroup.
(PDF)

S3 Fig. UK Biobank: Comparison of imputation quality methods. MACH \hat{r}^2 [46] (x-axis) versus IMPUTE's info measure used by *genotype imputation* (y-axis). To avoid clumping of dots, we used tiles varying from grey (few dots) to black (many dots). The identity line is dotted.
(PDF)

S4 Fig. UK Biobank: Comparison of imputation quality methods. IMPUTE's info measure used by *genotype imputation* (x-axis) vs $\hat{r}_{\text{pred,adj}}^2$ used by *summary statistics imputation* (y-axis). To avoid clumping of dots, we used tiles varying from grey (few dots) to black (many dots).

The identity line is dotted.
(PDF)

S5 Fig. UK Biobank (simulation): FPR versus power. This figure compares false positive rate (FPR) (x-axis on log10-scale) versus power (y-axis) for *genotype imputation* (blue) and *summary statistics imputation* (green) for different significance thresholds (α). It includes 95%-confidence intervals in both directions (vertically as a ribbon and horizontally as lines). This figure is a zoom into the bottom-left area of Fig 7 and shows FPR between 0 and 0.1. The coloured dots represent the $\alpha = 0.05$. The vertical, dashed line represents FPR = 0.05. Results are grouped according to MAF (columns) and imputation quality (rows) categories.
(PDF)

S6 Fig. GIANT: Concordance between genotyping and exome chip results. This graph shows the Z-statistics of the exome chip study on the x-axis versus the Z-statistics of SNP-array study on the y-axis. Each dot shows one of the 2'601 variants that had $LD_{\max} > 0.1$ (LD with one of the top variants in the exome [13] or HapMap study [12]). To make the density more visible, dots have been made transparent. The solid line indicates a linear regression fit, with the slope in the top right corner (including the 95%-confidence interval in brackets). The dashed line represents the ratio between the two median sample sizes $0.82 = \frac{\sqrt{N_{\text{HapMap-study}}}}{\sqrt{N_{\text{exome-study}}}} = \frac{\sqrt{251'647}}{\sqrt{370'529}}$.
(PDF)

S7 Fig. Locus-zoom plots of all 35 regions. Filename according to column 'filename' S1 Table. This figure shows three datasets: Results from the HapMap and the exome chip study, and imputed summary statistics. The top window shows HapMap P -values as orange circles and the imputed P -values (using *summary statistics imputation*) as solid circles, with the colour representing the imputation quality (only $\hat{r}_{\text{pred,adj}}^2 \geq 0.3$ shown). The bottom window shows exome chip study results as solid, grey dots. Each dot represents the summary statistics of one variant. The x-axis shows the position (in Mb) on a ≥ 2 Mb range and the y-axis the $-\log_{10}(P)$ -value. The horizontal line shows the P -value threshold of 10^{-6} (dotted) and 10^{-8} (dashed). Top and bottom window have annotated summary statistics: In the bottom window we mark dots as black if it is are part of the 122 reported hits of [13]. In the top window we mark the rs-id of variants that are part of the 122 reported variants of [13] in bold black, and if they are part of the 697 variants of [12] in bold orange font. Variants that are black (plain) are imputed variants (that had the lowest conditional P -value). Variants in orange (plain) are HapMap variants, but were not among the 697 reported hits. Each of the annotated variants is marked for clarity with a bold circle in the respective colour. The genes annotated in the middle window are printed in grey if the gene has a length $< 5'000$ bp or is an unrecognised gene (RP-).
(ZIP)

S8 Fig. Summary of exome results replication. This graph shows for all 111 variants the $-\log_{10}(p)$ -value of the exome chip study on the x-axis and the imputed $-\log_{10}(p)$ -value on the y-axis. The first row refers to the highest imputation quality (between 0.7 and 1), with the columns as the different allele frequency categories. The number of dots in each window is marked top left. The vertical and horizontal dotted lines mark the significance threshold of $-\log_{10}(0.05/111)$ (dashed). The width of the x-axis is proportional to the range of the y-axis. For MAF and $\hat{r}_{\text{pred,adj}}^2$ notation, the lower bound is excluded while the upper bound is included. For example, $1 - 5\%$ is equivalent to $1 < \text{MAF} \leq 5$.
(PDF)

S9 Fig. UK Biobank: Distribution of P -values from *summary statistics imputation*. These QQ-plots show the distribution of p -values resulting from *summary statistics imputation*, for associated variants (left window), null variants (right window). The colours refer to the imputation quality categories. Note that the P -value in these plots are not λ_{GC} corrected.
(PDF)

S10 Fig. Variable sample size in GIANT. In the GIANT meta-analysis (BMI, women over 50 years of age) the set of SNVs is different in each cohort, allowing us to create a binary ‘missingness’ vector for each SNV recording whether a given individual in the combined population was genotyped for this SNV. For 10’000 randomly selected pairs of nearby SNVs, we compute the correlation between these missingness vectors and plot the density plot. The correlations are usually greater than zero, and often quite close to one, confirming that a ‘missing independently at random’ assumption is not appropriate.
(PDF)

S11 Fig. *Summary statistics imputation* versus *genotype imputation* for associated variants colored by imputation quality. The x-axis shows the Z-statistics of the *genotype imputation* summary statistics, while the y-axis shows the Z-statistics from *summary statistics imputation*. The color of each point refers to the imputation quality of *summary statistics imputation*. Results are grouped according to MAF (columns) and imputation quality (rows) categories and the numbers top-right in each window refers to the number of SNVs represented. The identity line is indicated with a dotted line. The estimation for correlation and slope are noted in the bottom-right corner.
(PDF)

S12 Fig. *Summary statistics imputation* versus *genotype imputation* for non-associated variants colored by imputation quality. The x-axis shows the Z-statistics of the *genotype imputation* summary statistics, while the y-axis shows the Z-statistics from *summary statistics imputation*. The color of each point refers to the imputation quality of *summary statistics imputation*. Results are grouped according to MAF (columns) and imputation quality (rows) categories and the numbers top-right in each window refers to the number of SNVs represented. The identity line is indicated with a dotted line. The estimation for correlation and slope are noted in the bottom-right corner.
(PDF)

S13 Fig. *Summary statistics imputation* versus *genotype imputation* for associated variants colored by info measure. The x-axis shows the Z-statistics of the *genotype imputation* summary statistics, while the y-axis shows the Z-statistics from *summary statistics imputation*. The color of each point refers to the imputation quality of *genotype imputation*. Results are grouped according to MAF (columns) and imputation quality (rows) categories and the numbers top-right in each window refers to the number of SNVs represented. The identity line is indicated with a dotted line. The estimation for correlation and slope are noted in the bottom-right corner.
(PDF)

S14 Fig. *Summary statistics imputation* versus *genotype imputation* for non-associated variants colored by info measure. The x-axis shows the Z-statistics of the *genotype imputation* summary statistics, while the y-axis shows the Z-statistics from *summary statistics imputation*. The color of each point refers to the imputation quality of *genotype imputation*. Results are grouped according to MAF (columns) and imputation quality (rows) categories and the numbers top-right in each window refers to the number of SNVs represented. The identity line is

indicated with a dotted line. The estimation for correlation and slope are noted in the bottom-right corner.

(PDF)

S15 Fig. Accounting for missingness in GIANT. The x-axis shows the Z-statistics of the conventional estimate, the y-axis the Z-statistics when accounting for missingness (dependent approach). The dotted line marks the genome-wide threshold. There are 11'200'403 variants displayed in a binned fashion.

(PDF)

S1 Table. GIANT: Detailed results of 35 candidate loci. This table presents details of the 35 candidate loci discovered with *summary statistics imputation*. Within each candidate locus, we provide for the top variant the imputation results (`.imp`), along with conditional analysis results (`.cond`), the UK Biobank replication (`.ukbb`, whether it replicated or not (replication), and (if available) the exome chip study results (`.exome`). `filename` shows the filename of the locus-zoom plot in [S7 Fig](#). `SNP.cond.info` presents each HapMap SNV used for conditional analysis, including its MAF, LD between the HapMap SNV and the imputed SNV, and a reversed conditional analysis result (HapMap variant conditioned on the imputed SNV). The column `Group` classifies each row into candidate loci (i) that were reported by [13] already, (ii) that had no reported HapMap variant nearby, (iii) that had at least one reported HapMap variants nearby. The column `max.P.cond.hm` represents the maximum *P*-value from a conditional analysis performed with each HapMap variant nearby. *P* = *P*-value, *N* = sample size, *r*² = imputation quality, *eff* = effect size, *EAF* = effect allele frequency, *MAF* = minor allele frequency. If a candidate locus was not available in the UK Biobank, we provide a replication for a second variant that is in high LD with the primary variant, hence duplicated region numbers for some candidate loci.

(CSV)

S2 Table. GTEx annotation results for variants in eQTLs. This table shows SNVs which are significant eQTLs in GTEx [27]. We only report SNV-gene expression associations where the summary statistics pass the significance threshold of $\alpha = 10^{-6}$. The first four columns represent the region number, SNV, *P*-value from *summary statistics imputation* and the *P*-value in the UK Biobank. The three remaining columns are information extracted from GTEx, with the tissue name, gene name and the *P*-value of the association between the SNV and the gene expression. For each region, we order SNV-gene-tissue associations according to their *P*-value. # refers to the region number.

(CSV)

S3 Table. GIANT: Results of 122 exome variants. This table presents the summary statistics imputation results (`.imp`) for all 122 variants shown as “novel” in [13]. The right hand part of the table shows the original exome chip results for comparison (`.exome`). *P* = *P*-value, *N* = sample size, *r*² = imputation quality, *eff* = effect size, *EAF* = effect allele frequency. 11 variants were not referenced in UK10K or on chromosome X and therefore not imputed (see column ‘comment’). The position corresponds to hg19.

(CSV)

S1 Appendix. Simulation framework.

(PDF)

S2 Appendix. Imputation quality.

(PDF)

S3 Appendix. Summary statistics imputation accounting for varying sample size and missingness.
(PDF)

Acknowledgments

This research has been conducted using the UK Biobank Resource. Eleonora Porcu, Jonathan Sulc, Kaido Lepik and Ninon Mounier gave valuable comments on an earlier draft of the manuscript. We also thank the reviewers for their constructive comments, which improved the manuscript considerably.

Author Contributions

Conceptualization: Sina Rüeger, Aaron McDaid, Zoltán Kutalik.

Formal analysis: Sina Rüeger.

Methodology: Sina Rüeger, Aaron McDaid, Zoltán Kutalik.

Supervision: Zoltán Kutalik.

Visualization: Sina Rüeger.

Writing – original draft: Sina Rüeger, Aaron McDaid, Zoltán Kutalik.

Writing – review & editing: Sina Rüeger, Zoltán Kutalik.

References

- 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. 2010; 467(7319):1061–1073. <https://doi.org/10.1038/nature09534> PMID: 20981092
- Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*. 2016; 48(10).
- Howie B, Marchini J, Stephens M. Genotype Imputation with Thousands of Genomes. *G3*. 2011; 1(6):457–470. <https://doi.org/10.1534/g3.111.001198> PMID: 22384356
- Fuchsberger C, Abecasis GR, Hinds DA. Minimac2: Faster genotype imputation. *Bioinformatics*. 2015; 31(5):782–784. <https://doi.org/10.1093/bioinformatics/btu704> PMID: 25338720
- Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*. 2013; 37(7):658–665. <https://doi.org/10.1002/gepi.21758> PMID: 24114802
- Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Gen*. 2010; 42(7):565–569. <https://doi.org/10.1038/ng.608>
- Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*. 2015; 47(3):291–295. <https://doi.org/10.1038/ng.3211> PMID: 25642630
- Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American Journal of Human Genetics*. 2014; 94(4):559–573. <https://doi.org/10.1016/j.ajhg.2014.03.004> PMID: 24702953
- Pasaniuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*. 2014; 30(20). <https://doi.org/10.1093/bioinformatics/btu416> PMID: 24990607
- Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet*. 2017; 18(2):117–127. <https://doi.org/10.1038/nrg.2016.142> PMID: 27840428
- McDaid A, Rüeger S, Kutalik Z. SSIMP: Summary statistics imputation software; 2017. <http://wp.unil.ch/sgg/summary-statistic-imputation-software/>.
- Wood AR, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*. 2014; 46(11). <https://doi.org/10.1038/ng.3097>

13. Marouli E, et al. Rare and low-frequency coding variants alter human adult height. *Nature*. 2017;. <https://doi.org/10.1038/nature21039> PMID: 28146470
14. Lee D, Bigdeli TB, Riley BP, Fanous AH, Bacanu SA. DIST: Direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics*. 2013; 29(22). <https://doi.org/10.1093/bioinformatics/btt500>
15. Eaton ML. *Multivariate Statistics: A Vector Space Approach*. John Wiley & Sons Inc; 1983.
16. Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*. 2005; 4. <https://doi.org/10.2202/1544-6115.1175> PMID: 16646851
17. Wen X, Stephens M. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Annals of Applied Statistics*. 2010; 4(3):1158–1182. <https://doi.org/10.1214/10-AOAS338> PMID: 21479081
18. Lee D, Williamson VS, Bigdeli TB, Riley BP, Fanous aH, Vladimirov VI, et al. JEPEG: a summary statistics based tool for gene-level joint testing of functional variants. *Bioinformatics*. 2014; 31(8).
19. Kutalik Z, Johnson T, Bochud M, Mooser V, Vollenweider P, Waeber G, et al. Methods for testing association between uncertain genotypes and quantitative traits. *Biostatistics*. 2011; 12(1):1–17. <https://doi.org/10.1093/biostatistics/kxq039> PMID: 20543033
20. Theil H. *Economic Forecasts and Policy*. North-Holland Publishing Co.; 1961.
21. Gao X, Starmer AJ, Martin ER. A Multiple Testing Correction Method for Genetic Association Studies Using Correlated Single Nucleotide Polymorphisms. *Genetic Epidemiology*. 2008; 369:361–369. <https://doi.org/10.1002/gepi.20310>
22. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*. 2015; 12(3):1–10. <https://doi.org/10.1371/journal.pmed.1001779>
23. UK Biobank Phasing and Imputation Documentation; 2015. https://biobank.ctsu.ox.ac.uk/crystal/docs/impute_ukb_v1.pdf.
24. Slatkin M. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature reviews Genetics*. 2008; 9(6):477–85. <https://doi.org/10.1038/nrg2361> PMID: 18427557
25. Yang J, Ferreira T, Morris AP, Medland SE, Madden PaF, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*. 2012; 44(4):369–375. <https://doi.org/10.1038/ng.2213> PMID: 22426310
26. Abbott L, Anttila V, Aragam K, Bloom J, Bryant S, Churchhouse C, et al. Rapid GWAS of thousands of phenotypes for 337'000 samples in the UK Biobank; 2017. Available from: <http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank>.
27. Ardlie KG, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*. 2015; 348(6235):648–660. <https://doi.org/10.1126/science.1262110>
28. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, et al. PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics*. 2016; 32(20):3207. <https://doi.org/10.1093/bioinformatics/btw373> PMID: 27318201
29. Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*. 2011;. <https://doi.org/10.1093/bioinformatics/btr341>
30. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nature genetics*. 2013; 45(11):1274–83. <https://doi.org/10.1038/ng.2797> PMID: 24097068
31. Morris AP, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*. 2012; 44(9):981–990. <https://doi.org/10.1038/ng.2383> PMID: 22885922
32. Moayyeri A, Hammond CJ, Valdes AM, Spector TD. Cohort profile: Twinsuk and healthy ageing twin study. *International Journal of Epidemiology*. 2013; 42(1):76–85. <https://doi.org/10.1093/ije/dyr207> PMID: 22253318
33. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, et al. Cohort profile: The 'Children of the 90s'-The index offspring of the avon longitudinal study of parents and children. *International Journal of Epidemiology*. 2013; 42(1):111–127. <https://doi.org/10.1093/ije/dys064> PMID: 22507743
34. R Core Team. *R: A Language and Environment for Statistical Computing*; 2017. Available from: <http://www.R-project.org/>.
35. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*. 2007; 39(7):906–13. <https://doi.org/10.1038/ng2088> PMID: 17572673
36. Lango Allen H, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*. 2010; 467(7317):832–838. <https://doi.org/10.1038/nature09410> PMID: 20881960

37. Wood AR, Perry JRB, Tanaka T, Hernandez DG, Zheng HF, Melzer D, et al. Imputation of Variants from the 1000 Genomes Project Modestly Improves Known Associations and Can Identify Low-frequency Variant—Phenotype Associations Undetected by HapMap Based Imputation. *PLOS ONE*. 2013; 8(5):1–13. <https://doi.org/10.1371/journal.pone.0064343>
38. Kettunen J, Demirkan A, Würtz P, Draisma HHMM, Haller T, Rawal R, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nature Communications*. 2016; 7:11122. <https://doi.org/10.1038/ncomms11122> PMID: 27005778
39. Chen D, McKay JD, Clifford G, Gaborieau V, Chabrier A, Waterboer T, et al. Genome-wide association study of HPV seropositivity. *Human Molecular Genetics*. 2011; 20(23):4714–4723. <https://doi.org/10.1093/hmg/ddr383> PMID: 21896673
40. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*. 2014; 506(7488):376–381. <https://doi.org/10.1038/nature12873> PMID: 24390342
41. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics*. 2015; 47(9):979–989. <https://doi.org/10.1038/ng.3359> PMID: 26192919
42. Wu Y, Zheng Z, Visscher PM, Yang J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biology*. 2017; 18(1):86. <https://doi.org/10.1186/s13059-017-1216-0> PMID: 28506277
43. Rüeger S, McDaid A, Kutalik Z. Improved imputation of summary statistics for realistic settings. *bioRxiv*. 2017;
44. Wood AR, Tuke MA, Nalls MA, Hernandez DG, Bandinelli S, Singleton AB, et al. Another explanation for apparent epistasis. *Nature*. 2014; 514(7520):E3–E5. <https://doi.org/10.1038/nature13691> PMID: 25279928
45. Hemani G, Shakhbazov K, Westra HJ, Esko T, Henders AK, Mcrae AF, et al. transcription in humans. *Nature*. 2014; 508(7495):249–253. <https://doi.org/10.1038/nature13005> PMID: 24572353
46. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*. 2010; 34(8):816–834. <https://doi.org/10.1002/gepi.20533> PMID: 21058334

Subject Section

Improved imputation of summary statistics for mixed populations

Sina Rüeger^{1,2}, Aaron McDaid^{1,2} and Zoltán Kutalik^{1,2*}

¹Swiss Institute of Bioinformatics, Lausanne, 1015, Switzerland and

²Institute of Social and Preventive Medicine, Lausanne University Hospital, Lausanne, 1010, Switzerland

*Zoltán Kutalik: zoltan.kutalik@unil.ch, tel.: +41-21 314 6750

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: *Summary statistics imputation* can be used to infer association summary statistics of an already conducted, genotype-based meta-analysis to higher genomic resolution. This is typically needed when *genotype imputation* is not feasible for some cohorts. Oftentimes, cohorts of such a meta-analysis are variable in terms of (country of) origin or ancestry. This violates the assumption of current methods that an external LD matrix and the covariance of the Z-statistics are identical.

Results: To address this issue, we present *variance matching*, an extension to the existing *summary statistics imputation* method, which manipulates the LD matrix needed for *summary statistics imputation*. Based on simulations using real data we find that accounting for ancestry admixture yields noticeable improvement only when the total reference panel size is > 1000 . We show that for population specific variants this effect is more pronounced with increasing F_{ST} .

Contact: zoltan.kutalik@unil.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genotype data for genome-wide association studies (GWASs) are often collected using DNA chips, which cover only a small fraction of the variable genome. To be able to combine GWASs that measured different sets of genetic markers (due to differences in the content of commercial arrays), genetic information has to be inferred for a common set of markers. Such inference exploits the fact the neighbouring SNVs are often in linkage disequilibrium (LD), which has been well-quantified in different human populations. Statistical inference of these untyped SNVs in a study cohort, therefore, relies on an external reference panel of densely genotyped or sequenced individuals. The inference process is termed *imputation*, of which there are two main types. *Genotype imputation* (Marchini and Howie, 2010) first estimates all haplotypes both in the reference panel and the study cohort, then using a Hidden Markov Model every observed haplotype in the study cohort is assembled as a probabilistic mosaic of reference panel haplotypes. The reconstruction facilitates the computation of the probability of each genotype for every SNV of the reference panel in each individual of the study cohort. Having imputed the genotype data set, one can then run an association scan with an arbitrary trait and obtain

association summary statistics. *Summary statistics imputation* Pasiuniuc *et al.* (2014) on the other hand starts off with association summary statistics available for all genotyped markers and infers, combined with a reference panel, directly the association summary statistics of SNVs available in the reference panel. More specifically, estimating the local pair-wise linkage disequilibrium (LD) structure of each genetic region using the reference panel and combining it with association summary statistics allows to calculate a conditional expectation of normally distributed summary statistics. This latter approach is the central focus of our paper. Compared to *genotype imputation*, *summary statistics imputation* is much less demanding on computational resources, and requires no access to individual level genetic data.

Methods making use of summary statistics, such as calculating genetic correlation (Bulik-Sullivan *et al.*, 2015), approximate conditional analysis (Yang *et al.*, 2012) or causal inference (Burgess *et al.*, 2013), have gained interest in recent years, because they bypass the need of genotype data, but mimic it by making use of external reference panels. These methods could profit from summary statistics being available on an arbitrarily chosen panel of SNVs – provided by *summary statistics imputation*. However, it is not clear how to optimally combine different LD reference panels for summary statistics emerging from a meta-analysis of a large number of different studies (coming from different countries/regions),

with potentially different ancestries. To ensure accurate imputation of such “admixed” meta-analyses, we propose a method called *variance matching* that, for each genomic region, optimally combines reference panels to best match the local LD pattern of the underlying GWAS population. Using a simulation framework, we compare *variance matching* to a *benchmark* solution and one of previously proposed approaches (Lee et al., 2015a,b; Park et al., 2015).

2 Methods

2.1 Summary statistics imputation (SSimp)

We assume a set of univariate effect size estimates α_i are available for SNVs $i = 1, \dots, I$ from a linear regression between a continuous phenotype \mathbf{y} and the corresponding genotype \mathbf{g}^i measured in N individuals. Without loss of generality we assume that both vectors are normalised to have zero mean and unit variance. Thus $\alpha_i = \frac{(\mathbf{g}^i)' \cdot \mathbf{y}}{N}$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_I)' \sim \mathcal{N}(\boldsymbol{\alpha}, \boldsymbol{\Sigma})$. $\boldsymbol{\Sigma}$ represents the pairwise covariance matrix of effect sizes of all $i = 1, \dots, I$ SNVs.

To estimate the univariate effect size α_u of an untyped SNV u in the same sample, one can use the conditional expectation of a multivariate normal distribution. The conditional mean of the effect of SNV u can be expressed using the effect size estimates of the tag SNVs (Eaton, 1983; Pasaniuc et al., 2014):

$$\hat{a}_u = a_{u|\mathcal{M}} = \alpha_u + \boldsymbol{\Sigma}_{u\mathcal{M}} \boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1} (\boldsymbol{\alpha} - \boldsymbol{\alpha}), \quad (1)$$

where \mathcal{M} is a vector of marker SNVs, $\boldsymbol{\Sigma}_{u\mathcal{M}}$ represents the covariance between SNV u and all \mathcal{M} markers and $\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}$ represents the covariance between all \mathcal{M} markers.

We assume that estimates for the two covariances are available from an external reference panel with n individuals and denote them $\mathbf{s} = \hat{\boldsymbol{\Sigma}}_{\mathcal{M}u}$, $\mathbf{S} = \hat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}$. The corresponding correlation matrices are $\mathbf{c} = N \cdot \mathbf{s}$ and $\mathbf{C} = N \cdot \mathbf{S}$ (with $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$ as the corresponding true correlation matrices). Further, by assuming that SNV u and the trait are independent conditioned on the \mathcal{M} markers, i.e. $\alpha_u - \boldsymbol{\Sigma}_{u\mathcal{M}} \boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1} \boldsymbol{\alpha} = 0$, Eq. (1) becomes

$$\hat{a}_u = a_{u|\mathcal{M}} = \mathbf{s}' \mathbf{S}^{-1} \boldsymbol{\alpha} = \mathbf{c}' \mathbf{C}^{-1} \boldsymbol{\alpha} \quad (2)$$

One can also choose to impute the Z-statistic instead, as derived by Pasaniuc et al. (2014):

$$\hat{z}_{u|\mathcal{M}} = \mathbf{c}' \mathbf{C}^{-1} \mathbf{z} \quad (3)$$

with $\mathbf{z} = \boldsymbol{\alpha} \sqrt{N}$, when the effect size is small (as is the case in typical GWAS).

Similar to Pasaniuc et al. (2014), we chose \mathcal{M} to include all measured variants within at least 250 Kb of SNV u . To speed up the computation when imputing SNVs genome-wide, we apply a windowing strategy, where SNVs within a 1 Mb window are imputed simultaneously using the same set of \mathcal{M} tag SNVs the 1 Mb window plus 250 Kb flanking regions on each side.

Shrinkage of SNV correlation matrix

To estimate \mathbf{C} (and \mathbf{c}) we use an external reference panel of n individuals. Since the size of \mathbf{C} often exceeds the number of individuals ($q \gg n$), shrinkage of matrix \mathbf{C} is needed to guarantee that it is invertible. By applying shrinking, the modified matrix \mathbf{C} becomes

$$\mathbf{C}_\lambda = (1 - \lambda) \mathbf{C} + \lambda \mathbf{I} \quad (4)$$

Even though \mathbf{c} is not inverted, we still shrink it to curb random fluctuations in the LD estimation in case of no LD.

$$\mathbf{c}_\lambda = (1 - \lambda) \mathbf{c} \quad (5)$$

Inserting \mathbf{c}_λ and \mathbf{C}_λ , Eq. (2) then becomes

$$\hat{a}_u = a_{u|\mathcal{M}} = \mathbf{c}'_\lambda \mathbf{C}_\lambda^{-1} \boldsymbol{\alpha} \quad (6)$$

Note that λ can vary between 0 and 1, with $\lambda = 1$ turning \mathbf{C} to the identity matrix, while $\lambda = 0$ leaves \mathbf{C} unchanged. Here, we mainly focus on λ changing with the reference panel size n : $\lambda = 2/\sqrt{n}$ (Lee et al., 2014).

2.2 Optimal combination of reference panel subpopulations to match the GWAS sample

For *summary statistics imputation* we would like to estimate the local LD structure of each region in the GWAS population ($\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}$ and $\boldsymbol{\Sigma}_{u\mathcal{M}}$) and to do so we use a (sequenced) reference population, yielding estimates \mathbf{C} and \mathbf{c} . Clearly, the closer these estimates are to the real values, the better the imputation will be (i.e. smaller the estimation error in Eq. (6)). Our aim is to find a weighted mixture of the reference sub-populations that has an LD structure as similar as possible to the LD in the GWAS population.

Park et al. (2015) proposed an elegant, generalised approach to weight population LD structure. Their algorithm *Adapt-Mix* chooses weights (\mathbf{w}^{am}) based on optimising an objective function. In the case of imputation the objective function is the MSE of the (re-)imputed Z-statistics at observed SNVs. Lee et al. (2015a) developed *Distmix*, which minimises the Euclidean distance between allele frequencies of the reference panels and the GWAS study, but ignores the variance-bias trade-off.

While the true LD structure of the actual GWAS population is rarely known, the GWAS allele frequencies are routinely calculated (even if not always reported for out-dated privacy preserving reasons) in meta-analytic studies. In the following we show how this information can be exploited to improve *summary statistics imputation*.

First, suppose that the reference panel is made up of P subpopulations of sizes $n^{(1)}, n^{(2)}, \dots, n^{(P)}$. Next, we introduce a set of weights $\mathbf{w} = (w_1, w_2, \dots, w_P)$, which can be viewed as the collection of weights that determine the reference population mixture, i.e. $\sum_{p=1}^P w_p = 1$ and $w_p \geq 0$.

We can calculate the covariance s as a function of these weights ($s(\mathbf{w})$), i.e. for each subpopulation we calculate the covariance separately ($s^{(p)}$), and then combine them, weighted by their weights \mathbf{w} :

$$s_{kl}(\mathbf{w}) = \sum_{p=1}^P w_p \left[s_{kl}^{(p)} + t_{kl}^{(p)} \right], \quad (7)$$

where $t_{kl}^{(p)}$ is the between-group covariance for variants k and l in population p :

$$t_{kl}^{(p)} = (\bar{g}_k^p - \bar{g}_k)(\bar{g}_l^p - \bar{g}_l) \quad (8)$$

and $s_{kl}^{(p)}$ denotes the covariance for variants k and l in population p :

$$s_{kl}^{(p)} = \frac{1}{n^{(p)} - 1} \sum_{i \in I_p}^{n^{(p)}} (g_{i,k} - \bar{g}_k^p)(g_{i,l} - \bar{g}_l^p) \quad (9)$$

$g_{i,k}$ refers to genotype of variant k for individual i . The overall, mean population genotype dosage (i.e. twice the allele frequency) is naturally defined as the weighted mean sub-population genotype dosage $\bar{g}_k = \sum_p w_p \cdot \bar{g}_k^p$ and \bar{g}_k^p being the average genotype dosage in population p : $\bar{g}_k^p = \frac{1}{n^{(p)}} \sum_{i \in I_p} g_{i,k}$ and I_p refers to the indices of individuals contained in population p .

While the reference panel population sizes are being fixed at $n^{(p)}$ and we defined $w_p \propto n_p$, we could use any arbitrary weights \mathbf{w} in order to match a GWAS population, which has different population proportions than the reference panel. This manipulation of the covariance estimation

can be used to adapt the reference panel population structure towards the population structure that is observed in GWAS summary statistics.

The corresponding correlation between SNV k and l from a reference panel with specific chosen weights \mathbf{w} is

$$c_{kl}^{\mathbf{w}} = \frac{s_{kl}(\mathbf{w})}{\sqrt{s_{kk}(\mathbf{w}) \cdot s_{ll}(\mathbf{w})}} \quad (10)$$

Our goal is to quantify the mean squared error (MSE) between the true GWAS LD matrix ($\Sigma_{\mathcal{M},\mathcal{M}}$) and the LD matrix estimated from the reference panel ($\mathbf{C}_{\mathcal{M},\mathcal{M}}^{\mathbf{w}}$). Since we cannot estimate the off-diagonal values of the GWAS covariance matrix, we focus on its diagonal elements and estimate them from the GWAS allele frequencies. The MSE of Eq. (7) for SNV k can be written as

$$\text{MSE}[s_{kk}(\mathbf{w})] = \text{Bias}^2[s_{kk}(\mathbf{w})] + \text{Var}[s_{kk}(\mathbf{w})] \quad (11)$$

In short, the MSE of Eq. (7) depends on known quantities (mean genotype dosage \bar{g}_k^p for SNV k in population p ; sample sizes $n^{(p)}$ of the reference panel population p ; average genotype for SNV k in the GWAS study: \bar{g}_k^{obs}) and the unknown mixing parameter \mathbf{w} . Assuming Hardy-Weinberg equilibrium (HWE), we showed that the variance term is a sixth-degree, while the squared bias is a fourth-degree polynomial in \mathbf{w} . Details to derivations of the MSE are provided in Supplement A.2.

We aim to find a \mathbf{w} that minimises the MSE in Eq. (11) for a set of \mathcal{M} SNVs, for which we know the allele frequencies \bar{g}^{obs} in the GWAS population and can estimate \bar{g}^p from a reference panel:

$$\mathbf{w}^{\text{VM}} = \arg \min_{\mathbf{w}} \sum_{k=1}^{\mathcal{M}} \text{MSE}[s_{kk}(\mathbf{w})] \quad (12)$$

with $\sum_{p=1}^P w_p = 1$ and $w_p \geq 0$. We call this method *variance matching* (vm), as we are using the GWAS allele frequencies to match genotype variances. Parameter \mathbf{w}^{VM} gives us an estimation of the population weights used in Eq. (7).

Finally, we substitute \mathbf{w}^{VM} into Equation Eq. (10) and plug it into Eq. (6):

$$\hat{a}_u = a_{u|\mathcal{M}} = \mathbf{c}_{\lambda}^{\mathbf{w}^{\text{VM}}} (\mathbf{C}_{\lambda}^{\mathbf{w}^{\text{VM}}})^{-1} \mathbf{a} \quad (13)$$

Detailed derivations of Eqs. (7) to (12) can be found in Supplement A.

2.3 Reference panels

As reference panels we used genetic data from the 1000 Genomes Project Consortium (2010).

2.4 Simulation

2.4.1 Simulation of GWAS summary statistics

For simulation of GWAS summary statistics we used data from the five European subpopulations CEU, GBR, FIN, TSI and IBR of the 1000 Genomes project (1KG). We chose to up-sample chromosome 15 using HAPGEN2 (Su *et al.*, 2011) to 5'000 individuals for each subpopulation, yielding a total of 25'000 individuals. Of these 5000 individuals per population we used half each to generate a GWAS with an *in silico* phenotype. The remaining 12'500 individuals were used as reference panel for summary statistic imputation.

We split chromosome 15 into 74 disjoint regions of 1.5 Mb. Due to the sliding window imputation approach we did not include regions at the very start and end of the chromosome. In each region we chose a causal variant g randomly from all SNVs with minor allele frequency between 0.05 and 0.2. We simulated an *in silico* phenotype y using a normal linear model $y = \beta g + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 1 - \beta^2 \cdot 2q(1 - q))$, where q is the allele frequency of the causal SNV and β was selected such that the explained

variance $\beta^2 \cdot 2q(1 - q)$ is set to 0.02. To obtain the association summary statistics we ran linear regression for each variant k in the 1.5 Mb region, yielding effect size and standard error estimates $a_k, se(a_k)$, from which we calculated the standardised effect size estimate $a_k/se(a_k)/\sqrt{N}$ (N being the sample size).

2.4.2 Applying summary statistics imputation and comparing methods

We constructed GWAS genotype datasets with a fraction w^+ of Finnish individuals. The total number of individuals in the GWAS genotype dataset was constant at 2'500. Next, we calculated a re-weighted \mathbf{C} from our reference panel (with weight \mathbf{w}). We then created different scenarios by repeating this procedure for many different GWAS compositions (i.e. we varied the Finnish fraction w^+ between 0 and 1 in 0.2 increments) and weights \mathbf{w} of Finnish for the correlation matrix of the reference panel (which we varied between 0 and 1 in 0.05 increments). For each scenario, we calculated three MSE for a set of imputed SNVs (Eq. (14)): first, the MSE of the standardised effect size; second, for the *variance matching* approach we calculated the MSE of (the diagonal of) matrix \mathbf{C} estimated from the reference panel; third, for the *Adapt-Mix* approach we calculated the MSE of the standardised effect sizes of observed SNVs, as described in Park *et al.* (2015).

$$h(\mathbf{w}) = \sum_{m=1}^{\mathcal{M}} \left(\mathbf{c}^{\mathbf{w}^+} (\mathbf{C}^{\mathbf{w}})^{-1} \mathbf{a}_{-m} - a_m \right)^2 \quad (14)$$

By minimising each error measurement over all \mathbf{w} , the first MSE will determine \mathbf{w}^* , which gives the theoretically best possible solution.

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} h(\mathbf{w}) \quad (15)$$

In our approach, the estimated MSE of matrix \mathbf{C} will determine \mathbf{w}^{VM} . While in the best competing algorithm, *Adapt-Mix*, the minimised MSE of the reimputed Z-statistics determines the value of \mathbf{w}^{AM} . We chose to vary the proportion of the Finnish population, as it differs the most from other populations in Europe in terms of allele frequencies and LD structure Lim *et al.* (2014), McEvoy *et al.* (2009). The remaining four populations of the European 1000 Genomes populations share equal weights in all scenarios. In our simulation we are looking at HapMap SNVs only as tag SNVs. There are between 167 and 1103 tag SNVs per region with mean 635. We imputed on average 1'743 SNVs per region (out of 74 in total).

3 Results

Summary statistics imputation works through combining summary statistics from a set of SNVs with pairwise SNV LD information obtained from an external reference panel. We extended the most recent summary statistics method (Pasaniuc *et al.*, 2014) by an optimal assembly of the LD matrix from a mixture of reference panels. Because our approach optimises the diagonal of the covariance matrix, we term our method *variance matching*. We compare our it to *Adapt-Mix* by Park *et al.* (2015) and a *benchmark* solution.

3.1 Simulation framework

To assess *variance matching* we used upsampled datasets, yielding 25'000 European individuals in total. GWASs were simulated using *in silico* phenotypes. This semi-simulation framework allowed us to study the impact of the reference panel sizes (up to 12'500) and their composition.

In brief, for various ancestry compositions of simulated GWAS sample we computed association summary statistics, masked a fraction of SNVs and imputed them. When imputing a single SNV we used tag SNVs within at least 250 Mb. For the imputation of an entire region we used a sliding window of 1 Mb with 250 Kb flanking regions on each side.

More specifically, for each simulated GWAS, we fixed the proportion of the Finnish subpopulation of the European reference panel of 1000 Genomes Project Consortium (2010) in the GWAS, then let the proportion of this population vary in the reference panel used for LD estimation. We repeated this for different Finnish proportions in the GWAS and in the reference panel (varying from 0 to 1), calculated each time the MSE between the estimated and the imputed standardised effect sizes ($h(w)$, Eq. (14)) and determined the *benchmark* weight that yields a minimal MSE (denoted as w^* , Eq. (15)). In parallel, we applied for each fixed proportion of Finnish in the GWAS the *variance matching* and the *Adapt-Mix* approach to determine their optimal weight — w^{VM} and w^{AM} — in the reference panel (Figure S1). To identify other factors that influence the choice of weights, we grouped the 637'153 SNVs into population specific (76'013) and population non-specific (561'140) groups (based on $F_{ST} \geq 1\%$ vs $F_{ST} < 1\%$, respectively), and ran the simulation from small to large reference panels ($n = \{500, 1'000, 2'500, 5'000, 12'500\}$).

3.2 Improving summary statistics imputation via variance matching

Ultimately, we are interested in two comparisons. First: the optimisation of weights versus the ad-hoc reference panel (which has roughly equal weights in the European sub-panel, i.e. $w = 0.2$). Second: how *Adapt-Mix* and our novel method *variance matching* perform compared to the *benchmark* estimation (the best possible choice if we were to know the true effect size). These two comparisons are presented in Fig. 1, where we compare the MSE of the three optimal weights (w^{VM} , w^{AM} , w^*) determined by each method relative to the MSE when using equal weights: $MSE\text{-ratio} = h(w)/h(w = 0.2)$, with h denoting the MSE between the estimated and the imputed effect sizes described in Eq. (14).

From the extensive simulation results (Fig. 1) it is clear, that the ad-hoc reference panel with equal weights works best (i.e. MSE-ratio close to 1) in two scenarios: for an equally partitioned GWAS (independent of reference panel size or whether the variants are population specific) and when the reference panel is small in size ($n \leq 1000$). For all other scenarios, i.e. either $n > 1000$ or the fraction of Finnish in the GWAS is not 0.2, the MSE-ratio is well below 1, therefore indicating a smaller MSE for the optimisation scenario. Note that the *benchmark* MSE-ratio is, by definition, always lower than *Adapt-Mix* and *variance matching* (as it is the best theoretically possible MSE).

When comparing *variance matching* and *Adapt-Mix* to the *benchmark* solution, we find that both optimising methods show a similar trend (greater advantage of using specific weights for population specific markers, large reference panel and heterogeneous GWAS). Except for three instances (population non-specific variants when imputed with a reference panel with 5'000 individuals and a GWAS with 40% or 80% Finnish individuals, and specific variants when imputed with a reference panel with 12'500 individuals and a GWAS with 100% Finnish individuals), *variance matching* offers equal or lower MSE than *Adapt-Mix*. When comparing the median MSE-ratio of the optimisation methods to the *benchmark*, *Adapt-Mix* is performing worst among population specific variants, a reference panel size of 500 and a GWAS with Finnish individuals only. *Variance matching* is performing worst in similar conditions, but when the GWAS consists of no Finnish individuals.

For a reference panel of 500 individuals, population specific variants and a GWAS with 80% Finnish individuals we observe a median MSE-ratio for the theoretically best possible reference panel composition of 0.886, while it is 0.892 for *variance matching* and 0.926 for *Adapt-Mix*. When increasing sample size to 12'500 the MSE-ratio for the *benchmark* solution using becomes 0.648 and 0.658 for *variance matching* and *Adapt-Mix*, respectively.

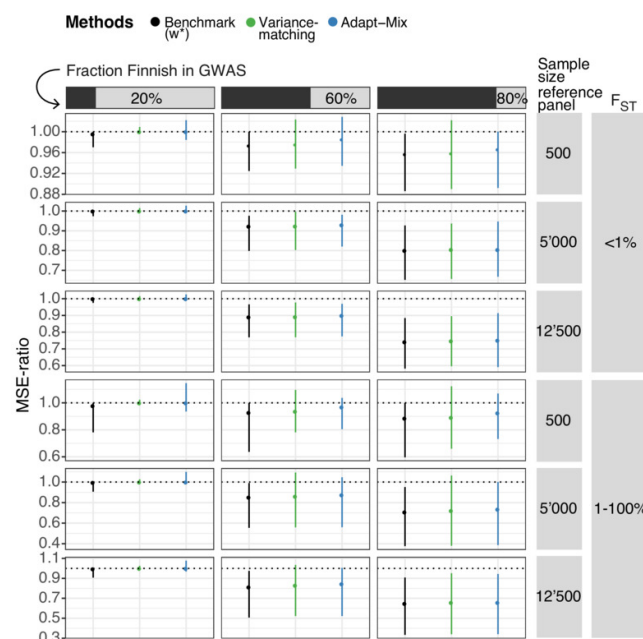


Fig. 1. Comparison of methods accounting for population structure. This figure shows the comparison of each w -optimisation method with respect to choosing the full European panel of 1000 Genomes project (which corresponds to equal weights). Each vertical line represents a summary of 74 simulated regions (the dot being the median, the line range representing 0.025 to 0.975 quantile). The x-axis shows the three different strategies: using theoretical best possible weights in black (if the estimated effect sizes were to be known), variance matching in green and Adapt-Mix in blue. The y-axis shows the MSE-ratio. The MSE-ratio represents the MSE when choosing the weights according to the respective optimisation relative to the MSE when choosing equal weights for all populations (hence a weight of 0.2 for all five populations), i.e. in black $h(w^*)/h(0.2)$, in green $h(w^{VM})/h(0.2)$, and in blue $h(w^{AM})/h(0.2)$. Function $h(w)$ is the MSE between the estimated and the imputed effect sizes described in Eq. (Eq. (14)). Values on the y-axis smaller than 1 show a smaller MSE in imputation with a specific w compared to the choice of an unadjusted reference panel with equal weights, while values larger than 1 indicate a higher MSE. Each row represents a subset of different sizes of reference panels, while a subset of the different Finnish fractions in the GWAS populations are grouped by column. Variants are also grouped according to F_{ST} , with population specific results being on the lower and population unspecific results on upper part of the graph. Figure S3 shows the same graph for all reference panel sizes and GWAS compositions. Table Table 1 provides the same information in a text file. Table B provides the results for all reference panel sizes and fraction of Finnish in the GWAS.

For variants that are not population specific we see a similar trend with increasing fraction of Finnish individuals and reference panel size, but as expected, less pronounced. For a reference panel of 500 individuals, population unspecific variants and a GWAS with 80% Finnish individuals we observe a median MSE-ratio of 0.957, while it is 0.959 for *variance matching* and 0.966 for *Adapt-Mix*. When increasing sample size to 12'500 it drops to 0.742, 0.747 and 0.751, for w^* , *variance matching* and *Adapt-Mix*, respectively.

For details to w^{VM} , w^* and w^{AM} check Fig. S1.

4 Discussion

Summary statistics are used more and more frequently for downstream analyses, but are not always available for all desired variants. These missing summary statistics can, however, be directly imputed from publicly available data using *summary statistics imputation*. The covariance matrices required for this are difficult to estimate from publicly available reference panels due to their size and population structure,

F_{ST}	Sample size reference panel	Method	Fraction Finnish in GWAS		
			20%	60%	80%
0-1%	500	w^*	0.996 (0.97-1)	0.974 (0.925-1)	0.957 (0.886-0.996)
0-1%	500	w^{VM}	1 (0.996-1.01)	0.975 (0.929-1.02)	0.959 (0.89-1.02)
0-1%	500	w^{AM}	1 (0.984-1.02)	0.985 (0.935-1.03)	0.966 (0.892-1)
0-1%	5000	w^*	1 (0.974-1)	0.923 (0.798-0.976)	0.8 (0.652-0.927)
0-1%	5000	w^{VM}	1 (0.999-1.02)	0.924 (0.803-0.997)	0.805 (0.656-0.937)
0-1%	5000	w^{AM}	1 (0.997-1.03)	0.93 (0.821-0.981)	0.805 (0.667-0.947)
0-1%	12500	w^*	1 (0.975-1)	0.891 (0.769-0.966)	0.742 (0.582-0.884)
0-1%	12500	w^{VM}	1 (0.996-1.02)	0.892 (0.77-0.977)	0.747 (0.596-0.895)
0-1%	12500	w^{AM}	1 (0.994-1.03)	0.9 (0.774-0.969)	0.751 (0.591-0.913)
1-100%	500	w^*	0.979 (0.781-1)	0.928 (0.637-1)	0.886 (0.596-1)
1-100%	500	w^{VM}	1 (0.983-1.02)	0.937 (0.781-1.1)	0.892 (0.66-1.12)
1-100%	500	w^{AM}	1 (0.936-1.15)	0.969 (0.805-1.04)	0.926 (0.733-1.07)
1-100%	5000	w^*	0.996 (0.906-1)	0.853 (0.555-0.99)	0.708 (0.377-0.951)
1-100%	5000	w^{VM}	1 (0.986-1.03)	0.861 (0.559-1.09)	0.722 (0.38-1.06)
1-100%	5000	w^{AM}	1 (0.982-1.1)	0.875 (0.56-1.05)	0.736 (0.386-1)
1-100%	12500	w^*	0.995 (0.908-1)	0.814 (0.507-0.975)	0.648 (0.334-0.908)
1-100%	12500	w^{VM}	1 (0.99-1.03)	0.83 (0.522-1.03)	0.658 (0.34-0.951)
1-100%	12500	w^{AM}	1 (0.968-1.08)	0.845 (0.522-1.01)	0.658 (0.341-0.944)

Table 1. This table presents the results corresponding to Figure Fig. 1: each entry represents the median MSE (in bold) and the 0.025 - 0.975 quantile in brackets for different Finnish fractions in the GWAS populations are (columns), F_{ST} , different methods (w^* , w^{VM} and w^{AM}) and reference panel size.

requiring their careful adjustment with shrinkage parameters. To address these limitations, we extended the *summary statistics imputation* method as presented in Pasaniuc *et al.* (2014) with an optimal combination of covariance matrices from reference panel subpopulation.

Choice of reference panel

Formulae for *summary statistics imputation* have two components: GWAS summary statistics and LD matrix estimates which represent the correlation between SNVs. The latter matrix is highly dependent on the reference panel composition: if the ancestry is different between the GWAS and the reference panel, the LD estimation will be biased and yield erroneous summary statistics. An adequate reference panel is therefore critical to the accuracy of *summary statistics imputation*, unlike *genotype imputation* where the Hidden Markov Model makes panel composition much less relevant. Most often, the reference panel for *summary statistics imputation* is often chosen ad-hoc, guessing the underlying GWAS population admixture.

Variance matching and Adapt-Mix

To tackle this problem, Park *et al.* (2015) proposed *Adapt-Mix* and we propose *variance matching* (Eqs. (12) and (13)). Both methods assume that the GWAS sample is composed of a mixture (or admixture) of populations and that we have a separate reference panel for each population. They then calculate the local LD structure as a linear combination of population-specific estimates, where the weight of each population depends either on the Z-statistics (*Adapt-Mix*) or the allele frequency (*variance matching*). *Variance matching* performed consistently, yet not significantly better than *Adapt-Mix* in 57 out of 60 subgroups that we explored (95% quantile ranges are overlapping in Figure S3). We also found, that *variance matching* offers performance very close to the best possible reference panel composition w^* .

Variance-bias tradeoff

Although the aim is to approximate the true mixing weights in the GWAS sample, the weights returned will usually deviate from that in an attempt to minimise the MSE. For example, given a GWAS performed in an exclusively Finnish population and using the 1000 Genomes project reference panel, we could either use only the 99 Finnish individuals (weight of 1 for the Finnish population, 0 for others), or select all 503 individuals of the European panel (weight of about 0.2 for the Finnish population). Using only Finnish individuals would more closely match the GWAS allele frequencies and reduce bias, however using the full panel would increase the precision of the estimated correlation matrix, reducing variance (Figure S2). Our approach aims to strike a balance between bias and variance by finding an optimal weight, somewhere between 0.2 and 1 in this example. We find that for smaller reference panels ($n = 500$) the optimal weight tends towards lower values, relying more on information from other populations, whereas for larger panels ($n = 12'500$) the optimal weights tend to be closer to the true underlying population composition in the GWAS (Fig. 2).

Limitations

Variance matching assumes that the population admixture that is reflected in the variance of tag SNVs (diagonal in matrix C') is the same as the covariance between tag SNVs (off-diagonal of C') as well as between tag SNVs and SNVs to impute (matrix c). Furthermore, our analytical solution to Eq. (11) involves approximations of the variance and the bias (Eq. (S7) and (S8)).

In general, finding a reference panel whose ancestry composition matches that of the GWAS is difficult because the mixture/admixture of populations is usually unknown. With *variance matching* we are addressing this by composing a matching LD matrix. However, there are other challenges too: publicly available reference panels have a limited number of populations with a limited number of individuals. To this end, we could not validate our approach in real data as diverse reference panels

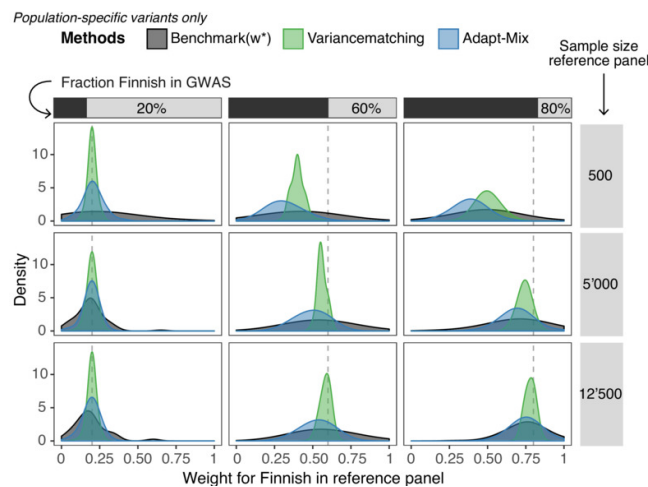


Fig. 2. Comparing w determined by different methods. This figure compares the weights chosen by all three optimisation methods for population specific variants: w^* as best possible weight (black), w^{VM} by variance matching (green), and w^{AM} by Adapt-Mix (blue). w^* represents the benchmark weight: the best possible choice if we were to know the true effect size, but given the same reference panel as for Adapt-Mix and variance matching. The x-axis displays the weights for the reference panel chosen by each method, and the y-axis shows the density. The results are split into columns and rows, with the rows for different reference panel sizes and the columns different Finnish fractions in the GWAS populations (also highlighted with the vertical dashed line). Each window contains w^* , w^{VM} and w^{AM} for each of the 74 regions.

with sample sizes > 500 per population are not publicly available at this time.

Due to lack of large, sequenced reference panels we used an upsampling technique called HAPGEN2 (Su *et al.*, 2011), which limits the lower bound of the global allele frequency to $1/(2 \cdot 503)$. Finally, the outcome used for the simulated GWAS is based on one causal variant with an explained variance of 0.02, therefore it might not be fully representative for a polygenic phenotype with more than one causal variant.

Finally, our method is not applicable to GWAS studies that decided not to share allele frequency information.

5 Conclusion

With *variance matching* we present an extension to the published *summary statistics imputation* method (Pasaniuc *et al.*, 2014) by allowing the LD structure to be adaptively estimated according to population admixture. To evaluate this extension, we performed GWAS on upsampled 1000 Genomes project data in combination with a simulated phenotypes. Due to the bias-variance trade-off, accounting for differences in population admixture between GWAS and reference panel yields better results with increasing panel size.

Acknowledgements

Ninon Mounier, Anthony Sonrel and Jonathan Sulc gave valuable comments on an earlier draft of the manuscript.

Funding

This work was supported by the Leenaards Foundation (<http://www.leenaards.ch>), the Swiss Institute of Bioinformatics (<http://www.isb-sib.ch/>), and the Swiss National Science Foundation [31003A-143914, 31003A-169929].

References

- 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319), 1061–1073.
- Bulik-Sullivan, B., Finucane, H., Anttila, V., Gusev, A., Day, F., Loh, P., ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan, L., and Perry, J. (2015). An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*, **47**(11), 1236–1241.
- Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, **37**(7), 658–665.
- Eaton, M. L. (1983). *Multivariate Statistics: A Vector Space Approach*. John Wiley & Sons Inc.
- Lee, D., Williamson, V. S., Bigdeli, T. B., Riley, B. P., Fanous, a. H., Vladimirov, V. I., and Bacanu, S.-A. (2014). JEPEG: a summary statistics based tool for gene-level joint testing of functional variants. *Bioinformatics*, **31**(8).
- Lee, D., Bigdeli, T. B., Williamson, V. S., Vladimirov, V. I., Riley, P., Fanous, A. H., and Bacanu, S.-a. (2015a). Genome analysis distmix : Direct imputation of summary statistics for unmeasured snps from mixed ethnicity cohorts. *Bioinformatics*.
- Lee, D., Williamson, V. S., Bigdeli, T. B., Riley, B. P., Webb, B. T., Fanous, A. H., Kendler, K. S., Vladimirov, V. I., and Bacanu, S.-A. (2015b). JEPEGMIX: gene-level joint analysis of functional SNPs in cosmopolitan cohorts: Table 1. *Bioinformatics*, **32**.
- Lim, E. T., Würtz, P., Havulinna, A. S., Palta, P., Tukiainen, T., Rehnström, K., Esko, T. o., Mägi, R., Inouye, M., Lappalainen, T., Chan, Y., Salem, R. M., Lek, M., Flannick, J., Sim, X., Manning, A., Ladenvall, C., Bumpstead, S., Hämäläinen, E., Aalto, K., Maksimow, M., Salmi, M., Blankenberg, S., Ardiissino, D., Shah, S., Horne, B., McPherson, R., Hovingh, G. K., Reilly, M. P., Watkins, H., Goel, A., Farrell, M., Girelli, D., Reiner, A. P., Stitzel, N. O., Kathiresan, S., Gabriel, S., Barrett, J. C., Lehtimäki, T., Laakso, M., Groop, L., Kaprio, J., Perola, M., McCarthy, M. I., Boehnke, M., Altshuler, D. M., Lindgren, C. M., Hirschhorn, J. N., Metspalu, A., Freimer, N. B., Zeller, T., Jalkanen, S., Koskinen, S., Raitakari, O., Durbin, R., MacArthur, D. G., Salomaa, V., Ripatti, S., Daly, M. J., and Palotie, A. (2014). Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. *PLoS Genetics*, **10**(7).
- Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, **11**(7), 499–511.
- McEvoy, B. P., Montgomery, G. W., McRae, A. F., Ripatti, S., Perola, M., Spector, T. D., Cherkas, L., Ahmadi, K. R., Boomsma, D., Willemsen, G., Hottenga, J. J.,

- Pedersen, N. L., Magnusson, P. K. E., Kyvik, K. O., Christensen, K., Kaprio, J., Heikkilä, K., Palotie, A., Widen, E., Muilu, J., Syvänen, A. C., Liljedahl, U., Hardiman, O., Cronin, S., Peltonen, L., Martin, N. G., and Visscher, P. M. (2009). Geographical structure and differential natural selection among North European populations. *Genome Research*, **19**(5), 804–814.
- Park, D. S., Brown, B., Eng, C., Huntsman, S., Hu, D., Torgerson, D. G., Burchard, E. G., and Zaitlen, N. (2015). Adapt-Mix: learning local genetic correlation structure improves summary statistics-based analyses. *Bioinformatics*, **31**(12).
- Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D. P., Patterson, N., and Price, A. L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, **30**(20).
- Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*.
- Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. a. F., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N., Loos, R. J., Frayling, T. M., McCarthy, M. I., Hirschhorn, J. N., Goddard, M. E., and Visscher, P. M. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, **44**(4), 369–375.