

RESEARCH ARTICLE

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

Sina Rüeger<sup>1,2</sup>✉, Aaron McDaid<sup>1,2</sup>✉, Zoltán Kutalik<sup>1,2</sup>\*

**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](mailto: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  $\mathbf{a}_i$  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  $\mathbf{a}_i = \frac{\mathbf{g}^i \mathbf{y}}{N}$  and  $\mathbf{a} = (a_1, a_2, \dots, a_I)' \sim \mathcal{N}(\boldsymbol{\alpha}, \Sigma)$ .  $\Sigma$  represents the pairwise covariance matrix of effect sizes of all  $i = 1, \dots, I$  SNVs.

To estimate the univariate effect size  $\alpha_u$  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]:

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

where  $\mathcal{M}$  is a vector of so-called *tag* 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  $\boldsymbol{\gamma}$  and  $\boldsymbol{\Gamma}$ , 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.  $\alpha_u - \boldsymbol{\Sigma}_{u\mathcal{M}} \boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1} \boldsymbol{\alpha} = 0$ , Eq (1) becomes

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

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

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

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  $\lambda$ . 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

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

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

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

Inserting  $\mathbf{c}_\lambda$  and  $\mathbf{C}_\lambda$ , Eq (2) then becomes

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

Note that  $\lambda$  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  $\lambda$  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  $\lambda$  values:  $\lambda$  fixed at 0.1 [9], and  $\lambda$  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

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

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  $q_{\text{eff}}$  [21]:

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

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,  $z_k^\circ$ , (with  $k \in \mathcal{M}$ ),

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

We define  $N_k$  as the sample size of SNV  $k$ ,  $\mathbf{N}$  as a vector recording the sample size of each tag SNV,  $N_{\text{max}}$  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  $N_{\text{max}}$  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:

$$\mathbf{D}_{kl} = c_{kl} \delta_{kl}. \tag{9}$$

We present two estimators of  $\delta_{kl}$ . 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  $\min(N_k, N_l)$ ,

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

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

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

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

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

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

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

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

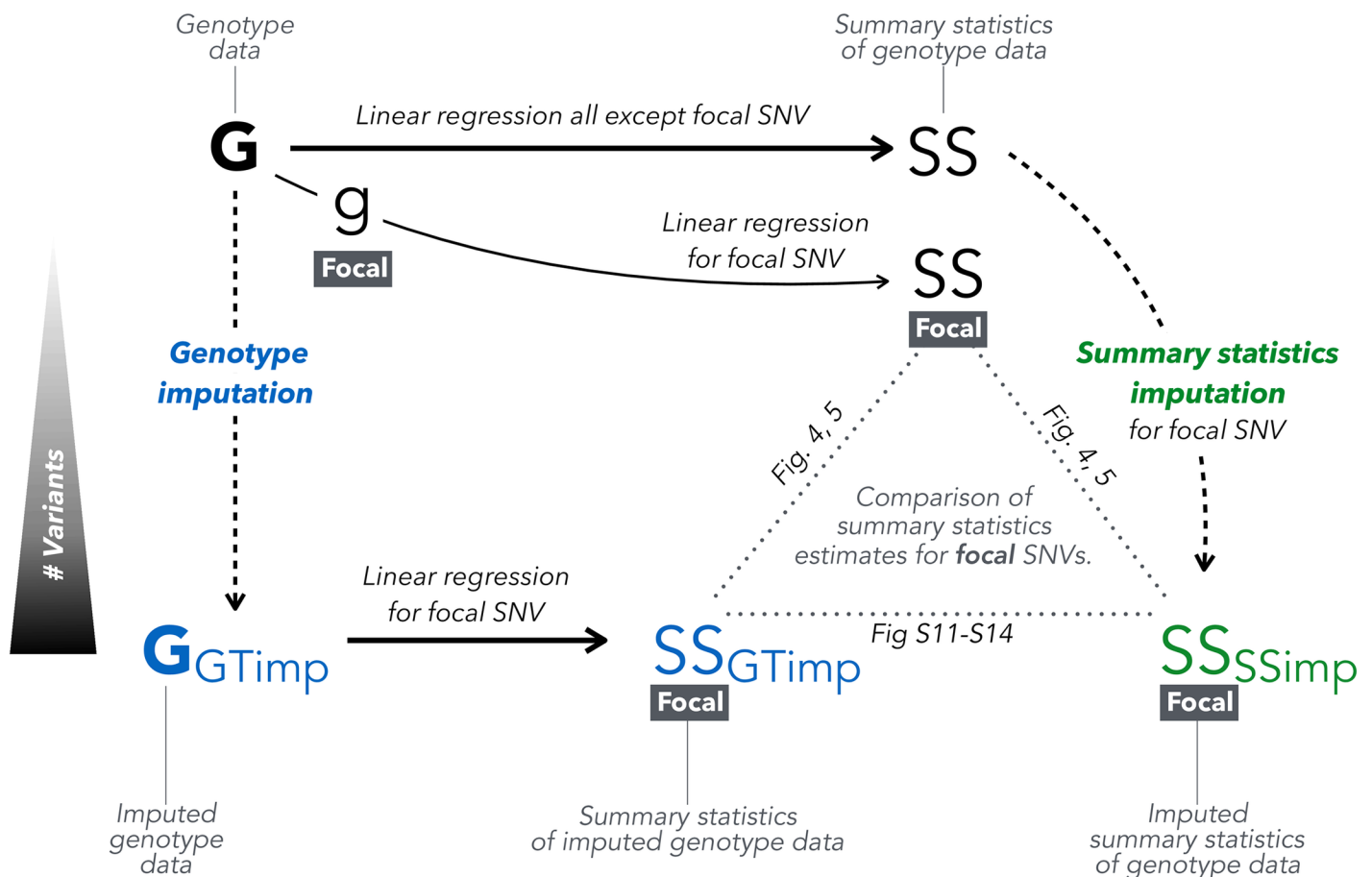
### Comparison of summary statistics imputation versus genotype imputation

**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-GTimp) using *genotype imputation* and then calculate summary statistics via linear regression (bottom-middle, SS-GTimp), 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 at the time and imputing it using summary statistics from neighbouring SNVs. We can then compare the three summary statistics calculated for a particular 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:

$$d_k = Z_k^{SSimp} - Z_k$$

$$RMSE = \sqrt{\frac{1}{K} \sum_{k=1}^K d_k^2}$$

$$Bias = \frac{1}{K} \sum_{k=1}^K d_k$$

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*.

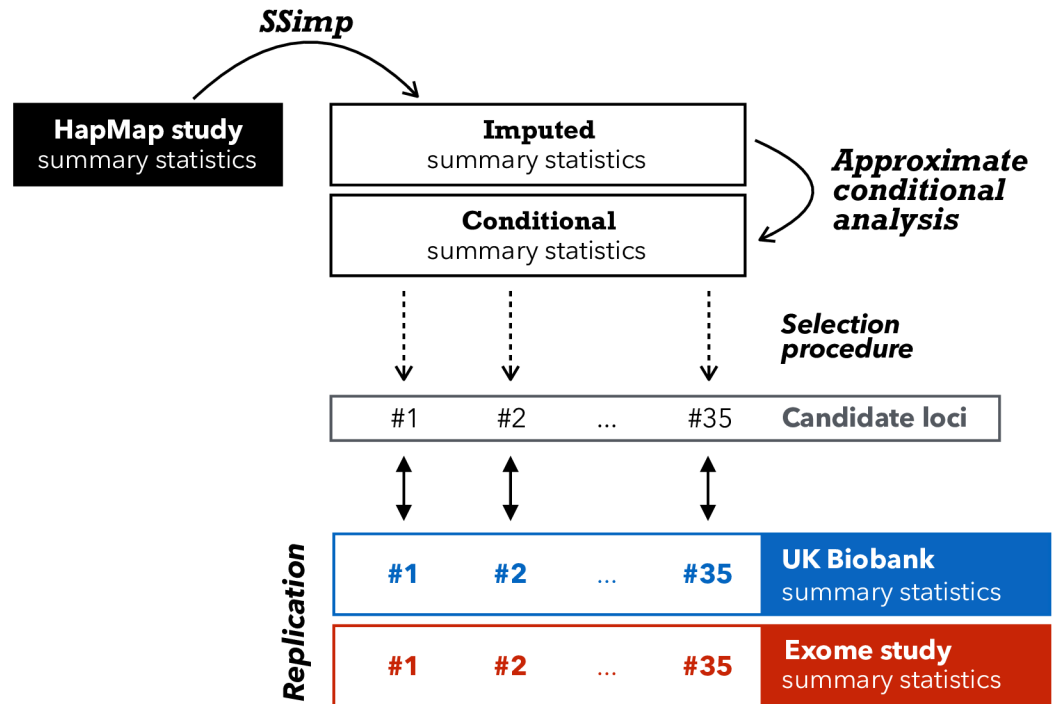
Note that for height-associated SNPs with missing genetic data we rescaled the association Z-statistic  $Z_u$  as follows  $Z_u^* = Z_u \cdot \sqrt{\frac{N_{max}}{N_u}}$  in order to make it comparable with its imputed version ( $Z^{GTimp}, Z^{SSimp}$ ), derived from the full sample.

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  $\pm 0.75$  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  $\alpha$  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:  $Var(q_i) = q_i(1 - q_i)/m_i$ .

**Stratifying results.** The obtained (summary statistics) imputation results were grouped based on the imputed SNVs (i) being correlated (LD > 0.3) to any height-associated SNV on the same chromosome or being a null SNV (LD < 0.05); (ii) low-frequency ( $1\% < MAF \leq 5\%$ ) or common SNV ( $MAF > 5\%$ ); (iii) being badly- ( $\hat{r}_{pred,adj}^2 \leq 0.3$ ), medium- ( $0.3 < \hat{r}_{pred,adj}^2 \leq 0.7$ ) or well-imputed ( $0.7 < \hat{r}_{pred,adj}^2 \leq 1$ ). 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  $8 \approx 0.001 \cdot 3781 \cdot 2$ ), 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  $\hat{r}^2_{pred,adj} \geq 0.3$  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 IBER 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  $\theta_{\text{miss}}$ , 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 SNPTEST-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  $N_{\mathcal{M}}$ . In this section, we used upsampled 1000 Genomes data [1] and simulated phenotype with known standardised effect  $\alpha$  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.

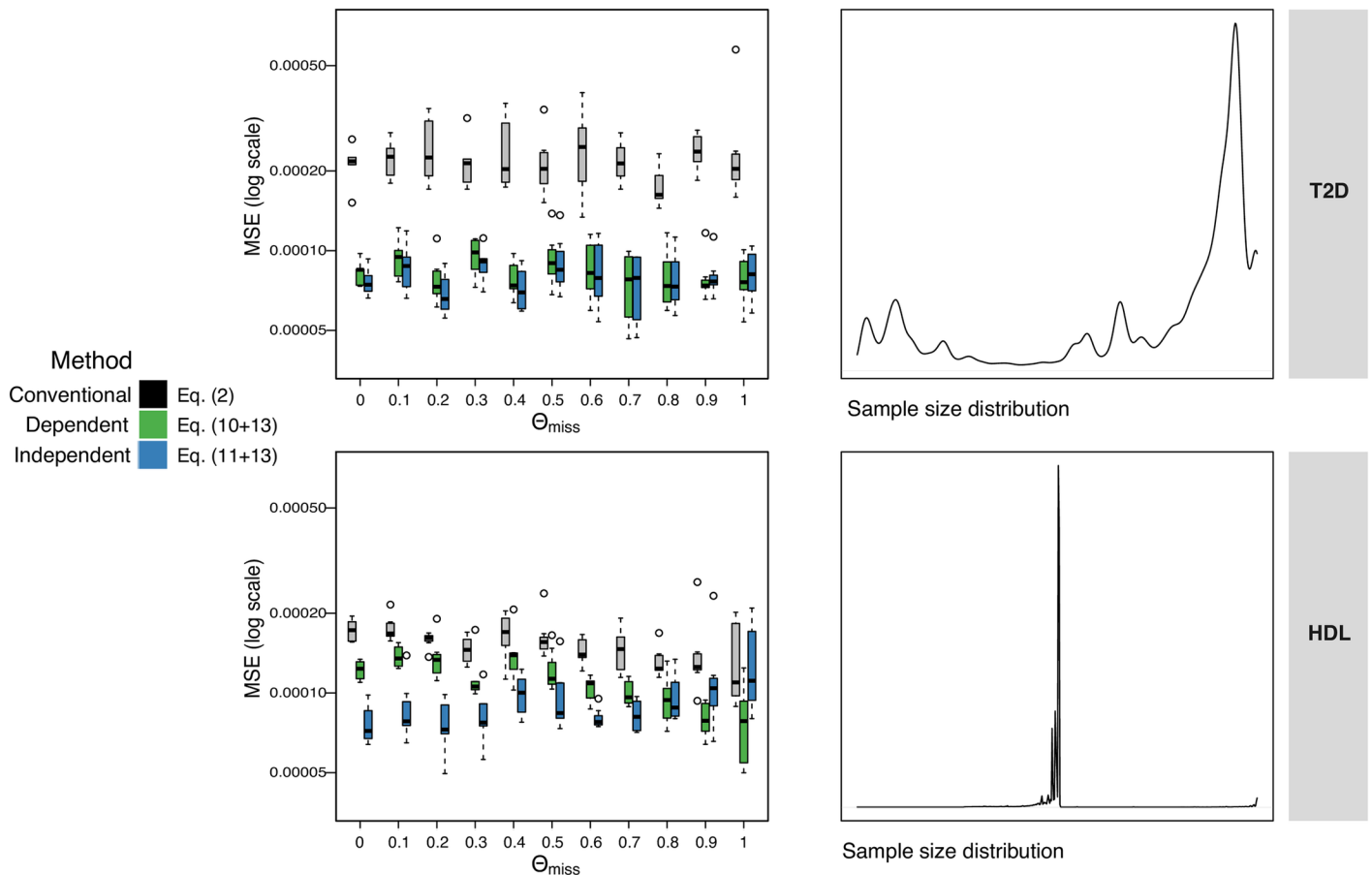
In general, the size of the overlap is unknown and we recommend using the assumption of maximum dependence ( $D^{(dep)}$ ) as it is the most conservative assumption. An alternative is to assume randomly distributed missingness ( $D^{(ind)}$ ). Most pairs of SNVs in GIANT attain close to the maximum possible 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’ ( $\theta_{miss}$ ), 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 ≤ 5% < common; 13’857 and 23’790 variants, respectively); and (iii) imputation quality based on *summary statistics imputation* ( $\hat{r}^2_{pred,adj}$ : low ≤ 0.3 < medium ≤ 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

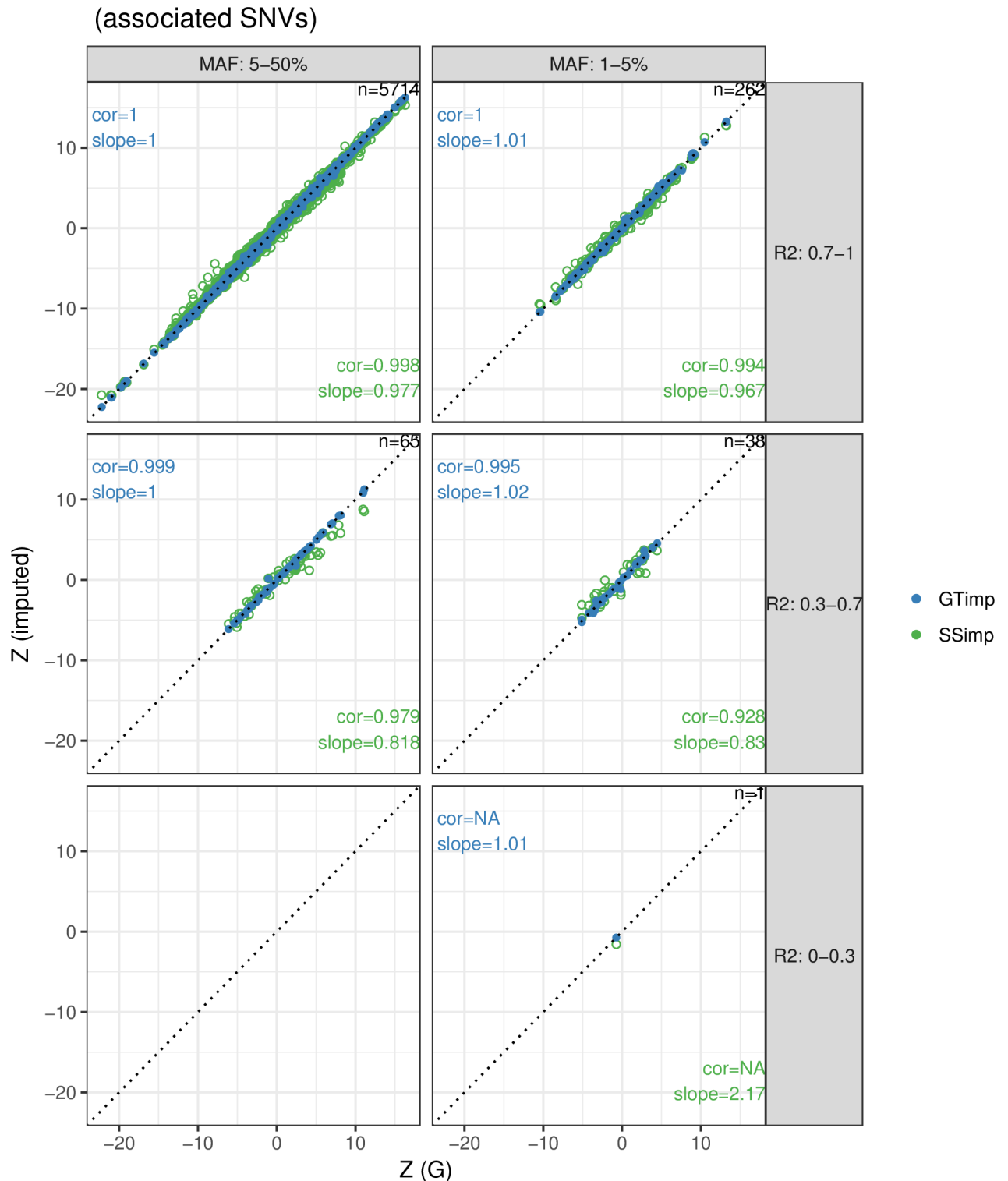
to the number of badly-imputed SNVs (724 variants, among these only one associated variant) to draw meaningful conclusions and hence we limited our analysis to common and low-frequency, and medium- and well-imputed variants.

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 (cor = 0.998 and slope = 0.98), but it drops to cor = 0.928 and a 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.

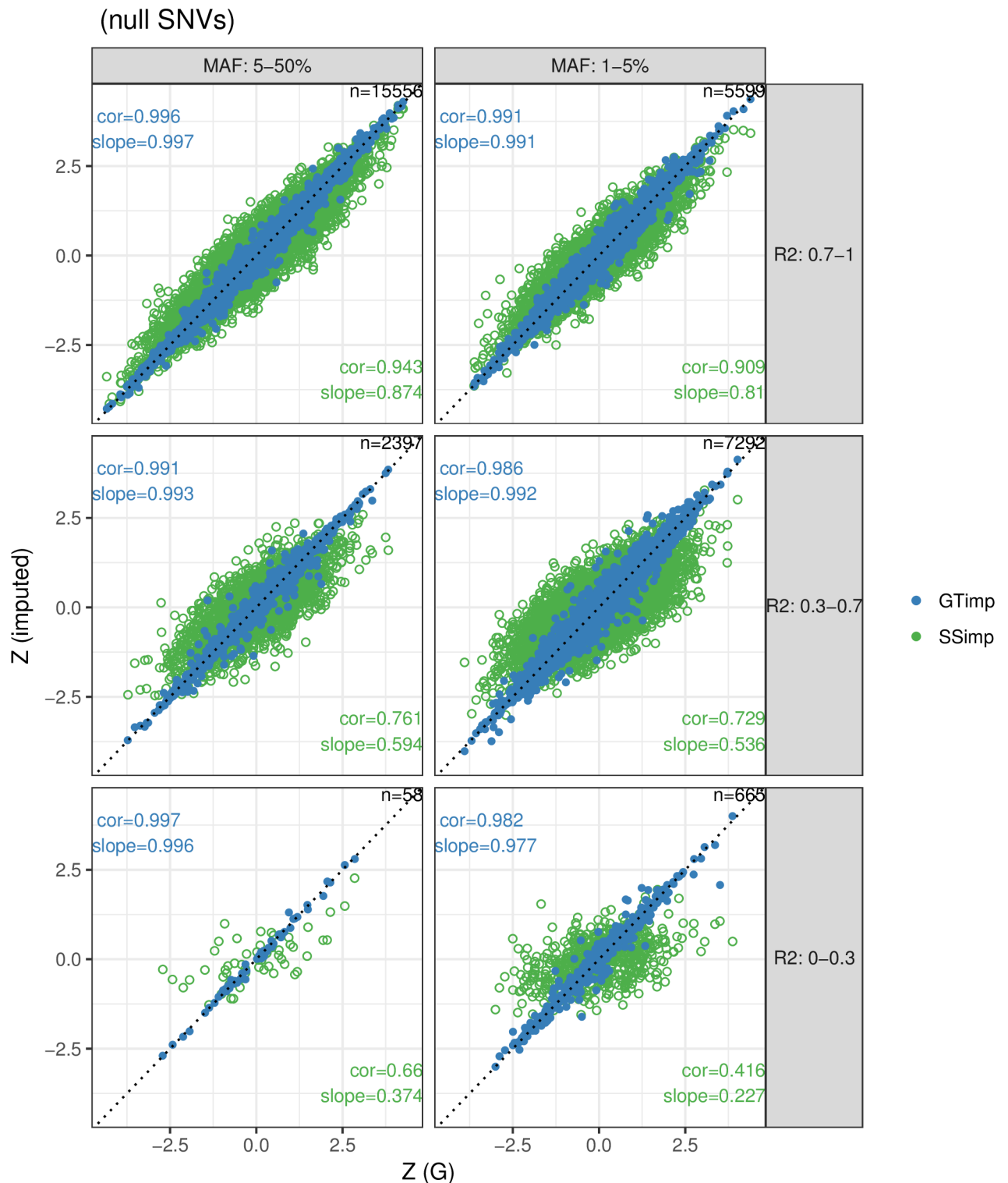
For the 31'567 null SNVs we present the same metrics as for associated SNVs. We analysed 13'556 low-frequency and 18'011 common variants. First, the green dots in Fig 5 show summary statistics from genotype data and *summary statistics imputation*. We find that both the correlation and slope gradually 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.

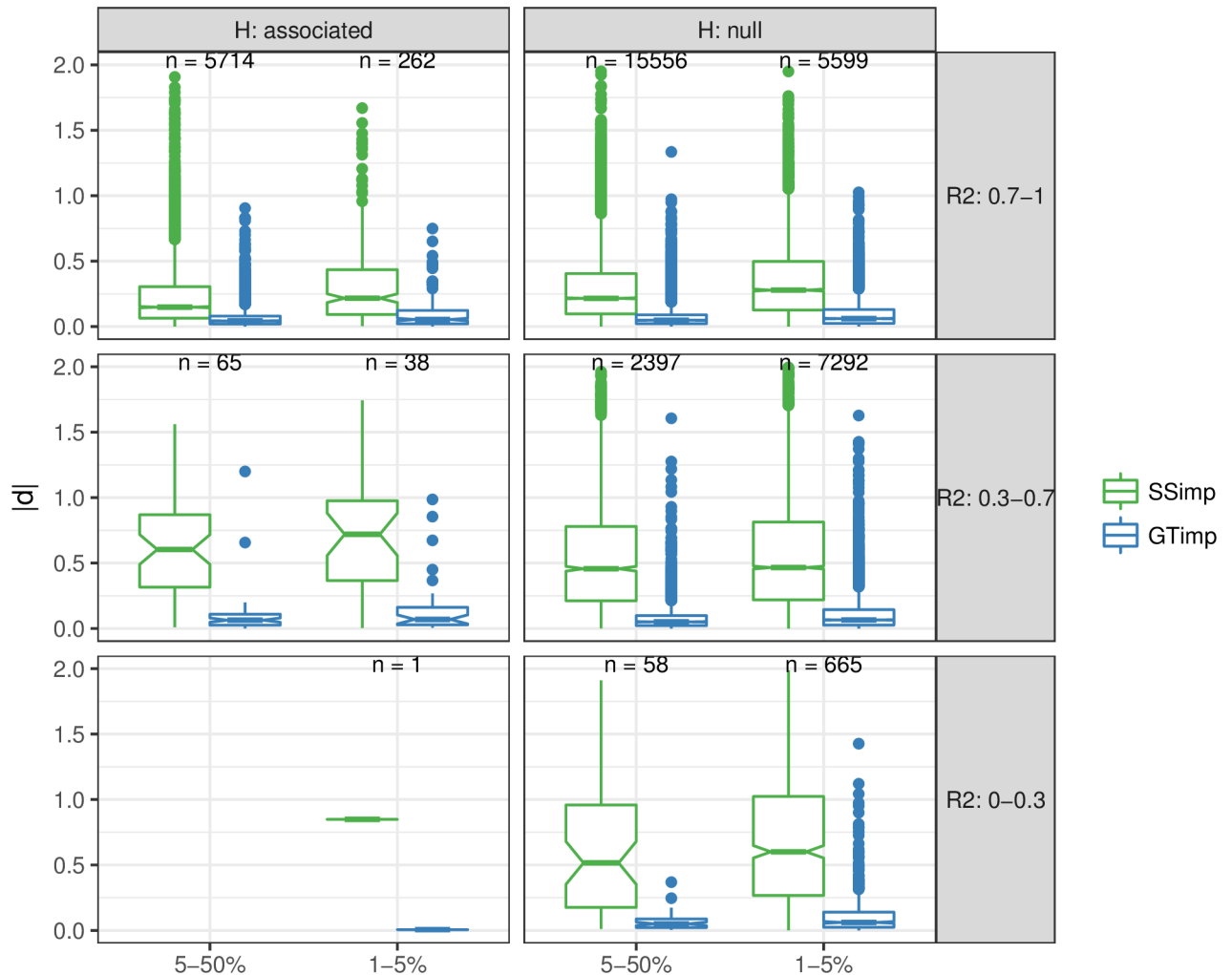
<https://doi.org/10.1371/journal.pgen.1007371.t001>

**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 ( $\alpha$  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

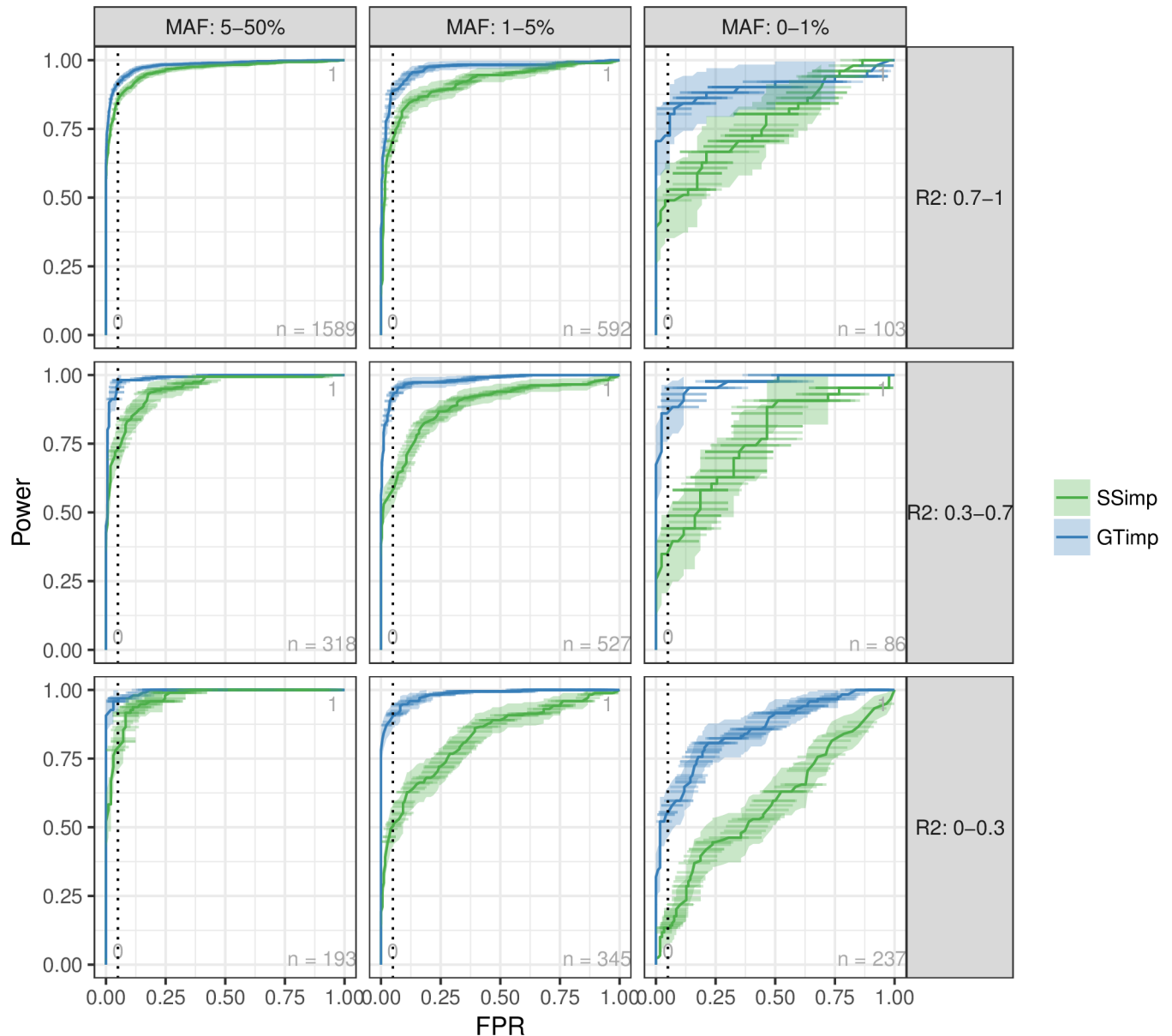


**Fig 6. Visualising RMSE of summary statistics imputation and genotype imputation.** This figure uses boxplots to compare the absolute difference  $|d|$  (used for calculation of RMSE) for each variant between Z-statistics of *summary statistics imputation* (SSimp, green) and *genotype imputation* (GTimp, 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  $\geq 0.3$ .

We subjected all 9'276'018 variants with an imputation quality  $\geq 0.3$  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 ( $\alpha$ ), 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 R/E	Gene <sup>(1)</sup>	MAF <sup>(2)</sup>	SSimp		UK Biobank		Group
							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	DCST1 [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	MYO9B [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	HIST1H2BD [I]	38.85%	2.97E-27	252327	8.64E-18	120086	(iii)
11	rs80171383	11	46084677	G/A	PHF21A [I]	14.72%	3.53E-16	247885	2.05E-16	336474	(iii)
12	rs13108218	4	3443931	A/G	HGFAC [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	MAP2K2 [I]	24.28%	8.17E-10	207571	7.11E-12	336474	(iii)
17	rs7955819	12	20677958	T/C	PDE3A [I]	23.23%	6.13E-10	250048	3.25E-08	336474	(iii)
18	rs7971674	12	1513526	A/T	ERC1 [I]	14.12%	8.10E-09	240270	2.19E-07	336474	(iii)
19	rs12939056	17	7754993	G/A	KDM6B [E]	43.26%	1.09E-12	245015	7.64E-07	336474	(iii)
20	rs58402222	1	46059835	T/TA	NASP [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.  $\hat{r}_{\text{pred,adj}}^2$  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.

(\*)  $r_{s28929474}$ , 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]).



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  $4.21 \times 10^{-14}$ ), 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		
				15	rs78566116	2.74E-19	9.80E-04	Chen D	21896673	Mixed	HPV8 seropositivity in cancer
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.**

$\hat{r}_{pred.adj}^2$	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 ( $\hat{r}_{\text{pred,adj}}^2 \geq 0.7$ ). 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:  $\mathbf{D}^{(ind)}$  for *independent* missingness, which is randomly distributed; and  $\mathbf{D}^{(dep)}$  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 ( $\mathbf{D}^{(dep)}$ ) and independent ( $\mathbf{D}^{(ind)}$ ) approaches. Overall, replacing  $\mathbf{C}$  and  $\mathbf{c}$  with  $\mathbf{D}$  and  $\mathbf{d}$  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  $\mathbf{D}^{(dep)}$  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 (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  $\hat{r}_{\text{pred,adj}}^2$  is conservative and probably underestimates the true imputation quality (S4 Fig). To calculate the imputation quality  $\hat{r}_{\text{pred,adj}}^2$ , 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)  $\hat{r}_{\text{pred,adj}}^2$  is an estimation that can either be erroneous due to choosing the wrong reference panel (and therefore  $\hat{r}_{\text{pred,adj}}^2$  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 ( $r_{s28929474}$ ) 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  $\lambda$  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  $\lambda$  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 ( $\hat{r}_{\text{ridge}}^2$ ). 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 ( $\alpha$ ). 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  $\geq 2$  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  $\lambda_{GC}$  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, `r2` = 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, `r2` = 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. JEPPEG: 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; 36(3):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](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. Rüeeger 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