Improved imputation of summary statistics for realistic settings

Sina Rüeger[1,2,*], Aaron McDaid[1,2,*], Zoltán Kutalik[1,2,◦]

[1] *Institute of Social and Preventive Medicine, Lausanne University Hospital, Lausanne, 1010, Switzerland*
[2] *Swiss Institute of Bioinformatics, Lausanne, 1015, Switzerland*
[◦] *Correspondance should be addressed to Zoltán Kutalik: zoltan.kutalik@unil.ch, tel.: +41-21 314 6750*

[*] *These authors contributed equally.*

October 16, 2017

Motivation: *Summary statistics imputation* can be used to impute association summary statistics of an already conducted, genotype-based meta-analysis to higher genomic resolution. This is typically needed when genotype imputation is not feasible. Oftentimes, cohorts of such a meta-analysis are variable in terms of ancestry and available sample size varies SNP-to-SNP. Both phenomena violate the assumption of current methods that an external LD matrix and the covariance of the Z-statistics are the same.
Results: To address these two issues, we present two extensions to the existing *summary statistics imputation* method. Both extensions manipulate the LD matrix needed for *summary statistics imputation*. Based on simulations using real data we find that accounting for ancestry admixture only becomes relevant with sample size > 1000, while for smaller reference panels the total sample size (irrespective of ancestry) seems more relevant. Furthermore, incorporating varying sample size across SNPs leads to a 2.6-fold reduction in MSE compared to the conventional approach.
Availability: SSIMP is available at
`https://wp.unil.ch/sgg/summary-statistic-imputation-software/`.
Contact: zoltan.kutalik@unil.ch

1 Introduction

Genotype data for genome-wide association studies (GWAS) are often collected using DNA chips, which cover only a small fraction of the variable genome. To be able to combine genome-wide association studies (GWAS) that measured different sets of genetic markers (due to variable commercial array content), genetic information has to be inferred for a common set of markers. The statistical inference of untyped single nucleotide variants (SNV) relies on a set of available genotype data and a set of densely genotyped or sequenced reference panel. Such inference is possible thanks to the fact that physically close SNVs are correlated due to recombination of chromosomes during meiosis. The inference process is termed *imputation* and it has two main types. *Genotype imputation* (Marchini and Howie 2010) first estimates all haplotypes both in the reference panel and the study cohort, then using a Hidden Markov Model every

observed haplotype in the study cohort is assembled as a probabilistic mosaic of reference panel haplotypes. The reconstruction facilitates the computation of the probability of each genotype for every SNV of the reference panel in each individual. Having imputed the genotype data set, one can then run an association scan with an arbitrary trait and obtain association summary statistics. *Summary statistics imputation* on the other hand starts off with association summary statistics available for all genotyped markers and uses the same reference panel to infer directly the association summary statistics of all SNVs available in the reference panel. This latter method starts off with estimating the local pair-wise linkage disequilibrium (LD) structure of each genetic region using the reference panel and then calculates a conditional expectation of normally distributed summary statistics. This latter approach is the central focus of our paper. Although genotype imputation followed by association scan yields in more reliable association statistics, in some situations summary statistic imputation may be preferred. For example in large collaborations, some studies may find genotype imputation challenging in terms of manpower, computational resources or legal restrictions barring data transfer. Oftentimes a study or a consortium has already provided association summary statistics imputed to an older reference panel, in which case it is much faster to re-impute summary statistics to a newer panel without going back to each study analyst and asking for genotype imputation and association analysis. In contrast to genotype imputation, summary statistics imputation is much less demanding on computational resources, hence much faster and requires no access to the actual genetic data.
Methods making use of summary statistics, such as calculating genetic correlation (Bulik-Sullivan et al. 2015) or approximate conditional analysis (Yang et al. 2012), have gained interest in recent years, because they bypass the need of genotype data, but mimic it by making use of external reference panels. These methods could profit from summary statistics being available on an arbitrarily chosen panel of SNVs – provided by summary statistic imputation. Similarly, Mendelian randomisation analysis can also be performed based on summary statistics (Burgess et al. 2013), which boost power substantially. To avoid violation of the assumptions (in particular the instrument being exogenous) it is safer to consider other possible confounders to test robustness of the causal effect estimation. To achieve this, summary statistics from a large number of different studies need to be imputed to the same set of reference SNVs.
Summary statistics imputation in the context of genomic data was first described by Wen and Stephens (2010), where they inferred allele frequencies for an untyped SNV, by a linear combination of observed allele frequencies. Lee et al. (2013, 2014) then further extended the method to the application of linear regression estimates and a covariance matrix shrinkage depending on the reference panel size. Later, Pasaniuc et al. (2014) included a sliding window, which allowed partitioning of the genome into smaller pieces to facilitate imputation on a larger scale, and introduced a different shrinkage approach. Since then a few extensions have been published, that mainly concentrate on summary statistics imputation for admixed populations (Lee et al. 2015a,b, Park et al. 2015), or account for covariates (Xu et al. 2015).
In this paper we present two improvements of the existing methods. First, we propose a method that is able to optimally combine reference panels to best match the GWAS population. Next, we extend the imputation formula to be able to accommodate variable sample size for SNVs with available summary statistics.

2 Methods

2.1 Summary statistics imputation (SSimp)

We assume a set of univariate effect size estimates $\boldsymbol{a}_i$ are available for SNVs $i = 1, \ldots, I$ from a linear regression between a continuous phenotype $\boldsymbol{y}$ and the corresponding genotype $\boldsymbol{g}^i$ measured in N individuals. Without loss of generality we assume that both vectors are normalised to have

zero mean and unit variance. Thus $a_i = \frac{(\boldsymbol{g}^i)' \cdot \boldsymbol{y}}{N}$ and $\boldsymbol{a} = (a_1, a_2, \ldots, a_I)' \sim \mathcal{N}(\boldsymbol{\alpha}, \boldsymbol{\Sigma})$. $\boldsymbol{\Sigma}$ represents the pairwise covariance matrix of effect sizes of all $i = 1, \ldots, I$ SNVs.
To estimate the univariate effect size α_u of an untyped SNV u in the same sample one can use the conditional expectation of a multivariate normal distribution, and express the conditional mean of the effect of SNV u using the effect size estimates of the tag SNVs, where $\mathcal{M}$ is a vector of length p identifying a set of SNVs (Eaton 1983, Pasaniuc et al. 2014):

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

where $\boldsymbol{\Sigma}_{u,\mathcal{M}}$ 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 $\boldsymbol{s} = \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}u}$, $\boldsymbol{S} = \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}$. The corresponding correlation matrices are $\boldsymbol{C} = N \cdot \boldsymbol{S}$ and $\boldsymbol{c} = N \cdot \boldsymbol{s}$ (with $\boldsymbol{\Gamma}$ and $\boldsymbol{\gamma}$ as the corresponding true correlation matrices). Further, by assuming that SNV u and the trait are independent conditioned on the $\mathcal{M}$ markers, i.e. $\alpha_u - \boldsymbol{\Sigma}_{u\mathcal{M}} \boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1} \boldsymbol{\alpha} = 0$, Eq. (1) becomes

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

One can also choose to impute the Z-statistic instead:

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

with $\boldsymbol{z} = \boldsymbol{a}\sqrt{N}$ and N being the respective sample size, when the effect size is small (as is the case in typical GWAS). Note that Pasaniuc et al. (2014) derived the same equation for the imputation of association Z-statistics.
Since LD between SNVs is minimal beyond 250 Kb, we choose $\mathcal{M}$ to include all measured variants at least within 250 Kb from SNV u. To speed up the computation when imputing SNVs genome-wide, we apply a sliding window 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 $\boldsymbol{C}$ (and $\boldsymbol{c}$) we use an external reference panel of n individuals. Since oftentimes the size of $\boldsymbol{C}$ will exceed the number of individuals ($p \gg n$), shrinkage of matrix $\boldsymbol{C}$ is needed to guarantee that it is invertible. But even if p and n are relatively close, or $n > p$, applying shrinkage might lower the error in *summary statistics imputation*. The reason for that is, that low values in $\boldsymbol{C}$, which are shrunk towards zero, may represent pure noise (and zero LD), which can be inflated when inverting the matrix.
To tackle this problem, off-diagonal values of $\boldsymbol{C}$ are shrunk towards zero and the extent of which is characterised by a shrinkage parameter λ (Schäfer and Strimmer 2005). The modified matrix becomes

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

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

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

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

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

Note that λ can vary between 0 and 1, with $\lambda = 1$ turning $\boldsymbol{C}$ to the identity matrix, while $\lambda = 0$ leaves $\boldsymbol{C}$ unchanged. Schäfer and Strimmer (2005) find an optimal λ by minimising the variance of matrix $\boldsymbol{C}$. Wen and Stephens (2010) propose to adjust matrix $\boldsymbol{C}$ in a way that they represent recombination hotspots correctly. A similar idea is, to set small absolute correlation values to 0. Here, we mainly focus on two commonly used λ values: λ fixed at 0.1 (Pasaniuc et al. 2014), and λ changing with the reference panel size n: $\lambda = 2/\sqrt{n}$ (Lee et al. 2014).

Imputation quality

Imputation quality, r^2, is defined as the the squared correlation between the imputed- and true genotype. An r^2 value of 1 means perfect imputation, whereas R^2 of 0 indicates poor imputation (Kutalik et al. 2011). In *summary statistics imputation* this quantity is the total explained variance of a linear model where the imputed genotype is regressed onto all measured markers. It was proposed by Pasaniuc et al. (2014) to be estimated as

$$\widehat{r}^2_{\text{pred,adj}} = \boldsymbol{c}'_\lambda \boldsymbol{C}_\lambda^{-1} \boldsymbol{c}_\lambda \tag{7}$$

with λ= 0.1. In line with shrinkage when estimating effect sizes, we propose to use λ= $2/\sqrt{n}$ and account for the number of parameters and sample size ratio (Theil 1961). 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 p_{eff} (Gao et al. 2008):

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

Negative values are set to zero.

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

For summary statistic imputation we would like to estimate the local LD structure of each region in the GWAS population ($\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}$ and $\boldsymbol{\Sigma}_{u\mathcal{M}}$) and to do so we use a (sequenced) reference population, yielding estimates $\boldsymbol{C}$ and $\boldsymbol{c}$. Clearly, the closer these estimates are to the real values, the better the imputation will be (i.e. smaller the estimation error in Eq. (6)). Our aim is to find a weighted mixture of the reference sub-populations that has an LD structure as similar as possible to the LD in the GWAS population. While the true LD structure of the actual GWAS population is rarely known, the GWAS allele frequencies are routinely calculated (even if not always reported) in meta-analytic studies. In the following we show how this information can be exploited to improve summary statistic imputation.
First, suppose that the reference panel is made up of P subpopulations of sizes $n^{(1)}, n^{(2)}, \ldots, n^{(P)}$. Next, we introduce a set of weights $\boldsymbol{w} = (n^{(1)}, n^{(2)}, \ldots, n^{(P)})/n$, which can be viewed as the collection of weights that give rise to the reference population mixture, i.e. $\sum_{p=1}^{P} w_p = 1$ and $w_p > 0$. We can calculate the covariance as a function of these weights, i.e. for each subpopulation we calculate the covariance separately, and then combine them, weighted by their weights $\boldsymbol{w}$

$$s_{kl}(\boldsymbol{w}) = \sum_{p=1}^{P} w_p \left[s_{kl}^{(p)} + t_{kl}^{(p)} \right] , \tag{9}$$

where $t^{(p)}$ is defined as

$$t_{kl}^{(p)} = \left(\bar{g}_k^p - \bar{g}_k\right)\left(\bar{g}_l^p - \bar{g}_l\right) \tag{10}$$

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

$$s_{kl}^{(p)} = \frac{1}{n^{(p)}-1} \sum_{i \in I_p}^{n^{(p)}} (g_{i,k} - \bar{g}_k^p)(g_{i,l} - \bar{g}_l^p) \tag{11}$$

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

While the reference panel population sizes being fixed at $n^{(p)}$ and we defined $w_p \propto n_p$, we could use any arbitrary weights $\boldsymbol{w}$ in order to match a GWAS population, which has different composition than the reference panel. This manipulation of the covariance estimation can be used to adapt the reference panel population structure towards the population structure that is observed in GWAS summary statistics.

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

$$c_{kl}^{\boldsymbol{w}} = \frac{s_{kl}(\boldsymbol{w})}{\sqrt{s_{kk}(\boldsymbol{w}) \cdot s_{ll}(\boldsymbol{w})}} \tag{12}$$

Our goal is to quantify the MSE between the true GWAS LD matrix ($\Sigma_{\mathcal{M}\mathcal{M}}$) and the LD matrix estimated from the reference panel ($\boldsymbol{C}_{\mathcal{M}\mathcal{M}}^{\boldsymbol{w}}$). Since we cannot estimate the off-diagonal values of the GWAS covariance matrix, we focus on its diagonal elements and estimate them from the GWAS allele frequencies. The MSE of Eq. (9) for SNV k can be written as

$$\text{MSE}\left[s_{kk}\right] = \text{Bias}^2\left[s_{kk}\right] + \text{Var}\left[s_{kk}\right] \tag{13}$$

In short, the MSE of Eq. (9) depends on known quantities (allele frequency $\bar{g}_k^p$ for SNV k in population p; sample sizes $n^{(p)}$ of the reference panel population p; allele frequency for SNV k in the GWAS study: $\bar{g}_k^{\text{obs}}$) and the unknown mixing parameter $\boldsymbol{w}$. More specifically, the variance term is a polynomial of degree eight that depends on $\bar{g}_k^p$, $n^{(p)}$ and $\boldsymbol{w}$, while the squared bias is a polynomial of degree four that depends on $\bar{g}_k^p$, $\bar{g}_k^{\text{obs}}$ and $\boldsymbol{w}$. Details to derivations of the MSE are provided in Supplement A.2.

We aim to find a $\boldsymbol{w}$ that minimises the MSE in Eq. (13) for a set of $\mathcal{M}$ SNVs, for which we know the allele frequencies $\bar{\boldsymbol{g}}^{\text{obs}}$ in the GWAS population and can estimate $\bar{\boldsymbol{g}}^p$ from a reference panel:

$$\boldsymbol{w}^{\text{vm}} = \arg\min_{\boldsymbol{w}} \sum_{k=1}^{\mathcal{M}} \text{MSE}\left[s_{kk}\right] \tag{14}$$

with $\sum_{p=1}^{P} w_p = 1$ and $w_p > 0$. We call this method *variance matching* (VM), as we are utilising the GWAS allele frequencies to match genotype variances. Parameter $\boldsymbol{w}^{\text{vm}}$ gives us an estimation of the population weights used in Eq. (9). A more thorough derivation of Eqs. (9) to (14) can be found in Supplement A.

Finally, we substitute $\boldsymbol{w}^{\text{vm}}$ into Equation Eq. (12) and plug it into Eq. (6):

$$\widehat{a}_u = a_{u|\mathcal{M}} = \boldsymbol{c}^{w'} (\boldsymbol{C}^{w})^{-1} \boldsymbol{a} \tag{15}$$

Park et al. (2015) proposed an elegant, generalised approach to weight population LD structure. Their algorithm *Adapt-Mix* choses weights $\boldsymbol{w}^{am}$ based on optimising an objective function. In the case of imputation the objective function is the MSE of imputed Z-statistics at observed SNVs. Lee et al. (2015a) developed *Distmix*, which is similar to parts of Eq. (S8), but does only

minimise the Euclidean distance between allele frequencies, which is highly suboptimal ignoring the variance-bias trade-off.

2.3 Varying sample size and missingness

All previously published summary statistic imputation methods assume that all effect estimates are based on the the same set of N individuals. This assumption does not always hold, for example when meta-analysing studies use different genotyping chips or different imputation 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.

$\boldsymbol{C}$, as defined above, is an estimate of $\Gamma_{\mathcal{M}\mathcal{M}}$, which is the correlation matrix due to LD among the tag SNVs in the current region. Define $\boldsymbol{N}$ as a vector recording the sample size of each tag SNV. Define N_{max} as the maximum in $\boldsymbol{N}$, and assume that every sample of individuals is a subset of a complete sample of N_{max} individuals.

For each tag SNV k, we have observed a 'partial' Z-statistic, z_k°, computed over the N_k individuals. Our goal is to impute, for every SNV, what the 'complete' Z-statistic, z_u, would be in the complete sample of N_{max} people. For any SNV u, where u may or may not be in $\mathcal{M}$, we wish to impute $z_u|\boldsymbol{z}_{\mathcal{M}}^\circ$. Note that if u is in $\mathcal{M}$, it is analogous to the situation when missing genotypes at typed SNVs are imputed.

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

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

where $c_{uk} = \widehat{\Gamma}_{uk}$ is our estimate of the LD-correlation between the two SNVs. We also require the correlations among the partial Z-statistics, $\boldsymbol{z}_{\mathcal{M}}^\circ$. Where $k \in \mathcal{M}$ and $l \in \mathcal{M}$, and $N_{k\cap l}$ is the number of individuals that contributed to the calculation of the effect estimates for both SNVs k and l,

$$\mathrm{Cor}[z_k^\circ, z_l^\circ] = c_{kl}\frac{N_{k\cap l}}{\sqrt{N_k N_l}}$$

By defining $\delta_{kl} = \frac{N_{k\cap l}}{\sqrt{N_k N_l}}$, we can then define the adjusted (estimated) correlation matrix $\boldsymbol{D}$, with cells defined as follows:

$$\boldsymbol{D}_{kl} := c_{kl}\delta_{kl}\ .$$

$\boldsymbol{D}$ and $\boldsymbol{d}$ are therefore defined as adjusted versions of $\boldsymbol{C}$ and $\boldsymbol{c}$ respectively. $\boldsymbol{C}$ and $\boldsymbol{D}$ are $p\times p$ matrices, where p is the number of tag SNVs, and they are recomputed in each region. $\boldsymbol{c}$ and $\boldsymbol{d}$ are vectors of length p, and are recomputed for each target SNV.

The conditional distribution is

$$(z_u - \mathbb{E}[z_u])|\boldsymbol{z}_{\mathcal{M}}^\circ \sim \mathcal{N}\big(\boldsymbol{d}'\boldsymbol{D}^{-1}(\boldsymbol{z}_{\mathcal{M}}^\circ - \mathbb{E}[\boldsymbol{z}_{\mathcal{M}}^\circ])\ ,\ 1 - \boldsymbol{d}'\boldsymbol{D}^{-1}\boldsymbol{d}\big).$$

Applying the simplifying assumption that $\mathbb{E}[z_u] \approx \boldsymbol{d}'\boldsymbol{D}^{-1}\mathbb{E}[\boldsymbol{z}_{\mathcal{M}}^\circ]$, similar to the assumption that took us from Eq. (1) to Eq. (2) above,

$$z_u|\boldsymbol{z}_{\mathcal{M}}^\circ \sim \mathcal{N}\big(\boldsymbol{d}'\boldsymbol{D}^{-1}\boldsymbol{z}_{\mathcal{M}}^\circ\ ,\ 1 - \boldsymbol{d}'\boldsymbol{D}^{-1}\boldsymbol{d}\big)$$

and therefore we impute $z_u|\boldsymbol{z}^{\circ}_{\mathcal{M}}$ as

$$\hat{z}_u = \mathbb{E}[z_u|\boldsymbol{z}^{\circ}_{\mathcal{M}}] = \boldsymbol{d}'\boldsymbol{D}^{-1}\boldsymbol{z}^{\circ}_{\mathcal{M}} \ . \tag{16}$$

In order to convert $\hat{z}_u$ into the corresponding estimate of the standardised effect, we consider the (hypothetical) process of imputing each individual genotype.
If we had the individual-level genetic data, with j to index individuals, each element of the N_{max}-element vector $\boldsymbol{g}^u$ for SNV u could be imputed using genotypes from the tag SNVs $G^{\mathcal{M}}$ via $\widehat{g}^u_j = \boldsymbol{c}'_{\mathcal{M}_{(j)},u}\boldsymbol{C}^{-1}_{\mathcal{M}_{(j)},\mathcal{M}_{(j)}}G^{\mathcal{M}}_j$, where the set $\mathcal{M}_{(j)}$ can be different for each individual as each individual has a different set of tagged SNVs. The corresponding standardised effect estimate, based on linear regression, would be

$$\widehat{a}_u = \mathbb{E}[a_u|\boldsymbol{a}_{\mathcal{M}}] = \frac{(\widehat{\boldsymbol{g}}^u)'\boldsymbol{y}}{(\widehat{\boldsymbol{g}}^u)'\widehat{\boldsymbol{g}}^u}$$

The denominator of this is $(\widehat{\boldsymbol{g}}^u)'\widehat{\boldsymbol{g}}^u = N_{max}\boldsymbol{d}'\boldsymbol{D}^{-1}\boldsymbol{d}$, as opposed to $(\boldsymbol{g}^m)'\boldsymbol{g}^m = N_{max}$, and we define the *effective sample size* as $N_{max}\boldsymbol{d}'\boldsymbol{D}^{-1}\boldsymbol{d}$. Therefore, even though we do not have the per-individual genetic data, we can impute the standardised effect a_u via

$$\widehat{a}_u = \mathbb{E}[a_u|\boldsymbol{z}^{\circ}_{\mathcal{M}}] = \frac{\hat{z}_u}{\sqrt{N_{max}\boldsymbol{d}'\boldsymbol{D}^{-1}\boldsymbol{d}}} \ . \tag{17}$$

Estimating overlap $N_{k\cap l}$ and δ

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 $\boldsymbol{N}$. Therefore, we must derive the sample overlap based on assumptions about the dependence structure of missingness. 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, $N_{k\cap l} = \min(N_k, N_l)$, and minimised when the sample overlap is at its minimum. If the missingness vectors are independent of each other, the expected overlap can be estimated as $N_{k\cap l} \approx \frac{N_k N_l}{N_{max}}$.
The most conservative assumption is maximum possible overlap, resulting in maximum dependence, as this minimises the imputation quality, $\boldsymbol{d}'\boldsymbol{D}^{-1}\boldsymbol{d}$. We therefore construct a $\boldsymbol{D}$ matrix using

$$\boldsymbol{D}^{(dep)}_{kl} = \boldsymbol{C}_{kl}\hat{\delta}^{(dep)}_{kl} = \boldsymbol{C}_{kl}\min\left(\frac{\sqrt{N_k}}{\sqrt{N_l}}, \frac{\sqrt{N_l}}{\sqrt{N_k}}\right) \ . \tag{18}$$

and plug $\boldsymbol{D}^{(dep)}$ into Eqs. (16) and (17). We will use $\boldsymbol{D}^{(dep)}$ as the primary method in the experiments in the missingness subsection in the results section.
We will also consider the following, which is based on the assumption of independent missingness vectors:

$$\boldsymbol{D}^{(ind)}_{kl} = \boldsymbol{C}_{kl}\hat{\delta}^{(ind)}_{kl} = \boldsymbol{C}_{kl}\frac{\sqrt{N_k N_l}}{N_{max}} \ . \tag{19}$$

2.4 Reference panels

As reference panels we used genetic data from the 1000 Genomes project (1KG)(1000 Genomes Project Consortium 2010).

2.5 Simulation

For simulation of GWAS summary statistics we used data from the five European subpopulations `CEU, GBR, FIN, TSI` and `IBR` of the 1KG (1000 Genomes Project Consortium 2010). We chose to up-sample chromosome 15 using HAPGEN2 (Su et al. 2011) to 5'000 individuals for each subpopulation, yielding a total of 25'000 individuals. Of these we used half to generate a GWAS with an *in silico* phenotype. The remaining 12'500 individuals were used as reference panel for summary statistic imputation. The simulation procedure is described in more detail in Supplement B.

3 Results

Summary statistics imputation works through combining summary statistics from a set of SNVs with pairwise SNV LD information obtained from an external reference panel. We extended the most recent summary statistics method (Pasaniuc et al. 2014) by the following two features: the estimation of standardised effect sizes robust to variable sample missingness, and optimisation of the reference panel composition.

To assess our extended *summary statistics imputation* method we used upsampled datasets, yielding $25'000$ European individuals in total[1]. GWASs were simulated using *in silico* phenotypes. This semi-simulation framework allows us to study reference panel sizes up to $12'500$ (half of each sub-population) as well as the impact of reference panel composition.

In principle, we accessed summary statistics, masked a fraction of SNVs and imputed them, then assessed imputation accuracy. When imputing a single SNV we used tag SNVs within at least 250 Mb. For the imputation of an entire region we used a sliding window of 1 Mb with 250 Kb flanking regions on each side. The imputation accuracy was measured by the MSE between the true and the imputed effect sizes. In the following we assess our new *variance matching* and varying sample size approach.

3.1 Improving summary statistic imputation via variance matching

To compose an optimal $\boldsymbol{C}$ and $\boldsymbol{c}$, we apply *variance matching* that utilises a function of allele frequencies as proxy for population structure. We compare this novel method to an existing approach called *Adapt-Mix* and the best possible reference panel composition.

We used an extended version of our simulation set up. In short, we fix the proportion of the Finnish subpopulation of 1KG-EUR in the GWAS, then let the proportion of this population vary in the reference panel used for LD estimation. We repeat this for different proportions in the GWAS and in the reference panel and calculate each time the MSE between the true and the imputed effect sizes ($h(\boldsymbol{w})$, Eq. S10) and determine the best possible w^* (Eq. (S11)). In parallel, we apply for each fixed proportion of Finnish in the GWAS the *variance matching* approach to determine their optimal weight in the reference panel (Fig. S2). To identify other factors that influence the choice of weights, we group SNVs into low and high F_{ST}, and run the simulation from small to large reference panels. For all Figures we include results from *Adapt-Mix*, which we implemented for comparison.

Ultimately, we are interested in a comparison to the standard reference panel, which has roughly equal weights in the European sub-panel. In Fig. 1 we compare MSE of the three optimal weights determined by each method (w^* best possible, w^{vm} *variance matching*, w^{am} *Adapt-Mix*) to the MSE when using equal weights (i.e. $w = 0.2$).

[1] Our complementary paper Rüeger et al. (2017) assesses the extended *summary statistics imputation* method with real data, using UK Biobank to compare *summary statistics imputation* to *genotype imputation*, and GIANT summary statistics to show how *summary statistics imputation* can be utilised to identify novel associations.

From the extensive simulation results it is clear, that *variance matching* offers performance very close to the theoretically best population mixing w^*. We find that both optimising methods show a similar trend (greater advantage of using specific weights for population specific markers, large reference panel and heterogeneous GWAS). Except for one instance, *variance matching* offers equal or bigger benefit than *Adapt-Mix*. Since the reference panel is equally partitioned, we observe for an equally partitioned GWAS a ratio of $h(w^*)/h(w = 0.2)$ close to one for all three methods, independent of reference panel or whether the variants are population specific (function $h(\boldsymbol{w})$ is the MSE between the true and the imputed effect sizes described in Eq. S10). For a reference panel of 500 individuals, population specific variants and a GWAS with 80% Finnish individuals we observe a theoretically best possible advantage of 0.886, while its 0.892 for *variance matching* and 0.926 for *Adapt-Mix*. When increasing sample size to $12'500$ the benchmark advantage becomes 0.648 and 0.658 *variance matching* and *Adapt-Mix*, respectively. For variants that are not population specific we see a similar trend with increasing fraction of Finnish individuals and reference panel size, but less pronounced. For a reference panel of 500 individuals, population specific variants and a GWAS with 80% Finnish individuals we observe a maximum advantage of 0.957, while its 0.959 for *variance matching* and 0.966 for *Adapt-Mix*. When increasing sample size to $12'500$ it drops to 0.742, 0.747 and 0.751, for w^*, *variance matching* and *Adapt-Mix*, respectively.

3.2 Varying sample size and missingness

The conventional estimate of the standardised effect of a SNV u, based on a vector estimated effects $\boldsymbol{a}_{\mathcal{M}}$, where $\mathcal{M}$ is a vector of nearby tag SNVs, is

$$\widehat{a}_u^{(conv)} = \boldsymbol{c}'\boldsymbol{C}^{-1}\boldsymbol{a}_{\mathcal{M}} \ ,$$

and $\boldsymbol{C}$ is the estimated correlation due to LD among those $\mathcal{M}$ SNVs and $\boldsymbol{c}$ is a vector of estimated correlations between SNV u and the set of $\mathcal{M}$ SNVs. This is unbiased, under certain assumptions, but can have large variance when there is variation in the sample sizes recorded in $\boldsymbol{N}_{\mathcal{M}}$.

In this section, we simulate data with known standardised effects, α, and apply various estimators to estimate the standardised effect. We compare the MSE of the conventional estimation $\widehat{a}_u^{(conv)}$ to the MSE of two other estimators, $\widehat{a}_u^{(dep)}$ and $\widehat{a}_u^{(ind)}$, derived in the method section. If two SNVs were observed in two different samples, the correlation between the Z-statistics (or between the estimates of standardised effect) will decrease with the number of individuals in common (the 'sample overlap') between the two samples. The latter two estimators take this into account, replacing $\boldsymbol{C}$ with matrix $\boldsymbol{D}$, where the correlation estimates have been adjusted.

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 GIANT (Genetic Investigation of ANthropometric Traits) study, we demonstrate in Fig. S7 that the size of the overlap is generally larger than would be the case under a strict 'missing independently at random' assumption. In general, we assume the size of the overlap is unknown and we recommend using the assumption of maximum dependence as it has lower imputation quality and is therefore the most conservative assumption. For many pairs of SNVs, the overlap in GIANT attains close to the maximum possible and therefore this conservative assumption is not overly-conservative.

We adjust the estimated LD-correlation matrix $\boldsymbol{C}$, where we know the vector of sample sizes $\boldsymbol{N}_{\mathcal{M}}$ for the tag SNVs $\mathcal{M}$, along with the corresponding z-statistics $\boldsymbol{z}_{\mathcal{M}}$ (related to standardised effect estimates $a_k = z_k/\sqrt{N_k}$), and assuming that all samples are a subset of a set of $N_{max} =$

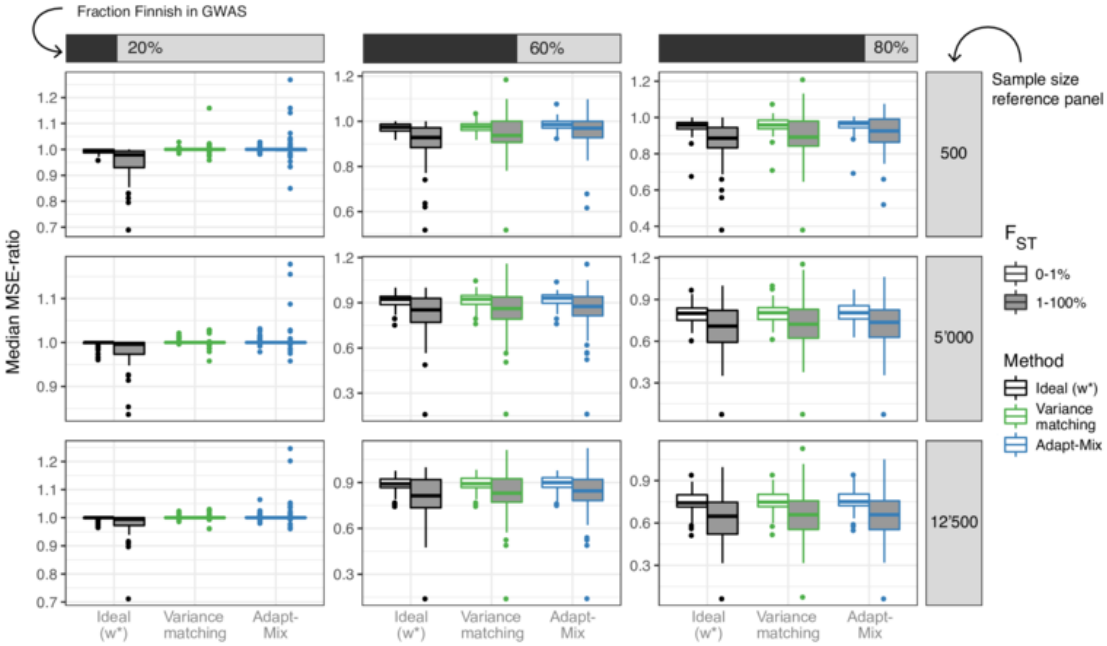

Figure 1: **Accounting for population structure.** This figure shows the comparison of each w-optimisation method with respect to choosing the full European panel of 1000 Genomes project (which corresponds to equal weights). Each boxplots represents 74 simulated regions. The x-axis shows the three different strategies: using theoretical best possible weights in black (if the true effect were to be known), *variance matching* in green and *Adapt-Mix* in blue. The y-axis shows the ratio between the MSE when choosing the weights according to the respective optimisation and the MSE when choosing equal weights for all populations (hence a weight of 0.2 for all five populations), i.e. in black $h(\boldsymbol{w}^*)/h(0.2)$, in green $h(\boldsymbol{w}^{\mathrm{vm}})/h(0.2)$, and in blue $h(\boldsymbol{w}^{\mathrm{am}})/h(0.2)$. Function $h(\boldsymbol{w})$ is the MSE between the true and the imputed effect sizes described in Eq. S10. Values on the y-axis smaller than 1 show a smaller MSE in imputation with a specific $\boldsymbol{w}$ compared to the choice of an unadjusted reference panel with equal weights, while values larger than 1 indicate a higher MSE. The results are split into columns and rows, with the rows for different reference panel sizes and the columns different Finnish fractions in the GWAS populations. Variants are also grouped according to F_{ST}, with population specific variants labeled as solid dots, and population unspecific variants as circles. The y-axis range varies for each window (for fixed y-axis range see Fig. S4). Table `stable_01.csv` provides the same information as a numeric table.

$\max(\boldsymbol{N})$ individuals. We use this adjusted matrix $\boldsymbol{D}$ in place of $\boldsymbol{C}$ to impute the Z-statistic of a target SNV u, and then divide it by the square root of an *effective sample size* to estimate the standardised effect (Eqs. (18) and (19)).
To evaluate these estimators simulated genetic data on 25'000 individuals were used as described in Supplement B. In brief, as before, half of the data were used to estimate the LD structure $\boldsymbol{C}$ and the other half to simulate the association study. 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 (Global Lipids Genetics Consortium 2013) and T2D (Morris et al. 2012)). Missingness was assigned at random positions while respecting the missingness correlation parameter θ_{miss}, with zero value reflecting missingness at random and one corresponding to the maximum possible sample overlap between SNVs.
The results in Fig. 2 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 2), 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. 2, 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. 2), i.e. the sample overlap between each pair of SNVs is as large as possible given their two sample sizes, Eq. (18) performs better (as expected). With lower overlap (first column) Eq. (19) performs better.

4 Discussion

Post-GWAS methods that take summary statistics as input have increasing demand, but often require denser summary statistics than available. To tackle this problem, *summary statistics imputation* can be used to directly impute (often publicly available) summary statistics. However, the drawback of methods that work on summary level, is, that they require the estimation of covariance matrices, which is challenging due to the current lack of large, openly available reference panels. The current situation of reference panels requires careful considerations of the choice of subpopulations, and controlled manipulations of the correlation matrix using a shrinkage parameter. To tackle some of these limitations, we extended *summary statistics imputation* as presented in Pasaniuc et al. (2014) by two features: optimal combination of $\boldsymbol{C}$ and $\boldsymbol{c}$ from reference panel subpopulations and adaptive estimation of $\boldsymbol{C}$ and $\boldsymbol{c}$ to account for varying sample size in GWAS.

Choice of reference panel and variance matching

Formulae for *summary statistics imputation* have two components: GWAS summary statistics (Z-statistics or effect size estimations) and LD matrix estimations from an external reference panel. The LD matrix is needed to represent the correlations among SNVs in the region of interest. If the LD estimation is poor (in extreme case using vastly different ancestry groups in GWAS and in the reference panel), the relationship between variants is wrongly estimated, and the imputed summary statistics become erroneous. Therefore, a crucial aspect of *summary statistics imputation* is the choice of the reference panel and the estimation of the LD structure. However, oftentimes the reference panel is chosen ad-hoc, guessing an underlying GWAS admixture. Unlike for genotype imputation, where the Hidden Markov Model (HMM) estimates the most probable underlying haplotypes and hence the proportion of the reference panel populations are much less relevant, for *summary statistics imputation* the composition of the reference

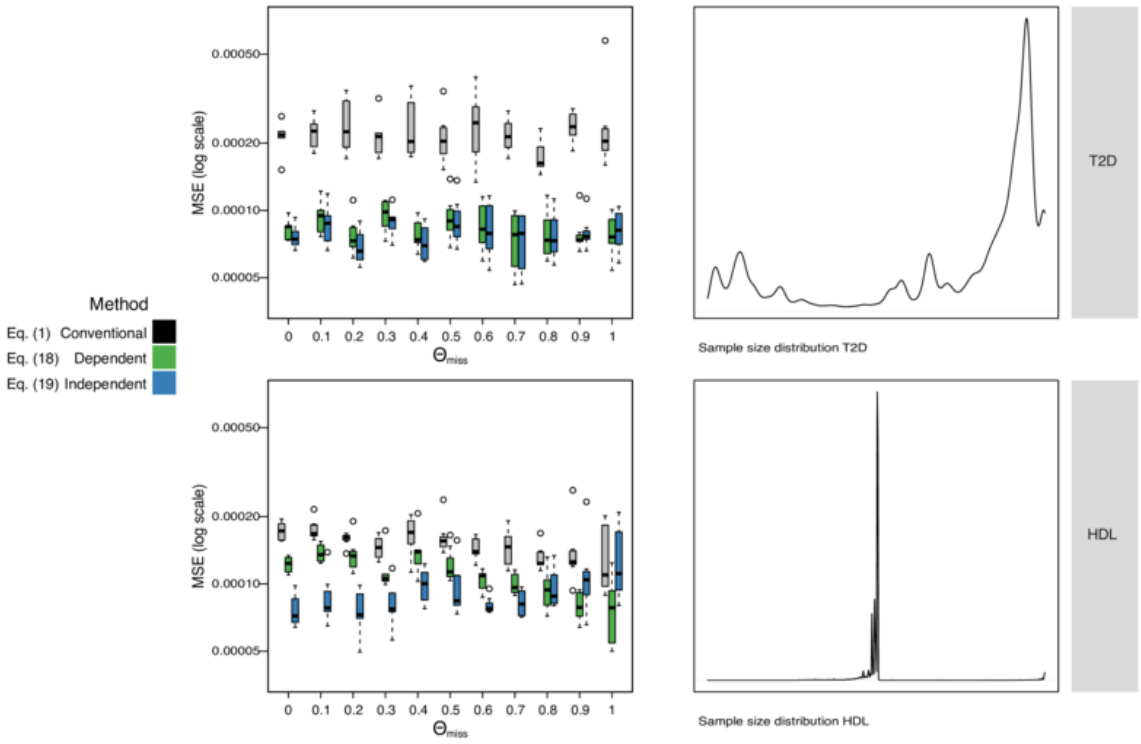

Figure 2: **Accounting for variable sample size.** Effect of missingness on accuracy of imputation of standardised effects, evaluated via simulations where true effect is known. The y-axis is the MSE between the true standardised effect and the conventional estimate which ignores missingness (Eq. (1), grey), our estimate assuming maximum sample overlap (Eq. (18), green), and our estimate assuming missingness-at-random (Eq. (19), blue). The x-axis is the 'missingness-correlation' (θ_{miss}), where a value of 1 means the number of individuals in the samples had maximum overlap with each other, and 0 means they were simulated independently leading to smaller overlap. Each boxplot shows the MSEs across the 40 regions simulated. Top row is where the $\boldsymbol{N}$'s (simulated sample sizes) are selected randomly from a study of T2D (Morris et al. 2012), with sample sizes varying between 13 and $110'219$ individuals. Bottom row is based on HDL (Global Lipids Genetics Consortium 2013), 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.

panel is crucial.
To tackle this problem in *summary statistics imputation* literature, Park et al. (2015) proposed *Adapt-Mix*, a method that utilises the GWAS summary statistics to improve the LD estimation. We developed *variance matching*, which leverages the knowledge of the GWAS allele frequencies. The ground principle of both methods is to build the local LD structure as a linear combination of population specific covariance estimations, with weights determined by the method. Both methods take a set of reference panels as input, and a vector containing observed statistics (e.g. Z-statistics for *Adapt-Mix* or allele frequency estimations for *variance matching*), and return a set of weights for each population contained in the reference panel. Both methods assume that the GWAS sample is composed of a mixture of the populations (whether it is an admixed population or a mixture of different populations), where we have a reference panel for each population. Such a mixture is defined as a set of weights, with one weight for each population. The optimal set of weights selected by *variance matching* or *Adapt-Mix* might approximate the true mixing weights in the GWAS sample. However, it will often deviate from that in an attempt to minimise the MSE. For example, given a GWAS of Finnish only, we could either choose the 99 individuals from 1000 Genomes project as reference panel (which corresponds to assigning a weight of 1 to the Finnish population), or select all 503 individuals of the European panel (which corresponds to a weight of roughly 0.2 for the Finnish population). The precision of the correlation matrix estimation $\boldsymbol{C}$ increases with the number of individuals available. On the other hand, allele frequencies match best when population proportions match (in this case: 1 for FIN, 0 for GBR, IBR, TSI and CEU). Our approach aims to strike a balance between the two aspects (bias and variance) to find an optimal weight somewhere between 0.2 and 1 for this specific example. We find that for an imbalanced GWAS this bias-variance trade-off leads for smaller reference panels ($n = 500$) to optimal weights pulled towards 0.2 and for larger reference panels ($n = 12'500$) to optimal weights that nearly match the weight in the GWAS (Fig. 3).
Variance matching performs better than *Adapt-Mix* in all subgroups that we explored, except one instance, where both method have an equal performance (Fig. 1). With *variance matching* we are exploring a similar approach as *Adapt-Mix* does, but use allele frequencies as population admixture proxies instead of Z-statistics.
Variance matching assumes that the population admixture that is reflected in the variance of tag SNVs (diagonal in matrix $\boldsymbol{C}$), is the same for the covariance between tag SNVs (off-diagonal of $\boldsymbol{C}$) and the covariance between tag SNVs and SNVs to impute (matrix $\boldsymbol{c}$). Furthermore, we provide an analytical solution to Eq. (13), which involves approximations of the variance and the bias (Eq. S7 and S8).
Finding a reference panel that has a matching ancestry composition with the GWAS summary statistics is difficult out of three reasons: 1) the degree of population admixture of populations is usually unknown (whether it is admixed ancestry, or a mixture of different populations), and publicly available reference panels have 2) a limited number of populations with a 3) limited number of individuals. With *variance matching* we are only addressing 1). To this end, we could not validate our approach in real data as diverse reference panels with sample size > 500 per population are not publicly available yet.

Varying sample size and missingness

To increase imputation accuracy when using tag SNPs with varying sample size, we propose an approach that shrinks the correlation matrix according to sample size overlap. We present two ways of estimating such a sample size overlap: one addressing randomly distributed missingness (termed *independent*) and one addressing highly correlated missingness (termed *dependent*). To evaluate the performance ot these two methods we used simulation data and induced two different distributions of missingness (a narrow and a wide range of sample sizes) and different

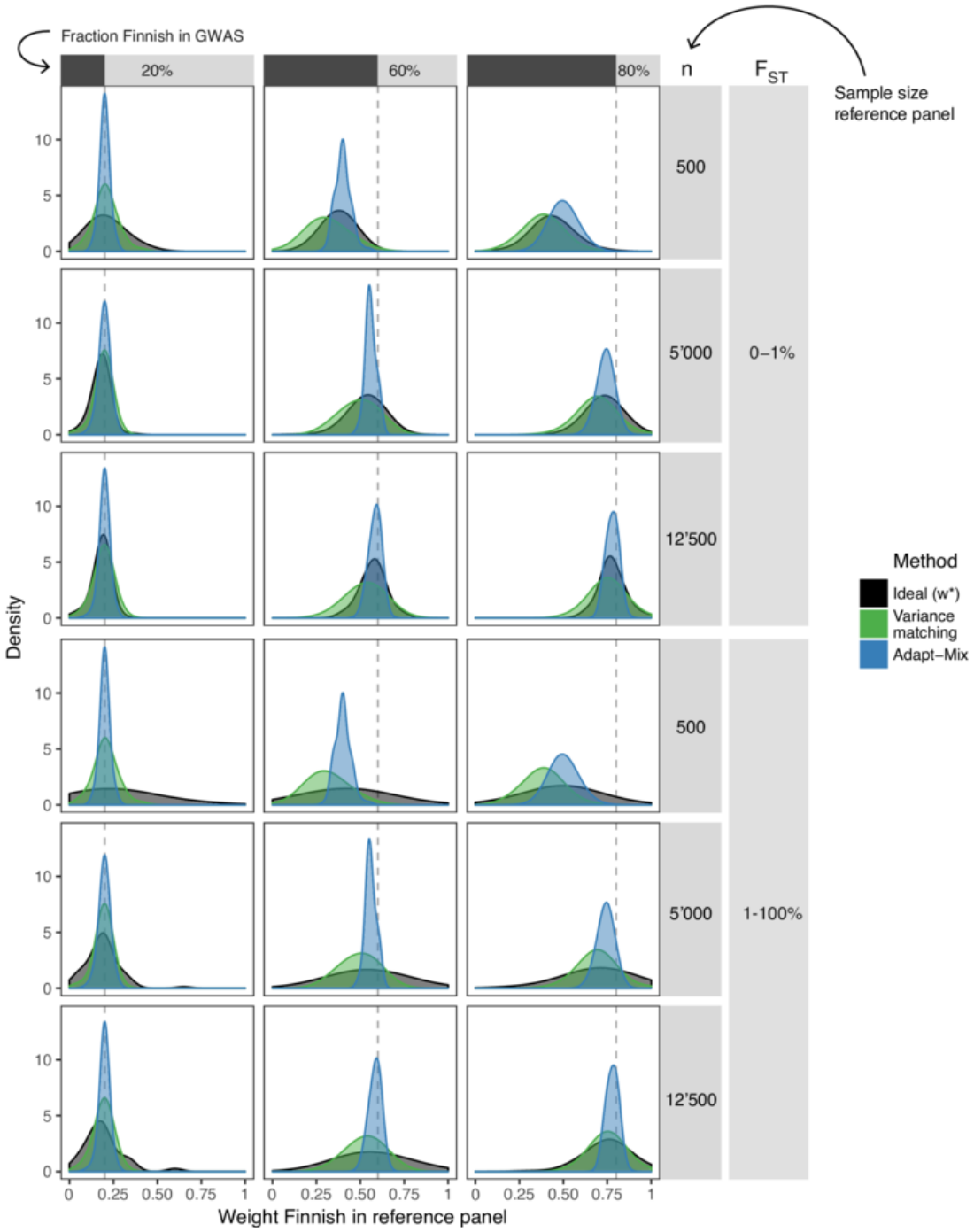

Figure 3: Comparing w determined by different methods. This figure compares the weights chosen by all three optimisation methods: w^* as best possible weight (black), w^{vm} by *variance matching* (green), and w^{am} by *Adapt-Mix* (blue). w^* represents the best possible method if we were to know the true effect size, but given the same reference panel as for *Adapt-Mix* and *variance matching*. The x-axis the weights for the reference panel chosen by each method, and the y-axis shows the density. The results are split into columns and rows, with the rows for different reference panel sizes and F_{ST} groups, and the columns different Finnish fractions in the GWAS populations (also highlighted with the vertical dashed line). Each window contains w^*, w^{vm} and w^{am} for each of the 74 regions.

missingness correlations between variants (missingness-at-random up to maximal possible overlap of missingness) (Fig. 2). We then applied the conventional *summary statistics imputation* approach, and the two alternative methods using the dependent approach ($\boldsymbol{D}^{(dep)}$) and the independent approach ($\boldsymbol{D}^{(ind)}$), respectively. Overall, replacing $\boldsymbol{C}$ and $\boldsymbol{c}$ with $\boldsymbol{D}$ and $\boldsymbol{d}$ yields a lower MSE. Furthermore, we note that the dependent approach has lower MSE when the sample size variance is low and the missingness correlation approaches one. Analysis of cohort level missingness data from the GIANT consortium demonstrates that 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, $N_{k\cap l} = \min(N_k, N_l)$ (Fig. S7). The reason for this is that studies can be grouped according to shared genotyping chips or imputation panels, which induces positive missingness correlation between markers.

Limitations

Due to lack of large sequencing reference panels we used an upsampling technique called HAPGEN2. Thus, the simulation study performed is limited to European subpopulation to chromosome 15 in 1000 Genomes project reference panel. Furthermore, the outcome used for the simulated GWAS is based on one causal variant with a explained variance of 0.02, therefore might not be fully representative for a polygenic phenotype with more than one causal variant. *Summary statistics imputation* relies on fine tuning of parameters, such as shrinkage of the correlation matrix. Any $\lambda > 0$ will turn the correlation matrix into an invertible one. Beyond that, an even stronger shrinkage can compensate for estimation error. We hypothesised that shrinkage depends on local LD structure, and attempted to optimise λ locally in each genomic region, using effect sizes of tag SNPs as training data set in a leave-one-out fashion. However, when looking at null variants only, $\lambda = 1$ will (on average) always lead to the smallest MSE. Therefore, when looking at a region with a mixture of null and associated SNVs, an optimally chosen λ will be shifted towards 1 and might not be an ideal choice for alternative variants (shrinking their estimation towards 0).
In *summary statistics imputation* the imputation quality 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 < 1000 the conventional estimation of imputation quality in Eq. (7) is biased towards overestimation. As an alternative imputation quality measure we calculated the out-of-sample prediction after model selection by fitting a ridge regression model for each unmeasured SNV ($\widehat{r^2}_{\text{ridge}}$). Although the ridge regression algorithm has a fast implementation, the calculation still takes longer than the actual imputation. The alternative measure presented in Eq. (8) offers an alternative that is based on Eq. (7), but takes the effective number of variants and sample size into account. We provide a more detailed analysis in Supplement D.

5 Conclusion

We present two additional approaches to the published *summary statistics imputation* method: allowing the LD structure to be adaptive according to ancestry admixture and varying sample size in summary statistics of tag SNPs. To evaluate these two extensions, we computed GWASs from upsampled 1000 Genomes project data in combination with a simulated phenotype. Our approach to account for ancestry admixture in the GWAS lies in between an already published approach and the best possible solution given the reference panel. Due to the bias-variance trade-off, accounting for ancestry admixture different from the given fraction in the reference panel only becomes relevant with increasing sample size. Our approach to account for varying sample size works better in all presented scenarios, with the dependent sub-approach being

more conservative. *Summary statistics imputation* that accounts for varying sample size is implemented in `https://wp.unil.ch/sgg/summary-statistic-imputation-software/`.

Acknowledgements

Funding

This work was supported by the Leenards Foundation (http://www.leenaards.ch), the Swiss Institute of Bioinformatics (http://www.isb-sib.ch/), the Swiss National Science Foundation [31003A-143914, 31003A-16992c].

References

1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

Bulik-Sullivan, B., Finucane, H., Anttila, V., Gusev, A., Day, F., Loh, P., ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan, L., and Perry, J. (2015). An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*, 47(11):1236–1241.

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, 37(7):658–665.

Eaton, M. L. (1983). *Multivariate Statistics: A Vector Space Approach*. John Wiley & Sons Inc.

Gao, X., Starmer, Ã. J., and Martin, E. R. (2008). A Multiple Testing Correction Method for Genetic Association Studies Using Correlated Single Nucleotide Polymorphisms. *Genetic Epidemiology*, 369:361–369.

Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. *Nature genetics*, 45(11):1274–83.

Kutalik, Z., Johnson, T., Bochud, M., Mooser, V., Vollenweider, P., Waeber, G., Waterworth, D., Beckmann, J. S., and Bergmann, S. (2011). Methods for testing association between uncertain genotypes and quantitative traits. *Biostatistics*, 12(1):1–17.

Lee, D., Bigdeli, T. B., Riley, B. P., Fanous, A. H., and Bacanu, S. A. (2013). DIST: Direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics*, 29(22).

Lee, D., Bigdeli, T. B., Williamson, V. S., Vladimirov, V. I., Riley, P., Fanous, A. H., and Bacanu, S.-a. (2015a). Genome analysis distmix : Direct imputation of summary statistics for unmeasured snps from mixed ethnicity cohorts. *Bioinformatics*.

Lee, D., Williamson, V. S., Bigdeli, T. B., Riley, B. P., Fanous, a. H., Vladimirov, V. I., and Bacanu, S.-A. (2014). JEPEG: a summary statistics based tool for gene-level joint testing of functional variants. *Bioinformatics*, 31(8).

Lee, D., Williamson, V. S., Bigdeli, T. B., Riley, B. P., Webb, B. T., Fanous, A. H., Kendler, K. S., Vladimirov, V. I., and Bacanu, S.-A. (2015b). JEPEGMIX: gene-level joint analysis of functional SNPs in cosmopolitan cohorts: Table 1. *Bioinformatics*, 32.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511.

Morris, A. P. et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, 44(9):981–990.

Park, D. S., Brown, B., Eng, C., Huntsman, S., Hu, D., Torgerson, D. G., Burchard, E. G., and Zaitlen, N. (2015). Adapt-Mix: learning local genetic correlation structure improves summary statistics-based analyses. *Bioinformatics*, 31(12).

Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D. P., Patterson, N., and Price, A. L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30(20).

Rüeger, S., McDaid, A., and Kutalik, Z. (2017). Evaluation and application of summary statistic imputation to discover new height-associated loci.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4.

Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*.

Theil, H. (1961). *Economic Forecasts and Policy*. North-Holland Publishing Co.

Wen, X. and Stephens, M. (2010). Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Annals of Applied Statistics*, 4(3):1158–1182.

Xu, Z., Duan, Q., Yan, S., Chen, W., Li, M., Lange, E., and Yun, L. (2015). Genome Analysis DISSCO : Direct Imputation of Summary Statistics allowing COvariates. *Bioinformatics*.

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. a. F., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N., Loos, R. J., Frayling, T. M., McCarthy, M. I., Hirschhorn, J. N., Goddard, M. E., and Visscher, P. M. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, 44(4):369–375.