Improved imputation of summary statistics for admixed populations

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.*

February 9, 2018

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. This violates the assumption of current methods that an external LD matrix and the covariance of the Z-statistics are identical.
Results: To address this issue, we present an extension to the existing *summary statistics imputation* method that manipulates 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.
Contact: zoltan.kutalik@unil.ch

1 Introduction

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

association summary statistics. *Summary statistics imputation* on the other hand starts off with association summary statistics available for all genotyped markers and infers, combined with a reference panel, directly the association summary statistics of SNVs available in the reference panel. More specifically, estimating the local pair-wise linkage disequilibrium (LD) structure of each genetic region using the reference panel and combining it with association summary statistics allows to calculate a conditional expectation of normally distributed summary statistics. This latter approach is the central focus of our paper.

Although genotype imputation followed by association scan yields more reliable association statistics, it also has its limitations. For example, in large collaborations where individual-level data cannot be shared across studies, genotype imputation requires each group to impute the genotype separately, which may be challenging in terms of manpower and computational resources. Furthermore, this costly process needs to be repeated when the reference panel is updated to improve imputation. Summary statistics imputation on the other hand is much less demanding on computational resources, and therefore much faster, and requires no access to the actual genetic data.

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).

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 statistics imputation*. Similarly, Mendelian randomisation analysis can also be performed based on summary statistics (Burgess et al. 2013), which substantially increases power. To avoid violation of the assumptions (in particular the instrument being exogenous) it is safer to consider possible confounders to test robustness of the causal effect estimation. However, achieving this requires summary statistics to be imputed from a large number of different studies, with potentially different ancestries. To ensure accurate estimation or such admixed meta-analyses, we propose a method that, depending on local genomic structure, optimally combines reference panels to best match the GWAS population. We test this with a simulation framework.

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. The conditional mean of the effect of SNV u can be expressed using the effect size estimates of the tag SNVs (Eaton 1983,

Pasaniuc et al. 2014):

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

where $\mathcal{M}$ is a vector of marker SNVs, $\boldsymbol{\Sigma}_{u\mathcal{M}}$ represents the covariance between SNV u and all $\mathcal{M}$ markers and $\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}$ represents the covariance between all $\mathcal{M}$ markers.
We assume that estimates for the two covariances are available from an external reference panel with n individuals and denote them $\boldsymbol{s} = \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}u}$, $\boldsymbol{S} = \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}$. 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, as derived by Pasaniuc et al. (2014):

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

with $\boldsymbol{z} = \boldsymbol{a}\sqrt{N}$, when the effect size is small (as is the case in typical GWAS).
Similar to Pasaniuc et al. (2014), we chose $\mathcal{M}$ to include all measured variants within at least 250 Kb of SNV u. To speed up the computation when imputing SNVs genome-wide, we apply a windowing strategy, where SNVs within a 1 Mb window are imputed simultaneously using the same set of $\mathcal{M}$ tag SNVs the 1 Mb window plus 250 Kb flanking regions on each side.

Shrinkage of SNV correlation matrix

To estimate $\boldsymbol{C}$ (and $\boldsymbol{c}$) we use an external reference panel of n individuals. Since the size of $\boldsymbol{C}$ often exceeds the number of individuals ($q \gg n$), shrinkage of matrix $\boldsymbol{C}$ is needed to guarantee that it is invertible.
By applying shrinking, the modified matrix $\boldsymbol{C}$ 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. Here, we mainly focus on λ changing with the reference panel size n: $\lambda = 2/\sqrt{n}$ (Lee et al. 2014).

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

For *summary statistics imputation* we would like to estimate the local LD structure of each region in the GWAS population ($\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}$ and $\boldsymbol{\Sigma}_{u\mathcal{M}}$) and to do so we use a (sequenced) reference population, yielding estimates $\boldsymbol{C}$ 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.

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 because it ignores the variance-bias trade-off.

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

First, suppose that the reference panel is made up of P subpopulations of sizes $n^{(1)}, n^{(2)}, \ldots, n^{(P)}$. Next, we introduce a set of weights $\boldsymbol{w} = (n^{(1)}, n^{(2)}, \ldots, n^{(P)})/n = w_1, w_2, \ldots, w_P$, 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 s as a function of these weights ($s(\boldsymbol{w})$), i.e. for each subpopulation we calculate the covariance separately ($s^{(p)}$), and then combine them, weighted by their weights $\boldsymbol{w}$:

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

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

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

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

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

$g_{i,k}$ 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: $\bar{g}_k^p = \frac{1}{n^{(p)}} \sum_{i \in I_p} g_{i,k}$ and I_p refers to the indices of individuals contained in 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

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

Our goal is to quantify the mean squared error (MSE) between the true GWAS LD matrix ($\Sigma_{\mathcal{M}\mathcal{M}}$) and the LD matrix estimated from the reference panel ($\boldsymbol{C}_{\mathcal{M}\mathcal{M}}^{\boldsymbol{w}}$). 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. (7) for SNV k can be

written as

$$\mathrm{MSE}\left[s_{kk}(\boldsymbol{w})\right] = \mathrm{Bias}^2\left[s_{kk}(\boldsymbol{w})\right] + \mathrm{Var}\left[s_{kk}(\boldsymbol{w})\right] \tag{11}$$

In short, the MSE of Eq. (7) 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: $\bar{g}_k^{\mathrm{obs}}$) 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 $\bar{g}_k^p$, $\bar{g}_k^{\mathrm{obs}}$ 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. (11) for a set of $\mathcal{M}$ SNVs, for which we know the allele frequencies $\bar{\boldsymbol{g}}^{\mathrm{obs}}$ in the GWAS population and can estimate $\bar{\boldsymbol{g}}^p$ from a reference panel:

$$\boldsymbol{w}^{\mathrm{vm}} = \arg\min_{\boldsymbol{w}} \sum_{k=1}^{\mathcal{M}} \mathrm{MSE}\left[s_{kk}(\boldsymbol{w})\right] \tag{12}$$

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

Finally, we substitute $\boldsymbol{w}^{\mathrm{vm}}$ into Equation Eq. (10) and plug it into Eq. (6):

$$\widehat{a}_u = a_{u|\mathcal{M}} = \boldsymbol{c}_\lambda^{w'} (\boldsymbol{C}_\lambda^w)^{-1} \boldsymbol{a} \tag{13}$$

A more thorough derivation of Eqs. (7) to (12) can be found in Supplement A.

2.3 Reference panels

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

2.4 Simulation

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 statistics 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 an optimisation of the reference panel composition.

3.1 Simulation framework

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* pheno-

[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

types. 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. 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.
To compose an optimal $\boldsymbol{C}$ and $\boldsymbol{c}$, we apply *variance matching* (vm) that utilises a function of allele frequencies as proxy for population structure. For a set of SNVs, we compare this novel method to an existing approach called *Adapt-Mix* (am) (Park et al. 2015). As a benchmark (*), we use the best possible reference panel composition if we were to know the estimated effect sizes of the untyped SNVs. The weights determined by these three methods are subsequently called $\boldsymbol{w}^{\mathrm{vm}}$, $\boldsymbol{w}^{\mathrm{am}}$ and $\boldsymbol{w}^{*}$.
We use an extended version of our simulation setup. 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 estimated and the imputed effect sizes ($h(\boldsymbol{w})$, Eq. S10) and determine the weight that yields a minimal MSE (denoted as $\boldsymbol{w}^{*}$, Eq. (S11)). In parallel, we apply for each fixed proportion of Finnish in the GWAS the *variance matching* and the *Adapt-Mix* approach to determine their optimal weight — $\boldsymbol{w}^{\mathrm{vm}}$ and $\boldsymbol{w}^{\mathrm{am}}$ — in the reference panel (Fig. S1). To identify other factors that influence the choice of weights, we group SNVs into population specific and population non-specific groups (high and low F_{ST}, respectively), and run the simulation from small to large reference panels.
Ultimately, we are interested in a comparison to the standard reference panel, which has roughly equal weights in the European sub-panel (i.e. $\boldsymbol{w} = 0.2$). In Fig. 1 we compare MSE of the three optimal weights determined by each method to the MSE when using equal weights.

3.2 Improving summary statistics imputation via variance matching

From the extensive simulation results it is clear, that *variance matching* offers performance very close to the best possible reference panel composition $\boldsymbol{w}^{*}$. 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 instances (population non-specific variants when imputed with a reference panel with $5'000$ individuals and a GWAS with 80% Finnish individuals), *variance matching* offers equal or bigger benefit than *Adapt-Mix*. When comparing the median MSE ratio of the optimisation methods to the benchmark, the maximum difference for *variance matching* is 0.021, while it is 0.041 for *Adapt-Mix*. We observe for an equally partitioned GWAS a ratio of $h(\boldsymbol{w}^{*})/h(\boldsymbol{w} = 0.2)$ close to one for all three methods, independent of reference panel size or whether the variants are population specific, since the reference panel is equally partitioned. For a reference panel of 500 individuals, population specific variants and a GWAS with 80% Finnish individuals we observe a theoretically best possible reference panel composition of 0.886, while it is 0.892 for *variance matching* and 0.926 for *Adapt-Mix*. When increasing sample size to $12'500$ the benchmark advantage using $\boldsymbol{w}^{*}$ becomes 0.648 and 0.658 for *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 as expected, 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 it is 0.959 for *variance matching* and 0.966 for *Adapt-Mix*. When increasing sample size to $12'500$ it drops to 0.742, 0.747 and 0.751, for $\boldsymbol{w}^{*}$, *variance matching* and *Adapt-Mix*, respectively.

---

summary statistics to show how *summary statistics imputation* can be utilised to identify novel associations.

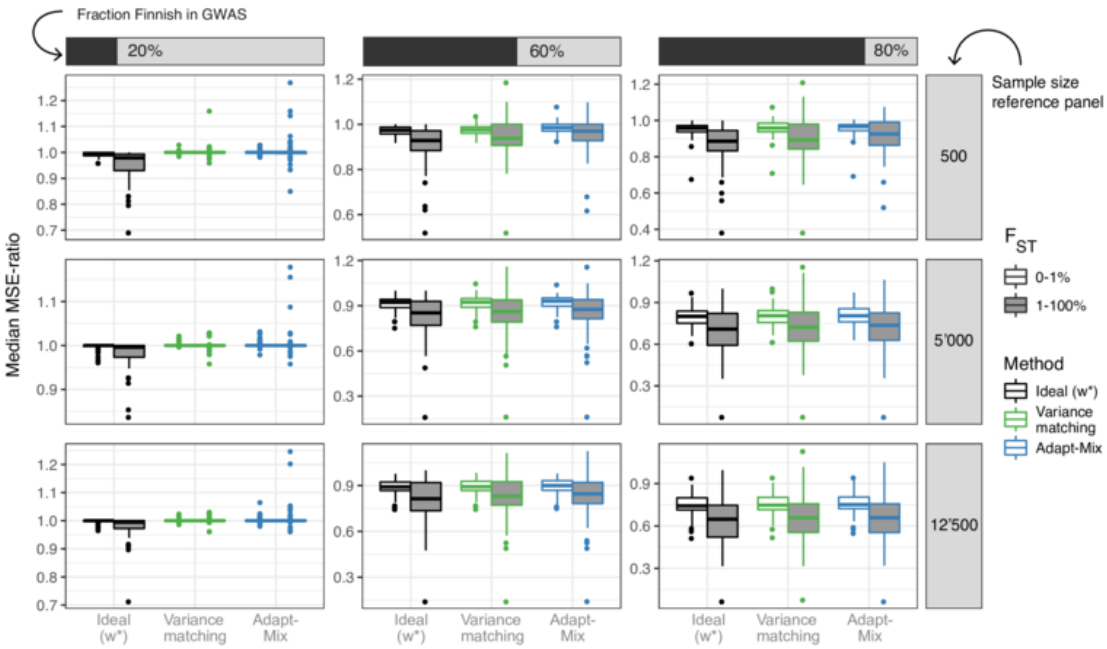

Figure 1: **Comparison of methods accounting for population structure.** This figure shows the comparison of each w-optimisation method with respect to choosing the full European panel of 1000 Genomes project (which corresponds to equal weights). Each boxplots represents 74 simulated regions. The x-axis shows the three different strategies: using theoretical best possible weights in black (if the estimated effect sizes were to be known), *variance matching* in green and *Adapt-Mix* in blue. The y-axis shows the 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 estimated 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. Each row represents different sizes of reference panels, while the different Finnish fractions in the GWAS populations are grouped by column. Variants are also grouped according to F_{ST}, with population specific results having a gray shaded, and population unspecific results having a plain boxplot. The y-axis range varies for each window (for fixed y-axis range see Fig. S3). Table `stable_01.csv` provides the same information as a numeric table.

4 Discussion

Summary statistics are used more and more frequently for downstream analyses but are not always available for all desired variants. These missing summary statistics can, however, be directly imputed from publicly available data using *summary statistics imputation*. The covariance matrices required for this are difficult to estimate from publicly available reference panels due to their size and population structure, requiring their careful adjustment with shrinkage parameters. To address these limitations, we extended the *summary statistics imputation* method as presented in Pasaniuc et al. (2014) with an optimal combination of covariance matrices form reference panel subpopulation.

Choice of reference panel

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

Variance matching and Adapt-Mix

To tackle this problem, Park et al. (2015) proposed *Adapt-Mix* and we propose *variance matching* (Eqs. (12) and (13)). Both methods assume that the GWAS sample is composed of a mixture (or admixture) of populations and that we have a separate reference panel for each population. They then calculate the local LD structure as a linear combination of population-specific estimates, where the weight of each population depends either on the Z-statistics (*Adapt-Mix*) or the allele frequency (*variance matching*). Although the aim is to approximate the true mixing weights in the GWAS sample, the weights returned will usually deviate from that in an attempt to minimise the MSE. For example, given a GWAS performed in an exclusively Finnish population and using the 1000 Genomes project reference panel, we could either use only the 99 Finnish individuals (weight of 1 for the Finnish population, 0 for others), or select all 503 individuals of the European panel (weight of about 0.2 for the Finnish population). Using only Finnish individuals would more closely match the GWAS allele frequencies and reduce bias, however using the full panel would increase the precision of the estimated correlation matrix, reducing variance. Our approach aims to strike a balance between bias and variance by finding an optimal weight, somewhere between 0.2 and 1 in this example. We find that for smaller reference panels ($n = 500$) the optimal weight tends towards lower values, relying more on information from other populations, whereas for larger panels ($n = 12'000$) the optimal weights tend to be closer to the population composition in the GWAS (Fig. 2).
Variance matching performed better than *Adapt-Mix* in all subgroups that we explored, except for one instance where both method performed equally (Fig. 1).

Limitations

Variance matching assumes that the population admixture that is reflected in the variance of tag SNVs (diagonal in matrix $\boldsymbol{C}$) is the same as the covariance between tag SNVs (off-diagonal of $\boldsymbol{C}$) as well as between tag SNVs and SNVs to impute (matrix $\boldsymbol{c}$). Furthermore, we provide

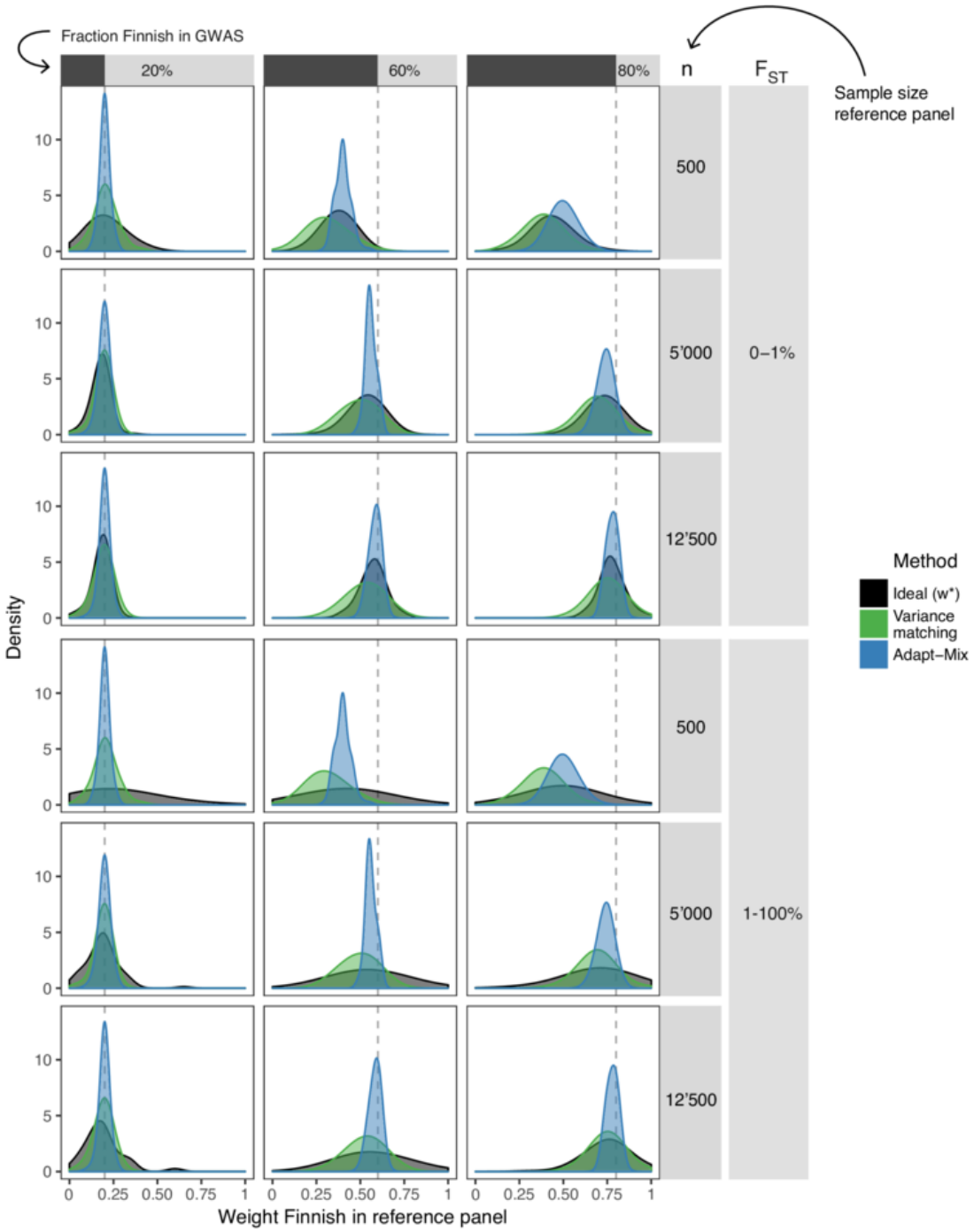

Figure 2: Comparing w determined by different methods. This figure compares the weights chosen by all three optimisation methods: w^* as best possible weight (black), w^{vm} by *variance matching* (green), and w^{am} by *Adapt-Mix* (blue). w^* represents the best possible choice if we were to know the true effect size, but given the same reference panel as for *Adapt-Mix* and *variance matching*. The x-axis displays the weights for the reference panel chosen by each method, and the y-axis shows the density. The results are split into columns and rows, with the rows for different reference panel sizes and 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.

an analytical solution to Eq. (11), which involves approximations of the variance and the bias (Eq. S7 and S8).
Finding a reference panel whose ancestry composition matches that of the GWAS is difficult for three reasons: 1) the mixture/admixture of populations is usually unknown, and 2) publicly available reference panels have 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 sizes > 500 per population are not publicly available at this time.
Due to lack of large, sequenced reference panels we used an upsampling technique called HAPGEN2 (Su et al. 2011), which limits the lower bound of the global allele frequency to $1/(2 \cdot 503)$. 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.

5 Conclusion

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

Acknowledgements

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

Funding

This work was supported by the Leenards Foundation (http://www.leenaards.ch), the Swiss Institute of Bioinformatics (http://www.isb-sib.ch/), and 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.

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.

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.

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

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.