JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 27, Number 3, 2020
© Mary Ann Liebert, Inc.
Pp. 418–428
DOI: 10.1089/cmb.2019.0449

A Unifying Framework for Imputing Summary Statistics in Genome-Wide Association Studies

YUE WU,[1] ELEAZAR ESKIN,[1–3] and SRIRAM SANKARARAMAN[1–3]

ABSTRACT

Methods to impute missing data are routinely used to increase power in genome-wide association studies. There are two broad classes of imputation methods. The first class imputes genotypes at the untyped variants, given those at the typed variants, and then performs a statistical test of association at the imputed variants. The second class, summary statistic imputation (SSI), directly imputes association statistics at the untyped variants, given the association statistics observed at the typed variants. The second class is appealing as it tends to be computationally efficient while only requiring the summary statistics from a study, while the former class requires access to individual-level data that can be difficult to obtain. The statistical properties of these two classes of imputation methods have not been fully understood. In this study, we show that the two classes of imputation methods yield association statistics with similar distributions for sufficiently large sample sizes. Using this relationship, we can understand the effect of the imputation method on power. We show that a commonly used approach to SSI that we term SSI with variance reweighting generally leads to a loss in power. On the contrary, our proposed method for SSI that does not perform variance reweighting fully accounts for imputation uncertainty, while achieving better power.

Keywords: genome-wide association studies, imputation, summary statistics.

1. INTRODUCTION

Genome-wide association studies (GWAS) have been successfully used to discover genetic variants, typically single-nucleotide polymorphisms (SNPs), that affect the trait of interest (Hakonarson et al., 2007; Sladek et al., 2007; Zeggini et al., 2007; Yang et al., 2011; Köttgen et al., 2012; Lu et al., 2013; Ripke et al., 2013). GWAS measure or type the genotypes of individuals at a chosen set of SNPs, and then perform a statistical test of association between a given SNP and the trait of interest. SNPs, at which the null hypothesis of no association between the genotype and the trait can be rejected, are said to be associated with the trait. The threshold that the absolute value of association statistics passes to reject null hypothesis is also referred as significance level.

In a typical GWAS, due to the cost considerations, only a subset of SNPs is genotyped (typed SNPs). Thus, a direct analysis of typed SNPs is likely to have reduced power to detect associations between

Departments of [1]Computer Science, [2]Human Genetics, and [3]Computational Medicine, University of California, Los Angeles, Los Angeles.

untyped SNPs and the trait. Imputation methods, which aim to fill in ''data'' at untyped SNPs, are commonly used to increase the power of GWAS. These methods all rely on the correlation or linkage disequilibrium (LD; Pritchard and Przeworski, 2001; Reich et al., 2001) between genotypes at untyped SNPs and those at typed SNPs (Browning and Browning, 2007; Marchini et al., 2007; Howie et al., 2009, 2012; Li et al., 2009, 2010; Marchini and Howie, 2010). Initial work on imputation focused on the problem of genotype imputation, that is, inferring the genotypes at untyped SNPs given the genotypes at typed SNPs. Genotype imputation methods rely on a reference panel, in which individuals are typed at all SNPs of interest, to learn the LD patterns across SNPs. Given a target data set in which genotypes are typed at a subset of the SNPs, these methods rely on the LD patterns learned from the reference panel to infer the genotypes at the remaining untyped SNPs.

In the context of GWAS, there are two broad classes of imputation methods to estimate the association statistics at untyped SNPs. The first class relies on genotype imputation to infer the genotypes at the untyped SNPs followed by computing association statistics at the imputed genotypes (Browning and Browning, 2007; Marchini et al., 2007; Howie et al., 2009, 2012; Li et al., 2009, 2010). We refer to this class of imputation methods as the two-step imputation methods. In practice, the most successful methods for the first step of genotype imputation are based on discrete hidden Markov models (HMMs; Browning and Browning, 2007; Marchini et al., 2007). The second class of methods directly imputes the association statistics at the untyped SNPs, given the association statistics at the typed SNPs. As shown in previous work (Han et al., 2009; Kostem et al., 2011), the joint distribution of marginal statistics at the typed SNPs and untyped SNPs follows a multivariate normal distribution (MVN; Han et al., 2009; Kostem et al., 2011; Hormozdiari et al., 2014, 2015, 2016). This class of methods utilizes the correlation between the association statistics induced by their dependence on the underlying genotypes (Lee et al., 2013; Pasaniuc et al., 2014). This class of methods is termed summary statistic imputation (SSI). SSI is appealing as it tends to be computationally efficient while only requiring the summary statistics from a study, while the first class requires access to individual-level data, which can be difficult to obtain in practice.

Current summary statistic-based imputation methods calibrate the imputed statistics using a technique we call *variance reweighting* (SSI-VR). Despite recent progress, the statistical properties of SSI methods (including the impact of variance reweighting) and the connection between the two classes of SSI methods have not been adequately understood.

In this study, we characterize the asymptotic distribution of the association statistics under each of the two classes of imputation methods, the two-step imputation and SSI. The resulting statistics are asymptotically multivariate normal with differences in the underlying covariance matrix that depend on the details of the HMM used for genotype imputation. Using this characterization, we can understand the effect of the imputation method on power. Our new method, SSI, performs SSI without variance reweighting. The resulting statistics do not then have unit variance as in traditional SSI, but instead correctly take into account the ambiguity of the imputation process. We compared the performance of the imputation methods on the Northern Finland Birth Cohort (NFBC) data set (Sabatti et al., 2009) to show that SSI increases power over no imputation, while SSI-VR can sometimes lead to lower power. Finally, we ran SSI, SSI-VR, and two-step imputation on the NFBC data set and show that the resulting statistics are close, thereby justifying the theory.

2. METHODS

2.1. Summary statistics

Under the null hypothesis, the joint distribution of the association statistics of the U untagged SNP s_U and the O tag SNPs s_O follows an MVN:

$$\begin{bmatrix} s_U \\ s_O \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\lambda}_U \\ \boldsymbol{\lambda}_O \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_U & \boldsymbol{\Sigma}_{UO} \\ \boldsymbol{\Sigma}_{UO}^T & \boldsymbol{\Sigma}_O \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_U & \boldsymbol{\Sigma}_{UO} \\ \boldsymbol{\Sigma}_{UO}^T & \boldsymbol{\Sigma}_O \end{bmatrix}\right) \tag{1}$$

Since none of the $M = (U + O)$ SNPs is associated, the noncentrality parameters (NCPs) of both $\boldsymbol{\lambda}_U$ and $\boldsymbol{\lambda}_O$ are $\mathbf{0}$. Furthermore, the statistics are standardized so that the diagonal elements of the covariance matrix are 1, that is, $\Sigma_{U_{i,i}} = \Sigma_{O_{j,j}} = 1$.

2.1.1. Summary statistic imputation. Under the null assumption where $\boldsymbol{s}_O$ and $\boldsymbol{s}_U$ are not associated, $\boldsymbol{\lambda}_U$ and $\boldsymbol{\lambda}_O$ are each $\mathbf{0}$. Using the joint distribution, we can compute the distribution of the true statistics at the untagged SNPs, $\boldsymbol{s}_U$ conditioned on the statistics observed at the tag SNP, $\boldsymbol{s}_O$. The conditional distribution follows an MVN, which is computed as follows:

$$P(\boldsymbol{s}_U|\boldsymbol{s}_O) \sim \mathcal{N}\left(\Sigma_{UO}\Sigma_O^{-1}\boldsymbol{s}_O, \Sigma_U - \Sigma_{UO}{\Sigma_O}^{-1}\Sigma_{OU}\right) \tag{2}$$

The observed statistics are denoted $\hat{\boldsymbol{s}}_O$. Thus, $\boldsymbol{s}_U$ is imputed using a function of observed statistics:

$$\hat{\boldsymbol{s}}_U(\hat{\boldsymbol{s}}_O) = \Sigma_{UO}\Sigma_O^{-1}\hat{\boldsymbol{s}}_O \tag{3}$$

Let $\boldsymbol{A} = \Sigma_{UO}{\Sigma_O}^{-1}$ and thus $\hat{\boldsymbol{s}}_U(\hat{\boldsymbol{s}}_O) = \boldsymbol{A}\hat{\boldsymbol{s}}_O$.

2.1.2. SSI with variance reweighting. From the previous result, we have $\hat{\boldsymbol{s}}_U(\hat{\boldsymbol{s}}_O) = \boldsymbol{A}\hat{\boldsymbol{s}}_O$. Notice that the underlying joint distribution over the test statistics assumes that each of the statistics at the observed as well as unobserved SNPs has variance one. On the contrary, Equation 3 shows that the variance of the imputed statistic is <1. Variance reweighting proposes standardizing the statistics at the untagged SNPs.

Let s_i be the statistic at the ith untagged SNP. Thus, instead of imputing s_i using $\hat{s}_i$, we impute using $\hat{z}_i = \frac{\hat{s}_i}{\sqrt{var(\hat{s}_i)}}$, so that all the imputed $\hat{z}_i$ have variance equal to 1. We have $var(\hat{s}_i) = \mathbb{E}[\Sigma_{U_i,O}{\Sigma_{O,O}}^{-1}$ $\hat{s}_O{\hat{s}_O}^T{\Sigma_O}^{-1}\Sigma_{OU_i}] = \Sigma_{U_i,O}{\Sigma_O}^{-1}\Sigma_{O,U_i}$. Thus we have

$$\hat{z}_i(\hat{\boldsymbol{s}}_O) = \frac{\Sigma_{UO}{\Sigma_O}^{-1}\hat{\boldsymbol{s}}_O}{\sqrt{\Sigma_{U_i,O}{\Sigma_O}^{-1}\Sigma_{O,U_i}}} \tag{4}$$

2.2. The impact of imputation on the rejection boundary

SSI uses the following function to impute statistics at the unobserved statistics: $\hat{\boldsymbol{s}}_U(\hat{\boldsymbol{s}}_O) = \boldsymbol{A}\hat{\boldsymbol{s}}_O$. Let $\boldsymbol{A}_i$ be the ith row of matrix $\boldsymbol{A}$, $\boldsymbol{A}_i = {\Sigma_{U_iO}}^T{\Sigma_O}^{-1}$, where $\Sigma_{U_iO}^T$ is the correlation vector between untagged variant snp_i and all the observed SNPs. We choose thresholds t for rejecting statistics at each of the observed and imputed SNP, that is, we reject the null hypothesis at observed SNP O_j if $|\hat{s}_{O_j}| > t$, while we reject the null hypothesis at unobserved SNP U_i if $|\hat{s}_{U_i}| > t$, where t is chosen to control the family-wise error rate (FWER). We would like to understand the conditions the threshold t for SSI relative to the threshold t when no imputation was performed, that is, we want to provide conditions when imputation changes the rejection boundary.

Theorem 1. *The imputed statistic at* snp_i *computed using SSI will change the rejection boundary iff the sum of the absolute values of all the entries of* $\boldsymbol{A}_i$, $\sum_j |A_{ij}| > 1$.

Proof. See Section S2 in Supplementary Material. ■

In SSI-VR, instead of using $\hat{s}_i$ as the imputed statistic for variant i, we use

$$\hat{z}_i = \frac{\hat{s}_i}{\sqrt{var(\hat{s}_i)}} = \frac{\sum\limits_j A_{ij}\hat{s}_{O_j}}{\sqrt{\sum\limits_j A_{ij}^2 + 2\sum\limits_{j \neq k} A_{ij}A_{ik}\Sigma_{O_j,O_k}}} \tag{5}$$

In SSI-VR, untagged variant i will effect the rejection boundary iff $\frac{\sum_j |A_{ij}|}{\sqrt{\sum_j A_{ij}^2 + 2\sum_{j \neq k} A_{ij}A_{jk}\Sigma_{O_j,O_k}}} > 1$.

2.3. Two-step imputation

The two-step approach to SSI first performs genotype imputation followed by testing for association using the imputed genotypes. Genotype imputation fills in the genotypes at the unobserved SNPs $\boldsymbol{G}_U$, given the genotypes at observed SNPs $\boldsymbol{G}_O$ (Marchini and Howie, 2010). Typically, this involves defining a probability distribution for the missing genotypes, given the observed genotypes $P(\boldsymbol{G}_u|\boldsymbol{G}_O)$. Let $p_i(\boldsymbol{g}) = P(\boldsymbol{G}_{U_i} = \boldsymbol{g}|\boldsymbol{G}_O)$ denote the posterior probability at unobserved SNP i. Given a vector $\boldsymbol{g}$ of N genotypes at an SNP, let the association statistic $s(\boldsymbol{g})$ be a function of the genotypes $\boldsymbol{g}$. We can then compute the association statistic at unobserved SNP i as the posterior mean of the association statistic: $\mathbb{E}[s(\boldsymbol{G}_{U_i})|\boldsymbol{G}_O] = \sum_g s(\boldsymbol{g})p_i(\boldsymbol{g})$.

In practice, instead of the posterior mean, association statistics are restricted to imputed SNPs, at which the imputation is confident (e.g., using the INFO score reported by software such as IMPUTE2; Marchini et al., 2007) followed by using the maximum *a posteriori* estimate of the genotype at each SNP. We focus on the posterior mean as it accounts for the uncertainty in imputation and is easier to analyze. We first consider a simple genotype imputation strategy that uses the pairwise correlation among SNPs in an MVN (Wen and Stephens, 2010; Section 2.3.1). In Section 2.3.2, we consider the use of HMMs for genotype imputation.

2.3.1. Genotype imputation using MVN. First, we consider an MVN with mean zero and covariance matrix given by the LD matrix to model the distribution of the genotype vector at the observed and unobserved SNPs for each individual (Wen and Stephens, 2010). We can then impute the genotypes for missing SNPs $\hat{\boldsymbol{G}}_U$ as a function of observed genotypes $\boldsymbol{G}_O$ using the conditional mean for the MVN (Eq. 2). Denoting the $N \times O$ matrix of standardized genotypes as $\boldsymbol{X}_O$ and the imputed genotype vector across N individuals at unobserved SNP i as $\hat{x}_{U_i}$, we have the following:

$$\hat{\boldsymbol{x}}_{U_i}(\boldsymbol{X}_O) = (\Sigma_{U_iO}\Sigma_O{}^{-1}\boldsymbol{X}_O{}^{\mathrm{T}})^{\mathrm{T}} = \boldsymbol{X}_O\Sigma_O{}^{-1}\Sigma_{OU_i} \tag{6}$$

where Σ_{U_iO} is the i^{th} row of matrix Σ_{UO}.

Given a vector of continuous phenotypes $\boldsymbol{y} \in \mathbb{R}^N$ measured across N individuals, the effect size $\hat{\beta}_j$ for observed SNP j can be estimated by a linear regression of $\boldsymbol{y}$ on the genotypes at SNP j: $\hat{\beta}_j = \frac{\boldsymbol{x}_{O_j}{}^{\mathrm{T}}y}{N}$ so that the association statistic s_j at this SNP j: $\hat{s}_j = \frac{\hat{\beta}_j}{\sqrt{var(\hat{\beta}_j)}} = \frac{\boldsymbol{x}_{O_j}{}^{\mathrm{T}}y}{\sigma\sqrt{N}}$. Here σ denotes the standard deviation of the phenotype. Analogously, the association statistic $\hat{s}_i$ at unobserved SNP i is $\hat{s}_i = \frac{\hat{\boldsymbol{x}}_{U_i}^{\mathrm{T}}\boldsymbol{y}}{\sqrt{var(\hat{\boldsymbol{x}}_{U_i}^{\mathrm{T}}\boldsymbol{y})}}$. From Equation 6, we have the following:

$$\hat{s}_i = \frac{\Sigma_{U_iO}\Sigma_O{}^{-1}\boldsymbol{X}_O{}^{\mathrm{T}}\boldsymbol{y}}{\sigma\sqrt{\Sigma_{U_iO}\Sigma_O{}^{-1}\boldsymbol{X}_O^{\mathrm{T}}\boldsymbol{X}_O\Sigma_O{}^{-1}\Sigma_{OU_i}}} = \frac{\Sigma_{U_iO}\Sigma_O{}^{-1}\boldsymbol{s}_O}{\sqrt{\Sigma_{U_iO}\Sigma_O{}^{-1}\Sigma_{OU_i}}} \tag{7}$$

Here we used $\frac{\boldsymbol{X}_O^{\mathrm{T}}\boldsymbol{X}_O}{N} = \Sigma_O$.

This function is identical to SSI-VR as seen in Equation 5. Thus, applying the imputation function in Equation 6 to directly impute genotypes is equivalent to SSI-VR.

2.3.2. Genotype imputation using HMMs. We consider the use of an HMM for genotype imputation. These models assume that a reference panel $\boldsymbol{M}$ is available that contains genotype data across $M = (U + O)$ SNPs (Scheet and Stephens, 2006; Marchini et al., 2007; Browning and Browning, 2007; Li et al., 2010). The HMM models the conditional distribution of each of the pair of haplotypes ($\boldsymbol{h}_n^{(1)}$, $\boldsymbol{h}_n^{(2)}$) in each of the N individuals in the study at the O observed and U unobserved SNPs by the conditional distribution $P(\boldsymbol{h}|\boldsymbol{M})$. Specifically, $h_n^{(a)} \sim^{iid} P(\boldsymbol{h}|\boldsymbol{M})$ for $n \in \{1, \ldots, N\}$, $h_n^{(a)} \in \{0, 1\}^M$ $a \in \{1, 2\}$.

The effect size estimate for SNP j: $\hat{\beta}_j = \frac{cov(\boldsymbol{h}_j, \boldsymbol{y})}{var(\boldsymbol{h}_j)}$ and the association statistic $s_j = \frac{cov(\boldsymbol{h}_j, \boldsymbol{y})}{\sigma\sqrt{var(\boldsymbol{h}_j)}}$.

We show in Section S1 of the Supplementary Material that the vector of association statistics asymptotically follows an MVN:

$$s \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \Sigma_S) \tag{8}$$

The asymptotic covariance matrix of the association statistics Σ_S depends on the specific HMM used. Under the commonly used Li–Stephens model (Li and Stephens, 2003), this covariance matrix is as follows:

$$\Sigma_{S,ij} = \begin{cases} (1-\theta)^2 + \frac{\theta}{2}\left(1 - \frac{\theta}{2}\right)\frac{1}{\sigma_i^2}, i = j \\ \qquad\qquad\qquad\qquad\quad , i \neq j \\ \exp\left(-\frac{\rho_{ij}}{2N}\right)\Sigma_{ij} \end{cases} \tag{9}$$

Here Σ_{ij} is the LD or the correlation between SNPs i and j, θ is a parameter related to the mutation rate, and ρ_{ij} is an estimate of the population-scaled recombination rate between SNPs i and j. Thus, the association statistic computed using genotypes imputed using an HMM follows an MVN with mean zero and covariance matrix equal to an LD matrix with shrinkage applied according to the recombination rate between SNPs.

3. RESULTS

3.1. Overview of summary statistics

Assume we have a total of $M = (U + O)$ SNPs that are partitioned into O observed (or tag) SNPs $\{snp_1,\ snp_2,\ snp_3 \ldots snp_O\}$ and U missing SNPs $\{snp_1,\ snp_2,\ snp_3,\ \ldots snp_U\}$ for N individuals. For the O tag SNPs, let $\boldsymbol{s}_O$ be a vector of association statistics of length O, $\boldsymbol{\lambda}_O$ be a vector of NCPs of length O, and let Σ_O be a $O \times O$ matrix of their pairwise correlation coefficients. For the U missing SNPs, let $\boldsymbol{s}_U$ be a vector of association statistics of length U, $\boldsymbol{\lambda}_U$ be a vector of NCPs also of length U, and let $\boldsymbol{\Sigma}_U$ be a $U \times U$ matrix of their pairwise correlation coefficients.

Let Σ_{UO} be a $U \times O$ matrix of the pairwise correlation, that is, LD, between missing SNPs and observed SNPs. Thus, we have an $M \times M$ LD matrix, Σ_{LD}. We can partition the LD matrix as follows: $\Sigma_{LD} = \begin{bmatrix} \Sigma_U & \Sigma_{UO} \\ \Sigma_{OU} & \Sigma_O \end{bmatrix}$. For large sample sizes, the association statistics follow an MVN,

$$\begin{bmatrix} \boldsymbol{s}_U \\ \boldsymbol{s}_O \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\lambda}_U \\ \boldsymbol{\lambda}_O \end{bmatrix}, \begin{bmatrix} \Sigma_U & \Sigma_{UO} \\ \Sigma_{OU} & \Sigma_O \end{bmatrix} \right) \tag{10}$$

Under the null where we assume that none of the SNPs is causal, $\boldsymbol{\lambda}_U$ and $\boldsymbol{\lambda}_O$ are equal to 0.

3.2. Example

We consider a simple example to illustrate how imputation affects the rejection threshold at a given set of SNPs. We consider three SNPs: snp_1, snp_2, and snp_3. In this example, $snp_1,\ snp_2$ are observed, and snp_3 is imputed. We assume the statistics of the tag SNPs ($snp_1,\ snp_2$), $\begin{bmatrix} s_1 \\ s_2 \end{bmatrix}$ follows $\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$ where $|\rho| \leq 1$ and we use $\pi(s_1,\ s_2)$ to denote this distribution. We also assume that the statistics of the tag SNPs $snp_1,\ snp_2$ and the unobserved SNP snp_3 jointly follow the distribution $\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho & \alpha \\ \rho & 1 & \alpha \\ \alpha & \alpha & 1 \end{bmatrix} \right)$ where $|\rho| \leq 1$, $|\alpha| \leq 1$.

Thus, having the joint distribution of the statistics s_1, s_2, and s_3, we can compute the conditional distribution of the untyped SNP conditioned on the marginal statistics of the typed SNPs s_1 and s_2:

$$P(s_3 | s_1,\ s_2) \sim \mathcal{N}\left(\begin{bmatrix} \alpha \\ \alpha \end{bmatrix}^T \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}, 1 - \begin{bmatrix} \alpha \\ \alpha \end{bmatrix}^T \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} \alpha \\ \alpha \end{bmatrix} \right)$$

Typically, SSI uses the posterior mean of the statistic s_3, given the observed values of $\hat{s}_1$ and $\hat{s}_2$ to estimate s_3. In our example, this leads to the statistic s_3 for snp_3 being imputed as a function of $\hat{s}_1,\ \hat{s}_2$:

$$\hat{s}_3(\hat{s}_1,\ \hat{s}_2) = \frac{\alpha}{1 + \rho}(\hat{s}_1 + \hat{s}_2)$$

We choose thresholds t for rejecting each of the statistics ($\hat{s}_1,\ \hat{s}_2,\ \hat{s}_3$) such that the FWER, that is, the probability of at least one false positive, is controlled at a level 0.05. For each tested SNP, we choose the threshold to be the same.

In the case where no imputation is performed, we only test two SNPs. We use the same threshold t for SNPs snp_1 and snp_2. Figure 1a shows the rejection boundary (the blue box) for two SNPs with correlation $\rho = 0.36$ where the region outside this box corresponds to the rejection region. Given the joint density $\pi(s_1,\ s_2)$ of the association statistics $(s_1,\ s_2)$, we determined the rejection boundary by computing the length of the side of the blue box such that the cumulative density in the rejection area, that is, the area under the density $\pi(s_1,\ s_2)$ outside the box is equal to 0.05. Mathematically, we need to find t such that $FWER(t) = 0.05$ where:

$$FWER(t) \equiv 1 - \int \pi(s_1,\ s_2)\mathbf{1}\{s_1 \in -[t, t]\}\mathbf{1}\{s_2 \in [-t, t]\}ds_1 ds_2$$

Here $\mathbf{1}\{s_1 \in -[t, t]\}\mathbf{1}\ \{s_2 \in [-t, t]\}$ defines the acceptance region, that is, the set of points $(s_1,\ s_2) \in \mathbb{R}^2$ where the null hypothesis at both SNPs is accepted.

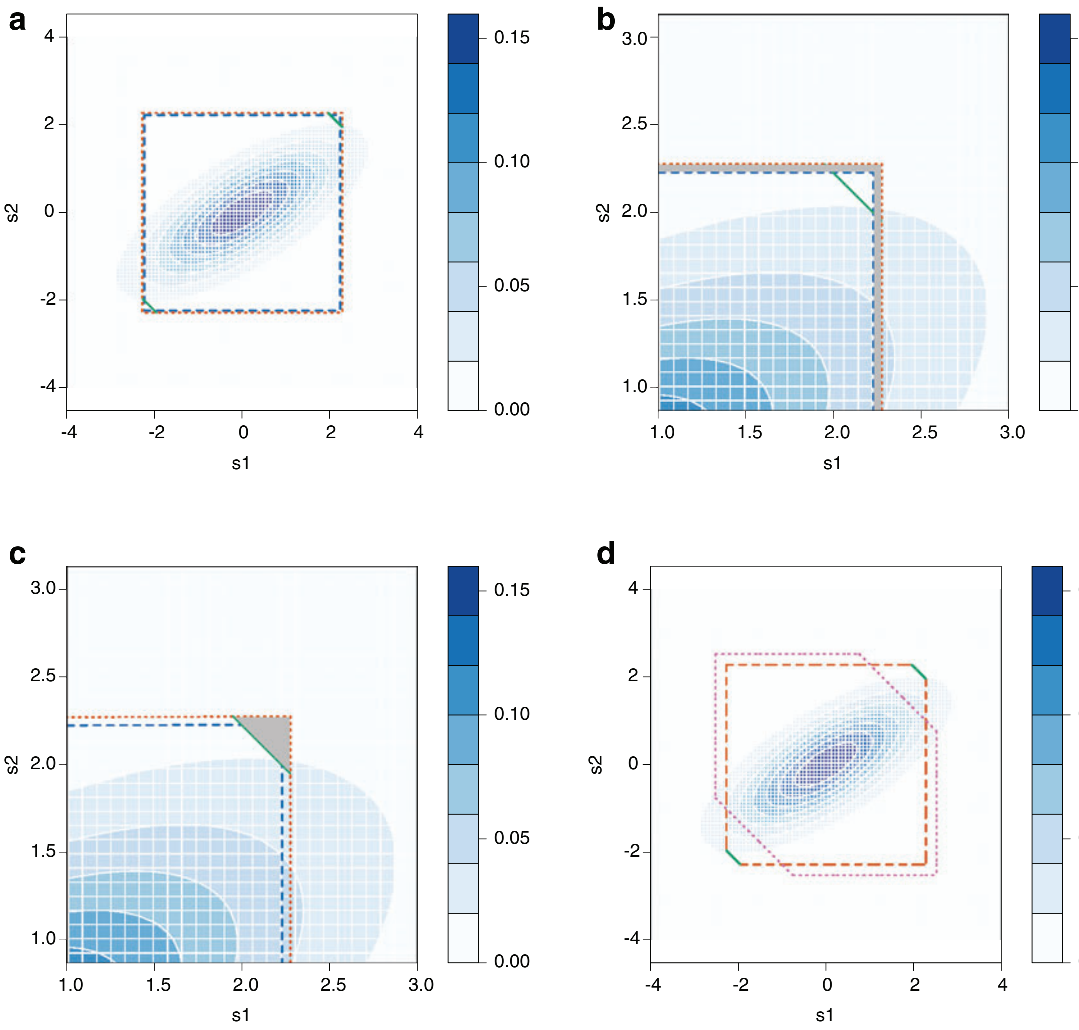

FIG. 1. The effect of imputation on the rejection boundary. This figure shows rejection boundary with no imputation, with imputation (SSI), and variance reweighted imputation (SSI-VR) for an example containing two observed SNPs snp_1, snp_2 and an unobserved SNP snp_3. The contours represent the probability density of the statistics for the observed SNPs: s_1 and s_2 projected in the plane. **(a)** The blue box is the rejection boundary with FWER 0.05 for snp_1 and snp_2 before imputation. The polygon with red- and green-colored boundaries is the rejection boundary after imputation. **(b, c)** A zoomed in version of **(a)** to show the rejection boundary changes. **(b)** The power change on two observed SNPs. **(c)** The power change on the imputed SNP and has three points corresponding to different scenarios. **(d)** The rejection boundary of imputation with SSI-VR in pink color in addition to the rejection boundary of imputation (SSI) seen in **(a)**. We observe that the variance reduction technique leads to power gain on imputed SNP while causing power loss on observed SNPs using SSI-VR. FWER, family-wise error rate; SNPs, single-nucleotide polymorphisms; SSI, summary statistic imputation.

We now consider the effect of testing imputed SNPs in addition to the tag SNPs. The rejection regions for snp_1, snp_2, snp_3 are the regions outside the intervals $R_1 = [-t, t]$, $R_2 = [-t, t]$, $R_3 = [-t, t]$, respectively. We can compute the FWER for a given t by determining the probability mass outside the rejection region. To do this, we note that the joint sampling distribution of $(s_1, s_2, \hat{s}_3)$ is determined only by the distribution of (s_1, s_2) since $\hat{s}_3$ is a deterministic function of s_1 and s_2.

$$
\begin{aligned}
FWER(t) &\equiv 1 - \int \pi(s_1, s_2)\mathbf{1}\,\{s_1 \in -[-t, t]\}\mathbf{1}\,\{s_2 \in [-t, t]\}\ \mathbf{1}\,\{s_3 \in [-t, t]\}ds_1ds_2ds_3 \\
&= 1 - \int \pi(s_1, s_2)\mathbf{1}\,\{s_1 \in [-t, t]\}\mathbf{1}\,\{s_2 \in [-t, t]\}\ \mathbf{1}\left\{\frac{\alpha}{1+\rho}(s_1 + s_2) \in [-t, t]\right\}ds_1ds_2
\end{aligned}
$$

Notice that, in the setting with imputation, the acceptance region $\mathbf{1}\{s_1 \in [-t, t]\}\ \mathbf{1}\{s_2 \in [-t, t]\}$ $\mathbf{1}\left\{\frac{\alpha}{1+\rho}(s_1+s_2) \in [-t, t]\right\}$ can never increase relative to the setting where only the tag SNPs are tested. Now consider the case where the null hypothesis at both the observed SNPs is accepted. This happens when $|\hat{s}_1| \leq t$ and $|\hat{s}_2| \leq t$. Then the statistic at the imputed SNP is as follows:

$$\begin{aligned}|\hat{s}_3(\hat{s}_1, \hat{s}_2)| &= |\frac{\alpha}{1+\rho}(\hat{s}_1+\hat{s}_2)| \\ &\leq |\frac{\alpha}{1+\rho}|(|\hat{s}_1|+|\hat{s}_2|) \quad \text{(triangle inequality)} \\ &\leq 2|\frac{\alpha}{1+\rho}|t\end{aligned}$$

Thus, if $2|\frac{\alpha}{1+\rho}| \leq 1$, then we have $|\hat{s}_3(\hat{s}_1+\hat{s}_2)| \leq t$. Thus, the imputed SNP will never be rejected when neither of the observed SNPs is rejected. Thus, the acceptance region remains the same as the setting when only the tag SNPs are tested. In other words, imputation does not change the rejection boundary.

On the contrary, when $\frac{\alpha}{1+\rho} > \frac{1}{2}$, then imputation will change the rejection region. Figure 1 shows the effect of imputation with $\alpha = 0.80$ and $\rho = 0.36$ so that $\hat{s}_3(\hat{s}_1, \hat{s}_2) = 0.5882(\hat{s}_1+\hat{s}_2)$. The rejection boundary of the observed SNPs snp_1 and snp_2 after imputation is shown by the red lines. The rejection region for snp_3 corresponds to the region where $|0.5882(s_1+s_2)| > t$, which corresponds to the green line. Thus, the cumulative density outside the polygon of red and green lines is the same as the rejection area outside the blue box. In Figure 1b, the shaded area indicates the power loss on the observed SNPs, and in Figure 1c, the shaded area is the power gained from imputation.

Thus assume we have three points, $p1$, $p2$, and $p3$ in Figure 1c, which are three different pairs of association statistics of observed SNPs $snp1$ and $snp2$. The first point is in both the blue rectangle and the polygon, which means we will accept null with or without imputation. The second point $p2$ is the case that without imputation we will reject null, and after imputation we will accept null because of the change of boundary on observed SNPs. The third point $p3$ is the special case. In this case, the observed SNP does not have a significant association because it lies inside the blue box, but after imputation, the imputed SNP has a significant association since it lies outside the polygon and thus we reject the null.

3.3. Simulation results

As shown in previous work on summary statistics (Lee et al., 2013), the marginal statistics at typed SNPs and untyped SNPs follow an MVN. With the assumption that none of the SNPs is significantly associated with train, the mean of the MVN is 0.

As in the previous simple case having three SNPs, snp_1, snp_2, and snp_3, under the null hypothesis of no association, the summary statistics follow the distribution $\mathcal{N}\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho & \alpha\\ \rho & 1 & \alpha\\ \alpha & \alpha & 1\end{bmatrix}\right)$.

Thus having the joint distribution of the statistics s_1, s_2, and s_3, we can compute the conditional distribution of the untyped SNP conditioned on the marginal statistics of the typed SNPs s_1 and s_2:

$$P(s_3|s_1,\ s_2) \sim \mathcal{N}\left(\begin{bmatrix}\alpha\\ \alpha\end{bmatrix}^T \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}^{-1}\begin{bmatrix}s_1\\ s_2\end{bmatrix}, 1-\begin{bmatrix}\alpha\\ \alpha\end{bmatrix}^T\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}^{-1}\begin{bmatrix}\alpha\\ \alpha\end{bmatrix}\right) \quad (11)$$

SSI estimates s_3 using the mean of the above distribution $\hat{s}_3$. The variance of the imputed statistic: $var(\hat{s}_3) = \begin{bmatrix}\alpha\\ \alpha\end{bmatrix}^T\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}^{-1}\begin{bmatrix}\alpha\\ \alpha\end{bmatrix}$ is smaller than 1 (since Eq. 11 shows that the variance of $s_3|s_1,\ s_2$ is $1-\begin{bmatrix}\alpha\\ \alpha\end{bmatrix}^T\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}^{-1}\begin{bmatrix}\alpha\\ \alpha\end{bmatrix}$ and the variance is non-negative). Thus, in most summary statistic imputations (Lee et al., 2013; Pasaniuc et al., 2014), snp_3 is imputed as $\hat{z}_3 = \frac{\hat{s}_3}{\sqrt{var(\hat{s}_3)}}$ so that all the association statistics have variance 1. Since the variance of $\hat{s}_3$ is ≤ 1, the new statistic $|\hat{z}_3| \geq |\hat{s}_3|$. As a result, for a given threshold, the acceptance region in SSI-VR is never greater than with SSI. In other words, to achieve a given FWER, the threshold t needs to be larger for SSI-VR than without, as shown in Figure 1d.

Now having snp_3 imputed using summary statistics, we want to find out how power is affected by SSI and SSI-VR. In Section S3 of the Supplementary Material, we analytically compute the average marginal power function for both methods. To assess power, we assume that three SNPs, $snp_1,\ snp_2$, and snp_3, are drawn from a region associated with a trait. We assume that the untagged variant, snp_3, is causal with NCP so that $(s_1,\ s_2,\ s_3)$ follow a nonzero mean MVN: $\mathcal{N}\left(\begin{bmatrix} 2.31\alpha \\ 2.31\alpha \\ 2.31 \end{bmatrix}, \begin{bmatrix} 1 & \rho & \alpha \\ \rho & 1 & \alpha \\ \alpha & \alpha & 1 \end{bmatrix}\right)$. We choose the NCP to be 2.31 so that the maximum power of no imputation will be around 0.5, which will happen when both α and ρ are 1. We let the correlation between untagged and tag SNPs α and the correlation between tag SNPs ρ vary across: $[0.1, 0.2, \ldots, 0.9, 1]$.

For each combination of $[\alpha, \rho]$, we determined a set of three thresholds (1) for no imputation, (2) for imputation, and (3) imputation with variance correction. We drew 10^8 samples from each distribution, and the power is defined as the probability that we reject the null hypothesis based on thresholds for each method.

In all the combinations except the cases that the LD matrix is no longer positive definite, we find the power of no imputation, SSI, and SSI-VR (Fig. 2). In Figure 2a, we compared SSI versus no imputation,

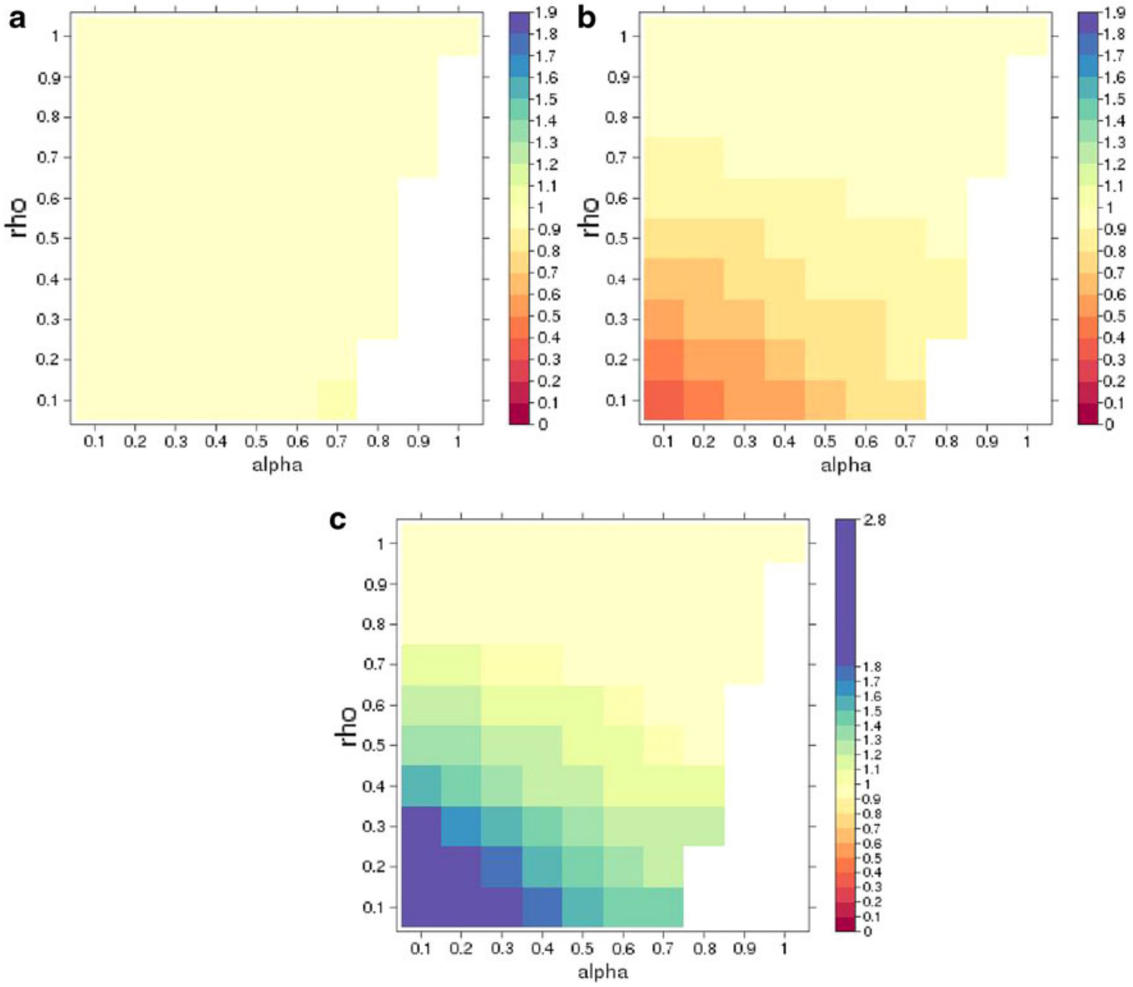

FIG. 2. A comparison of the power of imputation (SSI) versus no imputation **(a)**, SSI-VR versus no imputation **(b)**, and SSI versus SSI-VR in a simple example consisting of three SNPs, of which only two are observed. In each panel, we plot the ratio of the power of the two methods under all configurations of α and ρ. In each figure, the configuration of α and ρ that results in a covariance matrix that is not positive definite, for example, $\alpha = 1,\ \rho = 0.1$, is left empty. **(a)** Shows that for values of $\alpha \le \frac{1+\rho}{2}$, the ratio is near one since the rejection boundary is unchanged (as predicted by our theory). while for values of $\alpha > \frac{1+\rho}{2}$, the power of SSI is greater than that of no imputation. **(b, c)** Show that SSI-VR can lose power relative to both no imputation as well as SSI for a range of configurations of linkage disequilibrium.

and we show that SSI always increases power when $\frac{\alpha}{1+\rho} > \frac{1}{2}$ as the ratio is always larger in 1. Since the power of no imputation depends more on the correlation between tagged and untagged SNPs, we see the power being sensitive to α. For instance, if $\alpha = 0.7$ and $\rho = 0.3$, the average power of no imputation is 0.4918, while the average power of imputation with no correction is 0.6614. In Figure 2b, we compared SSI-VR versus no imputation. We see comparing with Figure 2a, the power increasing much less significantly. In fact, in some cases, we observe SSI-VR has less power than no imputation. For example, when $\alpha = 0.7$ and $\rho = 0.1$, the average power of imputation with variance correction is 0.4639, and null has an average power of 0.5154.

Then, we compare imputation and imputation with variance reweighting in Figure 2c and we notice that SSI-VR will always cause power loss. and in the figure, the values of ratio are all larger than 1. For instance, when $\alpha = 0.7$ and $\rho = 0.3$, the average power of imputation is 0.6614, and the average power of imputation with variance correction is 0.5403.

3.4. *SSI achieves better power compared with existing methods in NFBC*

To assess the power of imputation and the effect of SSI-VR on imputation in a real data set, we simulated marginal statistics utilizing the NFBC data set.

We assume that every other SNP on chromosome 22 is missing. Thus, we observe half of SNPs on chromosome 22 and perform imputation on the rest. We find the per-SNP threshold for only observed SNPs (i.e., no imputation), for SSI and for SSI-VR with the constraint that FWER is controlled at 0.05. We sampled association statistics from the multivariate distribution on the observed SNPs from the genome. Then we used the sampled statistics to find the per-SNP significance threshold on the observed SNPs. We found the threshold to be 4.59705. Having this threshold, we then assume that there are causal SNPs in the genome, that is, the mean of statistics on these SNPs is not 0, and assess the power with no imputation. For no imputation, we found an average power of 0.4946.

Table 1. We Show That the Two Classes of Imputation Method, Summary Statistic Imputation and Two-Step Imputation, Have Similar Imputation Statistics on the Northern Finland Birth Cohort Data Set

Phenotype	*Chr*	*rsID*	*True statistics*	*SSI*	*True SSI*	*SSI-VR*	*True SSI-VR*	*IMPUTE2*	*True IMPUTE2*
TG	2	rs673548	–5.444	–5.37	0.074	–5.37	0.074	–4.46	0.984
	8	rs10096633	–5.679	–5.63	0.049	–5.76	0.082	–5.17	0.509
	15	rs2624265	4.22	3.55	0.67	–3.85	0.37	3.60	0.62
HDL	15	rs1532085	7.13	5.59	1.54	6.33	0.8	6.47	0.66
	16	rs3764261	12.01	8.23	3.78	10.19	1.82	6.47	5.54
	16	rs255049	6.06	5.11	0.95	5.5	0.56	5.70	0.36
	17	rs9891572	4.25	3.99	0.26	4.02	0.23	4.40	0.15
LDL	1	rs646776	–7.70	–7.7	0	–7.81	0.11	–6.96	0.74
	2	rs693	6.81	6.27	0.54	6.34	0.47	5.91	0.9
	11	rs102275	–4.51	–4.43	0.08	–4.45	0.06	–4.54	0.03
	11	rs174546	–4.52	–4.43	0.09	–4.45	0.07	–4.58	0.06
	11	rs174556	–4.69	–4.73	0.04	–4.85	0.16	–4.62	0.07
	11	rs1535	–4.43	–4.46	0.03	–4.66	0.23	–4.45	0.02
	19	rs11668477	–5.96	–3.78	2.18	–4.4	1.56	–5.33	0.63
	19	rs157580	–5.161	–2.6	2.561	–3.11	2.051	–4.20	0.961
CRP	12	rs2650000	–7.08	–5.25	1.83	–6.54	0.54	–6.05	1.03
GLU	2	rs560887	–6.97	–6.21	0.76	–6.3	0.67	–5.69	1.28
	7	rs10244051	5.31	4.34	0.97	4.45	0.86	4.97	0.34
	7	rs2191348	5.30	4.33	0.97	4.47	0.83	4.97	0.33
	11	rs1447352	–6.35	–5.08	1.27	–5.21	1.14	–4.75	1.6
	11	rs7121092	–5.50	–4.93	0.57	–5.31	0.19	–4.60	0.9

We consider SNPs that were reported significant in a previous study (Sabatti et al., 2009). Then, we treat these SNPs as untyped and impute the marginal statistics using SSI, SSI-VR, and two-step imputation using IMPUTE2 to impute genotype of untyped SNPs.

Chr, chromosome; CRP, C-reactive protein; GLU, glutamate; HDL, high-density lipoprotein; LDL, low-density lipoprotein; SNPs, single-nucleotide polymorphisms; SSI, summary statistic imputation; TG, triglycerides.

For the imputation methods, SSI and SSI-VR, we impute the association statistics using the sample statistics. We impute in two ways, one utilizing the MVN of Equation (2), and the other one using the variance reweighting technique as Equation (3). Under the null, we found per-SNP thresholds for SSI and SSI-VR to be 4.5977 and 4.6891. We then assume that there are causal SNPs and used the thresholds to compute the power of each of the imputation methods. We found the average power to be 0.50124 for SSI and 0.4346 for SSI-VR. Notice that the threshold we found for no imputation, SSI, and SSI-VR is more accurate than Bonferroni correction and thus less conservative.

In Table 1, we also impute the most significantly associated SNPs reported in previous studies using SSI, SSI-VR, and a two-step imputation using IMPUTE2 to perform genotype imputation. We find the association statistics are similar across the three methods validating our theoretical results.

4. DISCUSSION

In this study, we have shown that the two broad classes of methods for imputating summary statistics in GWAS, two-step imputation and SSI, have identical asymptotic distributions. We also showed that a commonly used modification of SSI, variance reweighting, will cause power loss using simulation and real data. This leads us to conclude that SSI (with no variance re-weighting) is more powerful while retaining the computational efficiency of methods that rely on summary statistics alone. SSI assumes that statistics follow MVN: this assumption breaks down for small sample sizes and for rare SNPs. Compared with summary statistics, current HMM methods are likely to be more accurate for rare variation. A possible future direction is to improve accuracy on rare variants and small sample sizes.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

S. Sankararaman was supported in part by NIH grants RODGM111T44 and R35GM125055; NSF Grant III-1705121; an Alfred P. Sloan Research Fellowship; and a gift from Okawa Foundation.

SUPPLEMENTARY MATERIAL

Supplementary Data

REFERENCES

Browning, S., and Browning, B. 2007. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097.

Hakonarson, H., Grant, S.F., Bradfield, J.P., et al. 2007. A genome-wide association study identifies kiaa0350 as a type 1 diabetes gene. *Nature* 448, 591–594.

Han, B., Kang, H.M., and Eskin, E. 2009. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet.* 5, e1000456.

Hormozdiari, F., Kichaev, G., Yang, W.-Y., et al. 2015. Identification of causal genes for complex traits. *Bioinformatics* 31, i206–i213.

Hormozdiari, F., Kostem, E., Kang, E.Y., et al. 2014. Identifying causal variants at loci with multiple signals of association. *Genetics* 198, 497–508.

Hormozdiari, F., van de Bunt, M., Segre, A.V., et al. 2016. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* 99, 1245–1260.

Howie, B., Fuchsberger, C., Stephens, M., et al. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* 44, 955–959.

Howie, B.N., Donnelly, P., and Marchini, J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5, e1000529.

Kostem, E., Lozano, J.A., and Eskin, E. 2011. Increasing power of genome-wide association studies by collecting additional single-nucleotide polymorphisms. *Genetics* 188, 449–460.

Köttgen, A., Albrecht, E., Teumer, A., et al. 2012. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat. Genet.* 45, 145–154.

Lee, D., Bigdeli, T.B., Riley, B.P., et al. 2013. Dist: Direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics* 29, 2925–2927.

Li, N., and Stephens, M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165, 2213–2233.

Li, Y., Willer, C., Sanna, S., et al. 2009. Genotype imputation. *Annu Rev Genomics Hum Genet.* 10, 387–406.

Li, Y., Willer, C.J., Ding, J., et al. 2010. Mach: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* 34, 816–834.

Lu, Y., Vitart, V., Burdon, K.P., et al. 2013. Genome-wide association analyses identify multiple loci associated with central corneal thickness and keratoconus. *Nat. Genet.* 45, 155–163.

Marchini, J., and Howie, B. 2010. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11, 499–511.

Marchini, J., Howie, B., Myers, S., et al. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 39, 906–913.

Pasaniuc, B., Zaitlen, N., Shi, H., et al. 2014. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* 30, 2906–2914.

Pritchard, J.K., and Przeworski, M. 2001. Linkage disequilibrium in humans: Models and data. *Am. J. Hum. Genet.* 69, 1–14.

Reich, D.E., Cargill, M., Bolk, S., et al. 2001. Linkage disequilibrium in the human genome. *Nature* 411, 199–204.

Ripke, S., O'Dushlaine, C., Chambert, K., et al. 2013. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* 45, 1150–1159.

Sabatti, C., Hartikainen, A.-L., Pouta, A., et al. 2009. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* 41, 35–46.

Scheet, P., and Stephens, M. 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78, 629–644.

Sladek, R., Rocheleau, G., Rung, J., et al. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 445, 881–885.

Wen, X., and Stephens, M. 2010. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Ann Appl. Stat.* 4, 1158.

Yang, J., Manolio, T.A., Pasquale, L.R., et al. 2011. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* 43, 519–525.

Zeggini, E., Weedon, M.N., Lindgren, C.M., et al. 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 316, 1336–1341.

Address correspondence to:
Prof. Eleazar Eskin
Department of Computer Science
University of California, Los Angeles
296B Engineering VI
Los Angeles, CA 90095

E-mail: eeskin@cs.ucla.edu

Dr. Sriram Sankararaman
Department of Computer Science
University of California, Los Angeles
296B Engineering VI
Los Angeles, CA 90095

E-mail: sriram@cs.ucla.edu