JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 26, Number 11, 2019
© Mary Ann Liebert, Inc.
Pp. 1203–1213
DOI: 10.1089/cmb.2018.0139

Improving Imputation Accuracy by Inferring Causal Variants in Genetic Studies

YUE WU,[1,*] FARHAD HORMOZDIARI,[1–3,*] JONG WHA J. JOO,[4] and ELEAZAR ESKIN[1,5]

ABSTRACT

Genotype imputation has been widely utilized for two reasons in the analysis of genome-wide association studies (GWAS). One reason is to increase the power for association studies when causal single nucleotide polymorphisms are not collected in the GWAS. The second reason is to aid the interpretation of a GWAS result by predicting the association statistics at untyped variants. In this article, we show that prediction of association statistics at untyped variants that have an influence on the trait produces is overly conservative. Current imputation methods assume that none of the variants in a region (locus consists of multiple variants) affect the trait, which is often inconsistent with the observed data. In this article, we propose a new method, CAUSAL-Imp, which can impute the association statistics at untyped variants while taking into account variants in the region that may affect the trait. Our method builds on recent methods that impute the marginal statistics for GWAS by utilizing the fact that marginal statistics follow a multivariate normal distribution. We utilize both simulated and real data sets to assess the performance of our method. We show that traditional imputation approaches underestimate the association statistics for variants involved in the trait, and our results demonstrate that our approach provides less biased estimates of these association statistics.

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

1. INTRODUCTION

Genome-wide association studies (GWAS) have been used to discover the genetic variants 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 collect information on genetic variants, typically single nucleotide polymorphisms (SNPs), from two populations. In this case, the two populations comprise a large number of individuals who carry a specific disease (cases) and those who do not (controls). GWAS estimate correlations between disease status and collected genetic variants. After estimating the

[1] Department of Computer Science, University of California Los Angeles, Los Angeles, California.
[2] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts.
[3] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts.
[4] Department of Computer Science and Engineering, Dongguk University, Seoul, South Korea.
[5] Department of Human Genetics, University of California Los Angeles, Los Angeles, California.
*These authors contributed equally to this work.

correlations, we perform a statistical test to indicate whether each of the estimated correlations is statistically significant. The computed significant statistics are known as summary statistics or marginal statistics. In GWAS, due to cost considerations, only a subset of SNPs, called tag SNPs, are genotyped and SNPs that are not collected are referred to as untyped SNPs. Although genotypes of untyped SNPs are not collected, we can infer these variant genotypes using their correlations to the tag SNPs. The correlation between a pair of variants is referred to as linkage disequilibrium (LD) (Pritchard and Przeworski, 2001; Reich et al., 2001). Imputation is a process that uses LD to compute the genotypes of the missing variants (Marchini et al., 2007; Browning, 2008; Marchini and Howie, 2008, 2010; Howie et al., 2009, 2012; Li et al., 2009, 2010).

Genotype imputation requires two data sets. One data set is a set of individuals who are genotyped at all the SNPs, and this data set is referred to as the reference panel. The other data set, which is the data set of interest, consists of individuals who are only genotyped at the tag SNPs. We can impute the genotypes of untyped SNPs in the second data set by utilizing the correlations between SNPs that are learned from the reference panel. To use the imputed genotypes for GWAS, we compute the summary statistics of the imputed genotypes by applying the same statistical test as if the imputed SNPs are collected in the second data set. In this article, we use summary statistics and marginal statistics interchangeably. Summary statistics, such as z-scores, indicate the magnitude of the associations between genotypes and a phenotype of interest.

There are two methodologies for aiding GWAS analysis with imputation. The standard way of utilizing imputation in the GWAS analysis is to impute the genotypes and compute the summary statistics from the imputed genotypes (Marchini et al., 2007; Browning, 2008; Marchini and Howie, 2008, 2010; Howie et al., 2009, 2012; Li et al., 2009, 2010). More recently, a second class of methods has been developed that directly imputes the marginal statistics. These methods approximate the combined result of genotype imputation and association test results. It is shown that the statistics of tag SNPs and untyped SNPs follow a multivariate normal distribution (MVN) (Han et al., 2009; Kostem et al., 2011; Hormozdiari et al., 2014, 2015, 2016, 2017, 2018). Thus, given the LD between tag SNPs and untyped SNPs, we get a conditional distribution of statistics of untyped SNPs conditioning on the statistics of tag SNPs. Having the statistics of tag SNPs, we can impute the untyped SNPs with mean of the conditional distribution (Lee et al., 2013; Pasaniuc et al., 2014). These methods are shown to have similar accuracy of genotype imputation and are much faster to use for GWAS. Another benefit of the second class of methods is that these methods only require summary statistics to perform imputation while the first class of methods require individual's level genotype data that are not always available.

Genotype imputation has been widely utilized for two reasons in the analysis of GWAS. One reason is to increase the statistical power of association studies when the causal SNPs are not collected in the GWAS. The second reason is to aid the interpretation of GWAS results by predicting the association statistics at untyped variants. Unfortunately, all the existing methods assume a null-based model where all the variants are not causal. As a result, the computed summary statistics for untyped SNPs are lower than the true summary statistics when there exists a causal variant. Thus, the null-based imputation approach is conservative. These approaches are reasonable when the goal is to identify more genetic variants associated with the trait (Marchini et al., 2007; Browning, 2008; Marchini and Howie, 2008, 2010; Howie et al., 2009, 2012; Li et al., 2009, 2010). However, when the goal is to interpret the associated regions to identify the actual causal variants, this assumption will cause bias at variants that are actually causal.

In this article, we introduce a novel method for imputation of summary statistics under the assumption that some SNPs in a locus can be causal. Our approach uses the statistics at tag SNPs and LD patterns to infer which of the variants are causal, and performs imputation with this information taken into account. As shown in previous works (Han et al., 2009; Kostem et al., 2011; Hormozdiari et al., 2014, 2015), the joint distribution of marginal statistics follows MVN, and the mean of the distribution depends on which SNPs are causal. We compute the marginal statistics of the untyped SNPs conditional on the marginal statistics of tag SNP and the knowledge which SNPs are causal. Since we do not know which variants are causal within a region, we impute the marginal statistics of the untyped SNPs as a weighted average of all possible subsets of SNPs in the region to be causal. Unfortunately, considering all possible subsets of SNPs are intractable, so we assume that we have at most three causal SNPs in a locus. This assumption makes our approach applicable to larger loci in the genome without reducing the accuracy of our method. The idea of bounding the number of causal SNPs is widely used in fine-mapping literature (Hormozdiari et al., 2014, 2015, 2016).

We show that our method (CAUSAL-Imp) performs favorably in both simulated and real data. We apply our method to simulated data sets wherein we generated the marginal statistics. Then, we treat some of the

SNPs as untyped and other SNPs as tagged. We apply CAUSAL-Imp and DIST*, which is our implementation of DIST (Lee et al., 2013). We use simulated data to illustrate that CAUSAL-Imp tends to impute summary statistics that are closer to the true generated summary statistics than DIST*. Next, we evaluate our performance utilizing the Northern Finland Birth Cohort (NFBC) data set (Sabatti et al., 2008). We treat the previously reported significant SNPs as untyped and try to impute their summary statistics using CAUSAL-Imp and DIST*. We show that CAUSAL-Imp imputes the associated statistics more accurately than previous approaches.

2. RESULTS

2.1. Overview of CAUSAL-Imp

CAUSAL-Imp builds on methods that perform imputation on summary statistics. It is known that the statistics for a set of SNPs (SNPs in a locus) follow an MVN distribution with a variance–covariance matrix equal to the pairwise correlation between the genotypes (Han et al., 2009; Kostem et al., 2011; Hormozdiari et al., 2014, 2015). For simplicity, let us consider the case wherein one SNP is untyped and the rest are tag SNPs in a region; we have ℓ SNPs and the ℓ-th SNP is untyped. Let s_i be the marginal statistics of the i-th SNP. Let $S_{\neg\ell} = \{s_1, s_2 \cdots s_{\ell-1}\}$ and s_ℓ indicate the marginal statistics for the tag and untyped SNPs, respectively. In traditional methods that impute the summary statistics, the model of the joint distribution is as follows:

$$\left(\begin{bmatrix} S_{\neg\ell} \\ s_\ell \end{bmatrix}\right) \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{\neg\ell} & R_{\neg\ell\ell} \\ R^T_{\neg\ell\ell} & 1 \end{bmatrix}\right), \tag{1}$$

where $\Sigma_{\neg\ell}$ is a $((\ell-1)\times(\ell-1))$ matrix of LD for all the SNPs excluding the ℓ-th SNP and $R_{\neg\ell\ell}$ is a $((\ell-1)\times 1)$ vector that represents the correlation of all the variants with the ℓ-th SNP, excluding the ℓ-th SNP. We can obtain the variance–covariance matrix of the model utilizing the correlation of genotypes from a reference panel, such as the 1000 Genomes data (Durbin et al., 2010; McVean et al., 2012). Then, given the association statistics at observed variants, we can use the conditional form of the multivariate normal to estimate the association statistics at the untyped variants. In traditional methods, marginal statistics of untyped SNPs conditioned on the marginal statistics of tag SNP is as follows:

$$\left(s_\ell | S_{\neg\ell} = \hat{S}_{\neg\ell}\right) \sim \mathcal{N}\left(R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}\hat{S}_{\neg\ell}, 1 - R^T_{\neg\ell\ell}\Sigma_{\neg\ell}{}^{-1}R_{\neg\ell\ell}\right), \tag{2}$$

where $\hat{S}_{\neg\ell}$ is the observed marginals statistics for all the tag SNPs. We impute the untyped SNP with the mean of the mentioned distribution $R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}\hat{S}_{\neg\ell}$ (Lee et al., 2013; Pasaniuc et al., 2014).

Our method, CAUSAL-Imp, takes into account the fact that some variants can be causal. Let us assume we only have one causal SNP and the i-th SNP is causal. Then, the marginal statistics for this SNP follows a normal distribution as follows: $s_i \sim N(\lambda_i, 1)$ where λ_i is the noncentrality parameter (NCP) for the i-th SNP that depends on the true effect size of the SNP toward the phenotype. We extend this to the case where the j-th SNP is not causal and is in LD with the causal SNP i. Then the marginal statistics for the j-th SNP is as follows: $s_j \sim N(r_{ij}\lambda_i, 1)$, where r_{ij} is the LD (genotype Pearson's correlation) between SNPs i and j. To provide a simplified description of this section, we assume that all causal variants have the same NCP. However, CAUSAL-Imp takes into account that causal variants can have different NCP values. We define any subset of SNPs that are causal as the causal status. Causal status indicates which SNPs are causal and which are not. We use 1 to indicate the variants that are causal and 0 to indicate the variants that are not causal. Let $C_{\neg\ell}$ be a vector of size $\ell-1$ to represent the causal status of the first $\ell-1$ SNPs. Similarly, Let c_ℓ be a binary variable that indicates the causal status of the ℓ-th SNP. As shown in previous works (Han et al., 2009; Hormozdiari et al., 2014, 2015), the joint marginal statistics given the causal statistics is as follows:

$$\left(\begin{bmatrix} S_{\neg\ell} \\ s_\ell \end{bmatrix} | \begin{bmatrix} C_{\neg\ell} \\ c_\ell \end{bmatrix}\right) \sim \mathcal{N}\left(\lambda\sqrt{N}\begin{bmatrix} \Sigma_{\neg\ell} & R_{\neg\ell\ell} \\ R^T_{\neg\ell\ell} & 1 \end{bmatrix}\begin{bmatrix} C_{\neg\ell} \\ c_\ell \end{bmatrix}, \begin{bmatrix} \Sigma_{\neg\ell} & R_{\neg\ell\ell} \\ R^T_{\neg\ell\ell} & 1 \end{bmatrix}\right).$$

The summary statistics of untyped SNP (s_ℓ) conditioning on the statistics of the tag SNPs ($S_{\neg\ell}$) and the given causal status, $C = C^*$, are as follows:

$$\left(s_\ell | S_{\neg\ell} = \hat{S}_{\neg\ell}, C_= C^*\right) \sim \mathcal{N}\left(\underbrace{\lambda\sqrt{N}(1 - R^T_{\neg\ell\ell}\Sigma_{\neg\ell}{}^{-1}R_{\neg\ell\ell})c^*_\ell}_{\text{Contribution of causal status for the } \ell-\text{th SNP}} + \underbrace{R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}\hat{S}_{\neg\ell}}_{\text{Contribution of Null}}, 1 - R^T_{\neg\ell\ell}\Sigma_{\neg\ell}{}^{-1}R_{\neg\ell\ell} \right). \quad (3)$$

However, the true causal status is not known. Thus, CAUSAL-Imp considers all the possible causal statuses. We impute summary statistics as a weighted average of all the summary statistics computed for the unobserved variants for different causal status.

$$\sum_{C^*}\left(\lambda\sqrt{N}(1 - R^T_{\neg\ell\ell}\Sigma_{\neg\ell}{}^{-1}R_{\neg\ell\ell})c^*_\ell + R^T_{\neg\ell\ell}\Sigma^{-1}_{\neg\ell}\hat{S}_{\neg\ell}\right)\Pr\left(C = C^* | S_{\neg\ell} = \hat{S}_{\neg\ell}\right), \quad (4)$$

where $\Pr\left(C = C^* | S_{\neg\ell} = \hat{S}_{\neg\ell}\right)$ is the posterior probability of a causal status given the observed marginal statistics. Although we describe the method to consider all possible causal status, in practice, we allow up to three causal variants in a locus to reduce the computational complexity.

2.2. A motivating example

Figure 1 shows a simple region where we have 10 SNPs. In this example, we observe the statistics of three SNPs (SNP3, SNP7, and SNP10), which are indicated by the black arrows. The light triangles indicate the real marginal statistics for all the 10 SNPs. The rest of the SNPs are untyped. Given, the marginal statistics of these three SNPs, we want to impute the marginal statistics of other SNPs. In this example, as the marginal statistic of SNP10 is slightly inflated, we assume one of the SNPs in the region should be causal. In CAUSAL-Imp, we do not know the real causal SNPs, thus we consider all the possible causal statuses in this region. In this example, there are 2^{10} possible causal statuses. For a specific causal status, we impute the summary statistics of the seven unobserved SNPs utilizing the conditional MVN. The dark dots indicate the marginal statistics imputed by CAUSAL-Imp. The light dots indicate the marginal statistics imputed by DIST* [our implementation of DIST; Lee et al. (2013)], which assumes the null model wherein

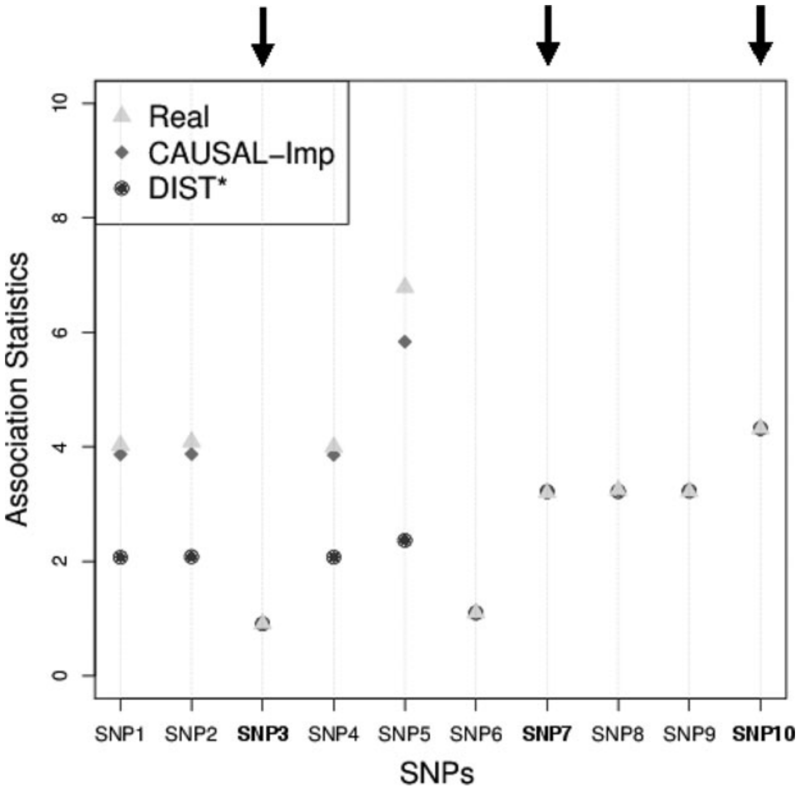

FIG. 1. Motivating example for CAUSAL-Imp. Black arrows indicate the observed (tag) SNPs. Utilizing the fact that the observed marginal statistics of SNP10 is inflated, we can assume one of the SNPs in this region is causal. SNP, single nucleotide polymorphism.

all variants are not causal. In this example, our imputed marginal statistics are closer to the true marginal statistics than those of DIST*.

Note that we perform our evaluations using our own implementation of the standard summary statistic method (DIST) (Lee et al., 2013), which we refer to as DIST*. The reason we used our own implementation is that these methods rely on many matrix operations that may result in numerical issues. The differences in linear algebra libraries dealing with numerical issues can cause differences in the results. By re-implementing DIST, our approach and DIST* share many parts of the implementation to eliminate this issue from the evaluation.

2.3. CAUSAL-Imp achieves better statistics than the existing methods in simulated data sets

To assess the performance of our method, we simulated marginal statistics utilizing the NFBC data set. The NFBC data set consists of 10 phenotypes and 331,476 genotypes measured in 5327 individuals. Since imputation is a regional analysis, we selected 20 regions from the NFBC and computed the LD between each pair of SNPs. In this setting, we use 100 SNPs for each locus. Then, we simulated the marginal statistics from the MVN distribution similar to the previous studies (Zaitlen et al., 2007; Hormozdiari et al., 2014, 2015), where we implant one causal SNP. We generated 1000 sets of summary statistics. We assume that 30% of the SNPs are tagged and that the rest of SNPs, including the causal SNP, are untyped. Then, we run CAUSAL-Imp and DIST* on the simulated data.

We compute the average distance between the imputed marginal statistics and the true simulated marginal statistics as a measure of accuracy. We use the ℓ_1 distance as a measure of accuracy, which is computed as follows: $d(x, y) = \frac{1}{N}\sum_{i=1}^{N} |x_i - y_i|$. We compute this distance for the causal SNP, shown in Figure 2A, and the other SNPs, shown in Figure 2B. We vary the power from 20% to 80%. We observe that the statistics imputed by our method are closer to the true statistics. We perform a similar experiment wherein we implant two causal variants in a locus. In this experiment, the imputed statistics from CAUSAL-Imp are closer to true statistics than those of DIST*. The results for this experiment are not shown due to space limitation.

2.4. CAUSAL-Imp controls Type I error

We illustrate that CAUSAL-Imp performs better than existing methods. In addition, we need to show these methods control the Type I error. Imputed summary statistics that are controlled for Type I error under the null (no variant is causal) are not inflated or deflated. Genomic inflation is a metric used to check

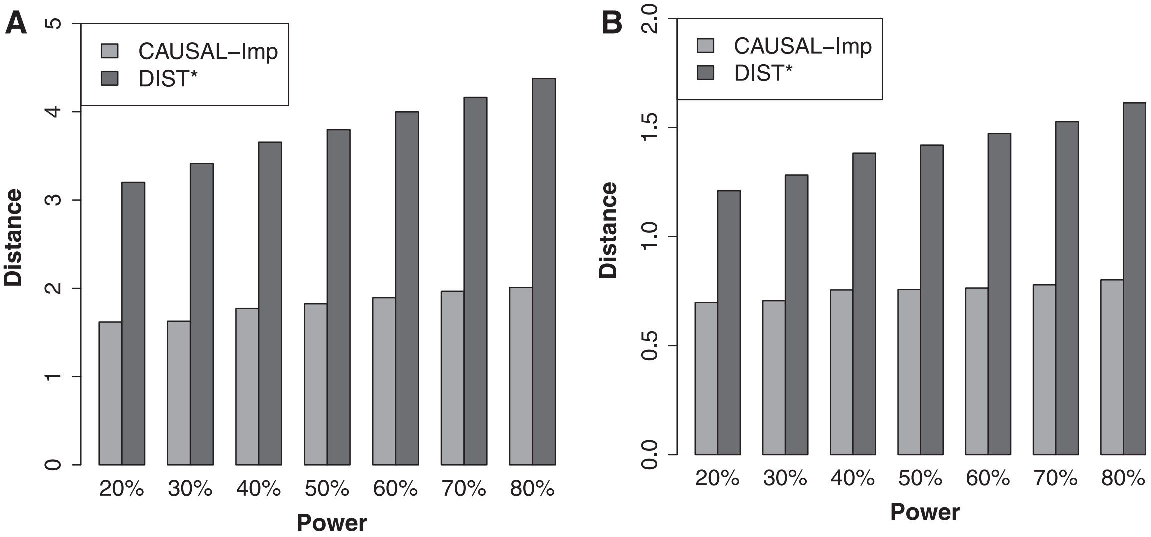

FIG. 2. CAUSAL-Imp achieves better statistics than the existing methods in simulated data sets. We simulated marginal statistics for regions that are obtained from the NFBC data. We compared the imputed marginal statistics of our method and DIST*. Our method tends to impute statistics that are closer to the true estimated marginal statistics both for causal and noncausal SNPs. We use ℓ_1 norm to compute the distance. We range the power on the causal SNPs from 20% to 80%. **(A)** Illustrates the results of the causal variants. **(B)** Illustrates the results of noncausal variants. NFBC, Northern Finland Birth Cohort.

whether the Type I error is controlled (Devlin and Roeder, 1999). We expect the genomic inflation to be close to 1 when there exists no inflation or deflation of statistics. We simulated data under the null where no variant is causal. We consider 30% of the variants to be missing, and then we impute their summary statistics. The genomic inflation for the true summary statistics is 0.98, and the genomic inflation for CAUSAL-Imp is 0.93. However, the genomic inflation of DIST* and IMPUTE2 (Howie et al., 2009) is 0.80 and 1.02, respectively. Thus, CAUSAL-Imp controls the Type I error.

2.5. *CAUSAL-Imp achieves better statistics than the existing methods in NFBC*

The actual utility of our approach is in examining regions that contain associations where the actual causal variants are not collected. We simulate this scenario by taking actual associated regions in the NFBC data set and removing the peak-associated SNPs from each associated regions [which were reported in a previous study; Sabatti et al. (2008)]. We then apply CAUSAL-Imp, DIST*, and IMPUTE2 (Howie et al., 2009) to evaluate the accuracies of these methods on the peak SNPs. The results are given in Table 1. We observe that the imputed summary statistics from CAUSAL-Imp are closer to the estimated summary statistics than those of DIST*.

3. METHODS

3.1. *A standard association statistics*

In this study, we have a quantitative phenotype collected for n individuals at m SNPs. Let Y be a $(n \times 1)$ vector of phenotypic values where y_j is the phenotypic values for j-th individual. Let G be an $(n \times m)$ matrix of minor allele counts, where $g_{ji} \in \{0, 1, 2\}$ is the minor allele count for j-th individual at i-th SNP, and X be the normalized allele counts matrix G. Define β to be an $(m \times 1)$ effect size vector, and β_i is the effect size of i-th SNP. For simplicity, we assume that both the phenotypic values and the allele counts at each SNP are normalized to have mean 0 and variance 1. Let $x_{ji} \in \{\frac{-2p_i}{\sqrt{p_i(1-p_i)}}, \frac{1-2p_i}{\sqrt{p_i(1-p_i)}}, \frac{2-2p_i}{\sqrt{p_i(1-p_i)}}\}$ that is the normalized value for g_{ji}, where p_i is the frequency of i-th SNP in the population. Assuming Fisher's polygenic model holds, we use the generative model, $Y = \mathbf{1}^T \mu + \Sigma_{i=1}^{m} X_i \beta_i + \boldsymbol{e}$, where μ is the phenotypic

Table 1. CAUSAL-Imp Achieves Better Statistics in Northern Finland Birth Cohort Data Set

Phenotype	*chr*	*rsID*	*True statistics*	*DIST**	*CAUSAL-Imp*	*IMPUTE2*
TG	2	rs673548	–5.444	–5.37	**–5.38**	–4.46
	8	rs10096633	–5.679	–5.63	**–5.64**	–5.17
	15	rs2624265	4.22	3.55	**4.15**	3.60
HDL	15	rs1532085	7.13	5.59	**7.17**	6.47
	16	rs3764261	12.01	8.23	**8.28**	6.47
	16	rs255049	6.06	5.11	5.61	**5.70**
	17	rs9891572	4.25	3.99	**4.02**	4.40
LDL	1	rs646776	–7.70	**–7.92**	**–7.92**	–6.96
	2	rs693	6.81	6.27	**6.63**	5.91
	11	rs102275	–4.51	–4.43	–4.44	**–4.54**
	11	rs174546	–4.52	–4.43	–4.45	**–4.58**
	11	rs174556	–4.69	–4.73	–4.75	**–4.62**
	11	rs1535	–4.43	–4.46	–4.46	**–4.45**
	19	rs11668477	–5.96	–3.78	–3.78	**–5.33**
	19	rs157580	–5.161	–2.6	**–5.24**	–4.20
CRP	12	rs2650000	–7.08	–5.25	**–7.36**	–6.05
GLU	2	rs560887	–6.97	–6.21	**–6.80**	–5.69
	7	rs10244051	5.31	4.34	4.67	**4.97**
	7	rs2191348	5.30	4.33	4.66	**4.97**
	11	rs1447352	–6.35	–5.08	**–5.39**	–4.75
	11	rs7121092	–5.50	–4.93	**–5.78**	–4.60

We run association on the NFBC data set. We consider the SNPs that are reported significant in a previous study (Sabatti et al., 2008). Then, we treat these SNPs as untyped and impute the marginal statistics using CAUSAL-Imp, DIST*, and IMPUTE2. Our method tends to produce summary statistics closer to the estimated marginal statistics than the two other methods.

TG, triglycerides; HDL, high-density lipoprotein; LDL, low-density lipoprotein; CRP, C-reactive protein; GLU, glucose.

Bold values indicate the best results.

mean of population, $\mathbf{1}$ is an $(n \times 1)$ vector of 1, X_i is normalized minor allele counts at i-th SNP, β_i is effect size of i-th SNP, and $\boldsymbol{e}$ is a vector of measurement noise and environment contributions. We assume $\boldsymbol{e}$ has a normal distribution with mean 0 and variance, $\sigma^2\boldsymbol{I}$ $(\boldsymbol{e} \sim N(0, \sigma^2\boldsymbol{I}))$.

In standard GWAS, effect size for each SNP is estimated one SNP at a time. Thus, to compute the marginal statistics for each SNP, we use the following model, $Y = \mathbf{1}^T\mu + X_i\beta_i + \boldsymbol{e}$. We note there is a discrepancy between the generative model and testing model; as long as there is no population structure in the data, the estimated effect size is unbiased and follows a normal distribution with mean equal to the true value of effect size. Thus, we have $\hat{\beta}_i = \frac{X_i^TY}{X_i^TX_i}$ and $\hat{\beta}_i \sim N(\beta_i, \sigma(X_i^TX_i)^{-1})$. We use "hat" for each variable to indicate the estimated value for that variable.

It is known that the marginal statistics for each SNP is computed as the ratio between the estimated effect size and the estimated variance. Let s_i indicate the marginal statistics estimated for the i-th SNP. As the marginal statistics follow a normal distribution, we can define the statistics as follows:

$$s_i = \frac{\hat{\beta}_i}{\hat{\sigma}}\sqrt{n} \sim N(\frac{\beta_i}{\sigma}\sqrt{n}, 1) = N(\lambda_i, 1),$$

where λ_i is the NCP for the i-th SNP and $\lambda_i = \frac{\beta}{\sigma}\sqrt{n}$.

3.2. *Indirect association statistics*

To show the indirect association statistics, we assume that i-th variant is associated with the phenotype and j-th variant is correlated with the i-th variant. Thus, the estimated effect size and the marginal statistics for the j-th variant are computed as $\hat{\beta}_j = \frac{X_j^TY}{X_j^TX_j}$, $\hat{\beta}_j \sim N(\beta_j, \sigma(X_j^TX_j)^{-1})$, $s_j \sim N(r_{ij}\lambda_i, 1)$, where r_{ij} is the correlation between genotypes of i-th and j-th SNPs. Moreover, we estimate the correlation between the genotypes as $\frac{1}{n}X_i^TX_j$. We compute the covariance between the estimated marginal statistics for the i-th and j-th SNPs as $\mathrm{Cov}(s_i, s_j) = r_{ij}$. Thus, the joint distribution of the marginal association statistics for the two SNPs given their NCPs follows an MVN:

$$\left(\begin{bmatrix} s_i \\ s_j \end{bmatrix} | \begin{bmatrix} \lambda_i \\ \lambda_j \end{bmatrix}\right) \sim \mathcal{N}\left(\begin{bmatrix} \lambda_i \\ \lambda_j \end{bmatrix}, \begin{bmatrix} 1 & r_{ij} \\ r_{ij} & 1 \end{bmatrix}\right).$$

3.3. *Traditional summary statistics imputation when one SNP is untyped*

In this section, we show how traditional summary statistics imputation approaches (Lee et al., 2013; Pasaniuc et al., 2014) work under the scenario when only one SNP is untyped in a locus. Let us say we have ℓ SNPs in a region where $\ell - 1$ of the SNPs are tagged and only the last SNPs is untyped. We select the ℓ-th SNP to be untyped just for simplicity. Let s_i indicate the marginal statistics of i-th SNP. Let $S_{\neg\ell} = \{s_1, s_2, \cdots s_{\ell-1}\}$ be an $(\ell - 1 \times 1)$ vector of association statistics, $\Lambda_{\neg\ell} = \{\lambda_1, \lambda_2, \cdots \lambda_{\ell-1}\}$ be an $(\ell - 1 \times 1)$ vector of NCPs, and $\Sigma_{\neg\ell}$ be an $(\ell - 1 \times \ell - 1)$ matrix of the pairwise correlation coefficients for the tag SNPs. For the untyped SNP, we use λ_ℓ to indicate the unknown NCP. We want to impute the association statistic s_ℓ, and let $R_{\neg\ell\ell}$ denote the $(\ell - 1 \times 1)$ vector of the correlation coefficients between s_ℓ and the $\ell - 1$ tag SNPs. Thus the joint distribution of the association statistics of the untyped SNP, s_ℓ, and the $\ell - 1$ tag SNPs, $S_{\neg\ell}$, follows a MVN, which can be expressed as follows:

$$\begin{bmatrix} S_{\neg\ell} \\ s_\ell \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \Lambda_{\neg\ell} \\ \lambda_\ell \end{bmatrix}, \begin{bmatrix} \Sigma_{\neg\ell} & R_{\neg\ell\ell} \\ R_{\neg\ell\ell}^T & 1 \end{bmatrix}\right). \tag{5}$$

Under the null assumption where s_ℓ and $S_{\neg\ell}$ are not associated, λ_ℓ and $\Lambda_{\neg\ell}$ are 0's. Using this equation, we can generate a distribution of the statistics of untyped SNP, s_ℓ condition on the observed summary statistics, $S_{\neg\ell} = \hat{S}_{\neg\ell}$. The conditional distribution follows a MVN, which is computed as follows: $\left(s_\ell | S_{\neg\ell} = \hat{S}_{\neg\ell}\right) \sim \mathcal{N}\left(R_{\neg\ell\ell}^T\Sigma_{\neg\ell}^{-1}\hat{S}_{\neg\ell}, 1 - R_{\neg\ell\ell}^T\Sigma_{\neg\ell}{}^{-1}R_{\neg\ell\ell}\right)$. Thus, utilizing this equation, the traditional summary statistics imputation approaches impute the statistics of the untyped SNP as $R_{\neg\ell\ell}^T\Sigma_{\neg\ell}^{-1}\hat{S}_{\neg\ell}$.

3.4. *Traditional summary statistics imputation when more than one SNP is untyped*

In this section, we show how traditional summary statistics imputation approaches (Lee et al., 2013; Pasaniuc et al., 2014) work under the scenario where more than one SNP is untyped in a locus. We use $\mathcal{U}$

and $\mathcal{T}$ to indicate the set of untyped and tag SNPs, respectively. Let $S_{\mathcal{U}}$ and $S_{\mathcal{T}}$ indicate the unobserved summary statistics of untyped SNPs and observe summary statistics of tag SNPs, respectively. We use $\Sigma_{\mathcal{U}}$ and $\Sigma_{\mathcal{T}}$ to denote $(p \times p)$ and $(\ell \times \ell)$ matrices of pairwise correlation coefficients obtained from the untyped SNPs and tag SNPs, respectively. We want to impute unobserved summary statistics $S_{\mathcal{U}}$ using both observed ℓ SNPs and p unobserved SNPs. In this case, $\Lambda_{\mathcal{U}}$ is a $(p \times 1)$ vector of NCPs of untyped SNPs and $\Sigma_{\mathcal{U},\mathcal{T}}$ denotes the $(p \times \ell)$ matrix of the correlation coefficients between the p untyped SNPs and the ℓ tag SNPs. The joint distribution of the association statistics of the untyped SNP $S_{\mathcal{U}}$ and the tag SNPs $S_{\mathcal{T}}$ follows an MVN, which can be expressed as follows:

$$\begin{bmatrix} S_{\mathcal{U}} \\ S_{\mathcal{T}} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \Lambda_{\mathcal{U}} \\ \Lambda_{\mathcal{T}} \end{bmatrix}, \begin{bmatrix} \Sigma_{\mathcal{U}} & \Sigma_{\mathcal{U},\mathcal{T}}^T \\ \Sigma_{\mathcal{U},\mathcal{T}} & \Sigma_{\mathcal{T}} \end{bmatrix} \right). \tag{6}$$

Under the null assumption that the untyped SNPs and tag SNPs are not associated, the NCPs of both $\Lambda_{\mathcal{U}}$ and $\Lambda_{\mathcal{T}}$ are 0's. Using Equation (6), we can generate a distribution of the statistics of the untyped SNPs, $S_{\mathcal{U}}$, conditioned on the observed statistics, $S_{\mathcal{T}} = \hat{S}_{\mathcal{T}}$. The conditional distribution follows an MVN, which is computed as follows:

$$\left(S_{\mathcal{U}} | S_{\mathcal{T}} = \hat{S}_{\mathcal{T}}\right) \sim \mathcal{N}\left(\Sigma_{\mathcal{U},\mathcal{T}}^T \Sigma_{\mathcal{T}}^{-1} \hat{S}_{\mathcal{T}}, \Sigma_{\mathcal{U}} - \Sigma_{\mathcal{U},\mathcal{T}}^T \Sigma^{-1} \Sigma_{\mathcal{U},\mathcal{T}}\right). \tag{7}$$

Thus, utilizing the mentioned equation, the traditional summary statistics imputation approaches impute the statistic of the untyped SNPs as $\Sigma_{\mathcal{U},\mathcal{T}}^T \Sigma_{\mathcal{T}}^{-1} \hat{S}_{\mathcal{T}}$.

3.5. CAUSAL-Imp summary statistics imputation with fixed NCP

Recall that having ℓ SNPs whose summary statistics are observed and p SNPs whose summary statistics are unobserved, we have a MVN expressed as Equation (6). Instead of assuming that all $\Lambda_{\mathcal{U}}$ and $\Lambda_{\mathcal{T}}$ are 0's, our method considers that any subset of SNPs are causal. We introduce C to denote the causal status of the SNPs. Causal status is an $((\ell + p) \times 1)$ vector of 0's and 1's where c_i indicates the causal status of the i-th SNP. Each SNP can have two possible causal statuses 0 or 1, where 0 indicates the SNP is not causal and 1 indicates the SNP is causal. For simplicity, we assume that the NCPs for all the causal variants are the same and equal to $\lambda\sqrt{N}$. Later, we will relax this assumption. There are $2^{\ell+p}$ possible causal statuses for C, which is denoted by the set $\mathcal{C}$ (in practice we only consider up to three causal variants in locus, thus CAUSAL-Imp needs to consider at most $(\ell + p)^3$ causal statuses). The causal status is consisted of two parts, the causal status of tag SNPs, which we denote by $C_{\mathcal{T}}$, and the causal status of untyped SNPs, which we denote by $C_{\mathcal{U}}$. The joint distribution of observed and unobserved summary statistics in Equation (7) can be expressed as follows: $\left(\begin{bmatrix} S_{\mathcal{U}} \\ S_{\mathcal{T}} \end{bmatrix} | \begin{bmatrix} C_{\mathcal{U}} \\ C_{\mathcal{T}} \end{bmatrix} \right) \sim \mathcal{N}\left(\lambda\sqrt{n} \begin{bmatrix} \Sigma_{\mathcal{U}} & \Sigma_{\mathcal{U},\mathcal{T}}^T \\ \Sigma_{\mathcal{U},\mathcal{T}} & \Sigma_{\mathcal{T}} \end{bmatrix} C, \begin{bmatrix} \Sigma_{\mathcal{U}} & \Sigma_{\mathcal{U},\mathcal{T}}^T \\ \Sigma_{\mathcal{U},\mathcal{T}} & \Sigma_{\mathcal{T}} \end{bmatrix} \right)$. Using this equation, we can compute the distribution of the untyped statistics, $S_{\mathcal{U}}$, conditional on the observed statistics, $S_{\mathcal{T}} = \hat{S}_{\mathcal{T}}$, and the known causal status, $C = c^\star$. This conditional distribution follows a multivariate normal that is expressed as follows:

$$\left(S_{\mathcal{U}} | S_{\mathcal{T}} = \hat{S}_{\mathcal{T}}, C = c^\star, \lambda\right) \sim \mathcal{N}\left(\lambda\sqrt{n}(\Sigma_{\mathcal{U}} - \Sigma_{\mathcal{U},\mathcal{T}} \Sigma_{\mathcal{T}}^{-1} \Sigma_{\mathcal{U},\mathcal{T}}) C_{\mathcal{U}} + \Sigma_{\mathcal{U},\mathcal{T}}^T \Sigma_{\mathcal{T}}^{-1} \hat{S}_{\mathcal{T}}, \Sigma_{\mathcal{U}} - \Sigma_{\mathcal{U},\mathcal{T}}^T \Sigma_{\mathcal{T}}^{-1} \Sigma_{\mathcal{U},\mathcal{T}}\right). \tag{8}$$

We want to compute the probability of summary statistics of untyped SNPs given the summary statistics of the tag SNPs, Pr $\left(S_{\mathcal{U}} | S_{\mathcal{T}} = \hat{S}_{\mathcal{T}}\right)$. Utilizing the total probability and Baye's rule, we have

$$\begin{aligned} \Pr\left(S_{\mathcal{U}} | S_{\mathcal{T}} = \hat{S}_{\mathcal{T}}\right) &= \sum_{C^* \in \mathcal{C}, \lambda} \Pr\left(S_{\mathcal{U}}, C = C^* | S_{\mathcal{T}} = \hat{S}_{\mathcal{T}}\right) \\ &= \sum_{C^* \in \mathcal{C}} \Pr\left(S_{\mathcal{U}} | S_{\mathcal{T}} = \hat{S}_{\mathcal{T}}, C = C^*\right) \Pr\left(C = C^* | S_{\mathcal{T}} = \hat{S}_{\mathcal{T}}\right), \end{aligned} \tag{9}$$

where Pr $\left(S_{\mathcal{U}} | S_{\mathcal{T}} = \hat{S}_{\mathcal{T}}, C = C^*\right)$ is computed from Equation (8), and Pr $\left(C = C^* | S_{\mathcal{T}} = \hat{S}_{\mathcal{T}}\right)$ is computed as follows:

$$\Pr\left(C = C^* | S_{\mathcal{T}} = \hat{S}_{\mathcal{T}}\right) = \frac{\Pr\left(S_{\mathcal{T}} = \hat{S}_{\mathcal{T}} | C = C^*\right) \Pr\left(C = C^*\right)}{\sum_{C^\dagger \in \mathcal{C}} \Pr\left(S_{\mathcal{T}} = \hat{S}_{\mathcal{T}} | C = C^\dagger\right) \Pr\left(C = C^\dagger\right)}, \tag{10}$$

where Pr $\left(C=C^{\dagger}\right)$ is the prior of the causal status. Similar to most of the fine-mapping methods, for the prior, we assume that SNPs are independent and the probability of an SNP to be causal is equal to 0.01 (Hormozdiari et al., 2014, 2015). This prior implies a sparsity prior on the causal status. Moreover, Pr $\left(S_{\mathcal{T}}=\hat{S}_{\mathcal{T}}|C=C^*\right)$ is the likelihood of observed summary statistics given the causal status C^*. The observed summary statistics, given the causal status, follows a normal distribution and is computed as follows:

$$(S_{\mathcal{T}}=\hat{S}_{\mathcal{T}}|C=C^*,\lambda)\sim\mathcal{N}\left(\lambda\sqrt{n}(\Sigma_{\mathcal{U},\mathcal{T}}C^*_{\mathcal{U}}+\Sigma_{\mathcal{T}}C^*_{\mathcal{T}}),\Sigma_{\mathcal{T}}\right). \tag{11}$$

Utilizing Equations (8), (10), and (11), we compute the value of Pr $(S_{\mathcal{U}}|S_{\mathcal{T}},\lambda)$ from Equation (9). Thus, we impute $S_{\mathcal{U}}$ as the mean of $(S_{\mathcal{U}}|S_{\mathcal{T}},\lambda)$ as follows:

$$\sum_{C^*\in\mathcal{C}}\left(\lambda\sqrt{n}(\Sigma_{\mathcal{U}}-\Sigma_{\mathcal{U},\mathcal{T}}\Sigma_{\mathcal{T}}^{-1}\Sigma_{\mathcal{U},\mathcal{T}})C_{\mathcal{U}}+\Sigma^{T}_{\mathcal{U},\mathcal{T}}\Sigma^{-1}_{\mathcal{T}}\hat{S}_{\mathcal{T}}\right)P(C=C^*|S_{\mathcal{T}}=\hat{S}_{\mathcal{T}}). \tag{12}$$

3.6. *CAUSAL-Imp summary statistics imputation*

In previous sections, we assume that the NCPs of the causal variants are fixed and their values are known. In this section, we relax this assumption. We utilize CAVIAR-model (Hormozdiari et al., 2014, 2015, 2016) that is used in fine-mapping frameworks. In CAVIAR-model, the joint distribution of marginal statistics (S) given the vector of NCPs (Λ) follows an MVN distribution that is expressed as $(S|\Lambda)\sim\mathcal{N}(\Lambda,\Sigma)$. In addition, the vector of NCPs given the causal status (C) follows an MVN distribution that is expressed as $(\Lambda|C)\sim\mathcal{N}(0,\Sigma\Sigma_C\Sigma)$, where $\Sigma_C=\sigma^2\,\mathrm{diag}(C)$ and diag(X) creates a diagonal matrix where the i-th diagonal element is assigned to x_i. Using the conjugate prior, we have the following:

$$(S|C)\sim\mathcal{N}(0,\Sigma+\Sigma\Sigma_C\Sigma). \tag{13}$$

Thus, utilizing the same statistical framework in CAUSAL-Imp, we have the following:

$$\left(\begin{bmatrix}S_{\mathcal{U}}\\ S_{\mathcal{T}}\end{bmatrix}\middle|\begin{bmatrix}C_{\mathcal{U}}\\ C_{\mathcal{T}}\end{bmatrix}\right)\sim\mathcal{N}\left(\begin{bmatrix}0\\ 0\end{bmatrix},\begin{bmatrix}V_{11} & V_{12}\\ V_{21} & V_{22}\end{bmatrix}\right), \tag{14}$$

where

$$\begin{aligned}
V_{11}&=\Sigma_{\mathcal{U}}+\Sigma_{\mathcal{U}}\sigma^2\,\mathrm{diag}(C_{\mathcal{U}})\Sigma_{\mathcal{U}}+\Sigma^{T}_{\mathcal{U},\mathcal{T}}\sigma^2\,\mathrm{diag}(C_{\mathcal{T}})\Sigma_{\mathcal{U},\mathcal{T}}\\
V_{12}&=\Sigma^{T}_{\mathcal{U},\mathcal{T}}+\Sigma_{\mathcal{U}}\,\mathrm{diag}(C_{\mathcal{U}})\Sigma^{T}_{\mathcal{U},\mathcal{T}}+\Sigma^{T}_{\mathcal{U},\mathcal{T}}\,\mathrm{diag}(C_{\mathcal{T}})\Sigma_{\mathcal{T}}\\
V_{21}&=\Sigma_{\mathcal{U},\mathcal{T}}+\Sigma_{\mathcal{T}}\,\mathrm{diag}(C_{\mathcal{T}})\Sigma_{\mathcal{U},\mathcal{T}}+\Sigma_{\mathcal{U},\mathcal{T}}\,\mathrm{diag}(C_{\mathcal{U}})\Sigma_{\mathcal{U}}\\
V_{22}&=\Sigma_{\mathcal{T}}+\Sigma_{\mathcal{T}}\sigma^2\,\mathrm{diag}(C_{\mathcal{T}})\Sigma_{\mathcal{T}}+\Sigma_{\mathcal{U},\mathcal{T}}\sigma^2\,\mathrm{diag}(C_{\mathcal{U}})\Sigma^{T}_{\mathcal{U},\mathcal{T}}.
\end{aligned}$$

Using the MVN conditional distribution, we have

$$\left(S_{\mathcal{U}}|S_{\mathcal{T}}=\hat{S}_{\mathcal{T}},C=C^{\star}\right)\sim\mathcal{N}\left(V_{12}V_{22}^{-1}\hat{S}_{\mathcal{T}},V_{11}-V_{12}V_{22}^{-1}V_{21}\right). \tag{15}$$

Thus, for a given causal status, the optimal value for the imputed marginal statistics is the mean of the mentioned distribution, which is $V_{12}V_{22}^{-1}\hat{S}_{\mathcal{T}}$. It is worth mentioning that both V_{12} and V_{22} depend on the vector of causal status $C=C^{\star}$. CAUSAL-Imp utilizes Equation (15) instead of Equation (8).

4. DISCUSSION

Genotype imputation is widely used to predict the genotypes of untyped SNPs that are not collected in a data set by utilizing the correlation (LD) between the untyped SNPs and the tag SNPs whose genotypes are collected. We propose a new method, CAUSAL-Imp, which combines the principle of fine mapping and summary statistics imputation. CAUSAL-Imp computes the summary statistics for unobserved SNPs by conditioning on the statistics of the observed SNPs and given causal status. CAUSAL-Imp considers all the possible causal statuses where any subset of SNPs can be causal. Thus, the imputed summary statistic is the weighted average of all the summary statistics computed for the unobserved variants for different causal statuses.

Our approach builds upon the recently developed summary statistics framework for imputation (Lee et al., 2013; Pasaniuc et al., 2014). Imputation methods utilizing hidden Markov models (HMMs) to impute individual level data were developed almost 10 years ago (Marchini et al., 2007; Browning, 2008; Marchini and Howie, 2008, 2010; Howie et al., 2009, 2012; Li et al., 2010) and have been improved ever since. In our approach, we incorporate idea of a causal variant and implicitly are then taking the phenotype into account when performing the imputation. It is theoretically possible to extend the HMM-based imputation approaches to take into account causal variants and phenotypes. However, the implementation of such an approach would be incredibly complicated.

ACKNOWLEDGMENTS

Y.W., F.H., J.W.J.J., and E.E. are supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448, 1320589, 1331176, and 1815624, and National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568, P01-HL28481, R01-GM083198, R01-ES021801, R01-MH101782, and R01-ES022282. E.E. is supported, in part, by the NIH BD2 K award, U54EB020403. We acknowledge the support of the NINDS Informatics Center for Neurogenetics and Neurogenomics (P30 NS062691).

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

Browning, S.R. 2008. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum. Genet*. 124, 439–450.

Devlin, B., and Roeder, K. 1999. Genomic control for association studies. *Biometrics* 55, 997–1004.

Durbin, R.M., Altshuler, D.L., Abecasis, G.R., et al. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073.

Hakonarson, H., Grant, S.F.A., 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., Gazal, S., van de Geijn, B., et al. 2018. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet*. 50, 1041–1047.

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., Segrè, A.V., et al. 2016. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet*. 99, 1245–1260.

Hormozdiari, F., Zhu, A., Kichaev, G., et al. 2017. Widespread allelic heterogeneity in complex traits. *Am. J. Hum. Genet*. 100, 789–802.

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.

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.

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.

Marchini, J., and Howie, B. 2008. Comparing algorithms for genotype imputation. *Am. J. Hum. Genet*. 83, 535–539; author reply 539–540.

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.

McVean, G.A., Altshuler, D.M., Durbin, R.M., et al. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65.

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., Service, S.K., Hartikainen, A.-L., et al. 2008. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet*. 41, 35–46.

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.

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.

Zaitlen, N., Kang, H.M., Eskin, E., et al. 2007. Leveraging the hapmap correlation structure in association studies. *Am. J. Hum. Genet*. 80, 683–691.

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:
Dr. Farhad Hormozdiari
Department of Computer Science
University of California Los Angeles
3532-J Boelter Hall
Los Angeles, CA
90095-1596

E-mail: hormozliari@hsph.harvard.edu

Prof. Eleazar Eskin
Department of Human Genetics
University of California Los Angeles
Los Angeles, CA

E-mail: eeskin@cs.ucla.edu