A Generic Coalescent-based Framework for the Selection of a Reference Panel for Imputation

Bogdan Paşaniuc,[1]* Ram Avinery,[2] Tom Gur,[2] Christine F. Skibola,[3] Paige M. Bracci,[4] and Eran Halperin[1,2,5]

[1]*International Computer Science Institute, Berkeley, California*
[2]*The Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel*
[3]*School of Public Health, University of California, Berkeley, California*
[4]*Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California*
[5]*Molecular Microbiology and Biotechnology Department, Tel-Aviv University, Tel-Aviv, Israel*

An important component in the analysis of genome-wide association studies involves the imputation of genotypes that have not been measured directly in the studied samples. The imputation procedure uses the linkage disequilibrium (LD) structure in the population to infer the genotype of an unobserved single nucleotide polymorphism. The LD structure is normally learned from a dense genotype map of a reference population that matches the studied population. In many instances there is no reference population that exactly matches the studied population, and a natural question arises as to how to choose the reference population for the imputation. Here we present a Coalescent-based method that addresses this issue. In contrast to the current paradigm of imputation methods, our method assigns a different reference dataset for each sample in the studied population, and for each region in the genome. This allows the flexibility to account for the diversity within populations, as well as across populations. Furthermore, because our approach treats each region in the genome separately, our method is suitable for the imputation of recently admixed populations. We evaluated our method across a large set of populations and found that our choice of reference data set considerably improves the accuracy of imputation, especially for regions with low LD and for populations without a reference population available as well as for admixed populations such as the Hispanic population. Our method is generic and can potentially be incorporated in any of the available imputation methods as an add-on. *Genet. Epidemiol.* 2010.

Key words: genotype imputation; coalescent; GWAS; linkage disequilibrium; weighted panel

*Correspondence to: Bogdan Paşaniuc, International Computer Science Institute, 1947 Center St., Berkeley, CA 94704.
E-mail: bogdan@icsi.berkeley.edu
Received 5 November 2009; Revised 25 February 2010; Accepted 12 March 2010
Published online in Wiley Online Library (www.wileyonlinelibrary.com).
DOI: 10.1002/gepi.20505

INTRODUCTION

In an effort to reveal the etiology of complex diseases, genome-wide association studies (GWAS) have been successfully applied to a wide range of diseases [Barrett et al., 2008; Wellcome Trust Case Control, 2007; Zeggini et al., 2008]. In a typical study, a set of cases and a set of controls is genotyped, and then each of the genotyped single nucleotide polymorphisms (SNPs) is tested for association using a statistical test such as the Armitage trend test [Armitage, 1955]. Recent advances in high-throughput genotyping allow such studies to examine about a million SNPs per sample [http://www.affymetrix.com/products_services/arrays/specific/genome_wide_snp6/genome_wide_snp_6.affx, 2009]. These account for about 10% of the total number of common SNPs in the genome [Kruglyak and Nickerson, 2001]. Because the causal SNP is often not typed within the study, it is important to interrogate SNPs that have not been genotyped directly which is normally done through imputation. Imputation methods infer the alleles of SNPs not directly genotyped in the study (or *hidden SNPs*) using the correlation structure between the SNPs (linkage disequilibrium, LD) in the region.

Imputation methods are also used in the case of meta-analysis of multiple studies; i.e. when more than one study has been performed on the same phenotype, a combination of the data sets of the multiple studies will result in increased power to detect association. However, the genotyping platforms often differ across the different studies, and thus, using a naive approach power is only increased for SNPs that are genotyped in at least two of the studies. Fortunately, applying imputation to meta-analysis can overcome this problem as the SNPs that were genotyped in one study can be imputed in the other studies. Such an approach has been shown to be useful in several instances [Barrett et al., 2008; Zeggini et al., 2008].

The starting point of imputation methods is a reference data set such as the HapMap [The International HapMap, 2005], for which the genotypes of a dense set of SNPs are provided. The underlying assumption is that the reference samples, the cases, and the controls are all sampled from the same population. Under this simplifying assumption, the three populations share the same LD structure. Thus, the structure of the LD in the reference

population, in conjunction with the structure of the LD of the observed SNPs within the cases and the controls, may be used to impute the alleles of a hidden SNP.

The assumption that the reference population matches the studied population cannot always be realized in practice. First, some studies involve populations that have no representation in the HapMap project or any other dense genotype panel and as a result the reference population has to be a different population than the studied one. In a first step towards addressing this, recent works [Egyud et al., 2009; Huang et al., 2009; Pemberton et al., 2008] have introduced the idea of weighting the HapMap populations using empirical estimates of ancestry to capture the variation in populations with no match in the HapMap. Second, it has been recently shown [Novembre et al., 2008; Price et al., 2006] that even homogeneous populations such as the European population do not correspond to one large cluster, but rather they correspond to a continuum across different axes of the principal component maps. This implies that the notion of ''one reference data set for all samples in the study'' may not be an optimal strategy. Potentially, the imputation accuracy may improve if the imputation procedure uses different reference populations for different individuals in the studied population sample.

Based on the above intuition, here we propose a method for the selection of a reference dataset that optimizes the overall imputation accuracy. Our method takes as an input the studied population sample (i.e., a set of genotypes of the case and control groups), and a reference dataset. Unlike Egyud et al. [2009] who construct a reference panel for the whole studied population to model untyped variation, we select a ''personalized'' reference data set that is constructed from a subset of the original reference data set, for each sample in the studied population. Intuitively, if we consider a principal component map of the studied and reference population, the individuals who are closer to the sample in that map are more likely to be included in the reference data set of that sample. Leveraging on the above intuition, the method of IMPUTE v2 [Howie et al., 2009] constrains phasing updates at each MCMC iteration in the estimation step of their model to condition on a subset of k haplotypes, ''closest'' in genealogical terms of the individual being updated, instead of all the available haplotypes (reference haplotypes and current-guess haplotypes for the study individuals). Unlike IMPUTE v2, we independently perform imputation at each study individual using a personalized tailored reference panel of haplotypes for that individual constructed only from the reference sample of haplotypes. In addition, in our framework the reference panel of haplotypes for a given sample is allowed to differ from one genomic region to the other. Specifically, we partition the genome into non-overlapping windows and select a *weighted reference* data set for each window in each sample, where the weights correspond to the contribution of each haplotype to the reference population (some haplotypes are ''more important'' than others). The weights are chosen to be the inverse of the expected distance on a random Coalescent realization of the studied sample with each of the reference samples. This makes our approach more robust to handle admixed populations. Indeed, we observe that our framework achieves a considerable improvement in imputation accuracy over the accuracy achieved by previous approaches. Furthermore, the decrease in imputation error rate achieved by our framework can potentially lead to large gains in statistical power of detecting associations at the imputed markers [Huang et al., 2009]. We note, however, that quantifying the gain in power depends on the imputation error rate as well as on how one conducts statistical tests on the imputed genotypes.

Importantly, our approach can be used as an add-on to any imputation method as it defines the reference population, which is a separate process from the imputation method itself. In principle, any existing imputation method can potentially implement our method as a subroutine. Thus, we are not proposing a new imputation method, but an approach that utilizes current imputation methods in a better way by providing it an optimized choice for the reference data set.

We evaluated our framework on a large set of populations both from the HapMap project and the Human Genome Diversity Project [Li et al., 2008]. We found that the weighting scheme improves the imputation accuracy of all genomic regions under any condition tested. Specifically, we found that our proposed framework achieves the greatest improvements for regions with low LD, where existing imputation methods fail to provide accurate results. We also demonstrate that our framework is generic such that it can be incorporated into any imputation method, including Beagle [Browning and Browning, 2009], IMPUTE [Howie et al., 2009; Marchini et al., 2007], and GEDI [Kennedy et al., 2009], to improve the imputation accuracy. The largest gains in accuracy are attained on admixed populations (e.g., Hispanics), and for populations where there was no available reference population. However, we also observed an improvement in the imputation accuracy for populations that do have available dense reference panels such as individuals with European ancestry.

METHODS

The general framework of most imputation methods involves using the typed makers as ''predictors'' for the untyped SNP in conjunction with a model of the LD structure observed among all the typed and untyped SNPs. The information about the correlation structure at the untyped SNPs is usually estimated from large repositories of SNP variation such as the HapMap [The International HapMap, 2005] project (*the reference panels*). The main underlying assumption employed here is that the reference panel and the study population share the same LD patterns. This assumption cannot always be realized in practice as some studies involve populations for which no reference panel is readily available. In such cases, a reference panel of a closely related population can be used for the imputation although this will incur additional imputation errors. It was recently shown [see Huang et al., 2009] that a better approach to impute such population is to use a mixture of populations for the reference panel. However, it is unclear how to choose the relative contribution of each population to the reference panel.

Here, we introduce a method for the selection of a reference panel from a set of populations. Once a reference panel has been selected, any of the existing imputation methods can be applied. The basic idea behind our

method is that the optimal reference population should be different for different samples in the study. Consider for example a study involving individuals of European descent. Intuitively, for individuals with northern European ancestry, it would be beneficial to increase the relative contribution of northern Europeans in the reference panel. Therefore, the imputation results can be improved by creating a separate reference population for each sample of the study. Furthermore, for a given individual, different regions in the genome may originate from different ancestral population. This is obviously the case in recently admixed populations such as Hispanics or African Americans, in which the genome can be divided into long genomic regions originating from one of a few ancestral populations. However, this also is the case for other populations where the genome can be divided into shorter contiguous genomic regions originating from one of a few ancestral populations. Therefore, a tailor-made reference population is constructed for each region of the genome of each of the samples.

THE GENERAL FRAMEWORK

Consider an untyped genotype of an individual in the study at marker i. Let H be the reference panel of haplotypes and g be the multi-locus genotype of the studied individual. We consider haplotypes that span a window of length n SNPs typed both in the reference panel as in the study individual around the untyped SNP, and thus, each haplotype spans $n/2$ typed SNPs downstream and upstream of the untyped SNP. Note that, since we impute every untyped SNP in every study individual independently, we need only to consider the neighboring SNPs that are typed both in the reference panel and in the study individual. That is, the neighboring SNPs untyped in the study individual bring no additional information to the imputation procedure at current untyped SNP. Current imputation methods treat all the haplotypes $h \in H$ equally, a priori. However, we instead give different weights w_h to each reference haplotype h, corresponding to the degree of similarity between h and g. The weights are used to decide the contribution of the specific haplotype to the reference data set for this individual at the SNP i. In practice, a new reference data set is constructed where this haplotype is represented w_h times.

Our suggested framework is generic in that for each individual, and for each genomic region, we find the weights w_h, as described below. We then construct a reference data set tailored to that individual in that genomic region by duplicating each haplotype in the reference data set w_h times. To impute the untyped marker, we can now use any of the existing imputation methods [Browning and Browning, 2009; Howie et al., 2009; Kennedy et al., 2009]. We note, however, that because the assigned weights w_h are not necessarily integer numbers, they can be rounded to the nearest whole integer when this method is included as an add-on to an existing imputation method. In some of the imputation methods, particularly those that are based on hidden Markov models, the weights can be easily incorporated directly, and therefore no rounding is needed. For example, the imputation method of GEDI [Kennedy et al., 2009] uses an Expectation Maximization (EM) procedure based on the reference haplotypes to estimate the parameters of the model which can be easily adapted to include weights for the reference haplotypes.

COALESCENT-BASED HAPLOTYPE WEIGHTS

Intuitively, the weights w_h are chosen to be proportional to the inverse of the time since one of the individual haplotypes and the reference haplotypes coalesced. If the window length is short enough (i.e., we use $n = 20$ for all our experiments which is equivalent to 15 kb on the average for the HapMap panels), we can assume that this region has not been subject to recombination events for a relatively large number of generations, and that the infinite site assumption holds. Thus, following standard coalescent theory [Donnelly and Tavare, 1995; Hudson, 1991; Hudson and Kaplan, 1995], any two haplotypes (h'_1, h'_2) have a unique most recent common ancestor (MRCA). The similarity measure we use is proportional to the time to the MRCA of the two haplotypes. Intuitively, if the time to MRCA is long, enough time has passed for the two haplotypes to drift away from each other, and therefore there is a higher chance that the reference haplotype will be less informative about the untyped SNP.

We now derive the calculation of the time to the MRCA of a pair of haplotypes (h'_1, h'_2). Let k be the number of SNPs where the two haplotypes match out of n SNPs and let t be the number of generations since their MRCA. Then the number of matches k follows a binomial distribution [Walsh, 2001], with

$$P(k) = \frac{n!}{(n-k)!k!}[(1-\mu)^{2t}]^k[1-(1-\mu)^{2t}]^{n-k},$$

where μ is the mutation probability per generation per marker. Since $(1-\mu)^{2t} \approx e^{-2\mu t}$ it follows that the likelihood of t generations since the MRCA given k out of n matches is:

$$L(t|k,n) = \frac{n!}{(n-k)!k!}[e^{-2\mu t}]^k[1-e^{-2\mu t}]^{n-k}.$$

The maximum likelihood of $2\hat{t}\mu$ is attained for $2\hat{t}\mu = \ln(n/k)$ giving $\hat{t} = (1/2\mu)\ln(n/k)$. This estimate is limited by its high variance and highly assymetric confidence intervals [Donnelly and Tavare, 1995; Walsh, 2001]. Particularly for our method, when the two haplotypes match across all markers, the estimate will be of zero generations. To account for this Walsh [2001] proposes a Bayesian posterior estimate, $\bar{t}$, for the time to MRCA using a prior of $p(t) = \lambda \exp(-\lambda t)$, where $\lambda = N_e^{-1}$, as follows:

$$\bar{t} = \frac{h(\mu,k,n,\lambda)}{I(\mu,k,n,\lambda)}$$

with variance

$$\sigma^2(t) = \frac{g(\mu,k,n,\lambda)}{I(\mu,k,n,\lambda)} - \bar{t}^2$$

where

$$I(\mu,k,n,\lambda) = \frac{2^{n-k}(n-k)!\mu^{n-k}}{\prod_{i=0}^{n-k}(\lambda+2\mu(n-i))},$$

$$h(\mu,k,n,\lambda) = \sum_{i=0}^{n-k}(-1)^i\frac{(n-k)!}{i!(n-k-i)!}\frac{1}{(2\mu((k+i)+\lambda)^2)}$$

and,

$$g(\mu,k,n,\lambda) = \sum_{i=0}^{n-k}(-1)^i\frac{(n-k)!}{i!(n-k-i)!}\frac{2}{(2\mu((k+i)+\lambda)^3)}$$

For our experiments we used the Bayesian posterior estimates assuming a flat prior ($\lambda = 0$, corresponding to very large populations) resulting in the following estimate for t:

$$\bar{t} = \frac{\sum_{i=0}^{n-k} (-1)^i \frac{(n-k)!}{i!(n-k-i)!} \frac{1}{(2\mu(k+i)^2)}}{\frac{(n-k)!}{\prod_{i=0}^{n-k}(n-i)}} = \frac{1}{2\mu} * f(n,k)$$

where

$$f(n,k) = \sum_{i=0}^{n-k} (-1)^i \frac{n!}{k!i!(n-k-i)!} \frac{1}{(k+i)^2}.$$

.

As haplotype weights, we chose

$$w_h = C \frac{2\mu}{f(n,k)}$$

where C is a constant. The smallest non-zero weight is 1 which is achieved for $k = \bar{k}$ matches (out of the n considered SNPs) by setting $C = f(n,\bar{k})/2\mu$. Due to computational requirements, in all our experiments below, we limited the total size of the reference panel by using $n = 20$ and $\bar{k} = 19$ resulting in weights of 1 and 2 for haplotypes having 19 and 20 matches out of the $n = 20$ neighboring typed SNPs considered. Any haplotype with less than 19 matches is given a weight of 0. In the unlikely case, when there is no haplotype in the reference panel with at least 19 matches, 30 reference haplotypes are randomly chosen and assigned weights of 1. When averaging across all regions and all individuals, weights of 0, 1 and 2 are assigned to 76.31, 12.01 and 11.68% of the reference HapMap haplotypes. Importantly, we note that, since we are using weights that are proportional to $f(n,k)$ (the term 2μ cancels out), the mutation rate has no impact on our final imputation results.

A technical issue to be addressed is that the true haplotypes of the imputed individuals are unknown. We therefore weight each reference haplotype h by the inverse of the *lower bound* on the time to the MRCA between h and any of the two haplotypes of the individual. To do this, we first compute the upper bound u on the number of matches between h and the unknown haplotypes of the individual. The upper bound can be easily computed based on the multi-locus genotype of the individual and a reference haplotype as n minus the number of markers where the individual genotype is homozygous for one allele and the reference haplotype h has the other allele. As a toy example over 4 SNPs (the notation represents the number of minor alleles at given SNP), if $g = 0102$ and $h = 1001$ then any haplotype compatible with g must have at least 1 mismatch with h (at first SNP) and thus any haplotype compatible with g cannot have more than $u = 3$ matches when compared to h. By plugging u into the Bayesian estimates of $\bar{t}$, we obtain an estimate for the lower bound on the time to MRCA between any haplotype h' compatible with g and the reference haplotype h. As discussed above, the final weight is the inverse of that estimate scaled by a constant factor, namely: $w_h = C(2\mu/f(n,u))$.

RESULTS

DATA SETS

We assessed the performance of the above framework on a wide range of data sets and populations. First, we took the data from the HapMap project [The International HapMap, 2005] composed of dense genotypes from four populations: Utah residents of European Ancestry (CEU), Yoruba people from Ibadan, Nigeria(YRI), Han Chinese individuals from Beijing China, and a Japanese population from the Tokyo area, Japan (CHB+JPT). Next, we used the genotypes provided by the Human Genome Diversity Project HGDP [Li et al., 2008] that consists of 938 unrelated individuals spanning 52 population groups from diverse worldwide locations genotyped at 650,000 SNPs using an Illumina platform. Finally, we assessed the performance of our framework on a Hispanic population composed of 89 samples genotyped at 370,000 SNPs using an Illumina Infinium platform as part of a larger GWAS of the molecular epidemiology of non-Hodgkin lymphoma (NHL) in the San Francisco Bay Area to investigate risk factors for NHL and included incident NHL cases ($N = 2{,}055$) identified from the cancer registry who were frequency-matched to Bay Area population-based controls ($N = 2{,}081$) [Skibola et al., 2008, 2009]. For all the analyses presented in this paper, we considered all non-monomorphic SNPs genotyped in Chromosome 1 (49,105 in total for the HGDP data sets and 23,809 for the Hispanic population).

ACCURACY OF IMPUTATION OF AN ADMIXED POPULATION

We first tested our framework on the Hispanic population collected in the San Francisco Bay Area. The Hispanic population is a recently admixed population; in the last 10–20 generations, three ancestral populations have been mixing to form that population: individuals of European descent, individuals of African descent, and Native Americans. This data set is interesting from an imputation perspective for two reasons. First, because this is a recently admixed population, each region in the genome may originate from a different ancestral population, and therefore the locality of our framework should play a key role in improving the accuracy of imputation. Second, the HapMap populations, which are used as the reference populations, do not include Hispanics or native Americans, and therefore the assumption of matching LD patterns made by current imputation methods is invalid.

We tested our proposed approach as follows. For each individual and each genomic region, we calculated the weights w_h for the different haplotypes, and created a reference set by duplicating each haplotype w_h times.

For the assessment of the accuracy of the imputation over the Hispanic population we used a masking methodology in which a random portion of the study population SNPs were masked as untyped (all the genotypes at the masked SNPs were set as missing) followed by imputation of the masked genotypes. We masked 15% of the SNPs as missing resulting in 3,326 (out of 23,809) SNPs in imputation. To measure the error rate of the imputation we used the standard *genotype imputation error rate* that computes the percentage of erroneously inferred genotypes as a percent of the total masked SNP

genotypes. Note that genotypes for which at least one allele is incorrectly inferred are counted as imputation errors. Following Huang et al. [2009] we also report the *squared correlation coefficient* between the imputed genotypes and the directly measured genotypes computed as:

$$\left(\frac{\sum(x_i-\bar{x}_i)(g_i-\bar{g}_i)}{\sqrt{\sum(x_i-\bar{x}_i)^2\sum(g_i-\bar{g}_i)^2}}\right)^2$$

where x_i and g_i are the imputed and the original genotype and $\bar{x}_i$ and $\bar{g}_i$ are the average genotypes at that SNP; the genotype encodes the number of minor alleles at a given location and has possible values 0,1, or 2.

We assessed the performance of our approach in choosing the reference panel to the "cosmopolitan" method of using all the HapMap populations as reference. We also included in the comparison the *global weighting* approach introduced in Egyud et al. [2009] that consists of creating a reference panel by weighting each HapMap population according to estimated ancestries; for this approach we used the estimated proportions for a self-reported sample of US Latinos from the Multiethnic Cohort of Los Angeles and Hawaii (MEC) data set, a collection of 215,251 adult men and women from Hawaii and Los Angeles County, California collected for the purpose of studying diet and cancer in the United States [Kolonel et al., 2000], resulting in weights of 13:1:6 for the CEU:YRI:JPT+CHB panels [Egyud et al., 2009]. Note that our approach also starts from all the original (no global weights) HapMap populations; however, our algorithm will decide to remove some of these individuals based on their local weights.

To impute genotypes at masked SNPs we used IMPUTE v2 [Howie et al., 2009], a highly accurate and widely used method for genotype imputation. IMPUTE v2 uses a hidden Markov model (HMM) to obtain an estimate of the genotypes at untyped markers in each individual. For all the results presented here we used the default parameters when running IMPUTE v2, namely a sliding window of 5 Mb across the genome with a buffer of 250 Kb, $k = 40$ clusters and 30 iterations with 10 burn-in steps for estimating the parameters of the model. For recombination rates, we used the combined genetic map inferred from all the HapMap panels. In the case of the global weighting approach IMPUTE v2 crashed on several windows, probably due to the large reference panel (the global weighting approach leads to a reference panel 20 times larger in the number of haplotypes than using all HapMap haplotypes), and managed to impute only 2,559 out of 3,326 masked SNPs; the results for the global weighting approach are given only for the 2,559 imputed SNPs. To maintain a meaningful comparison and keep a constant call rate across different scenarios, we used no threshold on the posterior probabilities of the imputed genotypes and thus the call rate across all scenarios was 100%.

Table I shows that our approach outperforms both in terms of error rate and squared correlation coefficient the other two compared methods. We note that, although IMPUTE v2 was run in windows of 21 markers only for our proposed approach, as opposed to thousand of markers at a time (5 Mb window length) as in the case of the un-weighted approach or the global haplotype weighted approach of Egyud et al. [2009], using weights over haplotypes of 21 markers attains the lowest imputation error rate. Indeed, our approach leads to an imputation error rate decrease of over 1% when compared to the other methods, which can potentially lead to a significant increase in statistical power at the imputed SNPs [as recently shown by Huang et al., 2009]. We also note that by using no weights at all we obtained similar results to the global weighting approach of Egyud et al. [2009], most probably because of the weighting procedure implemented in estimating the phase of every individual in each MCMC iteration of IMPUTE v2. Given that the discovery of rare variants has received increasing attention recently, it is important to quantify the results for various minor allele frequencies (MAF). Figure 1 plots the results obtained by IMPUTE v2 under the three approaches for the reference panel composition showing that our approach consistently outperforms the other approaches in terms of accuracy across all MAFs with the largest gains for the more common variants.

TABLE I. Imputation error rate and squared correlation coefficient of IMPUTE v2 method with or without weights averaged across all individuals of the Hispanic data set with 15% (3,326 out of 23,809) random SNPs masked

	Error rate	Squared correlation coeff.
IMPUTE v2	8.93%	0.764
IMPUTE v2 with local haplotype weights	7.78%	0.795
IMPUTE v2 with global haplotype weights [Egyud et al., 2009]	9.00%	0.758

All the 420 HapMap haplotypes (CEU+YRI+CHB+JPT) were used as reference panel using either no weighting, our proposed weighting or the global weighting of Egyud et al. [2009].

IMPUTATION ACCURACY AND LD

One of the major drawbacks of imputation methods is that the quality of imputed SNPs in regions of low LD is very poor. Importantly, we observed that when using our approach the regions of low LD were the regions with the largest gain in accuracy. Figure 2 displays the imputation error rates and squared correlation coefficient obtained by IMPUTE v2 as a function of the average LD (computed using the standard r^2 measure) between the masked SNPs and $n = 20$ SNPs in their flanking regions ($n/2$ upstream and $n/2$ downstream). Notably, there is a tight correlation between the gain in using weighted haplotype panels and the average LD in the imputed regions. Specifically, the largest gain is attained in regions with low LD (e.g. a reduction of more than 6% in terms of error rate and an increase of over 0.1 in terms of correlation coefficient for SNPs with an average $r^2 < 0.05$ to their neighboring SNPs) whereas in high LD regions the weighted and un-weighted versions achieve similar accuracies. This further underlines the capacity of our proposed weighting scheme to reduce the error rates of imputation methods, making reliable imputation applicable to more regions across the genome. For example, haplotype weighting increases the total number of SNPs imputed correlation greater than 0.6 from 76 to 83% on the Hispanic data set.

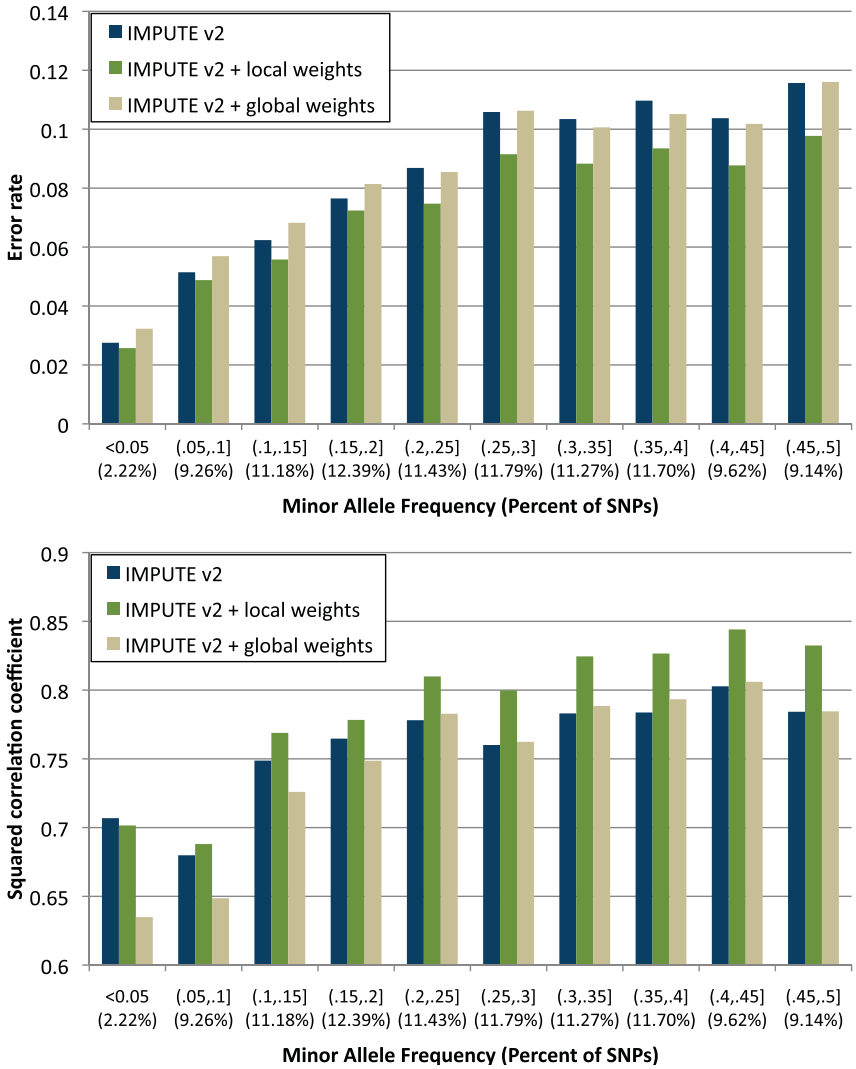

Fig. 1. Imputation error rates (top) and squared correlation coefficient (bottom) obtained by IMPUTE v2 with and without weights for SNPs with various minor allele frequencies (MAF) on the Hispanic dataset with 15% (3,326 out of 23,809) random SNPs masked. In parentheses below each bar is the frequency of SNPs in that category.

When looking at the accuracy of the imputed allele frequencies, we found that the imputed frequencies achieve a Pearson correlation of 99.24 (99.21) for the weighted (unweighted) versions of IMPUTE v2 when averaged over all masked SNPs.

GENERALITY OF THE APPROACH

In principle, the weighting scheme framework can be applied together with any imputation method. Therefore, in addition to IMPUTE v2, we tested our approach on two other imputation methods when applied with or without our proposed weighting scheme: BEAGLE [Browning and Browning, 2009] and GEDI [Kennedy et al., 2009]. BEAGLE makes use of a special class of HMM's called haplotype HMMs to obtain reliable estimates of haplotype phase and missing data while GEDI implements a model similar to the one used by Kimmel and Shamir [2005] and Rastas et al. [2008] trained using a standard EM procedure on the reference panel of haplotypes. Imputation at untyped SNPs is performed based on the conditional probability of the alleles at that SNP given the rest of the observed genotypes for that individual.

Results in Table II show that in this case, all three methods yield improved imputation accuracy when our proposed weighting framework was applied, although the compared methods were run on short windows of 21 markers when local haplotype weights were used, as opposed to whole chromosome data sets for the no weighting approach. Specifically, BEAGLE and IMPUTE v2 result in substantial improvements in terms of error rates (a reduction of approximately 2% out of 10% in the error rate).

IMPUTATION IN THE ABSENCE OF A REFERENCE POPULATION

As there are only a small number of populations for which a dense reference panel is available (i.e., the HapMap populations), we explored the effect of our framework on

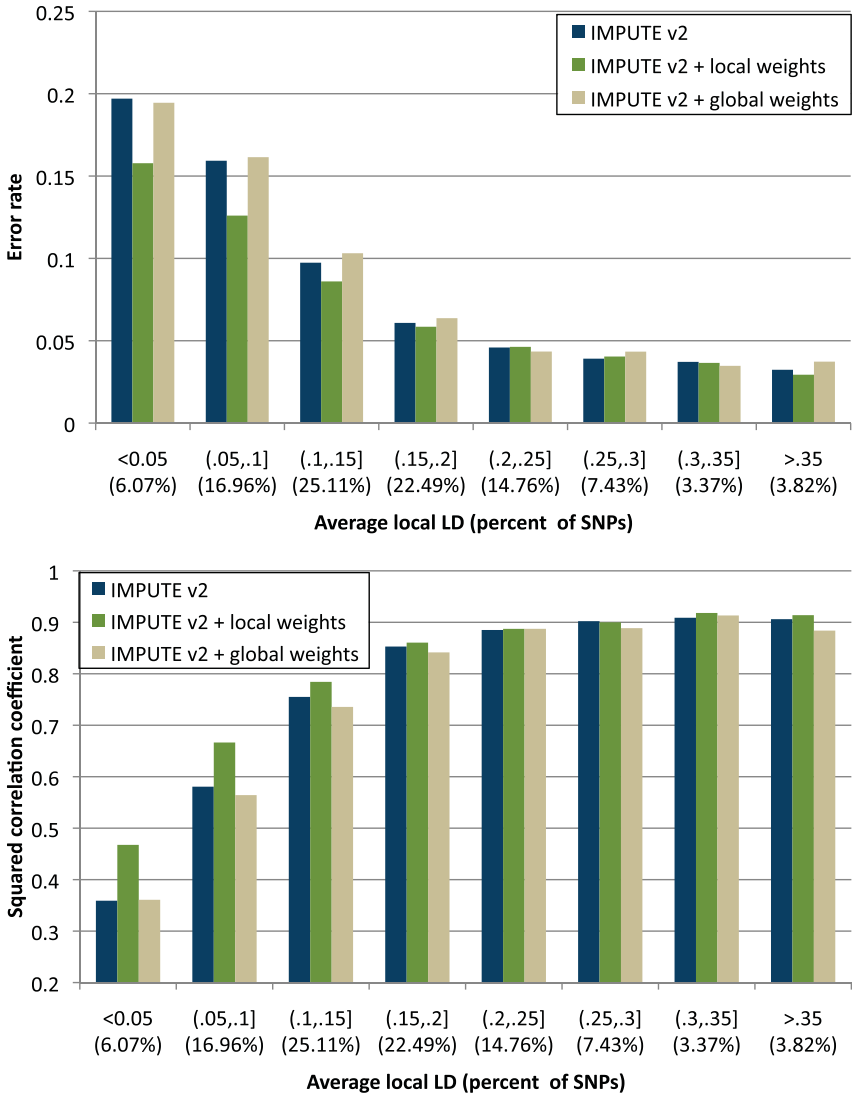

Fig. 2. Imputation error rates (top) and squared correlation coefficient (down) obtained by IMPUTE v2 with and without weights for SNPs with various amounts of local average LD on the Hispanic dataset with 15% (3,326 out of 23,809) random SNPs masked. In parentheses below each bar is the frequency of SNPs in that category.

TABLE II. Imputation error rate of all compared methods with or without weights averaged across all individuals of the Hispanic data set with 15% (3,326 out of 23,809) random SNPs masked

	BEAGLE	IMPUTE v2	GEDI
No weighting	10.39%	8.93%	11.62%
Weighted haplotypes	8.78%	7.78%	10.99%

All the 420 HapMap haplotypes (CEU+YRI+CHB+JPT) were used as reference panel.

populations that have no reference. For the following experiments we used GEDI because of its computational efficiency. Also, because we had access to the code we were able to improve the computation efficiency by implementing the full coalescent-based haplotype weighting scheme inside the code rather than as an external procedure. As demonstrated above, we expect the results to be transferable to other imputation methods.

We imputed HGDP genotypes using HapMap populations as reference. For each HGDP population, we varied the HapMap panels (we either used the CEU, YRI, CHB+JPT or All the HapMap haplotypes) to allow us to explore which HapMap panel best fits each individual from the HGDP. Due to computational issues for each individual we randomly masked only 1% of the 49k SNPs of Chromosome 1 as untyped resulting in 496 SNPs in imputation. We report the average error rate across the individuals of each continental group. Results in Table III show that the best choice for the reference panel consists of the HapMap population from the same continent if available, or all the HapMap haplotypes otherwise. This is consistent with the findings of Huang et al. [2009]. Data in Table III further show that the imputation error rate obtained when using haplotype weighting over all the HapMap haplotypes usually provides the best accuracy.

TABLE III. Imputation error rates averaged by continent when different HapMap populations were used as a reference haplotype panel

	Continent						
Reference panel	Middle East	Africa	Oceania	East Asia	Central Asia	Europe	America
CEU	8.29%	26.98%	8.99%	8.99%	7.48%	5.27%	7.31%
JPT+CHB	15.70%	28.41%	8.98%	4.40%	12.15%	13.39%	6.02%
YRI	16.32%	12.59%	14.32%	15.93%	16.53%	17.32%	14.51%
All	7.35%	14.34%	6.99%	4.64%	6.90%	6.33%	4.53%
All (weighted)	6.31%	12.74%	6.58%	4.05%	6.11%	5.22%	4.15%

"All" denotes the panel of haplotypes obtained by merging all the HapMap haplotypes. 1% of the 49k SNPs from Chromosome 1 were masked.

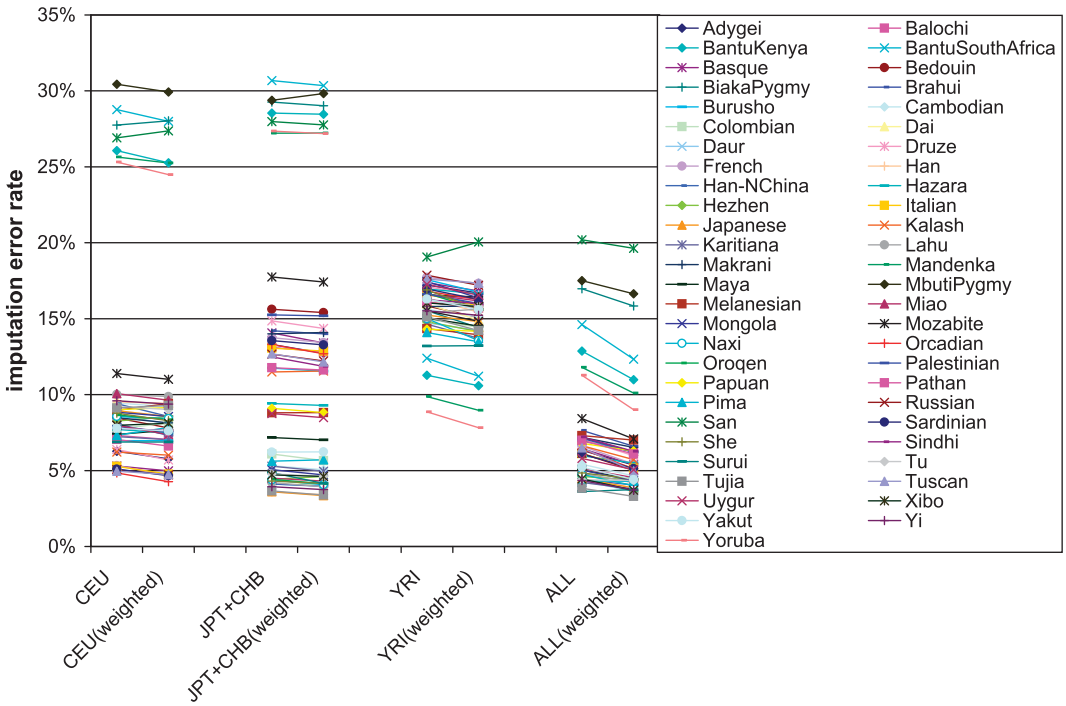

Fig. 3. Average imputation error rate for each population in the HGDP data with or without haplotype weights. The *X*-axis plots the HapMap population used as reference panel used while the *Y*-axis shows the error rate obtained by GEDI.

We also looked at the effect of haplotype weighting on the imputation error rates for each of the HGDP populations (as opposed to continents). In Figure 3 the error rate per population is plotted to show that in most scenarios haplotype weighting decreases the error rate of the imputation procedure.

PCA-BASED REFERENCE POPULATION

It has been demonstrated that human populations tend to display a continuum of genetic variation across different axes of the principal component maps, and that principal component analysis can be reliably used to detect differences between populations [Novembre et al., 2008; Price et al., 2006]. This suggests that PCA could be used as a guide to construct a reference population. In essence, a suitable reference population for the imputation of a genotype g should be physically close to g on the PCA map. In addition, our proposed weighting scheme can be used to improve the accuracy of the imputation by changing the PCA-based reference data sets according to the weights. Therefore we performed a principal component analysis on the HGDP genotypes (we used the complete genotype data over all the SNPs in the PCA) and for each genotype g we tested the imputation accuracy using the haplotypes of the closest 50, 100 or 200 individuals in the PCA map as reference. For the results presented here, we used the first two principal components although similar results (not shown) were obtained when more principal components are employed. We compared these scenarios to the cases where the haplotypes from the same population or continental ancestry of g were used as the reference. Haplotypes were estimated from genotype data using the BEAGLE [Browning and

TABLE IV. Imputation error rate of GEDI averaged across all individuals with or without haplotype weights when reference haplotype panels with various degrees of relatedness are used

	Reference haplotype panel					
	Population	Continental	All	PCA-50	PCA-100	PCA-200
No weighting	7.32%	5.11%	6.06%	5.46%	5.12%	5.10%
Weighted haplotypes	6.95%	4.42%	5.11%	4.92%	4.50%	4.42%

''Population'', ''Continental'' and ''All'' denotes the haplotypes panel inferred from genotypes from the same population, continent or all continents. PCA-50 (100,200) denotes the haplotypes of the closest 50 (100,200) individuals using PCA distance. 1% of the 49k SNPs from Chromosome 1 were masked.

TABLE V. Imputation error rates averaged for each continent with or without haplotype weights

	Middle East	Africa	Oceania	East Asia	Central Asia	Europe	America
No weighting	5.22%	10.85%	4.17%	3.44%	4.76%	3.83%	2.58%
Weighted haplotypes	4.53%	9.21%	3.86%	3.10%	4.11%	3.36%	2.31%

The results were obtained using the top 200 PCA haplotypes as reference. 1% of the 49 k SNPs from Chromosome 1 were masked.

Browning, 2007] phasing software. For each individual we randomly masked 1% of the SNPs as untyped.

Table IV presents the average imputation error rates across the HGDP individuals, with or without haplotype weighting. As expected, the PCA correctly inferred the ''closest'' genotypes with respect to genotype imputation. Indeed, using the closest 200 (100) genotypes yielded an error rate of 5.10% (5.12%) similar to that obtained when the continental haplotypes were used as the reference. This is of particular importance when imputation is performed on individuals whose origins are ambiguous as PCA can reliably infer the correct panel of reference haplotypes. Furthermore, Table IV shows that weighting the haplotypes significantly improved the imputation in all scenarios, regardless of the reference panel used. The greatest improvement was observed when all the HGDP haplotypes were provided as a reference panel, possibly due to a much larger set from which to choose. We also analyzed the distribution of improvements in imputation accuracy across continents. We found that the imputation error rate was improved consistently across all populations (see Table V). The largest improvement was observed for the African genotypes. These results demonstrate that the haplotype weighting scheme is beneficial regardless of the continental group of the individual under imputation and regardless of the reference haplotype panel chosen.

DISCUSSION

Imputation is a widely used tool for increasing the association power in GWAS, and therefore improving its performance is critical. It is particularly important to improve and analyze imputation methods in regions of the genome in which their accuracy is low (e.g., regions of low LD) as well as for recently admixed populations and for populations with no available reference panel. Here, we suggest a framework that considerably improved the accuracy of imputation, across all methods and scenarios studied, improvement that could lead to higher power for detecting association at the imputed markers. The main drawback of our framework is the one order of magnitude increase in runtime, which can greatly be reduced by incorporating the weighting scheme inside the method rather than an external add-on as well as by parallelizing the computations (which is trivial in our framework since every local window of every individual is imputed independently).

An important feature of our framework is that it is especially robust in regions of low LD. Imputation methods typically perform poorly in regions of low LD, and as we showed in the Results section, the weighting scheme reduced the error rates in such regions considerably, resulting in an increase in the number of reliably imputed markers that can be added to the association study.

Our framework is based on a personalized construction of a weighted reference population for each region in the genome. This decomposition of the genome into separate regions also leads to considerable improvements in the imputation accuracy on admixed populations, in which different regions of the genome may originate from different ancestries. Furthermore, the personalized construction of the reference population resulted in an improvement in the overall accuracy across all populations.

ACKNOWLEDGMENTS

E.H. is a faculty fellow of the Edmond J. Safra Bioinformatics program at Tel-Aviv University. E.H. was supported by National Science Foundation grant IIS-071325412 and by the Israel Science Foundation grant no. 04514831. B.P. was supported by National Science Foundation grant IIS-071325412. T.G. and R.A. were supported by the Israel Science Foundation grant no. 04514831.

REFERENCES

2009. http://www.affymetrix.com/products_services/arrays/specific/genome_wide_snp6/genome_wide_snp_6.affx.

Armitage P. 1955. Tests for linear trends in proportions and frequencies. Biometrics 11:375–386.

Barrett J, Hansoul S, Nicolae D, Cho J, Duerr R, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR,

Xavier RJ, NIDDK IBD Genetics Consortium, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40:955–962.

Browning SR, Browning BL. 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.

Browning BL, Browning SR. 2009. A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84:210–223.

Donnelly P, Tavare S. 1995. Coalescents and genealogical structure under neutrality. Annu Rev Genet 29:401–421.

Egyud M, Gajdos Z, Butler J, Tischfield S, Le Marchand L, Kolonel L, Haiman C, Henderson B, Hirschhorn J. 2009. Use of weighted reference panels based on empirical estimates of ancestry for capturing untyped variation. Hum Genet 125:295–303.

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

Huang L, Li Y, Singleton A, Hardy J, Abecasis G, Rosenberg N, Scheet P. 2009a. Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet 84:235–250.

Huang L, Wang C, Rosenberg NA. 2009b. The relationship between imputation error and statistical power in genetic association studies in diverse populations. Am J Hum Genet 85:692–698.

Hudson R. 1991. Gene genealogies and the coalescent process. Oxford Surv Evol Biol 7:1–44.

Hudson R, Kaplan N. 1995. The coalescent process and background selection. Philos Trans R Soc Lond B Biol Sci 349:19–23.

Kennedy J, Mandoiu I, Pasaniuc B. 2009. Gedi: scalable algorithms for genotype error detection and imputation. Technical Report 0911.1765, Cornell University.

Kimmel G, Shamir R. 2005. GERBIL: genotype resolution and block identification using likelihood. Proc Natl Acad Sci USA 102:158–162.

Kolonel LN, Henderson BE, Hankin JH, Nomura AMY, Wilkens LR, Pike MC, Stram DO, Monroe KR, Earle ME, Nagamine FS. 2000. A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. Am J Epidemiol 151:346–357.

Kruglyak L, Nickerson D. 2001. Variation is the spice of life. Nat Genet 27:234–236.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko A, Auton A, Indap A, King K, Bergmann S, Nelson M, Stephens M, CD B. 2008. Genes mirror geography within Europe. Nature 456:274.

Pemberton TJ, Jakobsson M, Conrad DF, Coop G, Wall JD, Pritchard JK, Patel PI, Rosenberg NA. 2008. Using population mixtures to optimize the utility of genomic databases: linkage disequilibrium and association study design in India. Ann Hum Genet 72:535–546.

Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909.

Rastas P, Koivisto M, Mannila H, Ukkonen E. 2008. Phasing genotypes using a hidden Markov model. In Mãndoiu I, Zelikovsky A, editors, Bioinformatics Algorithms: Techniques and Applications. New York: Wiley. p 355–372.

Skibola CF, Bracci PM, Halperin E, Nieters A, Hubbard A, Paynter RA, Skibola DR, Agana L, Becker N, Tressler P, Forrest MS, Sankararaman S, Conde L, Holly EA, Smith MT. 2008. Polymorphisms in the estrogen receptor 1 and vitamin c and matrix metalloproteinase gene families are associated with susceptibility to lymphoma. PLoS ONE 3:e2816.

Skibola CF, Bracci PM, Halperin E, Conde L, Craig DW, Agana L, Iyadurai K, Becker N, Brooks-Wilson A, Curry JD, Spinelli JJ, Holly EA, Riby J, Zhang L, Nieters A, Smith MT, Brown KM. 2009. Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. Nat Genet 41:873–875.

The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

Walsh B. 2001. Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a Pair of Individuals. Genetics 158:897–912.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

Zeggini E, Scott L, Saxena R, Voight B, Marchini J, et al. 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40:638–645.