Statistics and Its Interface Volume 4 (2011) 339–351

Practical consideration of genotype imputation: Sample size, window size, reference choice, and untyped rate

Boshao Zhang, Degui Zhi, Kui Zhang, Guimin Gao,
Nita A. Limdi and Nianjun Liu*

Imputation offers a promising way to infer the missing and/or untyped genotypes in genetic studies. In practice, however, many factors may affect the quality of imputation. In this study, we evaluated the influence of untyped rate, sizes of the study sample and the reference sample, window size, and reference choice (for admixed population), as the factors affecting the quality of imputation. The results show that in order to obtain good imputation quality, it is necessary to have an untyped rate less than 50%, a reference sample size greater than 50, and a window size of greater than 500 SNPs (roughly 1 MB in base pairs). Compared with the whole-region imputation, piecewise imputation with large-enough window sizes provides improved efficacy. For an admixed study sample, if only an external reference panel is used, it should include samples from the ancestral populations that represent the admixed population under investigation. Internal references are strongly recommended. When internal references are limited, however, augmentation by external references should be used carefully. More specifically, augmentation with samples from the major source populations of the admixture can lower the quality of imputation; augmentation with seemingly genetically unrelated cohorts may improve the quality of imputation.

Keywords and phrases: Genotype imputation, Genetic study, Admixed population, Untyped rate, Window size, Reference.

1. INTRODUCTION

Genome-wide association studies have successfully identified regions of the genome associated with disease risks and other heritable traits. One of the key steps for these studies is to collect the genotype data across the genome. Despite advances in genotyping technology, missing genetic information can arise from varieties of sources, such as errors in genotype calling, variation in depth and scope of assessment across different genotyping platforms. Recently, handling missing genotypes by design is drawing more attention (Marchini and Howie 2010; Spencer, et al. 2009a; The Wellcome Trust Case-Control Consortium 2007). Although ignoring the untyped loci in association analysis is a common practice, availability of complete information can significantly enhance power (Anderson, et al. 2008a; Hao, et al. 2009; Spencer, et al. 2009b).

Genotype imputation is the process of inferring genotypes that are not directly assessed in a cohort of individuals (Anderson, et al. 2008b; Li, et al. 2009; Marchini and Howie 2010; Servin and Stephens 2007; Spencer, et al. 2009b; The Wellcome Trust Case-Control Consortium 2007). Advances in the understanding of the genome structure, substructure and population admixture, and statistical genetic methodology have enabled imputation of the missing genetic information. Therefore imputation of genotypes has been widely adopted to boost power, fine-map associations and synchronize the genotype data from studies using different platforms (Hao, et al. 2009; Marchini and Howie 2010; Zeggini, et al. 2008). The need for robust imputation has led to development of novel imputation algorithms through a wide-array of software packages, such as IMPUTE (Marchini, et al. 2007b), MACH (Li, et al. 2010), FastPHASE (Scheet and Stephens 2006), and BEAGLE (Browning and Browning 2007). Although studies have compared various software packages and their underlying methods for imputation (Altshuler, et al. 2010; Aulchenko, et al. 2010; Browning 2008; Nothnagel, et al. 2009; Pei, et al. 2008; Pei, et al. 2010), a systematic investigation of the factors affecting the quality of imputation is needed. Ensuring quality of imputation is vital to the downstream analyses as illustrated by the recent study by Huang et al. (Huang, et al. 2009c) which highlighted how imputation errors can seriously compromise statistical power.

Herein we focus on the influence of four factors on the quality of genotype imputation: sample size, untyped rate, window-size, and reference choice. Guided by the work of Huang et al. (Huang, et al. 2009b; Huang, et al. 2009c) and Pei et al. (Pei, et al. 2008) we evaluate both accuracy and efficacy as measures of the quality of imputation. Although most algorithms specify a fairly large window

*Corresponding author. Correspondence address: Department of Biostatistics, Ryals Public Health Bldg, 420A, 1665 University Boulevard, University of Alabama at Birmingham, Birmingham, AL 35294, USA. Tel.: (205) 975-9190, fax: (205) 975-2540.

size (the length of the chromosomal region, for example, is at least 5MB for IMPUTE2 (Marchini, et al. 2007a)) for analysis, a smaller window size (< 5 MB) may be necessary for analysis of discreet regions such as candidate genes (Huang, et al. 2009c). We therefore investigate how window size smaller than 5 MB affects the quality of genotype imputation. Several studies have discussed reference choices (Huang, et al. 2009a; Shriner, et al. 2010; Zhao, et al. 2008). Huang et al. (Huang, et al. 2009a) proposed strategies to optimize the quality of imputation for admixed study samples using existing external references. They evaluated the "portability" of the HapMap data (The International HapMap Consortium 2003) as reference panels. Zhao et al. (Zhao, et al. 2008) used principal component analysis to stratify the African American samples into two groups: one group close to YRI (Yoruba in Ibadan, Nigeria) and the other group close to CEU (Utah residents with Northern and Western European ancestry from the CEPH collection). They found that the accuracy of genotype imputation for the group close to CEU with CEU as reference improved dramatically compared with the accuracy of imputation for this group with YRI as the reference. These studies, however, focused on external references, claiming that a "cosmopolitan" reference panel formed by pooling the available reference cohorts would work for practical use (de Bakker, et al. 2006; Huang, et al. 2009a). Intuitively, internal references (i.e., reference samples drawn from the same population as the study samples) would be preferred. But they are usually more expensive to obtain because the studies have to genotype a proportion of individuals as references with a much greater density. Thus, a study is necessary to evaluate internal references in terms of the gain in the quality of imputation against the cost for internal reference. In addition, when an internal reference panel is small, the researchers may want to augment the reference panel with some existing external references. In this case, researchers are interested in the portability of the "cosmopolitan rule".

These questions are relevant to our ongoing work in pharmacogenomics of warfarin, the most commonly used oral anticoagulant. Anticoagulant therapy with warfarin is challenging due to marked and often unpredictable variability in response (Aquilante, et al. 2006; Budnitz, et al. 2007). The recent years have witnessed a bourgeoning understanding of genetic regulation of warfarin response making warfarin the '*poster child*' for pharmacogenetics. Clearly, the bulk of available evidence supports a major influence of polymorphisms in two genes: vitamin K epoxide reductase complex 1(*VKORC1*) and cytochrome P450 2C9 (*CYP2C9)* in determining warfarin dose in populations of European, Asian descent (Aquilante, et al. 2006; Borgiani, et al. 2007; Caldwell, et al. 2007; Carlquist, et al. 2010; Cho, et al. 2007; D'Andrea, et al. 2005; Furuya, et al. 1995; Gage, et al. 2008; Klein, et al. 2009; Limdi, et al. 2008a; Limdi, et al. 2010a; Limdi, et al. 2010b; Limdi, et al. 2009; Takahashi, et al. 2006; Wadelius, et al. 2007; Wadelius, et al. 2009; Zhu, et al. 2007), and recently African descent (Limdi, et al. 2008b; Momary, et al. 2007; Schelleman, et al. 2007). Among patients of European descent, polymorphisms in *CYP2C9* and *VKORC1* explain 30–35% of the variability in warfarin dose while clinical and demographic factors account for an additional 20 to 25% (Crawford, et al. 2007; Klein, et al. 2009; Rieder, et al. 2007). However, among patients of African descent, a smaller proportion of variability is accounted by *CYP2C9* (2-5%) and *VKORC1* (5–7%). Recognizing that differences in *VKORC1* haplotype structure between persons of European versus African descent (Crawford, et al. 2004; Kuffner, et al. 2003; Przeworski, et al. 2000) may explain racial differences in warfarin requirements, the predictive ability of single *VKORC1* polymorphisms and *VKORC1* haplotypes has been evaluated (Limdi and Veenstra 2008; Limdi, et al. 2010b). Participants in the report by International Warfarin Pharmacogenetics Consortium (IWPC) (Limdi, et al. 2010b) were recruited from 11 countries. In total, seven *VKORC1* SNPs were studied. However, as all study sites did not assess the same SNPs, genotype information was incomplete. To ensure complete genotype information, imputation methods that perform robustly over short window sizes (i.e., only seven SNPs in one gene) are vital. Also, the choice of reference samples is vital for the quality of the imputation. Another ongoing study involves pooling genotype information (Illumina 550K versus Illumina1M duo) for African-American samples from two study sites wherein differences in genotyping density (600 K SNPs and 1.2 million SNPs, respectively) across study sites raise important questions. With the resource of about 4 million SNPs from HapMap data, can we impute our data to the same number of SNPs as HapMap data? With the large amount of genome-wide data, how can we take full advantage of high-performance parallel computing resources and conduct imputation on small chromosome regions? For the African American samples, the choice of reference is also an issue as we have to deal with in the first study. By focusing on how sample size, window-size, untyped rate and reference choice influence the quality of genotype imputation, we hope to provide guidelines and recommendations for such studies.

2. MATERIALS AND METHODS

2.1 Simulation of the study populations and reference populations

We used software HAPGEN (Spencer, et al. 2009b) to simulate study populations and reference populations, each with 1,000 individuals, based on haplotype data on chromosomes 15, 19, and 22 of unrelated individuals from two HapMap 3 cohorts, CEU (Utah residents with Northern and Western European ancestry) and ASW (African ancestry in Southwest USA) (The International HapMap Consortium 2003). Study samples were randomly drawn from the study populations. Similarly, reference samples of haplotypes were randomly drawn from the reference populations. For convenience, the first 5,000 SNPs on these chromosomes were used for analyses. The results presented in this work were mainly for chromosome 22.

2.2 The measures of imputation quality

We used accuracy and efficacy as the measures of imputation quality. IMPUTE 2 gives posterior probabilities for all three genotypes (e.g., aa, aA, AA) at each locus for each individual. In this study, we chose the confidence threshold of 0.90 for a genotype to be accepted (i.e. successfully imputed). Efficacy (call rate) (Nothnagel, et al. 2009) was calculated as the ratio of the number of successfully imputed SNPs (i.e., the SNPs whose highest posterior probabilities of the three possible genotypes are greater than or equal to 0.90) to the total number of SNPs being imputed in the study. Accuracy (Huang, et al. 2009a; Nothnagel, et al. 2009; Zhao, et al. 2008) was defined as the proportion of correctly imputed SNPs among all the successfully imputed SNPs in the study samples under investigation. In other words, efficacy measures the proportion of the imputed SNPs passing the chosen confidence threshold, thus it is similar to the genotyping call rate; accuracy measures the proportion of the correctly imputed SNPs among those that passed the confidence threshold.

A higher confidence threshold is associated with a lower efficacy and a higher accuracy. Although the choice of the threshold should depend on the study, a considerably high confidence threshold is necessary for high accuracy of imputation. In our study, the confidence threshold of 0.90 resulted in considerably high accuracy.

2.3 The factors affecting the quality of imputation

2.3.1 The untyped rate

To simulate the study samples, the SNPs were masked randomly according to the chosen untyped rates. An untyped rate higher than 70% would result in low efficacy. On the other hand, an untyped rate less than 10% may not be appealing from the cost saving point of view. Therefore the untyped rates considered in this study were 10%, 30%, 50%, and 70%.

2.3.2 Study sample size and reference sample size

To comprehensively evaluate the effects of study sample size and reference sample size on imputation quality, we conducted imputation for studies of large, medium, and small sizes, although we did not present the results in the same way. For a study of large sample size, we considered study sample sizes of 500 and 1,000, with reference sample sizes of 100, 300, 500, and 1,000. For medium studies, we considered study sample sizes of 50, 100, and 300, with reference sample sizes of 5, 10, 25, 50, 100, 250, 300, and 600. For small studies we considered study sample sizes from 1 to 15, with reference sample sizes of 5, 10, 15, and 20.

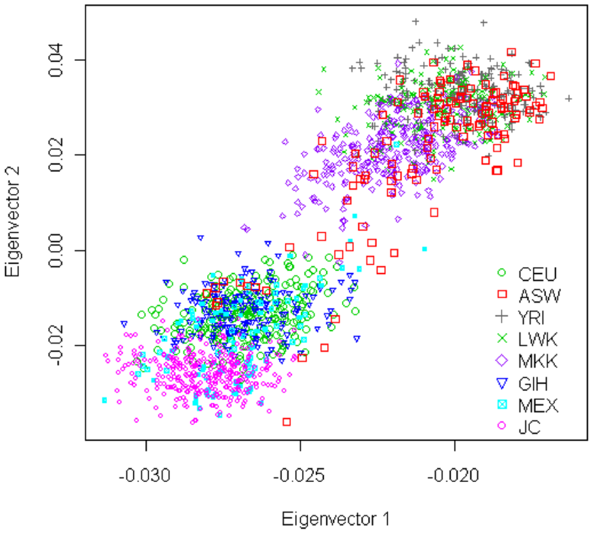

Figure 1. The distribution of haplotypes of different cohorts of HapMap 3 generated by the first two eigenvectors from SVD. The abbreviations and numbers of the haplotypes for unrelated individuals in the cohorts are: CEU (234), ASW (126), YRI (230), LWK (180), MKK (286), GIH (176), MEX (104), and JPT + CHB (JC) (340), respectively.

It is reasonable to expect that a large reference sample will improve the quality of imputation. But in some situations there may not be enough reference samples. Therefore, we included reference sample sizes as small as 10% of the study sample sizes for some combinations of scenarios, to see whether the effect of small reference sample size can be compensated by a larger study sample.

2.3.3 Reference choice for admixed study sample

The choice between the external references and internal references is critical for the quality of imputation. Internal reference refers to reference samples drawn from the same population as the study samples; external reference refers to reference samples drawn from a population other than the population from which the study samples are drawn. Publically available data, such as HapMap data (The International HapMap Consortium 2003) and the Wellcome Trust data (The Wellcome Trust Case-Control Consortium 2007), are usually used as external references. Internal references, however, usually have to be obtained for a specific study, adding extra cost when a certain number of individuals need to be genotyped at all loci of interest or genotyped with a higher density.

We chose ASW as an example of admixed study population. To choose the reference populations for the study samples from ASW, the relationship between ASW and other HapMap cohorts was examined using Singular-Value Decomposition (svd function in R (Team R Development Core 2009) was used). Figure 1 displays the locations of the haplotypes of the different cohorts along the first two principal components. The cohorts and their abbreviations are listed in the appendix. This figure shows that most of the ASW haplotypes are close to African populations (YRI, LWK, MKK), but some of them are close to the CEU and GIH cohorts. This is expected because it is well documented that ASW is an admixed population between Caucasian populations and African populations (Alexander, et al. 2009). We introduced two terms: the major source of admixture and the minor source of admixture. The former refers to the ancestry population that makes a major genomic contribution to the admixed population, and the latter refers to the ancestry population that makes a minor genomic contribution to the population. For example, in the case of ASW, we loosely designated YRI as the major source of admixture and CEU as the minor source of admixture. Intuitively, adding the cohorts genetically close to the reference panel should improve the quality of imputation; adding "irrelevant" cohorts may not improve imputation but should not compromise much of the quality of imputation. (de Bakker, et al. 2006)

We sequentially examined the external reference panel, the internal reference panel, and the internal reference panel augmented by external references to see whether this conjecture holds. Specifically, starting with external references, we chose YRI, LWK, and MKK as single reference separately, and we chose YRI + CEU, YRI + JPT + CHB, and CEU + YRI + MKK + GIH + MEX as mixed references. Parallel to external references, we examined the original ASW haplotypes in the HapMap project as internal references, followed by the augmented reference panels ASW + CEU, ASW + GIH, ASW + JPT + CHB, ASW + CEU + YRI, ASW + CEU + MKK, ASW + CEU + YRI + MKK, and ASW + CEU + YRI + MKK + GIH + MEX, respectively.

We compared internal references with external references to determine whether the effort of obtaining an internal reference is worth the gain in imputation quality. We also compared the imputation results from using original haplotypes of ASW as a reference, with those from using simulated haplotypes as references to evaluate the difference in imputation quality.

2.4 Piecewise vs. whole-region imputation

The IMPUTE 2 document recommends using chromosomal regions larger than 5 MB for an analysis for parallel computing purposes. We chose the first 5,000 SNPs on chromosomes 15, 19, and 22, regions greater than 10 MB, imputed the masked SNPs using whole-region imputation, and compared them with the ones obtained using piecewise imputation. Here, whole-region imputation runs over the entire region; piecewise imputation is implemented by breaking the whole region into smaller blocks and imputing the untyped SNPs within the blocks separately. Piecewise imputation was implemented on evenly spaced blocks of 5, 10, 20, 50, 100, 200, and 500 SNPs, and on blocks determined by a haplotype block partition scheme using the program Hapblock (Zhang, et al. 2005), to investigate how the partition of the whole region into smaller pieces affects the imputation quality.

2.5 Software, platform, and number of replicates

IMPUTE 2 was used for imputation because of its high accuracy and popularity (Nothnagel, et al. 2009; Spencer, et al. 2009b; The Wellcome Trust Case-Control Consortium 2007; Zhao, et al. 2008). The parameters for the package were set as: *buffer* 250 -*k* 70 -*iter* 40 -*burnin* 3 -*call_thresh* 0.90. The parameter *buffer* allows a buffer region (in KB) to ensure that the genotypes close to the ends of the imputation interval can be imputed with enough information and confidence. Parameter k is the number of "conditional states" for MCMC phasing updates. The default is 40. The parameter *iter* specifies the total number of iterations for MCMC. We increased these parameters to guarantee better performance. *Call_thresh* specifies the threshold for a genotype to be accepted (successfully imputed). For each individual at a specific locus, if the highest posterior probability of the three genotypes passes the threshold, IMPUTE2 uses that genotype as the imputed genotype. Otherwise the genotype is not successfully imputed. The details are available through the link listed in the appendix. IMPUTE 2 was run on a Linux-based cluster system.

Considering that the number of all combinations among untyped rate, study sample size, and reference sample size that we chose to simulate is enormously large, we generated 10 replicates for each combination scenario. For the evaluation of reference choice and comparison between piecewise imputation and whole-region imputation, 1,000 replicates were generated.

3. RESULTS

3.1 The effects of study sample size, reference sample size, and untyped rate

To investigate the effects of study sample size, reference sample size, and untyped rate on the quality of genotype imputation, we conducted simulations with 576 different combinations of these factors. For each combination we generated 10 replications, using internal reference samples. Concordant with prior report (Huang, et al. 2009a), for study sample size ranging from five to several hundreds the quality of imputation is similar for fixed reference sample sizes (data not shown). Because the effect of study sample size is negligible compared with the effect of reference sample size, the simulation results with the same reference size were averaged over different study sample sizes (Figure 2). As highlighted in Figure 2, both accuracy and efficacy increase as the reference sample size increases from 1 to 50, taper off for reference sample above 50, with little additional gains in accuracy and efficacy for reference samples over 300. Thus, for clarity, we only show the curves for the reference sample

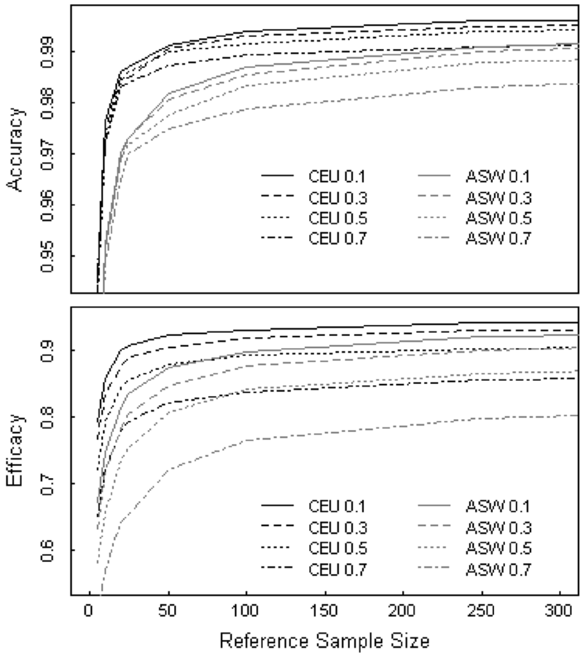

Figure 2. The mean values of accuracy and efficacy of imputation with varying reference sample size at different untyped rates for CEU and ASW. The untyped rates are 0.1, 0.3, 0.5, and 0.7, respectively. The study sample sizes vary for different points on the curves (refer to the main text).

size up to 300. The quality of imputation for ASW samples is considerably lower than that for CEU samples, with reference sample size exerting a stronger influence on the quality of imputation for the ASW study samples than for the CEU study samples.

Figure 3 shows the influence of different untyped rates on the quality of imputation with fixed study sample sizes and reference sample sizes. The quality does not change greatly when the untyped rate increases from 0.1 to 0.3, decreases modestly when the untyped rate increases from 0.3 to 0.5, but decreases considerably when the untyped rate increases from 0.5 to 0.7. For studies with a small reference sample size, both accuracy and efficacy are significantly compromised. This may be explained by the greater possibility that some subjects in the study sample may not find their matching references on some regions in a small reference sample.

3.2 The selection of reference for ASW study population

Each cell in Table 1 shows the mean values of efficacy and accuracy of imputation using the reference panel listed in the corresponding column for 1,000 random study samples, each consisting of 50 individuals randomly drawn from the ASW study population generated by HAPGEN using ASW haplotypes as the input. Untyped rate was chosen at 0.5 and 0.7.

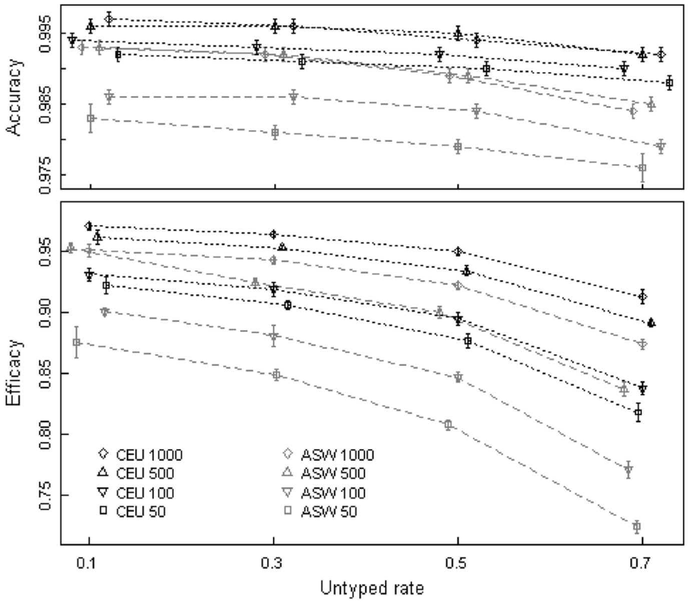

Figure 3. The effects of untyped rate on accuracy and efficacy. The points represent the accuracy and efficacy at four untyped rates: 0.1, 0.3, 0.5, and 0.7, with both study sample sizes and reference sample sizes of 50, 100, 500, and 1,000, respectively. The points with the same untyped rates are jittered for easy reading, and connected with dashed lines for clear presentation of the trends for different combinations of population and sample size.

3.2.1 External references

Based on their close relationship with ASW, each of the African cohorts LWK, MKK, and YRI was used as a single reference panel separately. We found that YRI performs the best among these three. When the cohorts YRI, MKK, GIH, MEX, and CEU were pooled together as a "cosmopolitan" reference panel (Ext-cosmo), the accuracy improves slightly compared with the one obtained from the best single-reference panel, YRI. According to Alexander et al. (Alexander, et al. 2009), ASW is an admixed population between African populations and CEU. Thus, a reference panel composed of YRI and CEU cohorts may be adequate. The results show that panel YRI + CEU (i.e., the reference panel consists of samples from YRI and CEU populations) is slightly better than all other external panels including Ext-cosmo. The reference panel YRI + JC (JPT + CHB) performs slightly worse than panel YRI + CEU, but better than YRI alone. We noticed that Ext-cosmo, as the largest reference panel in Table 1, does not outperform panel YRI + CEU.

3.2.2 Internal references augmented by external reference cohorts

The first reference panel, labeled as SIM in the first column in the lower panel of Table 1, is a reference sample

Table 1. Comparison of reference panels for quality of imputation for 50 individuals

External references									
	LWK	MKK	YRI			YRI + CEU	YRI + JC	Ext-Cosmo	
Untyped Rate 0.5									
Efficacy	0.778	0.768	0.793			0.801	0.797	0.793	
SD	0.007	0.006	0.007			0.006	0.007	0.006	
Accuracy	0.968	0.964	0.967			0.976	0.973	0.978	
SD	0.001	0.001	0.001			0.001	0.001	0.001	
Untyped Rate 0.7									
Efficacy	0.696	0.687	0.713			0.72	0.718	0.71	
SD	0.007	0.007	0.007			0.007	0.007	0.008	
Accuracy	0.961	0.957	0.96			0.97	0.965	0.972	
SD	0.001	0.001	0.002			0.001	0.001	0.001	
Internal references augmented									
	SIM	ASW	+ GIH	+ CEU	+ JC*	+ YRI + CEU	+ CEU + MKK	Cosmo	Cosmo1
Untyped Rate 0.5									
Efficacy	0.842	0.901	0.914	0.915	0.918	0.894	0.899	0.884	0.885
SD	0.006	0.005	0.005	0.005	0.005	0.005	0.005	0.006	0.006
Accuracy	0.984	0.996	0.994	0.994	0.994	0.994	0.994	0.994	0.994
SD	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Untyped Rate 0.7									
Efficacy	0.765	0.829	0.854	0.857	0.861	0.829	0.834	0.816	0.817
SD	0.008	0.007	0.006	0.006	0.006	0.007	0.007	0.008	0.008
Accuracy	0.98	0.994	0.992	0.991	0.99	0.991	0.991	0.99	0.99
SD	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Note:

1. Each of the cells in the table lists the mean or standard deviation of the quality measures (rounded to 3rd decimal) indicated in the first column for 1,000 replications for the combination of untyped rate and the reference panel.
2. The cohorts and their numbers of haplotypes (in parentheses): CEU (234), ASW (126), YRI (230), LWK (180), MKK (286), GIH (176), MEX (104), JPT + CHB (JC) (340).
3. The panel "Ext-Cosmo" is the pool of haplotyes from YRI, MKK, GIH, MEX, and CEU. The panel "SIM" was a random draw of 200 haplotypes (100 individuals) from 2000 haplotypes simulated by HAPGEN based on the original ASW haplotypes. " + " stands for the mixture of several cohorts. In the columns of the lower panel of the table, "ASW" is omitted from columns 4 to 8 because of the limited space. "Cosmo" means " + CEU + YRI + MKK + GIH + MEX", and "Cosmo1" means " + CEU + YRI + MKK."

drawn from the pseudo-population simulated by HAPGEN based on the ASW haplotypes obtained from the HapMap project. The second reference panel, labeled ASW, is the panel formed by the original ASW haplotypes. Because the ASW cohort (126 unrelated haplotypes) is not large, we wanted to augment the internal reference panel with external cohorts. Parallel to the external "cosmopolitan" panel, the internal panel was augmented by the external "cosmopolitan" panel to form a panel termed as expended panel (Cosmo), augmented by YRI + MKK + CEU to form the panel Cosmo1, etc. JC was the last reference panel used to augment the internal reference panel. All thus-formed panels were used as references to impute the randomly masked genotypes in the study samples. The imputed genotypes were compared with the actual genotypes to calculate the efficacy and accuracy of imputation for the whole region of 5,000 SNPs. Surprisingly, the largest augmented reference panel (Cosmo) performs the worst, but the panel augmented by the seemingly unrelated cohorts, JC (the column with "*" in the lower panel of Table 1), performs the best. Inclusion of fewer ASW related haplotypes for augmentation resulted in better imputation quality. Also, the simulated haplotypes did not perform as well as the original haplotypes as references. The results were similar with analyses on the first 5,000 SNPs on chromosomes 15 and 19.

3.3 Piecewise imputation vs. whole-region imputation

We evaluated four block-partitioning strategies. The first three, labeled as "even100", "even200," and "even500," are to divide the whole region into evenly spaced blocks with window size of 100 SNPs, 200 SNPs, and 500 SNPs, respectively. The fourth one defines the blocks based on haplotype block partition scheme using software HapBlock (Zhang, et al. 2005). Detailed information is freely available through the link listed in the appendix. The parameters were set as: blockpercent (the threshold for common haplotypes in block partitioning) at 0.85, and the fraction of strong pair-wise LD (D') at 0.002. With the above parameters, HapBlock partitioned the entire region into 53 blocks using the whole study population (2,000 haplotypes generated by HAPGEN from the original haplotypes). More than half of the blocks contain close to or more than 100 SNPs. Based on our prior observation that the quality of imputation for a region smaller than 100 SNPs is not very stable, the blocks with less than 100 SNPs were merged with the neighboring blocks until the new blocks contained more than 100 SNPs. The process started from the first block (SNP 1 to SNP 36) on the left and moved along the region. If a block was larger than 100 SNPs, the algorithm would simply move to the next block; if a block was smaller than 100 SNPs, the algorithm would merge the block with the block on its right until a new block larger than 100 SNPs was formed. The block partitioning strategy thus defined is labeled as "block85".

Figure 4 compares the performance of the piecewise imputation strategies even200 and even500 with that of whole-region imputation for the first 5,000 SNPs on chromosome 22. The imputed results from whole region imputation and even200 were re-grouped by the blocks that even500 used in order to have a common comparison basis among different partitioning strategies. Overall, the accuracy of whole-region imputation is better than, or at least as good as, that of piecewise imputation for most blocks for all of the strategies. The efficacy that strategy even500 achieves, however, is statistically significantly better at the level of 0.05 than the whole-region imputation for almost all of the blocks and scenarios along the whole region. The only exception is the block [2001, 2500] in the scenario of untyped rate 0.7 and reference sample size of 100, where the whole-region method is slightly better with a difference of 0.001 (p-value 0.26) (refer to the fifth point in the second graph on the top row in Figure 4). For the efficacy of imputation quality even200 performs slightly better the even500 in a few blocks when the reference sample is large, but performs worse than even500 in most of the blocks. However, it performs better than the whole region method for most of the intervals we investigated in terms of efficacy. Even100 and block85 also outperform whole region strategy in several blocks in terms of efficacy. For the legibility of the graphs, the results from these two strategies are not shown in Figure 4. We confirmed these comparison results on chromosome 19.

For strategy block85, we generated 100 replicates. Table 2 compares the means of accuracy and efficacy and their standard deviations of 100 simulations from piecewise method with those calculated based on the results obtained from whole-region method but regrouped for blocks partitioned using block85, for an untyped rate of 0.5, study sample size of 50 and reference sample size of 500. A bolded number indicates that one strategy outperforms the other for the measure by 1% in the block. We conducted *t*-test for the differences greater than 1%, finding that they are all significant at 0.05 except for the block [2804, 2925] which has a statistic value 1.57 (p-value 0.058). The difference in efficacy is not uniform among the blocks. In the underlined blocks, block85 outperforms the whole-region imputation by more than 4% in efficacy. *t*-test also shows that all of the differences in accuracy are highly significant, with much smaller standard deviations of accuracy than those of efficacy.

4. DISCUSSION

In this study, we evaluated several factors affecting the quality of genotype imputation through simulation based on real data. We examined different combinations of study sample size, reference sample size, and untyped rate for samples from CEU and ASW populations; we compared among external reference panels and internal reference panels augmented by different cohorts; finally we compared piecewise imputation with whole-region imputation for further divisibility of the region unit for imputation that IMPUTE recommends, and feasibility of imputation for studies focusing on short chromosomal regions.

Our results show that study sample size has little effect on the quality of genotype imputation. This is consistent with the report of conditional independency of imputation on study sample size given the reference panel (Huang, et al. 2009a). For the imputation of untyped loci, this conditional independence has a more intuitive explanation: none of the individuals are typed at the same set of loci, which means that there is no information to borrow from each other and a large study sample size cannot compensate much for a small reference sample size. This conclusion relieves the concern about the poor imputation quality for studies with small study sample sizes.

In Figure 3, the points at four chosen untyped rates are connected by straight lines to find patterns of the quality measures. We observed that the accuracy and efficacy of imputation are much lower at the untyped rate 0.7 than that at 0.5. There is a possibility that the decrease in genotype imputation quality could have occurred at any point larger than 0.3 if the decrease is not linear as shown in Figure 3. However, for practical consideration we recommend an untyped rate less than 0.5.

An appropriate reference sample size is crucial for the quality of imputation. Surely, the greater the reference sample size, the better the quality. The steep slopes of the curves to the left of reference sample size 50 in Figure 2 suggest that for both relatively homogeneous populations, such as CEU, and admixed populations, such as ASW, a reference panel size greater than 50 is necessary for quality of imputation, but a panel size greater than 300 may not be cost

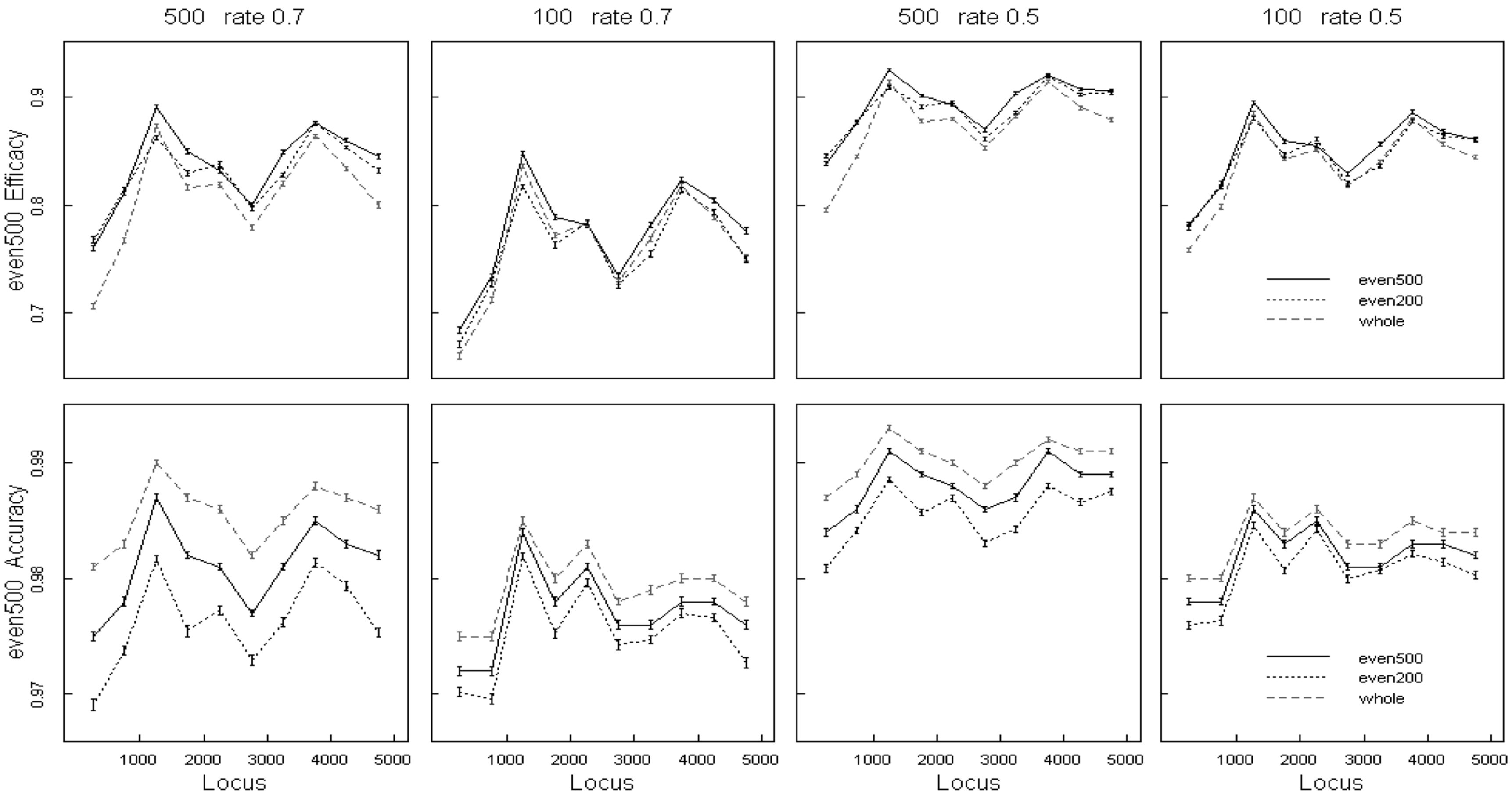

Figure 4. The mean values and their standard error bars of efficacy and accuracy achieved by piecewise imputation and whole-region imputation for the first 5,000 SNPs on chromosome 22 for the scenarios with different untyped rates (0.5 or 0.7) and reference sample sizes (100 and 500). The imputed results from the whole-region method and even200 were re-grouped by the blocks that even500 used in order to have a common comparison basis. The curves are generated by connecting the mean values for the blocks of 1,000 replications for even500 and even200.

Table 2. Comparison between Block85 and whole-region imputation

		Block85				Whole-region			
Start	End	Accuracy	SD	Efficacy	SD	Accuracy	SD	Efficacy	SD
1	140	0.981	0.006	**0.874**	0.029	0.987	0.004	0.836	0.034
141	284	0.987	0.005	**0.885**	0.026	0.993	0.003	0.865	0.026
285	397	0.979	0.007	**0.815**	0.036	0.986	0.005	0.736	0.04
398	520	0.97	0.008	**0.803**	0.034	**0.98**	0.006	0.729	0.039
521	729	0.984	0.005	**0.894**	0.02	0.989	0.003	0.872	0.021
730	899	0.986	0.005	**0.887**	0.025	0.991	0.003	0.861	0.026
900	1126	0.984	0.004	**0.865**	0.025	0.988	0.003	0.843	0.026
1127	1373	0.994	0.002	0.964	0.012	0.995	0.002	0.966	0.009
1374	1507	0.976	0.007	**0.822**	0.036	**0.986**	0.005	0.798	0.036
1508	1618	0.981	0.006	**0.894**	0.029	0.99	0.004	0.877	0.03
1619	1801	0.991	0.004	0.944	0.016	0.994	0.003	0.941	0.014
1802	2033	0.983	0.004	**0.864**	0.027	0.988	0.003	0.845	0.022
2034	2254	0.988	0.003	**0.904**	0.023	0.991	0.003	0.89	0.023
2255	2375	0.979	0.007	**0.857**	0.026	0.986	0.004	0.807	0.033
2376	2621	0.992	0.002	0.937	0.015	0.994	0.002	0.93	0.015
2622	2803	0.986	0.004	**0.893**	0.019	0.99	0.003	0.876	0.02
2804	2925	0.963	0.009	**0.725**	0.046	0.975	0.007	0.711	0.043
2926	3030	0.983	0.006	0.883	0.027	0.989	0.004	0.891	0.023
3031	3174	0.989	0.004	0.931	0.017	0.992	0.003	0.924	0.018
3175	3285	0.972	0.008	**0.814**	0.039	0.981	0.006	0.758	0.042
3286	3385	0.984	0.005	**0.891**	0.039	0.991	0.004	0.877	0.034
3386	3522	0.988	0.004	0.932	0.02	0.991	0.003	0.929	0.02
3523	3774	0.993	0.002	0.953	0.013	0.995	0.002	0.957	0.009
3775	3940	0.983	0.005	**0.89**	0.021	0.989	0.004	0.863	0.025
3941	4179	0.987	0.004	**0.904**	0.017	0.992	0.002	0.892	0.019
4180	4312	0.977	0.006	**0.838**	0.03	0.985	0.005	0.825	0.031
4313	4470	0.989	0.005	0.924	0.018	0.993	0.004	0.923	0.017
4471	4629	0.984	0.005	**0.9**	0.024	0.99	0.003	0.885	0.023
4630	4821	0.99	0.003	**0.92**	0.021	0.993	0.002	0.908	0.02
4822	5000	0.986	0.004	**0.896**	0.026	0.991	0.003	0.864	0.03

The means of accuracy and efficacy and their standard deviations of 100 simulations for blocks determined using the block85 and whole-region methods, with an untyped rate of 0.5, study sample size of 50, and reference sample size of 500. The first two columns are serial numbers of the start and end SNPs for blocks formed by block85 partitioning scheme. The bolded numbers indicate that one strategy outperforms the other for the measure by 1%. The underlined blocks are the ones in which block85 outperforms whole-region imputation by more than 4% in efficacy.

effective. However, Figure 2 indicates that reference sample size affects the imputation quality differentially for CEU and ASW. This may be due to smaller haplotype block sizes in African populations (Zhao, et al. 2003). Smaller block sizes implies more unique haplotype blocks in a region of a chromosome. Therefore, a larger reference sample is necessary for a larger number of distinct haplotypes. Recent admixture history of ASW between African populations and other ethnic populations also contributed to this ongoing process, leading to lower quality of imputation for ASW given the same reference sample size. In general, we recommend a reference sample size greater than 100 individuals for homogeneous populations, and 200 individuals for admixed populations, if possible.

Reference choice among available cohorts is another decision researchers have to make. Table 1 demonstrates that some prior information on the admixture of the study population is beneficial for the quality of imputation. The prior information should be used differentially depending on whether internal references are available. If a study does not have a budget for internal references, and there are no such existing sources, an external reference panel is the only choice. If possible, only the cohorts from the populations that contribute to the admixture of the study population should be included, although a "cosmopolitan" panel does not compromise much the quality of imputation. When internal references are available, however, the augmentation with external references should be careful.

Researchers may wrestle over whether internal references are necessary when planning the study budget. Based on our findings, the use of internal references provides a significant gain in quality that is worth the cost and effort. Further, a small internal reference panel can be augmented with the existing external cohorts. The choice of the external cohort for such augmentation is important for the quality of imputation. Our results demonstrate that compared with the external-reference-only panel, augmenting an internal reference panel with a cosmopolitan external panel can considerably lower the quality of imputation. More specifically, augmenting the reference panel with cohorts related to the study population is harmful. This phenomenon can be explained by the fact that the population represented by the reference panel may shift considerably from the internal reference population on inclusion of external references. For example, as more African cohorts, such as LWK, MKK, and YRI, are added to the reference panel it begins to more closely resemble the African population formed by LWK, MKK, and YRI. Consequently it will no longer serve as a good reference panel for the admixed ASW study population (formed by African populations and Caucasian populations). This led us to the other side of the spectrum of the seeming relationship among the cohorts. To our surprise, the panel augmented by JPT + CHB (JC) performs the best. This suggests that although ASW and JC are seemingly unrelated, they still share regions on chromosomes. Thus, the seemingly unrelated cohorts of appropriate sizes can provide additional reference information for the study sample from ASW, yet the augmented panel is not significantly different from the internal reference population ASW.

Zhao et al. (Zhao, et al. 2008) reported that the principal component-clustering method improved the quality of imputation greatly for the subgroup of the African American cohort close to CEU but did not improve the quality as much for the rest of the cohorts. The proportion of the subgroup close to CEU should be small, thus the overall improvement of the imputation quality for the African American cohorts would not be significant. By using an internal reference panel, however, the efficacy improved by 10% over the one achieved by using YRI + CEU as the reference panel for the whole study sample, and improved further by appropriate augmentation with external references, with the final efficacy of 0.918 for an untyped rate of 0.5 and 0.861 for an untyped rate of 0.7. These figures are similar to those reported for CEU cohorts (Figure 2). The internal reference also improves the accuracy by 2%. The best augmented reference panel performs slightly worse than the internal reference panel, but the magnitude of the difference is negligible compared with the gain in the efficacy. We expect that for the combinations with the untyped rate less than 0.5 the gain from internal reference and augmentation will be worth the cost. Therefore, we recommend augmenting the internal reference panel with unrelated cohorts. Although these results were obtained from an ASW population, the augmentation with unrelated cohorts may apply to other admixed populations and some homogeneous populations, such as African populations, which are known to be difficult to impute (Huang, et al. 2009a).

As shown in Table 1 the haplotypes simulated by HAPGEN based on the original haplotypes did not perform as well as the original haplotypes when used as the references. This may be due to the simplification of the underlying genetic model adopted by HAPGEN. The simplification may also affect the simulated genotypes and haplotypes differentially in terms of quality because haplotypes entail more genetic information than genotypes. This discrepancy in quality may also contribute to the worse performance of the simulated haplotypes as references.

In reality, some genetic studies may focus on short chromosomal regions, such as candidate gene studies. In addition to the block partitioning strategies presented in Figure 4, we evaluated window sizes as small as 5 SNPs to 100 SNPs. The imputation quality was not reliable for regions smaller than 50 SNPs (data not shown). Even100 does not perform as well as whole-region imputation except in some scenarios. Therefore we only present the results from using window size larger than 100 SNPs. Overall, even200 did not perform as well as even500 in terms of efficacy for most blocks and combinations of untyped rate and reference sample size. It uniformly performs worse than even500 and whole-region strategy in terms of accuracy.

The strategy block85 partitions the whole region into 30 blocks, more than the 25 blocks used by even200. The accuracy and efficacy achieved with block85 were higher than the ones achieved by even200 for most combinations of untyped rate and reference sample size (data not shown). As shown in Table 2, block85 outperforms the whole-region strategy by more than 4% in efficacy in some of the blocks. These results suggest that block-partitioning strategies based on haplotype blocks are promising for piecewise imputation. We investigated a limited number of combinations among the input parameters of the software HapBlock in this study. Further research may be needed to optimize the partition to further improve the imputation quality along the region of interest.

Our results show that the efficacy achieved by even500 is significantly better than the one achieved by whole-region imputation. Based on our results and extrapolation, a practical implication of these differences is that the strategy even500 may provide 2,000 to 8,000 more imputed SNPs than whole-region imputation for downstream analyses when we impute a 500,000-SNP panel using a 1,000,000-SNP panel as the reference. The differences in accuracy are quite small (0.004), although the mean differences of accuracy are also statistically significant. Thus, we conclude that the accuracy achieved by even500 is stable and acceptable compared with the one achieved by the whole-region strategy. For the general practice, therefore, we recommend even window sizes of 500 SNPs, with the best tradeoff between efficacy and accuracy among the strategies we investigated. A block of 500 SNPs is roughly equivalent to 1 MB, with one SNP every 2 KB (Gabriel, et al. 2002), a window size much smaller than 5 MB recommended by IMPUTE. For the researchers who need to focus on a shorter region of chromosome than this size, a window larger than 50 SNPs is highly recommended.

Overall, studies with a small study sample size and/or short (but not too short) chromosomal region are feasible in terms of the quality of genotype imputation. For studies of large chromosome regions, the efficacy gain and time saving through parallel computing indicates that the imputation of SNPs along chromosomes can be divided into pieces smaller than what IMPUTE recommends and can be conducted independently.

Our study mainly focused on the first 5,000 SNPs on chromosome 22. Results may differ on the rest of the chromosome and on other chromosomes. For this reason, we conducted the same simulations on chromosomes 15 and 19 and confirmed the results. However, we still advise that the results be used conservatively. For example, a study may increase the window size over 500 SNPs for piecewise imputation if high accuracy is desirable, and choose an untyped rate as small as the budget permits. In our study, we used a large number of replications for evaluation of piecewise imputation and reference choice. For other parts of the study, the numbers of the replications are not very large because of the large amount of computation. But merging of the data from combinations of simulation scenarios during the analyses made the amount of usable data much greater, resulting in the fairly small standard errors of the measures, which gave confidence about our results.

Another limitation is that the study was mainly conducted through simulation based on real data with a high density of SNPs. If a study population is not represented by one of the cohorts in the HapMap project, the internal reference panel that the study creates will be likely less dense than the ones in the HapMap project. In that case, we recommend using low untyped rates, large reference sample size, and appropriate augmentation. Hopefully, with the completion of the 1,000 Genome Project, Human Genome Diversity Project (HGDP) (Sanna, et al. 2008), studies will have more appropriate choices for reference.

The results obtained in this study will benefit the two studies we mentioned at the beginning. For the study of the seven *VKORC1* SNPs, we will try to retrieve as many SNPs of expanded chromosomal regions as possible and impute the missing SNPs using a window of more than 500 SNPs to achieve better imputation quality than the one using only the seven SNPs of interest. For the second study, we will impute the SNPs in HapMap data but not on the Illumina 550K/1M arrays. A rough check of these arrays against HapMapIII data gives an estimate of untyped rate between 0.5 and 0.7. If we use the ASW cohort in HapMap data as the internal reference augmented with JC, we will be able to impute 85% to 90% of the untyped SNPs. With the SNPs already genotyped in the arrays, we will have roughly 90% to 95% of the SNPs in the HapMap data with fairly good accuracy for subsequent analyses. We will also take advantage of the results on piecewise imputation to greatly facilitate the genome-wide imputation.

Imputation is very appealing in genetic studies. If used appropriately, imputation may greatly save time and money, and increase power. Our results may be helpful for the design and analysis of genetic studies. The design of cost-effective studies deserves further investigation.

APPENDIX

Populations and corresponding abbreviations (http://www.sanger.ac.uk/humgen/hapmap3/)

- ASW African ancestry in Southwest USA
- CEU Utah, USA residents with Northern and Western European ancestry from the CEPH collection
- CHB Han Chinese in Beijing, China
- CHD Chinese in metropolitan Denver, Colorado, USA
- GIH Gujarati Indians in Houston, Texas, USA
- JPT Japanese in Tokyo, Japan
- LWK Luhya in Webuye, Kenya
- MEX Mexican ancestry in Los Angeles, California, USA
- MKK Maasai in Kinyawa, Kenya
- TSI Toscani in Italy
- YRI Yoruba in Ibadan, Nigeria

Links to software packages:

IMPUTE 2: https://mathgen.stats.ox.ac.uk/impute/impute_v2.html.

HapBlock: http://www.soph.uab.edu/Statgenetics/People/KZhang/HapBlock/hapblock-index.html.

HAPGEN: http://www.well.ox.ac.uk/~zhan/hapgen/hapgen2.html.

ACKNOWLEDGEMENTS

We thank David B. Allison for his support of our work. This work was supported in part by grants GM081488 (NL), R00 RR024163 (DZ), GM074913 (KZ), GM073766 (GG), R01HL092173 (NAL), and K23NS45598 (NAL) from the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank Dr. Guo-Bo Chen for his suggestions for the tables and figures. We are very grateful to the reviewers and editors for their detailed comments and valuable suggestions, which have led to considerable improvement of our work.

Received 1 June 2010

REFERENCES

ALEXANDER, D. H., NOVEMBRE, J. and LANGE, K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**(9):1655–1664.

ALTSHULER, D., GIBBS, R., PELTONEN, L., DERMITZAKIS, E., SCHAFFNER, S., YU, F., BONNEN, P., DE BAKKER, P., DELOUKAS, P., GABRIEL, S., et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**(7311):52–58.

ANDERSON, C., PETTERSSON, F., BARRETT, J., ZHUANG, J. and RAGOUSSIS, J. 2008a. Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *Am. J. Hum. Genet.* **83**:112–119.

ANDERSON, C., PETTERSSON, F., BARRETT, J., ZHUANG, J. and RAGOUSSIS, J. 2008b. Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *Am. J. Hum. Genet.* **83**:112–119.

Aquilante, C., Langaee, T., Lopez, L., Yarandi, H., Tromberg, J., Mohuczy, D., Gaston, K., Waddell, C., Chirico, M. and Johnson, J. 2006. Influence of coagulation factor, vitamin K epoxide reductase complex subunit 1, and cytochrome P450 2C9 gene polymorphisms on warfarin dose requirements. *Clin. Pharmacol. Ther.* **79**(4):291–302.

Aulchenko, Y., Struchalin, M. and van Duijn, C. 2010. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* **11**:134.

Borgiani, P., Ciccacci, C., Forte, V., Romano, S., Federici, G. and Novelli, G. 2007. Allelic variants in the CYP2C9 and VKORC1 loci and interindividual variability in the anticoagulant dose effect of warfarin in Italians. *Pharmacogenomics* **8**(11):1545–1550.

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

Browning, S. R. and Browning, B. L. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**(5):1084–1097.

Budnitz, D., Shehab, N., Kegler, S. and Richards, C. 2007. Medication use leading to emergency department visits for adverse drug events in older adults. *Ann. Intern. Med.* **147**(11):755–765.

Caldwell, M., Berg, R., Zhang, K., Glurich, I., Schmelzer, J., Yale, S., Vidaillet, H. and Burmester, J. 2007. Evaluation of genetic factors for warfarin dose prediction. *Clin. Med. Res.* **5**(1):8–16.

Carlquist, J., Horne, B., Mower, C., Park, J., Huntinghouse, J., McKinney, J., Muhlestein, J. and Anderson, J. 2010. An evaluation of nine genetic variants related to metabolism and mechanism of action of warfarin as applied to stable dose prediction. *J. Thromb. Thrombolysis* **30**(3):358–364.

Cho, H., Sohn, K., Park, H., Lee, K., Choi, B., Kim, S., Kim, J., On, Y., Chun, M., Kim, H., et al. 2007. Factors affecting the interindividual variability of warfarin dose requirement in adult Korean patients. *Pharmacogenomics* **8**(4):329–337.

Consortium TWTC-C. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**:661–678.

Crawford, D., Carlson, C., Rieder, M., Carrington, D., Yi, Q., Smith, J., Eberle, M., Kruglyak, L. and Nickerson, D. 2004. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am. J. Hum. Genet.* **74**(4):610–622.

Crawford, D., Ritchie, M. and Rieder, M. 2007. Identifying the genotype behind the phenotype: a role model found in VKORC1 and its association with warfarin dosing. *Pharmacogenomics* **8**(5):487–496.

D'Andrea, G., D'Ambrosio, R., Di Perna, P., Chetta, M., Santacroce, R., Brancaccio, V., Grandone, E. and Margaglione, M. 2005. A polymorphism in the VKORC1 gene is associated with an interindividual variability in the dose-anticoagulant effect of warfarin. *Blood* **105**(2):645–649.

de Bakker, P. I., Burtt, N. P., Graham, R. R., Guiducci, C., Yelensky, R., Drake, J. A., Bersaglieri, T., Penney, K. L., Butler, J., Young, S., et al. 2006. Transferability of tag SNPs in genetic association studies in multiple populations. *Nat. Genet.* **38**(11):1298–1303.

Furuya, H., Fernandez-Salguero, P., Gregory, W., Taber, H., Steward, A., Gonzalez, F. and Idle, J. 1995. Genetic polymorphism of CYP2C9 and its effect on warfarin maintenance dose requirement in patients undergoing anticoagulation therapy. *Pharmacogenetics* **5**(6):389–392.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. 2002. The structure of haplotype blocks in the human genome. *Science* **296**(5576):2225–2229.

Gage, B., Eby, C., Johnson, J., Deych, E., Rieder, M., Ridker, P., Milligan, P., Grice, G., Lenzini, P., Rettie, A., et al. 2008. Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin. *Clin. Pharmacol. Ther.* **84**(3):326–331.

Hao, K., Chudin, E., McElwee, J. and Schadt, E. E. 2009. Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. *BMC Genet.* **10**:27.

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

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

Huang, L., Wang, C. and Rosenberg, N. A. 2009c. The relationship between imputation error and statistical power in genetic association studies in diverse populations. *Am. J. Hum. Genet.* **85**(5):692–698.

Klein, T., Altman, R., Eriksson, N., Gage, B., Kimmel, S., Lee, M., Limdi, N., Page, D., Roden, D., Wagner, M., et al. 2009. Estimation of the warfarin dose with clinical and pharmacogenetic data. *N. Engl. J. Med.* **360**(8):753–764.

Kuffner, T., Whitworth, W., Jairam, M. and McNicholl, J. 2003. HLA class II and TNF genes in African Americans from the Southeastern United States: regional differences in allele frequencies. *Hum. Immunol.* **64**(6):639–647.

Li, Y., Willer, C., Ding, J., Scheet, P. and Abecasis, G. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* **34**(8):816–834.

Li, Y., Willer, C., Sanna, S. and Abecasis, G. 2009. Genotype imputation. *Annual Review Genomics and Human Genetics* **10**:387–406.

Limdi, N., Arnett, D., Goldstein, J., Beasley, T., McGwin, G., Adler, B. and Acton, R. 2008a. Influence of CYP2C9 and VKORC1 on warfarin dose, anticoagulation attainment and maintenance among European-Americans and African-Americans. *Pharmacogenomics* **9**(5):511–526.

Limdi, N., Beasley, T., Crowley, M., Goldstein, J., Rieder, M., Flockhart, D., Arnett, D., Acton, R. and Liu, N. 2008b. VKORC1 polymorphisms, haplotypes and haplotype groups on warfarin dose among African-Americans and European-Americans. *Pharmacogenomics* **9**(10):1445–1458.

Limdi, N., Limdi, M., Cavallari, L., Anderson, A., Crowley, M., Baird, M., Allon, M. and Beasley, T. 2010a. Warfarin dosing in patients with impaired Kidney function. *Am. J. Kidney Dis.*

Limdi, N. and Veenstra, D. 2008. Warfarin pharmacogenetics. *Pharmacotherapy* **28**(9):1084–1097. MR2461922

Limdi, N., Wadelius, M., Cavallari, L., Eriksson, N., Crawford, D., Lee, M., Chen, C., Motsinger-Reif, A., Sagreiya, H., Liu, N., et al. 2010b. Warfarin pharmacogenetics: A single VKORC1 polymorphism is predictive of dose across 3 racial groups. *Blood* **115**(18):3827–3834.

Limdi, N., Wiener, H., Goldstein, J., Acton, R. and Beasley, T. 2009. Influence of CYP2C9 and VKORC1 on warfarin response during initiation of therapy. *Blood Cells Mol. Dis.* **43**(1):119–128.

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

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

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

Momary, K., Shapiro, N., Viana, M., Nutescu, E., Helgason, C. and Cavallari, L. 2007. Factors influencing warfarin dose requirements in African-Americans. *Pharmacogenomics* **8**(11):1535–1544.

Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M. and Franke, A. 2009. A comprehensive evaluation of SNP genotype imputation. *Hum. Genet.* **125**(2):163–171.

Pei, Y., Li, J., Zhang, L., Papasian, C. and Deng, H. 2008. Analyses and comparison of accuracy of different genotype imputation methods. *PLoS One* **3**(10):e3551.

Pei, Y., Zhang, L., Li, J. and Deng, H. 2010. Analyses and comparison of imputation-based association methods. *PLoS One* **5**(5):e10827.

Przeworski, M., Hudson, R. and Di Rienzo, A. 2000. Adjusting the focus on human variation. *Trends Genet.* **16**(7):296–302.

Rieder, M., Reiner, A. and Rettie, A. 2007. Gamma-glutamyl carboxylase (GGCX) tagSNPs have limited utility for predicting warfarin maintenance dose. *J. Thromb. Haemost.* **5**(11):2227–2234.

Sanna, S., Jackson, A. U., Nagaraja, R., Willer, C. J., Chen, W. M., Bonnycastle, L. L., Shen, H., Timpson, N., Lettre, G., Usala, G., et al. 2008. Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat. Genet.* **40**(2):198–203.

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

Schelleman, H., Chen, Z., Kealey, C., Whitehead, A., Christie, J., Price, M., Brensinger, C., Newcomb, C., Thorn, C., Samaha, F., et al. 2007. Warfarin response and vitamin K epoxide reductase complex 1 in African Americans and Caucasians. *Clin. Pharmacol. Ther.* **81**(5):742–747.

Servin, B. and Stephens, M. 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics* 3.

Shriner, D., Adeyemo, A., Chen, G. and Rotimi, C. N. 2010. Practical considerations for imputation of untyped markers in admixed populations. *Genet. Epidemiol.* **34**(3):258–265.

Spencer, C. C. A., Su, Z., Donnelly, P. and Marchini, J. 2009a. Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics* **5**(5):1–13.

Spencer, C. C. A., Su, Z., Donnelly, P. and Marchini, J. 2009b. Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics* **5**(5):1–13.

Takahashi, H., Wilkinson, G., Nutescu, E., Morita, T., Ritchie, M., Scordo, M., Pengo, V., Barban, M., Padrini, R., Ieiri, I., et al. 2006. Different contributions of polymorphisms in VKORC1 and CYP2C9 to intra- and inter-population differences in maintenance dose of warfarin in Japanese, Caucasians and African-Americans. *Pharmacogenet. Genomics* **16**(2):101–110.

Team R Development Core. 2009. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria.

The International HapMap Consortium. 2003. The international HapMap project. *Nature* **426**:789–796.

The 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.

Wadelius, M., Chen, L., Eriksson, N., Bumpstead, S., Ghori, J., Wadelius, C., Bentley, D., McGinnis, R. and Deloukas, P. 2007. Association of warfarin dose with genes involved in its action and metabolism. *Hum. Genet.* **121**(1):23–34.

Wadelius, M., Chen, L., Lindh, J., Eriksson, N., Ghori, M., Bumpstead, S., Holm, L., McGinnis, R., Rane, A. and Deloukas, P. 2009. The largest prospective warfarin-treated cohort supports genetic forecasting. *Blood* **113**(4):784–792.

Zeggini, E., Scott, L., Saxena, R., Voight, B. and Marchini, J. 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.

Zhang, K., Qin, Z., Chen, T., Liu, J.S., Waterman, M.S. and Sun, F. 2005. HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics* **21**(1):131–134.

Zhao, H., Pfeiffer, R. and Gail, M. H. 2003. Haplotype analysis in population genetics and association studies. *Pharmacogenomics* **4**(2):171–178.

Zhao, Z., Timofeev, N., Hartley, S. W., Chui, D. H., Fucharoen, S., Perls, T. T., Steinberg, M. H., Baldwin, C. T. and Sebastiani, P. 2008. Imputation of missing genotypes: an empirical evaluation of IMPUTE. *BMC Genet.* **9**:85.

Zhu, Y., Shennan, M., Reynolds, K., Johnson, N., Herrnberger, M., Valdes, R. J. and Linder, M. 2007. Estimation of warfarin maintenance dose based on VKORC1 (-1639 G>A) and CYP2C9 genotypes. *Clin. Chem.* **53**(7):1199–1205.

Boshao Zhang
Department of Biostatistics
School of Public Health
University of Alabama at Birmingham
Birmingham, AL 35294
USA

Degui Zhi
Department of Biostatistics
School of Public Health
University of Alabama at Birmingham
Birmingham, AL 35294
USA

Kui Zhang
Department of Biostatistics
School of Public Health
University of Alabama at Birmingham
Birmingham, AL 35294
USA

Guimin Gao
Department of Biostatistics
School of Medicine
Virginia Commonwealth University
Richmond, VA 23298-0032
USA

Nita A. Limdi
Department of Neurology
University of Alabama at Birmingham, AL
USA

Department of Epidemiology
University of Alabama at Birmingham
1719 6th Avenue South, CIRC-312
Birmingham, AL 35294
USA

Nianjun Liu
Department of Biostatistics
School of Public Health
University of Alabama at Birmingham
Birmingham, AL 35294
USA
E-mail address: nliu@ms.soph.uab.edu