Genetic Epidemiology 36 : 107–117 (2012)

Genotype Imputation of Metabochip SNPs Using a Study-Specific Reference Panel of ∼4,000 Haplotypes in African Americans From the Women's Health Initiative

Eric Yi Liu,[1] Steven Buyske,[2,3] Aaron K. Aragaki,[4] Ulrike Peters,[4] Eric Boerwinkle,[5] Chris Carlson,[4] Cara Carty,[4] Dana C. Crawford,[6] Jeff Haessler,[4] Lucia A. Hindorff,[7] Loic Le Marchand,[8] Teri A. Manolio,[7] Tara Matise,[3] Wei Wang,[1] Charles Kooperberg,[4] Kari E. North,[9] and Yun Li[10,11]*

[1]*Department of Computer Science, University of North Carolina, Chapel Hill, North Carolina*
[2]*Department of Statistics, Rutgers University, Piscataway, New Jersey*
[3]*Department of Genetics, Rutgers University, Piscataway, New Jersey*
[4]*Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington*
[5]*Human Genetics Center, University of Texas Health Science Center at Houston*
[6]*Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University, Nashville, Tennessee*
[7]*Office of Population Genomics, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland*
[8]*Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii*
[9]*Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina*
[10]*Department of Genetics, University of North Carolina, Chapel Hill, North Carolina*
[11]*Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina*

Genetic imputation has become standard practice in modern genetic studies. However, several important issues have not been adequately addressed including the utility of study-specific reference, performance in admixed populations, and quality for less common (minor allele frequency [MAF] 0.005–0.05) and rare (MAF < 0.005) variants. These issues only recently became addressable with genome-wide association studies (GWAS) follow-up studies using dense genotyping or sequencing in large samples of non-European individuals. In this work, we constructed a study-specific reference panel of 3,924 haplotypes using African Americans in the Women's Health Initiative (WHI) genotyped on both the Metabochip and the Affymetrix 6.0 GWAS platform. We used this reference panel to impute into 6,459 WHI SNP Health Association Resource (SHARe) study subjects with only GWAS genotypes. Our analysis confirmed the imputation quality metric Rsq (estimated r^2, specific to each SNP) as an effective post-imputation filter. We recommend different Rsq thresholds for different MAF categories such that the average (across SNPs) Rsq is above the desired dosage r^2 (squared Pearson correlation between imputed and experimental genotypes). With a desired dosage r^2 of 80%, 99.9% (97.5%, 83.6%, 52.0%, 20.5%) of SNPs with MAF > 0.05 (0.03–0.05, 0.01–0.03, 0.005–0.01, and 0.001–0.005) passed the post-imputation filter. The average dosage r^2 for these SNPs is 94.7%, 92.1%, 89.0%, 83.1%, and 79.7%, respectively. These results suggest that for African Americans imputation of Metabochip SNPs from GWAS data, including low frequency SNPs with MAF 0.005–0.05, is feasible and worthwhile for power increase in downstream association analysis provided a sizable reference panel is available. *Genet. Epidemiol.* 36:107–117, 2012.

Key words: genotype imputation; Metabochip; internal reference; African Americans; rare variants

Additional Supporting Information may be found in the online version of this article.
*Correspondence to: Yun Li, Department of Genetics, Campus Box 7264, University of North Carolina, Chapel Hill, NC 27599. E-mail: yunli@med.unc.edu
Received 20 September 2011; Revised 3 November 2011; Accepted 16 November 2011
Published online 17 January 2012 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.21603

INTRODUCTION

Genotype imputation has become standard practice to increase genome coverage and improve power in genome-wide association studies (GWAS) and meta-analysis [de Bakker et al., 2008; Li et al., 2009; Marchini and Howie, 2010]. The wealth of literature using genotype imputation has focused on using external reference panels (for example, phased haplotypes from the International HapMap Project [The International HapMap Consortium, 2007] or the 1000 Genomes Project [The 2010]), largely in individuals of European ancestry, for inference of genotypes at common (minor allele frequency [MAF] > 0.05) genetic markers.

GWAS have identified more than 4,300 genetic variants associated with human diseases and traits (http://www.genome.gov/gwastudies/) [Hindorff et al., 2009]. Investigators across the world have begun efforts to

fine map within regions where GWAS-identified SNPs reside, through dense genotyping (e.g., using region-centric or gene-centric chips such as the Metabochip for metabolic-related traits [http://www.sph.umich.edu/csg/kang/MetaboChip/], or the ITMAT-Broad-CARe [IBC] for cardiovascular related traits, or the immunochip for immune related diseases) or sequencing. Furthermore, multiethnic genetic association studies have been recognized as potentially more powerful for both gene discovery and fine mapping [McCarthy et al., 2008; Pulit et al., 2010; Rosenberg et al., 2010; Teo et al., 2010] and some initial efforts have been carried out [He et al., 2011; Keebler et al., 2010; Lanktree et al., 2009; Lettre et al., 2011; Smith et al., 2011; Waters et al., 2009]. In addition, because GWAS-identified SNPs (mostly common) explain only a small proportion of overall heritability for most complex diseases and traits [Eichler et al., 2010; Maher, 2008; Manolio et al., 2009], whole-genome or whole-exome sequencing for rare SNPs and genetic variants other than SNPs (e.g., copy number variations, structural variants) are under way.

So far, there has been relatively little research on the performance of genotype imputation in this new context. Our study provides a typical scenario where we had genotyped 8,421 African Americans from the Women's Health Initiative (WHI) [Anderson et al., 1998] SNP Health Association Resource (SHARe) using the Affymetrix 6.0 genotyping platform. In an attempt to generalize genetic effects across racial groups, the Population Architecture using Genomics and Epidemiology (PAGE) consortium genotyped a subset of 1,962 African American WHI participants with data on multiple metabolic-related phenotypes using the Metabochip [Matise et al., 2011]. To increase our power to detect moderate to small genetic effects, we sought to impute the Metabochip SNPs in the remaining 6,459 individuals in WHI SHARe with Affymetrix 6.0 data only. Imputing SNPs in the fine-mapping region tends to be more challenging because these SNPs tend to be rare and in low linkage disequilibrium (LD) with GWAS SNPs. Here, we describe a pipeline for constructing study-specific reference panels using individuals genotyped or sequenced at a larger set of genetic markers (in our case, individuals genotyped using both Affymetrix 6.0 and Metabochip) and for imputation into individuals with genotype data at a subset of markers (in our case, individuals genotyped using Affymetrix 6.0 only). We benchmark the quality of our imputation in an African American population, for SNPs on the Metabochip, a region-centric genotyping platform, with particular focus on low-frequency SNPs (MAF down to 0.001), using a large study-specific reference panel containing 3,924 haplotypes. An African American sample poses a greater challenge for genotype imputation due to more complex LD patterns in African Americans compared with individuals of European ancestry [Egyud et al., 2009; Shriner et al., 2009], and in which comparatively less discovery work has been done.

We first describe how we constructed our study-specific reference panel using the 1,962 African American individuals with genotypes for both Affymetrix 6.0 and Metabochip SNPs and how we performed imputation of the Metabochip-only SNPs into the remaining 6,459 individuals. We then show several approaches through which we estimated imputation quality for SNPs in different MAF categories, with a special focus on less common (MAF: 0.01–0.05) and rare (MAF < 0.01) variants. We provide practical guidelines regarding post-imputation quality control for different MAF categories, as well as for the inclusion of rare variants during imputation.

MATERIALS AND METHODS

PRE-IMPUTATION QUALITY CONTROL

Prior to phasing and imputation, quality control was applied to both the Metabochip data and the GWAS data. Specifically, for the GWAS dataset ($n = 6{,}459$) we removed Affymetrix 6.0 SNPs with genotype call rates $< 90\%$ ($m = 1{,}633$), or Hardy-Weinberg exact test [Wigginton et al., 2005] P-value $< 10^{-6}$ ($m = 16{,}327$), or MAF < 0.01 ($m = 14{,}014$), resulting in a 829,370 GWAS SNPs passing quality control criteria [Reiner et al., 2011]. Separate quality control criteria were applied to the Metabochip SNPs, leading to 182,397 QC + SNPs with genotype call rates >95% and Hardy-Weinberg P-value $> 10^{-6}$. Individuals were excluded if they had a call rate below 95%, showed excess heterozygosity, were part of an apparent first-degree relative pair, or were ancestry outliers as determined by Eigensoft [Price et al., 2006]. Details can be found in the PAGE Metabochip platform paper [Buyske et al., 2011].

GENERAL PIPELINE FOR REFERENCE CONSTRUCTION AND SUBSEQUENT IMPUTATION

Figure 1 shows schematically how imputation was performed. In the top left panel, we first merged genotypes from the Affymetrix GWAS panel (blue) and the Metabochip (yellow) SNPs genotyped as part of the PAGE study for the 1,962 reference individuals (i.e., individuals with genotype data from both platforms). We then reconstructed haplotypes encompassing both GWAS and Metabochip SNPs for the reference individuals, constituting the reference panel of 3,924 haplotypes. In the top right panel, haplotype reconstruction for target individuals (i.e., individuals with GWAS genotypes only) was carried out similarly, but at the GWAS markers only. Finally, a haplotype-to-haplotype (that is, data are in haplotype form for both the reference and target individuals) imputation was performed to generate estimated genotypes at the Metabochip SNPs for the 6,459 target individuals.

RESULTS

GENOMEWIDE IMPUTATION USING LARGE STUDY-SPECIFIC REFERENCE

After careful matching on strand (so that genotypes from both Affymetrix 6.0 and the Metabochip are on the same strand), SNP ID, genomic coordinates, and actual genotypes for SNPs in common, we had a merged set of 987,749 SNPs for the 1,962 reference individuals. The average concordance rate for the 23,703 SNPs in common was 99.7%. For discordant genotypes, we kept the GWAS genotypes to match those of the target individuals with GWAS data only. Haplotypes were reconstructed on the merged set using MaCH [Li et al., 2010b]. In parallel, we constructed haplotypes across the 829,370 QC + GWAS SNPs for all 8,421 individuals. Finally, we used the 3,924 haplotypes across the merged set of 987,749 SNPs as reference to impute into

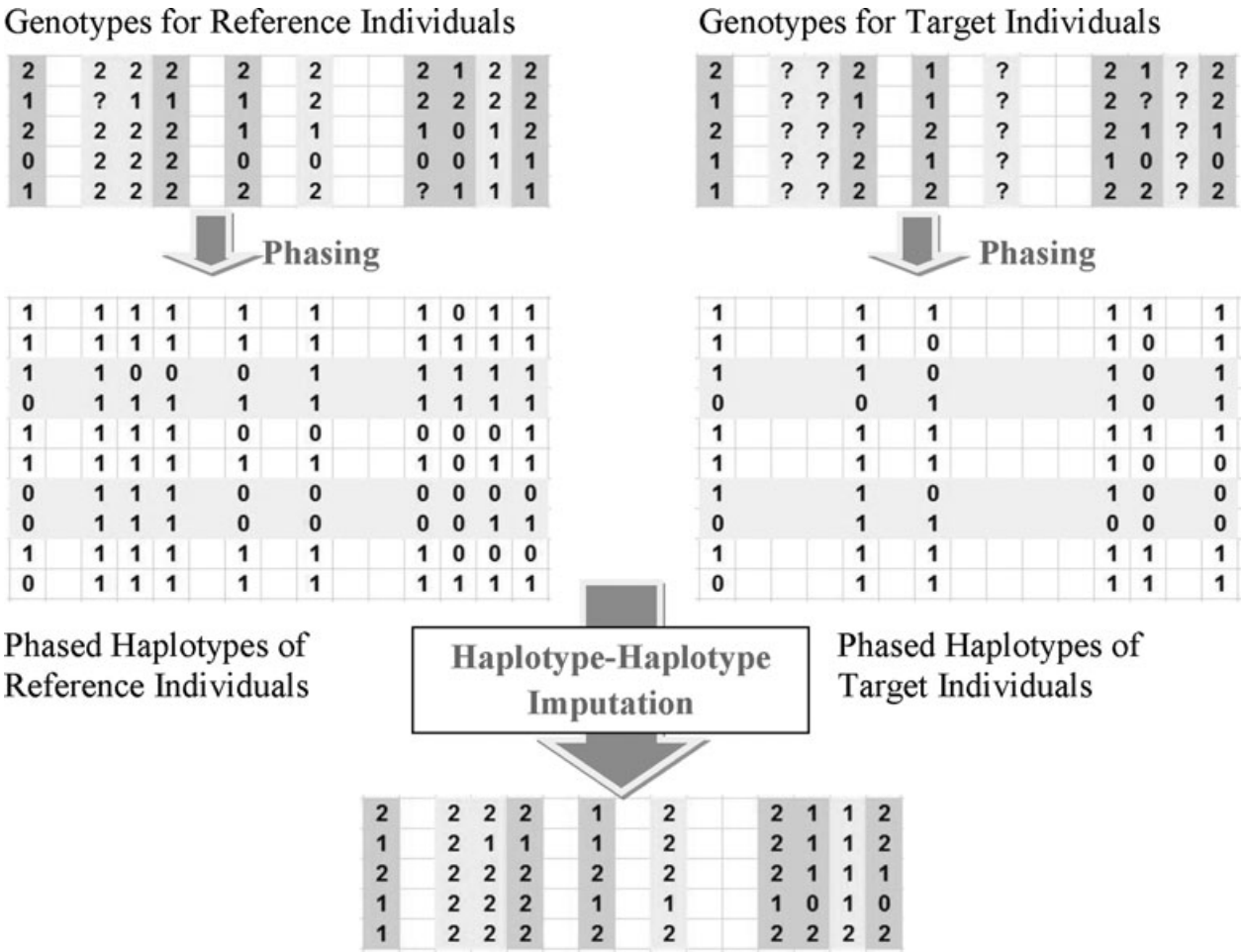

Fig. 1. Reference construction and imputation pipeline using a study-specific reference panel. This schematic cartoon shows how we constructed our study-specific reference panel using five individuals genotyped on both the Affymetrix 6.0 and the Metabochip platform and how we performed imputation into the remaining five individuals with Affymetrix 6.0 data only.

haplotypes across GWAS SNPs of the target individuals. The final haplotype-to-haplotype imputation was performed using the software package minimac, which generates the allele dosages (the fractional counts of an arbitrary allele at each SNP for each individual, ranging continuously from 0 to 2). Minimac also generates the SNP-level quality metric Rsq, which is the SNP-specific estimated r^2 between allele dosages and the unknown true genotypes. Rsq has been recommended as an efficient post-imputation quality control metric. Rsq, estimated r^2, and estimated imputation r^2 are used interchangeably in the literature [Browning and Browning, 2009; Li et al., 2009].

QUALITY ESTIMATE BY MASKING GENOTYPES AT 2% GWAS SNPs

Aside from production (actual imputation presented in the section above), we randomly masked 2% of the GWAS SNPs among the target individuals in the minimac imputation step to estimate the true imputation accuracy as well as to evaluate the utility of Rsq as a quality metric. By comparing imputed dosages with experimental genotypes, we and others have proposed several statistics to measure true imputation accuracy [Browning and Browning, 2009; Li et al., 2009; Lin et al., 2010; Marchini and Howie, 2010], measuring either the concordance rate, correlation, or degree of agreement. Here, we choose to report the dosage r^2, which is the squared Pearson correlation between the estimated allele dosages and the true experimental genotypes (recoded as 0, 1, and 2 corresponding to the number of minor alleles), because it is a more informative measure for low frequency variants by taking allele frequency into account and because it is directly related to the effective sample size for subsequent association analysis [Pritchard and Przeworski, 2001]. As dosage r^2 is calculated using the true genotypes (assuming the experimental genotypes are the true genotypes), people also call it true r^2. Like Rsq, dosage r^2 is also specific to each SNP.

Figure 2 shows the average dosage r^2 values for the 2% masked GWAS SNPs by chromosome. Genome-wide average is 93.68% (range 87.18 [chromosome 19] to 95.26% [chromosome 10]). As expected, larger chromosomes (in terms of physical length) tend to be slightly easier to impute due to slightly lower recombination rates and therefore higher level of LD [The International HapMap Consortium, 2005]. Chromosome 19, with the highest gene

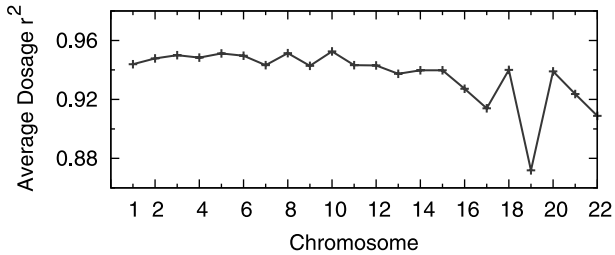

Fig. 2. Imputation accuracy by chromosome for 2% randomly masked GWAS SNPs. Imputation accuracy (as measured by average dosage r^2) for 2% GWAS SNPs masked at random is plotted by chromosome.

TABLE I. Average dosage r^2 by MAF, estimated by masking 2% GWAS SNPs

MAF	No. of SNPs	Average dosage r^2	Standard deviation dosage r^2
0.005–0.01	17	70.84%	18.23%
0.01–0.03	724	82.97%	16.07%
0.03–0.05	876	90.36%	11.03%
0.05–0.50	14,983	95.08%	7.70%

density, is most challenging for imputation. Table I shows the average dosage r^2 values by MAF. Not surprisingly, lower frequency variants are harder to impute due to poorer coverage by GWAS SNPs, lower degree of LD, and more challenging haplotype reconstruction. For example, the average dosage r^2 for SNPs with MAF > 0.05 is 95.08%; while the average for SNPs with MAF 0.005–0.01 is 70.84%.

While Figure 2 and Table I show the true imputation accuracy, in practice, we are more interested in how well imputation quality metrics can predict true imputation accuracy (measured by dosage r^2). Figure 3 assesses the quality metric Rsq by plotting it against dosage r^2. We can see that Rsq can predict dosage r^2 quite well, particularly for common SNPs and those with reasonable Rsq values. For example, the Pearson correlation is 0.938 for all SNPs (regardless of MAF and Rsq), 0.952 for SNPs with MAF > 0.03 (regardless of Rsq), and 0.955 for SNPs with MAF > 0.03 and Rsq > 0.3.

Whereas masking GWAS SNPs is a simple approach to estimate imputation accuracy, the approach estimates imputation quality for the "wrong" set of SNPs in that we are imputing genotypes for Metabochip SNPs, not GWAS SNPs. The two set of SNPs differ in two major aspects: MAF and physical density distribution. First, in terms of allele frequency distribution, while Affymetrix 6.0 SNPs, like most commercially available genome-wide genotyping platforms, contain SNPs that are mostly common, the Metabochip platform contains a much larger proportion of lower frequency variants. For example, while only 4.3% and 9.9% of the Affymetrix SNPs have MAF < 0.03 and < 0.05, respectively, the proportions are 29.8% and 37.8% for Metabochip SNPs. Supplementary Figure 1 shows the MAF distributions of the Affymetrix 6.0 SNPs and the Metabochip SNPs. Second, the physical distribution of the SNPs is quite different. The Affymetrix 6.0 SNPs are rather evenly spread across the genome. SNPs on the Metabochip, chosen for fine mapping of regions identified through

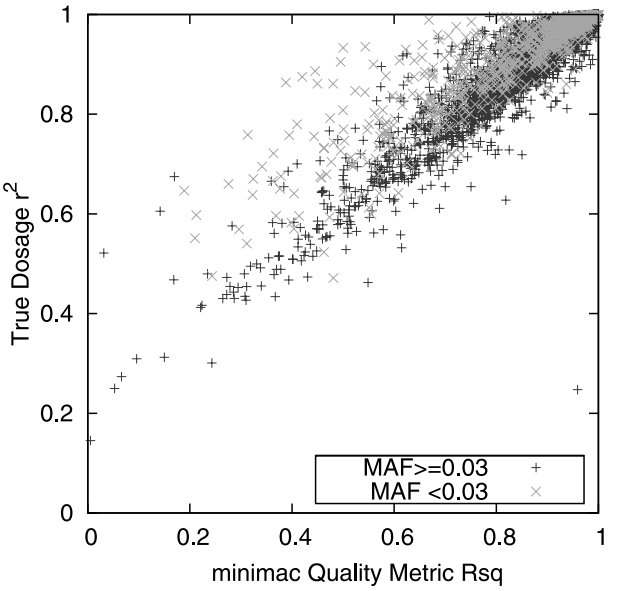

Fig. 3. Rsq by dosage r^2 for 2% randomly masked GWAS SNPs. Estimated imputation accuracy (minimac output Rsq) is plotted against the true dosage r^2, for the 2% GWAS SNPs masked at random.

GWAS to be associated with metabolic-related traits, scatter unevenly across the genome and are concentrated around GWAS-identified signals. Supplementary Figure 2 shows two typical regions where the GWAS SNP density (green) is quite uniform across the region, while Metabochip SNP density (red) peaks in a subregion chosen for follow-up but drops sharply outside the subregion of interest.

QUALITY ESTIMATE BY MASKING GENOTYPES AT METABOCHIP SNPs FOR A SUBSET OF REFERENCE INDIVIDUALS

To estimate the imputation quality for the actually imputed Metabochip SNPs, we masked Metabochip genotypes for 100 reference individuals, imputed them, and compared the estimated dosages with the masked experimental genotypes. Note that we used haplotypes constructed from GWAS data only for the 100 individuals. Supplementary Figure 3 shows the average dosage r^2 by chromosome. Again imputation quality is slightly higher for larger chromosomes and lowest for chromosome 19. Table II presents imputation accuracy by MAF, with and without post-imputation filtering according to Rsq. First, it is clear that lower frequency variants are harder to impute. We and others have shown earlier that imputation accuracy increases with the reference panel size, especially for the imputation of lower frequency variants [Li et al., 2009; Marchini and Howie, 2010; The International HapMap Consortium, 2010]. However, even with a reference panel of 3,924 haplotypes, we are not able to obtain reasonable imputed data for SNPs with MAF under 0.001. Without post-imputation filtering, the average dosage r^2 is merely 0.39%. If we apply a post-imputation filter of Rsq > 0.3 (>0.5), only 0.4% (0.3%) of the SNPs with MAF < 0.001 pass the filter with an average dosage r^2 of 24.85% (30.45%). For this rarest category of SNPs (MAF < 0.001), even at

TABLE II. Average Rsq and dosage r^2 by MAF, estimated by masking 100 reference individuals

	No Rsq filter			Rsq > 0.3			Rsq > 0.5		
MAF	No. of SNPs	Average Rsq	Average dosage r^2	Percent SNPs	Average Rsq	Average dosage r^2	Percent SNPs	Average Rsq	Average dosage r^2
0–0.001	18959	0.46%	0.39%	0.4%	72.31%	24.85%	0.3%	83.77%	30.45%
0.001–0.005	6925	21.80%	33.74%	23.8%	82.41%	73.94%	**20.5%**	**89.24%**	**79.71%**
0.005–0.01	7001	47.49%	64.87%	**52.0%**	**87.32%**	**83.05%**	48.2%	91.14%	86.00%
0.01–0.03	19894	77.57%	85.32%	**83.6%**	**91.72%**	**88.98%**	81.2%	93.21%	89.88%
0.03–0.05	13315	92.11%	91.73%	**97.5%**	**94.27%**	**92.11%**	96.3%	94.91%	92.57%
0.05–1.00	92597	96.94%	94.62%	**99.9%**	**97.05%**	**94.71%**	99.4%	97.30%	94.94%

Note: We evaluated a total of 158,691 of the total 182,397 QC+ Metabochip SNPs because 23,706 SNPs are both on the Metabochip and the Affymetrix 6.0 panel and were excluded from quality evaluation to avoid upward bias.

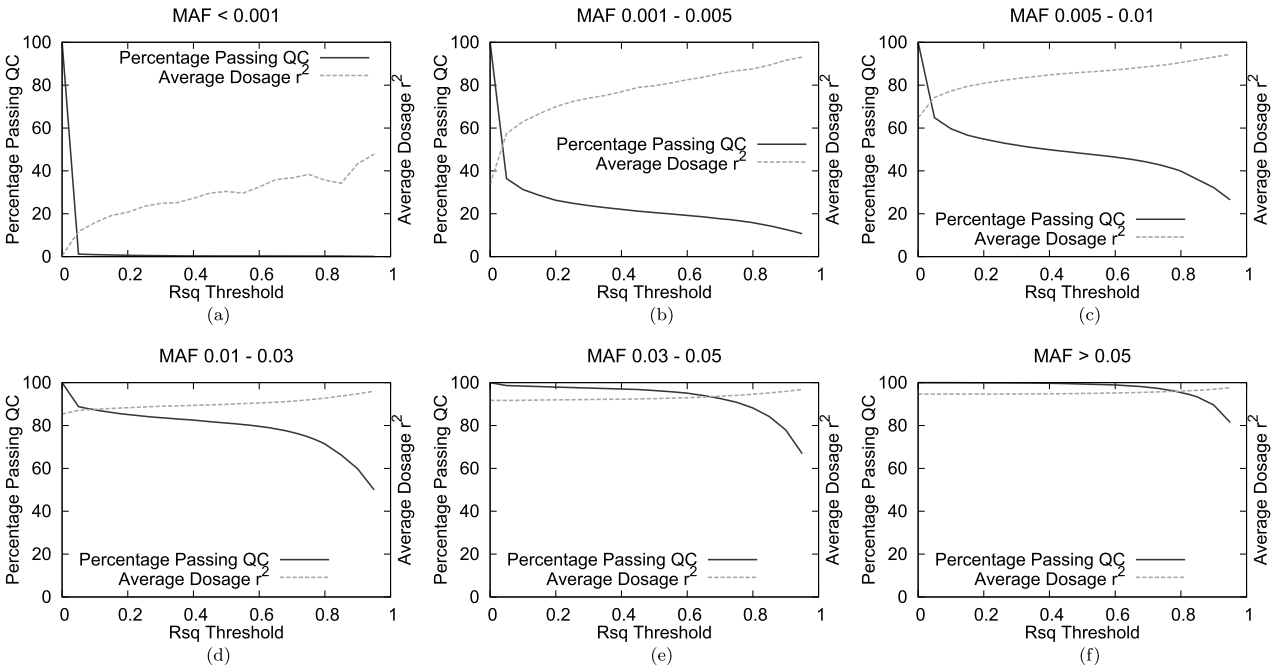

Fig. 4. Accuracy and calibration of imputation. Percentages of SNPs passing post-imputation QC (left *y* axis) and average dosage r^2 (right *y* axis) are plotted against Rsq threshold used for post-imputation QC for SNPs in different MAF categories.

an Rsq threshold of 0.95, which retains merely 23 of 18,959 SNPs, we can only achieve an average dosage r^2 of 47.82% (Fig. 4(a)). Second, SNPs with MAF > 0.01 can be imputed fairly well using a reference panel of this size. For example, even without any post-imputation quality control filter, the average dosage r^2 is 85.32%, 91.73%, and 94.62% for SNPs with MAF 0.01–0.03, 0.03–0.03, and >0.05, indicating that approximately 85–95% of the information can be recovered for SNPs in these MAF categories. Third, we are able to impute a considerable proportion of less common (MAF 0.001–0.01) variants reasonably well using a reference panel of this size along with post-imputation quality filtering according to Rsq. For example, we can obtain an average dosage r^2 of 79.71% for 20.5% of the SNPs with MAF 0.0010–0.005 by excluding SNPs with Rsq < 0.5; and an average dosage r^2 of 83.05% for 52.0% of the SNPs with MAF 0.005–0.01 by excluding SNPs with Rsq < 0.3, with both Rsq thresholds selected such that the average Rsq is above 80%.

OVERALL IMPUTATION PERFORMANCE AND PRACTICAL GUIDELINES

In practice, we recommend using Rsq as the post-imputation quality control metric. Figure 5 attests to the high correlation between Rsq and dosage r^2. We observe that the vast majority of SNPs are both imputed well and are predicted to be well imputed, corresponding to the biggest point masses (red to yellow range according to SNP frequency/count spectrum) with both high Rsq and high dosage r^2. Overall, we find that Rsq can predict dosage r^2 fairly well, particularly for common SNPs and those with reasonable Rsq values. For example, Pearson correlation between Rsq and dosage r^2 is 0.86 for SNPs with MAF 0.005–0.01 and Rsq > 0.5; and 0.93 for SNPs with MAF 0.01–0.03 and Rsq > 0.3. We also observe a noticeable point mass at the right bottom corner, corresponding to SNPs that are predicted to be poorly imputed (low Rsq) but are actually well imputed (high dosage r^2). Closer examination

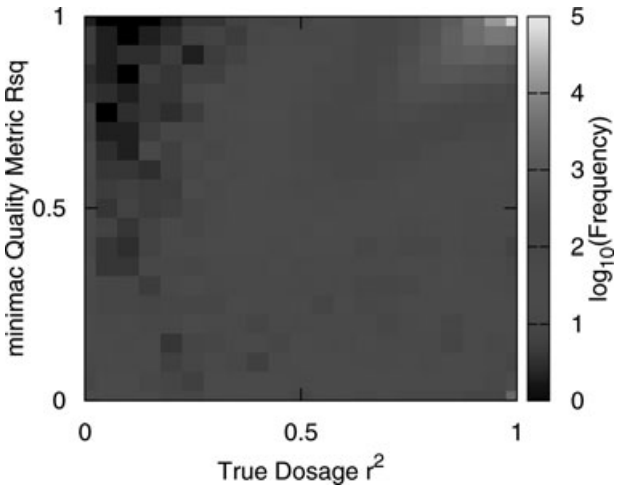

Fig. 5. Rsq by dosage r^2 Heatmap for Metabochip SNPs (estimated by masking 100 reference individuals). Estimated imputation accuracy (minimac output Rsq) is plotted against the true dosage r^2, for Metabochip SNPs by masking 100 reference individuals. Color scheme is defined by the number of underlying SNPs, specifically, log10 (Frequency).

revealed that most of these SNPs are of low frequency (95.4% have MAF < 0.03 and 99.7% have MAF < 0.05), for which the imputation model has low confidence in the estimated dosages that actually match the true dosages fairly well.

Furthermore, we recommend different Rsq thresholds for different MAF categories. Figure 4 presents the percentage of SNPs passing post-imputation QC (left y axis) and the average dosage r^2 (right y axis) as a function of Rsq threshold (x axis). To achieve an average dosage r^2 of at least 0.85 for example, one would have to use an Rsq threshold of 0.7 for SNPs with MAF 0.001–0.005 while an Rsq threshold of 0 suffices for SNPs with MAF > 0.03. Based on Table II and Figure 4, for our dataset, we chose an Rsq threshold of 0.5 for SNPs with MAF 0.001–0.005 and an Rsq threshold of 0.3 for SNPs with MAF > 0.005, resulting in a total of 127,132 SNPs (out of 158,691) passing post-imputation QC. The sample size for SNPs with MAF < 0.001 is too small for conclusions, but the pattern suggests that the few SNPs passing the post imputation filter of Rsq > 0.5 are well imputed. In general, we recommend selecting an Rsq threshold such that the average Rsq is above the desired average dosage r^2.

TO INCLUDE OR NOT TO INCLUDE: RARE SNPs DURING HAPLOTYPE RECONSTRUCTION

One open question concerns whether rare SNPs should be included for haplotype reconstruction, either for the reference individuals or for the target individuals. For the reference panel construction, on one hand, we would like to include as many variants as possible so that they can be subsequently imputed in the target individuals. On the other hand, inclusion of very rare SNPs may interfere with phasing (in the extreme case, for example, singletons cannot be phased), resulting in less accurately constructed haplotypes, and ultimately leading to inferior imputation quality, with little or no benefit in return because these very rare SNPs are unlikely to be accurately imputed into the target individuals. Similarly, for the target individuals, inclusion of rare SNPs may harm phasing quality, leading to less accurate imputation. On the other hand, as rare to-be-imputed SNPs are more likely to be tagged by rare GWAS SNPs than by common GWAS SNPs, inclusion of rare GWAS variants is expected to increase imputation quality for rarer SNPs. To evaluate this, we assessed the following 20 combinations by varying two parameters: MAF threshold used for the reference panel construction and MAF threshold used for phasing target individuals. For the reference panel construction, we evaluated the following four settings: (A) all MAF (i.e., no filtering by MAF); (B) no singletons (i.e., removing SNPs with only one copy of the minor allele among the 8,421 individuals with GWAS data); (C) MAF > 0.001; and (D) MAF > 0.005. For phasing target individuals, we evaluated the following five settings: (i) all MAF; (ii) no singletons (i.e., removing SNPs with only one copy of the minor allele among reference); (iii) MAF > 0.001; (iv) MAF > 0.005; and (v) MAF > 0.01. Note that for our production imputation, we used (v) MAF > 0.01. We picked a medium size chromosome, chromosome 12, for evaluation.

As the comparisons among the four settings for building the reference panel show similar patterns across the five settings for target haplotype reconstruction and vice versa, we present the average of all settings defined by the other parameter. For example, Table III shows the effect of including rare variants for reference panel construction, where the statistics (number of SNPs and average dosage r^2) for each of the four settings are averaged across the five settings for reconstructing target haplotypes. Among the four settings evaluated, setting B (no singletons) provides the best trade-off: noticeable gains for MAF categories 0.001–0.01 at little cost for common SNPs. For example, for SNPs with MAF 0.001–0.005, at an Rsq threshold of 0.3, setting B leads to 119 well-imputed SNPs with an average dosage r^2 of 84.0%, outperforming setting A which also results in 119 well-imputed SNPs but with a lower average dosage r^2 of 82.8%, setting C of 123 well-imputed SNPs with dosage r^2 of 82.8%, and setting D of 0 well-imputed SNPs (by design). For common SNPs with MAF > 0.01, all four settings have similar performance. On the other hand, there is no clear winner among the five settings for phasing GWAS data (Table IV). Removing SNPs with MAF < 0.001 or 0.005 (settings iii and iv) is slightly advantageous for imputing SNPs with MAF 0.001–0.01. For example, with an Rsq threshold of 0.3, average dosage r^2 for SNPs with MAF 0.001–0.005 is 85.6% and 84.0%, respectively, for settings iii and iv; while dosage r^2 for the other three settings is $\leq 83.0\%$. However, these settings result in slightly lower imputation quality for SNPs with MAF 0.01–0.05. For example, with an Rsq threshold of 0.3, average dosage r^2 for SNPs with MAF 0.01–0.03 is 90.4% (for 1,255 SNPs) and 90.6% (for 1,269 SNPs), respectively, for settings iii and iv; while dosage r^2 for the other three settings is $\geq 91.0\%$ for a larger number of SNPs (number of SNPs $\geq 1{,}289$).

DISCUSSION

As we are moving into the sequencing era, existing GWAS data provide an inexpensive opportunity to

TABLE III. Effect of including rare variants for reference panel construction

		A:All MAF		B:No Singletons		C:MAF > 0.1%		D:MAF > 0.5%	
MAF	Rsq threshold	No. of SNPs	Average dosage r^2	No. of SNPs	Average dosage r^2	No. of SNPs	Average dosage r^2	No. of SNPs	Average dosage r^2
0–0.001	0	22	44.0%	22	43.7%	0	NA	0	NA
0.001–0.005	0	266	70.9%	266	72.9%	266	72.3%	0	NA
0.005–0.01	0	494	85.7%	494	85.7%	494	84.8%	494	85.3%
0.01–0.03	0	1,521	90.4%	1,521	90.3%	1,521	90.3%	1,521	90.3%
0.03–0.05	0	955	93.4%	955	93.5%	955	93.4%	955	93.4%
0.05–1.00	0	5,494	95.5%	5,494	95.5%	5,494	95.5%	5,494	95.5%
0–0.001	0.3	2	100.0%	3	75.8%	0	NA	0	NA
0.001–0.005	0.3	119	82.8%	119	84.0%	123	82.8%	0	NA
0.005–0.01	0.3	333	87.6%	333	87.8%	328	87.6%	335	87.6%
0.01–0.03	0.3	1,307	91.1%	1,306	91.0%	1,307	91.0%	1,307	91.0%
0.03–0.05	0.3	941	93.6%	941	93.7%	940	93.8%	941	93.7%
0.05–1.00	0.3	5,486	95.6%	5,486	95.6%	5,487	95.5%	5,487	95.5%
0–0.001	0.5	2	100.0%	2	65.8%	0	NA	0	NA
0.001–0.005	0.5	105	85.6%	103	86.4%	105	85.9%	0	NA
0.005–0.01	0.5	310	89.1%	310	89.3%	308	89.1%	311	89.2%
0.01–0.03	0.5	1,268	92.1%	1,266	92.0%	1,268	92.0%	1,269	91.9%
0.03–0.05	0.5	931	94.2%	932	94.2%	932	94.2%	931	94.1%
0.05–1.00	0.5	5,460	95.9%	5,460	95.8%	5,461	95.8%	5,459	95.8%

TABLE IV. Effect of including rare variants for haplotype reconstruction among target individuals

		i:All MAF		ii:No singletons		iii:MAF > 0.1%		iv:MAF > 0.5%		v:MAF > 1%	
MAF	Rsq threshold	No. of SNPs	Average dosage r^2	No. of SNPs	Average dosage r^2	No. of SNPs	Average dosage r^2	No. of SNPs	Average dosage r^2	No. of SNPs	Average dosage r^2
0–0.001	0	22	45.3%	22	44.9%	22	44.2%	22	47.3%	22	37.5%
0–0.001	0.3	3	100.0%	2	100.0%	3	81.3%	2	75.0%	3	83.3%
0–0.001	0.5	3	100.0%	2	100.0%	3	81.3%	1	83.3%	3	83.3%
0.001–0.005	0	266	73.0%	266	72.7%	266	72.7%	266	71.8%	266	70.1%
0.001–0.005	0.3	102	83.0%	123	81.0%	122	85.6%	120	84.0%	133	82.3%
0.001–0.005	0.5	86	86.4%	104	84.5%	106	87.6%	107	86.5%	118	84.9%
0.005–0.01	0	494	85.5%	494	85.8%	494	86.6%	494	85.5%	494	83.4%
0.005–0.01	0.3	285	84.8%	332	88.8%	346	88.8%	350	88.3%	348	87.5%
0.005–0.01	0.5	264	86.4%	316	89.9%	325	90.1%	326	89.7%	317	89.8%
0.01–0.03	0	1,521	90.5%	1,521	90.6%	1,521	90.3%	1,521	90.1%	1,521	90.1%
0.01–0.03	0.3	1,289	91.4%	1,347	91.6%	1,255	90.4%	1,269	90.6%	1,373	91.0%
0.01–0.03	0.5	1,256	92.3%	1,293	92.7%	1,222	91.5%	1,231	91.6%	1,337	92.1%
0.03–0.05	0	955	93.4%	955	93.6%	955	93.2%	955	93.5%	955	93.5%
0.03–0.05	0.3	938	93.7%	943	93.8%	943	93.4%	933	93.8%	946	93.7%
0.03–0.05	0.5	932	94.1%	932	94.4%	934	93.9%	922	94.4%	938	94.1%
0.05–0.50	0	5,494	95.4%	5,494	95.5%	5,494	95.5%	5,494	95.5%	5,494	95.5%
0.05–0.50	0.3	5,486	95.5%	5,490	95.5%	5,487	95.5%	5,487	95.6%	5,484	95.7%
0.05–0.50	0.5	5,460	95.8%	5,463	95.8%	5,461	95.8%	5,457	95.9%	5,460	95.9%

leverage expensive sequencing data. Researchers across the world are becoming increasingly keen on imputation as a tool to infer genotypes at less common (MAF 0.01–0.05) and rare (MAF < 0.01) variants. We have previously shown that larger reference panels improve imputation accuracy for less common variants. In particular, enlarging a reference panel of 60 haplotypes to 1,000 haplotypes increases dosage r^2 for SNPs with MAF < 0.05 from 74% to 93% [Li et al., 2010a]. However, there has been little, if any, research on truly rare variants: it is not until recently that we have data available to assess imputation accuracy for these truly rare variants. Here, we used a reference panel of 3,924 reference haplotypes to demonstrate that it is indeed possible to impute a considerable proportion of rare variants reasonably well, even in a challenging admixed sample of African Americans. Specifically (as indicated in bold in Table II), we were able to impute 99.9% (97.5%, 83.6%, 52.0%, 20.5%) of SNPs with MAF > 0.05 (0.03–0.05, 0.01–0.03, 0.005–0.01, and 0.001–0.005) with average dosage r^2 94.7% (92.1%, 89.0%, 83.1%, and 79.7%).

In the previous section, we presented results from masking Metabochip genotypes for 100 reference individuals during minimac imputation, whom we also included along with the other 1,862 individuals during reference panel

TABLE V. Effect of including/excluding the 100 masked reference individuals during reference haplotype reconstruction

		$n = 1{,}862$ (Excluding)		$n = 1{,}962$ (Including)	
MAF	Rsq threshold	No. of SNPs	Average dosage r^2	No. of SNPs	Average dosage r^2
0–0.001	0	22	47.4%	22	40.3%
0–0.001	0.3	3	89.2%	2	85.2%
0–0.001	0.5	3	84.2%	2	77.1%
0.001–0.005	0	266	71.6%	266	72.6%
0.001–0.005	0.3	117	84.1%	123	82.3%
0.001–0.005	0.5	100	87.3%	108	84.7%
0.005–0.01	0	494	85.4%	494	85.4%
0.005–0.01	0.3	333	87.8%	332	87.5%
0.005–0.01	0.5	309	89.4%	310	89.0%
0.01–0.03	0	1,521	90.3%	1,521	90.4%
0.01–0.03	0.3	1,305	91.0%	1,308	91.1%
0.01–0.03	0.5	1,267	92.0%	1,268	92.1%
0.03–0.05	0	955	93.4%	955	93.4%
0.03–0.05	0.3	941	93.7%	940	93.7%
0.03–0.05	0.5	932	94.2%	931	94.2%
0.05–0.50	0	5,494	95.5%	5,494	95.5%
0.05–0.50	0.3	5,487	95.5%	5,487	95.6%
0.05–0.50	0.5	5,459	95.8%	5,462	95.8%

construction. One may reasonably argue that the inclusion of the 100 individuals during phasing results in local haplotype mosaics of other individuals better matching haplotypes of these 100 individuals (because constructed haplotypes of the 100 individuals are likely to serve as template to construct haplotypes of other individuals), and therefore over-estimated imputation accuracy. We evaluated this potential over-estimation of imputation accuracy by re-constructing the reference panel only on the other 1,862 individuals. Table V compares imputation accuracy at Metabochip SNPs for the 100 masked individuals with (phasing ref $n = 1{,}962$) or without (phasing ref $n = 1{,}862$) them during phasing. We observed no obvious over-estimation: the quality is either very close; or one has slightly smaller number of well-imputed SNPs with slightly higher dosage r^2 than the other. For example, for SNPs with MAF 0.001–0.005, when using Rsq > 0.3 as the post-imputation filter, the reference constructed using 1,862 individuals resulted in slightly fewer (117) SNPs passing the filter with a slightly better average dosage r^2 (84.1%), than the reference constructed using 1,962 individuals which had 123 SNPs passing the filter with an average dosage r^2 of 82.3%. The over-estimation may manifest itself if the reference panel were smaller because the 100 masked individuals would contribute more to the haplotype reconstruction of other reference individuals.

We would also like to note that masking 100 reference individuals, although allowing us to directly evaluate imputation quality at actually imputed Metabochip SNPs, still has limitations. For example, sample MAF cannot go below 0.005 and SNPs with "population" MAF (calculated based on $n = 1{,}962$ individuals) < 0.005 are either non-varying or have the minor allele over-represented among the 100 individuals (i.e., sample MAF > "population" MAF). Therefore, such SNPs are either not imputable (dosage r^2 un-defined and set to zero in our calculations) or tend to be easier to impute than a typical SNP in the population MAF category. The latter case leads to a winner's curse phenomenon such that the actual imputation quality tends to be over-estimated. In order to obtain more reliable estimates for the rarest MAF categories, we attempted a slightly more complicated experiment on chromosome 12 where we masked one reference individual at a time and imputed her genotypes at Metabochip SNPs using other reference individuals' haplotypes. This experiment allows us to examine a sample size of 1,962 instead of 100.

The overall recommendation of picking an Rsq threshold such that the average Rsq is at least 80% to achieve an average dosage r^2 of 80% or above still applies. However, compared with results based on 100 individuals, the actual Rsq thresholds selected for the rare MAF categories are considerably larger, but result in the passing of a larger proportion of SNPs. For example, an Rsq threshold of 0.75 (instead of 0.5 based on the 100 individuals) needs to be applied for SNPs with MAF 0.001–0.005 for the average Rsq to be above 80%, passing 38.8% (instead of 20.5% SNPs). The larger Rsq threshold and larger passing proportion are consistent with the winner's curse phenomenon we discuss above. For example, for SNPs with population MAF 0.001–0.005, the vast majority of SNPs are monomorphic among the 100 individuals and thus have Rsq close to zero, reflected by the fact that 68.7% of SNPs have Rsq < 0.1 (Fig. 4(b)). For the small proportion of SNPs that have reasonable Rsq (Rsq > 0.3), which is the proportion of SNPs with minor allele either over-represented or in more extensive LD with neighboring SNPs among the sample of 100 masked individuals, the distribution is highly skewed toward high values. For example, among the 20.5% SNPs with Rsq > 0.5, 16.9% (or 82.0% of the 20.5%) have Rsq > 0.75 such that the average Rsq is 89.24%. In contrast, a much larger proportion of SNPs are no longer monomorphic among the 1,962 individuals and better represent the full range of SNPs in these rare MAF categories, specifically by adding the more challenging SNPs (SNPs with less or no over-representation of the minor allele, and SNPs with less extensive LD with neighboring SNPs). For example, now only 1.8% (compared with

68.7% above based on 100 individuals) SNPs have Rsq < 0.1 for SNPs with MAF 0.001–0.005. Among the 73.5% (compared with 20.5% above) of SNPs with Rsq > 0.5, 38.8% (or 52.8% of the 73.5%) have Rsq > 0.75 (Supplementary Table 1).

Although this study examines an African American population genotyped using Affymetrix 6.0 platform, the recommendation to use Rsq threshold such that average Rsq is around but over the desired dosage r^2 value is generalizable to other populations and other GWAS genotyping platforms, based on similar experiments conducted in several European and Asian populations using different choices of genotyping platforms. For example, in a sample of Filipinos [Wu et al., 2010] genotyped using the Affymetrix 5.0 platform, we found applying a filter of Rsq > 0.6 for SNPs with MAF 0.01–0.02, the average dosage r^2 across the SNPs passing the filter was 0.8085 with an average Rsq of 0.8417. Additional assessment in other populations or using other GWAS platforms can be found in our earlier work [Li et al., 2011, 2010b]. Before more data become available, however, caution needs to be taken when applying the recommendation to rare variants. For example, although imputation in general is more difficult in African populations because of more combinations of the common alleles, recent work [Fumagalli et al., 2010; Gravel et al., 2011] argue that the more distinctive background of common alleles may benefit imputation of rare variants. In addition, tagSNPs on the Affymetrix 6.0 platform were selected largely based on physical positions, in contrast to those on the Illumina platforms that were selected largely to provide good coverage of the common SNPs according to HapMap-based LD. Therefore, the Affymetrix 6.0 platform may perform better for rare SNP imputation, particularly in samples of non-European ancestry.

Our sample consist of females only, therefore, it is straightforward to perform imputation on chromosome X. Even for samples including males, widely used imputation methods can now perform X chromosome imputation (see http://genome.sph.umich.edu/wiki/MaCH:_machX and [Marchini and Howie, 2010]). We did not attempt chromosome X in our dataset because there are only 93 QC+ Metabochip SNPs on chromosome X.

In summary, by constructing a study-specific reference panel of 3,924 haplotypes, we found it feasible to impute SNPs on the Metabochip, a region-centric dense genotyping platform, in a sample of African Americans, including less common SNPs with MAF 0.005–0.05. In addition, we confirmed Rsq as an effective imputation quality metric for these less common variants. In particular, we recommend different Rsq thresholds for different MAF categories such that the average Rsq is above 80%. Furthermore, we found it helpful to remove singleton SNPs when constructing reference haplotypes.

We view this work useful for investigators conducting fine-mapping studies using either dense genotyping or next generation sequencing, particularly for studies in non-European populations. Many efforts to fine map, especially in non-European ancestry participants, are limited by small sample sizes. Now that there are increasing numbers of GWAS studies conducted in non-European populations, imputation can provide a good solution to this sample size problem. For admixed samples like those in our study, new methods are being developed that both leverage the admixture for phenotype-genotype association mapping and take imputation uncertainty into account [Mao et al., 2011; Pasaniuc et al., 2011].

ACKNOWLEDGMENTS

We thank the WHI SHARe and PAGE consortia for generating the data. We also thank Dr. Karen Mohlke and other members of the CLHNS study for allowing us to use their data. Y.L. is partially supported by 5 UC2 HL102924-02 (awarded to K.E.N.), 3-R01-CA082659-11S1, and R01HG006292 (awarded to Y.L.). The PAGE program is funded by the National Human Genome Research Institute (NHGRI), supported by U01HG004803 (CALiCo), U01HG004798 (EAGLE), U01HG004802 (MEC), U01HG004790 (WHI), and U01HG004801 (Coordinating Center). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. The complete list of PAGE members can be found at http://www.pagestudy.org.

The WHI program is funded by the National Heart, Lung, and Blood Institute; NIH; and U.S. Department of Health and Human Services through contracts N01WH22110, 24152, 32100-2, 32105-6, 32108-9, 32111-13, 32115, 32118-32119, 32122, 42107-26, 42129-32, and 44221. The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: http://www.whiscience.org/publications/WHI_investigators_shortlist.pdf.

The PAGE coordinating center (U01HG004801-01) provides assistance with study design, phenotype harmonization, SNP selection and annotation, data cleaning, data management, integration and dissemination, and general study coordination. Genotype calling, genotype quality control, and statistical analyses are also performed by the coordinating center for some PAGE studies. The National Institute of Mental Health also contributes to the support for the coordinating center.

The PAGE consortium thanks the staff and participants of all PAGE studies for their important contributions.

REFERENCES

Anderson G, Cummings S, Freedman LS, Furberg C, Henderson M, Johnson SR, Kuller L, Manson J, Oberman A, Prentice RL, Rossouw JE. 1998. Design of the women's health initiative clinical trial and observational study. Controlled Clinical Trials 19(1):61-109.

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.

Buyske S, Wu Y, Carty CL, Cheng I, Assimes TL, Dumitrescu L, Hindorff LA, Mitchell S, Ambite JL, Boerwinkle E, Buzkova P, Carlson CS, Cochran B, Duggan D, Eaton CB, Fesinmeyer MD, Franceschini N, Haessler J, Jenny N, Hyun Min Kang, Lin Y, Le Marchand L, Matise TC, Robinson JG, Rodriguez C, Schumacher FR, Voight BF, Young A, Manolio TA, Mohlke KL, Haiman CA, Peters U, Crawford DC, North KE. 2011. Evaluation of the Metabochip genotyping array in African Americans and implications for fine mapping of GWAS-Identified loci: the PAGE Study. In preparation.

de Bakker PIW, Ferreira MAR, Jia XM, Neale BM, Raychaudhuri S, Voight BF. 2008. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet 17:R122–R128.

Egyud MR, Gajdos ZK, Butler JL, Tischfield S, Le Marchand L, Kolonel LN, Haiman CA, Henderson BE, Hirschhorn JN. 2009. Use of weighted reference panels based on empirical estimates of ancestry for capturing untyped variation. Hum Genet 125:295–303.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–450.

Fumagalli M, Cagliani R, Riva S, Pozzoli U, Biasin M, Piacentini L, Comi GP, Bresolin N, Clerici M, Sironi M. 2010. Population genetics of IFIH1: ancient population structure, local selection, and implications for susceptibility to type 1 diabetes. Mol Biol Evol 27:2555–2566.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci USA 108:11983–11988.

He J, Wilkens LR, Stram DO, Kolonel LN, Henderson BE, Wu AH, Le Marchand L, Haiman CA. 2011. Generalizability and epidemiologic characterization of eleven colorectal cancer GWAS hits in multiple populations. Cancer Epidemiol Biomarkers Prev 20:70–81.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 106:9362–9367.

Keebler ME, Deo RC, Surti A, Konieczkowski D, Guiducci C, Burtt N, Buxbaum SG, Sarpong DF, Steffes MW, Wilson JG, Taylor HA, Kathiresan S. 2010. Fine-mapping in African Americans of 8 recently discovered genetic loci for plasma lipids: the Jackson heart study. Circ Cardiovasc Genet 3(4):358-364.

Lanktree MB, Anand SS, Yusuf S, Hegele RA. 2009. Replication of genetic associations with plasma lipoprotein traits in a multiethnic sample. J Lipid Res 50:1487–1496.

Lettre G, Palmer CD, Young T, Ejebe KG, Allayee H, Benjamin EJ, Bennett F, Bowden DW, Chakravarti A, Dreisbach A, Farlow DN, Folsom AR, Fornage M, Forrester T, Fox E, Haiman CA, Hartiala J, Harris TB, Hazen SL, Heckbert SR, Henderson BE, Hirschhorn JN, Keating BJ, Kritchevsky SB, Larkin E, Li M, Rudock ME, McKenzie CA, Meigs JB, Meng YA, Mosley TH, Newman AB, Newton-Cheh CH, Paltoo DN, Papanicolaou GJ, Patterson N, Post WS, Psaty BM, Qasim AN, Qu L, Rader DJ, Redline S, Reilly MP, Reiner AP, Rich SS, Rotter JI, Liu Y, Shrader P, Siscovick DS, Tang WH, Taylor HA, Tracy RP, Vasan RS, Waters KM, Wilks R, Wilson JG, Fabsitz RR, Gabriel SB, Kathiresan S, Boerwinkle E. 2011. Genome-wide association study of coronary heart disease and its risk factors in 8090 african americans: the NHLBI CARe project. PLoS Genet. 7:e1001300.

Li L, Li Y, Browning SR, Browning BL, Slater Al, Kong X, Aponte JL, Mooser VE, Chissoe SL, Whittaker JC, Nelson MR, Ehm MG. 2011. Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. PLoS One 6:e24945.

Li Y, Willer C, Sanna S, Abecasis G. 2009. Genotype imputation. Annu Rev Genom Hum Genet 10:387–406.

Li Y, Byrnes AE, Li M. 2010a. To identify associations with rare variants, just WHaIT: weighted haplotype and imputation-based tests. Am J Hum Genet 87:728–735.

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. 2010b. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34:816–834.

Lin P, Hartz SM, Zhang Z, Saccone SF, Wang J, Tischfield JA, Edenberg HJ, Kramer JR, A MG, Bierut U, Rice JP. 2010. A new statistic to evaluate imputation reliability. PLoS One 5:e9697.

Maher B. 2008. Personal genomes: the case of the missing heritability. Nature 456:18–21.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy Ml, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM. 2009. Finding the missing heritability of complex diseases. Nature 461:747-753.

Mao X, Li Y, Li M. 2011. Detecting Genetic Association with Rare Variants in Admixed Populations. Abstract No. 660. Presented at the 12th International Congress of Human Genetics/61st Annual Meeting of The American Society of Human Genetics, October 13, 2011, Montreal, Canada. Annual Meeting of the American Society of Human Genetics.

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

Matise TC, Ambite JL, Buyske S, Carlson CS, Cole SA, Crawford DC, Haiman CA, Heiss G, Kooperberg C, Marchand LL, Manolio TA, North KE, Peters U, Ritchie MD, Hindorff LA, Haines JL. 2011. The Next PAGE in understanding complex traits: design for the analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. Am J Epidemiol 174(7):849-859.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356–369.

Pasaniuc B, Zaitlen N, Lettre G, Chen GK, Tandon A, Kao WH, Ruczinski I, Fornage M, Siscovick DS, Zhu X, Larkin E, Lange LA, Cupples LA, Yang Q, Akylbekova EL, Musani SK, Divers J, Mychaleckyj J, Li M, Papanicolaou GJ, Millikan RC, Ambrosone CB, John EM, Bernstein L, Zheng W, Hu JJ, Ziegler RG, Nyante SJ, Bandera EV, Ingles SA, Press MF, Chanock SJ, Deming SL, Rodriguez-Gil JL, Palmer CD, Buxbaum S, Ekunwe L, Hirschhorn JN, Henderson BE, Myers S, Haiman CA, Reich D, Patterson N, Wilson JG, Price AL. 2011. Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a breast cancer consortium. PLoS Genet. 7:e1001371.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909.

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

Pulit SL, Voight BF, de Bakker PI. 2010. Multiethnic genetic association studies improve power for locus discovery. PLoS One 5:e12600.

Reiner AP, Lettre G, Nalls MA, Ganesh SK, Mathias R, Austin MA, Dean E, Arepalli S, Britton A, Chen Z, Couper D, Curb JD, Eaton CB, Fornage M, Grant SF, Harris TB, Hernandez D, Kamatini N, Keating BJ, Kubo M, LaCroix A, Lange LA, Liu S, Lohman K, Meng Y, Mohler ER, 3rd, Musani S, Nakamura Y, O'Donnell G, Okada Y, Palmer CD, Papanicolaou GJ, Patel KV, Singleton AB, Takahashi A, Tang H, Taylor HA, Jr., Taylor K, Thomson C, Yanek LR, Yang L, Ziv E, Zonderman AB, Folsom AR, Evans MK, Liu Y, Becker DM, Snively BM, Wilson JG. 2011. Genome-wide association study of white blood cell count in 16388 African Americans: the continental origins and genetic epidemiology network (COGENT). PLoS Genet. 7:e1002108.

Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. 2010. Genome-wide association studies in diverse populations. Nat Rev Genet 11:356–366.

Shriner D, Adeyemo A, Chen GJ, Rotimi CN. 2009. Practical considerations for imputation of untyped markers in admixed populations. Genet Epidemiol 33:258–265.

Smith JG, Magnani JW, Palmer C, Meng YA, Soliman EZ, Musani SK, Kerr KF, Schnabel RB, Lubitz SA, Sotoodehnia N, Redline S, Pfeufer A, Muller M, Evans DS, Nalls MA, Liu Y, Newman AB, Zonderman AB, Evans MK, Deo R, Ellinor PT, Paltoo DN, Newton-Cheh C, Benjamin EJ, Mehra R, Alonso A, Heckbert SR, Fox ER. 2011. Genome-wide association studies of the PR interval in African Americans. PLoS Genet. 7:e1001304.

Teo YY, Small KS, Kwiatkowski DP. 2010. Methodological challenges of genome-wide association analysis in Africa. Nat Rev Genet 11:149–160.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.

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

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

The International HapMap Consortium. 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58.

Waters KM, Le Marchand L, Kolonel LN, Monroe KR, Stram DO, Henderson BE, Haiman CA. 2009. Generalizability of associations from prostate cancer genome-wide association studies in multiple populations. Cancer Epidemiol Biomarkers Prev 18:1285–1289.

Wigginton JE, Cutler DJ, Abecasis GR. 2005. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 76:887–893.

Wu Y, Li Y, Lange EM, Croteau-Chonka DC, Kuzawa CW, McDade TW, Qin L, Curocichin G, Borja JB, Lange LA, Adair LS, Mohlke KL. 2010. Genome-wide association study for adiponectin levels in Filipino women identifies CDH13 and a novel.uncommon haplotype at KNG1-ADIPOQ. Hum Mol Genet 19:4955-64.