Improved Imputation of Common and Uncommon Single Nucleotide Polymorphisms (SNPs) with a New Reference Set

Zhaoming Wang [1,2]
Kevin B. Jacobs [1,2]
Meredith Yeager [1,2]
Amy Hutchinson [1,2]
Joshua Sampson [2]
Nilanjan Chatterjee [2]
Demetrius Albanes [2]
Sonja I. Berndt [2]
Charles C. Chung [2]
W. Ryan Diver [3]
Susan M. Gapstur [3]
Lauren R. Teras [3]
Christopher A. Haiman [4]
Brian E. Henderson [4]
Daniel Stram [4]
Xiang Deng [1,2]
Ann W. Hsing [2]
Jarmo Virtamo [5]
Michael A. Eberle [6]
Jennifer L. Stone [6]
Mark P. Purdue [2]
Phil Taylor [2]
Margaret Tucker [2]
Stephen J. Chanock [2]

[1] Core Genotyping Facility, SAIC-Frederick, Inc., NCI-Frederick, Frederick, MD 21702, USA
[2] Division of Cancer Epidemiology and Genetics, NCI, NIH, Bethesda, MD 20892, USA
[3] Epidemiology Research Program, American Cancer Society, Atlanta, GA, 30303, USA
[4] Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, CA, 90089, USA
[5] Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland
[6] Illumina, Inc. San Diego, CA 92121, USA

Correspondence should be addressed to:
Stephen J. Chanock, M.D.
Laboratory of Translational Genomics
Division of Cancer Epidemiology and Genetics
National Cancer Institute
Advanced Technology Center- NCI
8717 Grovemont Circle
Bethesda, MD 20892-4605
Email: chanocks@mail.nih.gov
Tel: 301-435-7559
Fax: 301-402-3134

Abstract

Statistical imputation of genotype data is an important technique for analysis of genome-wide association studies (GWAS). We have built a reference dataset to improve imputation accuracy for studies of individuals of primarily European descent using genotype data from the Hap1, Omni1, and Omni2.5 human SNP arrays (Illumina). Our dataset contains 2.5-3.1 million variants for 930 European, 157 Asian, and 162 African/African-American individuals. Imputation accuracy of European data from Hap660 or OmniExpress array content, measured by the proportion of variants imputed with $R^2>0.8$, improved by 34%, 23% and 12% for variants with MAF of 3%, 5% and 10%, respectively, compared to imputation using publicly available data from 1,000 Genomes and International HapMap projects. The improved accuracy with the use of the new dataset could increase the power for GWAS by as much as 8% relative to genotyping all variants. This reference dataset is available to the scientific community through the NCBI dbGaP portal. Future versions will include additional genotype data as well as non-European populations.

Introduction

Genome-wide association studies (GWAS) have emerged as a successful strategy for the discovery of common single nucleotide polymorphism (SNP) markers associated with human diseases/traits[1]. The decreasing cost of commercial dense SNP arrays has enabled investigators to scan the genome and prioritize variants for confirmation of true signals amidst a large number of false positives[2]. Hundreds of different loci in the genome have been conclusively associated with more than 100 distinct complex diseases and traits, ushering in a new age of discovery in human genetics[3].

The first generation of commercial genotyping arrays was based on genetic variants annotated by the International Human HapMap project[4-6] as well as those in the NCBI dbSNP database. These arrays have targeted common SNP markers across the genome with the majority of variants with minor allele frequency (MAF) greater than 10%, though a small proportion of content included uncommon variants with MAF between 5 and 10%[7]. Between the bias due to array content and the larger sample sizes needed to detect uncommon variants, it is not surprising that the majority of markers identified in GWAS have been common variants. The proliferation of meta-analyses between existing scans, together with larger replication sets, have shifted the focus of the field from the identification of a few regions to the discovery of comprehensive sets of common variants associated with one or more human diseases or traits[8-11].

Until recently, genome-wide association studies have been conducted with either the "HumanHap" series of Illumina, Inc. or the Affymetrix 6.0, genotyping up to 1,000,000 SNPs. The 1000 Genome Project has laid the foundation for annotating many more common (MAF > 5%) and uncommon SNPs (MAF between 1% and 5%) in multiple continental reference populations[4-6]. Since commercial vendors have transitioned to new denser content arrays that can genotype up to 5,000,000 SNPs, the lack of overlap between commercial arrays will require GWAS investigators to depend upon statistical

genotype imputation techniques for analysis across vendor arrays as well as between different generations of assays by the same vendor. For instance, Illumina, Inc. has recently discontinued and replaced the "HumanHap" assays with a new series of "Omni" assays that contain 720,000-5,000,000 SNPs assays with content that is substantially different from the earlier series. Similarly, Affymetrix has released the Axiom assay, which allows similar genotyping of millions of SNPs based on population-customized content.

Statistical imputation of genotype data[12-15] is an important technique that uses patterns of linkage disequilibrium observed in a reference set to computationally predict additional genetic variants *in silico*. Imputed genotypes may be tested for association with phenotypes, but also enable the combining of genotype data typed on SNP arrays with different content, as well as the use of methods that are intolerant of missing data. The accuracy of genotype imputation depends on the sample size of the reference set, the comprehensive nature of the SNP coverage, data quality and the comparability of the reference set to the study population(s) with respect to the underlying population substructure[16-18]. Currently, the most popular reference sets are the publicly available International HapMap and 1000 Genomes datasets [4-6]. While these resources are valuable for imputing a sizeable fraction of common SNPs (MAF > 10%), they may not be optimal for imputing data for the next generation of GWAS arrays.

Since most studies have been conducted on first generation commercial arrays with less dense content (e.g., optimizing coverage for SNPs with MAF > 10%), we have developed a resource to enable use of older array data to impute new content as well as to enhance the ability to combine studies across platforms. Furthermore, the resource can address the performance of imputation of common and now uncommon variants not represented on first generation SNP arrays. The Division of Cancer Epidemiology and Genetics (DCEG) at the National Cancer Institute (NCI) has generated a dataset, now deposited in dbGaP, which is expected to expand to increase SNP genotype content and to incorporate additional data from non-European populations.

Results

The DCEG Imputation Reference Set includes; (1) 728 cancer-free individuals of European descent from three large prospectively sampled studies[19-22] genotyped on the Illumina Hap1, Omni1, Omni2.5 arrays; (2) 98 African-American individuals from the Prostate, Lung, Colon, and Ovary Cancer Screening Trial (PLCO) genotyped on Illumina Hap1 and Omni2.5 arrays; and (3) 74 Chinese individuals[23] genotyped on the Hap660 and the Omni 2.5 arrays from a study of upper gastrointestinal cancer in Shanxi, China (SHNX). We combined our dataset with 349 HapMap samples genotyped on the Omni 2.5 array to form a harmonized dataset of approximately 2.8 million autosomal polymorphic SNPs in 1,249 subjects after rigorous quality control metrics were applied (Table 1). During the quality control process, we removed 90 non-founders and two additional subjects forming related pairs from the HapMap set. The relationships and extent of overlapping content among the 2.8 million SNPs and the Illumina arrays (as portrayed in the commercial manifests) and 1000 Genomes and HapMap 3 data are shown in **Figure** 1**a**; the OmniExpress content simulated from 2.8 million SNPs of DCEG Reference Set and the content overlapping with 1000 Genomes and HapMap 3 data are shown in Figure 1**b**. In the DCEG Reference Set, the MAF distribution for the Omni2.5 array is shown per population in Figure 2. It is notable that over 1 million uncommon (or rare) SNPs (MAF < 10%) were observed across different populations on the Omni2.5 array; the largest number was observed in Asians while the smallest number is in African-Americans. Since the full content of the OmniExpress array is contained within the Omni1 array and the OmniExpress is less expensive, we have reported our analysis using the OmniExpress content only throughout the manuscript.

We compared the imputation performance of the DCEG Reference Set to the International HapMap and 1000 genome reference sets, which were available from the IMPUTE2 website (URL below). Our approach to measuring imputation performance is based on the comparison of directly

genotyped SNPs and probabilistic imputed genotypes (using both IMPUTE2[13] and BEAGLE[12]) using subsets of directly genotyped SNPs to simulate data genotyped on two prototypes of Illumina commercial arrays used in GWAS studies (Hap660 and OmniExpress). We investigated the utility of the DCEG Reference Set by comparing imputation accuracy with that of the publicly available set of HapMap and 1000 Genomes for SNPs with MAF estimated to be greater than 1%. We assumed that directly genotyped SNPs that pass quality control criteria are correct. We measured imputation accuracy between imputed SNPs and the previously masked directly genotyped SNPs as the squared-Pearson correlation coefficient (R^2) using imputed genotype probabilities without censoring and a trend/dosage model (See Methods).

To address the question of how well imputation can be applied to two commercially available SNP arrays, we simulated content for the Illumina Hap660 and OmniExpress arrays. Imputation was performed with the IMPUTE2 program with three reference data sets: (1) DCEG reference set (all samples excluding those used for the simulated Illumina datasets); (2) 1000 Genomes project June 2010 release and HapMap 3 release 2[13]; and (3) the combination of the DCEG reference set and the 1000 Genomes/HapMap 3 data sets. The reference sets included the 2.8 million SNPs genotyped as part of this dataset and/or the 7.8 million SNPs available from the 1000 Genomes project and HapMap 3 release. Accuracy was specifically assessed using the subset of 2 million SNPs with MAF > 1%.

Across the spectrum of MAFs for Illumina Hap660 and OmniExpress, we observed substantial improvement in imputation accuracy when using our reference panel compared to the combination of 1000 Genomes and HapMap data. We randomly selected 60 samples of European ancestry and imputed to the full set of 2.8 million SNPs, out of which 2 million SNPs with MAF > 1% were applicable for the evaluation of imputation performance (Figure 3 and also more detailed R^2 distribution by MAF is shown in Figure 4). Accuracy in the European data from Hap660 or OmniExpress array content, measured by the proportion of variants imputed with R^2>0.8, improved by 34%, 23% and 12% for variants with MAF of 3%, 5% and 10%, respectively. For the combined reference from DCEG, 1000 Genomes, and HapMap 3 data we observed slightly lower performance when imputing from the OmniExpress content to the full set

and when combined, we observed no appreciable improvement (Figure 5). We achieved similar results for all of these analyses using both the IMPUTE2 and BEAGLE imputation programs[12].

We suspected that the matching of populations between inference and reference sets could be an important factor in overall accuracy. We explored different European populations or US-based cohort studies with samples of European ancestry to assess the utility of the new DCEG reference set. The analysis imputed samples of European ancestry from the PLCO study based on an American prospective cohort the Cancer Prevention Study-II of the American Cancer Society (CPSII), a Finnish clinical trial, the Alpha Tocopherol Beta Carotene Cancer Prevention Study (ATBC), and the European HapMap samples (CEU+TSI) respectively. For each, the population substructure was evaluated by principal components analysis (Figure 6). Similar imputation performance was observed for each of the three references sets for SNPs > 10% (Figure 7); below this threshold, ATBC, performs surprisingly well for such a 'mismatched' subpopulation of the same continental origin. For common variants, performance is excellent even when population substructure exists, which suggests that for common variants, a reference set of sufficient size can adequately predict common SNPs when there is a discrepancy in population genetics history. In turn, this confirms the practical adaptation of this approach for many published imputation-based studies of European ancestry, in which the reference and inference sets for common SNPs differ. Since the accuracy of imputation is also based on sample size, we observed a gradual improvement in accuracy as sample size increased from 50 to 800 (Figure 8). Moreover, the effect was more pronounced for uncommon SNPs between 5 and 10%, whereas for SNPs above 10% the effect on the performance was less notable. The presumed return on investment for SNPs with lower MAF begins to diminish even at sample sizes of 600 and 800, an asymptotic behavior explored elsewhere [24].

Although the current build of the DCEG reference Set is primarily intended for use in European populations, we tested the accuracy of imputing OmniExpress data on an African-American set from the Multi-Ethnic Cohort (MEC) study [25]. Accuracy was less optimal for African Americans, but still superior to the publicly available reference data. As seen in Figure 9, there was an improvement in performance

for SNPs above 3-5% when the reference set included 98 African-American samples from PLCO and the HapMap YRI samples. It is also notable that the performance was better for the combined set than with only the African Americans in PLCO (Figure 9). For the two sets, the STRUCTURE plot shows a comparable distribution of admixture coefficients for the African-Americans along the axis between EUR and AFR (Figure 10). A comparison of the overall imputed SNPs (~2 million when using our reference set versus 1.38 million derived from 1000 Genome plus HapMap3) indicates an advantage of the DCEG Reference Set with MAF > 3%. The R^2 curve crossed over at 3% in MAF, which could be due to a substantial proportion of Omni2.5 SNPs with MAF < 3% that had to be imputed when using the DCEG set compared to the combined 1000 Genomes and HapMap set (which included more SNPs < 3% in total). When the common set of 1.38 million SNPs is compared, the performance improvement is consistent across all ranges of MAF, which was not clearly observed in the European populations (Figure 3). Additional studies are needed to investigate these observations for African-Americans and perhaps other admixed populations that also include a substantial European contribution.

We compared estimates of power to detect associations in GWAS when SNPs are imputed and when they are directly genotyped. With imputation based on the DCEG reference set and 1000 Genomes/HapMap 3, a GWAS can be expected to detect 92.9% and 84.7%, respectively, of those associations discovered by direct genotyping when using the Hap660. The relative power from the DCEG reference set and 1000 Genomes/HapMap 3 were 93.9% and 86.2%, respectively, when using the OmniExpress based on the model of Park et al.[10]. These results suggest that new reference set improves the power for GWAS by a small, but noticeable, margin.

Discussion

In this report we show that a new public resource, the DCEG Reference set, performs better than the standard reference data sets, 1000 Genome Project and HapMap 3, for the imputation of common and uncommon SNPs, particularly in populations of European background. The improvement in imputation is evident in both forward and backward scenarios, namely from the earlier generation arrays to more dense arrays and for complementary arrays. Our data suggest that there are distinct advantages for the use of genotyped data compared to low pass sequence coverage with respect to the accuracy of imputation, which has implications for efficient GWAS studies. While low pass sequencing data may capture more variants, the cumulative effects of both higher false-positive and -negative rates may have a suboptimal effect on the accuracy of imputation of common and uncommon SNPs. Still, we show that there are advantages to imputation using a dataset with validated assays.

Our data set is particularly useful for investigators to conduct GWAS with a hybrid of data generated across arrays or between distinct generations of the same vendor (Hap660 and OmniExpress) using genotype imputation techniques to discover promising regions that will require confirmation in follow-up studies[2]. Since many GWAS share cases and controls, the use of our dataset should facilitate the approach of shared controls of similar or comparable population background, perhaps even across platforms on economic and genetic grounds.

Since imputation accuracy depends on the similarity in population substructure between reference and study populations, we examined the interchangeability of the three similar, but not identical European populations. In our analysis of the DCEG Reference Set, we encountered two unexpected and notable results. First, we were surprised to observe that for common SNPs, there was little impact of a reference population from Finland, clearly mismatched, with two cohorts drawn from across the continental USA (e.g., CPSII and PLCO) (Figure 7). This suggests that for common variants a reference set of sufficiently

larger size can adequately capture common SNPs when there is a small but real discrepancy in population genetics history. This could enable investigators to proceed with imputation without additional genotyping in related but not identical populations. In the second instance, although overall performance of imputation was less optimal for African Americans, we observed that a sufficiently large sample set of European subjects can improve imputation in African-Americans, beyond the utility of the PLCO African American set, suggesting the value of the larger sample size of European subjects (Figure 9). This could have an impact on the design of future studies in other populations with distinct substructures. Despite the fact that a subset of MAFs may differ, longer haplotypes could still be useful for imputation of common and uncommon variants. Although frequencies of variants may differ among population groups and sub-populations, our findings suggest that genotype imputation is relatively robust to these differences provided that a sufficient number of matching haplotypes appear in the reference data. In the future, larger sample sizes should be useful to determine imputation performance, particularly across the spectrum of lower MAFs.

We tested imputing OmniExpress data from European individuals with a reference dataset that combined the DCEG and 1,000 Genomes and HapMap references and observed no increase in accuracy over that was achieved using the DCEG Reference alone. This finding suggests that there are distinct advantages to using SNP array data compared to low pass sequence coverage with respect to the accuracy of imputation and has implications for efficient GWAS studies. One possible explanation is that SNP array and low-pass sequence variant data have distinct patterns of non-random errors. While suboptimal for imputation, it may also be necessary to preserve these errors when combining directly genotyped and imputed data in order to recapitulate the patterns of differential misclassification and perhaps retain statistical validity for association testing. Thus, SNP array data should be superior for imputation of genotypes obtained from SNP arrays, though not necessarily superior at imputing the true genotypes. While low pass sequencing data may capture more variants, the cumulative effects of both higher false-positive and -negative rates may also decrease imputation accuracy.

In conclusion, our study has shown the utility of the new DCEG Reference Set and its advantage for imputation of common and uncommon SNPs, making it a valuable resource for next-generation GWAS with denser chips in new populations. In turn, it is likely that even larger sample sizes will be needed for reference sets to explore SNPs with MAF at or below 1%. The DCEG Reference set has been released to the Database of Genotypes and Phenotypes (dbGaP) as Build 1. We plan to release subsequent builds expanding the number of subjects from diverse populations and adding new genotype content from the Affymetrix 6.0, Omni5 arrays, and future commercial products.

Materials and Methods

DNA Samples and Other data sources

906 subjects were chosen from two clinical trials and 2 prospective cohorts, Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC), Cancer Prevention Study-II of the American Cancer Society (CPS II), the Prostate, Lung, Colon and Ovarian Cancer Prevention Trial (PLCO) and the Shanxi Upper Gastrointestinal Cancer Genetics Project (SHNX)[23,26]. These studies are notable because a large number of subjects have been scanned with first generation SNP arrays as components of more than a dozen multi-stage GWAS of cancer and cancer-related outcomes. All subjects were cancer-free and over the age of 55 at last ascertainment. Subjects of European background were selected from ATBC, CPSII and PLCO; African Americans from PLCO; and East Asians from SHNX. Illumina, Inc. provided data files for 446 Coriell individuals from HapMap3, namely, CEU, TSI, JPT, CHB and YRI populations genotyped on Illumina Omni 2.5 array. For 74 SHNX individuals, genotype data were available for the Illumina 660 array[26] as well as the Omni2.5 M array. 95 African American samples from the Multi Ethnic Cohort[27-28] were genotyped separately at USC with the Illumina HumanHap1 and the 2.5 arrays as a test set to evaluate the performance of the imputation of African American subjects in our reference set which includes 98 African American samples from the PLCO.

Genotyping and quality control

Genotype analysis of samples from the ATBC, CPSII, PLCO and SHNX studies were conducted at the NCI Core Genotyping Facility according to standard operating procedures. For each sample in ATBC, CPSII and PLCO, genotyping was attempted on three different Illumina arrays including Hap1, Omni1-Quad and Omni 2.5. Scanned intensities were clustered for each separate array per subject and genotypes were called using Gentrain2 algorithm within Illumina Genome Studio. 193 duplicates were included in the analysis (59, 63, 48, 21 and 2 for ATBC, CPSII, PLCO, Illumina set and SHNX

respectively). An established quality control (QC) process was applied to samples by study (Referred as "QC Groups") to ensure that only high-quality genotypes were retained for the analytic data set. QC metrics included completion rates by sample or locus, sample heterozygosity rate and duplicate concordance rate and standard thresholds for exclusion of data generated per array were applied. Overall, the results of 198 arrays from 153 different subjects were excluded (Table 2 and **3**). After sample-level QC was completed for each QC Group including data genotyped at Illumina, the average concordance rate for the 193 expected duplicates is greater than 99.9%. Subsequently, genotypes on distinct arrays were merged to subject-level. A total of 17 gender-discordant subjects were excluded on the basis of discrepancies in the mean heterozygosity for X chromosome SNPs. In addition, 98 PLCO African American samples previously genotyped on Hap1 and 74 SHNX samples previously genotyped on Hap660 were added in.

Ancestry was estimated based on a set of informative SNPs[29] using GLU *struct.admix* module; the HapMap build 27 CEU, YRI, ASA (JPT+CHB) samples were used as three continental reference populations. The detected ancestry is concordant with self-reported ethnicity except for two self-described African-American subjects, for whom the data indicate less than 15% non-CEU ancestry, and thus were considered to be CEU ancestry for this study (Figure 10). Identity by descent (IBD) was estimated using the GLU *qc.ibds* module for all pair-wise comparisons to search for both expected and unexpected relatedness among the data set.

We also excluded subjects and loci with discordance rates greater than 1% after merging the genotypes generated from different arrays, resulting in exclusion of five subjects (2 ATBC, 1 CPSII and 2 PLCO). The merged data of array per subject resulted in the exclusion of a number of loci: 9,662 from the ATBC, 6,134 from the CPSII and 10,526 from the PLCO data sets. Assays from Illumina Hap1, Omni1-Quad, Omni2.5 arrays were harmonized based on the locus meta-data of 1000 Genomes June 2010 release and HapMap 3 release 2. Also excluded are an additional 942 loci with incompatible alleles

(either matching directly or by reverse complementing) between our data and the public reference data, and additional 644 loci duplicated on Illumina arrays.

Imputation Scenarios

We chose IMPUTE2 to conduct all analyses because of its speed and the built-in sliding window user interface[13]. We conducted selected analyses with BEAGLE in parallel (using all default configuration and the same sliding window) [12]. Both programs performed very similarly, with IMPUTE2 slightly better than Beagle especially for the loci with MAF < 10%.

We investigated the effect of our new reference set (728 subjects of European ancestry from ATBC, CPSII and PLCO) on accuracy for genotyped SNPs with a MAF of > 1% in comparison to the 1000 Genomes and HapMap 3 dataset. To address the question of how well imputation can be applied to first generation SNP arrays, we simulated content for the Hap660 array in what we designate forward imputation, namely using the Hap660 data to impute content on the Omni arrays. A backward imputation assessed the ability to impute across the new data set using the OmniExpress**.** For this analysis, we randomly selected 60 subjects (20 each from ATBC, CPSII and PLCO), less than 10% of the 653 subjects genotyped on all three Illumina arrays to form the inference set of samples. SNP data for the 60 samples (20 from ATBC, CPSII and PLCO) were masked except for those in the inference set of loci. Random sampling was repeated at least twice more to ensure reproducibility and robustness. The reference loci sets were either the 2.8 million SNPs genotyped as part of this dataset or the 7.8 million SNPs available from the 1000 Genomes project June 2010 release and HapMap 3 February 2009 release. Analyses were done for: (1) DCEG reference set (all samples excluding those used as reference); (2) 1000 Genomes project and HapMap 3 (downloaded from the IMPUTE2 site: http://mathgen.stats.ox.ac.uk/impute/data_download_1000G_pilot_plus_hapmap3.html); and (3) the union of the DCEG reference set and the 1000 Genomes/HapMap 3 data sets. Accuracy was specifically assessed using the subset of 2.0 million SNPs with MAF >1%. In an exploration of the effect of population structure on the imputation accuracy, we evaluated imputation from content on Illumina

OmniExpress to the Omni 2.5 SNPs using a subset (202 ATBC, 202 CPSII or 202 CEU+TSI) to impute genotypes of 255 PLCO samples of European ancestry. For testing the effect of reference size on imputation, the overall combined set of 930 samples was used to assess imputation accuracy in comparison to subsets (50, 100, 200, 400, 600 and 800 randomly selected samples) against a fixed set of 100, randomly chosen. In a preliminary exploration of the utility of the data set in other populations, we evaluated imputation in 94 African Americans drawn from the Multi-Ethnic Cohort, using the OmniExpress content to impute the remaining SNPs on the Omni 2.5.

The metric, allelic dosage based R^2 value, was calculated for each locus by comparing the imputed genotype dosage with the actual assayed genotypes for the inference set. Dosage R^2 is a convenient measure of imputation accuracy since its inverse is related to the decrease in power for case-control association tests. Another advantage to this measure is that it is applicable to low frequency variants, where simpler genotype concordance based measurements become increasingly insensitive. For example, a SNP with a minor allele frequency of 1% achieves a 99% concordance rate by assigning all genotypes to homozygous major allele. In our comparisons, the focus is on GWAS study power rather than purely on imputation accuracy; consequently, directly genotyped SNPs within the Hap660 and OmniExpress subsets are assigned a squared-correlation coefficient $R^2=1$. The curve of $R^2 > 0.8$ for each MAF bin is represented in scatter plots.

Power Estimate

Power calculations assumed a case/control study with the 10,000 individuals divided equally between the two groups. Associations were assumed to be tested by the score statistic. For SNP j, under the null hypothesis of no association, we assumed the test statistic, S_j, was distributed according to a noncentral chi-square distribution with 1df. Let t_α be $1\text{-}10^{-7}$ quantile for this distribution. For SNP j, under the alternative hypothesis, we assumed that S_j was distributed according to a noncentral chi-square distribution,$\chi^2_{\eta j}$, with noncentrality parameter η_j and 1 df. The potential power for SNP j is the $P(\chi^2_{\eta j} >$

t_α) . The estimated power for a GWAS is the average of the potential power across all SNPs. To calculate η_j, we assumed that disease risk followed an additive genetic model and randomly selected an OR based on the distribution described in Park et al [10]. When calculating the power for SNP j in the scenario using imputation, the value of η_j presuming direct genotyping needed to be multiplied by R^2_j for those SNPs not on the Hap660 or OmniExpress.

Acknowledgments

The authors thank all of the participants who took part in this research. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Figure legends

Figure 1. Loci included in the analysis of the imputation reference set.

a) The DCEG Reference Set consists primarily of assays on three commercial arrays totaling ~2.8 million autosomal polymorphic loci that passed rigorous QC criteria. Note that Hap660 is entirely nested within Hap1 except for 14 assays (not depicted). Numbers are in 1,000 units of SNPs.
b) Approximately 700,000 OmniExpress assays (yellow) were simulated from the DCEG Reference Set (blue), of which 683,000 loci exist in the 1000 Genome + HapMap3 except for 22,000 loci. Imputation performance using either reference set was compared in the overlapping loci by both reference sets (1,421,000+683,000 = 2,104,000). The 1000 Genomes + HapMap3 reference is from the IMPUTE2 website, which includes CEU of the 1000 Genome low-coverage data (June 2010) and HapMap3 data (February 2009). Numbers are in 1,000 units of SNPs.

Figure 2. Minor Allele Frequency (MAF) distributions for SNPs on Illumina Omni2.5M array across study/populations.

Figure 3. Imputation accuracy for European-American data with DCEG and public reference set.

The figure depicts the proportion of SNPs with allelic dosage $R^2 > 0.8$ by MAF, shown on the log scale to emphasize differences at smaller values. Solid red depicts imputation of Hap660 data using the DCEG Reference Set. Dashed red depicts imputation of Hap660 using the 1000 Genome plus HapMap3 reference. Solid blue depicts imputation of OmniExpress data using the DCEG Reference Set. Dashed blue depicts imputation of OmniExpress using the 1000 Genome plus HapMap3 reference.

Figure 4. Imputation accuracy for European-American data with DCEG and public reference set (more detailed R^2 distribution by MAF).

The proportion of SNPs with five different allelic dosage R^2 ranges for each MAF bin. (a) Impute OmniExpress data using 1000 Genomes plus HapMap 3 reference; and (b) Impute OmniExpress data using DCEG reference.

Figure 5. The combination of the DCEG Reference Set with the 1000 Genomes and HapMap 3 results in no improvement in imputation performance.

The figure depicts the proportion of SNPs with allelic dosage $R^2 > 0.8$ by MAF, shown on the log scale to emphasize differences at smaller values. Solid red depicts imputation of Hap660 data using the DCEG Reference Set. Dashed red depicts imputation of Hap660 using the 1000 Genome plus HapMap3 reference. Solid blue depicts imputation of OmniExpress data using the DCEG Reference Set. Dashed blue depicts imputation of OmniExpress using the 1000 Genome plus HapMap3 reference.

Figure 6. Principal component analysis of populations in ATBC, CPSII, PLCO and HapMap CEU, TSI.

Figure 7. Imputation accuracy for European-American data with matched and mismatched reference sub-populations.

The proportion of SNPs with allelic dosage $R^2 > 0.8$ by MAF, is shown on the log scale to emphasize differences at smaller values. Each scenario measures accuracy of imputing OmniExpress data for 255 European-American individuals from the PLCO cohort with the following reference data: (ATBC) 202 ATBC individuals from Finland; (CPSII) 202 CPSII European-American individuals; (CEU+TSI) 202 HapMap individuals of European-descent from Utah and Northern Italy.

Figure 8. Imputation accuracy and size of the reference set.

The proportion of SNPs with allelic dosage $R^2 > 0.8$ by MAF, is shown on the log scale to emphasize differences at smaller values. The same set of randomly chosen 100 samples was used as the inference set whereas the reference set varied in size by 50, 100, 200, 400, 600 and 800 respectively. All scenarios were imputing from OmniExpress to the contents of Omni 2.5.

Figure 9. Imputation accuracy for African-American sample set.

The proportions of SNPs with allelic dosage $R^2 > 0.8$ by MAF, is shown on the log scale to emphasize differences at smaller values. Each scenario measures accuracy of imputing OmniExpress data for 94

African-American individuals from the MEC cohort[25] with the following reference data for 1.4 M SNPs in the 1000 Genome Yoruba and HapMap set; Solid blue corresponds to the DCEG Reference Set, Solid red corresponds to the 1000 Genome plus HapMap, Solid green corresponds to the subset only of the 98 PLCO African Americans in the DCEG Reference Set. The dashed blue corresponds to the set of 2 M SNPs with MAF > 1% in the DCEG Reference Set.

Figure 10. STRUCTURE plot of the DCEG Reference Set and MEC data

The analysis was conducted using HapMap CEU+TSI, JPT+CHB and YRI as three continental reference sets. The admixture coefficients for ATBC, CPSII, MEC, PLCO African American (PLCO_AA), PLCO European American (PLCO_EUR) and SHNX are shown along the edges of the triangle. African American samples from both PLCO (black) and MEC (cyan) show similar distribution along the AFR and EUR axis.

Table legends

Table 1. Samples included in the imputation reference set

Subjects passing quality control metrics for the SNP arrays indicated in the right hand columns. This table reports the content of Build 1.

Table 2. QC exclusion threshold

Table 3. Summary of excluded loci

URL

IMPUTE2: http://mathgen.stats.ox.ac.uk/impute/data_download_1000G_pilot_plus_hapmap3.html

References

1. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-7 (2009).

2. Chanock, S.J. *et al.* Replicating genotype-phenotype associations. *Nature* **447**, 655-60 (2007).
3. Wacholder, S. *et al.* Performance of common genetic variants in breast-cancer risk models. *N Engl J Med* **362**, 986-93 (2010).
4. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
5. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-61 (2007).
6. Altshuler, D.M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-8 (2010).
7. Barrett, J.C. & Cardon, L.R. Evaluating coverage of genome-wide association studies. *Nat Genet* **38**, 659-62 (2006).
8. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-8 (2010).
9. Voight, B.F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* **42**, 579-89 (2010).
10. Park, J.H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet* **42**, 570-5 (2010).
11. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).
12. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**, 210-23 (2009).
13. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
14. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**, 816-34 (2010).
15. Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M. & Franke, A. A comprehensive evaluation of SNP genotype imputation. *Hum Genet* **125**, 163-71 (2009).
16. Huang, L. *et al.* Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet* **84**, 235-50 (2009).
17. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
18. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511 (2010).
19. The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. The ATBC Cancer Prevention Study Group. *Ann Epidemiol* **4**, 1-10 (1994).
20. Calle, E.E. *et al.* The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. *Cancer* **94**, 2490-501 (2002).
21. Gohagan, J.K., Prorok, P.C., Hayes, R.B. & Kramer, B.S. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. *Control Clin Trials* **21**, 251S-272S (2000).
22. Prorok, P.C. *et al.* Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Control Clin Trials* **21**, 273S-309S (2000).
23. Ke, L. Mortality and incidence trends from esophagus cancer in selected geographic areas of China circa 1970-90. *Int J Cancer* **102**, 271-4 (2002).
24. Sampson, J. *et al.* A two-platform design for next generation genome-wide association studies. (2011).
25. Kolonel, L.N. *et al.* A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am J Epidemiol* **151**, 346-57 (2000).

26. Abnet, C.C. *et al.* A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nat Genet* **42**, 764-7 (2010).
27. Haiman, C.A. *et al.* Characterizing genetic risk at known prostate cancer susceptibility Loci in african americans. *PLoS Genet* **7**, e1001387 (2011).
28. Haiman, C.A. *et al.* Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. *Nat Genet* **43**, 570-3 (2011).
29. Yu, K. *et al.* Population substructure and control selection in genome-wide association studies. *PLoS One* **3**, e2551 (2008).

Table 1. Samples included in DCEG Reference Set

	Populations				Illumina Array			
Group	**European American**	**African American**	**African**	**Asian**	**Hap660**	**Hap1**	**Omni1**	**Omni2.5**
ATBC	246					✓	✓	✓
CPSII	227					✓	✓	✓
PLCO	255					✓	✓	✓
PLCO		98				✓		✓
SHNX				74	✓			✓
HapMap								
CEU	116							✓
CHB				44				✓
JPT				44				✓
TSI	86							✓
YRI			59					✓
Total	930	98	59	162				

Table 2. QC exclusion thresholds

QC group	allowed sample heterozygosity	max. sample missing rate	max. locus missing rate
ATBC Omni2.5	0.17 - 0.19	0.02	0.05
ATBC Hap1	0.25 - 0.27	0.03	0.06
ATBC Omni1	0.24 - 0.27	0.03	0.05
CPSII Omni2.5	0.17 - 0.20	0.04	0.06
CPSII Hap1	0.26 - 0.28	0.04	0.06
CPSII Omni1	0.25 - 0.27	0.04	0.06
HapMap Omni2.5	0.16 - 0.22	0.01	0.04
PLCO Omni2.5	0.17 - 0.22	0.04	0.05
PLCO Hap1	0.25 - 0.28	0.04	0.06
PLCO Omni1	0.24 - 0.27	0.02	0.05
SHNX Omni2.5	0.16 - 0.18	0.04	0.06

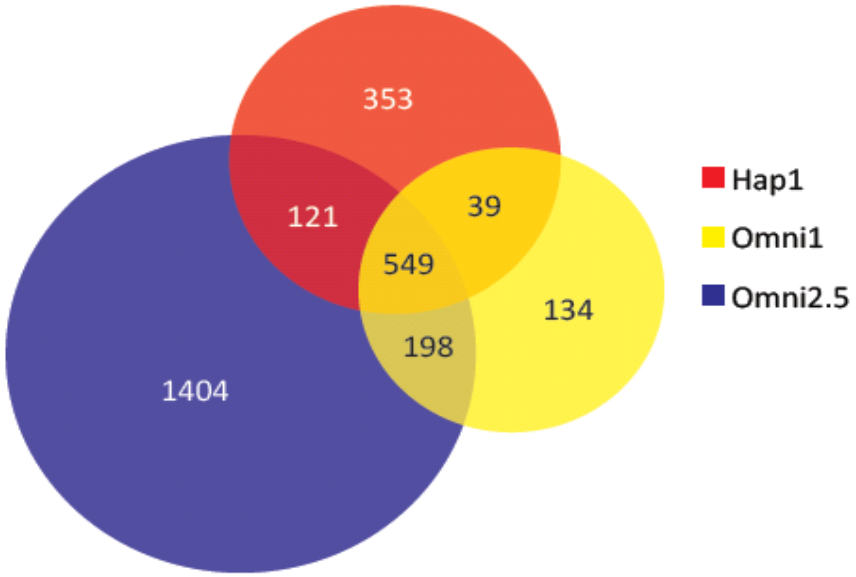
Table 3. Summary of excluded loci and samples

	Locus Exclusions	Sample Exclusions			
QC group	**Missing counts**	**Hetero-zygosity**	**Missing counts**	**Discordant duplicates**	**Total***
ATBC Omni2.5	20,224	1	6		6
ATBC Hap1	54,513	5	13		14
ATBC Omni1	132,017	1	10		11
CPSII Omni2.5	52,990	7	33		37
CPSII Hap1	64,691	8	20		25
CPSII Omni1	148,472	6	20	4	29
HapMap Omni2.5	7,551	1	0		1
PLCO Omni2.5	21,616	2	23		24
PLCO Hap1	66,124	10	33		35
PLCO Omni1	135,192	0	12		12
SHNX Omni2.5	53,971	0	4		4

* Count of unique samples excluded. Some samples were excluded for both excess heterozygosity and missing rates.

Figure 1. Loci included in the analysis of the imputation reference set.

(a)



(b)

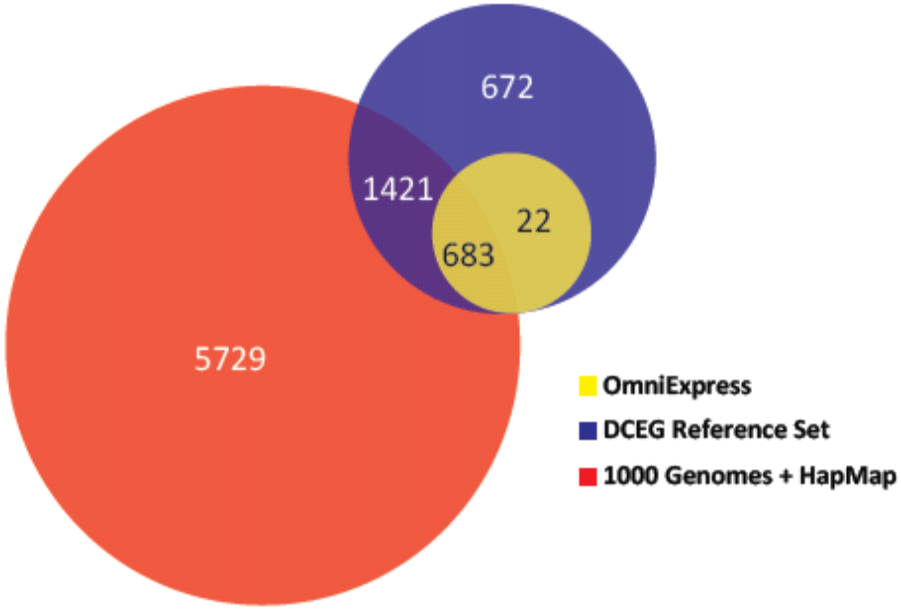


Figure 2. Minor Allele Frequency (MAF) distributions for SNPs on Illumina Omni2.5M array across study/populations

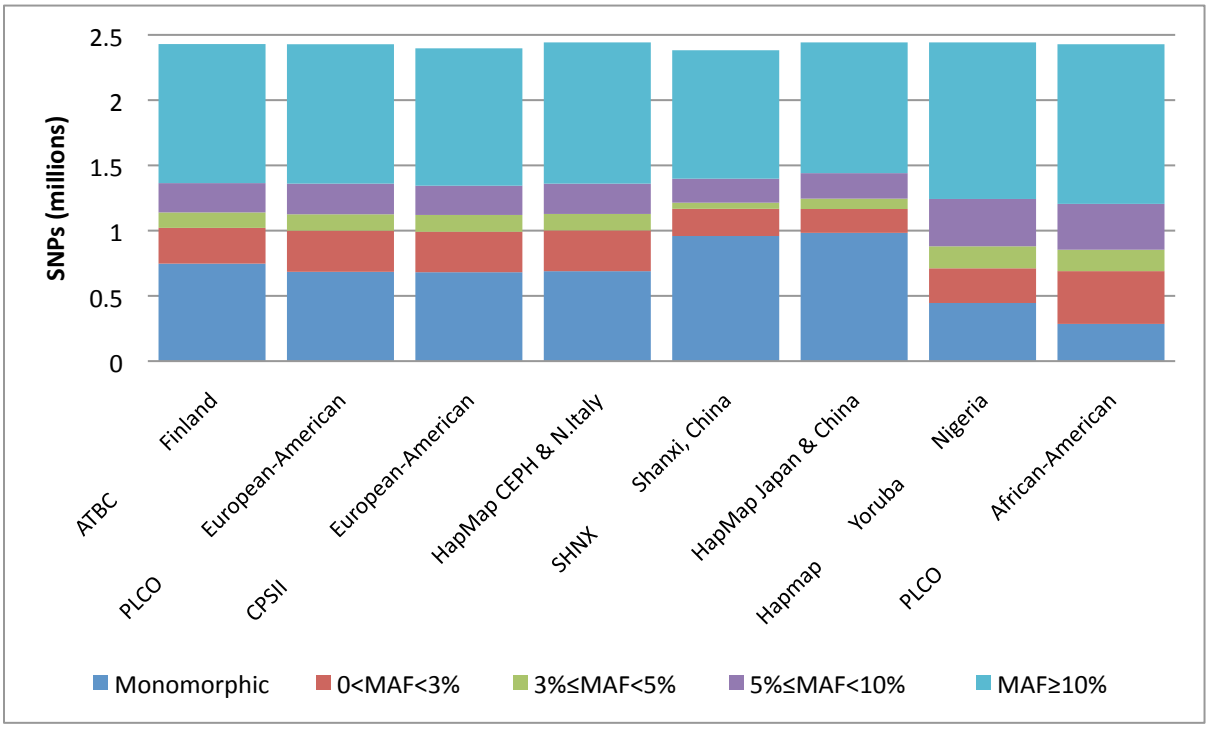


Figure 3. Imputation accuracy for European-American data with DCEG and public reference set

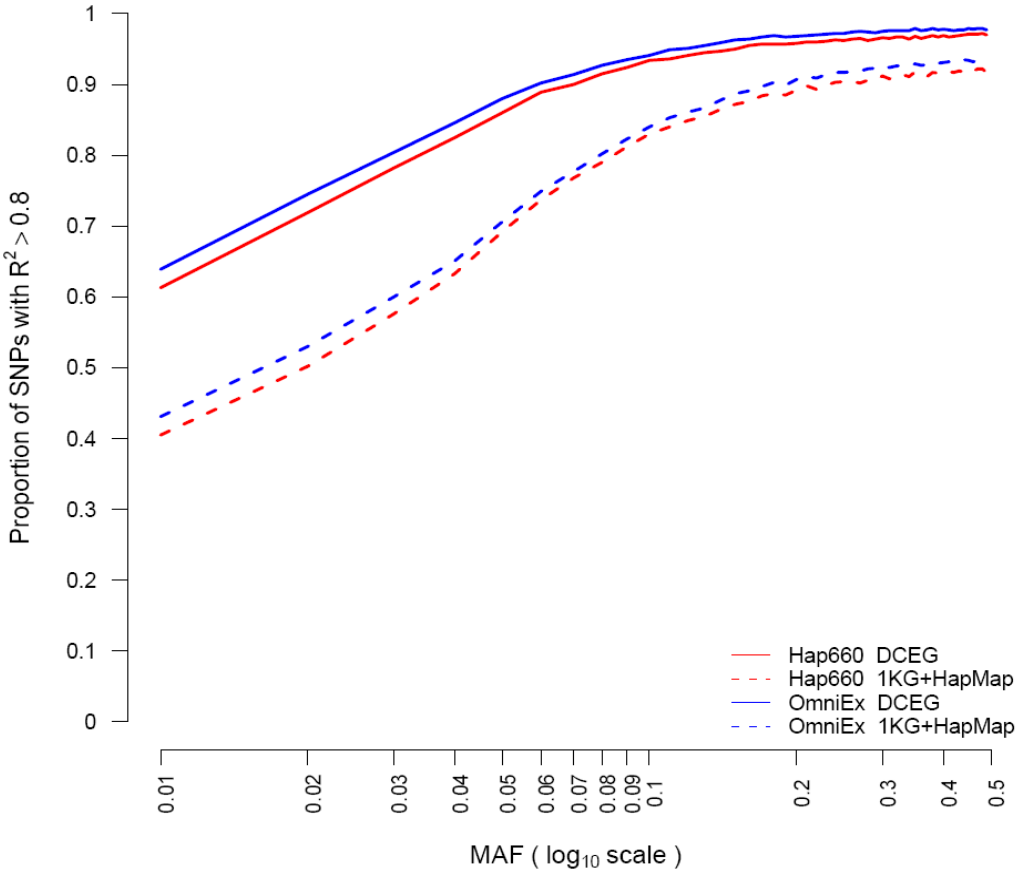
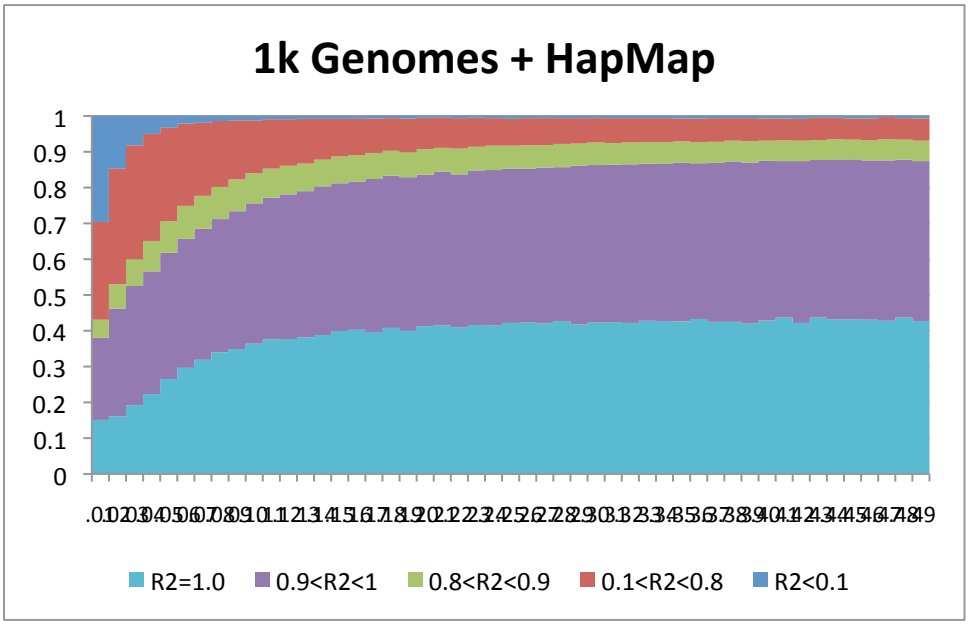


Figure 4. Imputation accuracy for European-American data with DCEG and public reference set (more detailed R^2 distribution by MAF)

a)



b)

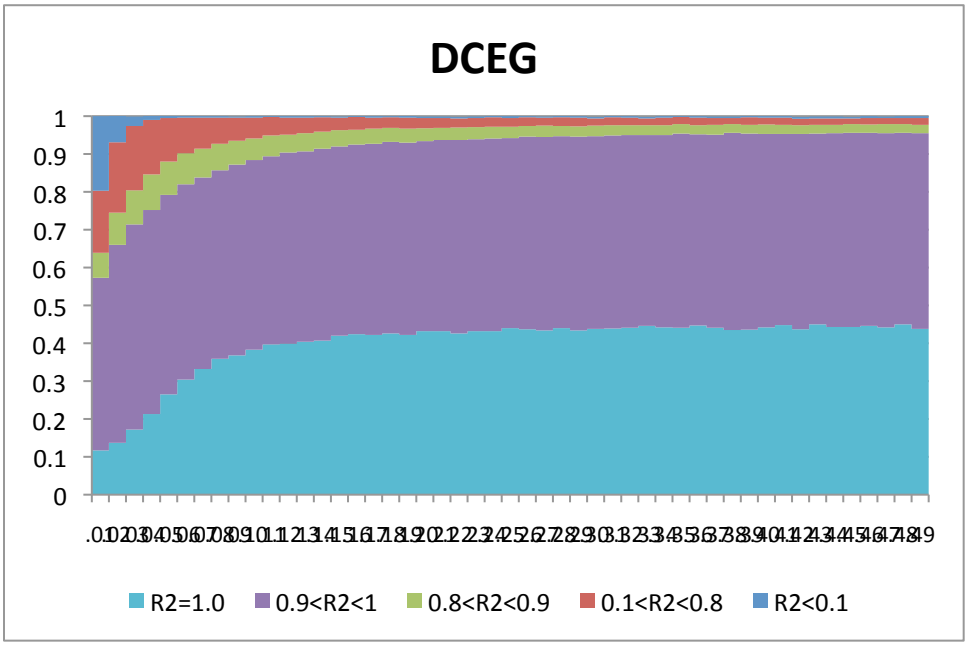


Figure 5. The combination of the DCEG Reference Set with the 1000 Genomes and HapMap 3 results in no improvement in imputation performance.

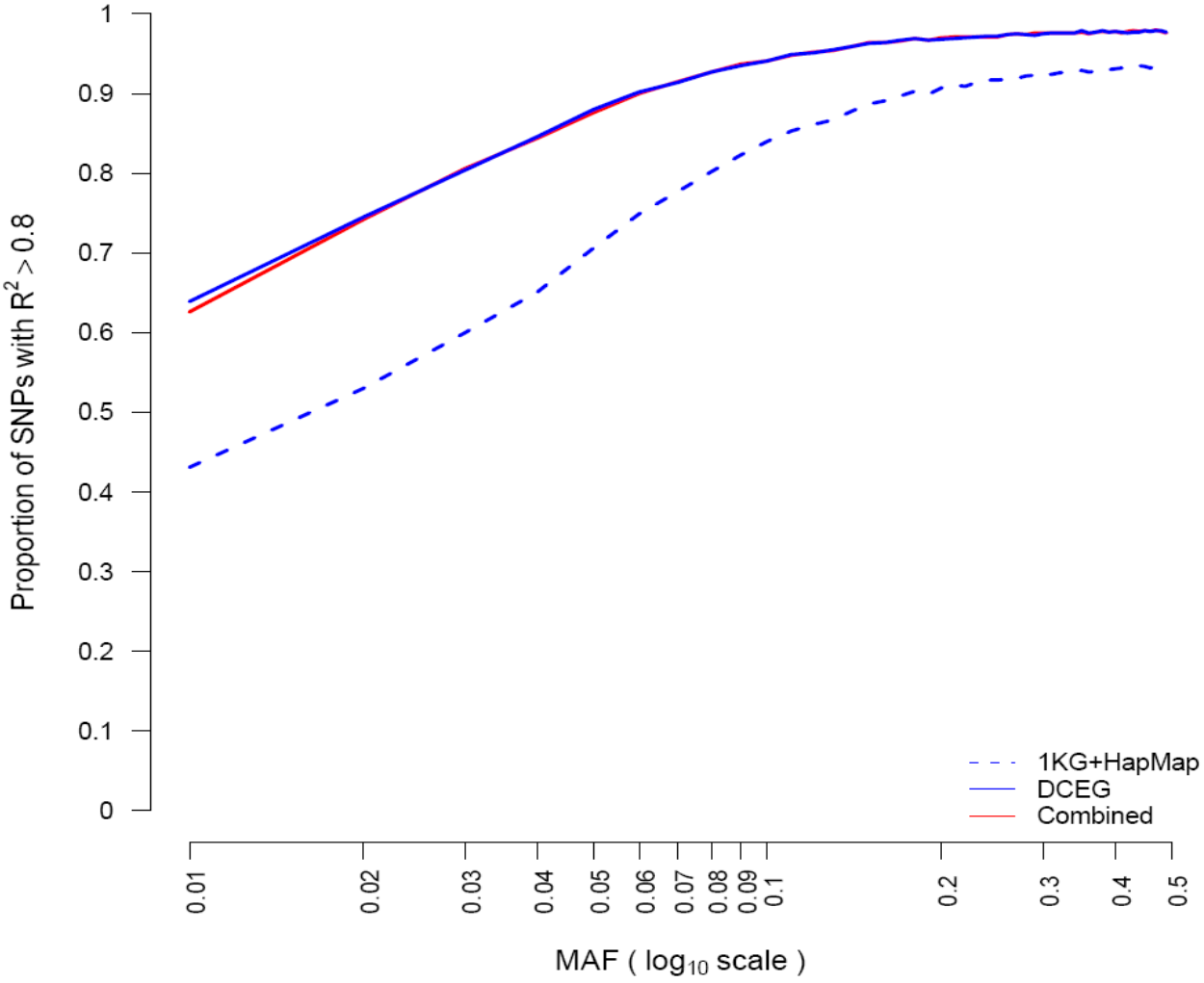


Figure 6. Principal component analysis of populations in ATBC, CPSII, PLCO and HapMap CEU, TSI.

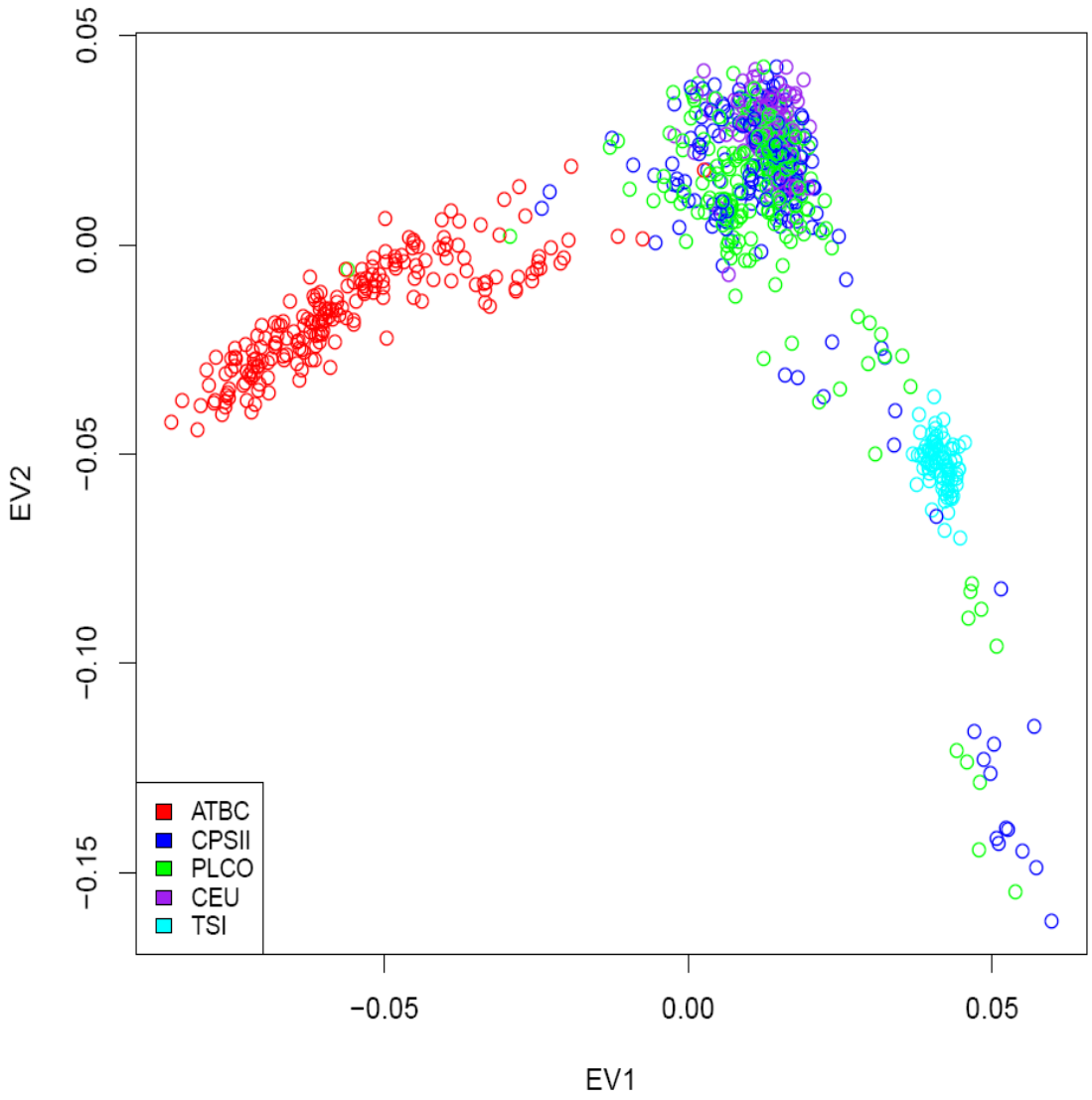


Figure 7. Imputation accuracy for European-American data with matched and mismatched reference sub-populations.

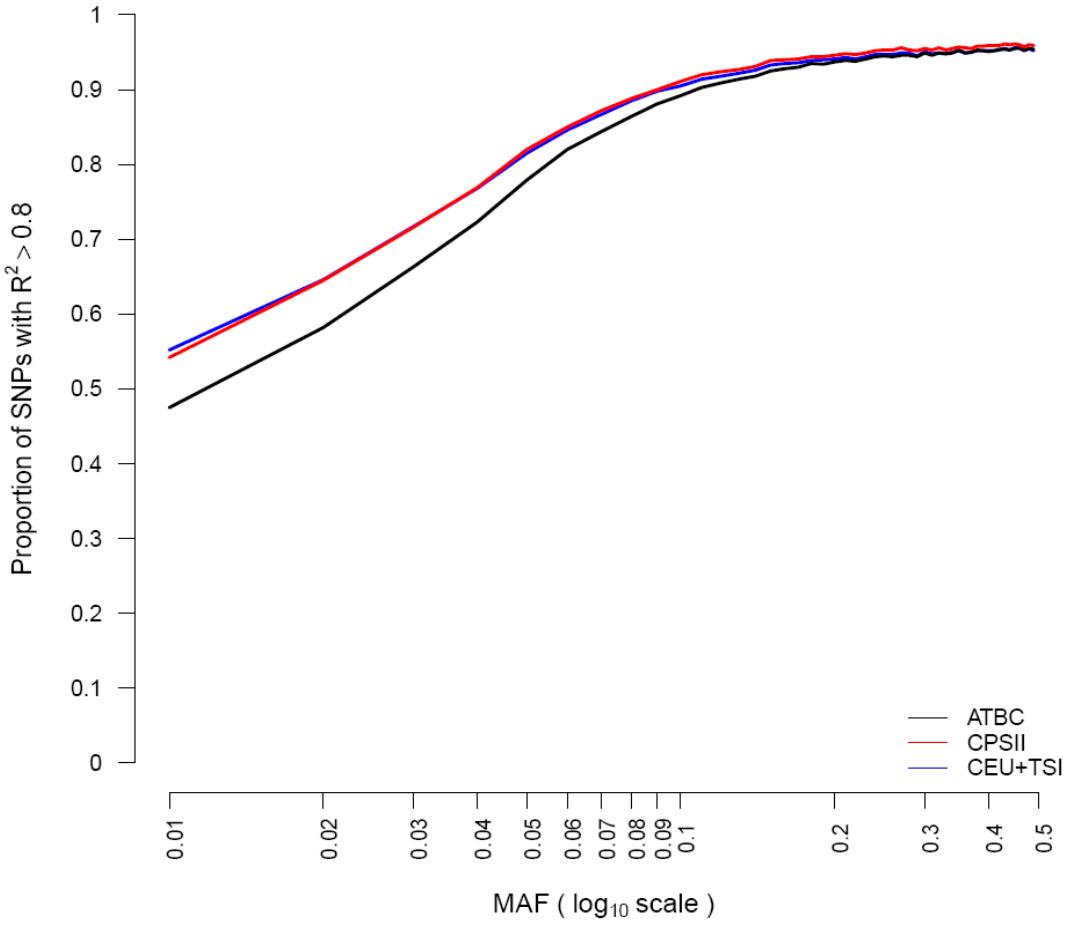


Figure 8. Imputation accuracy and size of the reference set.

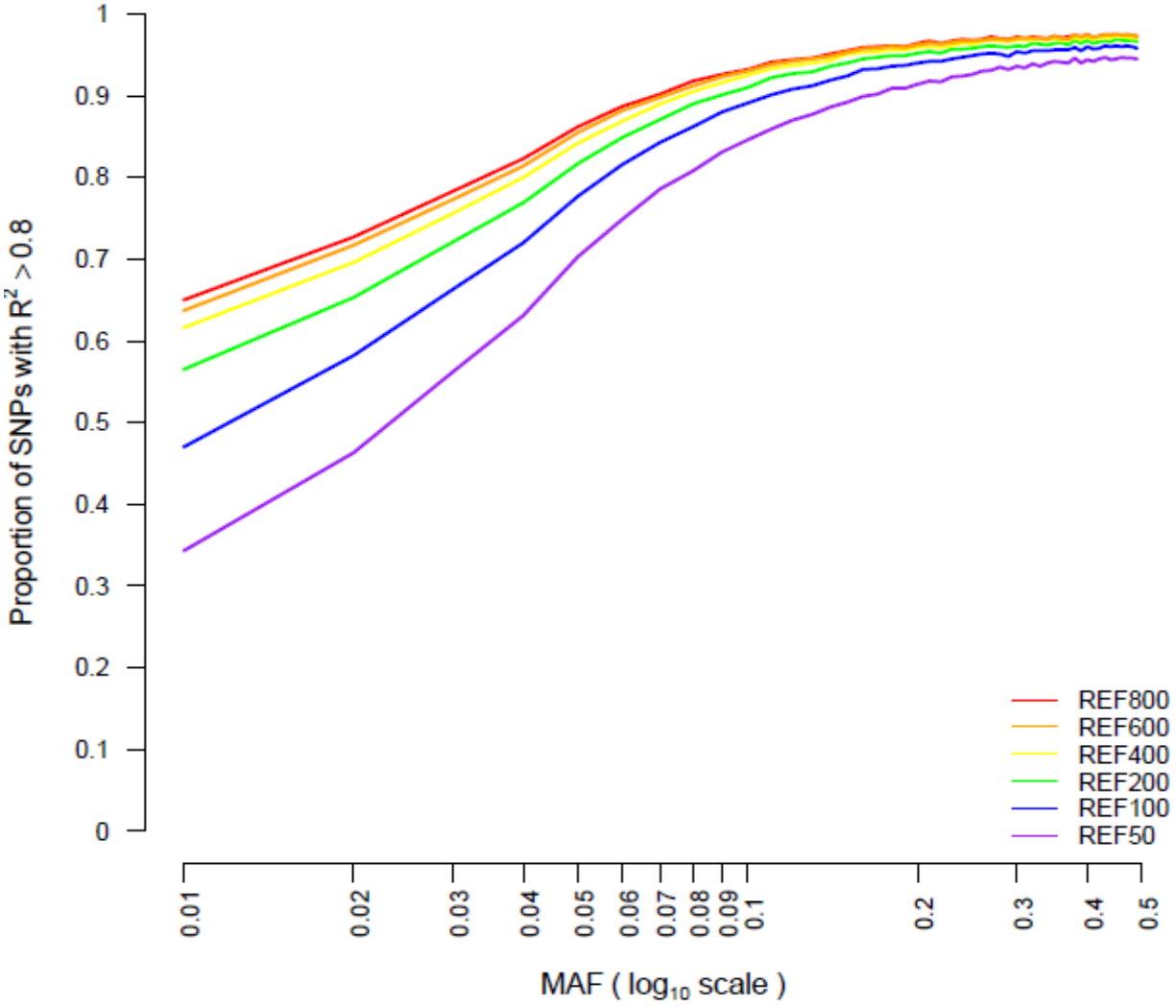


Figure 9. Imputation accuracy for an African-American sample set.

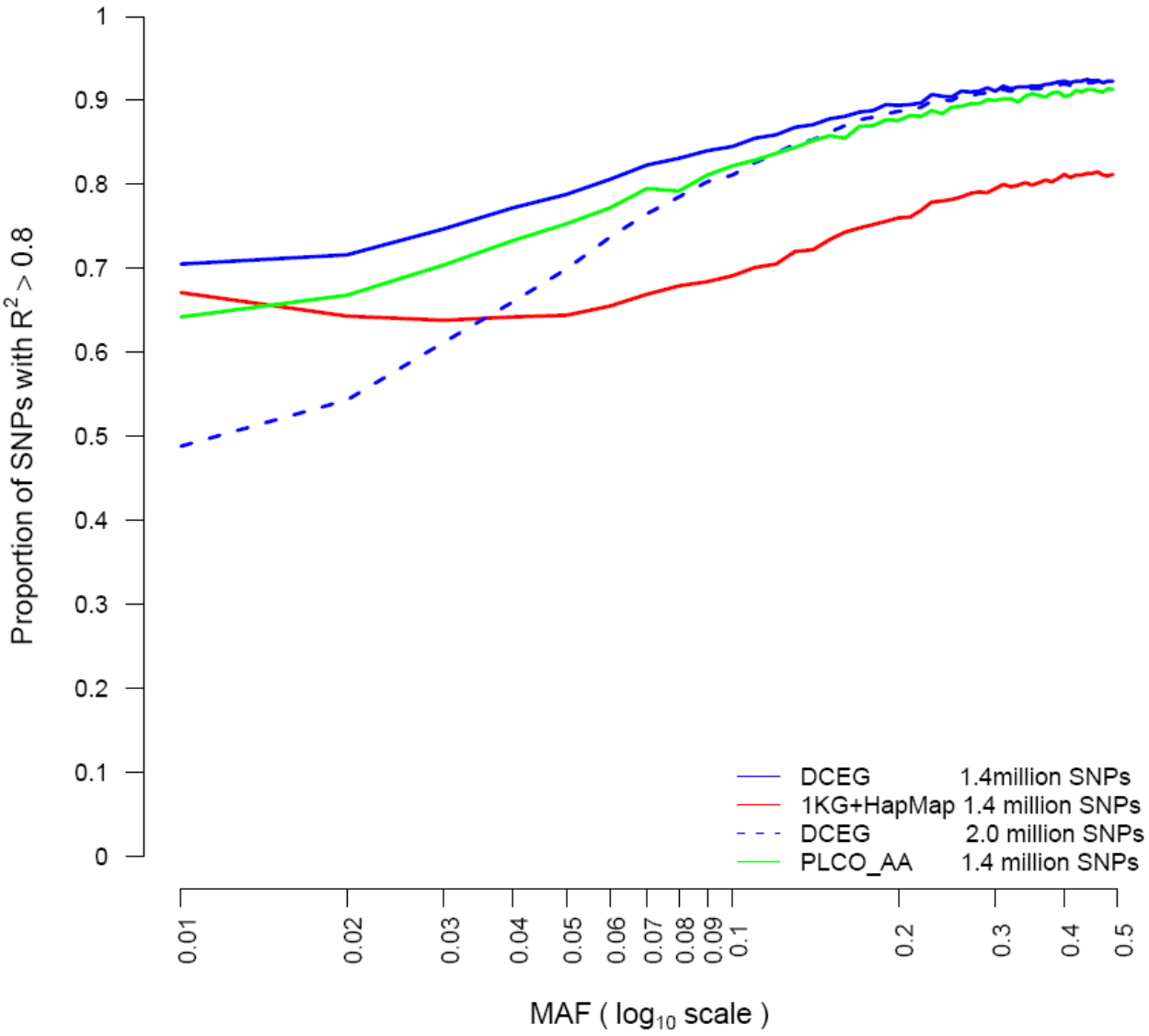


Figure 10. STRUCTURE plot of the DCEG Reference Set and MEC data