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

Supplementary Methods

Materials and methods

906 individuals 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) [7,12]. All individuals were cancer-free and over the age of 55 at last ascertainment. Individuals of European ancestry 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 Hap660 array [12] as well as the Omni 2.5 array. 95 African American samples from the Multi Ethnic Cohort (MEC) [11,13-14] were genotyped at USC with the Illumina Hap1 and the Omni 2.5 arrays.

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 Illumina arrays, Hap1, Omni1 and Omni2.5. 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 to 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 QC Group were applied. The results of 198 samples from 153 different individuals were excluded (**Supplementary Tables 1,2**). After sample-level QC was completed for each QC Group, the average concordance rate for the 193 expected duplicates is greater than 99.9%. Genotypes on distinct arrays were merged to subject-level. A total of 17 gender-discordant individuals were excluded on the basis of discrepancies in the mean heterozygosity for SNPs mapped to the chromosome X.

Ancestry was estimated based on a set of informative SNPs [15] using GLU *struct.admix* module; the HapMap build 27 CEU, YRI, ASA (JPT+CHB) samples were used as three continental reference populations. The estimated ancestry is concordant with self-reported ethnicity except for two self-described African-American individuals, for whom the data indicate less than 15% non-CEU ancestry, and thus were considered to be of European ancestry for this study (**Supplementary Figure 3**).

We also excluded individuals and loci with discordance rates greater than 1% after merging the genotypes generated from different arrays, resulting in exclusion of 5 individuals (2 ATBC, 1 CPSII and 2 PLCO). Assays from Illumina Hap1, Omni1, Omni2.5 arrays were harmonized based on the locus meta-data of 1000 Genomes June 2010 release and HapMap 3 February 2009 release. An additional 942 loci were excluded due to incompatible alleles (neither a direct match nor a reverse complement) between our data and the public reference data, and 644 loci were excluded due to duplication on Illumina arrays.

We investigated the effect of the DCEG Reference Set on imputation accuracy for SNPs with a MAF > 1% in comparison to the 1000 Genomes and HapMap 3 datasets. We simulated contents of either Hap660 or OmniExpress array for the 60 individuals randomly selected (20 each from ATBC, CPSII and PLCO), which were genotyped on all three Illumina arrays. SNP data for these 60 individuals were masked except for those in the inference set of loci. Random sampling was repeated at least twice more without re-sampling to ensure reproducibility and robustness. The reference loci sets were either the 2.8 million SNPs genotyped in this dataset or the 7.8 million SNPs available from the 1000 Genomes project June 2010 release CEU population and HapMap 3 February 2009 release. Analyses were done for: (1) DCEG Reference Set (all individuals excluding those used as inference); (2) 1000 Genomes project and HapMap 3; and (3) union of the DCEG Reference Set and the 1000 Genomes/HapMap 3 data sets. Accuracy was specifically assessed using 2.0 million common reference SNPs with MAF >1%. In an exploration of the effect of population structure on the imputation accuracy, we evaluated imputation from content on OmniExpress to the Omni2.5 SNPs using a subset of DCEG reference individuals (202 ATBC, 202 CPSII or 202 HapMap CEU+TSI) to impute genotypes of 255 PLCO individuals of European ancestry. In a preliminary exploration of the utility of the data set in other populations, we evaluated imputation in 94 African Americans drawn from the MEC (One subject was excluded due to QC), using the OmniExpress content to impute the remaining SNPs on the Omni 2.5.

The imputation accuracy metric, the squared-Pearson correlation coefficient (R^2) on allelic dosage, 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. We focused 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.

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 - 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 $P(\chi^2_{\eta j} \geq t_\alpha)$ 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.

References

12. 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).

13. Haiman, C.A. *et al.* Characterizing genetic risk at known prostate cancer susceptibility Loci in african americans. *PLoS Genet* **7**, e1001387 (2011).

14. 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).

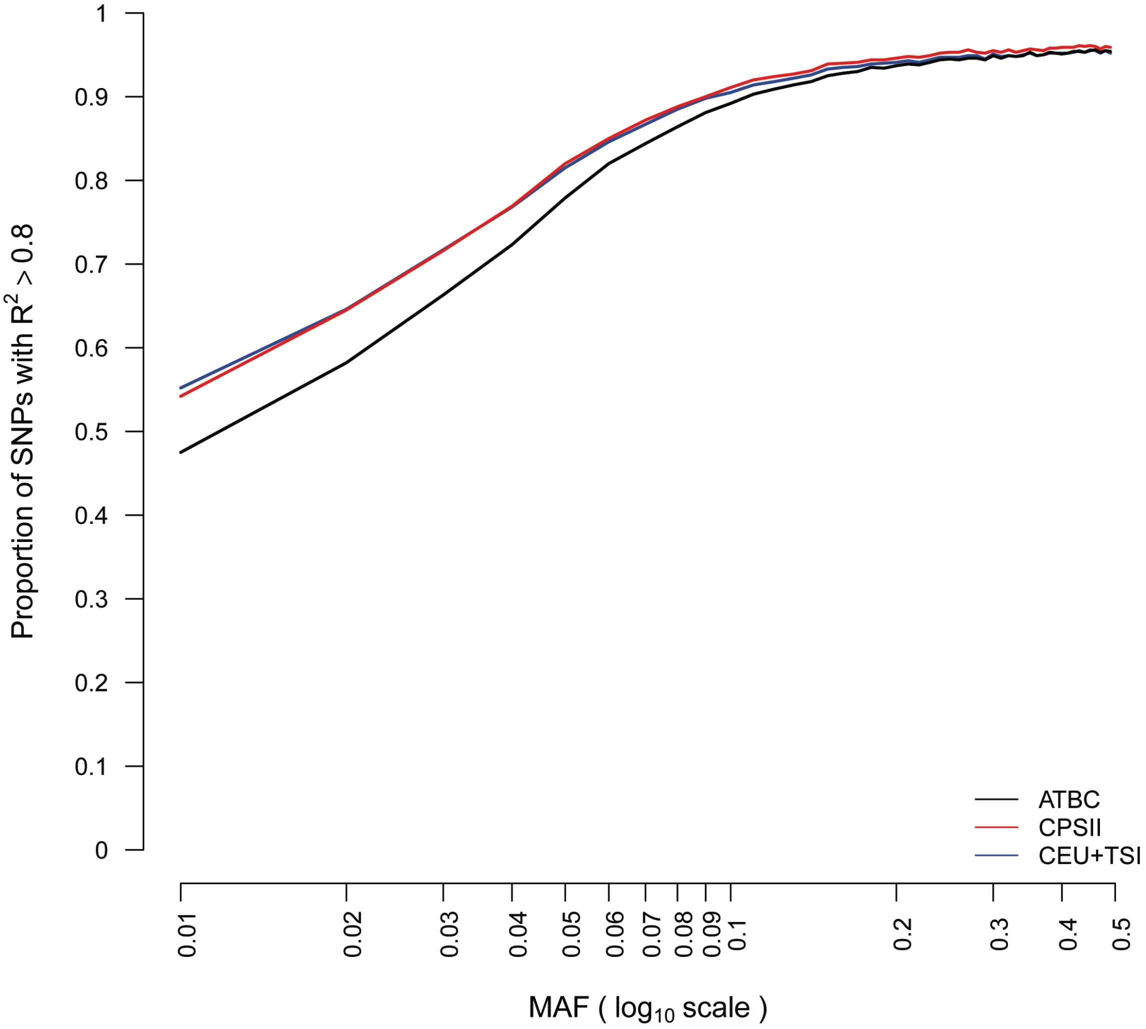
15. Yu, K. *et al.* Population substructure and control selection in genome-wide association studies. *PLoS One* **3**, e2551 (2008).

URLs

dbGAP	http://www.ncbi.nlm.nih.gov/gap
GLU	http://code.google.com/p/glu-genetics/
HapMap	http://hapmap.ncbi.nlm.nih.gov/
1000 Genomes Project	http://www.1000genomes.org/

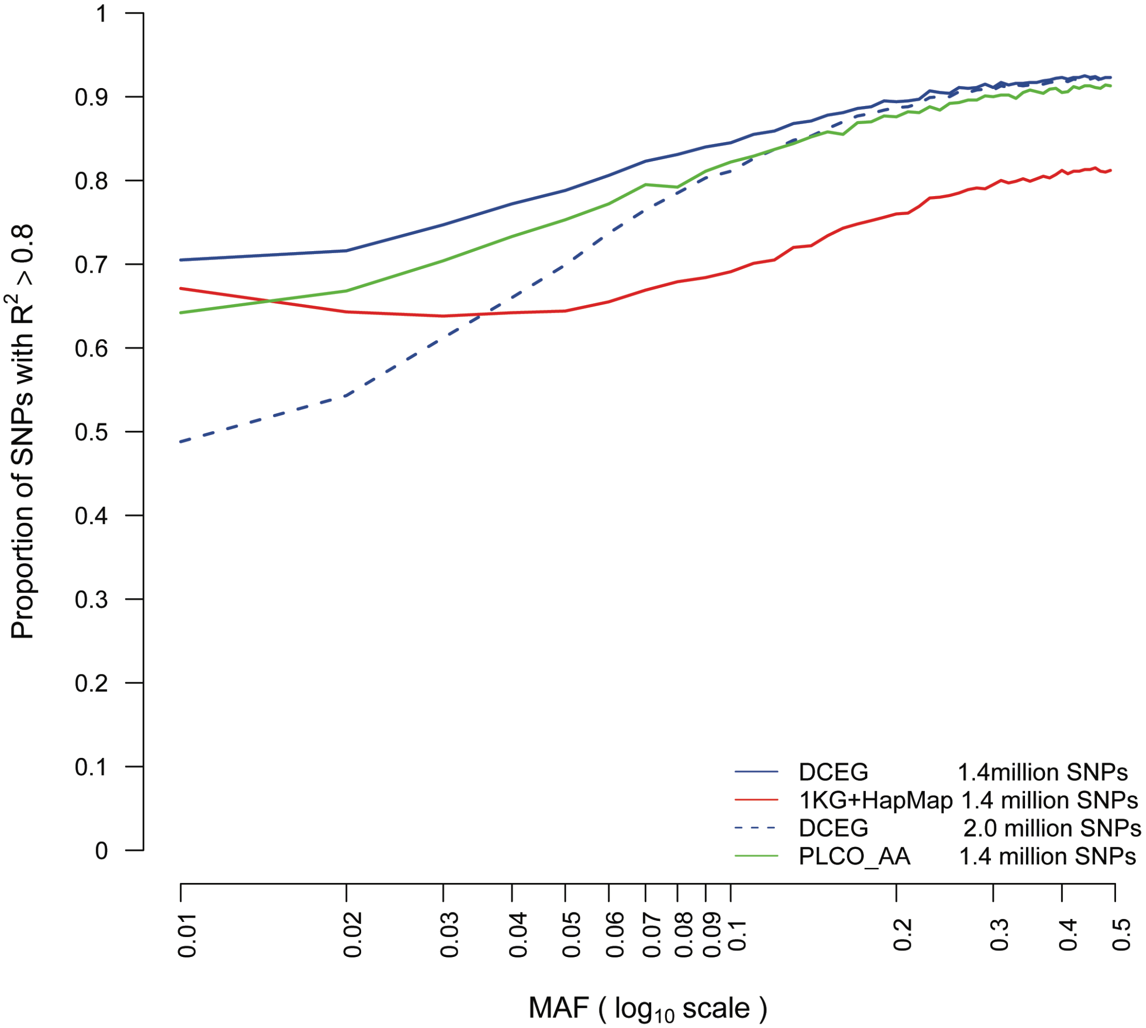
Supplementary Figures

Supplementary Figure 1. Imputation accuracy for individuals of European ancestry with matched and mismatched reference sub-populations.



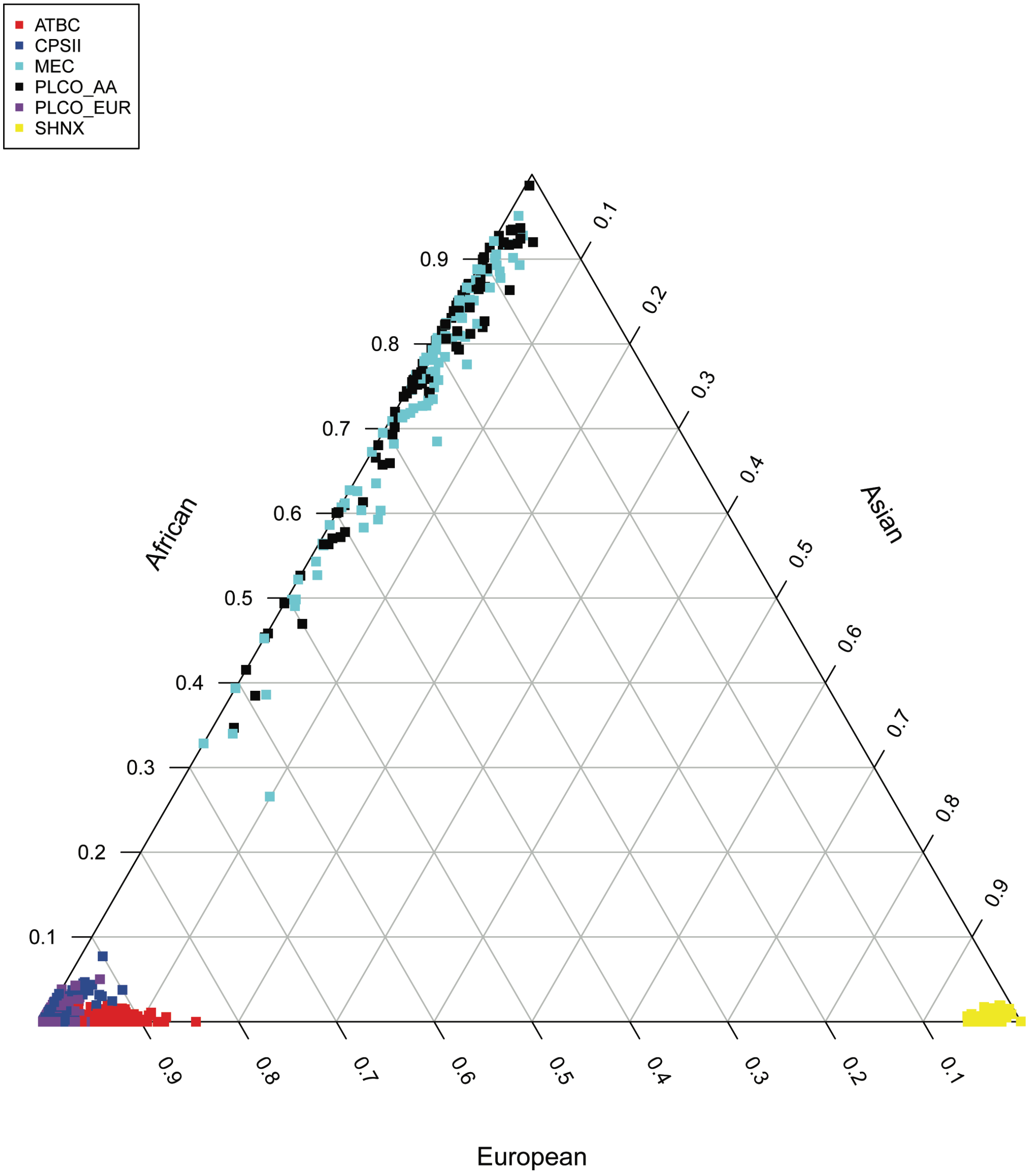
The proportion of SNPs with allelic dosage $R^2 > 0.8$ by MAF, is shown on the log scale to emphasize differences at smaller values. Each scenario measures accuracy of imputing OmniExpress data for 255 European-American individuals from the PLCO cohort with the following reference data: (ATBC) 202 ATBC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

Supplementary Figure 2. Imputation accuracy for an 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 with the following reference data for 1.4 million SNPs in the 1000 Genome Yoruba and HapMap 3 set; Solid blue corresponds to the DCEG Reference Set, Solid red corresponds to the 1000 Genome plus HapMap3, 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 million SNPs with MAF > 1% in the DCEG Reference Set.

Supplementary Figure 3. 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.

Supplementary Table 1. 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

Supplementary Table 2. Summary of excluded loci and samples

	Locus Exclusions	Sample Exclusions			
QC group	**missing rate**	**hetero-zygosity**	**missing rate**	**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.