BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 21 2013, pages 2744–2749
doi:10.1093/bioinformatics/btt477

Genetics and population analysis

Advance Access publication August 16, 2013

Imputation of coding variants in African Americans: better performance using data from the exome sequencing project

Qing Duan[1,†], Eric Yi Liu[2,†], Paul L. Auer[3,†], Guosheng Zhang[1,†], Ethan M. Lange[1,4], Goo Jun[5], Chris Bizon[6], Shuo Jiao[3], Steven Buyske[7,8], Nora Franceschini[9], Chris S. Carlson[3], Li Hsu[3], Alex P. Reiner[3], Ulrike Peters[3,10], Jeffrey Haessler[3], Keith Curtis[3], Christina L. Wassel[11], Jennifer G. Robinson[12], Lisa W. Martin[13], Christopher A. Haiman[14], Loic Le Marchand[15], Tara C. Matise[8], Lucia A. Hindorff[16], Dana C. Crawford[17], Themistocles L. Assimes[18], Hyun Min Kang[5], Gerardo Heiss[9], Rebecca D. Jackson[19], Charles Kooperberg[3], James G. Wilson[20], Gonçalo R. Abecasis[5], Kari E. North[9], Deborah A. Nickerson[21], Leslie A. Lange[1,†] and Yun Li[1,2,4,*,†,‡]

[1]Department of Genetics and [2]Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599, USA, [3]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA, [4]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA, [5]Department of Biostatistics and Center for Statistical Genetics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA, [6]Renaissance Computing Institute, University of North Carolina, Chapel Hill, NC 27599, USA, [7]Department of Statistics and [8]Department of Genetics, Rutgers University, Piscataway, NJ 08854, USA, [9]Department of Epidemiology, University of North Carolina, Chapel Hill, NC 27599, USA, [10]Department of Epidemiology, University of Washington, Seattle, WA 98195, USA, [11]Division of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA, [12]Department of Epidemiology and Medicine, University of Iowa, Iowa City, IA 52242, [13]Division of Cardiology, George Washington University School of Medicine and Health Sciences, Washington, DC 20037, USA, [14]Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, CA 90033, USA, [15]Epidemiology Program, University of Hawaii Cancer Center, HI 96813, USA, [16]Division of Genomic Medicine, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA, [17]Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University, Nashville, TN 37232, USA, [18]Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA, [19]Division of Endocrinology, Diabetes and Metabolism, Ohio State University, Columbus, OH 43210, USA, [20]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 39216, USA and [21]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: Although the 1000 Genomes haplotypes are the most commonly used reference panel for imputation, medical sequencing projects are generating large alternate sets of sequenced samples. Imputation in African Americans using 3384 haplotypes from the Exome Sequencing Project, compared with 2184 haplotypes from 1000 Genomes Project, increased effective sample size by 8.3–11.4% for coding variants with minor allele frequency <1%. No loss of imputation quality was observed using a panel built from phenotypic extremes. We recommend using haplotypes from Exome Sequencing Project alone or concatenation of the two panels over quality score-based post-imputation selection or IMPUTE2's two-panel combination.

Contact: yunli@med.unc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 24, 2013; revised on August 6, 2013; accepted on August 10, 2013

1 INTRODUCTION

Genotype imputation is a common practice for both genotyping (De Bakker *et al*., 2008; Li *et al*., 2009; Marchini and Howie, 2010) and sequencing studies (Fridley *et al*., 2010; Li *et al*., 2011). Increasingly large reference panels available in the public domain [e.g. those from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010, 2012) and UK10K project (Futema *et al*., 2012)] together with improved statistical methods (Howie

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first four and the last two authors should be regarded as joint First Authors.
‡On behalf of the National Heart, Lung, and Blood Institute GO Exome Sequencing Project.

et al., 2012; Liu *et al*., 2013) have enhanced imputation quality, especially for rare variants with minor allele frequency (MAF) <5%. These improvements have resulted in both discovery and refined mapping of association with complex traits (Auer *et al*., 2012; Holm *et al*., 2011; Huang *et al*., 2012). However, few studies have examined the use of large study-specific reference panels, particularly the use of exome sequencing-derived panels in admixed populations. Here, we present a new resource for imputation in African Americans, built from 1692 African Americans sequenced by the Exome Sequencing Project (ESP) (Tennessen *et al*., 2012). We assessed the use of the ESP data as an imputation reference panel and compared the results with those obtained using the 1000 Genomes Project Phase1 (1000G; version 3, March 2012 release) (The 1000 Genomes Project Consortium, 2012) data. Additionally, we evaluated the potential consequences of using a reference panel built from samples selected on the basis of phenotypic extremes or disease status instead of a population-based random sample. Lastly, we compared multiple approaches to combine the ESP and 1000G panels for the imputation of rare coding variants.

2 METHODS

2.1 Exome Sequencing Project

2.1.1 ESP and African American Participants The complete ESP dataset (Fu *et al*., 2012) consists of whole exome data for 6823 individuals. Samples were sequenced at the University of Washington (SeattleGO) and the Broad Institute (BroadGO). Among the 6823 individuals, 1692 participants were African Americans with genome-wide association data available for analysis. The 1692 African Americans ESP samples include 845 from the Women's Health Initiative (WHI) study (The Women's Health Initiative Study Group, 1998) as part of the WHI Sequencing Project (WHISP), and a total of 847 including Atherosclerosis Risk in Communities (ARIC; Muntaner *et al*., 1998) ($N=282$), Jackson Heart Study (JHS; Taylor *et al*., 2005) ($N=366$), Multi-Ethnic Study of Atherosclerosis (MESA; Bild *et al*., 2002) ($N=146$) and Coronary Artery Risk Development in Young Adults (CARDIA; Friedman *et al*., 1988) ($N=53$) as part of HeartGO. Most WHISP and HeartGO participants were selected on the basis of primary phenotypes for ESP, which included extremes of body mass index, blood pressure, low-density lipoprotein (LDL), cholesterol, early onset myocardial infarction (MI) cases and controls, ischemic stroke with either early onset or positive family history. Approximately 15% of samples were selected because of having non-missing data for a selected set of core phenotypes, but were not ascertained based on trait values.

2.2 Exome Sequencing

Exome sequencing was performed at the University of Washington (SeattleGO) and the Broad Institute (BroadGO). Initial quality control (QC) on all samples involved sample quantification (PicoGreen), confirmation of high-molecular weight DNA, fingerprint genotyping and sex determination. Samples were failed if total mass, concentration, integrity of DNA or quality of preliminary genotyping data was too low or sex typing was discordant. Following QC, 2 μg of extracted genomic DNA was subjected to shotgun library preparation and exome capture as previously described (Tennessen *et al*., 2012).

2.2.1 Genotype Calling For read mapping and variant analysis, samples were aligned to a human reference (hg19) using Burrows–Wheeler Aligner (Li and Durbin, 2009). Variant detection and genotyping were performed on both exomes and flanking 50 bp of intronic sequence. Typical mean coverage of the target was 60–80×. Variant data for each sample were formatted (variant call format) as 'raw' calls for all samples. Filters considered the total read depth, the number of individuals with coverage at the site, the fraction of variant reads in each heterozygote, the ratio of forward and reverse strand reads carrying reference and variant alleles and the average position of variant alleles along a read. Variant calling was performed across all 6515 samples at the University of Michigan (UMich). Only single nucleotide polymorphisms (SNPs) that passed the UMich support vector machine quality filter were retained for analysis. Details were previously described (Fu *et al*., 2012).

2.2.2 Reference Panel Construction A reference panel of 2163 individuals (including the 1692 African Americans used in this study and 471 European Americans) was constructed. All of the 2163 individuals have both Genome-wide association study (GWAS; Affymetrix 6.0) genotypes and whole exome sequencing data. When combining the two sources of data, a total of 375 024 bi-allelic autosomal SNPs with minor allele count ≥4 (in the 2163 reference panel subjects) did not overlap with the 702 205 GWAS SNPs. There were 10 130 SNPs that overlapped between ESP and the 702 205 GWAS markers. SNPs with concordance <95% were removed (65 SNPs). For overlapping SNPs that passed this concordance filter, GWAS genotype was retained for consistency with the target individuals. A total of 1 077 164 autosomal SNPs were included in the reference panel. These 1 077 164 markers were phased across all 2163 samples using BEAGLE v3.3.1 (Browning and Yu, 2009).

2.2.3 ESP 'Extreme' and 'Normal' Panel Construction The 1692 ESP African Americans were selected based on the following phenotypic traits: (i) LDL ($N=254$: 131 with high LDL and 123 with low LDL), (ii) blood pressure ($N=247$: 132 with high blood pressure and 115 with low blood pressure), (iii) body mass index (BMI, $N=609$: 429 with high BMI and 180 with normal to low BMI), (iv) early onset MI (EOMI, $N=324$: 39 EOMI cases and 285 EOMI controls), (v) stroke ($N=40$, all cases) and (vi) random samples ($N=218$). We constructed one ESP 'Extreme' panel and one ESP 'Normal' panel each with 853 individuals. The ESP 'Extreme' panel included (i) 254 individuals with high/low LDL (131 with high LDL and 123 with low LDL), (ii) 247 individuals with high/low blood pressure (132 with high blood pressure and 115 with low blood pressure), (iii) 40 stroke cases, (iv) 39 EOMI cases and (v) 273 individuals with high BMI. The ESP 'Normal' panel consists of 80% individuals with 'non-extreme' phenotypes and 20% with extreme phenotypes so as to better represent a population sample. Individuals with 'non-extreme' phenotypes ($N=683$) are from random sample, EOMI controls and low BMI group. Individuals with extreme phenotypes ($N=170$) are from high ($N=85$) and low LDL ($N=85$) group.

2.2.4 The 1000 Genomes Project (1000G) The 1000 Genomes Phase1 data were downloaded from http://www.sph.umich.edu/csg/yli/mach/download/1000G.2012-03-14.html. Details regarding the generation of the data can be found in the Phase 1 article (The 1000 Genomes Project Consortium, 2012).

2.3 Target African Americans

2.3.1 GWAS Data All of the 1661 target African Americans in this study were genotyped using the Affymetrix 6.0 genotyping platform as part of the WHI SNP Health Association Resource study. Before phasing and imputation, we removed Affymetrix 6.0 SNPs with genotype call rates <90%, or Hardy–Weinberg exact test (Wigginton *et al*., 2005) $P<10^{-6}$ or MAF <1%. QC details were described previously (Auer *et al*., 2012; Reiner *et al*., 2011).

2.3.2 Metabochip data All of the 1661 target African Americans in this study were also genotyped using the Metabochip (Voight *et al*., 2012) in an attempt to generalize genetic effects across racial groups by the

WHI Population Architecture using Genomics and Epidemiology (PAGE) study. Standard QC was performed, including removal of markers with genotype call rate <95% or Hardy–Weinberg $P < 10^{-6}$, as well as exclusion of individuals who 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 article (Buyske *et al*., 2012).

Genotypes at the Metabochip SNPs were not used for imputation but rather used for assessment of imputation quality. In total 5035 markers, which were on Metabochip, in 1000G and in ESP, but not on Affymetrix 6.0, were used for imputation quality assessment.

2.3.3 Overlap with ESP African Americans African Americans present in ESP were not included as target. In other words, individuals in the reference ESP and the target were mutually exclusive. In addition, we removed any target with PLINK (Purcell *et al*., 2007) estimated identity-by-descent (IBD) ≥ 0.2 with any reference individual such that our final target set did not contain any apparent first-degree relative with the reference ESP.

2.3.4 Imputation using IMPUTE2 In the main text, unless otherwise specified, we present results using minimac for imputation. Supplementary Figure S5 and Supplementary Table S7 showed that our recommendation of ESP alone or concatenation of ESP with 1000G (ESP_U_1000G) over 1000G still held when IMPUTE2 was used for imputation. We note that in the main text, our recommendation against IMPUTE2's two panel mode (option 3: ESP + 1000G) was confounded by software/method choice: ESP alone or ESP_U_1000G using minimac performed better than IMPUTE2's ESP + 1000G, but when using IMPUTE2 for all, ESP alone or ESP_U_1000G performed similarly as ESP + 1000G.

3 RESULTS

3.1 Comparison of imputation quality between ESP-based and 1000G-based imputation

We first performed imputation, using either ESP or 1000G as reference, into 1661 African Americans in the WHI study (the 'target' sample) who were genotyped by both the Affymetrix 6.0 (Auer *et al*., 2012) and the Illumina Metabochip array (Buyske *et al*., 2012; Liu *et al*., 2012). We used MaCH (Li *et al*., 2010), a hidden Markov model that leverages linkage disequilibrium information among samples of unrelated individuals, to pre-phase the 1661 WHI African Americans at the Affymetrix 6.0 markers. The ESP reference panel was built from 1692 African Americans with genotypes from both the Affymetrix 6.0 platform and whole exome sequencing. These genotypes were merged and phased using BEAGLE (Browning and Yu, 2009). Imputation into the 1661 target WHI African Americans was performed with minimac (Howie *et al*., 2012) (similar results were obtained with IMPUTE2; see Methods) using their Affymetrix 6.0 genotypes only; genotypes from the Metabochip genotyping were saved for evaluation. Following the literature (Browning and Yu, 2009; Li *et al*., 2010), we used dosage r^2 [squared Pearson correlation between imputed dosages (ranging continuously from 0 to 2) and experimental genotypes (coded as 0, 1 or 2)], which directly determines effective sample size for subsequent association analysis (Pritchard and Przeworski, 2001), to gauge imputation quality. We also use Rsq, the estimated dosage r^2 generated by minimac, as the post-imputation QC metric. We observed 8.3–11.4% increases in average dosage r^2 for variants with MAF < 1% using the ESP reference panel compared with the 1000G reference panel (paired Wilcoxon $P < 1.3 \times 10^{-4}$ – 4.1×10^{-16}). Such increases were observed without applying any post-imputation QC, that is, when every imputed variant was retained. Similarly increased dosage r^2 was observed across a broad range of post-imputation QC stringency (removing 0–90% of variants; Fig. 1 and Supplementary Fig. S1 and Supplementary Table S1). As imputation is routinely performed in 10 000–100 000 individuals (Auer *et al*., 2012; Cho *et al*., 2012; Dastani *et al*., 2012; Holm *et al*., 2011; Teslovich *et al*., 2010), such an increase would correspond to increasing the sample size for association testing by 1000–10 000 samples.

Because the ESP panel is larger and consists entirely of African Americans, we conducted more comparisons by assessing the performance of 10 random subsets from ESP of the same size as 1000G (both for the full 1000G panel [Number of haplotypes (H) = 1092×2; reference panels termed ESP.1092 and 1000G.1092] and the most relevant panel [AFR + EUR, H = 625×2; reference panels termed ESP.625 and 1000G.625]). The difference in effective sample size derived from the ESP and 1000G reference panels, although smaller, remains (Fig. 2 and Supplementary Table S2). For example, when comparing ESP.1092 with 1000G.1092 and retaining all imputed variants in the analysis (no post-imputation QC), we observed an average dosage r^2 increase of 11.3, 4.6 and 6.1% for variants with MAF <0.2%, 0.2–0.5% and 0.5–1%, respectively. The corresponding dosage r^2 increases for a comparison of ESP.625 with 1000G.625 were 13.9%, 1.0 and 3.1%, respectively. The superior performance of ESP over 1000G was likely driven by two primary factors. First, genotypes for rare variants from ESP were derived from high coverage sequencing, whereas those from 1000G were in part from low coverage sequencing (1000G data we used here are the integrated panel constructed from low coverage whole genome sequencing, deep exome sequencing and SNP array genotyping). Second, ESP African Americans (~50% also from WHI, detailed in Materials and Methods) were better matched to the 'target' WHI African Americans for ancestry than were the samples in the 1000G panel, which were pooled from several populations of European, African and African American ancestry.

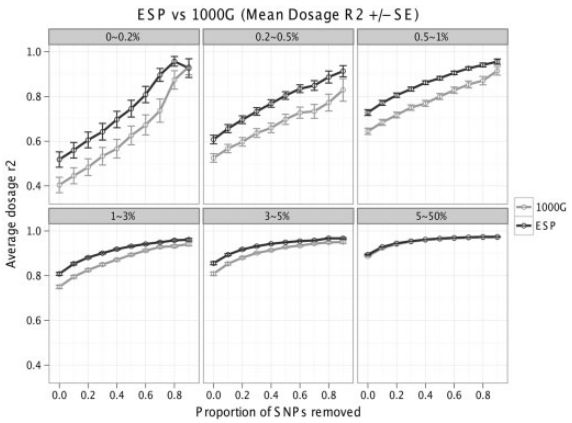

Fig. 1. Comparison of dosage r^2 between ESP-based and 1000G-based imputation. The *x*-axis is the proportion of SNPs that were removed based on elevated Rsq threshold (QC). The *y*-axis is the mean dosage r^2 (squared Pearson correlation between imputed dosages and experimental genotypes)

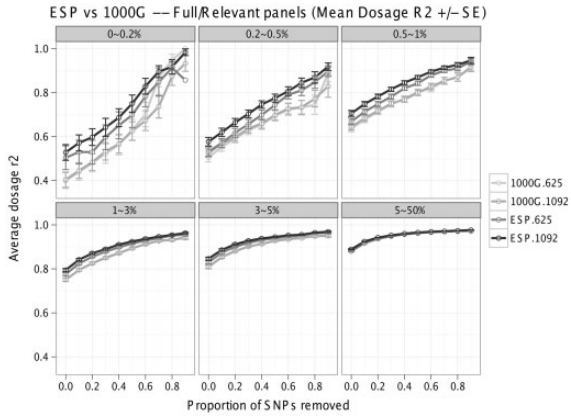

Fig. 2. Comparison of dosage r^2 between ESP and 1000G full/relevant panel imputation. The *x*-axis is the proportion of SNPs that were removed based on elevated Rsq threshold (QC). The *y*-axis is the mean dosage r^2 (squared Pearson correlation between imputed dosages and experimental genotypes)

As expected, better quality imputation using the ESP panel produces a larger number of well-imputed rare coding variants than using the 1000G panel (Rsq >0.6 for MAF <0.5%; detailed in Table 1). For example, the number of well-imputed variants was 2.28, 2.83, and 1.54 times greater than that from 1000G for MAF <0.2, 0.2–0.5 and 0.5–1%, respectively (Table 1). The boost in imputation quality as well as in the number of well-imputed markers is expected to enhance power for testing association with phenotypic traits. For example, out of the eight novel blood trait associated variants reported in Auer *et al.* (Auer *et al.*, 2012), two are not in 1000G but ESP only (Supplementary Table S3).

3.2 Impact of imputation reference panel constructed from subjects selected based on extreme phenotypes

Many subjects sequenced in ESP were selected on the basis of phenotypic extremes or disease status (detailed in Materials and Methods), an approach that has been shown to increase power for association testing of the specific phenotype (Barnett *et al.*, 2013; Guey *et al.*, 2011; Kryukov *et al.*, 2009). To our knowledge, the consequences of such a design for developing an imputation reference panel have not been previously evaluated. To this end, we constructed two ESP-derived reference panels: 'ESP.extreme' and 'ESP.normal' each of size $H = 853 \times 2$. The former included 254 African Americans from LDL cholesterol extremes, 247 from blood pressure extremes, 40 stroke cases, 39 early onset MI (EOMI) cases and 273 with extremely high BMI. The latter included 85 samples with high LDL, 85 with low LDL and 683 from the 'middle' of the phenotype distributions. We observed no loss of imputation quality using the 'Extreme' panel. (Fig. 3 and Supplementary Table S4).

3.3 Alternative options to use or combine ESP and 1000G reference panels

Although our results suggested that the ESP panel led to substantially improved imputation accuracy of rare coding variants

Table 1. Number and percentage of well-imputed exonic variants

MAF	Number (%) of well-imputed[a] markers		ESP:1000G ratio (Number of well-imputed)
	ESP	1000G	
0–0.2%	17 606 (31.8)	7713 (3.0)	2.28
0.2–0.5%	26 255 (70.0)	9283 (26.9)	2.83
0.5–1%	21 377 (92.1)	13 882 (62.9)	1.54
1–3%	29 784 (96.7)	26 466 (90.7)	1.13
3–5%	11 490 (96.9)	11 043 (96.0)	1.04
5–50%	40 500 (98.0)	39 849 (96.3)	1.02

[a]Well-imputed is defined such that the average Rsq of the QC + markers within each MAF category is >0.8.

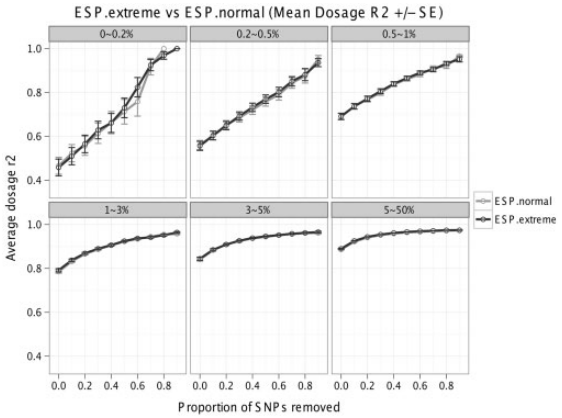

Fig. 3. Comparison of dosage r^2 between ESP.extreme and ESP.normal imputation. The *x*-axis is the proportion of SNPs that were removed based on elevated Rsq threshold (QC). The *y*-axis is the mean dosage r^2 (squared Pearson correlation between imputed dosages and experimental genotypes)

compared with the 1000G panel, the combination of the two panels could potentially result in even better performance than either one individually. We considered the following four options. The default option, Option 0, was to select a single panel *a priori* based on reference panel size, marker density and ancestry match. In this case, Option 0 would be the ESP reference panel alone, as it contains more haplotypes (3384 over 2184 in 1000G), greater marker density in exons and a better ancestry match with the target African Americans. Option 1 was to first impute using each panel separately, and then for each marker to select the one with higher Rsq. Option 2 was to impute using a concatenated panel of the two (ESP_U_1000G). Option 3 was to impute using IMPUTE2, which allows two separate reference panels (ESP + 1000G).

The best option among the four was the concatenation of the two panels (Option 2) with ESP alone (Option 0), a close second best. For example, the average dosage r^2 increased by 1.8%, 2.3% and 1.5%, respectively, for markers with MAF <0.2, 0.2–0.5 and 0.5–1% using Option 2 over Option 0 (Supplementary Fig. S2 and Supplementary Table S5). We observed no noticeable gains using Option 1 compared with

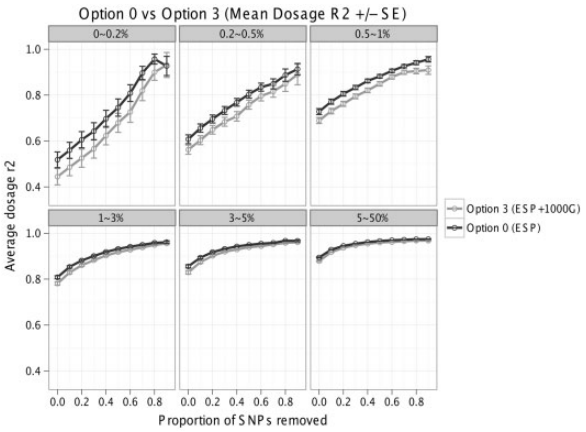

Fig. 4. Comparison of dosage r^2 between Option 0 (ESP) and Option 3 (ESP + 1000G) imputation. The *x*-axis is the proportion of SNPs that were removed based on elevated Rsq threshold (QC). The *y*-axis is the mean dosage r^2 (squared Pearson correlation between imputed dosages and experimental genotypes)

Option 0 with differences in dosage r^2 in the range of 0.02–1.5% (Supplementary Fig. S3 and Supplementary Table S6). Therefore, we would not recommend using Option 1, the Rsq-based selection, because higher Rsq does not guarantee better imputation quality. In fact a low quality reference panel could lead to poorly estimated Rsq values. Finally, IMPUTE2's ability to combine two reference panels (Option 3), led to decreased imputation quality compared with Option 0. For example, dosage r^2 decreased by an average of 7.3, 4.3 and 3.9% for markers with MAF 0.2, 0.2–0.5 and 0.5–1% (Fig. 4 and Supplementary Fig. S4 and Supplementary Table S7). Although less accurate, the convenience provided by IMPUTE2's approach warrants closer consideration. Decreases in quality could be due to software implementation because we used minimac for options 0–2 and IMPUTE2 for option 3. But importantly, our recommendation of concatenation of the two or ESP alone over 1000G alone or post-imputation Rsq-based selection holds when IMPUTE2 was used for all four options (see 'Imputation using IMPUTE2' in Materials and Methods, Supplementary Fig. S5 and Supplementary Table S8).

4 DISCUSSION

We note that ESP is heavily enriched for extremes from several phenotypes rather than a single phenotype. Thus, it is unclear whether these results generalize to a design where sequenced subjects are selected based on extremes for a single phenotype. We did not attempt to select one phenotype for evaluation, as doing so would reduce our reference size to below 300, which we view as of little value for the imputation of rare variants. We expect such 'Extreme' panels to make little difference for imputation overall and may affect imputation in the specific trait associated regions when the causal variant(s) exert large effect(s).

Although we recommend the concatenation of ESP and 1000G, we observed only modest gains in imputation quality by combining the two. Previous studies suggest that these gains may depend in part on the ethnic make-up of the study subjects (Browning and Yu, 2009) and whether 1000G data add substantial haplotype diversity. These gains should be weighed against the logistical challenges of combining data from multiple sources to avoid batch effects (e.g. mismatched strands, inconsistent marker naming schemes or systematic differences in genotype calling, QC or phasing).

In summary, we found that the ESP African American reference panel outperformed the 1000G reference panel for the imputation of rare coding variants in African Americans, both in terms of imputation quality, the number of imputable markers and consequently power for trait association testing. The finding was robust to adjustment of reference size and matching on ethnicity. We did not observe loss of imputation quality because of the ESP design for enriched sequencing of subjects selected for phenotypic extremes. Regarding the optimal way to combine the two panels, our evaluations suggested that ESP alone or concatenation of the ESP and 1000G reference panels was superior to either post-imputation selection based on Rsq or IMPUTE2's implementation of two separate reference panels. We focused here on imputation of coding variants from ESP. However, we believe that the conclusions drawn here apply to rare variants across the genome as recently reported by several whole-genome sequencing-based studies (Fuchsberger *et al.*, 2012; Sanna, 2012) in individuals of European ancestry. These studies and our present work strongly suggest that population matched samples, even in diverse populations such as African Americans, can clearly outperform 1000G imputation performance. Therefore, we recommend investigators routinely consider sequencing for the design (Kang *et al.*, 2013) and analysis of the study samples.

ACKNOWLEDGEMENTS

The authors thank Dr Michael L. Boehnke for helpful discussions. They also thank the ESP, 1000 Genomes Project, WHI SNP Health Association Resource and PAGE consortia for generating the data. The authors wish to acknowledge the support of the National Heart, Lung and Blood Institute (NHLBI) and the contributions of the research institutions, study investigators, field staff and study participants in creating this resource for biomedical research. Funding for GO ESP was provided by NHLBI grants RC2 HL-103010 (HeartGO), RC2 HL-102923 (LungGO) and RC2 HL-102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL-102925 (BroadGO) and RC2 HL-102926 (SeattleGO).

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 article 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: https://cleo.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Short%20List.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 QC 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.

Funding: National Institutes of Health (R01HG006292 and R01HG006703 to Y.L.).

Conflict of Interest: none declared.

REFERENCES

Auer,P.L. *et al.* (2012) Imputation of exome sequence variants into population-based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. *Am. J. Hum. Genet.*, **91**, 794–808.

Barnett,I.J. *et al.* (2013) Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genet. Epidemiol.*, **37**, 142–151.

Bild,D.E. *et al.* (2002) Multi-ethnic study of atherosclerosis: objectives and design. *Am. J. Epidemiol.*, **156**, 871–881.

Browning,B.L. and Yu,Z. (2009) Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am. J. Hum. Genet.*, **85**, 847–861.

Buyske,S. *et al.* (2012) Evaluation of the metabochip genotyping array in African Americans and implications for fine mapping of GWAS-identified loci: the PAGE study. *PLoS One*, **7**, e35651.

Cho,Y.S. *et al.* (2012) Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat. Genet.*, **44**, 67–72.

Dastani,Z. *et al.* (2012) Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet.*, **8**, e1002607.

De Bakker,P.I.W. *et al.* (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.*, **17**, R122–R128.

Fridley,B.L. *et al.* (2010) Utilizing genotype imputation for the augmentation of sequence data. *PLoS One*, **5**, e11018.

Friedman,G.D. *et al.* (1988) Cardia: study design, recruitment, and some characteristics of the examined subjects. *J. Clin. Epidemiol.*, **41**, 1105–1116.

Fu,W. *et al.* (2012) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, **493**, 216–220.

Fuchsberger,C. *et al.* (2012) The value of population-specific reference panels for genotype imputation in the age of whole-genome sequencing. In: *Presented at the 62nd Annual Meeting of The American Society of Human Genetics*. San Francisco, CA.

Futema,M. *et al.* (2012) Use of targeted exome sequencing as a diagnostic tool for Familial Hypercholesterolaemia. *J. Med. Genet.*, **49**, 644–649.

Guey,L.T. *et al.* (2011) Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. *Genet. Epidemiol.*, **35**, 236–246.

Holm,H. *et al.* (2011) A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat. Genet.*, **43**, 316–320.

Howie,B. *et al.* (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.*, **44**, 955–959.

Huang,J. *et al.* (2012) 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *Eur. J. Hum. Genet.*, **20**, 801–805.

Kang,J. *et al.* (2013) AbCD: arbitrary coverage design for sequencing-based genetic studies. *Bioinformatics*, **29**, 799–801.

Kryukov,G.V. *et al.* (2009) Power of deep, all-exon resequencing for discovery of human trait genes. *Proc. Natl Acad. Sci. USA*, **106**, 3871–3876.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,Y. *et al.* (2009) Genotype imputation. *Annu. Rev. Genomics Hum. Genet.*, **10**, 387–406.

Li,Y. *et al.* (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816–834.

Li,Y. *et al.* (2011) Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.*, **21**, 940–951.

Liu,E.Y. *et al.* (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. *Genet. Epidemiol.*, **36**, 107–117.

Liu,E.Y. *et al.* (2013) MaCH-Admix: genotype imputation for admixed populations. *Genet. Epidemiol.*, **37**, 25–37.

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

Muntaner,C. *et al.* (1998) Work organization and atherosclerosis: findings from the ARIC study. *Am. J. Prev. Med.*, **14**, 9–18.

Price,A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

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

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–75.

Reiner,A.P. *et al.* (2011) Genome-wide association study of white blood cell count in 16,388 African Americans: the continental origins and genetic epidemiology network (COGENT). *PLoS Genet.*, **7**, e1002108.

Sanna,S. (2012) Using low-pass whole genome sequencing to create a reference population for genome imputation in an isolated population: examples from the SardiNIA study. In: *Presented at the 62nd Annual Meeting of The American Society of Human Genetics*. San Francisco, CA.

Taylor,H.A. *et al.* (2005) Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. *Ethn. Dis.*, **15**, S6–4–17.

Tennessen,J.A. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.

Teslovich,T.M. *et al.* (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707–713.

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

The 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

The Women's Health Initiative Study Group. (1998) Design of the Women's Health Initiative clinical trial and observational study. *Control. Clin. Trials*, **19**, 61–109.

Voight,B.F. *et al.* (2012) The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.*, **8**, e1002793.

Wigginton,J.E. *et al.* (2005) A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.*, **76**, 887–893.