Human Molecular Genetics, 2014, Vol. 23, No. 24 **6607–6615**
doi:10.1093/hmg/ddu361
Advance Access published on July 15, 2014

Whole-exome imputation of sequence variants identified two novel alleles associated with adult body height in African Americans

Mengmeng Du[1,2,4,†,*], Paul L. Auer[1,5,†], Shuo Jiao[1], Jeffrey Haessler[1], David Altshuler[6], Eric Boerwinkle[7], Christopher S. Carlson[1], Cara L. Carty[1], Yii-Der Ida Chen[8], Keith Curtis[1], Nora Franceschini[9], Li Hsu[1], Rebecca Jackson[10], Leslie A. Lange[11], Guillaume Lettre[12], Keri L. Monda[9], National Heart, Lung, and Blood Institute (NHLBI) GO Exome Sequencing Project[‡], Deborah A. Nickerson[3], Alex P. Reiner[1], Stephen S. Rich[13], Stephanie A. Rosse[1], Jerome I. Rotter[8], Cristen J. Willer[14], James G. Wilson[15], Kari North[9], Charles Kooperberg[1], Nancy Heard-Costa[16] and Ulrike Peters[1,*]

[1]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [2]School of Public Health and [3]School of Medicine, University of Washington, Seattle, WA, USA, [4]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA, [5]University of Wisconsin-Milwaukee Joseph J. Zilber School of Public Health, Biostatistics, Milwaukee, WI, USA, [6]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA, [7]Human Genetics Center, The University of Texas Health Science Center at Houston, Houston, TX, USA, [8]Los Angeles Biomedical Research Institute, LABioMed at Harbor-UCLA Medical Center, Torrance, CA, USA, [9]Department of Epidemiology, University of North Carolina Gillings School of Global Public Health, Chapel Hill, NC, USA, [10]Division of Endocrinology, Diabetes and Metabolism, The Ohio State University Wexner Medical Center, Columbus, OH, USA, [11]Department of Genetics, University of North Carolina School of Medicine, Chapel Hill, NC, USA, [12]Medicine, Montreal Heart Institute and Université de Montréal, Montreal, QC, Canada, [13]Center for Public Health Genomics, University of Virginia School of Medicine, Charlottesville, VA, USA, [14]Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI, USA, [15]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA and [16]Department of Neurology, Boston University School of Medicine, Boston, MA, USA

Received April 24, 2014; Revised June 30, 2014; Accepted July 8, 2014

Adult body height is a quantitative trait for which genome-wide association studies (GWAS) have identified numerous loci, primarily in European populations. These loci, comprising common variants, explain <10% of the phenotypic variance in height. We searched for novel associations between height and common (minor allele frequency, MAF ≥5%) or infrequent (0.5% < MAF < 5%) variants across the exome in African Americans. Using a reference panel of 1692 African Americans and 471 Europeans from the National Heart, Lung, and Blood Institute's (NHLBI) Exome Sequencing Project (ESP), we imputed whole-exome sequence data into 13 719 African Americans with existing array-based GWAS data (discovery). Variants achieving a height-association threshold of $P < 5E-06$ in the imputed dataset were followed up in an independent sample of 1989 African Americans with whole-exome sequence data (replication). We used $P < 2.5E-07$ (=0.05/196 779 variants) to define statistically significant associations in meta-analyses combining the discovery and replication sets ($N = 15\,708$). We discovered and replicated three independent loci for association: 5p13.3/*C5orf22*/

*To whom correspondence should be addressed at: Fred Hutchinson Cancer Research Center, Cancer Prevention Program, 1100 Fairview Avenue N, M4-B402, Seattle, WA 98109-1024, USA. Tel: +1 2066672309; Fax: +1 2066677850; Email: mdu@fhcrc.org
†These authors contributed equally.
‡Exome Sequencing Project authorship banner is available in the Supplementary Methods.

rs17410035 (MAF = 0.10, β = 0.64 cm, P = 8.3E−08), 13q14.2/*SPRYD7*/rs114089985 (MAF = 0.03, β = 1.46 cm, P = 4.8E−10) and 17q23.3/*GH2*/rs2006123 (MAF = 0.30; β = 0.47 cm; P = 4.7E−09). Conditional analyses suggested 5p13.3 (C5orf22/rs17410035) and 13q14.2 (SPRYD7/rs114089985) may harbor novel height alleles independent of previous GWAS-identified variants (r^2 with GWAS loci <0.01); whereas 17q23.3/*GH2*/rs2006123 was correlated with GWAS-identified variants in European and African populations. Notably, 13q14.2/rs114089985 is infrequent in African Americans (MAF = 3%), extremely rare in European Americans (MAF = 0.03%), and monomorphic in Asian populations, suggesting it may be an African-American-specific height allele. Our findings demonstrate that whole-exome imputation of sequence variants can identify low-frequency variants and discover novel variants in non-European populations.

INTRODUCTION

Adult body height is a classic quantitative trait that involves numerous genetic loci (1). Because it is greatly determined by genetics (narrow-sense heritability $h^2 \sim 0.75–0.9$) (2–4) and is generally stable and well measured, height serves as a model for gaining insight into the genetic architecture of complex traits. Genome-wide association studies (GWAS) have identified hundreds of height-associated variants (1,5–23). These loci combined, however, explain only ∼10% of the phenotypic variation—suggesting many height-related variants remain unidentified (1). As GWAS are designed to capture common (>5%) genetic variation, identifying less frequent or rare variants that confer larger genetic effects may help account in part for the unexplained genetic component of height (24,25).

Whole-exome sequencing technology facilitates the comprehensive examination of genetic variation in or near protein-coding regions of the genome (26,27). These regions may harbor low-frequency variants with biological consequences that carry larger effects. For instance, exome sequencing has been used to identify rare variants that cause highly penetrant Mendelian disorders (28–30). For complex quantitative traits, however, detecting associations with low-frequency variation has been challenging as large populations must be sequenced, which is prohibitively expensive. This limitation can be addressed in part by using exome sequencing followed by imputation of sequence variants into those with existing genome-wide single nucleotide polymorphism (SNP) array data (31)—which increases the effective number of individuals with exome data and in turn increases statistical power. Recent studies have successfully applied this approach to identify novel low-frequency variants that contribute to phenotypic differences in complex traits, such as those related to blood cells (31) and von Willebrand factor (32).

The majority of height loci have been identified in European populations (1); however, GWAS conducted in Asian populations, including Chinese (7,14,16), Korean (9,14), Japanese (11) and Filipino (8), have reported height loci previously unidentified in European descent individuals (9,11). Similarly, two large GWAS in African Americans (5,6) replicated several height loci identified in European populations, but also reported three novel loci (2p14/*ANTXR1,* 17q23/*TMEM100/PCTP* and Xp22.3/*ARSE*). These data suggest some variants may be more common in non-European populations, which increases statistical power to identify certain height-related loci in these populations. African Americans show increased genomic diversity and overall lower linkage disequilibrium (LD) patterns compared with European and Asian populations (33). Applying exome sequencing in this population can help identify novel height-related variants and fine-map regions harboring GWAS-identified variants, which can help refine likely causal variant(s) and provide insight into the genetic basis of height and other complex traits (34–37).

In this study, we searched for novel genetic variants related to adult body height in African Americans by imputing exome sequence data from the National Heart, Lung, and Blood Institute's (NHLBI) Exome Sequencing Project (ESP) into 13 719 individuals with previously collected array-based GWAS data. In African Americans, imputation using ESP data outperformed imputation using 1000 Genomes Project data (38). We replicated variants showing the strongest associations in 1989 African Americans with whole-exome sequencing data. Using this combination of imputed and directly sequenced exome data, we examined comprehensively both common and less frequent genetic variation in or near protein-coding regions in relation to height.

RESULTS

Table 1 shows the mean age and height, by sex, for 15 708 African Americans in the study population. For participants in the Candidate Gene Association Resource (CARe) and Women's Health

Table 1 Age and height, separated by sex, for 15 708 African Americans included in the study population

	Discovery set CARe		WHI-SHARe		Replication set ESP	
	Female	Male	Female	Male	Female	Male
N	3778	2422	7519	0	1501	488
Age (years)[a]	49.9 (14.4)	50.3 (14.8)	61.7 (7.1)	–	57.5 (10.4)	53.2 (13.8)
Height (cm)[a]	163.4 (6.5)	176.7 (6.8)	162.5 (6.2)	–	162.5 (6.5)	176.1 (7.2)

[a]Values are mean (SD).

Initiative (WHI) SNP Health Association Resource (WHI-SHARe) cohorts, whole-exome association tests with height showed 35 SNPs (in 12 loci) with discovery $P < 5E-06$ (Supplementary Material, Table S1); these variants were subsequently carried forward for replication in ESP samples. We identified eight SNPs in three loci (5p13.3, 13q14.2 and 17q23.3) with replication *P*-values close to 0.05 and a statistically significant combined (discovery + replication) *P*-value ($P < 2.5E-07$) (Table 2). 13q14.2 and 17q23.3 each harbored more than one height association signal (Fig. 1). At each locus, we conditioned on the most strongly associated SNP (13q14.2/*SPRYD7*/rs114089985 and 17q23.3/*GH2*/rs2006123), and tested the remaining SNPs for association one-by-one. No additional SNPs at either locus demonstrated evidence for association in the conditional analyses, suggesting three independent signals [5p13.3 (*C5orf22*/rs17410035: MAF = 0.10, $\beta = 0.64$, $P = 8.3E-08$), 13q14.2 (*SPRYD7*/rs114089985: MAF = 0.03, $\beta = 1.46$, $P = 4.8E-10$) and 17q23.3 (*GH2*/rs2006123: MAF = 0.30; $\beta = 0.47$; $P = 4.7E-09$)] (bolded in Table 2).

In 17q23.3, rs2006123 and rs7223078 is located 523 kb apart and show weak correlation (Fig. 1C). After conditioning on rs2006123, the association with rs7223078 was slightly attenuated ($P = 3.6E-06$ to $8.5E-04$); likewise, after conditioning on rs7223078, the association with rs2006123 became weaker ($P = 1.6E-07$ to $4.2E-05$), but did not vanish. This suggests the possibility that multiple alleles contribute to height in 17q23.3.

GWAS have previously identified height-related variants in 5p13.3 (1,13,23) and 13q14.2 (1,20,23) in individuals of European ancestry, and in 17q23.3 in those of European (1,23) and African (5,6) ancestry. In 5p13.3, however, the distance between rs17410035 and the nearest GWAS SNP (rs645092) is 1.1 Mb; the association for rs17410035 remained very similar after conditioning on GWAS-identified SNPs in this locus (rs6450922, rs3792752 and rs1173727) (Supplementary Material, Table S2), suggesting that rs17410035 is independent of previously identified GWAS findings. In 13q14.2, rs114089985 has a MAF of 3% in African Americans (Exome Variant Server allele count: A = 38/G = 1346), but is extremely rare in European Americans (Exome Variant Server allele count: A = 1/G = 3181; MAF = 0.03%), and is monomorphic in Asian populations (1000 Genomes Project CHB + CHS + JPT allele count: A = 0/G = 572). As expected, in African ancestry individuals rs114089985 showed minimal correlation ($r^2 < 0.2$) with SNPs identified in European populations (rs3116602, rs3118905 and rs9596219) and conditional analyses did not alter the association (Supplementary Material, Table S2). In 17q23.3, conditioning on three GWAS-identified SNPs (rs7209435, rs2854160 and rs2941551), but not rs2665838, substantially attenuated the association with rs2006123 (Supplementary Material, Table S2). This, again, may suggest the existence of multiple effect alleles at this locus, including at least one allele tagged by rs2006123 and three of the GWAS variants. Alternatively, it is also possible that this set of modestly correlated SNPs tags a single causal variant that has not been genotyped or is poorly imputed.

It is important to note that our finding in 5p13.3/*C5orf22* should be interpreted with caution as locus-specific ancestry showed a statistically significant association with height at this locus ($P = 2.3E-03$). After additional adjustment for local

Table 2 Statistically significant (CARe + SHARe + ESP combined $P < 2.5E-07$) variants for height in African Americans

Locus	Position[a]	SNP	Gene	Function class	Effect allele	Other allele	MAF	Mean Rsq	CARe ($N = 6200$)		WHI-SHARe ($N = 7519$)		CARe + WHI-SHARe ($N = 13\,719$)			ESP ($N = 1989$)		CARe + WHI-SHARe + ESP ($N = 15\,708$)		
									β[b,c]	*P*	β[b,c]	*P*	β[b,c]	SE	*P*	β[b,c]	*P*	β[b,c]	SE	*P*
5p13.3	**31541142**	**rs17410035**	***C5orf22***	**Intron**	**T**	**G**	**0.10**	**0.97**	**0.45**	**3.0E−02**	**0.68**	**3.2E−05**	**0.59**	**0.13**	**4.1E−06**	**1.06**	**2.6E−03**	**0.64**	**0.12**	**8.3E−08**
13q14.2	**50501743**	**rs114089985**	***SPRYD7***	**Intron**	**A**	**G**	**0.03**	**0.94**	**1.43**	**6.4E−05**	**1.42**	**8.4E−06**	**1.42**	**0.24**	**2.1E−09**	**2.68**	**5.5E−02**	**1.46**	**0.23**	**4.8E−10**
	50623143	rs72631826	*DLEU2*	Intergenic	G	A	0.02	0.96	1.11	6.1E−03	1.38	1.0E−04	1.26	0.27	2.2E−06	2.22	2.1E−02	1.33	0.26	2.2E−07
17q23.3	**61958669**	**rs2006123**	***GH2***	**Intron**	**T**	**G**	**0.30**	**0.97**	**0.45**	**6.2E−04**	**0.44**	**7.5E−05**	**0.44**	**0.08**	**1.6E−07**	**0.72**	**4.9E−03**	**0.47**	**0.08**	**4.7E−09**
	61710010	rs6504171	*MAP3K3*	Intron	A	G	0.30	1.00	0.46	3.8E−04	0.41	2.2E−04	0.43	0.08	3.1E−07	0.72	2.4E−03	0.46	0.08	5.2E−09
	62481801	rs7223078	*POLG2*	Intron	G	A	0.32	0.99	0.26	4.8E−02	0.50	1.0E−05	0.39	0.08	3.6E−06	0.68	4.5E−03	0.43	0.08	1.0E−07
	62492582	rs1427463	*POLG2*	Missense	T	C	0.31	1.00	0.26	4.5E−02	0.50	8.9E−06	0.40	0.08	3.0E−06	0.61	8.8E−03	0.42	0.08	1.2E−07
	62476375	rs9908620	*POLG2*	Intron	C	A	0.32	0.98	0.27	3.8E−02	0.49	1.3E−05	0.40	0.08	3.1E−06	0.62	1.3E−02	0.42	0.08	1.9E−07

SNP, single nucleotide polymorphism; Ref, reference allele; Alt, alternate allele; MAF, minor allele frequency; Rsq, imputation quality.
Bolded SNPs represent independent height association signals.
[a]Based on NCBI build 37 data.
[b]Adjusted for age, sex, study and global ancestry (the first two principal components).
[c]Estimate calculated using the additive genetic model for each additional effect allele.

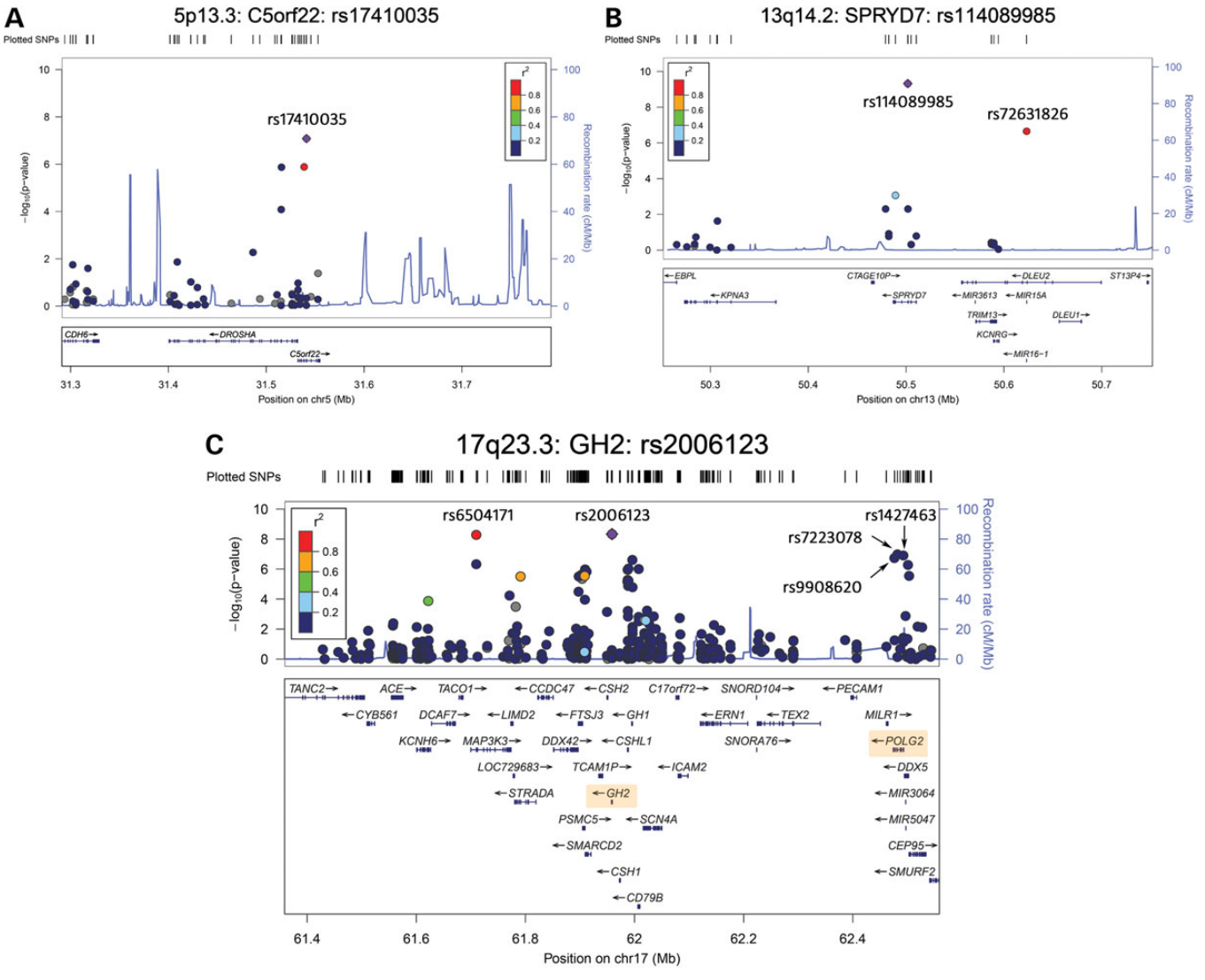

Figure 1 Regional association plots of height association signals in combined CARe + WHI-SHARe + ESP meta-analysis. Each dot reflects the −log10 *P*-value of one SNP in the region. The purple diamond indicates the index SNP showing an association signal. The color of the other dots reflects correlation (r^2) with the height-related SNP based on the 1000 Genomes Project African populations. Axes: left *Y*-axis shows −log10 of *P*-values; *X*-axis shows SNP genomic position based on NCBI build 37; and right *Y*-axis shows the estimated recombination rate from the 1000 Genomes Project African populations.

Table 3 Association results for height-related variants after adjusting for local ancestry in 13 719 African Americans in CARe and WHI-SHARe

Chr	Position[a]	SNP	Gene	Function class	Effect allele	Other allele	MAF	Mean Rsq	CARe + WHI-SHARe before adjustment[b]			CARe + WHI-SHARe after adjustment[c]		
									β[d]	SE	*P*	β[d]	SE	*P*
5p13.3	31541142	rs17410035	*C5orf22*	Intron	T	G	0.10	0.97	0.59	0.13	4.1E−06	0.53	0.14	1.1E−04
13q14.2	50501743	rs114089985	*SPRYD7*	Intron	A	G	0.03	0.94	1.42	0.24	2.1E−09	1.42	0.24	2.5E−09
17q23.3	61958669	rs2006123	*GH2*	Intron	T	G	0.30	0.97	0.44	0.08	1.6E−07	0.28	0.09	2.9E−03

SNP, single nucleotide polymorphism; Ref, reference allele; Alt, alternate allele; MAF, minor allele frequency; Rsq, imputation quality.
[a]Based on NCBI build 37 data.
[b]Adjusted for age, sex, study and global ancestry (the first two principal components).
[c]Adjusted for age, sex, study, global ancestry (the first two principal components) and local ancestry.
[d]Estimate calculated using the additive genetic model for each additional effect allele.

ancestry, the association with rs17410035 was slightly attenuated (before: $\beta = 0.59$, $P = 4.1E-06$; after: $\beta = 0.53$, $P = 1.1E-04$) (Table 3). In contrast, in 13q14.2/*SPRYD7*, local ancestry was not associated with height ($P = 7.0E-01$), and additional adjustment did not alter the association with rs114089985 (before: $\beta = 1.42$, $P = 2.1E-09$; after: $\beta = 1.42$,

$P = 2.5E-09$). In 17q23.3/*GH2*, local ancestry was associated with height ($P = 1.6E-04$). After adjustment for local ancestry, the association was attenuated for rs2006123 (before: $\beta = 0.44$, $P = 1.6E-07$; after: $\beta = 0.28$, $P = 2.9E-03$).

When investigating previous GWAS-identified height loci in our study, genotype data were available for only 10 of these height-related SNPs in our study as most of the ~200 GWAS-identified height-related SNPs are intronic or intergenic. For these 10 variants, 8 showed *P*-values ≤ 0.05 (Supplementary Material, Table S3). Further, β-estimates for 9 of the 10 SNPs showed directions consistent with previous findings.

DISCUSSION

We discovered and replicated eight SNPs in three loci (5p13.3, 13q14.2 and 17q23.3) that showed statistically significant height-associations after correction for multiple testing. These corresponded to three independent association signals: 5p13.3/*C5orf22*, 13q14.2/*SPRYD7* and 17q23.3/*GH2*. After conditioning on GWAS-identified SNPs in 5p13.3 and 13q14.2, we identified two potential novel alleles (*C5orf22*/rs17410035 and *SPRYD7*/rs114089985) associated with height in African Americans. After additional adjustment for local ancestry, however, the height association with rs17410035 was slightly attenuated, suggesting this relation should be interpreted cautiously and requires confirmation in other study populations.

C5orf22/rs17410035 is common in both African Americans (MAF = 0.10) and European populations (MAF = 0.30), and possesses a proxy SNP ($r^2 = 1$) genotyped on existing genome-wide arrays. It is therefore unclear why previous GWAS did not detect this variant. In over 180 000 individuals of European ancestry, for instance, rs17410035 showed $P = 2.3E-05$ (effect estimates were unavailable in public GIANT Consortium GWAS data) (1) (Supplementary Material, Fig. S1). One possibility is that the effect size in other populations is more modest than that observed in our study, necessitating even larger sample sets than that in the largest studies across populations (1,6,11). Another possibility is that, in African Americans, rs17410035 tags a different causal variant or is a better tag for the true causal variant. Finally, heterogeneity of effects across different study populations might also help explain the difference in findings. In contrast, *SPRYD7*/rs114089985 is infrequent in African Americans (MAF = 3%), extremely rare in European Americans (MAF = 0.03%), and monomorphic in Asian populations—suggesting this variant may be a population-specific allele undetected in studies conducted in non-African populations. Consistent with the hypothesis that low-frequency variation may confer larger effects (25), in our population, each rs114089985-A allele corresponded to a 1.46 cm increase in height (~0.2 SD). Despite this large effect size, however, rs114089985 was undetected in GWAS conducted in large African-American populations (5,6), likely because it was not identified in HapMap, not genotyped on previous genome-wide arrays, and did not possess a strong proxy (r^2 best proxy SNP = 0.43). In our study, imputation of exome sequence variants enabled the analysis of many low-frequency SNPs missed on previous genotyping arrays. These findings emphasize the importance of using these data to study low-frequency variants, which are less likely to be in LD with GWAS-identified variants (25), and to analyze non-European populations.

Although conditional analysis showed 17q23.3/*GH2*/rs2006123 was not independent of several GWAS-identified SNPs (1,5,6,23), in CARe and WHI-SHARe combined, rs2006123 ($P = 1.6E-07$) showed a smaller *P*-value compared with the GWAS variants (rs7209435, $P = 3.1E-07$; rs2665838, $P = 7.7E-01$; rs2854160, $P = 4.7E-05$; rs2941551, $P = 5.2E-03$) (Supplementary Material, Fig. S2), suggesting this SNP and rs7209435 may better tag the underlying functional variant. Whereas annotation by available bioinformatics sources (39–43) did not identify putative functions for these GWAS SNPs, including rs7209435, rs2006123 perfectly tags ($r^2 = 1$) several variants located within a secondary promoter for *GH2* (growth hormone 2, ~400 bp upstream of the transcription start site). Specifically, rs2955250, rs60832412 and rs5821405 are located within a region exhibiting open chromatin that showed histone modifications consistent with promoter and enhancer activity in several cell lines. In addition, these variants are located within a binding site for many transcription factors (e.g. MAFK, RUNX3, EGR1, RAD21 and CTCF). It is important to note, however, that bioinformatics-based annotation is intended to help identify biologically plausible functional candidates, and laboratory studies are needed to yield definitive evidence of the mechanisms driving SNP associations with height.

Adjustment for locus-specific ancestry resulted in a slight attenuation of the association in 5p13.3/*C5orf22* and did not alter the association in 13q14.2/*SPRYD7*; however, adjustment for local ancestry greatly attenuated the association in 17q23.3/*GH2*. These results are consistent with admixture mapping analyses in African Americans (5), which suggested regions on chromosome 17 may harbor variants that affect height and also show large allele frequency differences between European and African ancestral populations.

This study is the first to use exome imputation of sequence variants to examine frequent as well as less frequent variation in relation to height in African Americans. As we aimed to increase statistical power to comprehensively investigate exonic genetic variation, we imputed exome sequence variants into those with existing genome-wide SNP array data. Imputed genotypes can be called with varying accuracy, and we accounted for this using the genotype dosage, which yields unbiased effect size estimates (44). However, lower imputation accuracy may attenuate the estimated significance of association signals (45,46). In our study we observed good overall imputation accuracy as the information retained [estimated average dosage r^2, calculated using the squared Pearson correlation between imputed and experimental genotypes from the Metabochip (47)] was ~80%, even for low-frequency variants (see Supplementary Material, Methods for additional details). In a recent publication, we found that an imputation reference panel constructed from the ESP sequence data alone outperformed one constructed from the 1000 Genomes Project for imputation of rare-coding variants into African-American populations. The gain in imputation accuracy relative to 1000 Genomes was found for variants with MAF <1% (38).

The largest GWAS to date has reported hundreds of variants that showed weak or modest effects on height (1). Given this genetic architecture, limited statistical power may have accounted for the absence of rare height-related signals. For common genetic variants (allele frequency = 20%), the present analysis had 80% power to detect a per-allele change in height of 0.65 cm; for

rare variants (allele frequency = 1%), there was 80% power to detect a per-allele change in height of 2.6 cm (~0.4 SD) (Supplementary Material, Fig. S3). These estimates suggest that our data provided sufficient statistical power to detect less common SNPs with large effect sizes in regions with reasonably high imputation quality (Rsq = 0.80). However, it is notable that we found strong evidence for only one novel height-related allele (13q14.2/*SPRYD7*/rs114089985) and moderate evidence for another (5p13.3/*C5orf22*/rs17410035). These data emphasize that much larger populations are likely needed to detect remaining height associations with less frequent, but imputable, coding variation—particularly for variants with more modest associations. Finally, our study primarily tested for associations with genetic variation in protein-coding regions of the genome and provided minimal coverage of variants in non-coding regions. It remains important to study these, however, as ~88% of GWAS-identified variants for various diseases and traits have been located in non-coding regions (48), and these variants likely serve important regulatory purposes in the cell (40,41). This is supported by data in the present analysis showing that, even when using imputed exome sequence data, seven of the eight statistically significant variants were either intronic or intergenic.

In this African-American population, we used a combination of genome-wide SNP array and imputed whole-exome sequence data to comprehensively examine both common and less frequent variation in or near protein-coding regions of the genome. We identified two potentially novel height-related alleles in 5p13.3 (*C5orf22*/rs17410035) and 13q14.2 (*SPRYD7*/rs114089985), one of which (rs114089985) was uncommon and African American-specific. These findings warrant further study and show that whole-exome imputation of sequence variants is an effective strategy to investigate low-frequency variation as well as identify novel genetic associations in non-European populations.

MATERIALS AND METHODS

Study population

The discovery study population included 13 719 self-identified African Americans with GWAS data from five population-based cohort studies. This comprised 6200 participants from the NHLBI CARe consortium (Atherosclerosis Risk in Communities Study, ARIC; Coronary Artery Risk Development in Young Adults, CARDIA; Multi-Ethnic Study of Atherosclerosis, MESA and Jackson Heart Study, JHS), and 7519 participants from the WHI-SHARe project. The replication set included an independent population of 1989 individuals of African ancestry who underwent whole-exome sequencing as part of the NHLBI ESP as described previously (included studies can be found in Supplementary Material, Methods and by going to https://esp.gs.washington.edu/drupal/) (49–51). Study-specific descriptions and details are provided in Supplementary Material, Methods. Height in centimeters was collected by in-person examination, and clinical information was collected by self-report and in-person examination. We excluded participants missing height data ($N = 5$ in CARe, 44 in WHI-SHARe, 898 in ESP). All participants provided written informed consent as approved by local human-subjects committees.

Genome-wide genotyping and quality control

Genome-wide genotyping and quality control (QC) procedures have been described (31) and are available in the Supplementary Material, Methods. Briefly, DNA was genotyped using Affymetrix 6.0 arrays (Affymetrix, Santa Clara, CA, USA). CARe conducted a candidate gene survey using the ITMAT/Broad/CARe (IBC) candidate gene array in all participants, and a GWAS (Affymetrix 6.0) only in African Americans. Genotyped SNPs were excluded based on call rate (<98%), monomorphic SNPs, lack of Hardy–Weinberg Equilibrium in controls ($P < 1 \times 10-4$), and low allele frequency (MAF<1%). DNA samples were excluded based on genotyping success rate (<97%), duplicate discordance or sex mismatch, or genetic ancestry outliers as determined by principal component analysis (52). CARe and WHI-SHARe samples were combined into a single set for phasing using BEAGLE Version 3.3.1 (53).

Exome sequencing, variant calling and QC

Detailed information on exome sequencing, variant calling and QC procedures is provided in the Supplementary Material, Methods. Briefly, as part of the NHLBI ESP, whole-exome sequencing was performed in 6823 individuals by the University of Washington (Seattle, WA, USA) or the Broad Institute (Cambridge, MA, USA). Each institute used a unified sequencing framework that included the following steps: (i) initial QC of sample DNA quantity and quality, (ii) library construction and exome capture or in-solution hybrid selection, (iii) cluster amplification and sequencing of enriched libraries, (iv) read mapping and variant analysis using the Genome Analysis Toolkit (GATK, refv1.2905) (27) and (v) evaluation of exome sequencing data against standard QC metrics. Variants were called using the UMAKE software pipeline provided by the University of Michigan, Ann Arbor, MI, USA (http://www.sph.umich.edu/csg/kang/umake/download/), which allowed all samples to be analyzed simultaneously for variant calling and filtering. For variant-level QC we filtered out variants that failed a set of SNP quality metrics in a support vector machine (SVM), and those with a read depth <10 or >500, low call rate or lack of Hardy–Weinberg equilibrium. For sample-level QC, we filtered out DNA samples that exhibited relatedness based on a kinship analysis, high missing rates, high homozygosity, sex mismatch, low concordance with genome-wide SNP array data and genetic ancestry outliers as determined by principal component analysis.

Genotype imputation to the ESP

Detailed information on imputation of sequence data is available in the Supplementary Material, Methods. As the reference panel for imputation we used the haplotypes from 1692 individuals of African ancestry and 471 of European ancestry with whole-exome sequence data from ESP. Study-specific reference data from different ancestral populations help improve imputation accuracy of low-frequency variants (54). A total of 1 077 164 autosomal SNPs, pre-phased across all 2163 samples using BEAGLE (53), were included in the reference panel. The target panel comprised genome-wide genotype data in WHI-SHARe and CARe obtained using the methods described above. The target panel was phased using BEAGLE (53), and

the phased target panel was imputed to the ESP data using Minimac (55). We used Rsq as the imputation quality measure for imputed SNPs (56). Based on prior experimental validation of imputed low-frequency variants (31), we excluded imputed SNPs such that variants with lower MAFs required higher imputation quality: for SNPs with MAF > 0.01, we excluded those with Rsq ≤ 0.3; for MAFs of 0.005–0.01, we excluded Rsq < 0.7; SNPs with MAF < 0.005 were excluded because most of these SNPs are not imputed well, and our study had limited statistical power to detect associations with these SNPs. In total, we imputed 375 024 autosomal markers into WHI-SHARe and CARe (see Supplementary Material, Methods). After exclusion based on imputation Rsq (110 134 SNPs) and MAF (68 111 SNPs), 196 779 imputed variants (80 438 with MAF ≥ 0.05; 82 329 with MAFs of 0.01–0.05; and 34 012 with MAF ≤ 0.01) remained in the dataset for association analysis.

Statistical analysis

To reduce the influence of outliers, separately in WHI-SHARe, CARe and ESP, we set all height values below the 0.05th percentile of the overall height distribution to the 0.05th percentile ($N = 38$ in WHI-SHARe, 28 in CARe, 29 in ESP); all values above the 99.5th percentile were set to the 99.5th percentile ($N = 38$ in WHI-SHARe, 28 in CARe, 35 in ESP). Unless otherwise indicated, all analyses were adjusted for age, sex (in CARe and ESP), study and the first two principal components derived from EIGENSTRAT (52) to account for global population substructure.

Association analysis

Association tests were performed separately in WHI-SHARe and CARe. We estimated the association between each variant and height using multivariable linear regression assuming additive genetic effects. Each sequenced SNP was coded as 0, 1 or 2 copies of the alternate (variant) allele. For imputed variants, we used the expected number of copies of the alternate allele between 0 and 2 (the dosage), which gives unbiased effect estimates (44). To combine study-specific estimates across studies, we performed inverse-variance weighted fixed-effects meta-analysis using METAL (57). We tested for heterogeneity between WHI-SHARe and CARe results using Cochran's Q statistics (58). Quantile–quantile (Q–Q) plots were used to assess whether the distribution of P-values was consistent with the null distribution (except for the extreme tail).

Replication of height association signals

Based on combined estimates in WHI-SHARe and CARe (the discovery set), we selected SNPs showing $P < 5E-06$ for replication in an independent African ancestry population with whole-exome sequencing data. We tested for associations between individual variants and height using multivariable linear regression assuming additive genetic effects. This was followed by an inverse-variance weighted fixed-effects meta-analysis combining the association results in WHI-SHARe, CARe and ESP (discovery + replication). To maintain an overall Type 1 error rate of 5%, we used a conservative Bonferroni-corrected threshold of $P < 2.5E-07$ ($=0.05/196\,779$ variants) to identify statistically significant SNPs in the combined analysis (discovery + replication).

Conditional association testing

To test whether variants in the same locus were independently associated with height, we used conditional analyses that simultaneously included two or more variants in a single model. When testing whether associated variants were independent of GWAS-identified SNPs, we performed conditional analyses only in CARe and WHI-SHARe (87% of the total study population) as these participants had available genome-wide SNP array as well as imputed HapMap data (59).

Adjustment for locus-specific ancestry

In recently admixed populations such as African Americans, even within a single individual, specific genomic regions can originate with varying frequencies from different ancestral populations (60). Recent findings in African Americans suggest this locus-specific ancestry (i.e. local ancestry) may associate with height on several chromosomes (5). To determine the extent to which local ancestry affected the observed associations, we inferred the ancestry of every participant at each statistically significant SNP and additionally adjusted for these estimates in sensitivity analyses. To estimate local ancestry, we used the Hidden Markov model (61) implemented in genome-wide Affymetrix 6.0 data, as in Carty *et al.* (5). Local ancestry estimates were coded as 0 = two European alleles, 1 = one European/one African allele, 2 = two African alleles at each GWAS SNP. At each height-associated SNP, the two immediately adjacent local ancestry estimates were always identical. Thus, we used these as the estimated local ancestry at each height-associated SNP. We regressed height on the SNP of interest, local ancestry, age, sex, study and global ancestry based on the first two principal components. Local ancestry-adjusted analyses were limited to CARe and WHI-SHARe (the discovery set) as these participants had available GWAS data.

We used R (Version 2.15.1, R Foundation for Statistical Computing, Vienna, Austria) to conduct the statistical analysis, and LocusZoom (62) to visualize results. To determine the minimum detectable effect estimates in the present analysis, we estimated statistical power using Quanto Version 1.2.4 (http://hydra.usc.edu/gxe/).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the contributions of the research institutions, study investigators, field staff and study participants, as well as the support of the National Heart, Lung, and Blood Institute (NHLBI) in creating this resource for biomedical research. Further acknowledgements can be found in the Supplementary Material, Methods.

Conflict of Interest statement: The authors declare no competing financial interests.

FUNDING

This work was supported by the National Heart, Lung, and Blood Institute (RC2 HL-103010, RC2 HL-102923, RC2 HL-102924,

RC2 HL-102925, RC2 HL-102926); and the National Cancer Institute (R25 CA094880 to M.D.). Further funding information can be found in the Supplementary Methods.

REFERENCES

1. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S. *et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**, 832–838.
2. Hirschhorn, J.N., Lindgren, C.M., Daly, M.J., Kirby, A., Schaffner, S.F., Burtt, N.P., Altshuler, D., Parker, A., Rioux, J.D., Platko, J. *et al.* (2001) Genomewide linkage analysis of stature in multiple populations reveals several regions with evidence of linkage to adult height. *Am. J. Hum. Genet.*, **69**, 106–116.
3. Perola, M., Sammalisto, S., Hiekkalinna, T., Martin, N.G., Visscher, P.M., Montgomery, G.W., Benyamin, B., Harris, J.R., Boomsma, D., Willemsen, G. *et al.* (2007) Combined genome scans for body stature in 6,602 European twins: evidence for common Caucasian loci. *PLoS Genet.*, **3**, e97.
4. Visscher, P.M., Medland, S.E., Ferreira, M.A., Morley, K.I., Zhu, G., Cornes, B.K., Montgomery, G.W. and Martin, N.G. (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet.*, **2**, e41.
5. Carty, C.L., Johnson, N.A., Hutter, C.M., Reiner, A.P., Peters, U., Tang, H. and Kooperberg, C. (2012) Genome-wide association study of body height in African Americans: the Women's Health Initiative SNP Health Association Resource (SHARe). *Hum. Mol. Genet.*, **21**, 711–720.
6. N'Diaye, A., Chen, G.K., Palmer, C.D., Ge, B., Tayo, B., Mathias, R.A., Ding, J., Nalls, M.A., Adeyemo, A., Adoue, V. *et al.* (2011) Identification, replication, and fine-mapping of loci associated with adult height in individuals of African ancestry. *PLoS Genet.*, **7**, e1002298.
7. Hao, Y., Liu, X., Lu, X., Yang, X., Wang, L., Chen, S., Li, H., Li, J., Cao, J., Chen, J. *et al.* (2013) Genome-wide association study in Han Chinese identifies three novel loci for human height. *Hum. Genet.*, **132**, 681–689.
8. Croteau-Chonka, D.C., Marvelle, A.F., Lange, E.M., Lee, N.R., Adair, L.S., Lange, L.A. and Mohlke, K.L. (2011) Genome-wide association study of anthropometric traits and evidence of interactions with age and study year in Filipino women. *Obesity (Silver Spring)*, **19**, 1019–1027.
9. Kim, J.J., Lee, H.I., Park, T., Kim, K., Lee, J.E., Cho, N.H., Shin, C., Cho, Y.S., Lee, J.Y., Han, B.G. *et al.* (2010) Identification of 15 loci influencing height in a Korean population. *J. Hum. Genet.*, **55**, 27–31.
10. Liu, J.Z., Medland, S.E., Wright, M.J., Henders, A.K., Heath, A.C., Madden, P.A., Duncan, A., Montgomery, G.W., Martin, N.G. and McRae, A.F. (2010) Genome-wide association study of height and body mass index in Australian twin families. *Twin. Res. Hum. Genet.*, **13**, 179–193.
11. Okada, Y., Kamatani, Y., Takahashi, A., Matsuda, K., Hosono, N., Ohmiya, H., Daigo, Y., Yamamoto, K., Kubo, M., Nakamura, Y. *et al.* (2010) A genome-wide association study in 19 633 Japanese subjects identified LHX3-QSOX2 and IGF1 as adult height loci. *Hum. Mol. Genet.*, **19**, 2303–2312.
12. Tonjes, A., Koriath, M., Schleinitz, D., Dietrich, K., Bottcher, Y., Rayner, N.W., Almgren, P., Enigk, B., Richter, O., Rohm, S. *et al.* (2009) Genetic variation in GPR133 is associated with height: genome wide association study in the self-contained population of Sorbs. *Hum. Mol. Genet.*, **18**, 4662–4668.
13. Estrada, K., Krawczak, M., Schreiber, S., van Duijn, K., Stolk, L., van Meurs, J.B., Liu, F., Penninx, B.W., Smit, J.H., Vogelzangs, N. *et al.* (2009) A genome-wide association study of northwestern Europeans involves the C-type natriuretic peptide signaling pathway in the etiology of human height variation. *Hum. Mol. Genet.*, **18**, 3516–3524.
14. Cho, Y.S., Go, M.J., Kim, Y.J., Heo, J.Y., Oh, J.H., Ban, H.J., Yoon, D., Lee, M.H., Kim, D.J., Park, M. *et al.* (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.*, **41**, 527–534.
15. Soranzo, N., Rivadeneira, F., Chinappen-Horsley, U., Malkina, I., Richards, J.B., Hammond, N., Stolk, L., Nica, A., Inouye, M., Hofman, A. *et al.* (2009) Meta-analysis of genome-wide scans for human adult stature identifies novel loci and associations with measures of skeletal frame size. *PLoS Genet.*, **5**, e1000445.
16. Lei, S.F., Yang, T.L., Tan, L.J., Chen, X.D., Guo, Y., Guo, Y.F., Zhang, L., Liu, X.G., Yan, H., Pan, F. *et al.* (2009) Genome-wide association scan for stature in Chinese: evidence for ethnic specific loci. *Hum. Genet.*, **125**, 1–9.
17. Johansson, A., Marroni, F., Hayward, C., Franklin, C.S., Kirichenko, A.V., Jonasson, I., Hicks, A.A., Vitart, V., Isaacs, A., Axenovich, T. *et al.* (2009) Common variants in the JAZF1 gene associated with height identified by linkage and genome-wide association analysis. *Hum. Mol. Genet.*, **18**, 373–380.
18. Gudbjartsson, D.F., Walters, G.B., Thorleifsson, G., Stefansson, H., Halldorsson, B.V., Zusmanovich, P., Sulem, P., Thorlacius, S., Gylfason, A., Steinberg, S. *et al.* (2008) Many sequence variants affecting diversity of adult human height. *Nat. Genet.*, **40**, 609–615.
19. Lettre, G., Jackson, A.U., Gieger, C., Schumacher, F.R., Berndt, S.I., Sanna, S., Eyheramendy, S., Voight, B.F., Butler, J.L., Guiducci, C. *et al.* (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat. Genet.*, **40**, 584–591.
20. Weedon, M.N., Lango, H., Lindgren, C.M., Wallace, C., Evans, D.M., Mangino, M., Freathy, R.M., Perry, J.R., Stevens, S., Hall, A.S. *et al.* (2008) Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.*, **40**, 575–583.
21. Sanna, S., Jackson, A.U., Nagaraja, R., Willer, C.J., Chen, W.M., Bonnycastle, L.L., Shen, H., Timpson, N., Lettre, G., Usala, G. *et al.* (2008) Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat. Genet.*, **40**, 198–203.
22. Weedon, M.N., Lettre, G., Freathy, R.M., Lindgren, C.M., Voight, B.F., Perry, J.R., Elliott, K.S., Hackett, R., Guiducci, C., Shields, B. *et al.* (2007) A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nat. Genet.*, **39**, 1245–1250.
23. Berndt, S.I., Gustafsson, S., Magi, R., Ganna, A., Wheeler, E., Feitosa, M.F., Justice, A.E., Monda, K.L., Croteau-Chonka, D.C., Day, F.R. *et al.* (2013) Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat. Genet.*, **45**, 501–512.
24. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
25. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H. and Nadeau, J.H. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.
26. Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloglu, A., Ozen, S., Sanjad, S. *et al.* (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 19096–19101.
27. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
28. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A. *et al.* (2010) Exome sequencing identifies the cause of a Mendelian disorder. *Nat. Genet.*, **42**, 30–35.
29. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A. and Shendure, J. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.*, **12**, 745–755.
30. Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C. *et al.* (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.*, **42**, 790–793.
31. Auer, P.L., Johnsen, J.M., Johnson, A.D., Logsdon, B.A., Lange, L.A., Nalls, M.A., Zhang, G., Franceschini, N., Fox, K., Lange, E.M. *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.
32. Johnsen, J.M., Auer, P.L., Morrison, A.C., Jiao, S., Wei, P., Haessler, J., Fox, K., McGee, S.R., Smith, J.D., Carlson, C.S. *et al.* (2013) Common and rare von Willebrand factor (VWF) coding variants, VWF levels, and factor VIII levels in African Americans: the NHLBI Exome Sequencing Project. *Blood*, **122**, 590–597.
33. Parra, E.J., Marcini, A., Akey, J., Martinson, J., Batzer, M.A., Cooper, R., Forrester, T., Allison, D.B., Deka, R., Ferrell, R.E. *et al.* (1998) Estimating African American admixture proportions by use of population-specific alleles. *Am. J. Hum. Genet.*, **63**, 1839–1851.
34. Cooper, R.S., Tayo, B. and Zhu, X. (2008) Genome-wide association studies: implications for multiethnic samples. *Hum. Mol. Genet.*, **17**, R151–R155.

35. McCarthy, M.I. and Hirschhorn, J.N. (2008) Genome-wide association studies: potential next steps on a genetic journey. *Hum. Mol. Genet.*, **17**, R156–R165.
36. Gong, J., Schumacher, F., Lim, U., Hindorff, L.A., Haessler, J., Buyske, S., Carlson, C.S., Rosse, S., Buzkova, P., Fornage, M. *et al.* (2013) Fine mapping and identification of BMI loci in African Americans. *Am. J. Hum. Genet.*, **93**, 661–671.
37. Peters, U., North, K.E., Sethupathy, P., Buyske, S., Haessler, J., Jiao, S., Fesinmeyer, M.D., Jackson, R.D., Kuller, L.H., Rajkovic, A. *et al.* (2013) A systematic mapping approach of 16q12.2/FTO and BMI in more than 20,000 African Americans narrows in on the underlying functional variation: results from the Population Architecture using Genomics and Epidemiology (PAGE) study. *PLoS Genet.*, **9**, e1003171.
38. Duan, Q., Liu, E.Y., Auer, P.L., Zhang, G., Lange, E.M., Jun, G., Bizon, C., Jiao, S., Buyske, S., Franceschini, N. *et al.* (2013) Imputation of coding variants in African Americans: better performance using data from the exome sequencing project. *Bioinformatics*, **29**, 2744–2749.
39. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
40. Consortium, E.P., Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
41. Consortium, E.P., Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C. and Snyder, M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
42. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
43. Reese, M.G., Eeckman, F.H., Kulp, D. and Haussler, D. (1997) Improved splice site detection in Genie. *J. Comput. Biol.*, **4**, 311–323.
44. Jiao, S., Hsu, L., Hutter, C.M. and Peters, U. (2011) The use of imputed values in the meta-analysis of genome-wide association studies. *Genet. Epidemiol.*, **35**, 597–605.
45. Huang, L., Wang, C. and Rosenberg, N.A. (2009) The relationship between imputation error and statistical power in genetic association studies in diverse populations. *Am. J. Hum. Genet.*, **85**, 692–698.
46. Zheng, J., Li, Y., Abecasis, G.R. and Scheet, P. (2011) A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genet. Epidemiol.*, **35**, 102–110.
47. Liu, E.Y., Buyske, S., Aragaki, A.K., Peters, U., Boerwinkle, E., Carlson, C., Carty, C., Crawford, D.C., Haessler, J., Hindorff, L.A. *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.
48. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 9362–9367.
49. Emond, M.J., Louie, T., Emerson, J., Zhao, W., Mathias, R.A., Knowles, M.R., Wright, F.A., Rieder, M.J., Tabor, H.K., Nickerson, D.A. *et al.* (2012) Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis. *Nat. Genet.*, **44**, 886–889.
50. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.
51. Boileau, C., Guo, D.C., Hanna, N., Regalado, E.S., Detaint, D., Gong, L., Varret, M., Prakash, S.K., Li, A.H., d'Indy, H. *et al.* (2012) TGFB2 mutations cause familial thoracic aortic aneurysms and dissections associated with mild systemic features of Marfan syndrome. *Nat. Genet.*, **44**, 916–921.
52. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
53. Browning, S.R. and Browning, B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, **81**, 1084–1097.
54. Howie, B., Marchini, J. and Stephens, M. (2011) Genotype imputation with thousands of genomes. *G3*, **1**, 457–470.
55. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. and Abecasis, G.R. (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.*, **44**, 955–959.
56. Li, Y., Willer, C.J., Ding, J., Scheet, P. and Abecasis, G.R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816–834.
57. Willer, C.J., Li, Y. and Abecasis, G.R. (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, **26**, 2190–2191.
58. Cochran, W.G. (1954) The combination of estimates from different experiments. *Biometrics*, **10**, 101–129.
59. International HapMap, C. (2003) The International HapMap Project. *Nature*, **426**, 789–796.
60. Sankararaman, S., Sridhar, S., Kimmel, G. and Halperin, E. (2008) Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.*, **82**, 290–303.
61. Tang, H., Coram, M., Wang, P., Zhu, X. and Risch, N. (2006) Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet.*, **79**, 1–12.
62. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R. and Willer, C.J. (2010) LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, **26**, 2336–2337.