ARTICLE

Imputation of Exome Sequence Variants into Population-Based Samples and Blood-Cell-Trait-Associated Loci in African Americans: NHLBI GO Exome Sequencing Project

Paul L. Auer,[1,17] Jill M. Johnsen,[2,3,17] Andrew D. Johnson,[4,17] Benjamin A. Logsdon,[1,17] Leslie A. Lange,[5,17] Michael A. Nalls,[6] Guosheng Zhang,[5] Nora Franceschini,[5] Keolu Fox,[3] Ethan M. Lange,[5] Stephen S. Rich,[7] Christopher J. O'Donnell,[4] Rebecca D. Jackson,[8] Robert B. Wallace,[9] Zhao Chen,[10] Timothy A. Graubert,[11] James G. Wilson,[12] Hua Tang,[13,17] Guillaume Lettre,[14,17] Alex P. Reiner,[1,6,17,*] Santhi K. Ganesh,[15,17] and Yun Li[5,16,17,*]

Researchers have successfully applied exome sequencing to discover causal variants in selected individuals with familial, highly penetrant disorders. We demonstrate the utility of exome sequencing followed by imputation for discovering low-frequency variants associated with complex quantitative traits. We performed exome sequencing in a reference panel of 761 African Americans and then imputed newly discovered variants into a larger sample of more than 13,000 African Americans for association testing with the blood cell traits hemoglobin, hematocrit, white blood count, and platelet count. First, we illustrate the feasibility of our approach by demonstrating genome-wide-significant associations for variants that are not covered by conventional genotyping arrays; for example, one such association is that between higher platelet count and an *MPL* c.117G>T (p.Lys39Asn) variant encoding a p.Lys39Asn amino acid substitution of the thrombpoietin receptor gene ($p = 1.5 \times 10^{-11}$). Second, we identified an association between missense variants of *LCT* and higher white blood count ($p = 4 \times 10^{-13}$). Third, we identified low-frequency coding variants that might account for allelic heterogeneity at several known blood cell-associated loci: *MPL* c.754T>C (p.Tyr252His) was associated with higher platelet count; *CD36* c.975T>G (p.Tyr325*) was associated with lower platelet count; and several missense variants at the α-globin gene locus were associated with lower hemoglobin. By identifying low-frequency missense variants associated with blood cell traits not previously reported by genome-wide association studies, we establish that exome sequencing followed by imputation is a powerful approach to dissecting complex, genetically heterogeneous traits in large population-based studies.

Introduction

Blood cell counts are heritable traits that represent important intermediate phenotypes for a variety of cardiovascular, hematologic, oncologic, immunologic, and infectious diseases. Traits such as hemoglobin and leukocyte or white blood cell count (WBC) are known to differ by ancestry. Through admixture mapping, the *DARC* null allele has been identified as a major determinant of the lower WBC in African-descended individuals as compared with European Americans.[1] Recent genome-wide association studies (GWASs) have further contributed to our understanding of the variability of blood cell traits both within and across European-, African-, and Asian-descended populations. For example, the CHARGE, CARe, COGENT, and HaemGen consortia have identified ~100 loci associated with blood cell traits, including hemoglobin concentration, hematocrit, red blood cell indices, WBC, platelet count, and mean platelet volume (MPV).[2–10]

Lower-frequency or population-specific variants that are not well captured by current genome-wide genotyping platforms might account for inter-individual differences in blood cell counts, particularly among individuals of African ancestry. Alternatively, aggregations of rare variants might cause positive GWAS signals. Using the sickle cell anemia *HBB* Glu6Val variant (MIM 141900) as an example, Dickson et al. recently showed that multiple rare variants can account for some of the signals reported in GWASs for common, complex traits.[11] A number of other rare genetic variants are known to contribute to familial disorders involving the number, shape, and function of blood cells, but the extent to which these rare coding

[1]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle WA 98109, USA; [2]Research Institute, Puget Sound Blood Center, Seattle WA 98109, USA; [3]Departments of Medicine and Epidemiology and Genome Sciences, University of Washington, Seattle WA 98195, USA; [4]National Heart, Lung, and Blood Institute Center for Population Studies, The Framingham Heart Study, Framingham, MA 01702, USA; [5]Departments of Epidemiology, Genetics and Biostatistics, Bioinformatics and Computational Biology, University of North Carolina, Chapel Hill, NC 27599, USA; [6]Molecular Genetics Section, Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD 20892, USA; [7]Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia, Charlottesville, VA 22908, USA; [8]Division of Endocrinology, Diabetes and Metabolism, Ohio State University, Columbus, OH 43210, USA; [9]Department of Epidemiology, University of Iowa College of Public Health, Iowa City, IA 52242, USA; [10]Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, AZ 85724, USA; [11]Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA; [12]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 39216, USA; [13]Department of Statistics and Department of Genetics, Stanford University, Stanford, CA 94305, USA; [14]Montreal Heart Institute and Département de Médecine, Université de Montréal, Montréal, QC H1T 1C8, Canada; [15]Division of Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan 48108, USA; [16]On behalf of the National Heart, Lung, and Blood Institute GO Exome Sequencing Project
[17]These authors contributed equally to this work
*Correspondence: apreiner@u.washington.edu (A.P.R.), yunli@med.unc.edu (Y.L.)
http://dx.doi.org/10.1016/j.ajhg.2012.08.031.

variants or polymorphisms in these genes contribute to blood cell traits in the general population is unknown.

Exome sequencing has recently allowed the discovery of rare variants associated with highly penetrant Mendelian disorders.[12] The application of such next-generation sequencing technologies to more complex phenotypes in large population samples is currently limited by sequencing costs as well as technical and analytical challenges. The availability of genome-wide genotype data in many large, population-based cohort studies, combined with current genome-wide imputation approaches, offers the opportunity to extend the assessment of rare coding variants discovered through exome sequencing into larger population-based samples. As proof-of-principle, we apply this imputation approach to exome sequence data generated in a subset of African American cohort participants from the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP). Using a reference panel of 761 African Americans with exome sequence data, we imputed newly identified variants into the full African American cohorts comprising more than 13,000 individuals with blood cell phenotypes and genome-wide SNP genotyping data from four population-based studies. This allowed us to assess whether lower-frequency variants not well captured on current GWAS platforms are associated with blood cell traits in individuals of African ancestry and to identify potential genomic loci that might contribute to our understanding of hematopoiesis and blood cell biology.

Subjects and Methods

Subjects

Participants included a total of 15,829 self-identified African Americans with GWAS data from four population-based cohorts. Of these individuals, 761 had undergone whole-exome sequencing as part of the first phase of the NHLBI ESP. Detailed descriptions of each cohort are provided below. Clinical information was collected by self report and in-person examination. All participants provided written informed consent as approved by local human-subjects committees. We excluded study participants on the basis of pregnancy, cancer, or AIDS diagnosis at the time of the blood count.

Women's Health Initiative (WHI)

WHI is one of the largest ($n = 161,808$) studies of women's health ever undertaken in the U.S. There are two major components of WHI: (1) a clinical trial (CT) that enrolled and randomized 68,132 women ages 50–79 into at least one of three placebo-control clinical trials (hormone therapy, dietary modification, and supplementation with calcium and vitamin D); and (2) an observational study (OS) that enrolled 93,676 women of the same age range into a parallel prospective cohort study.[13] A diverse population including 26,045 (17%) women from minority groups was recruited from 1993–1998 at 40 clinical centers across the U.S. Of the CT and OS minority participants enrolled in WHI, 12,157 (including 8,515 self-identified African American and 3,642 self-identified Hispanic subjects) who had consented to genetic research were eligible for the WHI-SHARe GWAS project. Of these eligible individuals, 8,095 African Americans were included in the current study.

Atherosclerosis Risk in Communities Study (ARIC)

The ARIC study is a prospective population-based study of atherosclerosis and cardiovascular diseases in 15,792 men and women, including 11,478 non-Hispanic whites and 4,314 African Americans drawn from four U.S. communities (suburban Minneapolis, Minnesota; Washington County, Maryland; Forsyth County, North Carolina; and Jackson, Mississippi).[14] Only self-reported African American participants are included in this analysis. Participants were between ages 45 and 64 years at their baseline examination in 1987–1989, when blood was drawn for DNA extraction and participants consented to genetic testing. Blood for complete blood-count analysis was drawn at the baseline exam. After the availability of adequate amounts of high-quality DNA was taken into account and appropriate informed-consent and genotyping quality-control and assurance procedures were put in place, genome-wide genotype data were available for 2,989 African Americans, of whom 2,779 were included in the current analysis.

Coronary Artery Risk Development in Young Adults (CARDIA)

The CARDIA study is a prospective, multi-center investigation of the natural history and etiology of cardiovascular disease in a cohort of African Americans and whites who were 18–30 years of age at the time of initial examination.[15] The CARDIA sample was recruited at random during 1985–1986 primarily from populations based in Birmingham, Alabama; Chicago, Illinois; and Minneapolis, Minnesota; and Oakland, California from the membership of the Kaiser-Permanente Health Plan. The initial examination included 5,115 participants selectively recruited to represent proportionate racial, gender, age, and education groups from each of the four communities. From the time of initiation of the study in 1985–1986 (baseline examination), six follow-up examinations were conducted 2, 5, 7, 10, 15, 20, and 25 years later. DNA extraction for genetic studies was performed at the 10 year examination. After the availability of adequate amounts of high-quality DNA was taken into account and appropriate informed-consent and genotyping quality-control and assurance procedures were put in place, genome-wide genotype data were available for 955 African Americans, of whom 953 were included in the current analysis.

Jackson Heart Study (JHS)

The Jackson Heart Study (JHS) is a prospective population-based study aimed at identifying the reasons for the high prevalence of common complex diseases among African Americans in the Jackson, Mississippi metropolitan area. Such diseases include cardiovascular disease, type-2 diabetes, obesity, chronic kidney disease, and stroke.[16] During the baseline examination period (2000–2004), 5,301 self-identified African Americans were recruited from four sources, including (1) randomly sampled households from a commercial listing; (2) ARIC participants; (3) a structured volunteer sample that was designed to mirror the eligible population; and (4) a nested family cohort. Unrelated participants were between 35 and 84 years old, and members of the family cohort were $\geq$ 21 years old when consent for genetic testing was obtained and blood was drawn for DNA extraction. On the basis of DNA availability, appropriate informed consent, and genotyping results that met quality-control procedures, genome-wide genotype data

were available for 3,030 individuals, including 885 who are also ARIC participants. In the current study, JHS participants who were also enrolled in the ARIC study were analyzed with the ARIC data set—for this reason, the JHS data set analyzed here is defined as 2,145 individuals, of whom 2,132 are included in the current analysis.

Blood Cell Phenotype Data

Samples for complete blood count (CBC) analysis were collected at baseline by venipuncture into tubes containing ethylenediaminetetraaetic acid (EDTA). CBCs were performed at local clinical laboratories with automated hematology cell counters and standardized quality-assurance procedures.[17] Hematocrit was reported as a percentage, and hemoglobin was reported as g/dl. WBC and platelet count were reported in millions of cells per ml. The absolute numbers of each WBC subtype were calculated by multiplying the proportion of the WBC for each cell type by the total WBC for each individual. For each blood cell trait, we excluded individuals with (1) extreme outlying values (>10 standard deviations from the mean) or (2) any highly discordant values among the subset of ~3,000 participants who had serial CBC measurements available in the cohort database at a subsequent time point. After exclusions, data on blood-cell-count phenotypes were available for up to 13,959 ESP African American participants. Data on WBC subtype was available only in the ARIC, CARDIA, and JHS participants.

Exome Sequencing, Variant Calling, and QC

Through ESP, exome sequence data were available on a total of 761 African American participants. These include 362 from WHI as part of the WHI Sequencing Project (WHISP) and a total of 399 from ARIC (n = 216) and JHS (n = 166) as part of HeartGO. These WHISP and HeartGO participants were selected on the basis of primary phenotypes for ESP, which included extremes of body mass index, blood pressure, LDL cholesterol, early-onset myocardial infarction, and stroke. 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 the total mass, concentration, or integrity of DNA or the quality of preliminary genotyping data was too low or if sex typing was discordant. After QC, ~3 μg genomic, nonamplified DNA extracted from peripheral blood leukocytes was reformatted into 96-well plates for shotgun library preparation and exome capture.[18] Library construction steps included DNA fragmentation, end polishing and A-tailing, ligation of sequencing adaptors, and PCR amplification. Sample shotgun libraries were captured for exome enrichment with one of the following in-solution capture products: CCDS 2008 (~26 Mb), Roche/Nimblegen SeqCap EZ Human Exome Library v1.0 (~32 Mb; Roche Nimblegen EZ Cap v1), or EZ Cap v2 (~34 Mb). The fragment size distribution of the libraries was highly consistent (typically 125 ± 15 bp). Cluster amplification of denatured templates and hybridization was performed via bridge PCR with Genome Analyzer v3, Genome Analyzer v4, or HiSeq 2000 v2 cluster chemistry and flow cells (Illumina). Sequencing was performed on an Illumina GAIIx or HiSeq 2000 with paired-end 76 base runs or 50 base runs, respectively.

For read mapping and variant analysis, samples were aligned to a human reference (hg19) with BWA (Burrows-Wheeler Aligner).[19] Data were processed with the Genome Analysis ToolKit (GATK refv1.2905[20]). Reads were locally realigned (GATK IndelRealigner), and their base qualities were recalibrated (GATK Table Recalibration). Variant detection and genotyping were performed on both exomes and the flanking 50 bp of intronic sequence. Typical mean coverage of the target was 60×–80×. In brief, we took a two-step approach for discovering and genotyping candidate sites. First, we genertated genotype likelihood files (GLFs) by using samtools pileup on individual binary alignment map (BAM) files. Next, we used glfMultiples—a multi-sample variant caller—to generate initial single-nucleotide variant (SNV) calls.[21] This process allowed us to perform multi-sample calling of variants across thousands of samples. In brief, the distribution of observed bases and quality scores at each location (conditional on the true genotype) was modeled according to the MAQ model.[22] Using maximum likelihood, we then estimated an allele frequency for each site under the assumption that genotypes were segregating in Hardy-Weinberg proportions. For the initial variant calls, we assumed that the prior probability that a site was polymorphic was 0.0091, corresponding to an estimate of the prior for a segregating site in a simple population-genetics model and an estimated per-sample, per-base-pair heterozygosity of $\theta = 0.001$. We assumed that transition and transversion mutations were equally likely, thereby allowing the use of transition-transversion (T_i/T_v) ratio as a diagnostic of SNP call quality. Genotypes for each site and corresponding posterior probabilities were then calculated with the Bayes theorem, for which the Hardy-Weinberg proportions were used as priors and the MAQ error model was used for describing the conditional probability of bases and quality scores given the true genotype.

After these initial SNV calls were generated, we re-examined the BAM files to collect additional information about each variant site. 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. All individual exome-sequencing data were evaluated against the QC metrics of both bulk and per-sample properties, including the distribution of novel and known variants relative to dbSNP, fingerprint concordance, Ti/Tv ratio, Het/Hom ratio, library complexity, capture efficiency and uniformity, and coverage distribution (90% at ≥ 8× was required for completion at the University of Washington; 70% at > 20× was required for completion at the Broad Institute).

Variant data for each sample were formatted (variant call format [VCF]) as "raw" calls for all samples. The final SNP call set included variants that were called with posterior probability >99% (glfMultiples SNP quality >20), that were >5bp away from an indel detected in the 1000 Genomes Pilot Project, that were covered by at least one read in 85% of samples, and that had total depth across samples of between 2,500 and 2,500,000 reads (~1–100 reads per sample). Sites for which >65% of reads were heterozygotes carrying the variant allele or where the absolute squared correlation between allele (variant or reference) and strand (forward or reverse) was >0.15 were excluded.

Genotype-wide Genotyping and QC

The Affymetrix 6.0 platform was used for genome-wide genotyping in all participants, either at Affymetrix for the WHI-SHARe project or at the Broad Institute for the NHLBI Candidate Gene Association Resource (CARe) consortium (ARIC, CARDIA, and JHS). Genotyping and quality-control procedures for both WHI-SHARe and CARe have been described in detail.[6,10,23]

Principal-component analysis was implemented in EIGENSTRAT[24] within each African American cohort on cleaned genome-wide-association (GWA) genotype data. DNA samples with a genome-wide genotyping success rate of <90%, duplicate discordance or sex mismatch, genetic ancestry outliers (as determined by cluster analysis performed via principal-component analysis), SNPs with a genotyping success rate of <95%, monomorphic SNPs, SNPs with minor-allele frequency (MAF) <1%, and SNPs that mapped to several genomic locations were removed from the analyses. A total of 838,337 Affy6.0 SNPs were used for genotype imputation, as described below.

Imputation of "Virtual Exomes" and QC

We used the 761 African American ESP participants as a reference sample to impute variants identified by exome sequencing into the larger target sample of 15,829 African Americans with GWAS data from WHI, ARIC, CARDIA, and JHS. Specifically, we constructed, by using MaCH 1.0.18,[25] an internal reference of 1,522 phased haplotypes from the 761 African Americans with both whole-exome sequencing and Affymetrix 6.0 genotyping data. We also prephased the remaining ARIC, CARDIA, JHS, and WHI participants with genotyping data only. We then used minimac[26] to impute genotypes at markers discovered in ESP for the 15,829 African Americans. Minimac is a low-memory, computationally efficient implementation of the MaCH algorithm[25] for genotype imputation and is designed to work on phased genotypes. It can handle very large reference panels with hundreds or thousands of haplotypes. After careful QC and marker and strand matching, we imputed 420,876 markers from 838,337 markers on Affymetrix 6.0. Given the size of the reference panel, variants with MAF <0.1% were further excluded from association analyses, leaving 368,093 exome-sequencing-identified variants of MAF 0.1% or greater. Prior to data analysis, SNPs were additionally excluded if imputation quality metrics (Rsq, which is equivalent to the squared correlation between proximal imputed and genotyped SNPs) were less than the MAF-based thresholds chosen such that within each MAF category, SNPs passing the QC threshold had an average Rsq of 0.8 or more. After exclusion based on MAF and imputation quality scores, 178,954 imputed exome variants remained in the data set for association analysis. For post-imputation QC, we evaluated imputation quality by two methods: (1) comparing experimental genotypes from the Metabochip in a subset of 1,830 among the 8,059 WHI individuals;[27] and (2) imputing the 761 reference individuals one at a time: this involves masking genotypes at the exome SNPs for one reference individual at a time, imputing the masked genotypes by using the remaining 1,520 haplotypes, and comparing the imputed with the masked experimental genotypes. The two evaluation methods resulted similar conclusions, confirming that using an average Rsq of $\geq$ 0.8 is an effective postimputation filter for rare variants.

Statistical Analysis

Association analyses for quantitative blood cell traits were performed separately for the WHI and CARe cohorts (ARIC, CARDIA, and JHS) via linear regression as implemented in MACH2QTL v. 1.08. Allelic dosage at each SNP (a value between 0.0 and 2.0, calculated on the basis of the probability of each of the three possible genotypes) was used as the independent variable, adjusted for primary covariates of age, sex, cohort, and global ancestry, as represented by principal-component analysis of the GWA data. All WBC traits were natural-log-transformed so that the distributions of the phenotypic data would be normalized. The WBC analysis was additionally adjusted for genotype at rs2814778, the African-specific *DARC* or Duffy blood group null variant rs2814778, which is known to account for 15%–20% of the WBC variance among African Americans.[1,6]

For each blood cell phenotype, WHI and ARIC, CARDIA, and JHS study-level regression results were combined via inverse-variance weighted fixed-effects meta-analysis to derive an overall p value and effect estimate for each SNP. Meta-analyses were implemented with the METAL[28] software and corrected for genomic inflation factors (λ). Between-study heterogeneity of results was assessed with Cochran's Q statistic and the I^2 inconsistency metric. Because genome-wide association testing was performed for more than 1 million genotyped and imputed SNPs, we used a significance threshold of $\alpha = 2.5 \times 10^{-8}$, which has been suggested for African-ancestry populations so that an overall type 1 error rate of 5% can be maintained. We did not additionally correct for testing four phenotypes.

To detect transitions in ancestry along the genome, we used a Hidden Markov Model and local haplotype structure to estimate locus-specific ancestry (probabilities of whether an individual has 0, 1, or 2 alleles of African ancestry) at the *LCT* gene locus for each participant.[29,30] Phased haplotype data from the HapMap CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection) and YRI (Yoruba in Ibadan, Nigeria) individuals were used as reference panels. To assess the impact of local ancestry at the *LCT* locus on WBC phenotype association, we repeated SNP genotype-WBC linear regression analysis and adjusted for local ancestry proportion as a covariate. For annotation of coding variants, we used CONsensus DELeteriousness (Condel) scores to predict the impact of nonsynonymous single-nucleotide variants on protein function.[31]

Gene Expression Quantitative-Trait Loci (eQTL) Analysis

The associated *LCT* missense variants and their correlated proxies ($r^2 > 0.80$) were identified in the YRI in the 1000 Genomes Project data. We queried these SNPs against several web data sources to determine potential eQTLs linked to the WBC signal: SCAN (SNP and CNV Annotation Database) including YRI gene eQTLs in lymphoblastoid cell lines (LCLs)[32] and RNA-sequencing eQTLs in YRI LCLs from.[33] Additional *cis*-eQTL results were queried from an assembled database of associations in a wide variety of tissues derived primarily from European-ancestry populations: fresh lymphocytes,[34] fresh leukocytes,[35] leukocyte samples in individuals with Celiac disease,[36] LCLs derived from asthmatic children,[37] HapMap LCLs from three populations,[38] HapMap CEU LCLs from a separate study,[39] peripheral blood monocytes,[40,41] omental[42] and subcutaneous adipose cells,[42,43] stomach[42] and whole blood samples,[43,44] endometrial carcinomas,[45] brain cortex from two studies,[41,46] prefrontal cortex,[47] brain regions including prefrontal cortex, visual cortex, and cerebellum from three large studies, liver,[42,48–50] osteoblasts,[51] skin,[52] and additional fibroblast, T cell, and LCL samples.[53]

Assessment of Fine-Scale Population Structure at *LCT*

Recently, Mathieson and McVean[54] raised a concern that existing approaches (such as principal-component adjustment) might not adequately control for population stratification in rare-variant association studies. In such a situation, any rare variants that have higher frequencies within a sharply defined geographic region will appear to be associated with the phenotype. To

Table 1. Characteristics of ESP African American Participants[a]

Study	Atherosclerosis Risk in Communities (ARIC)	Coronary Artery Risk Development in Young Adults (CARDIA)	Jackson Heart Study (JHS)	Women's Health Initiative (WHI)
Sample size	2779	953	2132	8095
Age, years (SD)	53.3 (5.8)	24.4 (3.8)	50.0 (12.1)	61.6 (7.0)
% Female	63.2	61.3	60.8	100
Hemoglobin, g/dL (SD)	13.2 (1.5)	13.7 (1.5)	13.1 (1.5)	12.9 (1.0)
Hematocrit, % (SD)	40.2 (4.5)	41.1 (4.5)	39.5 (4.3)	39.1 (3.0)
WBC (SD) × 10^9/liter	5.69 (1.85)	5.93 (2.01)	5.66 (1.81)	5.63 (1.86)
Platelets (SD) × 10^9/liter	257.0 (66.2)	282.3 (68.5)	256.6 (65.7)	250.5 (62.6)
Neutrophils D) × 10^9/liter	2.89 (1.45)	3.11 (1.57)	3.15 (1.51)	NA
Lymphocytes (SD) × 10^9/liter	2.17 (0.75)	2.23 (0.84)	1.96 (0.66)	NA
Monocytes (SD) × 10^9/liter	0.344 (0.197)	0.314 (0.184)	0.390 (0.145)	NA
Eosinophils (SD) × 10^9/liter	0.165 (0.165)	NA	0.139 (0.126)	NA
Basophils (SD) × 10^9/liter	0.037 (0.043)	0.046 (0.044)	0.033 (0.022)	NA

Abbreviations are as follows: SD = standard deviation; NA = not available.
[a]n = 13,959.

formally test this possibility, we performed genome-wide association testing in which we treated carrier status at *LCT* rs35940156 as the "disease phenotype" or dependent variable in a regression model; variants that were enriched among carriers of rs35940156 were expected to show small p values. Specifically, we defined a quantitative phenotype by the number of minor alleles at *LCT* s35940156. We then performed genome-wide association analysis by adjusting for principal components and used METAL to combine the results from the CARe and WHI-SHARe cohorts. As references, we selected 58 control SNPs, which we selected to have the same minor-allele frequencies as rs35940156 in the 1000 Genomes CEU and YRI populations (~2.2% in YRI, 2% in the ESP-sequenced individuals, and 0% in CEU) and repeated the genome-wide scan for the phenotype or population stratum defined at each of these control SNPs. To eliminate linkage disequilibrium (LD) at linked regions, we removed variants residing on the same chromosomes as the respective control SNPs.

Results

We performed an exome-wide analysis of four hematologic traits (hemoglobin, hematocrit, WBC, and platelet count) in African Americans from four population-based cohorts (ARIC, CARDIA, JHS, and WHI). The characteristics of each cohort are summarized in Table 1. We imputed SNPs that we had identified by sequencing 761 African Americans from the NHLBI GO ESP to the full population-based sample of 13,959 African Americans. We used Affymetrix 6.0 genotype and blood-cell-count data for hemoglobin, hematocrit, WBC, and platelet-count phenotypes.

In contrast to conventional genome-wide SNP analyses, the variants interrogated in the current study were enriched in coding regions and had lower MAFs, and they included 178,954 high-quality, exome-sequencing-identified variants with a MAF of 0.1% or more. Of these 178,954 variants, 104,386 are present in dbSNP (build 131) and 1000 Genomes, 13,219 are in dbSNP only, and 19,979 are in 1000 Genomes only. Thus, 41,370, or 23%, of the imputed variants are "novel." The distribution of the 178,954 high-quality imputed exome variants by functional category is shown in Table S1 in the Supplemental Data available with this article online. The average depth of coverage for the 100,902 high-quality variants located within exons was 101.05. The average coverage depth for the 78,052 high-quality variants located outside exons was 61.89. Of the nonexonic variants, 94.4% were located within 50 bp of an exon, and the median distance from an exon was 22 bp.

Because lower-frequency variants are more challenging for imputation, we have applied more stringent postimputation quality filtering by using higher Rsq thresholds for rarer variants. Specifically, we chose an Rsq threshold of 0.9, 0.8, 0.6, 0.3, and 0.3 for SNPs with MAF 0.1%–0.5%, 0.5%–1%, 1%–3%, 3%–5%, and >5%, respectively. Under these criteria, 92.9% of SNPs with MAF > 5% passed QC, whereas only 7.3% of SNPs with MAF = 0.1%–0.5% passed QC (Table 2). By comparing imputed genotypes with experimental genotypes from the Metabochip,[27] we estimated that the information content retained (measured by dosage r^2, squared Pearson correlation between imputed and experimental genotypes) is > 81% across the MAF categories (Table 2). To further evaluate the sensitivity of our imputation, we examined the coverage of coding SNPs found on the Illumina Exome-Chip; 30.6% of the coding variants on the Exome-Chip were covered by the raw imputed data, whereas 15.4% were covered by

Table 2. Filtering of Exome Variants on the Basis of Allele Frequency and Imputation Quality

MAF	Rsq Threshold	Number of Imputed ESP SNPs	Number of QC+ Imputed ESP SNPs	Percent QC+ Imputed ESP SNPs	Average Dosage r^2
0.1%–0.5%	0.9	135,942	9,885	7.3%	83.17%
0.5%–1%	0.8	52,155	17,872	34.3%	84.01%
1%–3%	0.6	69,158	47,801	69.1%	81.88%
3%–5%	0.3	26,060	23,882	91.6%	81.04%
>5%	0.3	84,478	78,507	92.9%	84.40%

the imputed data that passed QC. These results met our expectations given that the imputation reference panel (761 African Americans) was less than 1/3 the size of the African American sample that was used in the design of the Exome-Chip array.

The genomic-control-corrected QQ plot for hemoglobin, WBC, and platelet count are shown in Figure S1, and the corresponding Manhattan plots are shown in Figure S2. Several regions reached genome-wide significance at the threshold of $p < 2.5 \times 10^{-8}$ (Table 3), as described in further detail below. GWAS findings for common variants derived from conventional genotyping arrays such as Affymetrix 6.0 and HapMap imputation have been reported in African Americans for WBC,[6] platelet count,[10] and hemoglobin and hematocrit.[8] Therefore, only loci newly identified through exome imputation are highlighted here.

Hemoglobin and Hematocrit

Globin Gene Variants are Significantly Associated with Hemoglobin and Hematocrit

The known rs334 sickle cell β-globin p.Glu6Val variant of *HBB* (MIM 141900) (MAF = 0.05), which is not covered by conventional GWA panels, was significantly associated with lower hematocrit ($p = 5.7 \times 10^{-11}$) and hemoglobin ($p = 9.7 \times 10^{-6}$). Ten SNPs located on the terminal portion of the long arm of chromosome 16 near the α-globin locus were significantly associated with hemoglobin (Table 3; see also Table S3). The SNP most strongly associated with hemoglobin was rs9924561 (MAF = 0.07), located in the next-to-last intron of *ITFG3* ($p = 6.6 \times 10^{-24}$) (Figure 1A). Another common variant (*LUC7L* [MIM 607782] rs1211375) on chromosome 16p13 was previously associated with hemoglobin in African Americans.[8] Of the remaining chromosome 16 α-globin-region SNPs significantly associated with hemoglobin, two are located within coding sequence (both synonymous): rs13335497 (c.423G>A [p.Ser141 =]; RefSeq NM_032039.2) of *ITFG3* (MAF = 0.07, $p = 1.8 \times 10^{-21}$) and rs11863726 (c.93A>G [p.Glu31 =]; RefSeq NM_005331.4) of *HBQ1* (MAF = 0.27, $p = 8.9x\ 10^{-11}$). *ITFG3* rs13335497 is in strong LD with *ITFG3* rs9924561 ($r^2 = 0.98$). In contrast, *HBQ1* (MIM 142240) rs11863726 is weakly correlated with other hemoglobin-associated SNPs in the region (maximum $r^2 = 0.35$ with rs9929571) and is 3 bp from a splice junction in the gene encoding fetal hemoglobin theta. Several lower-frequency (MAF = 0.02–0.03) nonsynonymous SNPs in the α-globin cluster were nominally associated with lower hemoglobin (Table S2): rs76613236 (c.155G>C [p.Gly52Ala]; RefSeq NM_005331.4) located in *HBQ1* (hemoglobin-theta), rs143256173 (c.45C>A [p.Asp15Glu]; RefSeq NM_001003938.3) in *HBM* (hemoglobin-mu [MIM 609639]), and rs144091859 (c.1600G>A [p.Asp534Asn]; RefSeq NM_032039.2) in *ITFG3* (p values in the range of 10^{-5}–10^{-7}). These three low-frequency coding variants of the fetal hemoglobin genes are in strong LD with one another (pair-wise $r^2 \approx 0.9$) but are largely independent of the other hemoglobin-associated variants in this region (maximum $r^2 \approx 0.2$ with rs1211375). Conditional regression analyses performed in a step-wise manner with WHI data (n = 8,087) suggest that this region might contain several independent hemoglobin association signals, tagged by the common rs13335497 and rs11863726 variants, in addition to the lower frequency *ITFG3* c.1600G>A (p.Asp534Asn) variant (Table S3). In a smaller subset of 1,800 WHI African American participants who underwent direct genotyping of rs144091859, we confirmed that the minor allele of *ITFG3* c.1600G>A (p.Asp534Asn) was associated with 0.248 ± 0.130 g/dl lower hemoglobin (p = 0.05).

Additional Loci for Hemoglobin and Hematocrit

Additional SNPs associated with red blood cell phenotypes in African Americans at nominally significant p values (range of 10^{-5}–10^{-7}) confirmed previous GWAS results in other ethnicities (Tables S2 and S4). Within *ABO* (MIM 110300) on chromosome 9, several common coding SNPs that tag the B blood group antigen (such SNPs include rs8176746, which confers a p.Met266Leu substitution in exon 7) were associated with higher hemoglobin in African Americans. The association between SNPs tagging the B blood group antigen and higher hemoglobin was recently reported in a Japanese GWAS.[4] A common rs3772219 missense variant (c.1021T>G [p.Leu341Val]; RefSeq NM_001128616.1) of *ARHGEF3* (MIM 612115), which encodes rho guanine-nucleotide exchange factor 3 (RhoGEF3), was associated with higher hematocrit. A distinct set of intronic *ARHGEF3* polymorphisms (tagged by rs12485738) has been associated with platelet count and mean platelet volume in European Americans,[9,55] and the *ARHGEF3* gene product was recently shown to be involved in iron uptake by erythroid cells.[56]

Table 3. Genome-wide-Significant SNPs for Blood Cell Traits

Variant ID	Rs Number	Chr.	Position (build 37)	Gene	Effect Allele	Other allele	EAF	Beta	SE	p Value	Rsq	Function	Amino Acids	Condel	Trait
snp.684276	rs334	11	5,248,232	*HBB*	T	A	0.0499	−0.688	0.105	5.70×10^{-11}	0.754	missense	p.Glu6Val	deleterious	hematocrit
snp.929281	rs9924561	16	314,780	*ITFG3*	G	T	0.0660	−0.3885	0.038	6.60×10^{-24}	0.535	intron	none	NA	hemoglobin
snp.929229	rs13335497	16	310,005	*ITFG3*	A	G	0.0685	−0.3435	0.036	1.80×10^{-21}	0.595	synonymous	none	NA	hemoglobin
snp.929121	rs11863726	16	230,578	*HBQ1*	G	A	0.2711	−0.1298	0.02	8.90×10^{-11}	0.518	synonymous	none	NA	hemoglobin
snp.177048	rs35837297	2	1,365,94439	*LCT*	G	A	0.0225	0.0836	0.0115	3.00×10^{-13}	0.982	missense	p.Ser101Gly	neutral	WBC
snp.177015	rs35940156	2	1,365,75300	*LCT*	A	G	0.0225	0.083	0.0115	4.30×10^{-13}	0.981	missense	p.Val440Ile	neutral	WBC
snp.41127	rs17292650	1	4,380,3807	*MPL*	T	G	0.0434	12.887	1.9089	1.50×10^{-11}	0.9274	missense	p.Lys39Asn	deleterious	platelet
snp.414923	rs513349	6	3,354,1719	*BAK1*	A	G	0.3673	−4.907	0.7783	2.90×10^{-10}	0.9882	intron	none	NA	platelet

Abbreviations are as follows: Chr. = chromosome; EAF = effect allele frequency; SE = standard error; Rsq = imputation quality score; NA = not applicable.

WBC

Associations of Nonsynonymous Variants of LCT *with higher WBC*

Several SNPs on chromosome 2q21 were significantly associated with a higher WBC (Table 3; see also Table S5). Most of these SNPs lie within a ~350 kb region containing *R3HDM1*, *MCM6* (MIM 601806), *LCT* (MIM 603202), and *DARS* (MIM 603084) (Figure 1B). The most strongly associated ($p = 3 \times 10^{-13}$–4×10^{-13}) were two lower-frequency *LCT* nonsynonymous coding variants, rs35837297 (c.301A>G [p.Ser101Gly]; RefSeq NM_002299.2) and rs35940156 (c.1318G>A [p.Val440Ile]; RefSeq NM_002299.2). These two *LCT* missense variants are in almost perfect LD with one another ($r^2 > 0.99$) and were also in strong LD ($r^2 > 0.75$) with another nonsynonymous SNP significantly associated with WBC, *RAB3GAP1* (MIM 602536) rs76927619 (c.669G>T, RefSeq NM_012233.2), which is located more than 700 kb downstream of *LCT* and encodes a p.Leu223Phe substitution in the catalytic subunit of Rab3 GTPase-activating protein. Upon adjustment for either of the *LCT* missense variants, the association between *RAB3GAP1* rs76927619 and WBC was markedly attenuated ($p = 0.1$).

Ancestry-Adjusted Analyses of LCT*-Containing Chromosomal Region 2q21*

Because ethnic background is a major determinant of WBC, careful assessment of population stratification is warranted by the presence of complex patterns of genetic variation on chromosomal region 2q21, which has undergone strong recent selective pressure in Europeans and Africans.[57] As expected, global African ancestry was strongly associated with *lower* WBC. In contrast, local African ancestry in the *LCT* region showed a weaker, but statistically significant, association with *higher* WBC ($p = 0.001$). We repeated the *LCT* missense SNP-WBC association analyses in the 8,095 WHI participants and adjusted for global ancestry, the first ten principal components derived from GWAS data, or local ancestry, as well as for four U.S. geographic regions (Table S6). The associations between WBC and the rare missense *LCT* variants were robust to all of these covariate adjustments, as well as to adjustment for ethnicity-specific *LCT* promoter SNPs (rs4988235, rs3806502, rs3087343, and rs749017) that have been associated with the lactase-persistence phenotype (MIM 223100).[57] When stratified by the estimated local number of European versus African chromosomes, the *LCT* rs35837297-WBC association in WHI was present among 4,841 African Americans who carry two African chromosomes ($\beta = 0.083 \pm 0.018$; $p = 3.7 \times 10^{-6}$), but not among the 3,462 African American who carry at least one European *LCT* chromosome ($\beta = 0.012 \pm 0.040$; $p = 0.76$).

Assessment of Fine-Scale Population Structure due to LCT *Missense Variants*

Recently, Mathieson and McVean[54] raised a concern that existing approaches (such as PC adjustment) might not adequately control for population stratification in rare-variant association studies. This argument assumes that

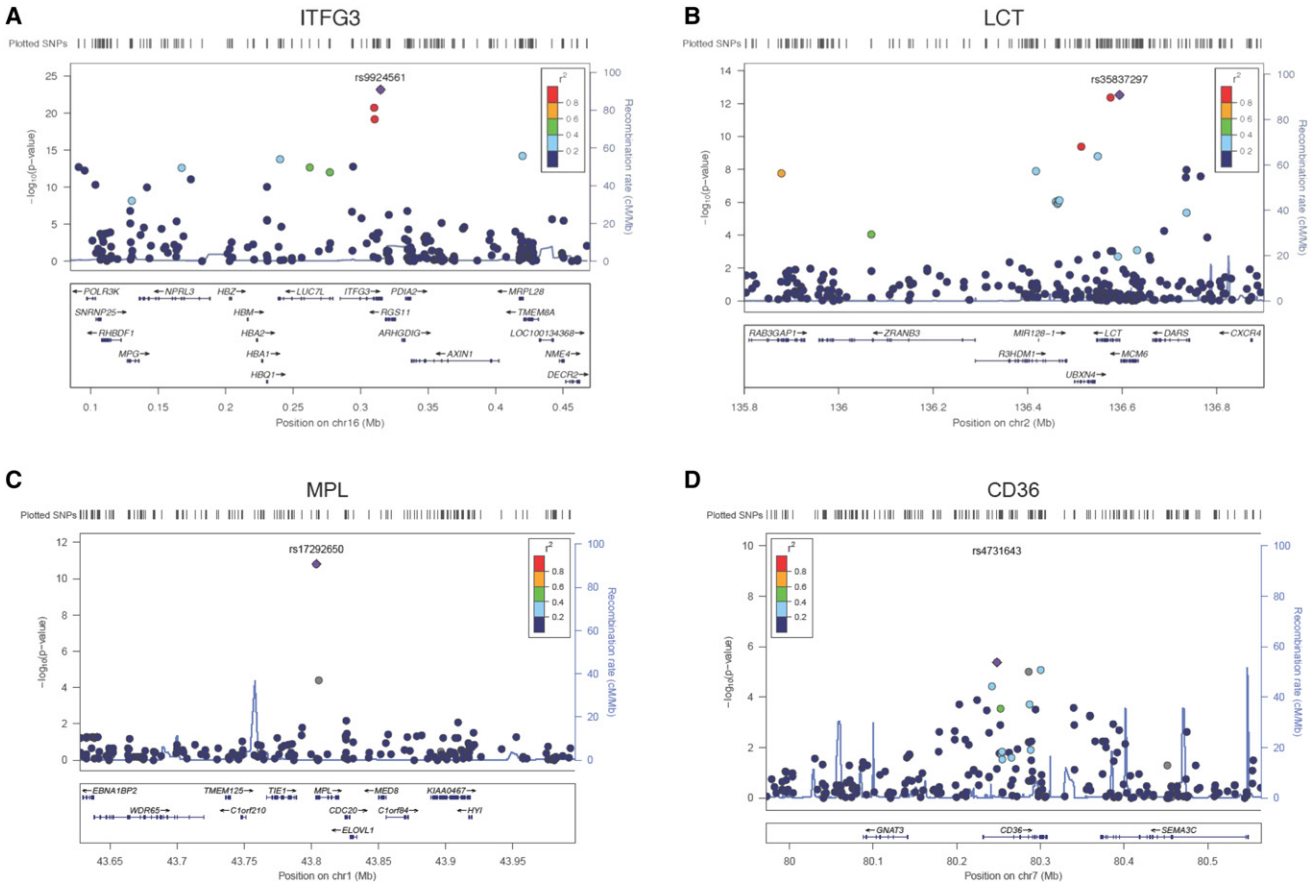

Figure 1. Regional Plots of Genetic Loci Associated with Blood Cell Traits
Shown are regional association plots generated with LocusZoom for (A) the chromosome 16 α-globin locus and hemoglobin; (B) the chromosome 2 *LCT* region and WBC; (C) *MPL* on chromosome 1 and platelet count; and (D) the CD36 locus on chromosome 7 and platelet count. The color of each single-nucleotide variant (SNV) indicates the level of pairwise linkage disequilibrium (LD) based on r-squared relative to the lead SNV in the region. r-squared values were calculated from 1000 Genomes YRI. SNVs with missing LD information are shown in gray. Included in these plots are either SNVs that were imputed from exome sequence data or any regional SNVs that were genotyped with the Affymetrix 6.0 array.

the phenotype distribution is substantially shifted within a "small, sharply defined [geographic] region"[54]. In such a situation, any rare variants that have higher frequencies in this geographic region will appear to be associated with the phenotype. If this were the case, we would expect additional (noncausal) rare variants to show similar geographic clustering; in other words, we would expect an excess of unlinked rare variants showing strong LD with the *LCT* missense variants. When we compared the *LCT* missense variant rs35940156 to 58 control SNPs selected on the basis of their having the same minor-allele frequencies as rs35940156 in the 1000 Genome's CEU and YRI populations, we found no evidence that it was associated with an excess of unlinked markers across the genome. The p value distribution when rs35940156 was used as the stratum definition is not significantly different from the distribution when the control SNPs defined the stratum (Figure S3). The lower extremes of the p value distributions for association with unlinked markers are also indistinguishable between the *LCT* and the control SNPs; for example, the 0.01% quantile ranges from 4.3×10^{-6}–5.1×10^{-5} and is 2.0×10^{-5} for the *LCT*-defined phenotype. These results argue against the notion that the observed *LCT*-WBC associations represent an artifact of a geographically isolated population stratum in which WBC is substantially different but is due to variants unlinked to *LCT*.

Validation and Functional Genomic Analysis of the Association between the LCT *Missense Variant and WBC*

The imputation quality for the *LCT* missense variants was excellent (Rsq > 0.98). Nonetheless, we performed validation of this finding by genotyping rs35940156 in a subset of 5,229 African American participants (624 from WHI, 2,795 from ARIC, 824 from CARDIA, and 1,756 from JHS). The minor allele was associated with a 0.168 ± 0.034 increase in the natural log unit in WBC ($p = 6.3 \times 10^{-7}$), confirming the association. We also performed genotyping in 1,029 independent African American samples (305 from WHI and 724 from the Cardiovascular Health Study), further validating this WBC association

($\beta = 0.102 \pm 0.043$; $p = 0.019$). The overall p value combining discovery and validation samples was 2.3×10^{-14}. Analysis of the *LCT* missense SNPs in the three cohorts (ARIC, CARDIA, JHS) with data available on WBC subtypes (n = 5,105) showed the associations with neutrophil and monocyte counts were strongest (Table S7).

Using HapMap and 1000 Genomes YRI data for further examination of the *LCT* region revealed extensive LD between the WBC-associated missense variants and numerous other SNPs in the chromosomal region 2q21. Interestingly, the extended region of LD spans ~1 Mb in Africans and includes *CXCR4* (MIM 162643), which encodes a leukocyte chemokine receptor involved in hematopoietic stem cell migration and leukocyte recruitment.[58] To further assess the functional significance of the *LCT* variants associated with WBC (particularly whether these SNPs have any regulatory effect on expression of *CXCR4*), we queried these SNPs and correlated proxies ($r^2 > 0.8$) against a database of >30 gene eQTL data sets, as well as RNA-sequencing eQTLs and SCAN. No significant *cis*-eQTLs were identified with the *LCT* missense variants or their correlated proxies. In summary, we have identified an association between a low-frequency African-ancestry-specific missense variant of *LCT* and higher WBC, but the biologic mechanism underlying this genetic association requires further study.

Platelet Count

Association of African-Specific MPL *Lys39Asn Missense Variant with Higher Platelet Count*

The strongest association signal for platelet count was rs17292650 (MAF = 0.04, $p = 1.5 \times 10^{-11}$). This African-ancestry-specific missense variant (c.117G>T [p.Lys39Asn]; RefSeq NM_005373.2) is located in exon 2 of *MPL* (MIM 159530), which encodes the thrombopoietin receptor on platelets and megakaryocytes (Table 3, Figure 1C). The c.117G>T [p.Lys39Asn] *MPL* variant has not been identified in prior GWASs, although it was previously discovered through a candidate-gene sequencing study of African Americans with essential thrombocytosis.[59] In a subset of 1,872 WHI African American participants who underwent genotyping of rs17292650, we confirmed that the minor allele of c.117G>T was associated with a 25.5 ± 5.31 higher platelet count ($p = 1.7 \times 10^{-6}$). A rare (MAF = 0.005) nonsynonymous variant rs141311765 in exon 5 of *MPL* (c.754T>C [p.Tyr252His]; RefSeq NM_005373.2) was independently associated with higher platelet count ($p = 4 \times 10^{-5}$).

Additional Suggestive Loci for Platelet Count

Variants within *CD109* (MIM 608859), *CD36* (MIM 173510), and *ITGB3* (MIM 173470), which encode several platelet surface glycoproteins involved in platelet adhesion and aggregation as well as platelet alloantigen formation, had p values for association with platelet count in the range of 10^{-5}–10^{-7}. (Table S8) GWAS recently showed that rs17154155, a common intronic variant of *CD36*, is associated with higher platelet count in African Americans.[10] In the current exome-imputed data set, two lower-frequency coding variants of *CD36* were associated with *lower* platelet count: these variants were rs3211938 (c.975T>G [p.Tyr325*]; RefSeq NM_000072.3) and rs3211862 (c.158_159insG [p.Asn53Lysfs*4]; RefSeq NM 000072.3) (Figure 1D). In conditional regression analyses, the platelet-count associations for these three *CD36* variants (intronic, nonsense, and frameshift) all appeared to be conditionally independent of one another (Table S9).

Discussion

We performed exome sequencing in a reference panel of 761 African Americans and imputed these samples into a much larger target sample of more than 13,000 from the original African American cohorts. We confirmed the association of a lower-frequency *MPL* receptor missense variant, c.117G>T (p.Lys39Asn), with higher platelet count in African Americans. We also identified a genome-wide association between two low-frequency *LCT* nonsynonymous coding variants (c.301A>G [p.Ser101Gly] and c.1318G>A [p.Val440Ile]) and higher WBC in African Americans. Lastly, we identified associations of low-frequency coding variants at existing blood-cell-associated genetic loci: *MPL* c.754T>C (p.Tyr252His) was associated with higher platelet count; *CD36* c.975T>G (p.Tyr325*) was associated with lower platelet count; and several missense variants at the α-globin locus were associated with lower hemoglobin. Together, these results demonstrate the utility of an "exome imputation" approach for identifying low-frequency coding variants associated with complex blood cell phenotypes in large population-based samples. The implications of these findings for particular blood cell phenotypes and related disorders are discussed below.

Red Blood Cell Disorders

Using a cardiovascular gene-based-tag SNP genotyping array in the population-based CARe consortium (which includes the cohorts JHS, ARIC, and CARDIA), Lo et al. reported a common variant that was located within the α-globin locus on chromosomal region 16p13 (*LUC7L* rs1211375) and was associated with lower hemoglobin, mean corpuscular hemoglobin, and mean corpuscular volume in African Americans.[8] This SNP might tag an African α-globin copy-number polymorphism (CNVR6569) that is absent in Europeans,[8] consistent with the known worldwide distribution of α-thalassemia copy-number variants.[60] We find evidence for additional genetic complexity and allelic heterogeneity at the α-globin locus at chromosomeal region 16p13 in African Americans; such evidence includes the association of hemoglobin with rs13335497 (c.423G>A, p.Ser141 =), a common synonymous variant of *ITFG3*, which encodes integrin-α FG-GAP repeat-containing-3. The function of the *ITFG3* gene product is unknown, although it is expressed in an erythroleukemia cell line[61]

and other common polymorphisms of this gene have been associated with red blood cell indices in European and Japanese GWASs.[2,4] Several of the lower-frequency nonsynonymous hemoglobin-associated variants (*HBQ1* c.155G>C [p.Gly52Ala], *HBM* c.45C>A [p.Asp15Glu], and *ITFG3* c.1600G>A [p.Asp534Asn]) are predicted to have deleterious effects on protein function. Nonetheless, the functions of the embryonic theta-globin and mu-globin genes are poorly understood; these genes are expressed at low levels in adult erythroid cells but are not translated into detectable globin protein.[62,63] Sequence analysis of the 1000 Genomes data has revealed that several common and rare structural variants at the α-globin locus segregate in African populations (R. Handsaker and S. McCarroll, personal communication), consistent with the clinical observation that many different DNA rearrangements can cause α-thalassemia. Thus, it is possible that the independent single-nucleotide variants associated with lower hemoglobin at the α-globin locus in our data set in fact capture these different structural variants. Alternatively, one or more of these single nucleotide variants could directly disrupt normal α-globin gene expression, as has been previously reported.[64] Establishing whether any of these African American hemoglobin-associated single-nucleotide variants are themselves "causal" or are proxies for additional α-thalassemia-related copy-number variants will require additional functional and population-genetics study.

WBC

The WBC-associated *LCT* missense variants rs35837297 (c.301A>G; p.Ser101Gly) and rs35940156 (c.1318G>A; p.Val440Ile) are distinct from the cluster of functional promoter variants that cause lactase persistence. The lactase-persistence SNPs are located ~14 kb upstream of *LCT* and have arisen on different population-specific haplotype backgrounds (−13910T in Europeans and −13907G, −13915G, and −14010C in East Africans) through convergent evolution as a result of positive selection.[57,65] The WBC association we observed for the low-frequency *LCT* coding variants was robust to adjustment of the European lactase-persistence variant −13910T, whereas the major East African lactase-persistence variant −14010C is absent in Nigerian Yoruba and African Americans. Hence, the observed WBC association at the *LCT* missense variants is unlikely to be due to population structure related to natural selection of the lactase-persistence phenotype. The WBC association observed for the low-frequency *LCT* coding variants was also robust to adjustment for genome-wide, locus-specific ancestry, as well as leading principal components. The *LCT*-WBC association was present only among African Americans who carried two African chromosomes in the *LCT* genomic region, further reducing the likelihood of confounding due to continental ancestry or gross-scale geographic structure. Finally, we found no evidence suggesting that the association was due to confounding of a small and undetected population stratum, as stipulated by a model in Mathieson and McVean.[54]

The mechanisms behind the association of *LCT* variants with higher WBC are not known. We reviewed the Mouse Phenome Database[66] and found *Lct* expression in the mouse is associated with CD8 count, suggesting a link to intestinal immunity and an effect of dietary exposure to lactose. It is possible that the WBC-associated *LCT* coding variants are proxies for a regulatory variant that controls expression or function of the nearby gene, *CXCR4*. Alternatively, it is possible that there exist functional *CXCR4* coding variant(s) that are in LD with the *LCT* missense variants and that these causal *CXCR4* variant(s) were not captured (or failed QC) during the original exome-sequencing study. Rare activating mutations in *CXCR4* are associated with WHIM syndrome (MIM 193670), a congenital immunodeficiency disorder characterized by infections and severe leukopenia.[67] From an evolutionary standpoint, the complex haplotype pattern and extensive LD in the *LCT-CXCR4* region among African populations suggests the possibility that additional selective sweeps might have occurred within Africa for reasons unrelated to milk consumption, perhaps due to host interactions between blood cells and pathogens.[57] Common variants of two other chemokine-related genes, *DARC* (MIM 613665) and *CXCL2* (MIM 139110), are associated with WBC.[5,6] Both the *DARC* null variant and low WBC are highly prevalent in Africa, where the *DARC* variant confers resistance to malarial infection (MIM 611162).

Platelet Count

Through a candidate-gene sequencing study, the African-specific *MPL* c.117G>T (p.Lys39Asn) variant was associated with reduced platelet protein MPL expression and mild thrombocytosis in heterozygotes and with more severe thrombocytosis in homozygotes.[59] We observed a similar allele-dose-dependent effect on platelet count in African Americans. The *MPL* c.117G>T (p. Lys39Asn) platelet-count association was not detected previously through European or African American GWASs[9,10] because the variant is low frequency, specific to African populations, and not well tagged on the Affy6.0 platform ("tag-r^2" = 0.076). In a recent small study of 245 black South Africans, the *MPL* c.117G>T (p.Lys39Asn) variant was nominally associated with platelet count ($p = 1.6 \times 10^{-5}$), thus confirming the association in an independent African-ancestry sample.[68]

MPL is the platelet and megakaryocyte receptor for thrombopoietin, or TPO (MIM 600044), an essential regulator of megakaryocyte differentiation and platelet production. Primary thrombocytosis, defined as a platelet count $>400 \times 10^9$/liter, can be due to rare *MPL* activating mutations, such as those encoding p.Pro106Leu[69] or p.Ser505-Asn[70] amino acid substitutions. Furthermore, acquired *MPL* activating mutations, such as those encoding p.Trp515Leu and p.Trp515Lys substitutions, are found in a subset of individuals with myeloproliferative disorders

involving high platelet count.[71,72] In the case of *MPL* c.117G>T (p.Lys39Asn), the mechanism relating *reduced* MPL receptor expression to *higher* platelet count might seem paradoxical. Reduced amounts of platelet MPL might shift the normal negative-feedback loop between circulating platelet count and thrombopoietin levels to higher TPO and increased platelet production.[73,74] In the current study, we identified a rare *MPL* variant associated with higher platelet count, c.754T>C (p.Tyr252His); this variant lies immediately adjacent to the minimum MPL binding domain of thrombopoietin (amino acids 206–251).[75] Together with recent GWAS findings for *THPO* in Europeans[9] and Japanese,[4] our results highlight the contribution of common and rare variants of megakaryopoiesis genes to inter-individual differences in platelet count in the general population.

We identified genetic variants of platelet surface glycoprotein receptors suggestively associated with platelet count. Such variants included those of *CD109*, *CD36*, and *ITGB3*. *ITGB3* rs61736876 (c.557C>T [p.Pro186Leu]; RefSeq NM_000212.2), which encodes a rare p.Pro186Leu substitution of platelet glycoprotein IIIa, was associated with lower platelet count. Glycoprotein IIIa p.Pro186Leu was previously identified in a person with Glanzmann thrombasthenia (MIM 273800), a congenital platelet disorder.[76] This amino acid substitution interferes with glycoprotein IIb and IIIa transport to the cell surface, binding to soluble fibrinogen, and α-v/β-3-mediated cell spreading on immobilized fibrinogen.[76] CD36, or platelet glycoprotein IV, is a multi-ligand scavenger receptor found on many cell types. *CD36* deficiency (MIM 608404; Nak^a-negative blood group) on platelets and monocytes as a result of null *CD36* mutations is common in African and Asian populations, where CD36 probably also serves as an erythrocyte malarial receptor.[77] Two of these *CD36* null mutations, rs3211938 (c.975T>G [p.Tyr325*]) and rs3211862 (c.158_159insG [p.Asn53Lysfs*4]), along with the higher-frequency intronic variant rs4731643, were each associated with lower platelet count in our African American population. Consistent with these results, rs3211938 has been associated with decreased CD36 expression on platelets and monocytes in African Americans.[78,79]

Large sample sizes will be required for robustly associating most rare coding variants with complex traits. Power calculations are shown for each variant in Tables S2, S4, S5, and S8. Even with an effective sample size of ~13,000, power to detect an association with platelet count is only ~10% for the functional variants *ITGB3* rs61736876 (c.557C>T [p.Pro186Leu]; MAF 0.007) and *CD36* rs3211862 (c.158_159insG [p.Asn53Lysfs*4]; MAF 0.02), and the effect size is in the range of ~15,000 platelets/ul (equivalent to approximately 0.25 SD unit). Therefore, even larger African American samples may be required for confirmation of associations that did not meet the threshold of genome-wide significance in the current study.

As the cost of whole-genome sequencing continues to decline, it might soon become feasible to directly genotype rare variants across the genome in tens of thousands of samples. However, because the vast majority of the variants captured through sequencing are exceedingly rare and outside the protein-coding region, it is not clear how such data would be analyzed for associations with complex traits. For variants of intermediate frequency (e.g., 0.1% < MAF < 5%), we have shown that high-coverage sequencing for variant discovery followed by imputation into large GWAS cohorts is a successful and cost-effective approach for testing associations with complex traits. An alternative approach involves a combination of low-coverage exome sequencing in all samples and imputation with 1000 Genomes Project reference panels.[80,81] Similar to our approach, low-coverage exome sequencing combined with 1000 Genomes imputation has the capability to capture common and intermediate frequency variation (1%–5% MAF) with increased power and cost efficiency in comparison to conventional GWAS.[80] Nonetheless, researchers have not yet used the low-coverage sequencing approach to discover associations with complex traits, and our high-coverage sequencing approach might capture additional imputable lower-frequency coding variants (in the range of 0.1% to 1% MAF) that would not be discovered through low-coverage exome sequencing. Moreover, because many rare variants are population specific, imputation reference panels derived from the same target population might be important for detecting lower-frequency alleles and might be less vulnerable to population stratification.

By sequencing the entire exomes or protein-coding regions of the genome in well-phenotyped population-based samples, the NHLBI GO Exome Sequencing Project has provided a large exome data set for identifying rare coding variation underlying a variety of complex traits related to heart, lung, and blood diseases. The ability to impute low-frequency coding variants with reasonable accuracy into larger data sets enables the inclusion of a greater number of causal coding-sequence variants that would otherwise not be captured by standard GWASs. Exome imputation should permit increased power to detect associations with low-frequency variants, greater characterization of allelic heterogeneity at previously known susceptibility loci, and fine mapping of causal variants.[82]

Perhaps not surprisingly, all of the newly identified coding variants explain <0.5% of the variance for any given trait. For example, the *LCT* missense variant(s) explain only 0.2% of WBC variance. Therefore, these rare coding variants do not appear to contribute substantively to overall heritability, which has been estimated in the range of 50% for most blood cell traits.[6,83] In prior published GWASs of blood cell traits in individuals of European descent, the proportion of variance explained by common SNPs has been ~1% for hemoglobin and hematocrit,[2] 5% for platelet count,[9] and ~7% for WBC,[5] whereas

the *DARC* null variant explains ~15%–20% of the WBC variance in African Americans.[1,6] Therefore, by comparison, the newly identified low-frequency coding variants explain a very small proportion of variance for each of the traits examined. Nonetheless, the evidence we found for allelic heterogeneity at several GWAS loci would not have been apparent without extending these analyses to include rare coding variants. Further studies involving larger samples will be required if we are to assess in a more comprehensive manner the hypothesis set forth by Dickson et al.,[11] who suggested that multiple rare variants account for some of the signals for common, complex traits reported in GWASs.

In summary, imputation of lower-frequency coding variants identified by whole-exome sequencing into large, population-based cohorts via population-specific reference panels offers the opportunity to identify variants that are associated with complex traits such as blood cell counts but that have not been previously targeted by genome-wide genotyping arrays. Studying the genetics of these quantitative blood traits in diverse populations might reveal new biologic pathways that ultimately not only contribute to our understanding of hematopoiesis but also increase our understanding of the association of blood counts with cardiovascular, inflammatory, and other systemic diseases.

Supplemental Data

Supplemental Data include Supplemental Acknowledgments, three figures, and nine tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

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). CARe was supported by a contract from NHLBI (HHSN268200960009C) to create a phenotype and genotype database for dissemination to the biomedical research community. Eight parent studies contributed phenotypic data and DNA samples through the Broad Institute (N01-HC-65226): the Atherosclerosis Risk in Communities study (ARIC), the Cleveland Family Study (CFS), the Coronary Artery Risk Development in Young Adults study (CARDIA), the Jackson Heart Study (JHS), the Multi-Ethnic Study of Atherosclerosis (MESA) study, the Cardiovascular Health Study (CHS), the Framingham Heart Study (FHS), and the Sleep Heart Health Study (SHHS). Support for CARe also came from the individual research institutions, investigators, field staff, and study participants. This research was supported in part by grants from the National Human Genome Research Institute to Y.L. (R01HG006292 and R01HG006703). This research was also supported in part by the Intramural Research Program of the National Institutes of Health, National Institute on Aging (Z01-AG000932-04), and this study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health (see Web Resources).

Received: May 4, 2012
Revised: June 12, 2012
Accepted: August 27, 2012
Published online: October 25, 2012

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes Project, http://www.1000genomes.org/
Biowulf Linux cluster, http://biowulf.nih.gov
Condel, http://bg.upf.edu/condel/home
Eigenstrat, http://genepath.med.harvard.edu/~reich/Software.htm
eQTL Resources, http://eqtl.uchicago.edu
Exome chip, http://genome.sph.umich.edu/wiki/Exome_Chip_Design
MaCH, http://www.sph.umich.edu/csg/yli/mach/
Minimac, http://genome.sph.umich.edu/wiki/Minimac
Mouse Phenome Database, http://phenome.jax.org
NHLBI Exome Variant Server (ESP), http://evs.gs.washington.edu/EVS
NHLBI GO Exome Sequencing Project, http://esp.gs.washington.edu/drupal
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org
SCAN: SNP and CNV Annotation Database, http://www.scandb.org

References

1. Nalls, M.A., Wilson, J.G., Patterson, N.J., Tandon, A., Zmuda, J.M., Huntsman, S., Garcia, M., Hu, D., Li, R., Beamer, B.A., et al. (2008). Admixture mapping of white cell count: genetic locus responsible for lower white blood cell count in the Health ABC and Jackson Heart studies. Am. J. Hum. Genet. *82*, 81–87.
2. Ganesh, S.K., Zakai, N.A., van Rooij, F.J., Soranzo, N., Smith, A.V., Nalls, M.A., Chen, M.H., Kottgen, A., Glazer, N.L., Dehghan, A., et al. (2009). Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nat. Genet. *41*, 1191–1198.
3. Soranzo, N., Spector, T.D., Mangino, M., Kühnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., et al. (2009). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat. Genet. *41*, 1182–1190.
4. Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., Nakamura, Y., and Kamatani, N. (2010). Genome-wide association study of hematological and biochemical traits in a Japanese population. Nat. Genet. *42*, 210–215.
5. Nalls, M.A., Couper, D.J., Tanaka, T., van Rooij, F.J., Chen, M.H., Smith, A.V., Toniolo, D., Zakai, N.A., Yang, Q., Greinacher, A., et al. (2011). Multiple loci are associated with white blood cell phenotypes. PLoS Genet. *7*, e1002113.
6. Reiner, A.P., Lettre, G., Nalls, M.A., Ganesh, S.K., Mathias, R., Austin, M.A., Dean, E., Arepalli, S., Britton, A., Chen, Z., 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.
7. Okada, Y., Hirota, T., Kamatani, Y., Takahashi, A., Ohmiya, H., Kumasaka, N., Higasa, K., Yamaguchi-Kabata, Y., Hosono, N., Nalls, M.A., et al. (2011). Identification of nine novel loci associated with white blood cell subtypes in a Japanese population. PLoS Genet. *7*, e1002067.
8. Lo, K.S., Wilson, J.G., Lange, L.A., Folsom, A.R., Galarneau, G., Ganesh, S.K., Grant, S.F., Keating, B.J., McCarroll, S.A., Mohler, E.R., 3rd., et al. (2011). Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans. Hum. Genet. *129*, 307–317.
9. Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A.H., Labrune, Y., et al. (2011). New gene functions in megakaryopoiesis and platelet formation. Nature *480*, 201–208.
10. Qayyum, R., Snively, B.M., Ziv, E., Nalls, M., Liu, Y., Tang, W., Yanek, L.R., Lange, L., Evans, M., Ganesh, S., et al. (2012). A meta-analysis and genome-wide association study of platelet count and mean platelet volume in African Americans. PLoS Genet. *8*, e1002491.
11. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D.B. (2010). Rare variants create synthetic genome-wide associations. PLoS Biol. *8*, e1000294.
12. 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.
13. 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.
14. The ARIC investigators. (1989). The Atherosclerosis Risk in Communities (ARIC) Study: Design and objectives. Am. J. Epidemiol. *129*, 687–702.
15. Friedman, G.D., Cutter, G.R., Donahue, R.P., Hughes, G.H., Hulley, S.B., Jacobs, D.R., Jr., Liu, K., and Savage, P.J. (1988). CARDIA: Study design, recruitment, and some characteristics of the examined subjects. J. Clin. Epidemiol. *41*, 1105–1116.
16. Taylor, H.A., Jr., Wilson, J.G., Jones, D.W., Sarpong, D.F., Srinivasan, A., et al. (2005). Toward resolution of cardiovascular health disparities in African Americans: Design and methods of the Jackson Heart Study. Ethn. Dis. *15*(*Suppl. 6*), 4–17.
17. Williams, W.J., and Schneider, A.S. (1972). Examination of the peripheral blood. In Hematology, W.J. Williams, E. Beutler, A.J. Erslev, and R.W. Rundles, eds. (New York: McGraw-Hill), pp. 10–22.
18. Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T.M., Young, G., Fennell, T.J., Allen, A., Ambrogio, L., et al. (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol. *12*, R1.
19. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
20. 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.
21. Li, Y., Sidore, C., Kang, H.M., Boehnke, M., and Abecasis, G.R. (2011). Low-coverage sequencing: Implications for design of complex trait association studies. Genome Res. *21*, 940–951.
22. Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. *18*, 1851–1858.
23. Lettre, G., Palmer, C.D., Young, T., Ejebe, K.G., Allayee, H., Benjamin, E.J., Bennett, F., Bowden, D.W., Chakravarti, A., Dreisbach, A., et al. (2011). Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project. PLoS Genet. *7*, e1001300.
24. 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.
25. 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.
26. 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.
27. 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.
28. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191.
29. Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. *5*, e1000519.
30. 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.
31. González-Pérez, A., and López-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am. J. Hum. Genet. *88*, 440–449.
32. Gamazon, E.R., Zhang, W., Konkashbaev, A., Duan, S., Kistner, E.O., Nicolae, D.L., Dolan, M.E., and Cox, N.J. (2010). SCAN: SNP and copy number annotation. Bioinformatics *26*, 259–262.
33. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–772.
34. Göring, H.H., Curran, J.E., Johnson, M.P., Dyer, T.D., Charlesworth, J., Cole, S.A., Jowett, J.B., Abraham, L.J., Rainwater, D.L., Comuzzie, A.G., et al. (2007). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat. Genet. *39*, 1208–1216.
35. Idaghdour, Y., Czika, W., Shianna, K.V., Lee, S.H., Visscher, P.M., Martin, H.C., Miclaus, K., Jadallah, S.J., Goldstein,

D.B., Wolfinger, R.D., and Gibson, G. (2010). Geographical genomics of human leukocyte gene expression variation in southern Morocco. Nat. Genet. *42*, 62–67.
36. Heap, G.A., Trynka, G., Jansen, R.C., Bruinenberg, M., Swertz, M.A., Dinesen, L.C., Hunt, K.A., Wijmenga, C., Vanheel, D.A., and Franke, L. (2009). Complex nature of SNP genotype effects on gene expression in primary human leucocytes. BMC Med. Genomics *2*, 1.
37. Dixon, A.L., Liang, L., Moffatt, M.F., Chen, W., Heath, S., Wong, K.C., Taylor, J., Burnett, E., Gut, I., Farrall, M., et al. (2007). A genome-wide association study of global gene expression. Nat. Genet. *39*, 1202–1207.
38. Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., et al. (2007). Population genomics of human gene expression. Nat. Genet. *39*, 1217–1224.
39. Kwan, T., Benovoy, D., Dias, C., Gurd, S., Provencher, C., Beaulieu, P., Hudson, T.J., Sladek, R., and Majewski, J. (2008). Genome-wide analysis of transcript isoform variation in humans. Nat. Genet. *40*, 225–231.
40. Zeller, T., Wild, P., Szymczak, S., Rotival, M., Schillert, A., Castagne, R., Maouche, S., Germain, M., Lackner, K., Rossmann, H., et al. (2010). Genetics and beyond—The transcriptome of human monocytes and disease susceptibility. PLoS ONE *5*, e10693.
41. Heinzen, E.L., Ge, D., Cronin, K.D., Maia, J.M., Shianna, K.V., Gabriel, W.N., Welsh-Bohmer, K.A., Hulette, C.M., Denny, T.N., and Goldstein, D.B. (2008). Tissue-specific genetic control of splicing: implications for the study of complex traits. PLoS Biol. *6*, e1.
42. Greenawalt, D.M., Dobrin, R., Chudin, E., Hatoum, I.J., Suver, C., Beaulaurier, J., Zhang, B., Castro, V., Zhu, J., Sieberts, S.K., et al. (2011). A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. Genome Res. *21*, 1008–1016.
43. Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. Nature *452*, 423–428.
44. Inouye, M., Silander, K., Hamalainen, E., Salomaa, V., Harald, K., Jousilahti, P., Männistö, S., Eriksson, J.G., Saarela, J., Ripatti, S., et al. (2010). An immune response network associated with blood lipid levels. PLoS Genet. *6*, e1001113.
45. Kompass, K.S., and Witte, J.S. (2011). Co-regulatory expression quantitative trait loci mapping: Method and application to endometrial cancer. BMC Med. Genomics *4*, 6.
46. Webster, J.A., Gibbs, J.R., Clarke, J., Ray, M., Zhang, W., Holmans, P., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., et al.; NACC-Neuropathology Group. (2009). Genetic control of human brain transcript expression in Alzheimer disease. Am. J. Hum. Genet. *84*, 445–458.
47. Colantuoni, C., Lipska, B.K., Ye, T., Hyde, T.M., Tao, R., Leek, J.T., Colantuoni, E.A., Elkahloun, A.G., Herman, M.M., Weinberger, D.R., and Kleinman, J.E. (2011). Temporal dynamics and genetic control of transcription in the human prefrontal cortex. Nature *478*, 519–523.
48. Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., et al. (2008). Mapping the genetic architecture of gene expression in human liver. PLoS Biol. *6*, e107.
49. Innocenti, F., Cooper, G.M., Stanaway, I.B., Gamazon, E.R., Smith, J.D., Mirkov, S., Ramirez, J., Liu, W., Lin, Y.S., Moloney, C., et al. (2011). Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. PLoS Genet. *7*, e1002078.
50. Schröder, A., Klein, K., Winter, S., Schwab, M., Bonin, M., Zell, A., and Zanger, U.M. (2011). Genomics of ADME gene expression: mapping expression quantitative trait loci relevant for absorption, distribution, metabolism and excretion of drugs in human liver. Pharm J. Published online October 18, 2011. http://dx.doi.org/10.1038/tpj.2011.44.
51. Grundberg, E., Kwan, T., Ge, B., Lam, K.C., Koka, V., Kindmark, A., Mallmin, H., Dias, J., Verlaan, D.J., Ouimet, M., et al. (2009). Population genomics in a disease targeted primary cell model. Genome Res. *19*, 1942–1952.
52. Ding, J., Gudjonsson, J.E., Liang, L., Stuart, P.E., Li, Y., Chen, W., Weichenthal, M., Ellinghaus, E., Franke, A., Cookson, W., et al. (2010). Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. Am. J. Hum. Genet. *87*, 779–789.
53. Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., et al. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. Science *325*, 1246–1250.
54. Mathieson, I., and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nat. Genet. *44*, 243–246.
55. Meisinger, C., Prokisch, H., Gieger, C., Soranzo, N., Mehta, D., Rosskopf, D., Lichtner, P., Klopp, N., Stephens, J., Watkins, N.A., et al. (2009). A genome-wide association study identifies three loci associated with mean platelet volume. Am. J. Hum. Genet. *84*, 66–71.
56. Serbanovic-Canic, J., Cvejic, A., Soranzo, N., Stemple, D.L., Ouwehand, W.H., and Freson, K. (2011). Silencing of RhoA nucleotide exchange factor, ARHGEF3, reveals its unexpected role in iron uptake. Blood *118*, 4967–4976.
57. Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. *39*, 31–40.
58. Zou, Y.R., Kottmann, A.H., Kuroda, M., Taniuchi, I., and Littman, D.R. (1998). Function of the chemokine receptor CXCR4 in haematopoiesis and in cerebellar development. Nature *393*, 595–599.
59. Moliterno, A.R., Williams, D.M., Gutierrez-Alamillo, L.I., Salvatori, R., Ingersoll, R.G., and Spivak, J.L. (2004). Mpl Baltimore: a thrombopoietin receptor polymorphism associated with thrombocytosis. Proc. Natl. Acad. Sci. USA *101*, 11444–11447.
60. Beutler, E., and West, C. (2005). Hematologic differences between African-Americans and whites: The roles of iron deficiency and alpha-thalassemia on hemoglobin levels and mean corpuscular volume. Blood *106*, 740–745.
61. Resing, K.A., Meyer-Arendt, K., Mendoza, A.M., Aveline-Wolf, L.D., Jonscher, K.R., Pierce, K.G., Old, W.M., Cheung, H.T., Russell, S., Wattawa, J.L., et al. (2004). Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. Anal. Chem. *76*, 3556–3568.
62. Albitar, M., Care, A., Peschle, C., and Liebhaber, S.A. (1992). Developmental switching of messenger RNA expression from the human alpha-globin cluster: fetal/adult pattern of theta-globin gene expression. Blood *80*, 1586–1591.

63. Goh, S.H., Lee, Y.T., Bhanu, N.V., Cam, M.C., Desper, R., Martin, B.M., Moharram, R., Gherman, R.B., and Miller, J.L. (2005). A newly discovered human alpha-globin gene. Blood *106*, 1466–1472.
64. De Gobbi, M., Viprakasit, V., Hughes, J.R., Fisher, C., Buckle, V.J., Ayyub, H., Gibbons, R.J., Vernimmen, D., Yoshinaga, Y., de Jong, P., et al. (2006). A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science *312*, 1215–1217.
65. Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., and Hirschhorn, J.N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. *74*, 1111–1120.
66. Maddatu, T.P., Grubb, S.C., Bult, C.J., and Bogue, M.A. (2012). Mouse Phenome Database (MPD). Nucleic Acids Res. *40*(Database issue), D887–D894.
67. Kawai, T., and Malech, H.L. (2009). WHIM syndrome: Congenital immune deficiency disease. Curr. Opin. Hematol. *16*, 20–26.
68. Ramsuran, V., Kulkarni, H., He, W., Mlisana, K., Wright, E.J., Werner, L., Castiblanco, J., Dhanda, R., Le, T., Dolan, M.J., et al. (2011). Duffy-null-associated low neutrophil counts influence HIV-1 susceptibility in high-risk South African black women. Clin. Infect. Dis. *52*, 1248–1256.
69. El-Harith el-HA, Roesl, C., Ballmaier, M., Germeshausen, M., Frye-Boukhriss, H., von Neuhoff, N., Becker, C., Nürnberg, G., Nürnberg, P., and Ahmed, M.A. (2009). Familial thrombocytosis caused by the novel germ-line mutation p.Pro106Leu in the MPL gene. Br. J. Haematol. *144*, 185–194.
70. Ding, J., Komatsu, H., Iida, S., Yano, H., Kusumoto, S., Inagaki, A., Mori, F., Ri, M., Ito, A., Wakita, A., et al. (2009). The Asn505 mutation of the c-MPL gene, which causes familial essential thrombocythemia, induces autonomous homodimerization of the c-Mpl protein due to strong amino acid polarity. Blood *114*, 3325–3328.
71. Pardanani, A.D., Levine, R.L., Lasho, T., Pikman, Y., Mesa, R.A., Wadleigh, M., Steensma, D.P., Elliott, M.A., Wolanskyj, A.P., Hogan, W.J., et al. (2006). MPL515 mutations in myeloproliferative and other myeloid disorders: a study of 1182 patients. Blood *108*, 3472–3476.
72. Pikman, Y., Lee, B.H., Mercher, T., McDowell, E., Ebert, B.L., Gozo, M., Cuker, A., Wernig, G., Moore, S., Galinsky, I., et al. (2006). MPLW515L is a novel somatic activating mutation in myelofibrosis with myeloid metaplasia. PLoS Med. *3*, e270.
73. Tiedt, R., Coers, J., Ziegler, S., Wiestner, A., Hao-Shen, H., Bornmann, C., Schenkel, J., Karakhanova, S., de Sauvage, F.J., Jackson, C.W., and Skoda, R.C. (2009). Pronounced thrombocytosis in transgenic mice expressing reduced levels of Mpl in platelets and terminally differentiated megakaryocytes. Blood *113*, 1768–1777.
74. Lannutti, B.J., Epp, A., Roy, J., Chen, J., and Josephson, N.C. (2009). Incomplete restoration of Mpl expression in the mpl-/- mouse produces partial correction of the stem cell-repopulating defect and paradoxical thrombocytosis. Blood *113*, 1778–1785.
75. Chen, W.M., Yu, B., Zhang, Q., and Xu, P. (2010). Identification of the residues in the extracellular domain of thrombopoietin receptor involved in the binding of thrombopoietin and a nuclear distribution protein (human NUDC). J. Biol. Chem. *285*, 26697–26709.
76. Morel-Kopp, M.C., Melchior, C., Chen, P., Ammerlaan, W., Lecompte, T., Kaplan, C., and Kieffer, N. (2001). A naturally occurring point mutation in the beta3 integrin MIDAS-like domain affects differently alphavbeta3 and alphaIIIbbeta3 receptor function. Thromb. Haemost. *86*, 1425–1434.
77. Fry, A.E., Ghansa, A., Small, K.S., Palma, A., Auburn, S., Diakite, M., Green, A., Campino, S., Teo, Y.Y., Clark, T.G., et al. (2009). Positive selection of a CD36 nonsense variant in sub-Saharan Africa, but no association with severe malaria phenotypes. Hum. Mol. Genet. *18*, 2683–2692.
78. Love-Gregory, L., Sherva, R., Schappe, T., Qi, J.S., McCrea, J., Klein, S., Connelly, M.A., and Abumrad, N.A. (2011). Common CD36 SNPs reduce protein expression and may contribute to a protective atherogenic profile. Hum. Mol. Genet. *20*, 193–201.
79. Ghosh, A., Murugesan, G., Chen, K., Zhang, L., Wang, Q., Febbraio, M., Anselmo, R.M., Marchant, K., Barnard, J., and Silverstein, R.L. (2011). Platelet CD36 surface expression levels affect functional responses to oxidized LDL and are associated with inheritance of specific genetic polymorphisms. Blood *117*, 6355–6366.
80. Pasaniuc, B., Rohland, N., McLaren, P.J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B.M., Daly, M.J., Sklar, P., et al. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat. Genet. *44*, 631–635.
81. Flannick, J., Korn, J.M., Fontanillas, P., Grant, G.B., Banks, E., Depristo, M.A., and Altshuler, D. (2012). Efficiency and power as a function of sequence coverage, SNP array density, and imputation. PLoS Comput. Biol. *8*, e1002604.
82. Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nat. Rev. Genet. *11*, 499–511.
83. Evans, D.M., Frazer, I.H., and Martin, N.G. (1999). Genetic and environmental causes of variation in basal levels of blood cells. Twin Res. *2*, 250–257.