Supplementary Materials for

Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes

Jacob A. Tennessen, Abigail W. Bigham, Timothy D. O'Connor, Wenqing Fu, Eimear E. Kenny, Simon Gravel, Sean McGee, Ron Do, Xiaoming Liu, Goo Jun, Hyun Min Kang, Daniel Jordan, Suzanne M. Leal, Stacey Gabriel, Mark J. Rieder, Goncalo Abecasis, David Altshuler, Deborah A. Nickerson, Eric Boerwinkle, Shamil Sunyaev, Carlos D. Bustamante, Michael J. Bamshad,* Joshua M. Akey,* Broad GO, Seattle GO, on behalf of the NHLBI Exome Sequencing Project

*To whom correspondence should be addressed. E-mail: akeyj@uw.edu (J.M.A.); mbamshad@u.washington.edu (M.J.B.)

Published 17 May 2012 on *Science* Express
DOI: 10.1126/science.1219240

This PDF file includes:

Supporting Online Material

Table of Contents

Study sample

The 2,440 exomes used in the analysis were generated from samples ascertained from 15 different cohorts.

Phenotypes associated with samples and cohort contributions:

Phenotype	Cohort	Number of samples contributed
Early onset myocardial infarction (n=1094)	Atherosclerosis Risk in Communities	258
	Cleveland Clinic GeneBank	3
	Framingham Heart Study	177
	Heart Attack Risk in Puget Sound	13
	Jackson Heart Study	112
	Massachusetts General Hospital-Premature Coronary Artery Disease	147
	Penn-Cath	34
	Translational Research Investigating Underlying Disparities in Acute Myocardial Infarction Patients' Health Status	77
	Women's Health Initiative	273
Cystic Fibrosis (n=91)	Early Pseudomonas Infection Control cystic fibrosis cohort	74
	University of North Carolina cystic fibrosis cohort	17
Chronic obstructive pulmonary disease (n=89)	Lung Health Study	88
	COPD Gene	1
Pulmonary hypertension (n=79)	Pulmonary Arterial Hypertension	79
LDL (n=393)	Atherosclerosis Risk in Communities	292
	Cardiovascular Health Study	24
	Framingham Heart Study	25
	Jackson Heart Study	52
Body mass index and type 2 diabetes (n=450)	Women's Health Initiative	450
Early-onset stroke (n=244)	Women's Health Initiative	244

Cohort descriptions

Women's Health Initiative (WHI): The WHI is a long-term national health study that has focused on strategies for preventing heart disease, breast and colorectal cancer, and osteoporotic fractures in postmenopausal women. The original WHI study included 161,808 postmenopausal women enrolled between 1993 and 1998. The Fred Hutchinson Cancer Research Center in Seattle, WA serves as the WHI Clinical Coordinating Center for data collection, management, and analysis of the WHI.

Framingham Heart Study (FHS): FHS is a study to identify the common factors or characteristics that contribute to cardiovascular disease (CVD) by following its development over a long period of time in a large group of participants who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke. In 1948, researchers recruited 5,209 men and women between the ages of 30 and 62 from the town of Framingham, Massachusetts, and began the first round of extensive physical examinations and lifestyle interviews that they would later analyze for common patterns related to CVD development. Since 1948, the subjects have returned to the study every two years for an examination consisting of a detailed medical history, physical examination, and laboratory tests. In 1971, the study enrolled a second-generation cohort, 5,124 of the original participants' adult children and their spouses, to participate in similar examinations. The second examination of the offspring cohort occurred eight years after the first examination, and subsequent examinations have occurred approximately every four years thereafter. In April 2002, the Study entered a new phase: the enrollment of a third generation of participants, the grandchildren of the original cohort. The first examination of the Third Generation Study was completed in July 2005 and involved 4,095 participants. Thus, the FHS has evolved into a prospective, community-based, three-generation family study. The FHS is a joint project of the National Heart, Lung and Blood Institute and Boston University.

Jackson Heart Study (JHS): JHS is a large, community-based, observational study whose participants were recruited from urban and rural areas of the three counties (Hinds, Madison, and Rankin) that make up the Jackson, Mississippi, metropolitan statistical area (MSA). Participants were enrolled from each of four recruitment pools: random, 17%; volunteer, 30%; currently enrolled in the Atherosclerosis Risk in Communities (ARIC) Study, 22%; and secondary family members, 31%. Recruitment was limited to non-institutionalized adult African Americans 35-84 years old, except in the family cohort where those 21 to 34 years of age were eligible. The final cohort of 5,301 participants includes 6.59% of all African American Jackson MSA residents aged 35-84 (N=76,426, US Census 2000). Major components of each exam include medical history, physical examination, blood/urine analyses and interview questions on areas such as: physical activity; stress, coping and spirituality; racism and discrimination; socioeconomic position; and access to health care. At 12-month intervals after the baseline clinic visit (Exam 1), participants are contacted by telephone to update information, confirm vital statistics, document interim medical events, hospitalizations, and functional status, and obtain additional sociocultural information. Questions about medical events, symptoms of cardiovascular disease, and functional status are

repeated annually. Ongoing cohort surveillance includes abstraction of medical records and death certificates for relevant International Classification of Diseases (ICD) codes and adjudication of nonfatal events and deaths.

Atherosclerosis Risk in Communities (ARIC): The ARIC Study, sponsored by the National Heart, Lung and Blood Institute (NHLBI), is a prospective epidemiologic study conducted in four U.S. communities. The four communities are Forsyth County, NC; Jackson, MS; the northwest suburbs of Minneapolis, MN; and Washington County, MD. ARIC is designed to investigate the etiology and natural history of atherosclerosis, the etiology of clinical atherosclerotic diseases, and variation in cardiovascular risk factors, medical care and disease by race, gender, location, and date.

ARIC includes two parts: the Cohort Component and the Community Surveillance Component. The Cohort Component began in 1987, and each ARIC field center randomly selected and recruited a cohort sample of approximately 4,000 individuals aged 45-64 from a defined population in their community. A total of 15,792 participants received an extensive examination, including medical, social, and demographic data. These participants were reexamined every three years with the first screen (baseline) occurring in 1987-1989, the second in 1990-1992, the third in 1993-1995, and the fourth and last exam was in 1996-1998. Follow-up occurs yearly by telephone to maintain contact with participants and to assess health status of the cohort.

Cardiovascular Health Study (CHS): The CHS is a study of risk factors for development and progression of coronary heart disease (CHD) and stroke in people aged 65 years and older. The objectives of the Cardiovascular Health Study are to: 1) quantify associations of conventional and hypothesized risk factors with CHD and stroke; 2) assess the associations of non-invasive measures of subclinical disease with the incidence of CHD and stroke; 3) quantify the associations of risk factors with subclinical disease; 4) characterize the natural history of CHD and stroke, and identify factors associated with clinical course; and 5) describe the prevalence and distributions of risk factors for CHD and stroke, non-invasive measures of subclinical disease, and clinical endpoints of CHD and stroke. The study originated in 1988 from the recommendations of an NHLBI workshop on the management of CHD in the elderly. This is the most extensive study undertaken by the NHLBI to study CVD exclusively in an elderly population. Initially funded for six years, it was renewed for a second six-year period in 1994, and recently was renewed for continued morbidity and mortality follow-up. The 5,888 study participants were recruited from four U.S. communities and have undergone extensive clinic examinations for evaluation of markers of subclinical cardiovascular disease. The original cohort totaled 5,201 participants. A new cohort was recruited in 1992. The 687 participants in the new cohort are predominately African-American and were recruited at three of the four field centers.

Cleveland Clinic GeneBank: The Cleveland Clinic GeneBank is a prospective cohort based study that enrolled patients undergoing elective diagnostic coronary angiography.

Heart Attack Risk in Puget Sound (HARPS): HARPS is a population-based case-

control study that enrolled cases with incident myocardial infarction (MI) presenting to a network of hospitals in Washington State. In HARPS, eligible cases were men <50 and women < 60 years of age.

Massachusetts General Hospital - Premature Coronary Artery Disease (MGH-PCAD): MGH-PCAD is a hospital-based case-control study that enrolled cases who were hospitalized with MI at MGH. In MGH-PCAD, eligible cases were men <50 and women < 60 years of age.

Penn-Cath: Penn-Cath is a catheterization-lab based cohort study from the University of Pennsylvania Medical Center and enrolled subjects at the time of cardiac catheterization and coronary angiography.

Translational Research Investigating Underlying Disparities in Acute Myocardial Infarction Patients' Health Status (TRIUMPH): TRIUMPH is an observational, multi-center prospective registry that enrolled subjects presenting with MI at participating medical centers.

Lung Health Study (LHS): The LHS is a randomized multicenter clinical trial with 5,887 participants carried out from October 1986 to April 1994, designed to test the effectiveness of smoking cessation and bronchodilator administration in smokers aged 35 to 60 with mild lung function impairment. Participants were randomly assigned to one of three groups: (1) usual care, who received no intervention; (2) smoking intervention with the inhaled bronchodilator ipratroprium bromide; or (3) smoking intervention with an inhaled placebo. The effect of intervention was evaluated by the rate of decline of forced expiratory volume in one second (FEV1).

Pulmonary Arterial Hypertension (PAH): The PAH cohort was comprised of idiopathic PAH (IPAH) and scleroderma-associated PAH (PAH-SSc) cases and healthy controls in the Hopkins SCCOR program. Pulmonary hypertension was defined in IPAH or PAH-SSc patients as a mean pulmonary artery pressure greater than 25 mm Hg proven by right heart catheterization. For patients with scleroderma, the presence of disease was defined as systemic sclerosis with diffuse or limited scleroderma meeting the American College of Rheumatology criteria.

Cystic Fibrosis (CF) cohorts (2): CF cases were ascertained from two separate cohorts. The first cohort consisted of 1,704 cases that represent participants in the Early *Pseudomonas* Infection Control (EPIC) Observational Study, a multicenter, longitudinal, prospective cohort of early lung disease in young CF patients. The second CF cohort consisted of 1,208 ΔF508 homozygotes who are at the extremes of lung disease severity ("severe", worst 25^{th} percentile of birth cohort vs. "mild", best 25^{th} percentile) based on ~22 measures of lung function for each patient (> 5 years of age) developed at the University of North Carolina.

Informed consent

All study participants in each of the component studies provided written informed consent for the use of their DNA in studies aimed at identifying genetic risk variants for disease and for broad data sharing. Institutional certification was obtained for each sample to allow deposition of phenotype and genotype data in dbGaP and BAM files in the short-read archive.

Exome sequencing

University of Washington exome sequencing / analysis

QC of sample DNA

Initial quality control (QC) performed on all samples included sample quantification (PicoGreen; Life Technologies, Grand Island, NY), confirmation of high-molecular weight DNA, test PCR amplification (four amplicons), and sex determination using a Taq-man assay (31). Prior to preparation for exome sequencing, all samples were genotyped (Illumina BeadXpress; Illumina, San Diego, CA) for 96 high frequency (30-50% minor allele frequency) exome-specific SNPs derived from the content found on large-scale (GWAS) genotyping chips from both Illumina and Affymetrix. Exome data at these variant sites were used to ensure sample tracking integrity through sample preparation and the sequencing pipeline. Samples failed QC if: the total mass, concentration, or integrity of DNA was low; low (<90%) genotype call rates were observed; sex-typing was inconsistent with the sample manifest. Following QC, all 3.5 ug of genomic DNA was reformatted into 96 well plates for library preparation and exome capture.

Library production and exome capture

All protocols for library construction and exome capture were automated on a Perkin-Elmer Janus II liquid handling robot or multi-channel pipettors, and performed in 96-well plate format. Samples were prepared by subjecting ~3.5 ug of genomic DNA to a series of shotgun library construction steps, including fragmentation through acoustic fragmentation (Covaris, Woburn, MA), end-polishing and A-tailing, ligation of sequencing adaptors, and PCR amplification. Sample shotgun libraries were captured for exome enrichment using one of three 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). Briefly, 1 µg of shotgun library was hybridized to biotinylated capture probes for 72 hours and recovered via streptavidin beads. Unbound DNA was washed away, and the captured DNA PCR amplified. Following capture, washing, and PCR, libraries were assessed again on the Bioanalyzer (Agilent, Santa Clara, CA) for concentration, molecular weight distribution, and the presence of PCR artifacts. The fragment size distributions of the libraries were highly consistent (typically 125 ± 15 bp).

Clustering and sequencing

Library concentration and flow-cell loading cluster densities were determined using a standardized qPCR protocol (Kapa Biosystems, Woburn, MA). Using the automated Illumina cBot cluster station, non-multiplexed samples were processed in batches of eight (one for each lane of the flow-cell), diluted, and denatured to their final effective loading concentrations. Hybridization was followed by cluster generation via bridge PCR as per standard protocols (Illumina). Enriched libraries were sequenced on an Illumina GAIIx or HiSeq 2000 using either paired-end 76 base or 50 base runs, respectively.

Read mapping and variant analysis

Samples were processed from real-time base-calls (RTA 1.7 software [Bustard], converted to qseq.txt files, and aligned to a human reference (hg19) using BWA (Burrows-Wheeler Aligner; 32). Read-pairs not mapping within ±2 standard deviations of the average library size (~125 ± 15 bp for exomes) were removed. Data was processed using the Genome Analysis ToolKit (GATK refv1.2905; 33). All aligned read data was subjected to "duplicate removal", indel realignment (GATK IndelRealigner), and base quality recalibration (GATK TableRecalibration). Variant detection and genotyping were performed using the UnifiedGenotyper (UG) tool from GATK, and only performed on the targeted exome regions. Variant data for each sample was formatted (variant call format [VCF]) as "raw" calls for all samples, and lower quality/false positives sites were flagged using the filtration walker (GATK) (low quality scores (≤50), allelic imbalance (≥0.75), long homopolymer runs (>3), and/or low quality by depth <5). Samples were considered complete when exome targeted read coverage was >8x over >90% of the exome target.

Data analysis QC

All individual exome sequencing data were evaluated against QC metrics including assessment of: (1) total reads: a minimum of 30M PE reads; (2) library complexity: the ratio of unique reads to total reads mapped to target; (3) capture efficiency: the ratio of reads mapped to target versus the reads mapped to human; (4) coverage distribution: 90% at ≥8x required for completion; (5) capture uniformity; (6) raw error rates; (7) Ti/Tv ratio (3.2 for known sites and 2.9 for novel sites); (8) distribution of known and novel variants relative to dbSNP; (9) fingerprint concordance >99%; (10) homozygosity; (11) heterozygosity. All QC metrics for both single-lane and merged data were reviewed to identify data deviations from known or historical norms. Lanes/samples that failed QC were re-queued for library prep or further sequencing.

Broad Institute exome sequencing / analysis

Receipt/QC of sample DNA

Samples were shipped to the Biological Samples Platform laboratory at the Broad Institute of MIT and Harvard. DNA concentration was determined by the Picogreen assay before storage in 2D-barcoded 0.75 mL Matrix tubes at −20°C in the SmaRTStore™ (RTS, Manchester, UK) automated sample handling system. We performed initial QC on all samples involving sample quantification (PicoGreen), confirmation of high-molecular weight DNA and fingerprint genotyping, and sex determination (Illumina iSelect). Samples failed if the total mass, concentration, integrity of DNA or quality of preliminary genotyping data was too low.

Library construction and in-solution hybrid selection
Starting with 3µg of genomic DNA, library construction and in-solution hybrid selection were performed as described previously (34). A subset of samples were, however, prepared using the Fisher et al. protocol (34) with slight modifications. Initial genomic DNA input into shearing was reduced from 3µg to 100ng in 50µL of solution. In addition, for adapter ligation, Illumina paired end adapters were replaced with palindromic forked adapters with unique eight base index sequences embedded within the adapter.

Preparation of libraries for cluster amplification and sequencing
After in-solution hybrid selection, libraries were quantified using qPCR (KAPA Biosystems) with probes specific to the ends of the adapters. This assay was automated using Agilent's Bravo liquid handling platform. Based on qPCR quantification, libraries were normalized to 2nM and then denatured using 0.1 N NaOH using Perkin-Elmer's MultiProbe liquid handling platform. A subset of the samples prepared using forked, indexed adapters was quantified using qPCR, normalized to 2nM using Perkin-Elmer's Mini-Janus liquid handling platform, and pooled by equal volume using the Agilent Bravo. Pools were then denatured using 0.1 N NaOH. Denatured samples were diluted into strip tubes using the Perkin-Elmer MultiProbe.

Cluster amplification and sequencing
Cluster amplification of denatured templates was performed according to the manufacturer's protocol (Illumina) using either Genome Analyzer v3, Genome Analyzer v4, or HiSeq 2000 v2 cluster chemistry and flowcells. After cluster amplification, SYBR Green dye was added to all flowcell lanes, and a portion of each lane visualized using a light microscope, in order to confirm target cluster density. Flowcells were sequenced either on Genome AnalyzerII using v3 and v4 Sequencing-by-Synthesis Kits, then analyzed using RTA v1.7.48, or on HiSeq 2000 using HiSeq 2000 v2 Sequencing-by-Synthesis Kits, then analyzed using RTA v1.10.15. All samples were run on 76 cycle, paired end runs. For samples prepared using forked, indexed adapters, Illumina's Multiplexing Sequencing Primer Kit was also used.

Read mapping and variant analysis
Samples were processed from real-time base-calls (RTA 1.7 software [Bustard], converted to qseq.txt files, and aligned to a human reference (hg19) using BWA (32). Aligned reads duplicating the start position of another read were flagged as duplicates and not analyzed ("duplicate removal"). Data were processed using the Genome Analysis ToolKit (GATK v1.1.3; (33). Reads were locally realigned (GATK IndelRealigner) and their base qualities were recalibrated (GATK TableRecalibration). Variant detection and genotyping were performed on both exomes and flanking 50 bp of intronic sequence using the UnifiedGenotyper (UG) tool from the GATK. Variant data for each sample was formatted (variant call format [VCF]) as "raw" calls for all samples. SNP and Indel sites were flagged using the Variant Filtration walker (GATK) to mark sites of low quality that were likely false positives. SNPs were marked as potential errors if they exhibited strong strand bias (SB >= 0.10), low average quality (QD <5.0), or fell in a homopolymer run (HRun > 4). Indels were marked as potential errors for low quality (QUAL < 30.0), low average quality (QD < 2.0), or if the site exhibited strong

strand bias (SB > -1.0). Samples were considered complete when exome targeted read coverage was ≥20x over ≥80% of the exome target.

Data analysis QC
Sequence data that was processed were required to match known fingerprint genotypes for their respective samples, and to achieve the sequence coverage noted above (20x over 80% target). Variant calls were evaluated on both bulk and per-sample properties: novel and known variant counts, Ti/Tv ratio, Het/Hom ratio, and Deletion/Insertion ratio. Both bulk and sample metrics were compared to historical values for exome sequencing projects at the Broad. No significant deviation of the ESP calls or ESP samples from historical values were noted.

Variant calling

Input data
Variant discovery and genotyping was performed simultaneously across the initial set of 2,520 exomes sequenced by the NHLBI Exome Sequencing Project. These consisted of 1,384 samples sequenced at the University of Washington and 1,136 samples sequenced at the Broad Institute. Most sequenced samples were of self-identified European or African American Ancestry, although self-identified ancestry was not available for a subset of samples. Sequencing centers carried out initial alignment and pre-processing for each sample. Briefly, read data were aligned to NCBI build 37 reference using BWA (32) and indel-realigned and recalibrated using the Genome Analysis Toolkit (GATK; 33). The initial pre-processing and alignment steps carried out at each center were broadly similar.

As indicated above, three different target solutions were used for exome capture across the 2,520 samples: Broad used Agilent capture solution (Agilent SureSelect Human All Exon Kit v2 [31 Mb]) and UW used two different targets of Nimblegen-designed solutions (custom RefSeq/CCDS design [28 Mb]; SeqCap EZ v1 [32 Mb]). In a preliminary analysis of chromosome 20, we found variants called within 50 bp of target regions had similar transition-transversion ratios, variant density, and overlap with dbSNP compared to variants called on-target; thus we attempted to call variants within 50bp of target regions in each sample. When X chromosome variant calls were generated, sex information was available only for a subset of samples; hence we used normalized read depths across the sex chromosomes to infer sex for the remaining samples, identifying 774 samples as male and 1,746 samples as female.

Calling pipeline overview
Briefly, we took a two-step approach for discovering variants and genotyping candidate sites. First, genotype likelihood files (GLFs) were generated using SAMtools pileup on individual BAM files. Next, we used glfMultiples - a multi-sample variant caller - to generate initial SNV calls. This two-step approach is very scalable and is especially beneficial for exome sequence data, because the size of GLF files largely depends on the number of sites targeted for analysis (~52.6 M sites are within 50 bp of one of the

three ESP capture reagents). This process allowed us to perform multisample calling of variants simultaneously across thousands of samples to avoid batch effects.

For the initial variant calls, we assumed that the prior probability that a site was polymorphic was $Prior(SNP) = \theta \sum_{i=1}^{2n} \frac{1}{i} = 0.0091$, 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 θ = 0.001. We assumed transition and transversion mutations were equally likely – while this assumption is not optimal, it makes it easier to use transition-transversion (T_i/T_v) ratio as a diagnostic of SNP call quality. Details of the likelihood model implemented in glfMultiples are given in (35) in the section entitled "Identifying Potential Polymorphic Sites". Briefly, the distribution of observed bases and quality scores at each location (conditional on the true genotype) is modeled using the MAQ model (36). Using maximum likelihood, we then estimated an allele frequency for each site under the assumption that genotypes were segregating in Hardy-Weinberg proportions. Genotypes for each site and corresponding posterior probabilities were then calculated via Bayes theorem, using the Hardy-Weinberg proportions as priors and the MAQ error model to describe the conditional probability of bases and quality scores given the true genotype.

Initial variant filtering

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 for reads carrying reference and variant alleles and the average position of variant alleles along a read.

The final SNV call set included variants that were called with posterior probability >99% (glfMultiples SNP quality >20), were >5bp away from an indel detected in the 1000 Genomes Pilot Project (37), were covered by at least one read in 85% of samples, and had total depth across samples of between 2,500 and 2,500,000 reads (~1-100 reads per sample). Sites having >65% of reads as 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. Additional quality metrics were examined, including the fraction of unexpected alleles (those that did not match either the reference or alternate allele) and the average location of variant calls along each read. These metrics were not used to filter this initial dataset. All processing steps were automated using GNU make, which facilitates parallelization of data processing steps, dependency checking, and repeats of failed runs.

Data filtering

The lightly filtered data set described above was further subjected to more stringent filtering in order to obtain high quality genotypes suitable for population genetics analyses. Because three distinct capture targets were used in sequencing, we restricted all analyses to the intersection of the capture targets; that is, sites not appearing on all

three targets were excluded. We removed individual genotypes if quality (GQ) was under 30 and/or filtered depth (DP or GD) was under 10. After such filtering, if more than 10% of individuals had a missing genotype, we excluded the entire site from analysis. We identified clusters of African Americans and European Americans based on ancestry-informative markers and mean heterozygosity per individual (see below). We used these clusters to remove sites violating Hardy-Weinberg equilibrium (chi-square p-value < 10^{-6}) with heterozygote or homozygote excess within either cluster. Similarly, we removed sites with overall heterozygote excess in the full dataset, but not sites with overall homozygote excess as these could represent population structure. After filtering Ti/Tv ratios for synonymous, missense, nonsense, and splice variants were 5.60, 2.31, 2.13, and 1.69, respectively.

Identifying related individuals

We used KING (38) to identify duplicate samples and kin with a third degree (e.g. first cousin) or closer relationship (Figure S8). Five exomes were removed due to low coverage and 15 exomes showed high and presumably spurious relatedness to a large number of other samples, likely due to low quality or contamination; we conservatively removed these exomes from subsequent analyses as well. For all duplicate (N=22) and kinship pairs (N=48), the individual with lower mean filtered depth was removed. We removed 60 samples as some individuals have multiple relationships (Figure S8).

Assignment of ancestry labels

Designated ancestry was determined using heterozygosity and ancestry-informative markers (AIMs), which we defined as autosomal SNVs estimated from preexisting datasets to exhibit large allele frequency differences between individuals of European and African ancestry (Figure S19). We chose 29 exomic AIMs varying between Europeans and Africans in the HapMap datset, and 17 exomic AIMs varying between those continents and Native Americans in the HGDP dataset. The one self-identified Native American in our dataset was a clear outlier, being the only individual matching over 40% of Native American AIMs and fewer than 65% European AIMs, and was designated as Native American. The remaining samples formed two very distinct clusters: 97% of samples had either >65% European AIMs and 13-15% of SNVs heterozygous, or else < 50% European AIMs and 16.5-18.5% of SNVs heterozygous. In order to assign the few ambiguous individuals falling between clusters, we designated all individuals with over 50% European AIMs and <16% of SNVs heterozygous as European American, and all other individuals as African American. This approach resulted in less ambiguous ancestry designation compared to a simple PC analysis.

Data annotation

We used SeattleSeq Annotation 131 to annotate all SNVs, with respect to site type, presence in existing databases, and other factors.

Prediction of functional variation

To identify putatively functional variation, we used we used PolyPhen2 (39), SIFT (40), a likelihood ratio test (41), Mutation Taster (42), GERP (43), PhyloP (44), and a novel population genetics approach that combines conservation information with the SFS that we designate SFS-Del (see below). Thresholds in determining whether a given metric predicted a SNV to be functional were as follows: PolyPhen2 “Probably Damaging”, SIFT “Damaging”, likelihood ratio test “Deleterious”, MutationTaster “disease causing automatic” or “disease causing”, GERP ≥ 5, PhyloP ≥ 3, SFS-Del probability ≥ 0.90.

Analyses of batch and cohort effects

Exome sequences were assessed for batch and cohort effects using several strategies. First, among individual exomes all samples showed similar properties of Ti/Tv, heterozygosity and frequency of missing genotypes (Figure S1). Second, no substantial differences were observed among sets of exomes sequences defined by target definition, target group, and sequencing center vs. heterozygote / homozygous non-reference (Figure S2) and Ti/Tv (Figure S3). Third, we assessed variant QC metrics including depth of coverage, frequency of missing genotypes, and minor allele frequency (MAF). Ti/Tv was relatively constant among a broad range of coverage, frequencies of missing data < 10%, and MAF (Figure S4). Last, to confirm that patterns of population structure and other population parameters were not significantly biased by sequencing strategy, we relabeled the PCA analyses of population structure by: exome capture target (Figure S5), sequencing center (Figure S6), and cohort/phenotype information (Figure S7) for each sample. Plots are based on variants with a MAF > 5% and pruned by LD.

Validation

Selection of variants for validation

Single nucleotide variants (SNVs) were selected for validation if they were novel – i.e. not present in dbSNP 131, a single base change (i.e. non-indel), biallelic, and functionally annotated as a missense or nonsense variant. This resulted in a total set of 247,906 variants over a frequency range from 1 observation (i.e., a singleton; 1/4880 chromosomes) to more common variants at a frequency >10%. From this set, we selected novel variants for validation including: 400 singletons, 768 variants <10%, and all 52 variants identified with >10% MAF. SNVs selected for validation were not allowed to be within 100,000bp of each other, nor within 100bp of any other known SNV. These variants were evenly split between the sequencing centers for testing.

University of Washington genotyping/validation

Directed PCR and Sanger sequencing were used to confirm all singleton variants (n=200). Primer design, optimization, and testing resulted in 146/200 passing amplicons. For each variant position, we selected the specific genotype positive singleton sample, along with six Coriell negative control samples. All chromatograms were generated on an ABI 3700XL, and aligned to a genomic reference sequence. Polyphred was used to identify putative variant positions within an amplicon and the specific position of interest was directly examined and validated by a data analyst. All samples were genotyped using PolyPhred or manually. Comparison of exome genotypes against the Sanger data resulted in 2/146 (1.3% discordant) genotypes. Of the singletons that failed confirmation, one variant was annotated by the filtration walker (GATK) as failing allelic balance (GATK filter status was not originally used as a selection criteria). The second singleton that could not be confirmed was supported by manual review of the exome data, however multiple PCR amplicons and sequencing did not confirm this sample genotype.

Singleton variants were validated by Sanger sequencing. Of 145 singleton variants that were assessed, 143 (99%) were validated. Non-singleton variants were validated using the Illumina BeadXpress (n=384 plex) assay. All 820 variants (768 < 10% MAF + 52 > 10% MAF) were submitted for design scores and the highest ranking variants used in a 384 plex genotyping pool (372 < 10% MAF + 12 > 10% MAF). All BeadXpress sample pools (10 pools x 96 samples) were scanned and initially clustered/genotyped using the BeadStudio software, with subsequent manual review. We selected 960 samples from the genotype positive sample set to be assayed across all 384 variants. 934/960 (98%) samples passed QC (>95% of variants called) and were used for validation. We found 332/384 (86%) variants passing QC (>90% of samples reporting a genotype) of which 316/332 (95%) of the variants were confirmed in the genotype positive samples. Of 332 non-singleton variants tested, 323 had a MAF < 10%. We validated 316/323 (98%) non-singleton variants with 1191/1255 (95%) confirmed genotypes. The observed confirmation failures (5%) appeared to be due to genotype clustering, with unconfirmed samples falling at or near cluster boundaries. Nine SNVs with MAF>10% were tested and all nine were validated with 1914/1925 (99%) confirmed genotypes.

Broad genotyping/validation

Singleton and non-singleton variants were validated using iplex technology (Sequenom; John Hopkins Court, San Diego, CA). In all, 97 putative singletons identified in 94 DNA samples were genotyped in 8 SNP pools. Each SNP was genotyped on the specific genotype positive singleton samples and the other 93 DNA samples from which the initial pool of singletons were selected. Four of the sites demonstrated high missingness (a no-call rate of >5%) and were excluded from further analysis. Of the remaining 93 singletons, 91 (98%) were validated. The two putative singletons that could not be detected were determined to be monomorphic and this was confirmed by visual analysis of the raw intensity plots. For validation of non-singletons, 402 mis- or non-sense SNPs with MAF $\leq$ 10% were selected for testing. All 402 SNPs were sequenced in 1024 samples from the genotype positive sample set. 352 of the 402 SNPs (88%) tested passed initial QC with 331 confirmed in the genotype positive samples. Of 391 non-singleton variants tested, 370 had a MAF < 10%. We validated 362/370 (98%) non-

singleton variants. Twenty-one SNVs with MAF>10% were tested and 20/21 (95%) were validated.

Power of variant discovery

We used two simple approaches for calculating the power of variant discovery. First, the probability that a variant with a population frequency f is observed in a sample of n individuals is $1-(1-f)^{2n}$ (45). Second, we used the hypergeometric distribution to calculate the probability of sampling one or more alleles of a variant drawn from a population of size N that consists of m minor alleles, N-m major alleles (i.e., the more common allele), in a sample size of k chromosomes is 1 – hyper(0, m, N-m, k). We evaluated the probabilities of variant discovery for a range of allele frequencies and population sizes, N. Note, both of these approaches implicitly assume a randomly mating population and therefore should be regarded as rough approximations on the power of variant discovery.

Comparison of variants discovered in the ESP with the 1000 Genomes Project

SNVs were compared between the ESP and 1000 Genomes Project (1KGP) for individuals of European ancestry (1000 Genomes Project Consortium; 7). In total, 1,351 ESP EA individuals were compared to 381 1KGP European individuals. The 1KGP European populations were: 87 Utah residents with Northern and Western European ancestry (CEU), 93 Finnish in Finland (FIN), 89 British from England and Scotland (GBR), 14 Iberian populations in Spain (IBS), and 98 Toscani in Italia (TSI). Data was downloaded from the 1KGP-Interim Phase I data (June 2011 release ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/).
The comparison focused on 140,323 autosomal regions (~23.92 Mb), with sequence coverage rates > 90% in the ESP. Among these regions, 261,251 SNVs were identified in the 1,351 EA ESP individuals and 90,961 SNVs were identified in the 381 European individuals from 1KGP (Table S1). Among these variants, 59,759 were identified in both the ESP and in 1KGP. For SNVs with a MAF ≤ 5%, 201,379 variants were uniquely observed in the ESP, compared with 25,641 variants uniquely observed in 1KGP. For SNVs with a MAF > 5%, 113 SNPs were uniquely observed in ESP, compared with 5,561 variants uniquely observed in 1KGP.

Among the 5,561 common SNVs that were not observed in ESP, 4,893 (87.98%) sites were filtered out in the ESP because >10% of individuals were called as missing (GQ<30 and/or DP<10), denoted as C90; 322 (5.79%) sites were filtered out in ESP because of departure from HWE in EAs (chi-square > 24), denoted as HWD; 151 (2.72%) sites were invariant (i.e., genotypes of all individuals are reference homozygous) in ESP, noted as INV; 120 (2.16%) sites were not called or were not targeted, noted as NoC; and 75 (1.35%) sites were filtered out for other reasons in ESP, such as allele balance, nearby indels, etc, denoted as OTH.

Figure S10 illustrates characteristics, such as reference allele frequency and HWE, of common SNVs specific to the 1KGP data. SNVs filtered out of the ESP data because of HWD or NoC (Figure S10B and D) also exhibit extreme departures from HWE in 1KGP. This departure from HWE in 1KGP is not explained by population structure compared with Figure S9A. SNPs filtered out of the ESP data because they were invariant (INV) show a higher frequency of the reference allele in 1KGP (Figure S9C), some of which were only observed in specific European populations, such as FIN, GBR and IBS (Table S2).

Thus, of the 5,561 common SNVs in 1KGP that were not observed in ESP, all but 151 variants can be accounted for by being filtered out of the ESP data or not present in the capture target. The remaining 151 SNPs can largely be accounted for by European population stratification (Table S2), as the 1KGP included a broader sampling of European populations.

Demographic Inference

Demographic inference based on the site frequency spectrum was performed using the package *dadi version 1.5.2* (46), using a demographic model similar to a previously described three-population model (46,47), but with three added parameters to allow for a recent growth, namely a time for growth change and two recent growth rates (one in the European and one in the African population).

Because we want to resolve the site frequency spectrum for population frequencies of the order of 0.0005, and dadi calculates the site-frequency spectrum by numerically solving diffusion equations, we require a very fine discretization of frequency space, which is computationally demanding and makes parameter inference challenging. As we expect the extra information provided by the current data to be particularly informative of recent migration, we fixed the parameters of ancient demographic events to those obtained in (47), allowing only the recent European and African demography to change, namely the two recent growth parameters in Europeans and Africans, the time for the accelerated growth, and the growth rate in Europeans since the split with the ancestors to the Asian population. Even with these constraints, the requirements for simulating a 3-population model at this level of accuracy are high. Since the current data does not include individuals of reported Asian ancestry, we did not simulate the Asian population in this model after the split from the European population. Because the original model includes migration between all populations, we verified that this omission did not overly affect the joint African-American/European site-frequency spectrum--the maximum difference between two-population and three-population models in any of the bins was within 21.2% of the expected Poisson sampling noise in the data, when the simulation was carried with 100 grid points, and this number decreased to 16% with 500 grid points (we expect this number to decrease further when convergence is reached).

For the optimization process, we used grids of 500, 750, and 1000 points per population and extrapolated to infinite sample size using the method described in (46). Once we

obtained optimal parameters, we used an even finer grid (1000, 1500, and 2000) to generate plots of frequency spectra and perform comparisons with the real data.

To obtain confidence intervals for the recent growth parameters, we performed a bootstrap analysis over exons. More specifically, we first defined 8239 target regions separated by more than 50kb, by merging together target regions separated by a smaller distance. Then, we generated 24 bootstrapped samples by selecting 8239 genomic regions with replacement among the initial regions. For each bootstrap sample, we calculated the allele frequency spectrum, and optimized the likelihood as a function of the four recent demographic parameters, as for the initial point estimates. Convergence of the optimization was ensured by performing six optimizations from different starting positions and verifying that the highest-likelihood values were attained for at least two initial conditions. For computational efficiency, we used coarser grids than in the point estimate (extrapolating from results 100, 200, and 400 grid points per population), as we expect the width of the confidence intervals to be less affected than point estimates by finite grid effects. We use the obtained standard deviation to estimate confidence interval for the point estimate reported in the text.

Analysis of natural selection

Inference of positive selection

All positive selection analyses were restricted to the autosomes, so that all compared genes shared the same demographic history. To find genes showing evidence of positive selection in humans, we first identified human-chimp fixed differences by comparison to panTro3, excluding sites that were polymorphic in our dataset. We retained only human-specific substitutions by excluding sites where the human allele was identical to macaque (rheMac2); sites where neither the chimp nor the human allele matched macaque were assigned a 50% probability of being a human-specific substitution. We used these sites to identify genes with high ratios of human-specific divergence to within-human polymorphism, as well as genes with high ratios of nonsynonymous/synonymous divergence relative to their ratio of nonsynonymous/synonymous polymorphism. For all measures of within-human polymorphism, we used π rather than counts of segregating sites in order to minimize the signal from rare and likely deleterious variants. To find genes showing evidence of population-specific adaptation, we identified candidates with unusually high or low ratios of AA polymorphism to EA polymorphism. We also identified genes with at least one SNV showing F_{ST} at least 0.3 and either AA polymorphism or EA polymorphism in the lowest 20% for that sample.

Inferences of purifying selection based on the SFS

We inferred purifying selection by examining the joint SFS (in EA and AA) of different site classes. We classified sites according to four basic type categories (synonymous, missense. nonsense/splice, and other) and GERP score rounded to the nearest integer. We assumed than synonymous sites with a GERP score of -5 or less (site class S) represented a neutral class of sites with a joint SFS reflecting only demographic history. We assumed that the effects of positive and balancing selection on segregating

polymorphism are negligible for the exome as a whole. We compared joint SFS for sites S against all non-S site classes, assuming that any differences are due to the presence of deleterious variants (D) in the non-S class, in addition to the neutral non-S variants (N) which behave just like S. We assumed that all sites which reach a frequency of 50% in both EA and AA are effectively neutral, so all non-S variants are N. For a given joint frequency rarer than 50% (F_{EA} and F_{AA}) the ratio of non-S deleterious SNVs (D+N) to selectively neutral SNVs (S) is can be approximated by the ratio at 50% (N/S) multiplied by several constants which scale with the logs of the frequencies and their interaction term:

$$\ln((D+N)/S) = \ln(N/S) + \beta_1(\ln(0.5/F_{EA})) + \beta_2(\ln(0.5/F_{AA})) + \beta_3((\ln(0.5/F_{EA}))*(\ln(0.5/F_{AA})))$$

We used multiple regression to solve for the β values for all site classes, combining some similar site classes to ensure a large sample size for all regression calculations. Thus, for a site in a given site class at a given frequency, we could estimate the probability that it was D rather than N. By summing these probabilities across the exome, once for every heterozygous or derived hemizygous variant and twice for every derived homozygous variant, we estimated the number of deleterious variants per individual. In effect, we estimated the number of variants that would be prevented by selection from reaching a 50% frequency in both populations if the populations continued to experience the same selective pressures and demographic processes as they have in the past. The advantage of this approach is that we have to make very few *a priori* assumptions about which site types are likely to be deleterious; rather we can empirically estimate a probability for each site class based on how disproportionately prevalent that site class is at increasingly rarer frequencies.

Purifying selection and protein structure

We analyzed six features that are potentially important for protein structural stability: normalized solvent-accessible surface area, change in amino acid side chain volume, change in amino acid hydrophobic potential, average number of hydrogen bonds formed by the residue, and secondary structure element. We also analyzed three features that are potentially important for protein biochemical function: number of active site contacts, number of heteroatom contacts, and number of side chain contacts. Variants with normalized accessible surface area above the median were classified as surface variants; variants with normalized accessible surface area below the median were classified as buried variants. Variants with side chain volume change above the top quintile (large positive volume change) were classified as overpacking variants; variants with side chain volume change below the bottom quintile (large negative volume change) were classified as cavity-forming variants. Variants with hydrophobic potential change above the top quintile were classified as hydrophobic potential-changing variants. Secondary structure elements were grouped into helices, strands, and loops. Variants in sites that formed any hydrogen bonds were classified as hydrogen-bonding; variants in sites that made any active site contacts were classified as active site variants; variants in sites that made any heteroatom contacts were

classified as ligand-binding variants; and variants in sites that made any interchain contacts were classified as variants contacting another protein chain.

Nonsynonymous variants were mapped to PDB structures using PolyPhen2, using a minimum sequence identity of 50% and a minimum alignment length of 100 amino acids. Each variant that mapped to a structure was then scored for the nine features described above. For each of the categories described above, we performed a random permutation test to determine whether the category contains a statistically significant enrichment of rare variants. For each category, we chose 100,000 random sets of variants that mapped to PDB structures and counted the number of variants with a MAF < 0.005.

Purifying selection and synonymous variation

The w scores of all possible human codons were taken from (48). The change in w score was calculated as the difference between that of the codon with the derived allele and that of the codon with the ancestral allele. The hypothesis that the correlation between the derived allele frequency (DAF) and the functional or w scores is equal to zero was tested using Fisher's transformation and a z score test.

Coordinated purifying selection on synonymous and nonsynonymous variation within genes

For each gene we estimated Spearman's rank correlation coefficients (RCCs) between the average w score changes of synonymous variants and the average functional prediction scores (SIFT, Polyphen2, LRT and MutationTaster) of the nonsynonymous variants. We observed a weak but statistically significant negative correlation between w score changes and functional prediction scores from SIFT, Polyphen2 and MutationTaster (Table S7), which suggests potential coordinated purifying selection acting on both the synonymous variants and the non-synonymous variants within a gene.

Rare variant association power analysis

The power of association studies on individual rare variants is extremely low, and numerous methods have been developed to pool, or aggregate, rare variants at a gene to improve power (28). Here, we used a simple Fisher's Exact Test (FET) to explore power on a gene by gene basis In order to simulate phenotypes we used a logistic model of the form:

$$\text{logit}(p) = \beta*(x_1+x_2\ldots x_n) + \varepsilon$$

where the x's are each individual variant in the genic region, and β was set to either 0.405 or 1.609, which leads to an OR of 1.5 and 5, respectively. ε is the random error term drawn from aa standard normal distribution. The logistic model yields a continuous probability of being a case or control, which is subsequently determined using a uniformly distributed random number compared to the probability. The data were

preprocessed to include only 1000 EA and AA, which will be sub-selected for 400 cases and 400 controls; sites with a MAF < 1%, calculated at the 1000 individual level for each sample (ignoring those greater than 1% from random sampling); and sites with a functional designation of missense, nonsense, splice-site, or GERP ≥ 5.

In an effort to remove spurious associations due to population demography and structure, the 1000 individuals for each population were selected in a PCA aware manner. For both samples, we use a global PCA of sites with a MAF > 5% and pruned for LD (plink --indep-pairwise 50 5 0.5) (see Figure S17). For the AA sample, we selected the 1,000 individuals with the lowest PC1 values, which is indicative of level of European admixture. With each analysis of AA, PC1 will be included as a covariate. For the EA sample, we selected the 1,000 individuals with the lowest PC2 values, which is indicative of a North/South European cline. This removes all the individuals that cluster with Eastern European Jewish populations and the most extreme individuals that cluster with Tuscans from HapMap3 (see below). The power results per gene for an OR = 5 are shown in the main text (Figure 4A) and for an OR = 1.5 in Figure S18.

Further acknowledgements

HeartGO:

Atherosclerosis Risk in Communities (ARIC): NHLBI (N01 HC-55015, N01 HC-55016, N01HC-55017, N01 HC-55018, N01 HC-55019, N01 HC-55020, N01 HC-55021); **Cardiovascular Health Study (CHS)**: NHLBI (N01-HC-85239, N01-HC-85079 through N01-HC-85086, N01-HC-35129, N01 HC-15103, N01 HC-55222, N01-HC-75150, N01-HC-45133, and grant HL080295), with additional support from NINDS and from NIA (AG-023629, AG-15928, AG-20098, and AG-027058); **Coronary Artery Risk Development in Young Adults (CARDIA)**: NHLBI (N01-HC95095 & N01-HC48047, N01-HC48048, N01-HC48049, and N01-HC48050); **Framingham Heart Study (FHS)**: NHLBI (N01-HC-25195 and grant R01 NS17950) with additional support from NIA (AG08122 and AG033193); **Jackson Heart Study (JHS)**: NHLBI and the National Institute on Minority Health and Health Disparities (N01 HC-95170, N01 HC-95171 and N01 HC-95172); **Multi-Ethnic Study of Atherosclerosis (MESA)**: NHLBI (N01-HC-95159 through N01-HC-95169 and RR-024156).

Lung GO:

Cystic Fibrosis (CF): Cystic Fibrosis Foundation (GIBSON07K0, KNOWLE00A0, OBSERV04K0, RDP R026), the NHLBI (R01 HL-068890, R02 HL-095396), NIH National Center for Research Resources (UL1 RR-025014), and the National Human Genome Research Institute (NHGRI) (5R00 HG-004316). **Chronic Obstructive Pulmonary Disease (COPDGene):** NHLBI (U01 HL-089897, U01 HL-089856), and the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, Novartis, Pfizer, and Sunovian. The COPDGene clinical centers and investigators are available at www.copdgene.org. **Acute Lung Injury (ALI)**: NHLBI (RC2 HL-101779). **Lung Health Study (LHS)**: NHLBI (RC2 HL-066583), the NHGRI (HG-004738), and the NHLBI Division of Lung Diseases (HR-46002). **Pulmonary Arterial Hypertension (PAH):** NIH (P50 HL-084946, K23 AR-52742), and the NHLBI (F32 HL-083714). **Asthma:** NHLBI (RC2 HL-101651), and the NIH (HL-077916, HL-69197, HL-76285, M01 RR-07122).

SWISS and ISGS:

Siblings with Ischemic Stroke Study (SWISS): National Institute of Neurological Disorders and Stroke (NINDS) (R01 NS039987); Ischemic Stroke Genetics Study (ISGS): NINDS (R01 NS042733)

WHISP:

Women's Health Initiative (WHI): The WHI Sequencing Project is funded by the NHLBI (HL-102924) as well as the National Institutes of Health (NIH), U.S. Department of Health and Human Services through contracts N01WH22110, 24152, 32100-2, 32105-6, 32108-9, 32111-13, 32115, 32118-32119, 32122, 42107-26, 42129-32, and 44221. The authors thank the WHI investigators and

staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: http://www.whiscience.org/publications/WHI_investigators_shortlist.pdf.

NHLBI GO Exome Sequencing Project

BroadGO

Stacey B. Gabriel (Broad Institute)[4, 5, 11, 16, 17], David M. Altshuler (Broad Institute, Harvard Medical School, Massachusetts General Hospital)[1, 5, 7, 17], Gonçalo R. Abecasis (University of Michigan)[3, 5, 9, 13, 15, 17], Hooman Allayee (University of Southern California)[5], Sharon Cresci (Washington University School of Medicine)[5], Mark J. Daly (Broad Institute, Massachusetts General Hospital), Paul I. W. de Bakker (Broad Institute, Harvard Medical School, University Medical Center Utrecht)[3, 15], Mark A. Depristo (Broad Institute)[4, 13, 15, 16], Ron Do (Broad Institute)[5, 9, 13, 15], Peter Donnelly (University of Oxford)5, Deborah N. Farlow (Broad Institute)[3, 4, 5, 14, 12, 16, 17], Tim Fennell (Broad Institute), Kiran Garimella (University of Oxford)[4, 16], Stanley L. Hazen (Cleveland Clinic)[5], Youna Hu (University of Michigan)[3, 9, 15], Daniel M. Jordan (Harvard Medical School, Harvard University)[13], Goo Jun (University of Michigan), Sekar Kathiresan (Broad Institute, Harvard Medical School, Massachusetts General Hospital)[5, 8, 9, 14, 12, 15, 17, 20], Steven Kawut (University of Pennsylvania)[10], Adam Kiezun (Massachusetts Institute of Technology)[5, 13, 15], Grigoriy Kryukov (Broad Institute), Guillaume Lettre (Broad Institute, Montreal Heart Institute, Université de Montréal)[1, 2, 13, 15], Bingshan Li (University of Michigan)[3], Mingyao Li (University of Pennsylvania)[5], Christopher H. Newton-Cheh (Broad Institute, Massachusetts General Hospital, Harvard Medical School)[3, 8, 15], Sandosh Padmanabhan (University of Glasgow School of Medicine)[3, 12, 15], Sara Pulit (Broad Institute)[3, 15], Daniel J. Rader (University of Pennsylvania)[5], David Reich (Broad Institute, Harvard Medical School)[15], Muredach P. Reilly (University of Pennsylvania)[5], Manuel A. Rivas (Broad Institute, Massachusetts General Hospital)[5], Steve Schwartz (Fred Hutchinson Cancer Research Center)[5, 12], Laura Scott (University of Michigan)[1], Johnathon A. Spertus (University of Missouri Kansas City)[5], Nathaniel O. Stitziel (Brigham and Women's Hospital)[5, 15], Nina Stoletzki (Brigham and Women's Hospital, Broad Institute, Harvard Medical School)[13], Shamil R. Sunyaev (Brigham and Women's Hospital, Broad Institute, Harvard Medical School)[1, 3, 5, 13, 15], Benjamin F. Voight (Broad Institute, Massachusetts General Hospital), Cristen J. Willer (University of Michigan)[1, 9, 13, 15]

HeartGO

Stephen S. Rich (University of Virginia)[2, 4, 7, 8, 9, 14, 11, 15, 17, 18, 31], Ermeg Akylbekova (Jackson State University, University of Mississippi Medical Center)[29], Larry D. Atwood (Boston University)[1, 11, 28], Christie M. Ballantyne (Baylor College of Medicine, Methodist DeBakey Heart Center)[9, 22], Maja Barbalic (University of Texas Health Science Center Houston)[9, 14, 15, 17, 22], R. Graham Barr (Columbia University Medical Center)[10, 31], Emelia J. Benjamin (Boston University)[14, 20, 28], Joshua Bis (University of Washington)[15, 23], Chris Bizon (University of North Carolina Chapel Hill)[3, 9, 13, 15, 23, 34], Eric Boerwinkle (University of Texas Health Science Center Houston)[3, 5, 9, 13, 15, 17, 22], Donald W. Bowden (Wake Forest University)[1, 31], Jennifer Brody (University of Washington)[3, 5, 15, 23], Matthew Budoff (Harbor-UCLA Medical Center)[31], Greg Burke (Wake Forest

University)[5, 31], Sarah Buxbaum (Jackson State University)[3, 13, 15, 29], Jeff Carr (Wake Forest University)[25, 29, 31], Donna T. Chen (University of Virginia)[6, 11], Ida Y. Chen (Cedars-Sinai Medical Center)[1, 31], Wei-Min Chen (University of Virginia)[13, 15, 18], Pat Concannon (University of Virginia)[11], Jacy Crosby (University of Texas Health Science Center Houston)[22], L. Adrienne Cupples (Boston University)[1, 3, 5, 9, 13, 15, 18, 28], Ralph D'Agostino (Boston University)[28], Anita L. DeStefano (Boston University)[13, 18, 28], Albert Dreisbach (University of Mississippi Medical Center)[3, 29], Josée Dupuis (Boston University)[1, 28], J. Peter Durda (University of Vermont)[15, 23], Jaclyn Ellis (University of North Carolina Chapel Hill)[1], Aaron R. Folsom (University of Minnesota)[5, 22], Myriam Fornage (University of Texas Health Science Center Houston)[3, 18, 25], Caroline S. Fox (National Heart, Lung, and Blood Institute)[1, 28], Ervin Fox (University of Mississippi Medical Center)[3, 9, 29], Vincent Funari (Cedars-Sinai Medical Center)[1, 11, 31], Santhi K. Ganesh (Johns Hopkins University, National Human Genome Research Institute, National Heart, Lung, and Blood Institute, University of Michigan)[2, 22, 28], Julius Gardin (Hackensack University Medical Center)[25], David Goff (Wake Forest University)[25], Ora Gordon (Cedars-Sinai Medical Center)[11, 31], Wayne Grody (University of California Los Angeles)[11, 31], Myron Gross (University of Minnesota)[1, 5, 14, 25], Xiuqing Guo (Cedars-Sinai Medical Center)[3, 15, 31], Ira M. Hall (University of Virginia), Nancy L. Heard-Costa (Boston University)[1, 11, 28], Susan R. Heckbert (University of Washington)[10, 14, 15, 23], Nicholas Heintz (University of Vermont), David M. Herrington (Wake Forest University)[5, 31], DeMarc Hickson (Jackson State University, University of Mississippi Medical Center)[29], Jie Huang (National Heart, Lung, and Blood Institute)[5, 28], Shih-Jen Hwang (Boston University, National Heart, Lung, and Blood Institute)[3, 28], David R. Jacobs (University of Minnesota)[25], Nancy S. Jenny (University of Vermont)[1, 2, 23], Andrew D. Johnson (National Heart, Lung, and Blood Institute)[2, 5, 11, 28], Craig W. Johnson (University of Washington)[15, 31], Richard Kronmal (University of Washington)[31], Raluca Kurz (Cedars-Sinai Medical Center)[11, 31], Ethan M. Lange (University of North Carolina Chapel Hill)[3, 5, 9, 13, 34], Leslie A. Lange (University of North Carolina Chapel Hill)[1, 2, 3, 5, 9, 12, 13, 15, 17, 18, 20, 25, 34], Martin G. Larson (Boston University)[3, 15, 28], Mark Lawson (University of Virginia), Daniel Levy (National Heart, Lung, and Blood Institute)[3, 15, 17, 28], Dalin Li (Cedars-Sinai Medical Center)[11, 15, 31], Honghuang Lin (Boston University)[20, 28], Chunyu Liu (National Heart, Lung, and Blood Institute)[3, 28], Jiankang Liu (University of Mississippi Medical Center)[1, 29], Kiang Liu (Northwestern University)[25], Xiaoming Liu (University of Texas Health Science Center Houston)[15, 22], Yongmei Liu (Wake Forest University)[2, 5, 31], William T. Longstreth (University of Washington)[18, 23], Cay Loria (National Heart, Lung, and Blood Institute)[25], Thomas Lumley (University of Auckland)[9, 23], Kathryn Lunetta (Boston University)[28], Aaron J. Mackey (University of Virginia)[16, 18], Rachel Mackey (University of Pittsburgh)[1, 23, 31], Ani Manichaikul (University of Virginia)[8, 15, 18], Taylor Maxwell (University of Texas Health Science Center Houston)[22], Barbara McKnight (University of Washington)[15, 23], James B. Meigs (Brigham and Women's Hospital, Harvard Medical School, Massachusetts General Hospital)[1, 28], Alanna C. Morrison (University of Texas Health Science Center Houston)[3, 15, 17], Solomon K. Musani (University of Mississippi Medical Center)[3, 29], Josyf C. Mychaleckyj (University of Virginia)[13, 15, 31], Jennifer A. Nettleton (University of Texas Health Science Center Houston)[9, 22], Kari North (University of North Carolina Chapel Hill)[1, 3, 9, 10, 13, 15, 17, 34], Christopher J. O'Donnell (Massachusetts General Hospital, National Heart, Lung, and

Blood Institute)[2, 5, 9, 14, 11, 12, 15, 17, 20, 28], Daniel O'Leary (Tufts University School of Medicine)[25, 31], Frank Ong (Cedars-Sinai Medical Center)[3, 11, 31], Walter Palmas (Columbia University)[3, 15, 31], James S. Pankow (University of Minnesota)[1, 22], Nathan D. Pankratz (Indiana University School of Medicine)[15, 25], Shom Paul (University of Virginia), Marco Perez (Stanford University School of Medicine), Sharina D. Person (University of Alabama at Birmingham, University of Alabama at Tuscaloosa)[25], Joseph Polak (Tufts University School of Medicine)[31], Wendy S. Post (Johns Hopkins University)[3, 9, 14, 11, 20, 31], Bruce M. Psaty (Group Health Research Institute, University of Washington)[3, 5, 9, 14, 11, 15, 23], Aaron R. Quinlan (University of Virginia)[18, 19], Leslie J. Raffel (Cedars-Sinai Medical Center)[6, 11, 31], Vasan S. Ramachandran (Boston University)[3, 28], Alexander P. Reiner (Fred Hutchinson Cancer Research Center, University of Washington)[1, 2, 3, 5, 9, 11, 12, 13, 14, 15, 20, 25, 34], Kenneth Rice (University of Washington)[15, 23], Jerome I. Rotter (Cedars-Sinai Medical Center)[1, 3, 6, 8, 11, 15, 31], Jill P. Sanders (University of Vermont)[23], Pamela Schreiner (University of Minnesota)[25], Sudha Seshadri (Boston University)[18, 28], Steve Shea (Brigham and Women's Hospital, Harvard University)[28], Stephen Sidney (Kaiser Permanente Division of Research, Oakland, CA)[25], Kevin Silverstein (University of Minnesota)[25], David S. Siscovick (University of Washington)[5, 1, 25], Nicholas L. Smith (University of Washington)[2, 15, 20, 23], Nona Sotoodehnia (University of Washington)[3, 15, 23], Asoke Srinivasan (Tougaloo College)[29], Herman A. Taylor (Jackson State University, Tougaloo College, University of Mississippi Medical Center)[5, 29], Kent Taylor (Cedars-Sinai Medical Center)[31], Fridtjof Thomas (University of Texas Health Science Center Houston)[3, 22], Russell P. Tracy (University of Vermont)[5, 9, 14, 11, 12, 15, 17, 20, 23], Michael Y. Tsai (University of Minnesota)[9, 31], Kelly A. Volcik (University of Texas Health Science Center Houston)[22], Chrstina L Wassel (University of California San Diego)[9, 15, 31], Karol Watson (University of California Los Angeles), Gina Wei (National Heart, Lung, and Blood Institute)[25], Wendy White (Tougaloo College)[29], Kerri L. Wiggins (University of Vermont)[23], Jemma B. Wilk (Boston University)[10, 28], O. Dale Williams (Florida International University)[25], James G. Wilson (University of Mississippi Medical Center)[1, 2, 5, 8, 9, 14, 11, 12, 17, 20, 29], Phillip Wolf (Boston University)[28], Neil A. Zakai (University of Vermont)[2, 23]

ISGS and SWISS

John Hardy (Reta Lila Weston Research Laboratories, Institute of Neurology, University College London)[18], James F. Meschia (Mayo Clinic)[18], Michael Nalls (National Institute on Aging)[2, 18], Stephen S. Rich (University of Virginia)[2, 4, 7, 8, 9, 14, 11, 15, 17, 18, 31], Andrew Singleton (National Institute on Aging)[18], Brad Worrall (University of Virginia)[18]

LungGO

Michael J. Bamshad (Seattle Children's Hospital, University of Washington)[4, 6, 7, 8, 10, 11, 13, 15, 17, 27], Kathleen C. Barnes (Johns Hopkins University)[2, 10, 14, 12, 15, 17, 20, 24, 30, 32], Ibrahim Abdulhamid (Children's Hospital of Michigan)[27], Frank Accurso (University of Colorado)[27], Ran Anbar (Upstate Medical University)[27], Terri Beaty (Johns Hopkins University)[24, 30], Abigail Bigham (University of Washington)[13, 15], Phillip Black (Children's Mercy Hospital)[27], Eugene Bleecker (Wake Forest University)[33], Kati Buckingham (University of Washington)[27], Anne Marie Cairns (Maine Medical Center)[27], Daniel Caplan (Emory University)[27], Barbara Chatfield (University of Utah)[27], Aaron Chidekel

(A.I. Dupont Institute Medical Center)[27], Michael Cho (Brigham and Women's Hospital, Harvard Medical School)[13, 15, 24], David C. Christiani (Massachusetts General Hospital)[21], James D. Crapo (National Jewish Health)[24, 30], Julia Crouch (Seattle Children's Hospital)6, Denise Daley (University of British Columbia)[30], Anthony Dang (University of North Carolina Chapel Hill)[26], Hong Dang (University of North Carolina Chapel Hill)[26], Alicia De Paula (Ochsner Health System)[27], Joan DeCelie-Germana (Schneider Children's Hospital)[27], Allen Dozor (New York Medical College, Westchester Medical Center)[27,] Mitch Drumm (University of North Carolina Chapel Hill)[26], Maynard Dyson (Cook Children's Med. Center)[27], Julia Emerson (Seattle Children's Hospital, University of Washington)[27], Mary J. Emond (University of Washington)[10, 13, 15, 17, 27], Thomas Ferkol (St. Louis Children's Hospital, Washington University School of Medicine)[27], Robert Fink (Children's Medical Center of Dayton)[27], Cassandra Foster (Johns Hopkins University)[30], Deborah Froh (University of Virginia)[27], Li Gao (Johns Hopkins University)[24, 30, 32], William Gershan (Children's Hospital of Wisconsin)[27], Ronald L. Gibson (Seattle Children's Hospital, University of Washington)[10, 27], Elizabeth Godwin (University of North Carolina Chapel Hill)[26], Magdalen Gondor (All Children's Hospital Cystic Fibrosis Center)[27], Hector Gutierrez (University of Alabama at Birmingham)[27], Nadia N. Hansel (Johns Hopkins University, Johns Hopkins University School of Public Health)[10, 15, 30], Paul M. Hassoun (Johns Hopkins University)[10, 14, 32], Peter Hiatt (Texas Children's Hospital)[27], John E. Hokanson (University of Colorado)[24], Michelle Howenstine (Indiana University, Riley Hospital for Children)[27], Laura K. Hummer (Johns Hopkins University)[32], Jamshed Kanga (University of Kentucky)[27], Yoonhee Kim (National Human Genome Research Institute)[24, 32], Michael R. Knowles (University of North Carolina Chapel Hill)[10, 26], Michael Konstan (Rainbow Babies & Children's Hospital)[27], Thomas Lahiri (Vermont Children's Hospital at Fletcher Allen Health Care)[27], Nan Laird (Harvard School of Public Health)[24], Christoph Lange (Harvard School of Public Health)[24], Lin Lin (Harvard Medical School)[21], Tin L. Louie (University of Washington)[13, 15, 27], David Lynch (National Jewish Health)[24], Barry Make (National Jewish Health)[24], Thomas R. Martin (University of Washington, VA Puget Sound Medical Center)[10, 21], Steve C. Mathai (Johns Hopkins University)[32], Rasika A. Mathias (Johns Hopkins University)[10, 13, 15, 30, 32], John McNamara (Children's Hospitals and Clinics of Minnesota)[27], Sharon McNamara (Seattle Children's Hospital)[27], Deborah Meyers (Wake Forest University)[33], Susan Millard (DeVos Children's Butterworth Hospital, Spectrum Health Systems)[27], Peter Mogayzel (Johns Hopkins University)[27], Richard Moss (Stanford University)[27], Tanda Murray (Johns Hopkins University)[30], Dennis Nielson (University of California at San Francisco)[27], Blakeslee Noyes (Cardinal Glennon Children's Hospital)[27], Wanda O'Neal (University of North Carolina Chapel Hill)[26], David Orenstein (Children's Hospital of Pittsburgh)[27], Brian O'Sullivan (University of Massachusetts Memorial Health Care)[27], Rhonda Pace (University of North Carolina Chapel Hill)[26], Peter Pare (St. Paul's Hospital, University of Washington)[30], H. Worth Parker (Dartmouth-Hitchcock Medical Center, New Hampshire Cystic Fibrosis Center)[27], Mary Ann Passero (Rhode Island Hospital)[27], Elizabeth Perkett (Vanderbilt University)[27,] Adrienne Prestridge (Children's Memorial Hospital)[27], Nicholas M. Rafaels (Johns Hopkins University)[30], Bonnie Ramsey (Seattle Children's Hospital, University of Washington)[27], Elizabeth Regan (National Jewish Health)[24], Clement Ren (University of Rochester)[27], George Retsch-Bogart (University of North Carolina Chapel Hill)[27],

Michael Rock (University of Wisconsin Hospital and Clinics)[27], Antony Rosen (Johns Hopkins University)[32], Margaret Rosenfeld (Seattle Children's Hospital, University of Washington)[27], Ingo Ruczinski (Johns Hopkins University School of Public Health)[13, 15, 30], Andrew Sanford (University of British Columbia)[30], David Schaeffer (Nemours Children's Clinic)[27], Cindy Sell (University of North Carolina Chapel Hill)[26], Daniel Sheehan (Children's Hospital of Buffalo)[27], Edwin K. Silverman (Brigham and Women's Hospital, Harvard Medical School)[24, 30], Don Sin (Children's Medical Center of Dayton)[30], Terry Spencer (Children's Hospital Boston)[27], Jackie Stonebraker (University of North Carolina Chapel Hill)[26], Holly K. Tabor (Seattle Children's Hospital, University of Washington)[6, 10, 11, 17, 27], Laurie Varlotta (St. Christopher's Hospital for Children)[27], Candelaria I. Vergara (Johns Hopkins University)[30], Robert Weiss [30], Fred Wigley (Johns Hopkins University)[32], Robert A. Wise (Johns Hopkins University)[30], Fred A. Wright (University of North Carolina Chapel Hill)[26], Mark M. Wurfel (Harvard School of Public Health, Massachusetts General Hospital, University of Washington)[10, 14, 21], Robert Zanni (Monmouth Medical Center)[27], Fei Zou (University of North Carolina Chapel Hill)[26]

SeattleGO

Deborah A. Nickerson (University of Washington)[3, 4, 5, 7, 8, 9, 11, 15, 17, 18, 19], Mark J. Rieder (University of Washington)[4, 11, 13, 15, 16, 17, 19], Phil Green (University of Washington), Jay Shendure (University of Washington)[1, 8, 14, 16, 17], Bryan Paeper (University of Washington), Joshua M. Akey (University of Washington)[14, 13, 15], Michael J. Bamshad (Seattle Children's Hospital, University of Washington)[4, 6, 7, 8, 10, 11, 13, 15, 17, 27], Carlos D. Bustamante (Stanford University School of Medicine)[3, 13, 15], David R. Crosslin (University of Washington)[2, 9], Evan E. Eichler (University of Washington)[19], P. Keolu Fox[2], Adam Gordon (University of Washington)[11], Simon Gravel (Stanford University School of Medicine)[13, 15], Gail P. Jarvik (University of Washington)[9, 15], Jill M. Johnsen (Puget Sound Blood Center, University of Washington)[2], Eimear E. Kenny (Stanford University School of Medicine)[3, 13, 15], Jeffrey M. Kidd (Stanford University School of Medicine)[13, 15], Fremiet Lara-Garduno (Baylor College of Medicine)[15], Suzanne M. Leal (Baylor College of Medicine)[1, 13, 15, 16, 17, 19, 20], Dajiang J. Liu (Baylor College of Medicine)[13, 15], Sean McGee (University of Washington)[13, 15, 19], Peggy D. Robertson (University of Washington)[4], Joshua D. Smith (University of Washington)[4, 16, 19], Jeffrey C. Staples (University of Washington), Emily H. Turner (University of Washington), Gao Wang (Baylor College of Medicine)

WHISP

Rebecca Jackson (Ohio State University)[1, 2, 4, 5, 8, 12, 14, 15, 17, 18, 20, 34], Kari North (University of North Carolina Chapel Hill)[1, 3, 9, 10, 13, 15, 17, 34], Ulrike Peters (Fred Hutchinson Cancer Research Center)[1, 3, 11, 12, 13, 15, 17, 18, 34], Christopher S. Carlson (Fred Hutchinson Cancer Research Center, University of Washington)[1, 2, 3, 5, 14, 12, 13, 15, 16, 17, 18, 19, 34], Garnet Anderson (Fred Hutchinson Cancer Research Center)[34], Hoda Anton-Culver (University of California at Irvine)[34], Themistocles L. Assimes (Stanford University School of Medicine)[5, 9, 11, 34], Paul L. Auer (Fred Hutchinson Cancer Research Center)[1, 2, 3, 5, 11, 12, 13, 15, 16, 18, 34], Shirley Beresford (Fred Hutchinson Cancer Research Center)[34], Chris Bizon (University of North Carolina Chapel Hill)[3, 9, 13, 15, 23, 34], Henry

Black (Rush Medical Center)[34], Robert Brunner (University of Nevada)[34], Robert Brzyski (University of Texas Health Science Center San Antonio)[34], Dale Burwen (National Heart, Lung, and Blood Institute WHI Project Office)[34], Bette Caan (Kaiser Permanente Division of Research, Oakland, CA)[34], Cara L. Carty (Fred Hutchinson Cancer Research Center)[18, 34], Rowan Chlebowski (Los Angeles Biomedical Research Institute)[34], Steven Cummings (University of California at San Francisco)[34], J. David Curb (University of Hawaii)[9, 18, 34], Charles B. Eaton (Brown University, Memorial Hospital of Rhode Island)[12, 34], Leslie Ford (National Heart, Lung, and Blood Institute, National Heart, Lung, and Blood Institute WHI Project Office)[34], Nora Franceschini (University of North Carolina Chapel Hill)[2, 3, 9, 10, 15, 34], Stephanie M. Fullerton (University of Washington)[6, 11, 34], Margery Gass (University of Cincinnati)[34], Nancy Geller (National Heart, Lung, and Blood Institute WHI Project Office)[34], Gerardo Heiss (University of North Carolina Chapel Hill)[5, 34], Barbara V. Howard (Howard University, MedStar Research Institute)34, Li Hsu (Fred Hutchinson Cancer Research Center)[1, 13, 15, 18, 34], Carolyn M. Hutter (Fred Hutchinson Cancer Research Center)[13, 15, 18, 34], John Ioannidis (Stanford University School of Medicine)[11, 34], Shuo Jiao (Fred Hutchinson Cancer Research Center)[34], Karen C. Johnson (University of Tennessee Health Science Center)[3, 34], Emond Kabagambe (University of Alabama at Birmingham)[34], Charles Kooperberg (Fred Hutchinson Cancer Research Center)[1, 5, 9, 14, 13, 15, 17, 18, 34], Lewis Kuller (University of Pittsburgh)[34], Andrea LaCroix (Fred Hutchinson Cancer Research Center)[34], Kamakshi Lakshminarayan (University of Minnesota)[18, 34], Dorothy Lane (State University of New York at Stony Brook)[34], Ethan M. Lange (University of North Carolina Chapel Hill)[3, 5, 9, 13, 34], Leslie A. Lange (University of North Carolina Chapel Hill)[1, 2, 3, 5, 9, 12, 13, 15, 17, 18, 20, 25, 34], Norman Lasser (University of Medicine and Dentistry of New Jersey)[34], Erin LeBlanc (Kaiser Permanente Center for Health Research, Portland, OR)[34], Cora E. Lewis (University of Alabama at Birmingham)[34], Marian Limacher (University of Florida)[34], Danyu Lin (University of North Carolina Chapel Hill)[1, 3, 9, 13, 15, 34], Benjamin A. Logsdon (Fred Hutchinson Cancer Research Center)[2, 34], Shari Ludlam (National Heart, Lung, and Blood Institute WHI Project Office)[34], JoAnn E. Manson (Brigham and Women's Hospital, Harvard School of Public Health)[34], Karen Margolis (University of Minnesota)[34], Lisa Martin (George Washington University Medical Center)[9, 34], Joan McGowan (National Heart, Lung, and Blood Institute WHI Project Office)[34], Keri L. Monda (Amgen, Inc.)[1, 15, 34], Jane Morley Kotchen (Medical College of Wisconsin)[34], Lauren Nathan (University of California Los Angeles)[34], Judith Ockene (Fallon Clinic, University of Massachusetts)[34], Mary Jo O'Sullivan (University of Miami)[34], Lawrence S. Phillips (Emory University)[34], Ross L. Prentice (Fred Hutchinson Cancer Research Center)[34], Alexander P. Reiner (Fred Hutchinson Cancer Research Center, University of Washington)[1, 2, 3, 5, 9, 11, 12, 13, 14, 15, 20, 25, 34], John Robbins (University of California at Davis)[34], Jennifer G. Robinson (University of Iowa)[9, 11, 18, 34], Jacques E. Rossouw (National Heart, Lung, and Blood Institute, National Heart, Lung, and Blood Institute WHI Project Office)[5, 14, 17, 20, 34], Haleh Sangi-Haghpeykar (Baylor College of Medicine)[34], Gloria E. Sarto (University of Wisconsin)[34], Sally Shumaker (Wake Forest University)[34], Michael S. Simon (Wayne State University)[34], Marcia L. Stefanick (Stanford University School of Medicine)[34], Evan Stein (Medical Research Labs)[34], Hua Tang (Stanford University)[2, 34], Kira C. Taylor (University of Louisville)[1, 3, 13, 15, 20, 34], Cynthia A. Thomson (University of Arizona)[34], Timothy A. Thornton (University of Washington)[13, 15, 18, 34],

Linda Van Horn (Northwestern University)[34], Mara Vitolins (Wake Forest University)[34], Jean Wactawski-Wende (University of Buffalo)[34], Robert Wallace (University of Iowa)[2, 34], Sylvia Wassertheil-Smoller (Boston University)[18, 34]

NHLBI

Deborah Applebaum-Bowden (National Heart, Lung, and Blood Institute)[4, 7, 12, 17], Michael Feolo (National Center for Biotechnology Information)[12], Weiniu Gan (National Heart, Lung, and Blood Institute)[7, 8, 16, 17], W. Keith Hoots (National Heart, Lung, and Blood Institute)[17], James Kiley (National Heart, Lung, and Blood Institute)[17], Michael Lauer (National Heart, Lung, and Blood Institute)[17], Hilary Leeds (National Heart, Lung, and Blood Institute), Alan Michelson (National Heart, Lung, and Blood Institute)[17], Dina N. Paltoo (National Heart, Lung, and Blood Institute)[4, 6, 11, 17], Phyliss Sholinsky (National Heart, Lung, and Blood Institute)[4, 17], Sonia Skarlatos (National Heart, Lung, and Blood Institute)[17], Anne Sturcke (National Center for Biotechnology Information)[12]

ESP Groups

[1]Anthropometry Project Team, [2]Blood Count/Hematology Project Team, [3]Blood Pressure Project Team, [4]Data Flow Working Group, [5]Early MI Project Team, [6]ELSI Working Group, [7]Executive Committee, [8]Family Study Project Team, [9]Lipids Project Team, [10]Lung Project Team, [11]Personal Genomics Project Team, [12]Phenotype and Harmonization Working Group, [13]Population Genetics and Statistical Analysis Working Group, [14]Publications and Presentations Working Group, [15]Quantitative Analysis Ad Hoc Task Group, [16]Sequencing and Genotyping Working Group, [17]Steering Committee, [18]Stroke Project Team, [19]Structural Variation Working Group, [20]Subclinical/Quantitative Project Team

ESP Cohorts

[21]Acute Lung Injury (ALI), [22]Atherosclerosis Risk in Communities (ARIC), [23]Cardiovascular Health Study (CHS), [24]Chronic Obstructive Pulmonary Disease (COPDGene), [25]Coronary Artery Risk Development in Young Adults (CARDIA), [26]Cystic Fibrosis (CF), [27]Early Pseudomonas Infection Control (EPIC), [28]Framingham Heart Study (FHS), [29]Jackson Heart Study (JHS), [30]Lung Health Study (LHS), [31]Multi-Ethnic Study of Atherosclerosis (MESA), [32]Pulmonary Arterial Hypertension (PAH), [33]Severe Asthma Research Program (SARP), [34]Women's Health Initiative (WHI)

Supplementary tables

Table S1. Comparison of the number of SNVs with different minor allele frequency (MAF) in individuals of European ancestry between ESP and 1KGP.

MAF	Exome Project	1000 Genomes Project
Singletons	161,906 (62.0%)	28,696 (31.5%)
≤0.001	186,602 (71.4%)	Not Available
≤0.005	220,086 (84.2%)	40,677 (44.7%)
≤0.01	227,364 (87.0%)	49,540 (54.5%)
≤0.05	239,824 (91.8%)	64,191 (70.6%)
>0.05	21,427 (8.2%)	26,770 (29.4%)
Total	261,251	90,961

Table S2. The distribution of MAF for the 151 common SNPs in different European ancestral populations of 1KGP, but invariant in ESP.

MAF (%)	CEU	FIN	GBR	IBS	TSI
Invariant	105 (69.5%)	0	0	16 (10.6%)	107 (70.9%)
≤0.01	3 (2.0%)	0	0	0	4 (2.6%)
≤0.05	17 (11.2%)	6 (4.0%)	5 (3.3%)	12 (7.9%)	23 (15.2%)
>0.05	29 (19.2%)	145 (96.0%)	146 (96.7%)	123 (81.4%)	21 (13.9%)

Table S3. Genes with unusually high proportions of rare variation.

Gene	Chrom	Pi	Tajima's D	S	CommonS	RareS	Length
SAMD11	1	1.08E-04	-2.42	49	7	42	743
KLHL17	1	3.84E-04	-2.35	67	6	61	1322
PLEKHN1	1	6.54E-04	-2.04	53	8	45	966
SCNN1D	1	2.31E-04	-2.45	67	7	60	1003
TAS1R3	1	8.04E-05	-2.51	62	3	59	1154
ATAD3B	1	1.02E-03	-2.09	83	15	68	1434
PRAMEF2	1	3.19E-03	-1.60	96	31	65	1324
TMEM82	1	6.01E-04	-2.22	47	7	40	670
CROCC	1	1.10E-04	-2.52	100	14	86	2111
CELA3B	1	1.52E-03	-2.01	52	12	40	788
GJB4	1	7.49E-04	-2.25	48	11	37	765
HRNR	1	9.62E-04	-2.17	311	53	258	4503
FLG	1	1.40E-03	-2.29	761	148	613	11166
OR2T12	1	1.44E-03	-1.94	50	11	39	820
PRR21	2	1.29E-03	-1.75	43	14	29	514

ANO7	2	6.00E-04	-2.27	98	12	86	2030
TMEM175	4	6.84E-04	-2.29	72	8	64	1424
KIAA1530	4	7.76E-04	-2.20	85	8	77	1791
CRIPAK	4	2.73E-03	-2.17	150	30	120	1299
DOK7	4	1.01E-04	-2.24	27	2	25	331
WFS1	4	1.51E-03	-2.21	175	25	150	2260
SH3TC1	4	4.74E-04	-2.42	145	21	124	2822
GPR78	4	6.49E-04	-2.00	35	5	30	570
CPZ	4	6.91E-04	-2.37	103	13	90	1537
ZDHHC11	5	6.12E-04	-2.30	62	8	54	1090
SRA1	5	4.39E-04	-2.44	58	11	47	499
HIST1H3A	6	3.82E-04	-2.23	28	3	25	378
HIST1H4B	6	1.61E-03	-2.03	30	2	28	315
HIST1H2AB	6	1.77E-03	-1.79	26	2	24	397
HIST1H1C	6	6.90E-04	-2.19	37	5	32	389
HIST1H1T	6	1.78E-03	-2.01	47	5	42	628
HIST1H1D	6	8.81E-04	-2.14	39	2	37	604
MUC21	6	1.95E-03	-2.10	126	40	86	1703
HLA-DQA2	6	2.85E-03	-1.70	56	15	41	781
MICALL2	7	8.04E-04	-2.04	83	19	64	1408
AMZ1	7	1.31E-04	-2.56	74	5	69	905
SDK1	7	7.71E-04	-2.33	275	41	234	6141
SLC29A4	7	1.49E-04	-2.47	58	2	56	1213
CD36	7	1.06E-04	-2.55	67	4	63	1464
MUC17	7	9.12E-04	-2.43	753	126	627	13448
ERICH1	8	6.94E-04	-2.28	66	8	58	1184
MYOM2	8	1.11E-03	-2.33	260	47	213	4529
AMAC1L2	8	2.03E-03	-1.91	66	13	53	1013
CTSB	8	8.78E-04	-2.19	55	7	48	1055
TSNARE1	8	1.04E-03	-2.05	66	8	58	1281
IFNA14	9	9.54E-05	-2.42	39	1	38	573
FCN1	9	1.36E-03	-2.00	54	5	49	1015
FBXW5	9	7.23E-05	-2.47	50	3	47	775
SLC34A3	9	3.30E-04	-2.29	52	3	49	952
FAM166A	9	3.14E-04	-2.39	49	11	38	784
PPYR1	10	3.16E-04	-2.47	66	7	59	1119
MUPCDH	11	8.35E-04	-2.12	90	14	76	1857
OR4A5	11	5.71E-04	-2.28	48	7	41	794
OR4C46	11	2.56E-03	-1.82	67	18	49	844
OR4C16	11	2.75E-03	-1.63	59	15	44	926
KRT82	12	5.10E-04	-2.34	68	6	62	1426
FAM70B	13	5.80E-04	-2.28	50	6	44	903
DHRS2	14	1.52E-04	-2.46	48	4	44	833
DHRS4L2	14	2.01E-03	-1.80	44	12	32	626
OR4M2	15	4.75E-04	-2.41	60	12	48	916
C16ORF38	16	3.12E-03	-1.05	63	19	44	961
NUBP2	16	1.50E-03	-1.91	45	6	39	734
RPL3L	16	2.80E-04	-2.46	62	6	56	1134
PKD1	16	1.32E-04	-2.45	146	15	131	2723

SRRM2	16	3.27E-04	-2.58	313	22	291	7866
DNASE1	16	9.90E-04	-2.28	63	7	56	878
TRAP1	16	6.01E-04	-2.39	103	13	90	2038
PPL	16	7.45E-04	-2.29	198	24	174	4644
FAM86A	16	6.26E-04	-2.10	38	7	31	571
TEKT5	16	9.62E-04	-2.21	76	12	64	1471
PDILT	16	1.45E-03	-1.96	84	10	74	1799
NECAB2	16	1.25E-03	-2.13	70	9	61	978
SLC38A8	16	9.65E-04	-2.28	81	12	69	1344
KIAA1609	16	1.65E-03	-1.96	76	16	60	1399
KLHDC4	16	9.44E-04	-2.16	73	12	61	1415
CDT1	16	1.03E-03	-1.74	47	8	39	757
KCNJ12	17	2.50E-04	-2.49	66	9	57	1187
ACTG1	17	1.03E-03	-2.03	50	8	42	885
THEG	19	5.27E-04	-2.36	61	9	52	1171
POLRMT	19	4.30E-04	-2.18	79	13	66	1535
PRSSL1	19	7.21E-04	-1.78	24	3	21	309
TJP3	19	1.09E-03	-1.98	103	13	90	2072
CAPS	19	9.26E-04	-2.19	41	7	34	558
PPAN-P2RY11	19	3.46E-04	-2.40	83	6	77	1596
P2RY11	19	6.59E-04	-2.26	55	4	51	837
CYP2A6	19	1.13E-03	-2.24	92	22	70	1497
CYP2B6	19	6.48E-04	-2.40	88	6	82	1483
PSG8	19	2.33E-03	-1.96	96	14	82	1264
PSG1	19	1.94E-03	-2.01	88	18	70	1236
PSG6	19	1.11E-03	-2.38	108	18	90	1294
PSG11	19	2.28E-03	-1.76	61	12	49	1008
PSG2	19	2.03E-03	-1.92	67	14	53	1028
PSG9	19	1.39E-03	-2.14	81	20	61	1303
LILRA2	19	9.31E-04	-2.31	87	24	63	1415
TRIB3	20	1.00E-03	-2.06	50	6	44	946
LAMA5	20	4.05E-04	-2.38	270	38	232	5428
TPTE	21	3.00E-04	-2.52	87	13	74	1651
TRPM2	21	2.71E-04	-2.58	178	17	161	4007
C21ORF29	21	1.23E-04	-2.57	84	5	79	1776
COL6A2	21	8.75E-04	-2.20	136	18	118	2541
IGLL1	22	1.35E-03	-1.97	38	12	26	443
IGLL3	22	3.90E-03	-1.75	68	22	46	640
CYP2D6	22	7.50E-04	-2.18	58	6	52	810
SERHL2	22	5.54E-04	-2.16	39	5	34	537
LOC553158	22	2.10E-03	-2.12	117	18	99	1402
ARHGAP8	22	2.20E-03	-2.11	115	18	97	1340
LMF2	22	5.64E-04	-2.19	65	6	59	898
PLCXD1	X	5.18E-04	-2.19	58	5	53	995
IL3RA	X	2.10E-04	-2.38	57	3	54	1181
VCX3A	X	5.49E-04	-2.00	23	7	16	171

Column headers are Gene ID, chromosome, nucleotide diversity, Tajima's D, number of segregating sites (S), number of common variants (CommonS), number of rare variants (RareS), and observed gene length.

Table S4. Summary information of 114 genes with signatures of positive selection.

Gene	Chr	F_{ST}	Pi (%)	AA Pi (%)	EA Pi (%)	Nsyn Pi (%)	Syn Pi (%)	Extreme Pattern
C1ORF63	1	0.16	0.049	0.017	0.068	0.065	0.001	LowAA
EIF2C1	1	0.36	0.012	0.021	0.002	0	0.049	HighFst
CDC14A	1	0.34	0.033	0.062	0.003	0.003	0.125	LowEA
BCL9	1	0.22	0.016	0.007	0.021	0.016	0.016	LowAA
OR6K6	1	0.26	0.275	0.505	0.028	0.26	0.32	LowEA
DARC	1	0.25	0.065	0.024	0.086	0.085	0.005	LowAA
KLHL20	1	0.31	0.025	0.046	0.002	0.001	0.097	HighFst
LAMC2	1	0.21	0.069	0.085	0.049	0.025	0.203	MK
PRG4	1	0.17	0.046	0.05	0.04	0.029	0.098	LowPi
LRRN2	1	0.27	0.065	0.04	0.079	0.036	0.154	LowAA
SLC41A1	1	0.22	0.046	0.021	0.057	0.001	0.182	LowAA
C1ORF116	1	0.36	0.079	0.147	0.004	0.099	0.018	LowEA
C1ORF101	1	0.21	0.063	0.098	0.025	0.016	0.206	MK
OR2W5	1	0.17	0.164	0.196	0.131	0.045	0.519	MK
OR2T10	1	0.02	0.015	0.025	0.008	0.012	0.025	LowPi
NT5C1B	2	0.26	0.066	0.072	0.045	0.003	0.245	MK
MAP4K3	2	0.45	0.04	0.04	0.021	0.002	0.154	MK
STK17B	2	0.28	0.087	0.072	0.087	0.001	0.346	MK
ARMC9	2	0.08	0.077	0.078	0.074	0.006	0.289	MK
DHFRL1	3	0.11	0.091	0.036	0.128	0.12	0.006	LowAA
UGT2B28	4	0.03	0.015	0.023	0.008	0.009	0.032	LowPi
UGT2A1	4	0.08	0.07	0.04	0.092	0.081	0.037	LowAA
BRD9	5	0.20	0.072	0.133	0.013	0.012	0.252	LowEA
UGT3A1	5	0.04	0.008	0.016	0.001	0.003	0.021	LowPi
C5ORF36	5	0.36	0.03	0.052	0.001	0.04	0	HighFst
C5ORF48	5	0.40	0.07	0.116	0.002	0.003	0.272	HighFst
LARS	5	0.15	0.075	0.037	0.096	0.024	0.228	LowAA
RARS	5	0.21	0.074	0.029	0.102	0.016	0.248	LowAA
FBXW11	5	0.35	0.022	0.04	0.001	0	0.087	HighFst
FAM153A	5	0.08	0.074	0.15	0.01	0.099	0.001	LowPi
HLA-G	6	0.11	0.352	0.379	0.326	0.068	1.205	MK
WDR46	6	0.07	0.032	0.015	0.044	0.023	0.059	LowAA
TMEM63B	6	0.05	0.02	0.01	0.028	0.007	0.061	LowAA
FUT9	6	0.32	0.023	0.042	0.001	0.03	0.002	HighFst
OR2A4	6	0.00	0	0	0.001	0	0.001	LowPi
C7ORF20	7	0.11	0.115	0.142	0.089	0.001	0.455	MK
MUC17	7	0.17	0.092	0.121	0.063	0.089	0.101	LowPi
TAS2R60	7	0.04	0.059	0.064	0.054	0.011	0.204	MK
SCARA3	8	0.01	0.003	0.005	0.001	0.002	0.006	LowPi
UNC5D	8	0.40	0.04	0.065	0.006	0.002	0.152	LowEA
PTK2	8	0.42	0.012	0.02	0.002	0.001	0.046	HighFst
C9ORF123	9	0.03	0.17	0.16	0.174	0.002	0.672	MK
CDC14B	9	0.41	0.025	0.043	0.002	0.005	0.086	HighFst
ZNF510	9	0.21	0.067	0.131	0.007	0.058	0.095	LowEA
ORM2	9	0.42	0.173	0.312	0.008	0.109	0.364	LowEA
OR1L3	9	0.39	0.086	0.162	0.007	0.102	0.039	LowEA

NMT2	10	0.30	0.021	0.039	0.002	0.001	0.081	HighFst
ZNF248	10	0.05	0.025	0.01	0.035	0.001	0.095	LowAA
CYP17A1	10	0.01	0.065	0.062	0.066	0	0.258	MK
PNLIPRP3	10	0.21	0.081	0.14	0.023	0.07	0.113	LowEA
OR51F2	11	0.11	0.054	0.066	0.04	0.03	0.126	MK
OR56A4	11	0.23	0.111	0.092	0.116	0.043	0.316	MK
KCNJ11	11	0.23	0.106	0.052	0.133	0.094	0.142	LowAA
PACSIN3	11	0.32	0.024	0.043	0.001	0.03	0.005	HighFst
FAM111A	11	0.05	0.034	0.042	0.026	0.013	0.099	MK
VWCE	11	0.21	0.021	0.043	0.002	0.005	0.07	MK
GPR152	11	0.47	0.139	0.227	0.005	0.047	0.414	LowEA
MLL	11	0.07	0.005	0.006	0.004	0.001	0.017	LowPi
C11ORF63	11	0.11	0.025	0.028	0.02	0.006	0.081	MK
OR6T1	11	0.14	0.146	0.126	0.146	0.032	0.486	MK
TAS2R20	12	0.16	0.291	0.154	0.36	0.383	0.016	LowAA
SLCO1B1	12	0.30	0.123	0.115	0.112	0.069	0.279	MK
ZNF641	12	0.20	0.054	0.023	0.069	0.036	0.108	LowAA
AMHR2	12	0.49	0.021	0.032	0.001	0.001	0.083	HighFst
OR6C70	12	0.16	0.118	0.089	0.131	0.107	0.149	LowPi
OR6C4	12	0.31	0.081	0.151	0.007	0.052	0.167	LowEA
POLR3B	12	0.13	0.022	0.012	0.029	0.004	0.035	LowAA
HNF1A	12	0.11	0.065	0.037	0.082	0.033	0.163	LowAA
DIABLO	12	0.32	0.035	0.062	0.002	0.002	0.133	HighFst
IRS2	13	0.00	0.001	0.002	0.001	0.002	0	LowPi
PACS2	14	0.42	0.049	0.081	0.004	0.021	0.133	LowEA
PGBD4	15	0.07	0.007	0.015	0.001	0.005	0.014	LowPi
SLC24A5	15	0.77	0.032	0.027	0.002	0.042	0.002	HighFst
CD19	16	0.16	0.062	0.031	0.079	0.042	0.121	LowAA
KIF22	16	0.28	0.034	0.065	0.002	0.009	0.109	LowEA
ABCC12	16	0.47	0.078	0.121	0.026	0.064	0.122	LowEA
CES2	16	0.13	0.007	0.013	0.002	0.001	0.024	LowPi
LRRC36	16	0.43	0.064	0.097	0.014	0.058	0.085	LowEA
CDRT15	17	0.06	0.058	0.048	0.063	0.062	0.046	LowPi
TBC1D29	17	0.11	0.062	0.075	0.049	0.003	0.17	LowPi
WNK4	17	0.24	0.029	0.054	0.003	0.018	0.061	LowEA
CNTD1	17	0.35	0.049	0.096	0.001	0.046	0.061	HighFst
IMP5	17	0.12	0.195	0.113	0.244	0.171	0.27	LowAA
MAPT	17	0.12	0.125	0.066	0.163	0.005	0.181	LowAA
KIAA1267	17	0.12	0.091	0.044	0.121	0.048	0.22	LowAA
C17ORF64	17	0.61	0.357	0.492	0.05	0.097	1.138	MK, LowEA
APPBP2	17	0.41	0.026	0.044	0.002	0.022	0.036	HighFst
LRRC37A3	17	0.17	0.014	0.029	0.002	0.017	0.006	LowPi
BPTF	17	0.36	0.011	0.02	0.002	0.004	0.029	HighFst; LowEA
GPS1	17	0.02	0.006	0.007	0.006	0.001	0.021	LowPi
BRUNOL4	18	0.06	0.038	0.03	0.044	0.001	0.151	MK
FECH	18	0.17	0.074	0.037	0.094	0.026	0.215	LowAA
OR7G2	19	0.10	0.138	0.115	0.146	0.134	0.15	LowPi
ICAM5	19	0.20	0.084	0.045	0.102	0.088	0.071	LowAA

RGL3	19	0.01	0.003	0.006	0.002	0.003	0.003	LowPi
ZNF442	19	0.20	0.056	0.103	0.01	0.065	0.028	LowEA
USHBP1	19	0.24	0.102	0.108	0.083	0.066	0.207	MK
HIPK4	19	0.03	0.014	0.016	0.013	0.016	0.01	LowPi
LIPE	19	0.23	0.03	0.058	0.005	0.033	0.016	LowEA
VRK3	19	0.19	0.114	0.217	0.012	0.093	0.176	LowEA
ZNF473	19	0.20	0.074	0.145	0.004	0.067	0.095	LowEA
POLD1	19	0.34	0.061	0.108	0.015	0.008	0.221	LowEA
KIR2DL1	19	0.04	0.028	0.049	0.011	0.034	0.011	LowPi
KIR2DS4	19	0.04	0.021	0.042	0.004	0	0	LowPi
NLRP11	19	0.19	0.06	0.068	0.05	0.026	0.163	MK
DEFB127	20	0.17	0.232	0.113	0.292	0.306	0.009	LowAA
C20ORF70	20	0.26	0.096	0.182	0.012	0.093	0.104	LowEA
R3HDML	20	0.04	0.118	0.106	0.128	0.005	0.458	MK
TPTE	21	0.07	0.031	0.045	0.018	0.028	0.015	LowPi
SLC5A1	22	0.03	0.033	0.016	0.046	0.019	0.076	LowAA
APOL1	22	0.32	0.155	0.141	0.143	0.152	0.162	LowPi
TXN2	22	0.01	0.006	0.007	0.005	0.001	0.02	LowPi
APOBEC3G	22	0.33	0.082	0.085	0.059	0.058	0.152	MK
TRMU	22	0.23	0.047	0.094	0.003	0.02	0.128	LowEA

Column headers are Gene ID, chromosome, F_{ST}, nucleotide diversity in all ESP samples, nucleotide diversity in the AA sample, nucleotide diversity in the EA sample, nucleotide diversity of nonsynonymous variants, nucleotide diversity of synonymous variants, and the characteristic that makes them an outlier (Extreme Patter = MK for McDonald-Kreitman; LowPi = low nucleotide diversity in the combined sample; LowEA or LowAA = low nucleotide diversity in EA or AA, respectively; HighFst = extreme levels of differentiation as assessed by F_{ST}).

Table S5. Significantly enriched KEGG pathways among the 114 genes with signatures of positive selection.

KEGG Pathway	Number of Genes	Adjusted P-value
Olfactory transduction	10	1.79×10^{-7}
Drug Metabolism -other enzymes	3	5×10^{-4}
Antigen processing and presentation	3	2.2×10^{-3}
Natural killer cell mediated cytotoxicity	3	4.3×10^{-3}
Metabolic pathways	8	4.3×10^{-2}

Column headers are KEGG pathway ID, Number of putatively selected genes in the pathway, and FDR adjusted p-value. Enrichment analysis was performed using the web-based tool WebGestalt(*13*) (http://bioinfo.vanderbilt.edu/webgestalt)

Table S6. Summary statistics of polymorphism for each of the 15,585 genes analyzed.

This is a large table that is available as a separate file. Column headers are:

Gene (gene name)
Chrom (chromosome)
Start (start site)
End (end site)
S (number of segregating sites)
EA_S (number of segregating sites in European Americans)
AA_S (number of segregating sites in African Americans)
Length (real exonic length of the gene)
ObservedLength (number of sites that passed all filters, including invariant sites)
Pi (mean number of pairwise differences between two randomly selected chromosomes, divided by ObservedLength)
EA_Pi (Pi in European Americans)
AA_Pi (Pi in African Americans)
Rare (number of segregating sites with frequency under 0.5%)
Common (number of segregating sites with frequency at least 0.5%)
Missense (number of missense segregating sites)
PolyPhen2 (number of segregating sites with "probably damaging" PolyPhen2 designation)
HighGERP (number of segregating sites with GERP score >= 5)
HighSFS (number of segregating sites with >=90% probability of functional importance using SFS-based method)
Nons (number of nonsense segregating sites)
Splice (number of splice segregating sites)
MissenseEA (number of missense segregating sites in European Americans)
NonsEA (number of nonsense segregating sites in European Americans)
SpliceEA (number of splice segregating sites in European Americans)
MissenseAA (number of missense segregating sites in African Americans)
NonsAA (number of nonsense segregating sites in African Americans)
SpliceAA (number of splice segregating sites in African Americans)
Sprop (segregating sites divided by observed length)
NS_Sprop (missense, nonsense, and splice segregating sites divided by observed length)
SpropEA (segregating sites in European Americans divided by observed length)
NS_SpropEA (missense, nonsense, and splice segregating sites in European Americans divided by observed length)
SpropAA (segregating sites in African Americans divided by observed length)
NS_SpropAA (missense, nonsense, and splice segregating sites in African Americans divided by observed length)

Table S7. Evidence for coordinated purifying selection on synonymous and nonsynonymous variation within genes.

Method	AA	EA
SIFT	-0.031 (0.0002)	-0.020 (0.0194)
Polyphen2	-0.032 (0.0001)	-0.018 (0.0070)
LRT	0.010 (0.1247)	-0.001 (0.4590)
MutationTaster	-0.020 (0.0098)	-0.026 (0.0013)

Values show the rank correlation coefficients between average functional prediction scores for different methods and average w score changes of a gene. P-values from a two-sided z test are shown in parantheses.

Supplementary figures

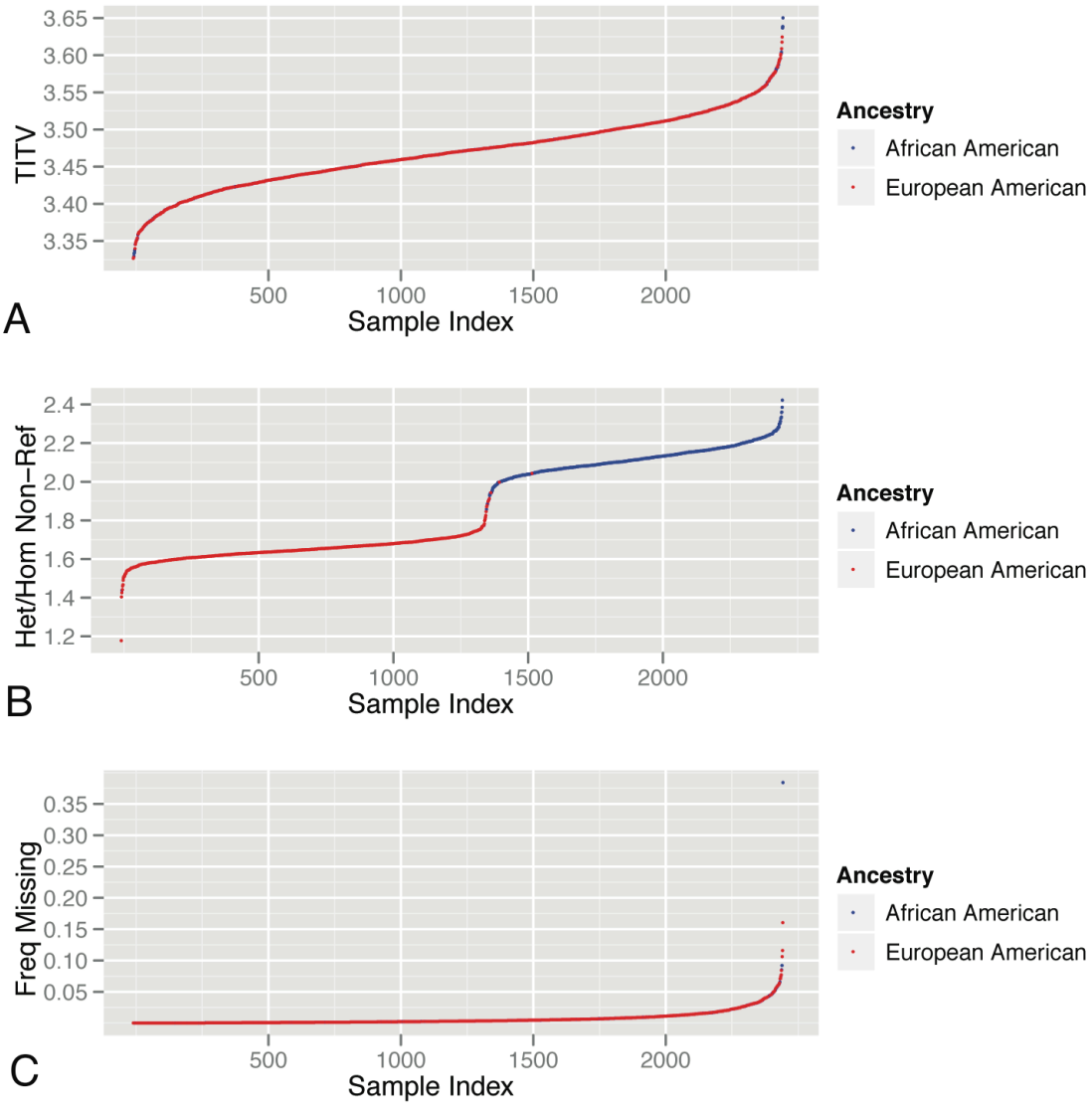

Figure S1. Assessment of Sample Outliers using Ti/Tv, Heterozygosity and Missingness (i.e., sites not called). A) Ti/Tv and B) Het/Hom Non-Ref and C) Frequency of Missingness vs. Sample Index. Each dot represents a single sample. Samples showed similar properties of Ti/Tv, heterozygosity and frequency of missing genotypes.

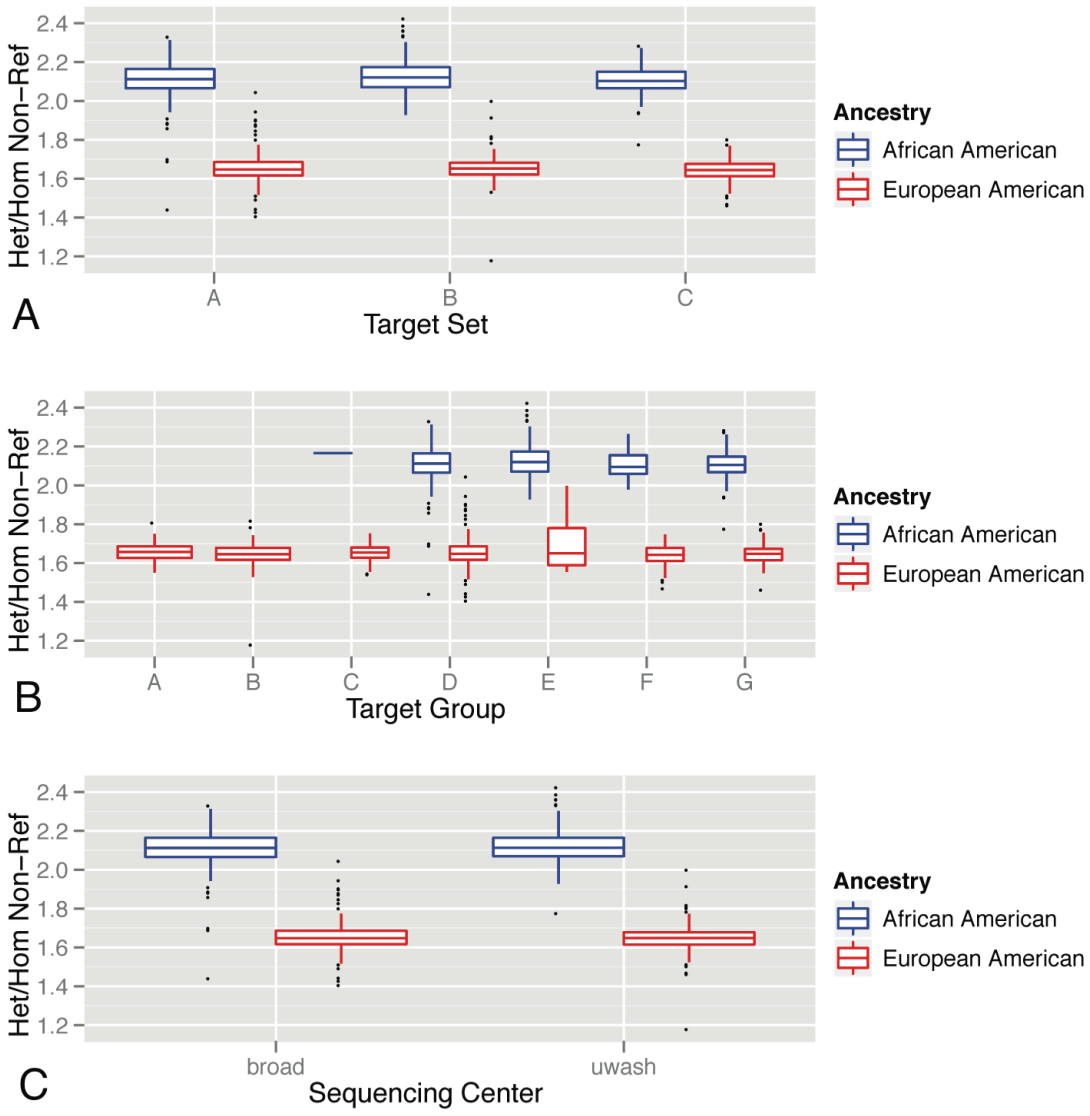

Figure S2. Assessment of Batch Effects between Samples Examining Heterozygosity. A) target definition and B) target Group and C) sequencing center vs. Het / Hom Non-Ref.

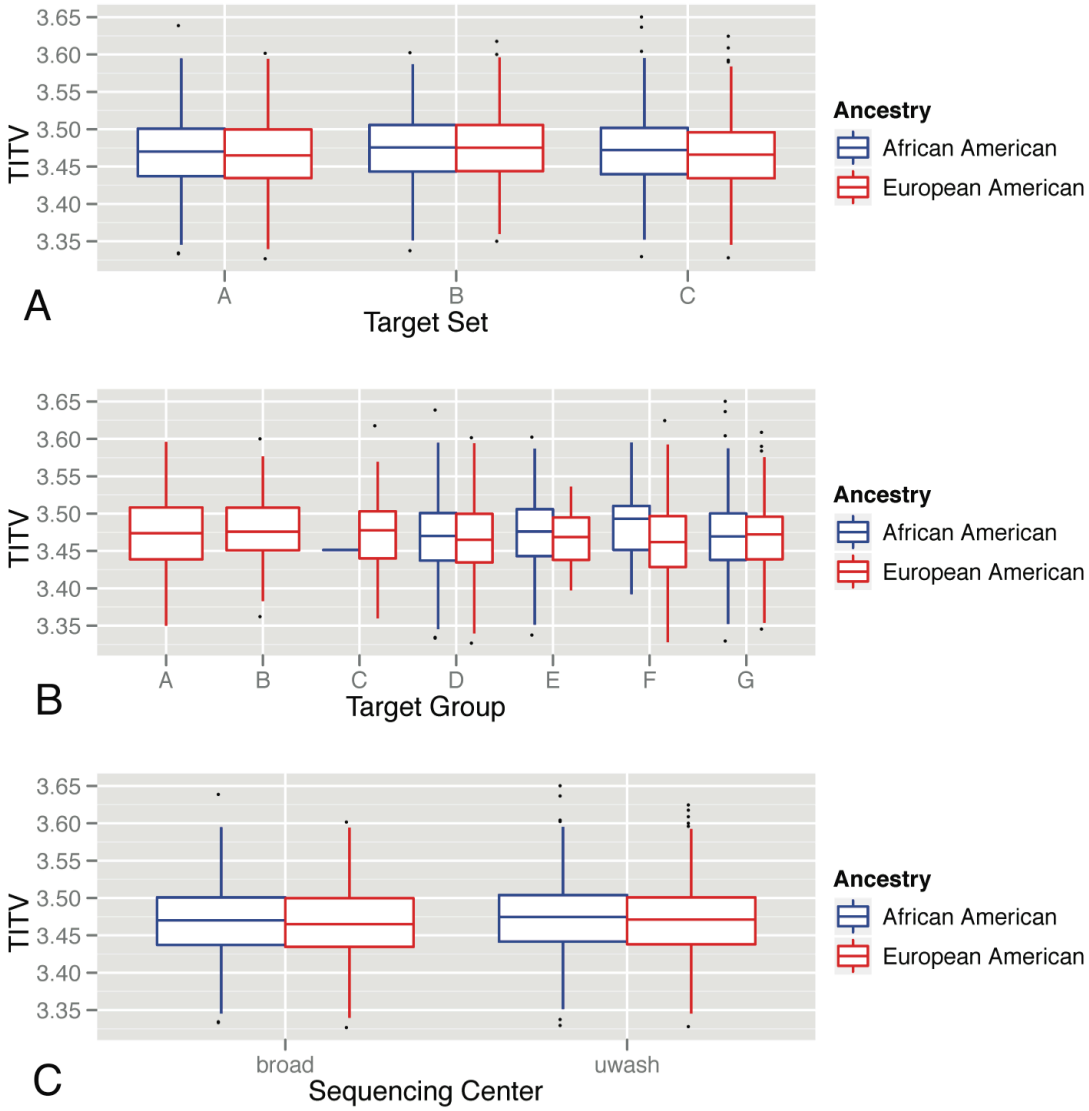

Figure S3. Assessment of Batch Effects between Samples using Ti/Tv. A) Target definition. B) Target group. C) Sequencing center vs. Ti/Tv.

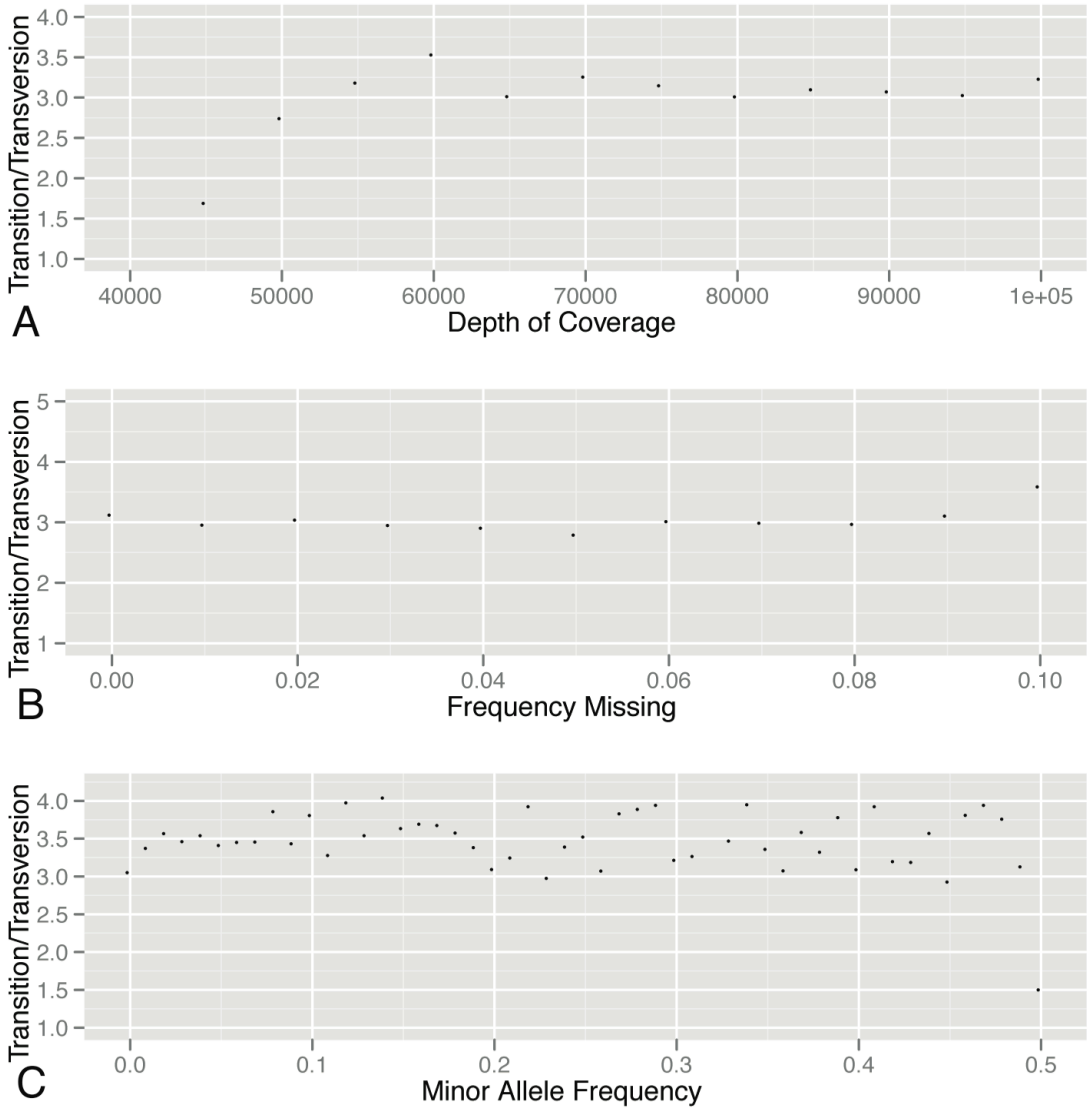

Figure S4. Variant QC metrics as a function of TITV. A) depth of coverage and B) frequency missing and C) minor allele frequency. Ti/Tv is relatively constant over a broad range of coverage. Ti/Tv is similar across frequencies missing < 10% and across all minor allele frequencies.

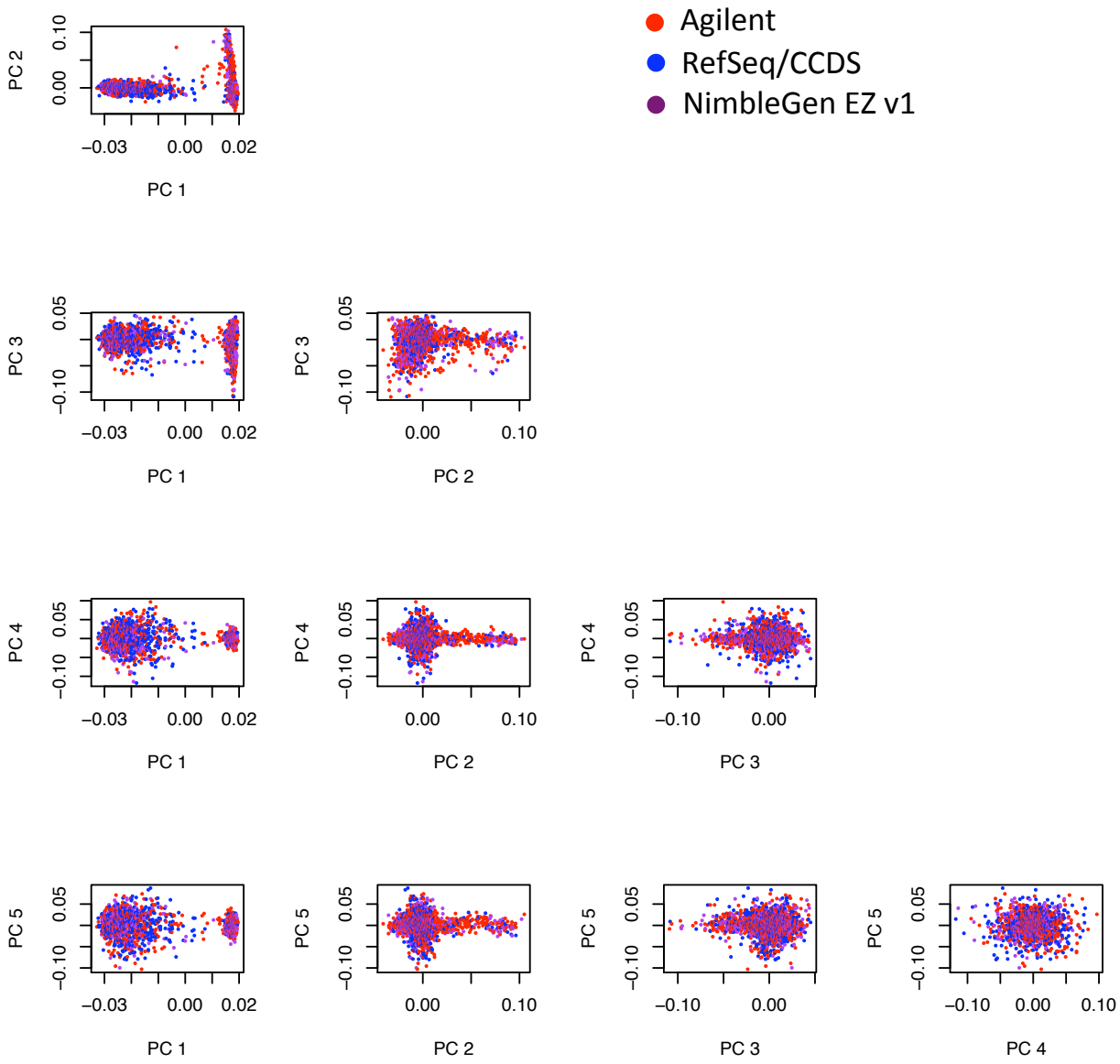

Figure S5. PCA plots with samples labeled by exome capture target.

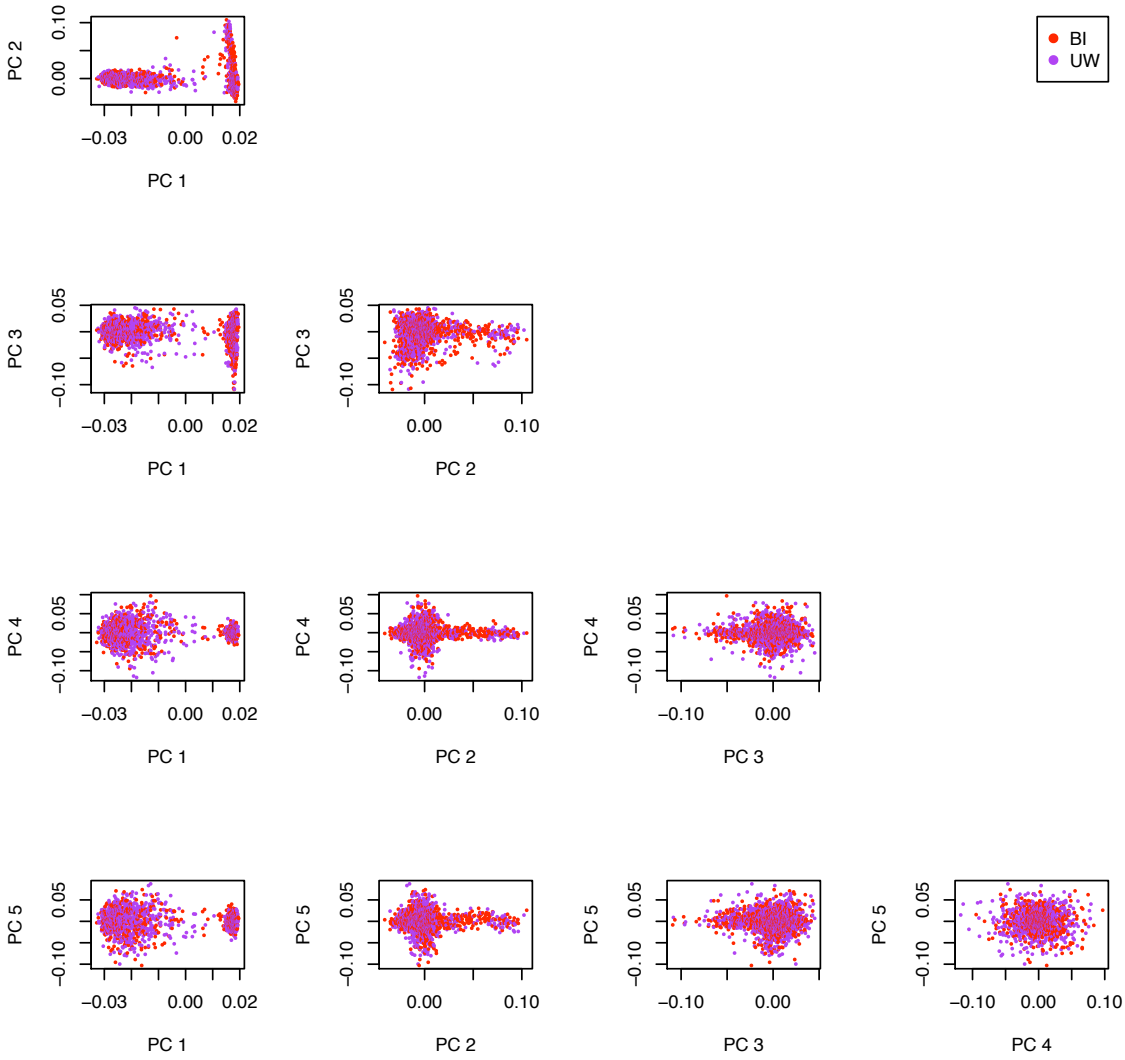

Figure S6. PCA plots with samples labeled by sequencing center. BI and WI denote the Broad Institute and University of Washington, respectively.

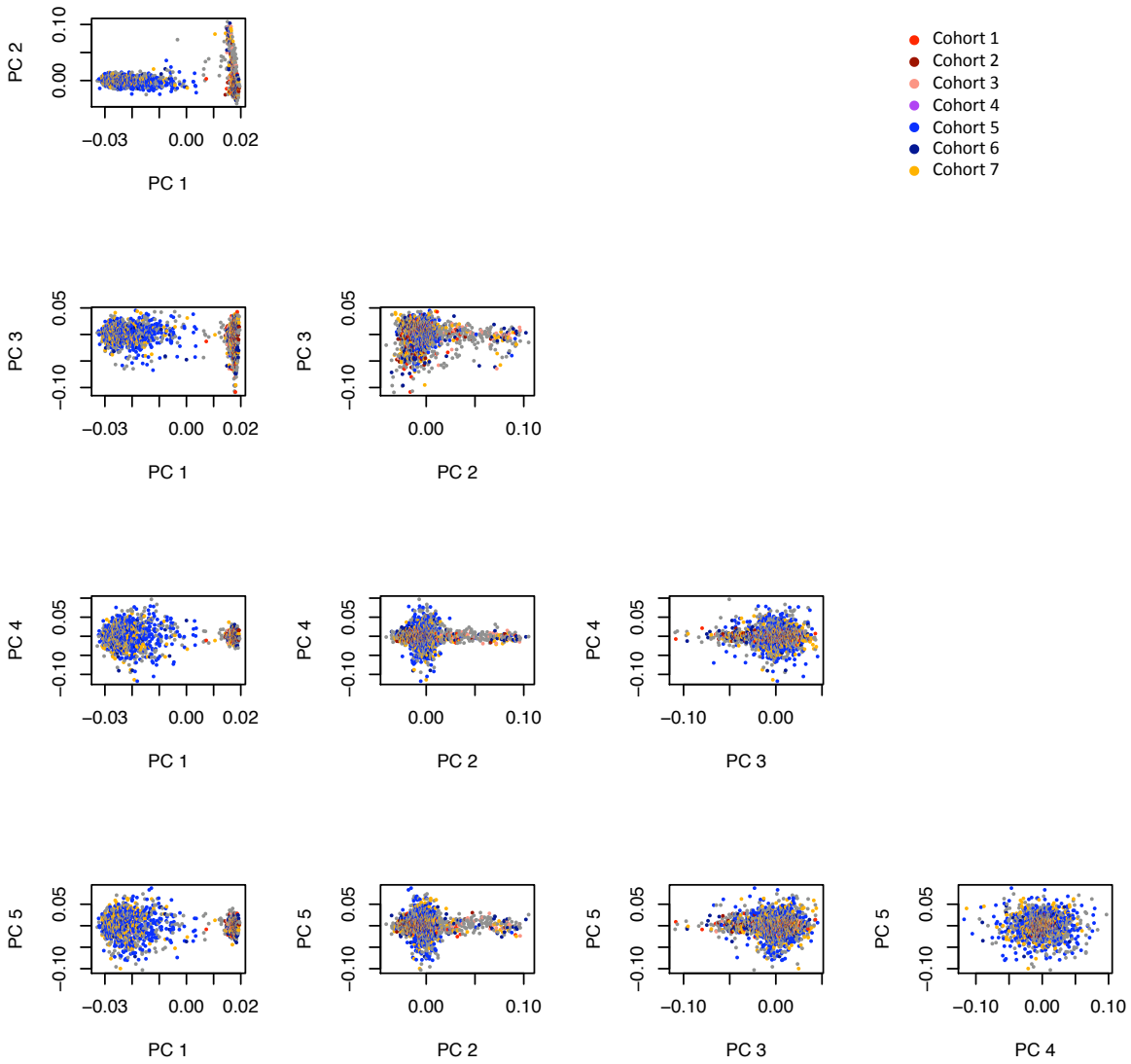

Figure S7. PCA plots with samples labeled by phenotype cohort.

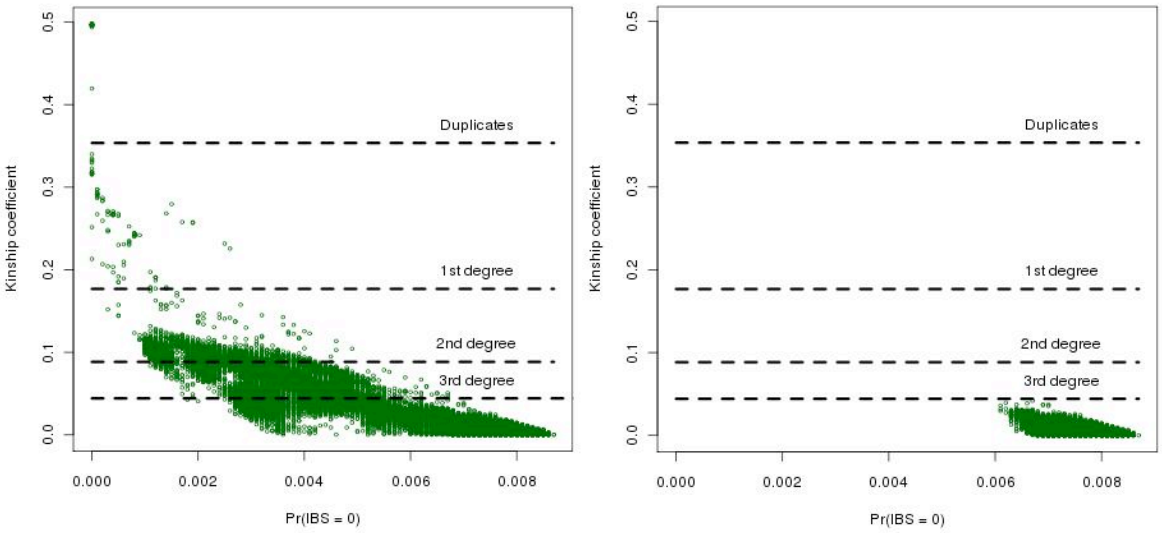

Figure S8. Kinship vs. IBS before (left) and after (right) filtering of related individuals and low quality samples.

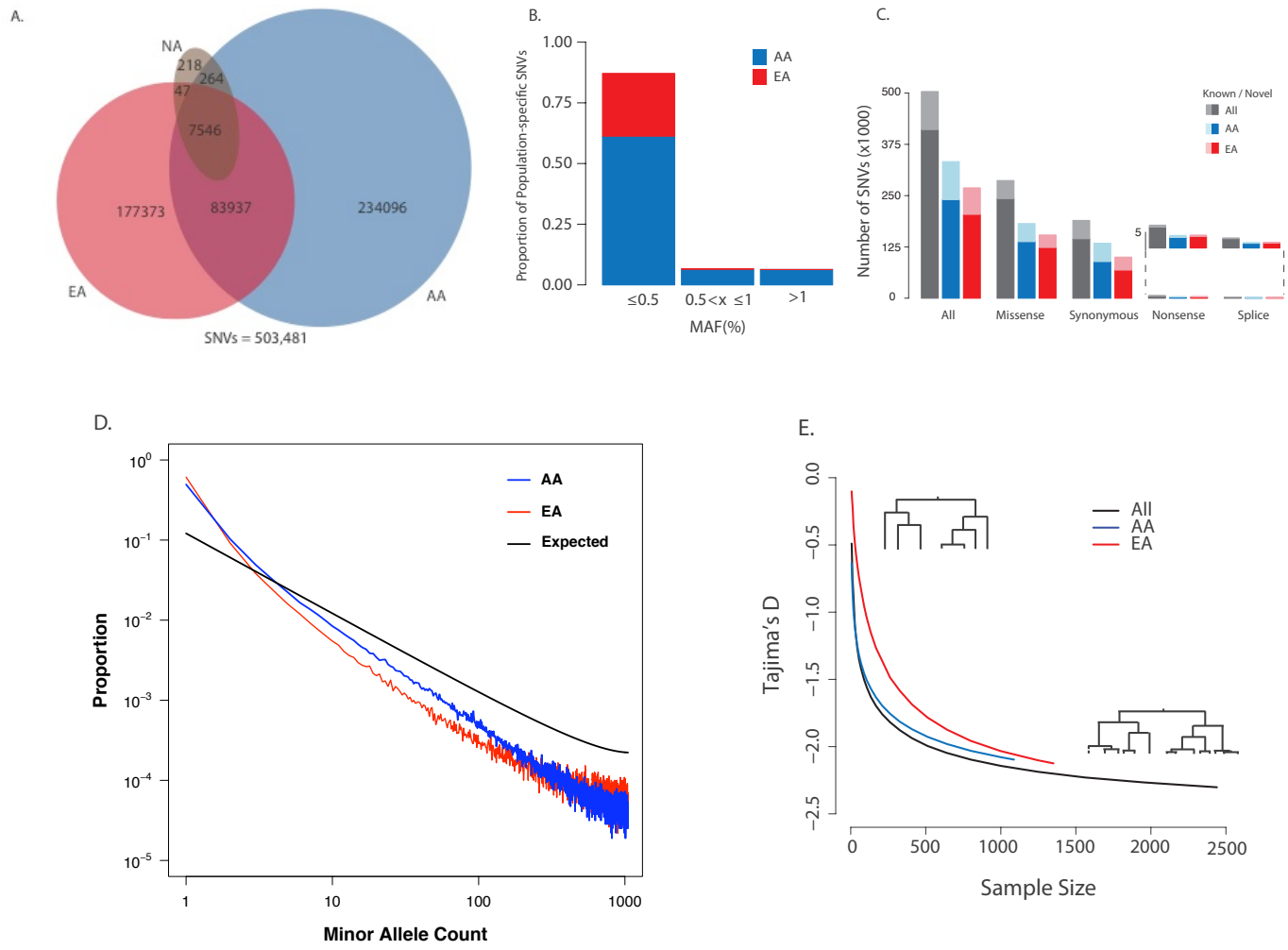

Figure S9. Characteristics of human protein-coding variation. A) Venn diagram showing the number of SNVs discovered in each sample, including the single Native American (NA) individual. Note, circles are not drawn to scale. B) Proportion of all non-singleton population-specific SNVs for different bins of MAF. Each bar also shows the relative proportion of population-specific SNVs observed in the AA and EA samples. C) Number of known and novel SNVs discovered for different site types. D) SFS of 1000 randomly selected individuals per sample. Note the large excess of rare SNVs relative to what is expected under a standard neutral model. E) The value of Tajima's *D* as a function of sample size. As sample size increases, the value of Tajima's *D* becomes sharply negative in each sample due to the increased resolution of observing recent population expansions (notice the primary consequence of increased sample size on genealogies is the addition of recent coalescent events).

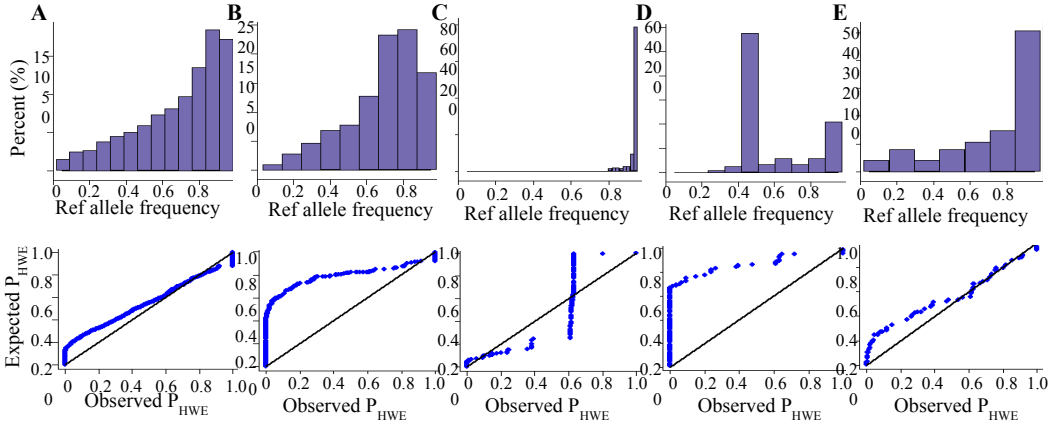

Figure S10. The performance of 5,561 SNPs that were only observed in European ancestral populations from 1KGP but not observed in EAs from ESP. The frequency distribution for reference allele were illustrated in the upper row; and the Q-Q plot for *P* values based upon HWE test were illustrated in the bottom row. **A** for C90, **B** for HWD, **C** for INV, **D** for NoC, **E** for OTH.

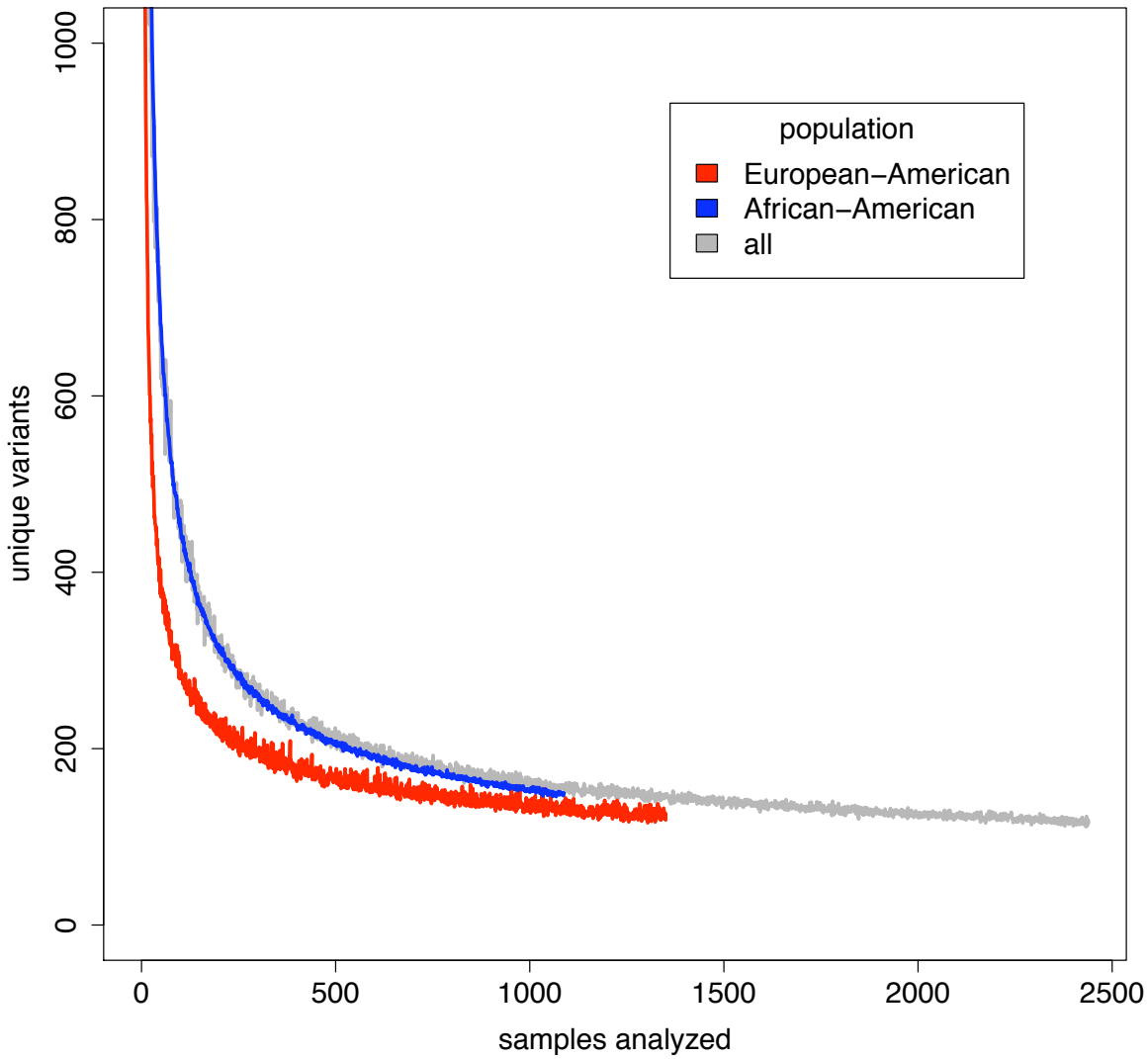

Figure S11. Number of unique variants identified as a function of sample size. Each point is the mean across 100 random samples of the desired number of individuals. Calculation of the number of unique variants was done for EA and AA populations individually and with both populations combined.

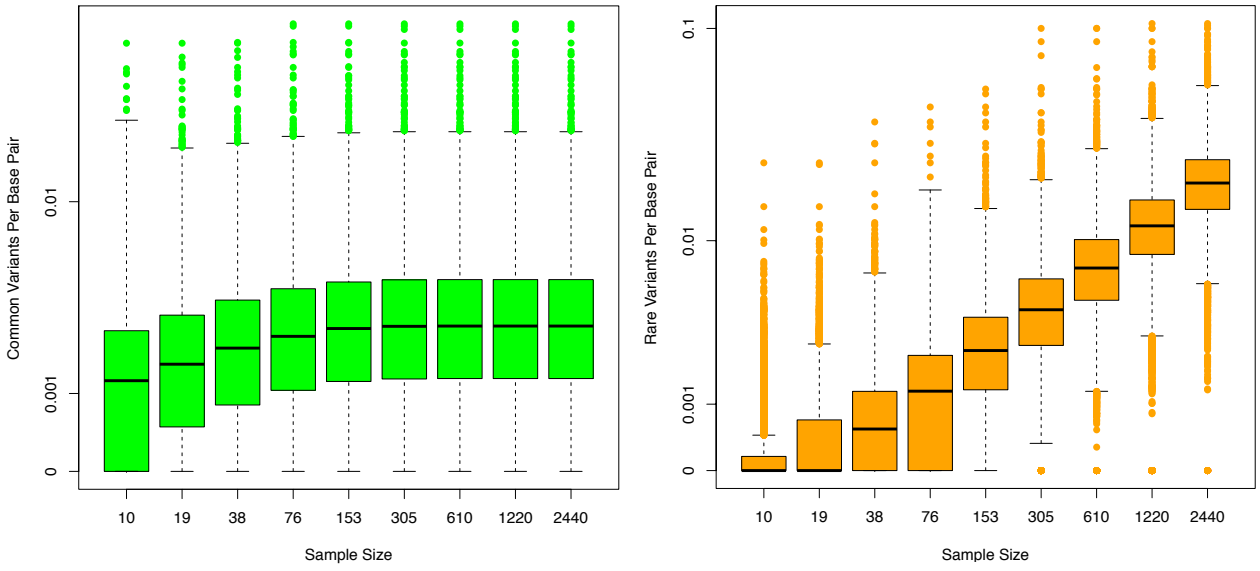

Figure S12. Number of common (left) and rare (right) variants per base pair as a function of sample size. A) Box plots showing the number of variants per bp with a MAF >0.5% as a function of sample size, estimated for each gene. Box edges represent the first and third quartile, while dark horizontal lines represent the median. Whiskers represent 1.5 times the interquartile range. Points represent outlier genes beyond the range of the whiskers. The number of common variants per base pair increases from a median of 0.0012 for a sample size of ten to a median of 0.0026 in the full dataset of 2440 individuals. Samples of approximately 100 individuals or more capture essentially all common variants. B) Box plots showing the number of rare (<0.5%) variants per bp as a function of sample size. Layout is the same as in (A). The number of rare variants per base pair increases from a median of zero for sample sizes under twenty to a median of 0.019 in the full dataset of 2440 individuals. Each increase in sample size reveals a linear increase of rare variants, with no sign of leveling off.

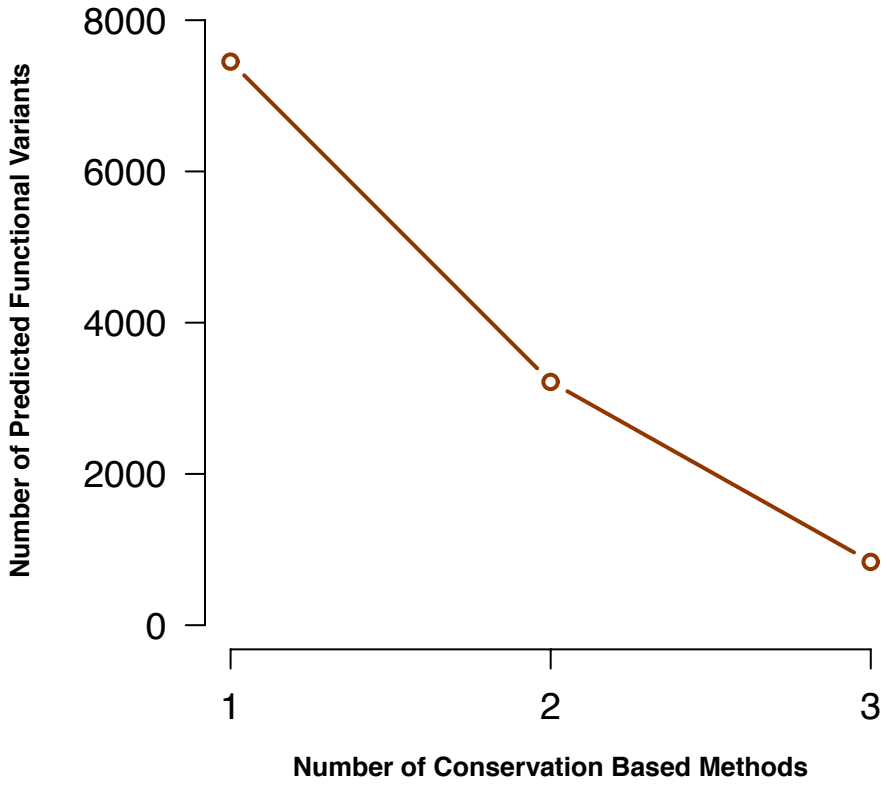

Figure S13. Number of synonymous SNVs predicted to be functionally important as a function of three different methods.

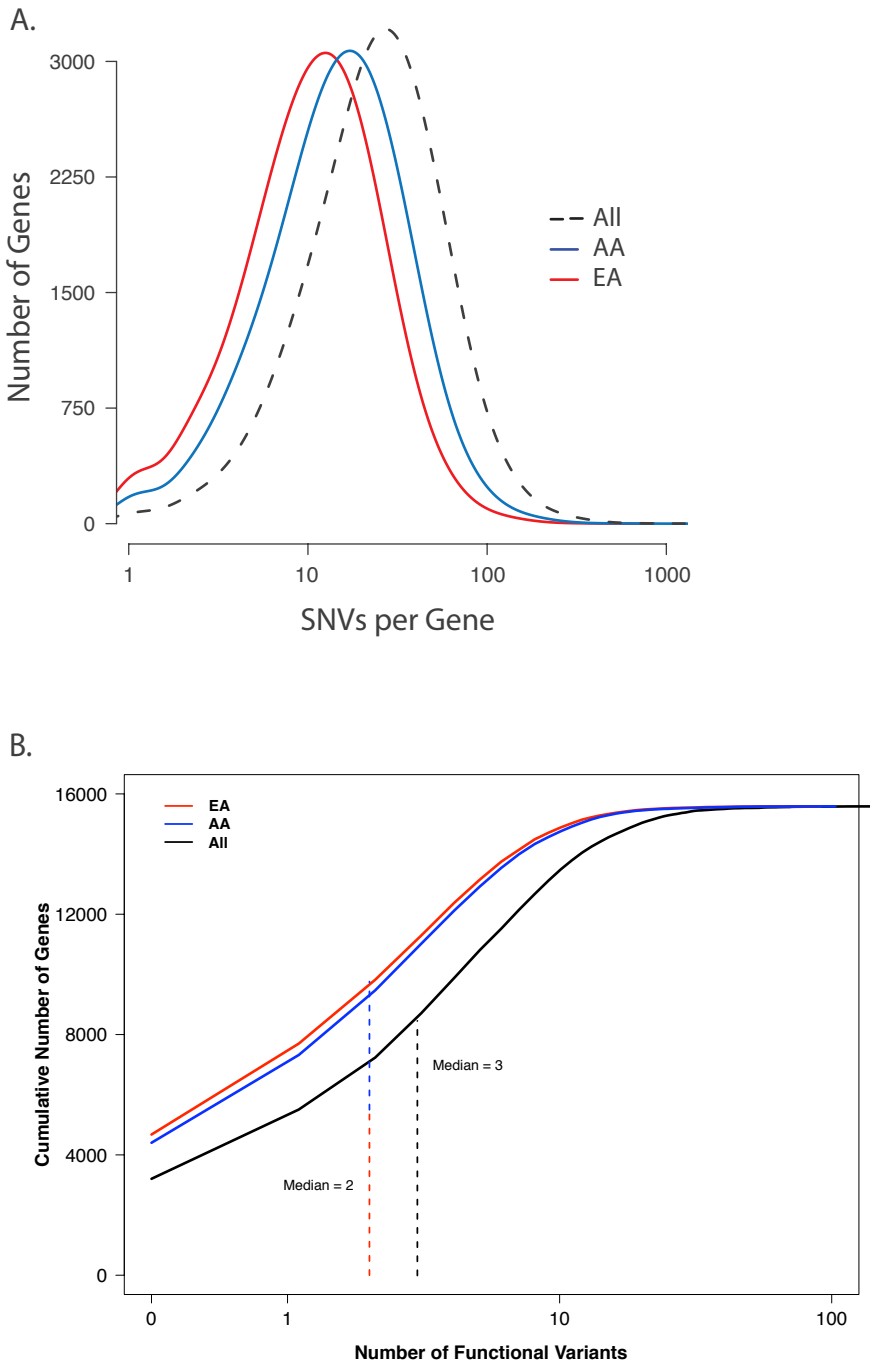

Figure S14. Distribution of protein-coding variation in genes, pathways, and across the genome. A) Number of SNVs per gene (note, to facilitate comparisons 1000 individuals were selected at random for each sample). B) Number of predicted functionally important SNVs per gene.

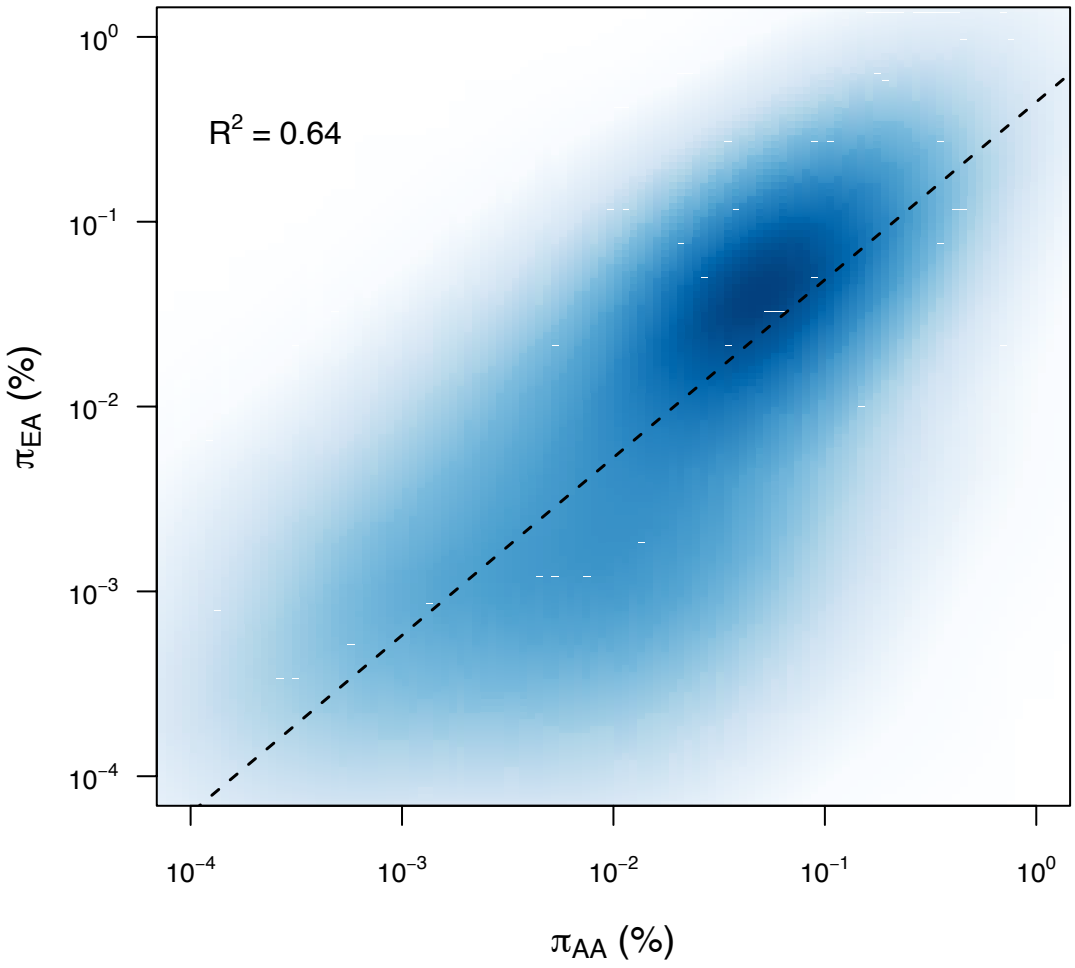

Figure S15. Smoothed scatter plot of nucleotide diversity in the EA versus AA samples. Darker shading increased a higher density of points.

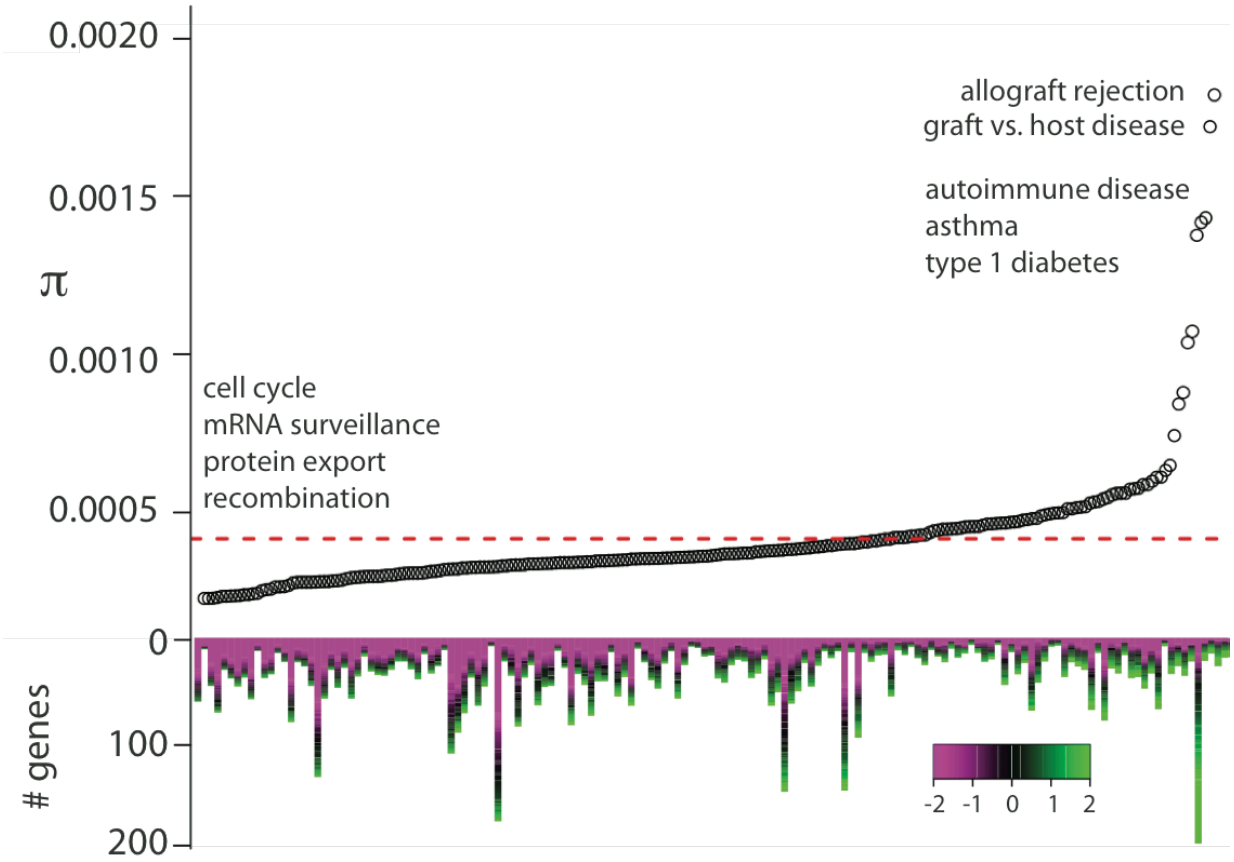

Figure S16. Average π for KEGG Pathways. The dashed red line denotes the average π across all pathways. The rows of the heatmap correspond to individual genes in each pathway, which are colored by their value of π (expressed as the $\log_2$ fold change relative to the average π across pathways).

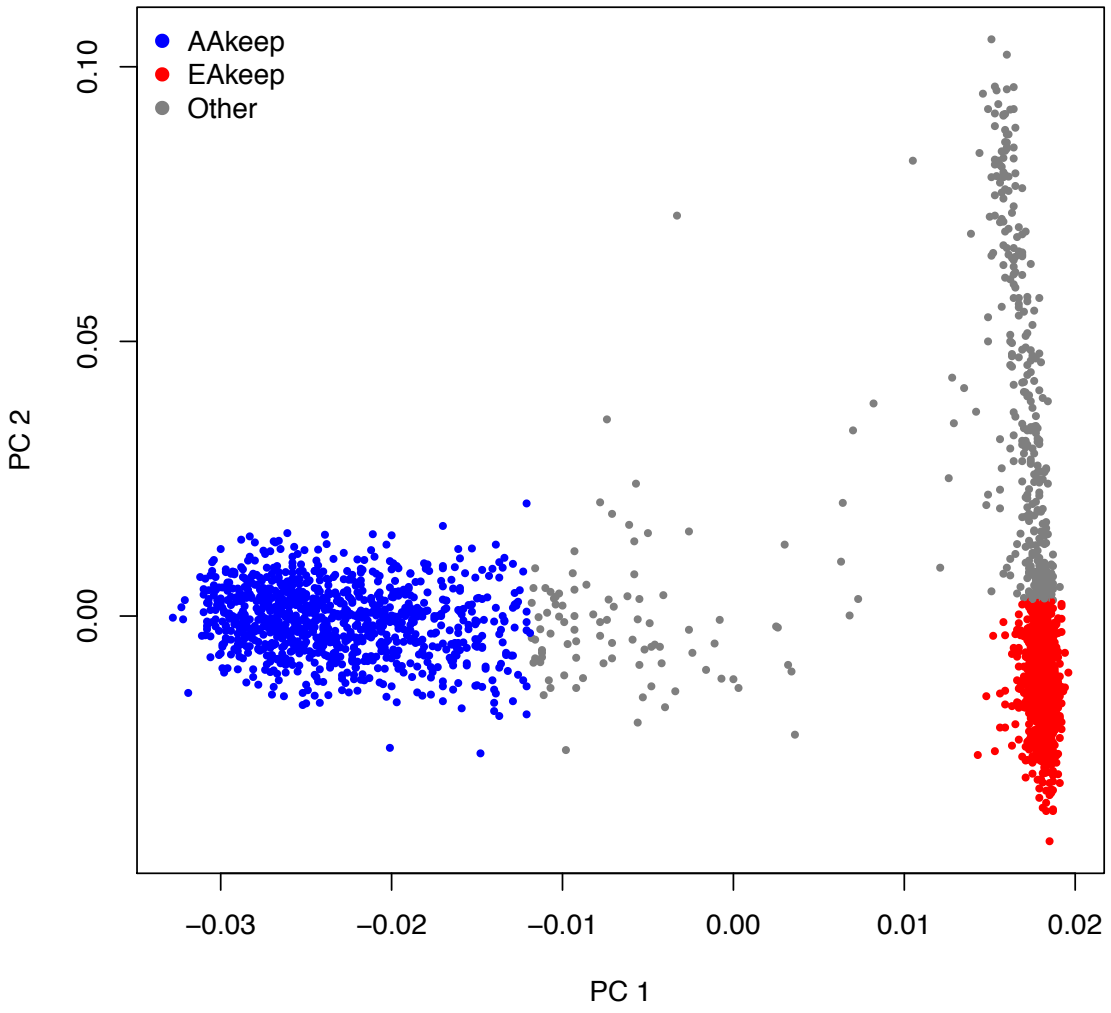

Figure S17. Summary of the 1,000 individuals selected in each sample (denoted EAkeep and AAkeep) for the simulations to evaluate the power of rare variant association studies. The goal of selecting these individuals was to mitigate spurious associations due to population stratification.

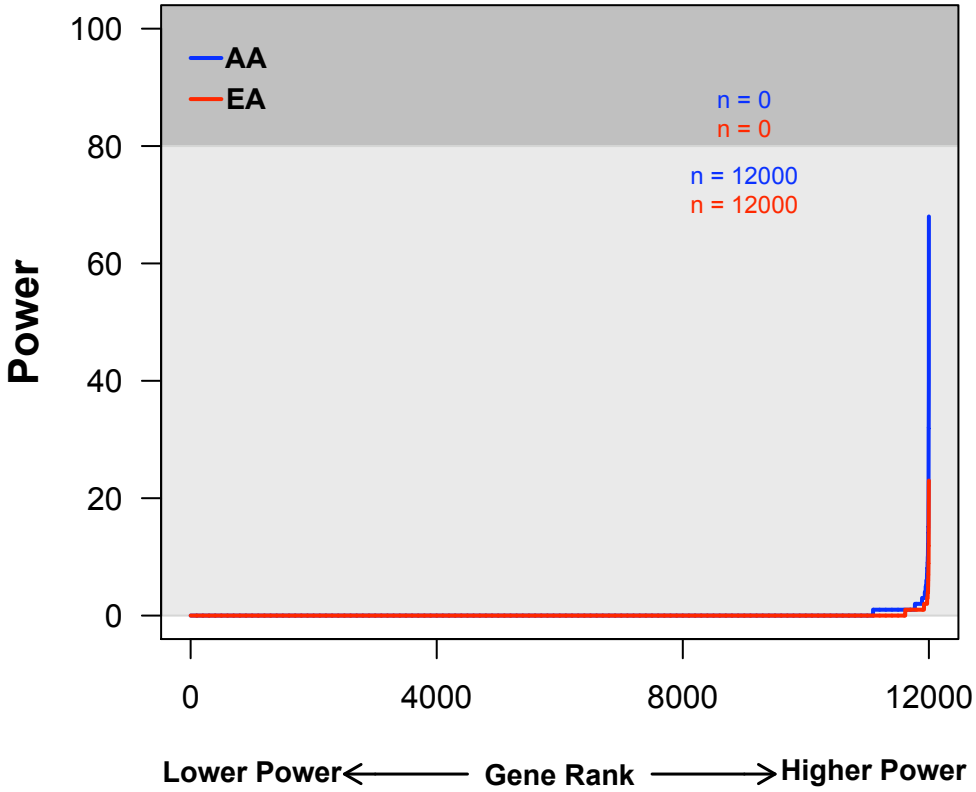

Figure S18. Distribution of gene-specific estimates of power to map causal rare variants across 12,000 protein-coding genes with at least three SNVs in the EA or AA samples assuming an OR for causal variants of 1.5.

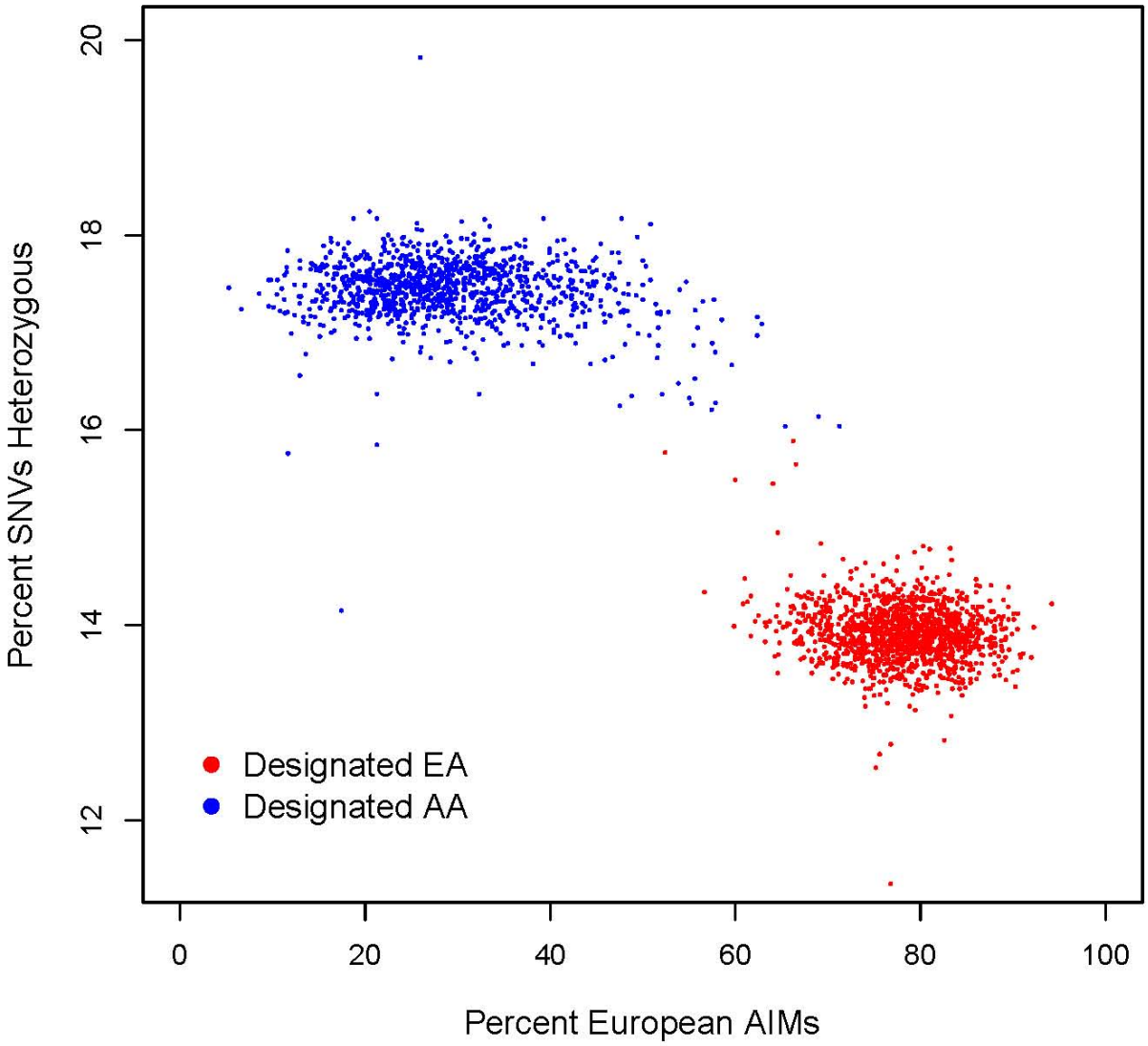

Figure S19. Assignment of AA and EA ancestry based on heterozygosity and percent European AIMs.

References and Notes

1. M. J. Bamshad *et al*., Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745 (2011). doi:10.1038/nrg3031 Medline

2. S. S. Ajay, S. C. Parker, H. O. Abaan, K. V. Fajardo, E. H. Margulies, Accurate and comprehensive sequencing of personal genomes. *Genome Res.* **21**, 1498 (2011). doi:10.1101/gr.123638.111 Medline

3. N. L. Sobreira *et al*., Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS Genet.* **6**, e1000991 (2010). doi:10.1371/journal.pgen.1000991 Medline

4. International HapMap Consortium, A haplotype map of the human genome. *Nature* **437**, 1299 (2005). doi:10.1038/nature04226 Medline

5. K. A. Frazer *et al*.; International HapMap Consortium, A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851 (2007). doi:10.1038/nature06258 Medline

6. J. Z. Li *et al*., Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100 (2008). doi:10.1126/science.1153717 Medline

7. Y. X. Fu, Statistical properties of segregating sites. *Theor. Popul. Biol.* **48**, 172 (1995). doi:10.1006/tpbi.1995.1025 Medline

8. G. T. Marth *et al*.; the 1000 Genomes Project, The functional spectrum of low-frequency coding variation. *Genome Biol.* **12**, R84 (2011). doi:10.1186/gb-2011-12-9-r84 Medline

9. S. B. Ng *et al*., Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272 (2009). doi:10.1038/nature08250 Medline

10. B. J. O'Roak *et al*., Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.* **43**, 585 (2011). doi:10.1038/ng.835 Medline

11. S. B. Ng *et al*., Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42**, 790 (2010). doi:10.1038/ng.646 Medline

12. S. B. Ng *et al*., Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30 (2010). doi:10.1038/ng.499 Medline

13. J. A. Tennessen, J. Madeoy, J. M. Akey, Signatures of positive selection apparent in a small sample of human exomes. *Genome Res.* **20**, 1327 (2010). doi:10.1101/gr.106161.110 Medline

14. X. Yi *et al*., Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75 (2010). doi:10.1126/science.1190371 Medline

15. J. McClellan, M. C. King, Genetic heterogeneity in human disease. *Cell* **141**, 210 (2010). doi:10.1016/j.cell.2010.03.032 Medline

16. T. A. Manolio *et al*., Finding the missing heritability of complex diseases. *Nature* **461**, 747 (2009). doi:10.1038/nature08494 Medline

17. G. Gibson, Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135 (2011). doi:10.1038/nrg3118 Medline

18. Supporting material is available on *Science* Online.

19. M. Kimura, Evolutionary rate at the molecular level. *Nature* **217**, 624 (1968). doi:10.1038/217624a0 Medline

20. A. Ramírez-Soriano, R. Nielsen, Correcting estimators of theta and Tajima's D for ascertainment biases caused by the single-nucleotide polymorphism discovery process. *Genetics* **181**, 701 (2009). doi:10.1534/genetics.108.094060 Medline

21. J. M. Akey *et al*., Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* **2**, e286 (2004). doi:10.1371/journal.pbio.0020286 Medline

22. A. Coventry *et al*., Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* **1**, 131 (2010). doi:10.1038/ncomms1130 Medline

23. S. Gravel *et al*.; 1000 Genomes Project, Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11983 (2011). doi:10.1073/pnas.1019276108 Medline

24. M. dos Reis, R. Savva, L. Wernisch, Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**, 5036 (2004). doi:10.1093/nar/gkh834 Medline

25. J. H. McDonald, M. Kreitman, Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**, 652 (1991). doi:10.1038/351652a0 Medline

26. J. M. Akey, Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.* **19**, 711 (2009). doi:10.1101/gr.086652.108 Medline

27. J. Asimit, E. Zeggini, Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.* **44**, 293 (2010). doi:10.1146/annurev-genet-102209-163421 Medline

28. G. V. Kryukov, A. Shpunt, J. A. Stamatoyannopoulos, S. R. Sunyaev, Power of deep, all-exon resequencing for discovery of human trait genes. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 3871 (2009). doi:10.1073/pnas.0812824106 Medline

29. K. E. Lohmueller *et al*., Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994 (2008). doi:10.1038/nature06611 Medline

30. L. Yu, F. D. Martinez, W. T. Klimecki, Automated high-throughput sex-typing assay. *Biotechniques* **37**, 662 (2004). Medline

31. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754 (2009). doi:10.1093/bioinformatics/btp324 Medline

32. M. A. DePristo *et al*., A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491 (2011). doi:10.1038/ng.806 Medline

33. S. Fisher *et al*., A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* **12**, R1 (2011). doi:10.1186/gb-2011-12-1-r1 Medline

34. Y. Li, C. Sidore, H. M. Kang, M. Boehnke, G. R. Abecasis, Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* **21**, 940 (2011). doi:10.1101/gr.117259.110 Medline

35. H. Li, J. Ruan, R. Durbin, Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851 (2008). doi:10.1101/gr.078212.108 Medline

36. R. M. Durbin *et al*.; 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061 (2010). doi:10.1038/nature09534 Medline

37. A. Manichaikul *et al*., Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867 (2010). doi:10.1093/bioinformatics/btq559 Medline

38. I. A. Adzhubei *et al*., A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248 (2010). doi:10.1038/nmeth0410-248 Medline

39. P. Kumar, S. Henikoff, P. C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073 (2009). doi:10.1038/nprot.2009.86 Medline

40. S. Chun, J. C. Fay, Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553 (2009). doi:10.1101/gr.092619.109 Medline

41. J. M. Schwarz, C. Rödelsperger, M. Schuelke, D. Seelow, MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* **7**, 575 (2010). doi:10.1038/nmeth0810-575 Medline

42. G. M. Cooper *et al*., Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods* **7**, 250 (2010). doi:10.1038/nmeth0410-250 Medline

43. G. M. Cooper *et al*.NISC Comparative Sequencing Program, Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901 (2005). doi:10.1101/gr.3577405 Medline

44. M. E. Zwick, D. J. Cutler, A. Chakravarti, Patterns of genetic variation in Mendelian and complex traits. *Annu. Rev. Genomics Hum. Genet.* **1**, 387 (2000). doi:10.1146/annurev.genom.1.1.387 Medline

45. R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009). doi:10.1371/journal.pgen.1000695 Medline

46. S. Gravel *et al*.; 1000 Genomes Project, Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11983 (2011). doi:10.1073/pnas.1019276108 Medline

47. Y. Y. Waldman, T. Tuller, T. Shlomi, R. Sharan, E. Ruppin, Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages. *Nucleic Acids Res.* **38**, 2964 (2010). doi:10.1093/nar/gkq009 Medline