Supplementary Materials for

An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People

Matthew R. Nelson*‡, Daniel Wegmann*, Margaret G. Ehm, Darren Kessner, Pamela St. Jean, Claudio Verzilli, Judong Shen, Zhengzheng Tang, Silviu-Alin Bacanu, Dana Fraser, Liling Warren, Jennifer Aponte, Matthew Zawistowski, Xiao Liu, Hao Zhang, Yong Zhang, Jun Li, Yun Li, Li Li, Peter Woollard, Simon Topp, Matthew D. Hall, Keith Nangle, Jun Wang, Gonçalo Abecasis, Lon R. Cardon, Sebastian Zöllner, John C. Whittaker, Stephanie L. Chissoe, John Novembre†‡, Vincent Mooser†

*† These authors contributed equally to this work
‡correspondence to: matthew.r.nelson@gsk.com; jnovembre@ucla.edu

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S15
Tables S1 to S17
Captions for databases S1 to S3

Other Supplementary Materials for this manuscript includes the following:

Databases S1 to S3 as zipped archives: S1, target regions of sequenced genes; S2, variants and their annotations; S3, European site frequency spectra.

Materials and Methods

Sample

The study sample of 14,002 individuals included 10,621 individuals from 12 case-control studies of common disease and 3,381 individuals from two population samples (table S4). Cases were selected from 12 common disease collections: coronary artery disease, metabolic syndrome, multiple sclerosis, osteoarthritis, rheumatoid arthritis, irritable bowel syndrome, epilepsy, Alzheimer's disease, unipolar depression, bipolar disorder, schizophrenia and chronic obstructive pulmonary disease. Population controls were included from two samples: CoLaus from Lausanne, Switzerland and LOLIPOP from London, UK, both having extensive cardiovascular trait measurements and the former including extensive psychiatric assessments. The two CEU and YRI trios sequenced at high depth by the 1000 Genomes Project (*30*) were also included. In total, 14,204 unique samples and 143 randomly selected sample duplicates were prepared for sequencing. The primary selection criteria applied to each study was the availability of at least 10 µg of DNA with a concentration at least 195 ng/µl from a primary blood sample. An overview of the samples selected from each collection is included below.

CoLaus Study. A population-based study of 6,188 European white subjects aged 35-75 years drawn from Lausanne Switzerland, through the CHUV University Hospital (*31*). Subjects included in the current study include 1,774 participants in the follow-on study of psychiatric traits (PsyCoLaus) (*32*) and 772 extremes of several selected cardiovascular disease-associated traits. There was an overlap of 460 subjects between these two selections.

LOLIPOP Study. A population-based study of 21,915 subjects, primarily of Indian Asians and Northwestern Europeans aged 35–75 years, identified from the lists of 58 general practitioners in West London (*33*). Subjects included in the current study include a random selection of 499 Indian Asians, 400 European whites selected for overlap with previous genome-wide genotyping studies, 149 European whites selected as extremes from several cardiovascular disease-related traits and 285 subjects of other non-European ancestry.

Metabolic Syndrome GEMS Study. The GEMS Study of Metabolic Syndrome and related traits included two types of samples; families and a set of unrelated cases and controls. Families (3,384 individuals from 535 families) were recruited from six study sites located in Australia, Canada, Finland, Switzerland, Turkey and the United States. Eligible families consisted of a minimum of two siblings (an affected sib-pair) with atherogenic dyslipidemia (ADL). In the case-control arm, a set of approximately 1,000 cases with ADL and 1,000 normolipidemic controls were recruited from the same GEMS sites. Details of the recruitment procedures, subject characteristics, and inclusion/exclusion criteria for both the family and case control studies have been previously described (*34, 35*). The current study includes 1,570 unrelated cases and controls and 30 parent-offspring trios for assessing sequence data quality, selected from all sites except the US and Turkey.

Coronary Artery Disease (CAD) MedStar Study. A premature CAD collection designed to investigate the genetics of plaque stability in acute coronary syndrome (ACS). The full study is comprised of 452 ACS CAD cases, 491 non-ACS CAD cases, and 483 non-CAD controls (*36*). Subjects were identified prospectively from the patient population of Cardiovascular Research Institute (MedStar/Washington Hospital Center). Standard criteria were used to identify cases with myocardial infarction and cases diagnosed with clinically significant coronary atherosclerosis without myocardial infarction. Subjects included in the current study include a selection of 609 ACS and non-ACS CAD cases.

Osteoarthritis GOGO Study. A large multicenter family-based study of 1,155 families from 5 United States and 2 United Kingdom sites with multiple joint osteoarthritis characterized both clinically and radiographically (*37*). The current study includes 836 cases.

Irritable Bowel Syndrome Study. A population-based study of 678 cases and 539 controls from 3 recruitment sites in Canada and the United States. Deeply phenotyped cases with a history of irritable bowel syndrome (IBS) for at least 6 months confirmed by a physician and according to the Rome II criteria and either a colonoscopy/barium enema with normal results supporting IBS diagnosis. Controls were matched to IBS cases and had no previous IBS diagnosis. The current study includes 317 cases.

Genetics of Rheumatoid Arthritis (GORA) Study. Patients with rheumatoid arthritis were recruited from Sheffield, United Kingdom, as described previously (*38*). Cases (~1,000) were of Northern European descent and all fulfilled the 1987 American College of Rheumatology classification criteria. A similar number of healthy controls were recruited. The current study includes 615 cases.

Multiple Sclerosis geneMSA Study. A study of 1,005 multiple sclerosis (MS) cases and 1,012 matched controls primarily of European ancestry from three sites in the United States, the Netherlands and Switzerland (*39*). The current study includes 673 cases.

Multiple Sclerosis African American Study. A study of African American cases and controls with subjects recruited from 39 states (*40, 41*). Cases were characterized through a systematic medical record review. Controls were invited to participate in the study by the probands and constitute primarily non-consanguinous spouses or friends of MS patients. All study participants were self-reported African-Americans and ancestry was documented based on genotyping results of 186 informative SNPs (*42*). The current study included 340 cases and 260 controls.

Epilepsy HitDIP Study. A study of 719 cases and 687 controls recruited from Norway and Finland. All patients had a definite diagnosis of epilepsy according to International League Against Epilepsy (ILAE) definitions. Controls had no neuropsychiatric conditions (*43*). The current study includes 185 Finnish cases.

Epilepsy GenEpa Study. A study of 318 cases and 348 controls from Swiss Epilepsy Centre, Zurich (*43, 44*). All patients had a definite diagnosis of epilepsy according to ILAE definitions. Controls had no neuropsychiatric condition. The current study includes 125 cases.

Alzheimer's Disease genADA Study. Study includes individuals with Alzheimer's disease (AD) diagnosed by the National Institute of Neurological and Communicative Diseases and Stroke/ Alzheimer's Disease and Related Disorders Association criteria. Subjects were recruited from nine memory referral clinics in Canada (*45*). The current study includes 705 cases.

Unipolar Depression Study. A study of 1,000 cases recruited from three ascertainment sites in Southern Germany (Munich, Augsburg and Ingolstadt) and 1,029 controls ascertained by the Max Plank Institute of Psychiatry in Munich. Cases diagnosed with recurrent major depressive disorder and controls were age and gender-matched non-affected controls (*46*). The current study includes 775 cases.

Schizophrenia Study. A study of approximately 1,600 cases and 850 controls collected from four sites in Aberdeen, UK, Greenock, UK, Munich, Germany and Quebec City, Canada. Cases were diagnosed with schizophrenia according to DSM-IV or ICD-10 criteria and healthy volunteers were randomly selected from the general population (*47*). The current study includes 1,109 cases.

Bipolar Disorder Study. A study of 965 bipolar cases and 933 controls from a multicenter study subjects of European ancestry from three different sites the Centre for Addiction and Mental Health in Toronto, Canada, the Institute of Psychiatry in London, UK and the University of Dundee, UK. Each case was assessed when euthymic and was diagnosed (lifetime) with the DSM-IV/ICD-10 bipolar I or bipolar II disorder (*47*). The current study includes 786 cases.

Chronic Obstructive Pulmonary Disease ECLIPSE Study. ECLIPSE (Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points) is a three-year non-interventional longitudinal study being conducted at 46 centers in 12 countries and is comprised of clinically relevant COPD cases with Global Initiative for Chronic Obstructive Lung Disease (GOLD) stage 11_IV COPD with a number of smoking and non-smoking and non-disease controls (*48*). The current study includes 1,002 cases from ten countries. Samples from the New Zealand site were excluded.

COPD HitDIP Study. A study of approximately 1,000 cases and 1,000 controls from Bergen, Norway. Cases consist of α1-antitrypsin deficiency-negative individuals with moderate to severe COPD according to GOLD criteria. (*49*). The current study includes 782 cases.

Informed consent

All study participants in the component studies provided written informed consent for the use of their DNA in genetic studies. A careful review was conducted to verify that the consents were consistent with the activities of this study. In selected instances further Institutional Review Board approval was sought and obtained where the appropriateness of the informed consent for the current study was not clear.

Self-reported ancestry information

We assigned each sampled individual to one of four ancestry groups: African-American (N = 594), European (N = 12,514), Southern Asian (N = 566, mostly from India) and other (N = 327) based of self-reported ancestry. Europe was further subdivided into eight geographic regions. These groupings do not reflect discrete structure in the data, rather the practical need to create sub-groups with reasonable sample sizes for more detailed analyses.

The demographic information available was variable across subjects. The most complete information contained self-identified ethnicity, country of birth and first language for the subject, two parents and four grandparents. Based on this information, we first attributed a best-guess geographic label to each of the family members based on the following rules: 1) missing data was ignored; 2) if ethnicity conflicted with birthplace or first language data, only ethnicity was considered; 3) if birthplace and first language disagreed, a higher level container label was chosen (e.g. an individual who was born in France but reported his first language to be Norwegian was labeled European); and 4) white individuals born in the US or Canada were attributed according to the first language information alone, if other than English. The geographic label for a sampled individual was then based upon the labels attributed to 1) the four grandparents, 2) two grandparents and one parent, 3) the two parents, or 4) the individual, based upon data availability. Conflicting labels of ancestors resulted in an attribution to a higher level label.

We divided Europe into geographic groups based on the UN geo-scheme for Europe, which has the four regions UN Northern, UN Western, UN Southern and UN Eastern Europe. We then further subdivided the regions with more than 500 individuals sampled (all but UN Eastern

Europe). UN Northern Europe was split into North-Western Europe (Great Britain & Ireland) and Northern Europe, which includes all other UN Northern countries but Finland. The Finnish population is known to be unique in its genetic diversity due to a strong, recent population bottleneck (*23*) and was thus treated as an independent unit. UN Western Europe was split into Western Europe (Belgium, France, Luxembourg, and the Netherlands) and Central Europe (Austria, Germany, and Switzerland). Finally, UN Southern Europe was split into South-Western Europe (Spain, Portugal, and Andorra) and South-Eastern Europe (all others). See table S10 for the European regions considered and the number of samples per group.

Target genes

The overriding objective of the experiment was to characterize a selection of target genes of interest to GlaxoSmithKline and conduct genotype-phenotype association analysis to identify potential drug repositioning opportunities (*50, 51*). Genes were selected from drug target genes across the pipeline, and for scale and feasibility reasons, was limited to 202 genes. The selected genes included 12 genes encoding targets of currently marketed drugs (Phase IV), 44 genes encoding targets of drugs which had been terminated after administration to humans (Phases I-III), and 76 genes encoding targets of drugs under active clinical development (Phases I-III). Drugs known to target multiple genes were omitted. In addition, 70 genes encoding targets of interest for pre-clinical development were included. The names and sequence characteristics of the genes are presented in table S1. The non-overlapping target regions are provided in database S1.

We compared several characteristics of the 202 genes selected in this study to the rest of the protein coding genome defined by GENCODE release 6 (*52*). There were 20,593 total protein coding GENCODE genes and 20,369 that overlapped with Ensembl Genes version 61 in GRCh37.p2. Of those, there were Gene Ontology terms (*53*) available for 20,340 genes downloaded from Ensembl BioMart on February 2, 2011 with a median length of 1,434 bp. The genes selected for this study had significantly longer coding regions than the rest of the coding genome, with medians of 1,756 and 1,434 bp, respectively (Wilcoxon $p = 7.2\times10^{-5}$).

The genes in this study differed from the rest of those in the genome in several common Gene Ontology terms (table S2). Significant terms were selected that were present in at least 5% of genes, either in GENCODE overall or within the study genes, and the differences in frequencies were statistically significant (p-value < 0.01 here). There were 16 cellular component terms that differentiated the set of genes under study, including substantial enrichment for proteins locating to the external side of the plasma membrane, membrane raft, integral to the plasma and postsynaptic membranes (all with odds ratio [OR] > 10). The study genes were significantly enriched for 27 different biological processes, including positive regulation of peptidyl-tyrosine phosphorylation, elevation of cytosolic calcium ion concentration and chemotaxis (OR > 15). There were 12 molecular functions that differed significantly between the two gene sets, including enrichment for G-protein coupled receptor activity, ion channel activity, receptor activity and cytokine activity (OR > 6) and a near absence of nucleic acid binding genes (OR = 0.08). These characteristics (table S2) that differentiate the genes in this study from the genome overall are expected for genes encoding drug targets.

Comparison of nonsynonymous:synonymous ratios with the protein-coding genome. Analysis of the number of sites at which sequenced subjects carried non-reference nonsynonymous (NS) alleles were fewer than expected based on projections from the 1000 Genomes Project pilot (*30*), as described in the main text. To determine if this could have been

due to experimental differences in the studies or differences in the rates of NS variants in the selected drug target genes, we assessed the ratio of nonsynonymous:synonymous (NS:S) variant alleles carried by each CEU subject using the published genotypes from the low coverage genome-wide 1000 Genomes Project sequence data, CEU.low_coverage.2010_09.genotypes.vcf.gz, accessed from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/paper_data_sets/a_map_of_human_variation/low_coverage/snps on March 15, 2011. Annotation of NS and S variants was obtained from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/technical/working/20100511_snp_annotation on March 15, 2011.

The results, shown in fig. S9, demonstrate that NS:S differs dramatically between the drug target genes in this study compared to the rest of the coding genome. To compare how different these genes are from other genes associated with human health and disease, we repeated this analysis using the genes reported from genome-wide association studies from the NHGRI Catalog of Genome-wide Association Studies (N = 3,736) (*22*), accessed on March 18, 2011, genes included in OMIM (N = 1,895), prepared as described below, and genes involved in drug absorption, distribution, metabolism and excretion (ADME; see http://www.pharmaadme.org for the gene listing; N = 299). We similarly found lower ratios in these selected genes, though not to the extent observed amongst the selected drug target genes.

DNA sequencing

DNA libraries were prepared from each sample by fragmenting 3 μg of genomic DNA to around 200 bp, followed by the addition of an 8 bp index sequence to each end to uniquely identify each sample and quantified in preparation for combining into 48-sample pools. Coding and noncoding exon boundaries were obtained via an Ensembl BioMart query of human genes against NCBI genome build 36.3. Fifty bases of flanking sequence were added, covering a total of 863,883 bases. The target gene regions were enriched for sequencing using a custom Roche (Madison, Wisconsin, USA) Nimblegen HD2.1M sequence capture array. Amplification of the eluted libraries was carried out with 12 PCR cycles. Loading volumes were determined by qPCR. Paired end sequencing was conducted for each 48-sample indexed pool on a single Illumina (San Diego, CA, USA) Genome Analyzer 2x lane. Over 93% of target bases were successfully sequenced in at least half of the study samples. The median sequencing depth was 27×.

SNP calling

Paired-end short reads were aligned with SOAP (*54*) and variants were called using SOAPsnp (*55*). Candidate SNV sites were identified for each sample where a genotype including non-reference allele was called with a minimum sequencing depth of four, a minimum consensus quality of 20 and no other SNV satisfying these criteria located within four base pairs in the same sample. Candidate SNVs were aggregated across all sequenced samples and consensus genotypes were called at these bases on each sample that had a minimum depth of seven and a minimum consensus quality of 20. There were a total of 50,432 such candidate variant sites identified. Additionally, variants were excluded if 1) singleton heterozygote variants had less than ten reads (N = 1,158), 2) fewer than 50% of sequenced subjects yielded a successful genotype (N = 1,373), or 3) duplicate genotype discordance was greater than 2% (N = 226). Samples were excluded from analysis if 1) their average sequencing depth was less than 10, 2) sequence-based genotypes were more than 15% discordant with genome-wide panel

genotypes (possible for 9,346 samples that had previous genotype data available to exclude possible sample mix-ups) or 3) the sample was sequenced multiple times and had lower average sequencing depth.

Missing genotype rates

Overall, missingness increased with allele frequency. Common variants (MAF > 0.05) had a median missing genotype rate of 2.1% compared to variants with MAF ≤ 0.001 with median missing rates less than 0.7%. As expected, we found a strong relationship between the distance of a sequenced base from the end of the target region and the subsequent depth, quality and missing genotype rates, as illustrated in fig. S10. Average sequencing depth is at its greatest approximately 100 bases from the end of the target region, and hence will generally have higher average genotype data quality and lower missing genotype rates. Hence it is more likely that rare variants will be missed (false negative) near the exon boundaries (50 bases from target start) than those that are further interior to the exon.

Transition:transversion ratios

The ratio of SNVs caused by transitions to transversions provides a qualitative assessment of the false positive rate as transition mutations are more than twice as common as transversions, whereas the two are equally likely as a result of sequencing errors. The transition:transversion ratios in our data are shown in table S11 for all variants, singletons, doubletons and a subset of the highest quality variants (MAF > 0.1% and missing less than 10% of genotype calls). The ratios are shown separately for NS, S, UTR and intronic SNVs as the expectations can differ markedly between variants of different types, particularly for S SNVs. Overall, the ratios are consistent with a high sequence data quality, though they are noticeably lower for singletons compared to doubletons and the highest quality variants.

Proximity to known insertions and deletions

The presence of insertions or deletions (indels) can result in incorrectly calling SNVs with calling algorithms that do not simultaneously carry out local realignments around known or suspected indels, as is the case for SOAPsnp in this study. We assessed the impact of this on variant quality of SNVs located around indels reported by the recent whole-genome sequencing of 179 individuals by the 1000 Genomes Project (*30*). The distance of each SNV to the nearest indel was calculated, where a distance of zero was given for variants located at or between the two reference bases spanning known indels, or the number of bases from those flanking positions.

Of 245 indels reported by the 1000 Genomes Project located within target regions, 206 included one or more SNVs located within 20 bases. There was a substantial excess of SNVs located within indel regions, a total of 150 SNVs within 106 unique indels, compared to surrounding SNVs (fig. S11). The majority of these are located within the UTR (96) and intronic (46) compared to coding (5) regions owing to the dearth of common coding indels. The average depth, quality and duplicate concordance of SNVs called within or near indels were significantly lower than those more distant. This pattern was noticeable for SNVs up to ten bases away. Similarly, transition:transversion ratios were significantly lower for the 267 SNVs within five bases of known indels, but indistinguishable for SNVs more distant. These 267 SNVs represent less than 1% of all variants observed, and only 0.2% (15) of NS SNVs. Although these results

emphasize the value of local realignment around known or suspected indels for genotype calling, they would have a minimal impact on the inferences of this study.

Determination of sequencing accuracy

Duplicate concordance. DNA samples from approximately one percent of subjects included in this study were randomly selected to be sequenced in duplicate. Duplicate samples were placed on separate microtiter plates and subsequently sequenced in separate indexed pools. We evaluated the duplicate sequence of 130 samples that passed subject-level quality control. We tabulated the number of discordant genotypes between duplicate pairs and estimated the overall and heterozygote discordance rates as well as the underlying error rates via maximum likelihood that gave rise to them (*56*). Table S12 contains counts of concordant and discordant genotypes for all variable base positions and stratified by whether the variant is included in dbSNP (release 126). Amongst singleton variants with genotypes called in both sample duplicates, 204 were observed to be heterozygous in both duplicates whereas 3 were heterozygous in only one. This gives a singleton duplicate heterozygous discordance rate of 0.015. Corresponding estimates of genotype error rates, assuming a single-allele error model (i.e. excluding the possibility that a genotype homozygous for one allele could be called as homozygous for another allele) are presented in table S13. The duplicate concordance reported here follows the exclusion of 226 SNVs with overall discordance rates >2%, which is only possible for variants with MAF >1% (at least 2% heterozygous calls). Most excluded variants were quite common. As a result, the error rate estimate for common variants is somewhat biased downward. However, independent methods (see below) were applied to further characterize the variant calling and genotype data quality.

1000 Genomes Project concordance. We included the CEU and YRI trios that were sequenced to high depth in pilot 2 of the 1000 Genomes Project (*30*) in this study to allow direct comparison of variants and genotypes called in independent experiments. We relied on the conservative genotype calls provided by the 1000 Genomes Project, that included only variants that passed stringent quality criteria, including Mendelian segregation, and had genotypes that were concordant between Broad and University of Michigan Genotype calls (*30*). Of these, there were 658 and 854 variants with genotype calls from both studies in the CEU and YRI samples, respectively. The genotype confusion table is shown in table S14. Combining the results from the two trios, we estimated an overall discordance rate of 0.42% and a heterozygote discordance rate of 0.95%. Although the overall discordance rate is strongly influenced by the frequencies of the variants available for comparison, the heterozygote discordance rate can be directly compared to the within-study duplicate heterozygote discordance rate described above. We find the two to be nearly identical at 0.95% versus 0.92%, respectively. No singleton discordances were observed.

Mendel errors. Thirty parent-offspring trios from the GEMS collection were sequenced. An analysis of genotype transmission patterns identified 37 Mendel errors involving 35 SNVs from 22 trios with one to three errors per pedigree. 32 of 37 errors involved homozygous parents and a heterozygous child. The SNVs involved were predominantly common, 24 of 35 with MAF greater than 0.5%. One singleton (of 21 carriers; 4.8% with exact 95% confidence interval = 0.12 to 24%) and no doubleton (of 73 carriers) variants resulted in Mendel errors. The overall genotyping error rate estimated from the Mendelian errors, from among 2,256 SNVs polymorphic in this sample of pedigrees, was estimated to be 0.06% using the method of Saunders et al. (*57*).

Capillary sequence concordance of singletons. Data from two standard Sanger capillary sequencing experiments were available to assess the singleton false discovery rate in this study. In the first, 985 of the subjects included in this study were sequenced in eight overlapping genes — *GPR119*, *GPBAR1*, *MLNR*, *PLA2G7*, *SIRT1*, *SIRT2*, *SIRT3* and *SIRT6* — covering approximately 10,000 coding and 24,000 noncoding bases. All amplicons were sequenced in both directions under standard conditions and resolved on ABI 3730xl automated sequencing instruments. Amplicons that passed quality control were analyzed to identify single base differences relative to the NCBI 36.3 reference sequence using PolyPhred v.5.04 and v.6.0. Genotype calls for all subjects at each coding variant position were manually reviewed. Sequencing, variant calling and manual review was carried out at Beckman Coulter Genomics (Danvers, MA).

We observed 40 singleton SNVs amongst the 985 subjects within the overlapping sequenced regions from the SOAPsnp data in the current study. When matched against the capillary sequencing results we found 35 of 40 SNVs were called heterozygous by both methods. All 22 of the coding singletons in the current study, subjected to manual review for genotype calling, were completely concordant between the two methods. Of the five remaining singleton SNVs that were not identified by the automated genotype calling software from the capillary sequence data, three were not successfully sequenced, one was found to show a clear double peak corresponding to the heterozygous genotype called in this study and one was undetermined (the read on one strand appeared clearly heterozygous while the other strand homozygous). Thus, of 37 singleton genotypes available for independent validation none were found to disagree between the two forms of sequencing, including two singleton trialleles.

In a second experiment specifically designed to assess the accuracy of the singleton calls from the short read sequence in this study, we randomly selected 125 singleton variants found amongst the 2,059 sequenced CoLaus subjects. As a typical capillary sequencing reaction would capture approximately 450 bases, we further identified any additional singleton variants carried by CoLaus subjects located within 200 bases of the randomly selected singleton. From among these variants we selected 225 singleton variants, sequencing three subjects for each singleton region. This design provided two negative controls (homozygous for the reference allele) and one heterozygous carrier for each singleton variant. Oligonucleotides were ordered from IDT (Integrated DNA Techologies, Coralville, Iowa). PCRs were set up using ABI GeneAmp FastPCR Master Mix (Applied Biosystems, Foster City, CA), and DNA sequencing reactions were set up using ABI BigDye Terminator v3.1 (1:10). Sequencing products were purified using CleanSEQ paramagnetic beads (Agencourt, Beverly, MA) automated on the Beckman FX (Beckman Coulter, Brea, CA) and sequenced on the ABI 3730xl DNA Analyzer. Sequence chromatograms were edited and aligned to the human genomic reference sequences using Sequencher (GeneCodes, Ann Arbor, MI) (v4.9) software. Secondary peak detection threshold was set as a minimum of 20 percent of major peak height to detect heterozygous peaks, and alleles were confirmed by automated and visual peak inspection for each polymorphism location.

Of the 225 sequencing reactions attempted, 15 failed. Of 210 successfully sequenced singleton carriers, six expected heterozygotes were found to be homozygous reference. All eight triallelic variants called with SOAPsnp were validated. Combining these results with those above from eight genes, capillary sequencing validated 240 out of 245 singletons identified in this study. The estimated false discovery rate is 2.0% with a 95% confidence interval of 0.7% to 4.7% (Pearson-Klopper exact method implemented from the binom package in R (*58*)). There

were no instances of calling non-reference homozygotes at singleton positions in the non-carriers from this study.

False negative rates. There are several potential reasons why true variants and variant carriers may be missed, including low genotype call rates at a variant site or genotyping errors that may result from allele-specific amplification biases, errors in short read sequence alignment, or other biases against calling non-reference alleles inherent in the genotype calling method employed here. The first source of false negatives is expected to disproportionately affect very rare variants, as there is a greater chance that they would be among the variants overlooked due to reduced genotype call rates. We approached assessment of this source by examining the relationship between SNV rates (number of SNVs observed per bp of sequence) over a range of call rates (fig. S12). Assuming that the false negative rate would be negligible amongst bases with call rates >95%, we used this subset of bases comprising 84% of all sequenced bases (>660 kb) to estimate the expected SNV rate. The SNV rate clearly decreases with decreasing call rates, though less than 4% of sequenced bases (passing the 0.5 call rate QC threshold) fall below 80% call rates where a significant drop in the SNV rate is observed. As expected, most of the undercalling is attributable to missed singletons (fig. S12B). Overall, we estimate that only 1.02% of all and as many as 2.72% of singleton variant sites were uncalled due to missing genotypes.

Other sources of uncalled variants can be difficult to identify and quantify. We carefully reviewed the capillary sequence data from the coding regions in eight genes described above. Of the bases that were successfully sequenced in this experiment, we identified 52 singleton variants carried by one of the 985 sequenced subjects. Of those, four were not identified among the SOAPsnp genotypes of the same subjects. One was due to an uncalled genotype (depth = 7, quality = 19; failing QC on both measures), in line with the expectations estimated in the previous paragraph. The remaining three show strong support in the capillary sequence traces, but no indication of the presence of the indicated allele in the short read data, in spite of high read depth (19, 32 and 76×) and genotype quality (91, 99 and 99). With both sources of data of such high quality, the true genotype cannot be determined. Assuming these are true singleton variants, the estimated false negative rate is 7.7%, though with fairly low precision (95% exact CI = 2.1-18.5%). In contrast, for the SNVs found in common between the capillary and short read sequence data, the probability of calling a reference homozygote in the short read data given a subject was found to be heterozygous by the capillary sequence data is 1.3% (28 of 2194 heterozygotes, 95% exact CI = 0.85–1.8%). This smaller value and non-overlapping confidence intervals suggests that some polymorphic sites may fail to be identified due to genomic context that may affect sequencing, alignment, and genotype calling.

SNP annotation

A substantial fraction of exonic SNPs will exert their effects on the protein function by altering protein structure. Starting with a table of NCBI build 36.3 SNP coordinates and annotation from the SNP calling pipeline, a Perl-based high-throughput protein analysis pipeline was developed to automate the use of two functional prediction tools. The primary output was a table of annotations containing the PolyPhen (*59*) and SIFT (*60*) predictions for the NS SNPs and the Ensembl consequence prediction for all SNPs.

The first step of the pipeline converted the build 36.3 SNP coordinates to those of NCBI build 37 using the Ensembl API (*61*). The build 37 SNP coordinates, reference base pair and mutation base pair were then used to query the SIFT web site. The next steps were required to

generate a list of UniProt identifiers and the residue changes as input for the PolyPhen program. For each SNP the Ensembl API was used to output the corresponding prediction and also a 200 bp RNA flanking sequence of SNP coordinates from all Ensembl transcripts which could contain that SNP. The RNA flanking sequence was aligned against UniProt protein sequences using BLAST. These alignments allowed the automated determination of the corresponding protein and any amino acid residue changes. Inconsistencies were automatically flagged for manual checking, for example synonym mismatches between the UniProt/HUGO Gene Symbol and the Gene Symbol provided by the SNP calling pipeline, alignments to gap regions (indicating alternative transcripts) and for mutations that would affect an existing start codon. Neither PolyPhen nor SIFT accounted for mutations in start codons. PolyPhen input files were generated containing the UniProt identifier, the amino acid residue position and the reference and mutant amino acid residues; the canonical amino acid residue found in the UniProt protein was used for the analysis even when it differed from that predicted by the build 37 reference genome. The PolyPhen input files were then submitted to the PolyPhen server grid via the PolyPhen Perl interface in batches of up to 50 residue changes at a time. The results were then extracted via the PolyPhen Perl interface and combined into a single file.

The final step in the process was to aggregate the Ensembl consequence predictions, warnings, the PolyPhen predictions and the SIFT predictions together with the original annotations. We note that our methods did not include all possible transcripts and protein isoforms, but rather the canonical forms.

The 46-way placental alignment phyloP conservation scores (*62*) were retrieved from the UCSC Genome archive (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP46way/) on October 10, 2010. Genome build 37 chromosome positions were converted to NCBI build 36 using the Ensembl API.

We assessed which variants were novel based on the overlap with dbSNP and 1000 Genomes Project variants. Overlapping variants were found by use of the SeattleSeq Annotation server (http://gvs.gs.washington.edu/SeattleSeqAnnotation/) using dbSNP build 131 and 1000 Genomes variants released in March 2010 using NCBI build 37 reference positions.

Triallelic variants

In total we found 745 triallelic and 12 tetraallelic sites, which corresponds to 2% of all variable sites to include at least a third allele. As expected, the rarest allele at most multi-allelic sites was usually very rare, seen only once or twice in 99% of instances. The expected number of triallelic variants increased almost linearly with sample size, reaching 1.2-1.9 variants per kilobase, depending on functional class (fig. S2).

Here we explored two complementary approaches to better characterize to what extent this observation is influenced by genotyping errors. First, we computed a liberal upper bound on the fraction of positions at which we expected to see a third allele based on known genotyping error rates from the duplicate analysis (see above). In addition, we assessed the expected number of triallelic sites expected given the observed mutation rate.

Expected tri-alleles due to genotyping error. As reported earlier, the rates of genotyping errors depend on the underlying genotype. We therefore computed the probability that a third allele is inferred due to genotyping error at a site by weighting the different error rates assuming Hardy-Weinberg equilibrium. Let us denote by N_l the total number of individuals successfully genotyped at locus $l = \{1, \ldots, L\}$ and by r_l and n_l the frequencies of the reference and non-

reference alleles at locus *l*, respectively. The fraction *f* of bi-allelic sites wrongly inferred as triallelic sites was then estimated as

$$f = \frac{1}{L}\sum_{l=1}^{L} N_l \left(1 - \left(r_l{}^2(1-\varepsilon_{RR}) + 2r_l n_l(1-\varepsilon_{RN}) + n_l{}^2(1-\varepsilon_{NN})\right)\right),$$

where the sum runs over *L* bi-allelic sites observed in the data sets and the genotyping error rates that lead to a calling a third allele are denoted by ε_{RR}, ε_{RN} and ε_{NN}, depending on the underlying true genotype homozygous reference (RR), heterozygous (RN) and homozygous non-reference (NN), respectively.

We obtained estimates of genotyping error rates from individuals for whom duplicates were sequenced (see above). A key observation was that error rates to miscall both alleles of an individual (for instance, calling a NN genotype RR) are virtually zero and we are ignoring such errors in the following. The rate at which a reference homozygous genotype (RR) is wrongly called heterozygous (RN or Rx) was also directly estimated from the duplicate analysis at 2.36×10^{-11}. However, assuming a transition:transversion ratio of 2:1 (which SOAPsnp does), on average only 7/12 of such errors lead to a third allele. We thus assumed $\varepsilon_{RR} = 1.38 \times 10^{-11}$. For the other two error rates (ε_{RN} and ε_{NN}) the error rates inferred from the duplicate analysis serve as liberal upper bounds. For instance, due to the strong reference bias of SOAPsnp, the rate at which homozygous non-reference genotypes (NN) are called heterozygous (RN) is much larger than the rate at which the same genotype is called heterozygous with a new non-reference allele (Nx). A liberal approach is thus to assume $\varepsilon_{NN} = \frac{7}{12} 8.34 \times 10^{-5} = 1.45 \times 10^{-5}$. Given that the most common error at heterozygous genotypes is to call a homozygous genotype instead, we assumed ε_{NN} serves as a liberal upper bound on ε_{RN} and thus assume $\varepsilon_{RN} = \varepsilon_{NN}$.

Based on those error rates, a liberal upper bound of the expected fraction of diallelic sites at which a third allele is observed due to error was estimated at 0.8%. This strongly suggests that a considerable fraction of the 2% polymorphic sites observed to be triallelic are not due to genotyping error.

Expected tri-alleles due to repeated mutation. A complementary approach is to estimate the fraction of polymorphic sites that are expected to be triallelic given the observed mutation rate. We assumed that the number of mutations *M* falling on the coalescent tree for a given site is Poisson distributed such that

$P(M = k) = \frac{e^{-\lambda}\lambda^k}{k!}$.

Then we could estimate λ using the proportion of monomorphic sites we observed (~20/21) obtaining $\lambda = -\ln\left(\frac{20}{21}\right) = 0.0487$. Here we assumed a homogenous mutation rate across the whole sequence. The true mutation rate at polymorphic sites is likely to be above average, and hence this approach will slightly underestimate the expected fraction of triallelic sites.

We now derived the probability that a given site is triallelic. Due to the very small probability that $M > 2$, we assumed triallelic sites arise only when $M = 2$, in which case the site may be monomorphic, diallelic, or triallelic, depending on where the mutations fall on the coalescent tree. We distinguished three cases:

a) Different lineages: no single lineage (from root to extant individual) contains both mutations, i.e. neither location is a descendant of the other.

b) Same lineage, different edges: the two mutations fall in a single lineage, but with a node between them, i.e. one is a descendant of the other, but there has been branching between them.

c) Same lineage, same edge: the two mutations fall on the same edge of the coalescent tree.

We assumed for simplicity that the probability that a given mutation is a transition is independent of the ancestral state, and denoted this probability by α. Similarly, we assumed the transversion probabilities to be independent of both the ancestral and derived state, and denoted this common probability by β. At any site, a mutation can be a transition or one of two possible transversions, so we had $\alpha + 2\beta = 1$.

Case (a) [different lineages]: The site is diallelic if we have parallel mutation, i.e. both mutations have the same derived base. Otherwise, the site is triallelic:

$P(\text{diallelic}|\text{different lineages}) = \alpha^2 + 2\beta^2$

$P(\text{triallelic}|\text{different lineages}) = 1 - (\alpha^2 + 2\beta^2)$

Case (b) [same lineage, different edges]: The site is diallelic if the second mutation reverts the first mutation, and triallelic otherwise.

$P(\text{diallelic}|\text{same lineage, different edges}) = \alpha^2 + 2\beta^2$

$P(\text{triallelic}|\text{same lineage, different edges}) = 1 - (\alpha^2 + 2\beta^2)$

Note that the probabilities are the same as in case (a).

Case(c) [same lineage, same edge]: The site is monomorphic if the second mutation reverts the first, and diallelic otherwise. We assumed that this case is rare.

We considered the case when $\alpha \approx 4\beta$, which corresponds to a transition:transversion ratio of 2:1, in which we had:

$\alpha = 2/3$

$\beta = 1/6$

$\alpha^2 + 2\beta^2 = 1/2$

An estimate on the proportion of triallelic sites was thus given by:

$P(\text{triallelic} \mid M=2) \times P(M=2) / P(M=1 \text{ or } 2) = 1.2\%$

In summary, these calculations strongly suggested a considerable fraction of the observed triallelic sites are expected to be seen and are not due to genotyping error. Further, all ten singleton triallelic variants subjected to Sanger capillary sequencing were validated (see above). Finally, there was also functional evidence for the triallelic variants to be real: sites at which we observed more than two alleles are on average less conserved among mammals that singleton, diallelic sites (fig. S3).

Overlap with HGMD and OMIM

We investigated the SNVs that overlapped between the current study and the HGMD and Online Mendelian Inheritance in Man (OMIM) variants to determine which and what fraction of variants reported to impact human health we observed, at what frequencies and how they relate to the predicted functional impact as assessed by SIFT, PolyPhen and phyloP. HGMD variants were queried from Professional version 2010.3. HGMD variants were merged with the current study based on their chromosome and map positions using NCBI human genome build 36.3. OMIM variants cross-referenced to SwissProt, curated by Dr. Andrew C. R. Martin, were downloaded from http://www.bioinf.org.uk/omim, last updated on August 20, 2010. OMIM variants were merged with the current study based on the amino acid position and UniProt

identifier. Due to the relatively high rate of misreports in OMIM (*63*, *64*), we carefully reviewed each entry in table S6 and assessed the evidence for a causal effect based on the information in the entry, or if needed, in the primary publication. This led to a categorization of the evidence supporting their involvement in the disorder as low, medium, or high. Top level observations in Europeans are described in the Supplementary Text below.

Frequency spectra

To generate frequency spectra for 2*n* chromosomes, we first excluded all sites for which less than *n* individuals had been genotyped. For sites with greater than *2n* chromosomes observed, we downsampled by calculating the expected number of sites with minor allele count *i*, which is given by the hyper geometric distribution (*65*). The sample size *n* was chosen for each population such that 80% of all targeted sites were used, with the exception of the European sample, for which the number was rounded down to an even number of 11,000 (which resulted in 84.6% of targeted sites to be included). We further generated two-dimensional frequency spectra for all population pairs using the same technique.

We summarized the frequency spectra by computing two estimators of $4N_e\mu$: θ_π, which is based on pairwise nucleotide diversity and θ_w, Watterson's estimator based on the number of segregating sites (*66*). We normalize these statistics by gene length by applying the same QC filters to monomorphic sites to determine the total number of fully observed base pair sites. Only autosomal genes were included in calculating the site frequency spectra.

Variant discovery curves

The fraction of sites expected to be found in a sample of size *n* was computed from the frequency spectra using the hyper-geometric distribution if *n* was smaller than the sample size (*65*) of the original spectrum, and using a jackknife approach for upward predictions for *n* larger than the observed sample size (*7*).

Allele sharing

Following (*7*) we computed sharing ratios between pairs of populations for each variant site as the probability that two randomly drawn carriers of the pooled sample are from different populations, normalized by the panmictic expectation. Computations were based on the expected two-dimensional frequency spectra with 474 chromosomes per population to have comparable values across population pairs while including all European populations. Reported values are averages across all variant sites in a given minor allele frequency bin, where minor allele frequencies were computed on the pooled, pairwise samples.

Expected ratio of NS to S variants

To compute the expected ratio of NS to S mutations in a given coding sequence we used the following method. We assumed known rates of mutation from a given nucleotide (e.g. C) to each of the other three nucleotides (e.g. A, G, T), conditioned on whether or not the nucleotide is within a CpG dinucleotide. Let $S = b_1\, b_2 \ldots b_L$ be a coding sequence of L nucleotide bases for a single gene. For two bases *x* and *y*, we let μ_{xy} be the rate of mutation from *x* to *y* if *x* is not in a CpG, and we let μ_{xy}^* be the rate of mutation from *x* to *y* if *x* is in a CpG.

For each nucleotide base b_i, we aimed to calculate the NS mutation rate r_i^N and S mutation rate r_i^S. This calculation is best illustrated by an example. Suppose b_i is the C nucleotide in the codon AAC, which codes for asparagine. We first looked at both the previous and next

nucleotide in the reference sequence to determine whether b_i is in a CpG (*67, 68*). We also consider the result of each possible mutation in the standard genetic code – in this case, AAT also codes for asparagine, while AAA and AAG code for lysine, i.e. C→T is S, while C→A and C→G are NS. In this case, we set $r_i^S = \mu_{CT}$ and $r_i^N = (\mu_{CA} + \mu_{CG})$. Note that if b_i is in a CpG, we used the rates μ_{xy}^* instead of μ_{xy}.

Then the overall NS:S mutation rate ratio for a single gene was calculated as:

$$\sum_{i=1}^{L} r_i^N \Big/ \sum_{i=1}^{L} r_i^S$$

To calculate a rate across all genes, we computed the sum over all nucleotides across all genes. For genes which have multiple transcripts, we chose the longest transcript, and suspect the overall results are robust to this choice as we found little variation from transcript to transcript in computed NS:S ratios.

For the mutation rates assumed by the calculation, we used values from two published studies on human-chimp nucleotide differences. Ebersberger et al. (*67*) (Tables 1-2) report frequencies of each possible nucleotide substitution observed at CpG and non-CpG sites based on 1.9Gb of human-chimp aligned sequence. Nachman and Crowell (*68*) (Table 4) report estimated transition and transversion mutation rates at CpG and non-CpG sites from pseudogenes. In the latter case, we needed to assume the two possible transversions for a given nucleotide are equally likely (and hence we took the rate for each of the two possible transversions from a given nucleotide to be one half of the overall transversion rate). To assess the robustness of estimated ratio of NS to S mutations to assumed rates, we used both sets of reported values, and we found they gave very similar predicted ratios (2.01 using Ebersberger et al's numbers vs. 2.08 using Nachman and Crowell's numbers).

Demographic history and mutation rate inference

Model and data. We followed the basic approach of Coventry *et al.* (*8*) to infer the current effective size of Europeans, *N*, the recent growth rate in the European population *r* and gene specific mutation rates $\boldsymbol{\mu}=\{\mu_1, ..., \mu_g\}$. This approach extends the demographic model of Schaffner *et al.* (*69*) to include a period of exponential growth in European population size that is parameterized by the current effective size of Europeans, *N*, the recent growth rate in the European population *r* and gene specific mutation rates $\boldsymbol{\mu}=\{\mu_1, ..., \mu_g\}$. In this model, the European expansion time is determined by solving for the time at which the ancestral European population of size 7,700 (from the Schaffner model) would need to start growing at rate *r* to reach a current size of *N*.

Likelihood approximation via Monte Carlo. The likelihood for a single gene is given by

$$L(N, r, \boldsymbol{\mu}|S) = \sum_{G \in \Psi} P(S|G, \mu) \cdot P(G|N, r),$$

where $\boldsymbol{S}$ is the site frequency spectrum for the gene, Ψ is the set of all possible genealogies G. Since Ψ is prohibitively large, the likelihood is approximated via Monte-Carlo. To be specific, we first generated *M* random genealogies G_i, $i=\{1, ...,400\}$ for each combination of demographic parameters *N* and *r* using fastsimcoal (*70*), a coalescent simulator that allows for multiple coalescent events per generatio. We approximate $L(N, r, \mu|S)$ as the average of $P(S|G_i, \mu)$ across these samples:

$$L(N, r, \boldsymbol{\mu}|S) \approx \frac{1}{M} \sum_{i=1}^{M} P(S|G_i, \mu).$$

Assuming that mutations follow a Poisson process on the genealogy, $P(S|G_i, \mu)$ can be calculated as the product of a "shape likelihood" and "rate likelihood" *(44)*.

$$P(S|G_i, \boldsymbol{\mu}) = P(S|G_i, S_{tot})P(S_{tot}|G_i, \mu).$$

The shape likelihood, $P(S|G_i, S_{tot})$ is specified by a multinomial with $S_{tot} = \sum_{i=1}^{m} S_m$ total observations, observed counts S, and success probabilities for a count of sites with *x* minor allele counts given by the relative length of all branches of the genealogy G_i with *x* descendants. The rate likelihood, $P(S_{tot}|G_i, \mu)$, is specified by a Poisson distribution for S_{tot} with a rate that depends on the total length of the genealogy and the mutation rate times the total number of sites considered:

$$P(S_{tot}|G_i, \mu) = e^{-nL\mu} \cdot \frac{(nL\mu)^S}{S!},$$

where L is the total length of the genealogy G_i and n the number of sites considered.

The above likelihood calculation can be done for each gene g, giving a likelihood $L_g(N, r, \mu_g)$. To extend to multiple genes, we let $\bar{\mu} = (\mu_1, \ldots, \mu_Z)$ where Z is the number of genes, and the likelihood for all genes was calculated by taking the product over all genes:

$$L(N, r, \bar{\mu}) = \prod_{z=1}^{Z} L_g\,(N, r, \mu_z)$$

For the model with a single μ value for all genes, we let $\mu_g = \mu$ for all g.

We restricted our inference to frequency spectra generated for genes on autosomes only and we only included four-fold degenerate synonymous sites. This allowed us to include a total of 188 genes in this analysis. Following Coventry *et al*. (*8*) we also pooled variants with minor allele counts > 17 into a single type. The frequency spectra are available in database table S3.

Initial grid approximation to the likelihood surface and comparison to Coventry *et al.* We initially used the same grid, marginal likelihood and posterior mean calculations as in Coventry *et al.* However, when we investigated the effect of changing the parameter grid, we found that the posterior mean estimates were dependent on the choice of parameter grid points. In effect, the Coventry method assigns a uniform prior over the parameter grid points, which will give different posterior distributions when, for example, the grid range is extended, or when the grid points for a parameter are placed on a logarithmic scale. For this reason, we chose to follow a strict maximum likelihood approach (retaining the Monte Carlo approximation to the likelihood of Coventry *et al.*), which gives estimates of *N*, *r*, and per-gene mutation rates that are more robust to the choice of grid points.

Expanded grid and strict maximum likelihood inference. Our initial per-gene mutation rate mutation rate estimates were very close to 1e-8, which is at the edge of our initial grid, so we expanded the grid over mutation rate values from 1e-9 to 1e-7 on a logarithmic scale (101 steps). Similarly, we extended the range of possible values of *N* to extend from 10,000 to 100 million on a logarithmic scale (41 steps). We chose our grid for *r* such that $r - 1$ extended from .005 to .14 on a logarithmic scale (57 steps), and we also included $r = 1.0$ (no growth). This grid spans the domains of parameter space with non-trivial likelihood and yet allows relatively fine-scale calculation of the surface in the regions of highest likelihood. To speed up the calculations, we recycled the coalescent trees for all grid points sharing the same N_e and *r* values.

The maximum likelihood estimate for N was 4.0 million with a 2 log-likelihood profile likelihood confidence interval of (2.5×10^6, 5.0×10^6) and for r was 1.017 (CI = 1.012, 1.023). The profile likelihood surface and per-gene mutation rate estimates are shown in Fig. 1C and 1D of the main text. Profile likelihood surfaces of N_e and a single global μ (or of r and μ, fig. S15) confirm that μ is an identifiable parameter in this inference scheme (i.e. the likelihood surface has a single point maximum as opposed to the ridge along fixed values of $N_e\mu$ expected in traditional population genetic inference).

Robustness to false negatives. We next checked the robustness of our inference to variants that were undetected due to variable coverage. Our quality control indicated a false negative rate of ~2%-8% for singletons plausible, so we spiked in extra singletons to mimic missing 2% and 8% of the total number of singletons.

Our maximum likelihood estimate of r was 1.018 in both cases and thus slightly higher than the estimate from the raw data. The maximum likelihood estimate for N were very similar with 4.0×10^6 and 5.0×10^6 for 2% and 8% singletons added, repectively. The resulting median per gene mutation rates were 1.45×10^{-8} in both cases and thus only marginally larger than our initial estimate of 1.38×10^{-8}.

Robustness to conservation of functional synonymous sites. One concern with our analysis is that some synonymous sites may be functionally constrained or experiencing background selection, which may lead to an artifical variance in inferred mutation rates across genes. We note that the average phyloP for the four-fold degenerate sites considered in this analysis is close to zero (0.08) and only 10% of all four-fold degenerate sites have phyloP scores above the median phyloP score observed at non-synonymous sites. More importantly, when we correlate the MLEs for μ for each gene with the average phylop score at all coding sites in a gene we find no correlation (p=0.08).

Ratio-based estimates of deleteriousness conditional on frequency

Given the ratio of nonsynonymous to synonymous variants r_i within minor allele frequency i and the same ratio r_f observed among variants of frequency f, an estimator of the fraction of nonsynonymous variants with frequency i that are deleterious enough to never reach frequency f is given by $1-(r_f/r_i)$ (*2*). While we note this inference procedure does not account for changing population size and/or the effects of background selection on synonymous variants, we used this approach to estimate the fraction of nonsynonymous mutations deleterious enough to never get fixed in humans ($r_f = 0.266$ from (*2*)) and the fraction to never reach high frequencies ($r_f = 0.516$ estimated from all variants > 5% in our European sample). The same rationale can be used to infer the fraction of new mutations never to be found polymorphic in a sample of 11,000 individuals by contrasting the expected ratio of nonsynonymous to synonymous variants of random mutations (r_i=2.01, see above) and the same ratio observed among singletons (r_f = 1.743). We thus estimate that about 13% of newly arising NS SNVs are deleterious enough not to be discovered in a sample of 11,000 individuals.

This last calculation provides an upper bound for the fraction of novel nonsynonymous mutations that are dominant lethal. We can thus estimate an upper bound on the number of de novo dominant lethal mutations arising per generation in a single individual. Given the total length of the human exome of 35.2 Mb (*71*), using our inferred median mutation rate of 1.38×10^{-8} as a proxy for the whole coding genome and assuming that 2/3 of all de novo mutations in coding regions are non-synonymous we arrive at ~0.32 de novo non-synonymous mutations per generation per individual. Hence, we estimate that no more than 0.32×0.13 =

0.042 dominant lethal mutations arise de novo per generation per individual, which corresponds to less than one per 23 generations.

Association analysis

Each of the sequenced collections includes a wide range of clinical and laboratory measures that merit careful analysis and interpretation, which are under way. However, to illustrate an analysis strategy and investigate possible associations of common coding and rare variants with the diseases represented in this collection, we present a standardized and simplified analysis for each of the 12 diseases.

As shown in table S4, very few disease-matched controls were included in this study. A population control strategy was employed to test for association with disease status. Control subjects were available from two population samples (CoLaus and LOLIPOP) and from other case groups where consent and/or the approval of an ethics review board was granted. There were a total of 10,114 subjects of European origin identified as possible controls. In this analysis, the only criteria used to select controls was genetic similarity assessed by principal component analysis (PCA) in the entire European sample using available genome-wide genotype data on a common set of SNPs present on the variety of genotyping platforms, including Affymetrix 5.0 and 6.0 and Illumina 550K. The genetic similarity between each case-control pair was based on the weighted Euclidean distance between each case and control (*72*) using the first five principal components. For each case collection, a range of case to matching control ratios was explored, selecting the maximum number of controls that resulted in a median distance less than 0.02. The resulting numbers of cases and controls and median Euclidean distance between them is listed in table S7 for each study.

Association analysis with case status was carried out using logistic regression, including the first five principal components as covariates in the model and assuming an additive genetic model. Given the use of population controls, no study-specific covariates were included. The one exception is the GEMS dyslipidemia study where normolipidemic study controls were available; age, sex, body mass index, physical activity and alcohol use were included as covariates. In the analyses presented here, only coding variants were included. Single variant analyses were carried out for all SNPs with a European allele frequency greater than 0.5% (606 "common" SNPs in this analysis). Coding variants, including those in splice sites, were analyzed as an aggregate indicator of rare variant carriage status, taking on a value of zero for no rare variants and one where one or more rare variants were present in a subject. Two aggregate tests were carried out and reported here: 1) all rare variants that lead to a change in the amino acid sequence (NS, nonsense and splice site variants) and 2) all amino acid-changing variants that are predicted to be functional by SIFT or PolyPhen or occur at a highly conserved base position (phyloP ≥ 2).

Power analysis

Statistical power to test the null hypothesis of no difference in cumulative rare variant frequencies in cases versus controls was carried out asymptotically by computing the noncentrality parameter of a chi-square distribution with one degree of freedom. The noncentrality parameter was derived using the expected genotype frequencies in cases and controls given the cumulative minor allele frequency observed for each gene, the number of cases and population controls, and assuming Hardy-Weinberg equilibrium and a disease

prevalence of 5% (see (*73*) for details). Power was computed at a test-wise significance level of 0.05/202.

Supplementary Text

Gene-to-gene variation of common and rare variants and mutation rates

Although the number of common (MAF > 0.5%) and rare (MAF ≤ 0.5%) NS variants (fig. S13) and the cumulative NS rare variant frequencies (fig. S4) are correlated with the number of successfully sequenced coding bases, we observed a substantial amount of gene-to-gene heterogeneity. Ordinary least squares regression with sequenced coding length as a predictor explained only 15% of the intergenic variation in the number of common NS variants, but 71% of the rare variants, and 53% of the cumulative MAF (cMAF) of rare variants predicted to affect protein function by SIFT or PolyPhen or occur at a highly conserved base (phyloP ≥ 2) (table S15). Mutation rate was not associated with coding length.

We investigated several possible explanations for the variation remaining after adjusting for sequenced coding length, including average coding sequence conservation scores, GC content, recombination rate, as well as Gene Ontology and Interpro terms, and embryonic lethality of mouse knock outs of the homologous gene. We estimated the average phyloP and GC content for all successfully sequenced coding bases for each gene and average sex-adjusted recombination rate within genes. Adjusting for coding length by division (except for mutation rate), we tested the association of these variables with a likelihood ratio test and estimated the amount of variation it explained (table S15). Average phyloP score was strongly associated with all dependent variables except mutation rate. GC content was strongly associated with the number of rare NS variants and mutation rate. Recombination rate was not significantly associated in any of the tests conducted and was dropped from the final models. The overall amount of variation explained, after adjusting for coding length, was 12%, 20%, 3%, and 5% for the number of common NS variants, number of rare NS variants, cMAF, and mutation rate, respectively.

We investigated whether the differences in the coding length-adjusted measures of genetic variation were explained by differences in gene activity or function using Gene Ontology and Interpro terms. Each gene may have several GO terms in each of three categories. We selected for analysis any term that was observed in at least 5% of the selected genes. For each term, we evaluated its association with each of the three length-adjusted measures of genetic variation by comparing those genes with the selected term to all others with a Wilcoxon sign rank test. As there are many terms, we only considered those with p-values less than 0.05 divided by the number of terms within the GO class to be statistically significant (Bonferroni adjustment). We found no statistically significant associations to report. We conducted a similar analysis using the Interpro (*74*) version 30.0 accessed from Ensembl BioMart on February 2, 2011. Again, no statistically significant associations were identified.

Finally, we investigated whether knocking out the gene resulted in embryonic lethality in a mouse knockout model provided any additional explanation for the intergenic variability in NS variation. Mouse knockout phenotypes were downloaded from the Jackson Labs MGI Biomart (*75*) on March 1, 2011. We identified genes with any type of lethal phenotype. Only 17 (8%) of our genes had documented lethal mouse knockout phenotypes (*ADAM10*, *ADIPOQ*, *BRD4*, *EDNRA*, *EDNRB*, *FGF10*, *GSK3B*, *HHIP*, *HTR4*, *IKBKB*, *MAPK14*, *PIK3CA*, *PPARD*, *PSEN1*, *PSEN2*, *STIM1*, and *TNFRSF1A*) compared to 27% of all genes with documented phenotypes. We found the lethality phenotype was associated with the number of rare NS variants after

adjusting for coding length ($p = 0.0017$), but not with the number of common variants or cumulative MAFs. However, lethality was also significantly associated with average phyloP. After adjusting for average phyloP, lethality was no longer significantly associated.

Hence, amongst the possible explanations for the intergenic variability investigated in this study, only average sequence conservation was consistently associated. However, although statistically significant, average phyloP did not explain a sufficient amount of variation to prove particularly useful in predicting the amount of rare NS variation expected.

Overlap with OMIM and HGMD

Fifty three of the 202 genes in this study are reported to have disease-causing mutations in HGMD. A total of 170 of 1,460 (11.6%) disease causing (DM) variants in 35 genes were observed in the combined sample. Of those, 149 were observed in Europeans, 40 in South Asians and 51 in African Americans. In Europeans, all disease-causing mutations had MAF less than 5%, and 23.5% had MAF greater than 0.1%, with nearly half (48.3%) being observed in only one or two subjects.

A total of 46 OMIM variants in 25 genes were observed in the combined sample. Of those, 44 were observed in Europeans, 20 in South Asians and 26 in African Americans. In total, 17.0% of disease-causing variants in these genes were observed in our sample but were not clustered within any particular disease cohort with the exception that the Alzheimer's disease variants were enriched in the Alzheimer's cases (1.0% in cases versus 0.28% in others; Fisher's $p = 0.005$, odds ratio = 3.8). There were 35 variants in 17 genes after excluding relatively common SNVs (table S6). The combined European frequency of those variants with medium to high evidence that they cause the corresponding indicated disorder in a dominant fashion was 2.7%. However, most are exceedingly rare. After excluding two variants with MAF >0.5% yields a combined frequency of 0.35%. However, caution is needed interpreting this result as little is known about the penetrance of most of these variants, having been reported in a single study or pedigree. Many of these variants may have relatively low penetrance in the general population.

Comparison of SIFT, PolyPhen and phyloP

Of 10,995 total NS SNVs called in the entire sample, 97.7% and 98.7% resulted in PolyPhen and SIFT predictions, respectively (table S16). Of those variants called by both PolyPhen and SIFT, 43.3% were called as benign/tolerated by both and 12.7% as probably damaging/damaging by both (i.e. 77.7% concordant excluding possibly damaging and low confidence damaging groups; table S17). A similar percentage were called damaging by SIFT but benign by PolyPhen (14.0%); however, only 2.2% were called probably damaging by PolyPhen but tolerated by SIFT.

The relationship between sequence conservation assessed by phyloP score and predicted functionality by SIFT and PolyPhen were very similar (fig. S14), though they differed significantly between methods. SIFT tolerated NS SNVs had significantly lower phyloP scores compared to PolyPhen benign ($p = 3.4\times10^{-8}$). The same was true of the low confidence damaging compared to the possibly damaging classes. The distribution of phyloP scores between SIFT damaging and PolyPhen probably damaging NS SNVs were not statistically significantly different ($p = 0.13$). The difference between the median phyloP scores between tolerated/benign and damaging/probably damaging NS SNVs was 1.1 and 0.92 for SIFT and PolyPhen, respectively. This difference was similar to that observed between common and

singleton NS SNVs (Fig. 2E). A total of 5957 NS variants (63%) were predicted to be damaging by PolyPhen or SIFT, or occurred at a nucleotide position with a phyloP conservation score greater than 2.0.

Rare NS SNVs were more often predicted to be damaging by SIFT ($p < 10^{-4}$) (*60*) and PolyPhen ($p < 10^{-3}$) (*59*) than common, NS SNVs (Fig. 2D). Similar, patterns of evolutionary conservation as measured by phyloP score (*62*) were negatively correlated with the frequency of NS variants ($p < 10^{-12}$), but not S variants ($p = 0.62$). Such negative correlation is expected if long-term conservation and on-going purifying selection act on the same sites. We also saw a negative correlation for UTR variants ($p < 10^{-9}$) and a weaker, but still significant relationship for intronic variants ($p < 0.005$).

The 297 SNVs from the current study found in the Human Gene Mutation Database (HGMD) include five classifications: disease-associated and putatively functional polymorphisms (DP, N = 45), disease-associated polymorphisms with additional support (DFP, N = 40), in vitro or in vivo functional polymorphisms (FP, N = 41), frameshift or truncating variant (FTV, N = 1) and disease-causing mutations (DM, N = 170). We explored the relationship between HGMD class and SIFT and PolyPhen predictions and phyloP sequence conservation. We found no relationship between HGMD class and functional predictions (Fisher $p > 0.05$). However, we did observe that phyloP scores differed significantly among classes. DP and DFP classes had median scores of 0.36 and 0.48, respectively, with their third quartiles falling below the medians of the other three classes. FP and DM were distributed similarly with medians of 1.2 and 1.4 (Wilcoxon $p = 0.90$). The single FTV SNV had a phyloP score of 2.9, nearly the maximum achievable in the placental alignment.

We observed similar patterns in the analysis of the 46 HGMD SNVs observed in this study that were also reported in OMIM compared to those that were not. SIFT and PolyPhen predictions were not associated with OMIM inclusion (Fisher $p > 0.05$), but phyloP score was significantly associated with medians of 2.1 and 1.0 (Wilcoxon $p = 0.025$) for those SNVs that were and were not in OMIM, respectively.

Association analysis results

An alternative strategy to uncover the contribution of a gene to traits of interest is the analysis of rare variants in aggregate (*5*). Two metrics of rare variant burden are the number of rare variants and the cumulative minor allele frequency (cMAF) of rare and potentially deleterious SNVs within each gene. As we saw for the number of variants, the values of rare cMAF across our whole sample irrespective of disease were strongly correlated with the number of sequenced bases per gene ($r^2 = 0.54$). The cMAF ranged from 0 to 3.9% (Figs. 1F, S4). Among genes with the lowest cMAF, singletons and doubletons accounted for 71% of the cMAF (Fig. 1E); among genes with the highest rare variant cMAF singletons and doubletons accounted for 25% of the cMAF.

The high level summary of the association results are presented as quantile-quantile plots for each disease in fig. S6. These plots and corresponding genomic control λs illustrate that even with principal components in the model, some type I error inflation remains in several of the common and rare variant results. The common and rare variant λs were significantly correlated (>0.45) and the average λ was very similar amongst the three tests presented (1.21, 1.31 and 1.22 for common, rare amino acid changing and rare functional, respectively), suggesting that the effect of population structure was less well controlled for rare as opposed to common variant

tests with this study design. The drop in average λ for the rare functional test is likely due to decreased power tied to lower cumulative minor allele frequencies (cMAFs).

Considering the number of tests conducted for each disease (1010), and excluding osteoarthritis due to the severe type I error inflation, there were five associations that were statistically significant ($\alpha_{Test} = 0.05/1010 = 5\times10^{-5}$; not accounting for testing multiple diseases), all amongst common variants. Common variants in *BRD2* located with the major histocompatibility complex (MHC) of chromosome 6 were significantly associated with multiple sclerosis and rheumatoid arthritis, both known to have relatively large MHC risk factors. Subsequent analysis with HLA alleles demonstrated that the *BRD2* associations were the result of linkage disequilibrium within the locus. The same was true of an association observed between *NFKBIL1* and multiple sclerosis. Three common variants within *CHRNA3* and *CHRNA5*, occurring at the same locus on chromosome 15 previously associated with nicotine dependence and smoking behavior, were significantly associated with COPD. Previous studies of COPD suggest that this association reflects the lack of control for smoking behavior in the choice of population controls. A common S variant within *KCNMA1* was also significantly associated with COPD status (rs45527834, MAF = 0.9%, $p = 1.9\times10^{-6}$, OR = 2.9). The depth and quality of this variant were high with very few missing genotypes. Given the low frequency of this variant, and its absence from HapMap sample genotypes, it is not likely to be well tagged by current genotyping platforms and could have escaped detection from previous GWAS. However, no other variants in this gene, common or rare, showed any evidence of association with COPD status. None of the rare variant tests satisfied the disease-specific statistical significance threshold.

We next investigated any insight that may be gained from having sequenced genes that had been previously shown to have common variant associations with the diseases studied here. We identified the set of candidate genes by filtering the NHGRI GWAS Catalog(*22*) (accessed on August 3, 2011) for associations with genes and diseases that overlap with the current study, shown in table S8. There was an overlap of 13 genes in six diseases. Using these gene-trait pairs as a set of a priori candidate genes, we reevaluated the rare variant association tests adjusting for the number of genes selected within each of the six diseases. There were four resulting associations that were found to be statistically significant ($p < 0.05$ after adjusting for the number of tests for the given disease): the test of amino acid-changing variants in *IL6* and *TNFRSF1A* with multiple sclerosis and *ITGB1* with unipolar depression and the test of functionally damaging variants in *IL6* and multiple sclerosis (table S9).

Of these, the association of variants predicted to be functionally damaging in *IL6* with multiple sclerosis is the most statistically compelling. The cumulative frequency of amino acid-changing variants in IL6, a 212 amino acid protein, was 0.15% overall and 0.06% for those predicted to affect protein function. Cases were five times more likely to carry a NS variant than controls and 12 times more likely to carry one predicted to be functionally damaging. Only one variant, carried by a case, was predicted to be functionally damaging by SIFT and PolyPhen. These results suggest that in addition to the modest effect of at least one common variant located near *IL6* (MAF = 0.05, OR = 0.57) (*76*), carriage of rare variants that affect the amino acid structure increase susceptibility to multiple sclerosis. Given the nature of these very rare variants carried by cases, future replication of this association will be reliant on sequencing the coding regions of additional cases and controls. To have 80% power to replicate the observed association, approximately 700 cases and controls would need to be sequenced.

The absence of compelling rare variant associations led us to estimate the magnitudes of genetic effects that would be required with available sample sizes to have high statistical power to identify significant gene-disease associations. Given the range of cMAFs of rare variants predicted to affect gene function observed (Fig. 1F), in an association study of 1,000 cases and 4,000 population controls (a size approached by half the studies here), only 8% of genes would have enough rare alleles to result in at least 80% power to detect an average odds ratio of 2.5 across selected variants, and only 56% of genes could detect odds ratios of 5 (fig. S5). Sample sizes of more than 10,000 cases and controls would be needed to have adequate power to detect odds ratios on the order 2.5 in at least half of the genes. Currently, little is known about the range of effect sizes that rare variants might have on common disease or to what extent rare variants functional may be enriched in extreme cases or tails of a quantitative distribution. However, the results here suggest the sample sizes required to study them will have to be very large, particularly for genes with small coding regions critical to gene function.

Further acknowledgements

We would like to thank the collaborating clinicians and their research teams, academic consultants and others who have contributed to the recruitment and characterization of subjects in the disease areas included in this study listed below.

We thank the co-primary investigators of the CoLaus study, Gerard Waeber and Peter Vollenweider, and the PI of the PsyColaus Study Martin Preisig. We gratefully acknowledge Yolande Barreau, Anne-Lise Bastian, Binasa Ramic, Martine Moranville, Martine Baumer, Marcy Sagette, Jeanne Ecoffey and Sylvie Mermoud for their role in the CoLaus data collection. The CoLaus study was supported by research grants from GlaxoSmithKline and from the Faculty of Biology and Medicine of Lausanne, Switzerland. The PsyCoLaus study was supported by grants from the Swiss National Science Foundation (#3200B0–105993) and from GlaxoSmithKline (Drug Discovery - Verona, R&D).

We thank the co-primary investigators of the LOLIPOP study: Jaspal Kooner, John Chambers and Paul Elliott. The LOLIPOP study is supported by the National Institute for Health Research Comprehensive Biomedical Research Centre Imperial College Healthcare NHS Trust, the British Heart Foundation (SP/04/002), the Medical Research Council (G0700931), the Wellcome Trust (084723/Z/08/Z) and the National Institute for Health Research (RP-PG-0407-10371).

We thank Metabolic Syndrome GEMs investigators: Scott Grundy, Jonathan Cohen, Ruth McPherson, Antero Kesaniemi, Robert Mahley, Tom Bersot, Philip Barter and Gerard Waeber. We gratefully acknowledge the contributions of the study personnel at each of the collaborating sites: John Farrell, Nicholas Nikolopoulos and Maureen Sutton (Boston); Judy Walshe, Monica Prentice, Anne Whitehouse, Julie Butters, and Tori Nicholls (Australia); Heather Doelle, Lynn Lewis, and Anna Toma (Canada); Kari Kervinen, Seppo Poykko, Liisa Mannermaa, and Sari Paavola (Finland); Claire Hurrel, Diane Morin, Alice Mermod, Myriam Genoud, and Roger Darioli (Switzerland); Guy Pepin, Sibel Tanir, Erhan Palaoglu, Kerem Ozer, Linda Mahley, and Aysen Agacdiken (Turkey); and Deborah A. Widmer, Rhonda Harris, and Selena Dixon (United States). Funding for the GEMS study was provided by GlaxoSmithKline.

We thank the primary investigator, Stephen E. Epstein, and all the staff who were involved in recruitment of patients for MedStar cohort: Mary Susan Burnett, Joseph M. Devaney, Kenneth M. Kent, Joseph M. Lindsay, Augusto D. Pichard, Lowell Satler, Ron Waksman. Recruitment of the MedStar cohort was supported by a research grant from GlaxoSmithKline.

We thank the co-primary investigators of the Genetics of Generalized Osteoarthritis ('GOGO') Study: V.B Kraus (Duke University Medical Center, Durham, NC); J.M Jordan, JB Renner (Thurston Arthritis Research Center, University of North Carolina, Chapel Hill, NC); M. Doherty (University of Nottingham, Nottingham, UK); A.G. Wilson (University of Sheffield, Sheffield, UK); R. Moskowitz, M. Hooper (Case Western Reserve University, Cleveland, OH); M Hochberg (University of Maryland, Baltimore, MD); R Loeser (Wake Forest University of Medicine, Winston-Salem, NC), and U. Atif (Genetics GlaxoSmithKline).

We thank Theadore Ptak, Sandra French (Toronto Digestive Disease Associates, Toronto, ON, Canada), Mark Silverberg, Lori Badadjay (Mount Sinai Hospital, Toronto, ON, Canada), and Yehuda Ringel, Sarah Causey, Sarah Yeskel (University of North Carolina at Chapel Hill, Chapel Hill, NC) for their work on the Irritable Bowel Syndrome Study.

We thank the primary investigator, A.G. Wilson (University of Sheffield, Sheffield, UK), for his work on the Genetics of Rheumatoid Arthritis (GORA) Study.

We thank the following people from the GeneMSA study: Jorge R. Oksenberg, Stephen L. Hauser, Joanne Wang, Sergio E. Baranzini, Daniel Pelletier, Pamela Qualley, Robin R. Lincoln, Refujia Gomez, Michaela F. George, Hourieh Mousavi, Rosa Guerrero, Wendy Chin, Ari Green, Emmanuelle Waubant, Darin T. Okuda, Bruce A. C. Cree (University of California at San Francisco, California, USA); Ludwig Kappos, Yvonne Naegelin, Ernst-Wilhelm Radue, Raija L.P. Lindberg, Jens Kuhle, Achim Gass (University Hospital Basel, Basel, Switzerland); and Chris H. Polman, Frederik Barkhof, Bernard Uitdehaag, Jeroen Geurts, Madeleine Sombekke, Jolijn Kragt, Hugo Vrenken (Vrije Universiteit Medical Centre, Amsterdam, The Netherlands).

We thank Jorge Oksenberg (UCSF) for DNA samples from African Americans. Sample collection was funded by grants from the National Institute of Health (R01 NS046297) and the National Multiple Sclerosis Society (RG3060C8).

We thank Reetta Kalviainen, Anne-Mari Kantanen (Kuopio Epilepsy Center, Kuopio University Hospital, Kuopio, Finland) and Kai Eriksson (Pediatric Neurology Unit, Tampere University Hospital, Tampere, Finland) for their work on the Epilepsy HitDIP study.

We thank Gunter Kraemer, Thomas Dorn, Jorg Hansen (Swiss Epilepsy Center, Zurich, Switzerland); Bernard Steinhoff (Kork Epilepsy Center, Kehl-Kork, Germany); and Heinz-Gregor Wieser, Dominik Zumsteg, Marcos Ortega (Department of Neurology, University Hospital Zurich, Zurich, Switzerland) for their work on Epilepsy GenEpa Study.

We thank Michael Borrie (Parkwood Hospital, Lawson Research Institute, London, ON, Canada); Andrew Kertesz (St. Joseph's Healthcare Centre, London, ON, Canada); Richard Delisle (Clinique de Neurologie et des Neurochirugie de Trois-Rivieres, Trois-Rivieres, QC, Canada); Luis Fornazzari (St. Michael's Hospital, Department of Psychiatry, Behavioral Neurology, University of Toronto, Toronto, ON, Canada); D. Antonio Guzman, Inge Loy-English (SCO Health Service, Ottawa, ON, Canada); Peter St. George-Hyslop, Ron Keren, John Wherrett (University Health Network, University of Toronto, Toronto, ON, Canada); Howard Feldman, Robin Hsiung (University of British Columbia/Vancouver Hospital and Health Sciences Centre, Vancouver, BC, Canada); and Serge Gauthier (McGill Center for Studies in Aging, Montreal, QC, Canada) for their work on the Alzheimer's Disease GenADA Study.

We thank S. Lucae, B. Müller-Myhsok, and F Holsboer (Max Planck Institute of Psychiatry, Munich, Germany) for their work on the Unipolar Depression Study.

We thank D. St Clair (Department of Mental Health, University of Aberdeen, Aberdeen, United Kingdom); D. Rujescu, I. Giegling (Division of Molecular and Clinical Neurobiology, Department of Psychiatry, Ludwig-Maximilians-University, Munich, Germany); and M.

Maziade (Centre de Recherché, Universite Laval Robert-Giffard, Quebec, Canada) for their work on the Schizophrenia Study.

We thank R. Day, K. Matthews (Centre for Neuroscience, Division of Medical Sciences, University of Dundee, Dundee, United Kingdom); J Strauss, J.L. Kennedy, J.B. Vincent (Centre for Addiction and Mental Health, Toronto, Neurogenetics Section, Toronto, ON, Canada); P. McGuffin and A. Farmer (MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, Denmark Hill, London, United Kingdom) for their work on the Bipolar Disorder Study.

We thank the ECLIPSE steering and scientific committees and investigators. The ECLIPSE Steering Committee: Per Bakke (Norway), Courtney Crim (GlaxoSmithKline USA) Harvey Coxson (Canada), Lisa Edwards (GlaxoSmithKline, USA), David Lomas (UK), William MacNee (UK), Edwin Silverman (USA), Ruth Tal-Singer (Co-chair, GlaxoSmithKline, USA), Jørgen Vestbo (Co-chair, Denmark), Julie Yates (GlaxoSmithKline, USA).

The ECLIPSE Scientific Committee: Alvar Agusti (Spain), Peter Calverley (UK), Bartolome Celli (USA), Courtney Crim (GlaxoSmithKline, USA), Bruce Miller (GlaxoSmithKline, UK), William MacNee (Chair, UK), Stephen Rennard (USA), Ruth Tal-Singer (GlaxoSmithKline, USA), Emiel Wouters (The Netherlands), Julie Yates (GlaxoSmithKline, USA).

The ECLIPSE Investigators: Bulgaria: Yavor Ivanov, Pleven; Kosta Kostov, Sofia. Canada: Jean Bourbeau, Montreal, Que Mark Fitzgerald, Vancouver, BC; Paul Hernandez, Halifax, NS; Kieran Killian, Hamilton, On; Robert Levy, Vancouver, BC; Francois Maltais, Montreal, Que; Denis O'Donnell, Kingston, On. Czech Republic: Jan Krepelka, Praha. Denmark: Jørgen Vestbo, Hvidovre. Netherlands: Emiel Wouters, Horn-Maastricht. New Zealand: Dean Quinn, Wellington. Norway: Per Bakke, Bergen. Slovenia: Mitja Kosnik, Golnik. Spain: Alvar Agusti, Jaume Sauleda, Palma de Mallorca. Ukraine: Yuri Feschenko, Kiev; Vladamir Gavrisyuk, Kiev; Lyudmila Yashina, Kiev; Nadezhda Monogarova, Donetsk. United Kingdom: Peter Calverley, Liverpool; David Lomas, Cambridge; William MacNee, Edinburgh; Dave Singh, Manchester; Jadwiga Wedzicha, London. United States of America: Antonio Anzueto, San Antonio, TX; Sidney Braman, Providence, RI; Richard Casaburi, Torrance CA; Bart Celli, Boston, MA; Glenn Giessel, Richmond, VA; Mark Gotfried, Phoenix, AZ; Gary Greenwald, Rancho Mirage, CA; Nicola Hanania, Houston, TX; Don Mahler, Lebanon, NH; Barry Make, Denver, CO; Stephen Rennard, Omaha, NE; Carolyn Rochester, New Haven, CT; Paul Scanlon, Rochester, MN; Dan Schuller, Omaha, NE; Frank Sciurba, Pittsburgh, PA; Amir Sharafkhaneh, Houston, TX; Thomas Siler, St. Charles, MO, Edwin Silverman, Boston, MA; Adam Wanner, Miami, FL; Robert Wise, Baltimore, MD; Richard Zu Wallack, Hartford, CT. The ECLIPSE Study was funded by GlaxoSmithKline (Clinicaltrials.gov identifier NCT00292552; GSK study code SCO104960).

We would like to thank the GenKOLS Investigators: Amund Gulsvik and Per Bakke, Bergen Norway. The GenKOLS study was funded by GlaxoSmithKline.

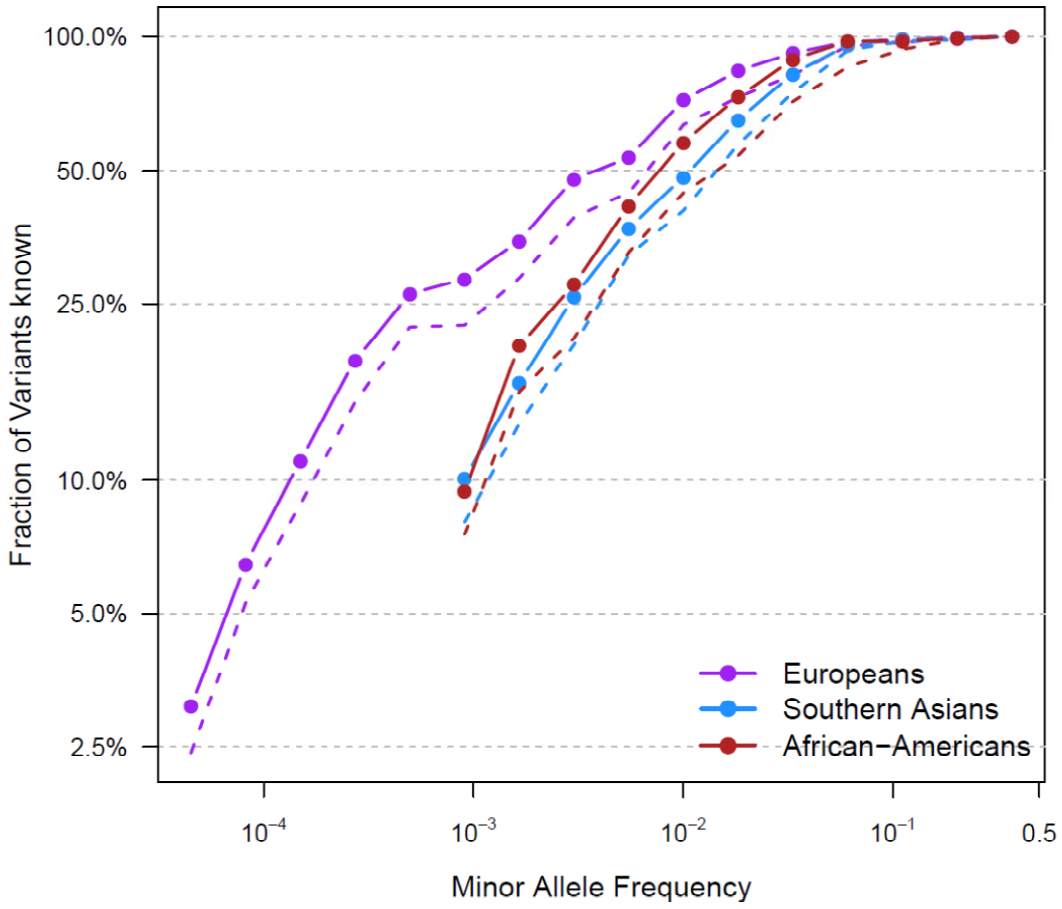

Fig. S1.

The fraction of variants previously reported. The fraction of the variants found in this study already reported in dbSNP 132 (solid line) and after excluding all variants only reported by the 1000 Genomes Project (dashed lines; reflecting the contribution of that study to the catalog of known variants) for Europeans, African Americans and Southern Asians.

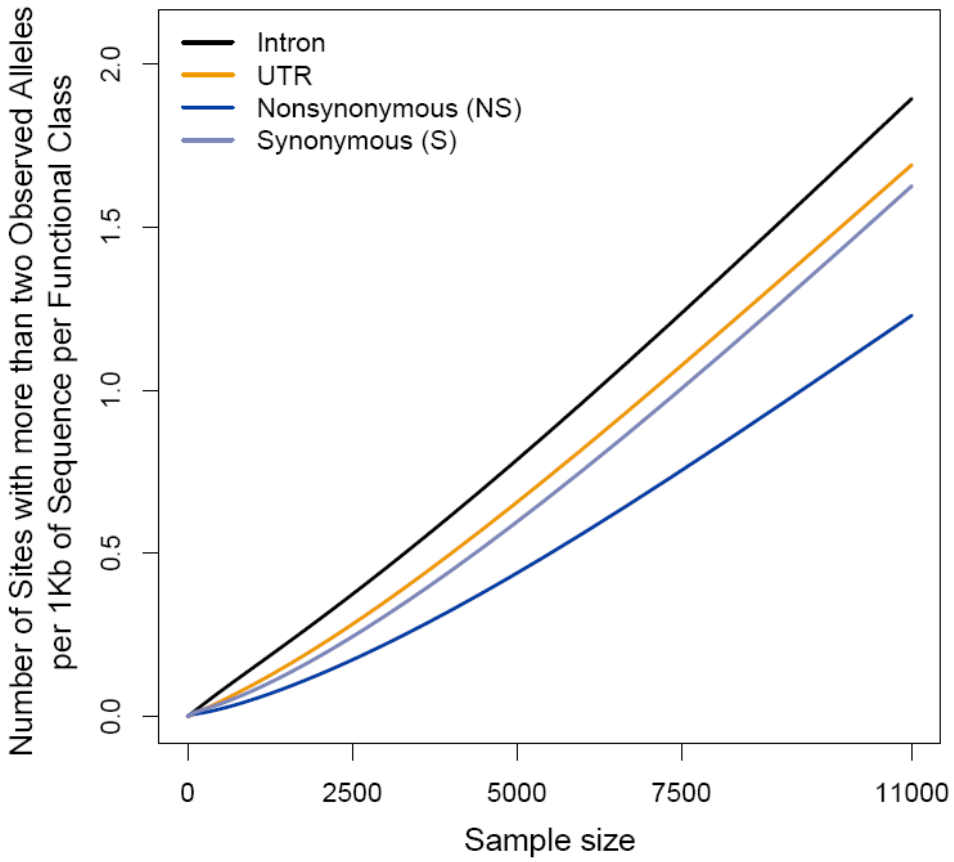

Fig. S2.

Rate of triallele variant discovery in Europeans. Number of triallelic variants discovered per kilobase sequenced with increasing sample size, stratified by variants found within introns, UTR and coding exons. NS and S coding lengths are adjusted by the number of base mutations that could give rise to their respective changes.

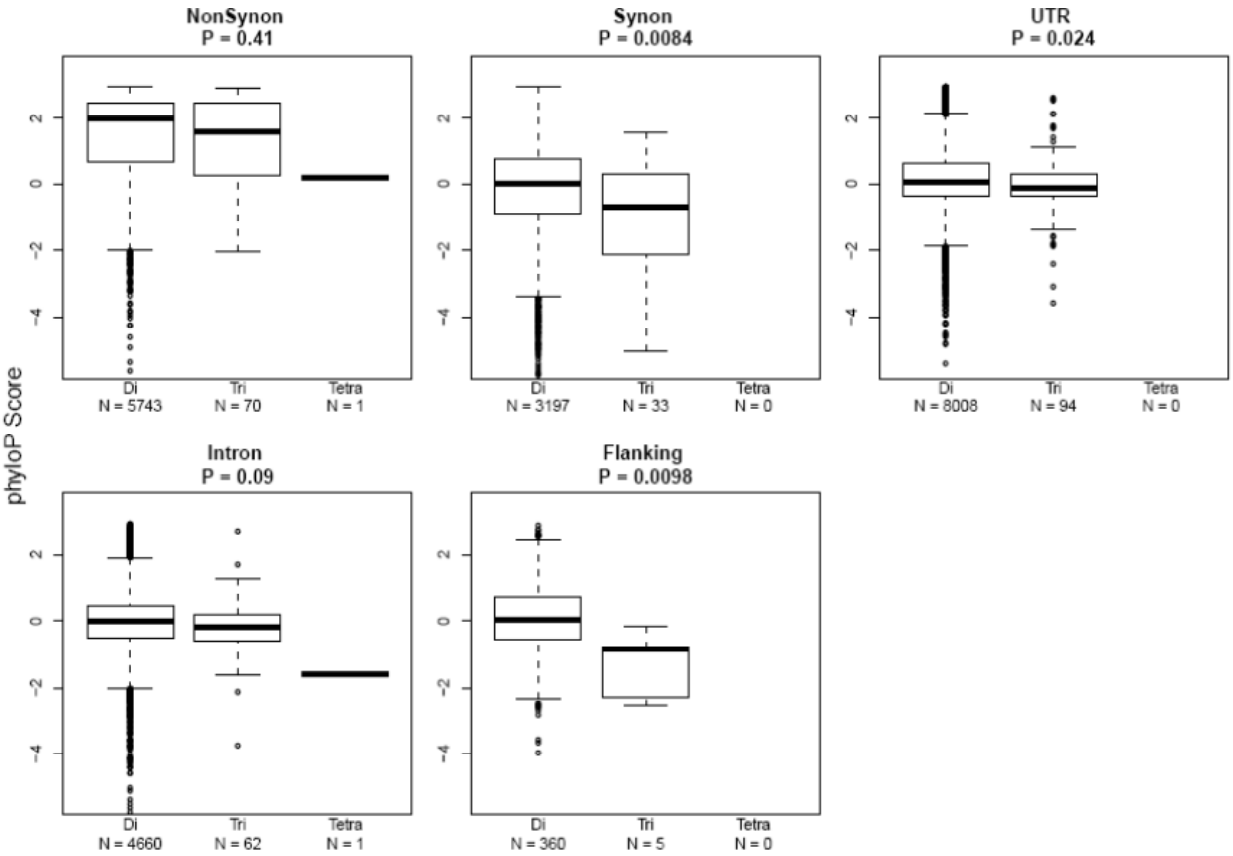

Fig. S3.

Distribution of vertebrate sequence conservation at singleton diallelic, triallelic and tetraallelic positions in Europeans. P-values are the result of the nonparametric Wilcoxon test of homogeneity of location comparing di- and trialleles phyloP scores.

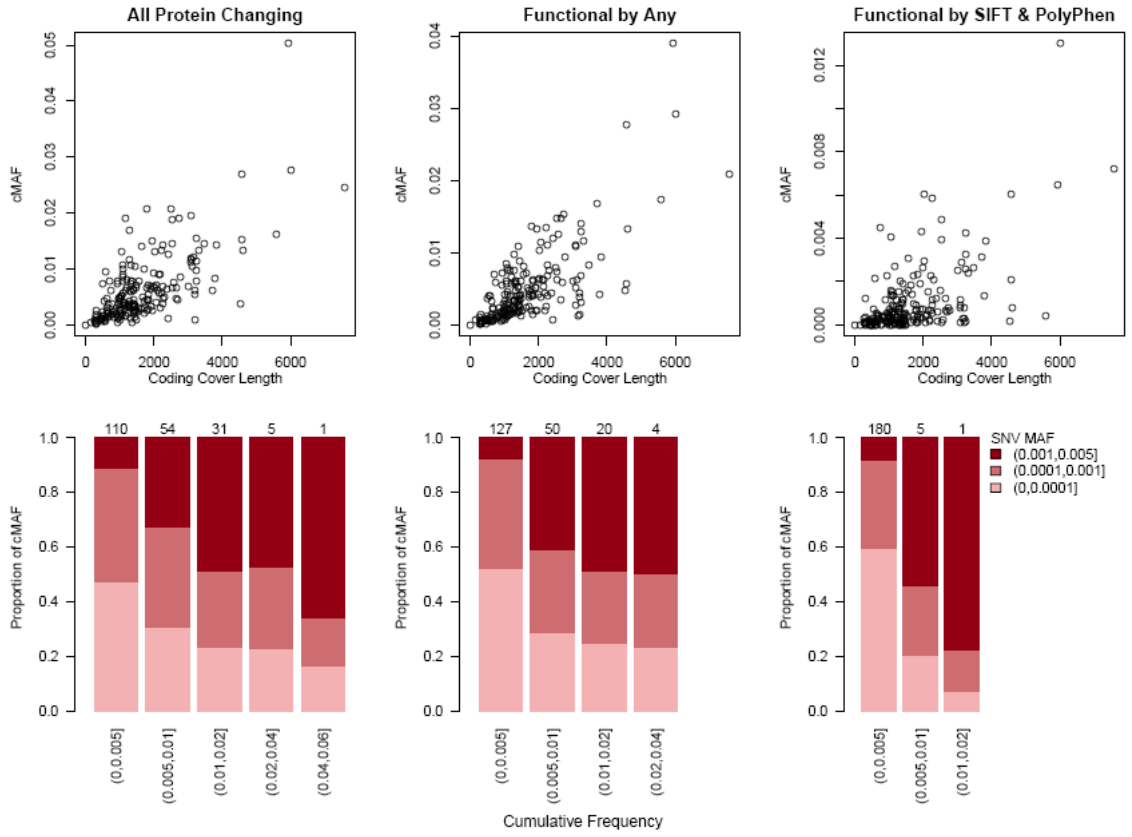

Fig. S4.

Cumulative minor allele frequencies (cMAFs) of rare protein-changing minor alleles. Upper row : Cumulative MAFs of uncommon (MAF ≤ 0.5%) minor alleles are shown for each gene, including all protein-changing variants, variants predicted to be functional by SIFT, PolylPhen or have a phyloP score ≥ 2, or are predicted to be functional by both SIFT and PolyPhen versus their coding length (successfully sequenced). Lower row: The proportion of rare cumulative MAFs shown in the upper row accounted for by variants with frequencies less than or equal to 0.0001, (0.0001,0.001] and (0.001,0.005] from light to dark red. Gene-level cumulative MAFs are binned into five groups as shown on the x axis. The number of genes with cumulative frequencies falling into each bin is shown above each stacked bar. Genes with cumulative MAFs equal to zero are excluded.

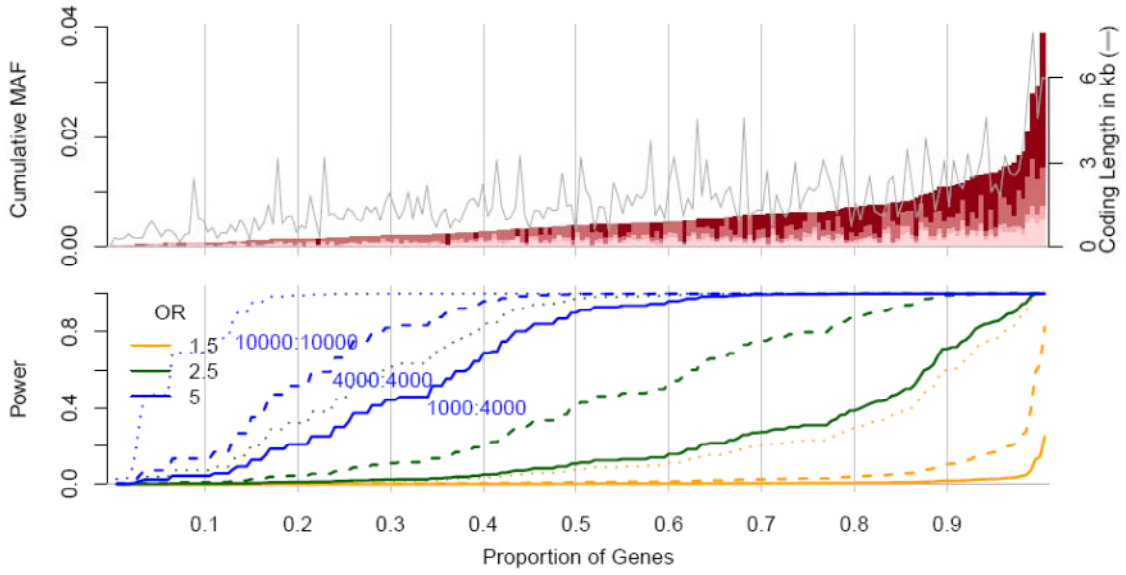

Fig. S5.

Statistical power to conduct a burden test of association. Cumulative rare variant MAF (cMAF) as described in Fig. 1 and the corresponding statistical power to test each gene for association with a binary outcome given the observed cMAF, a condition with a prevalence of 5%, and a significance threshold of 0.05/202 (Bonferroni adjustment for testing all genes in this study). Power is shown with case:control sample sizes of 1000:4000, 4000:4000 and 10,000:10,000.

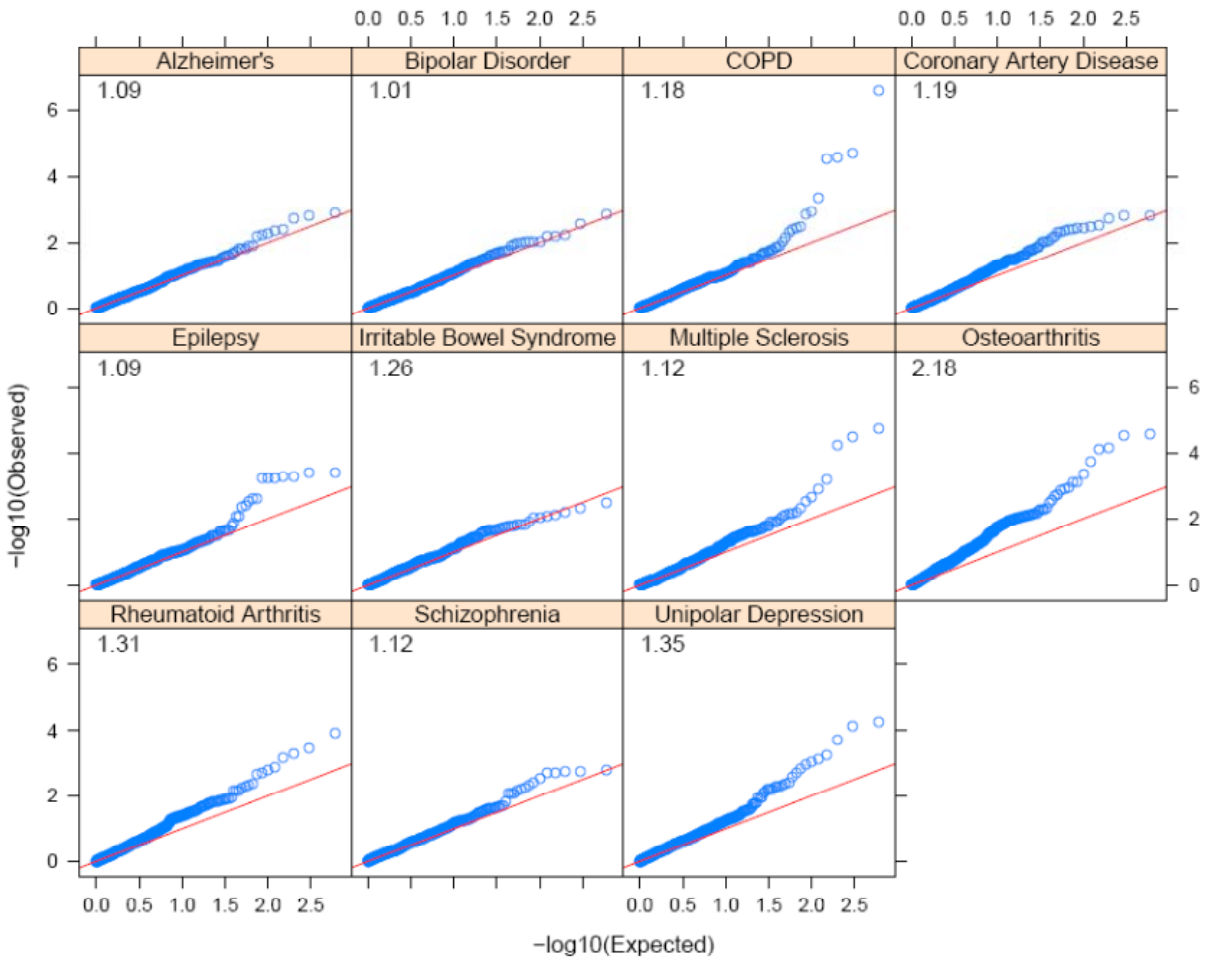

Fig. S6A.

Distribution of case control association p-values for common variants (MAF > 0.5%). Genomic control values (λ) are shown in the upper left corner of each panel.

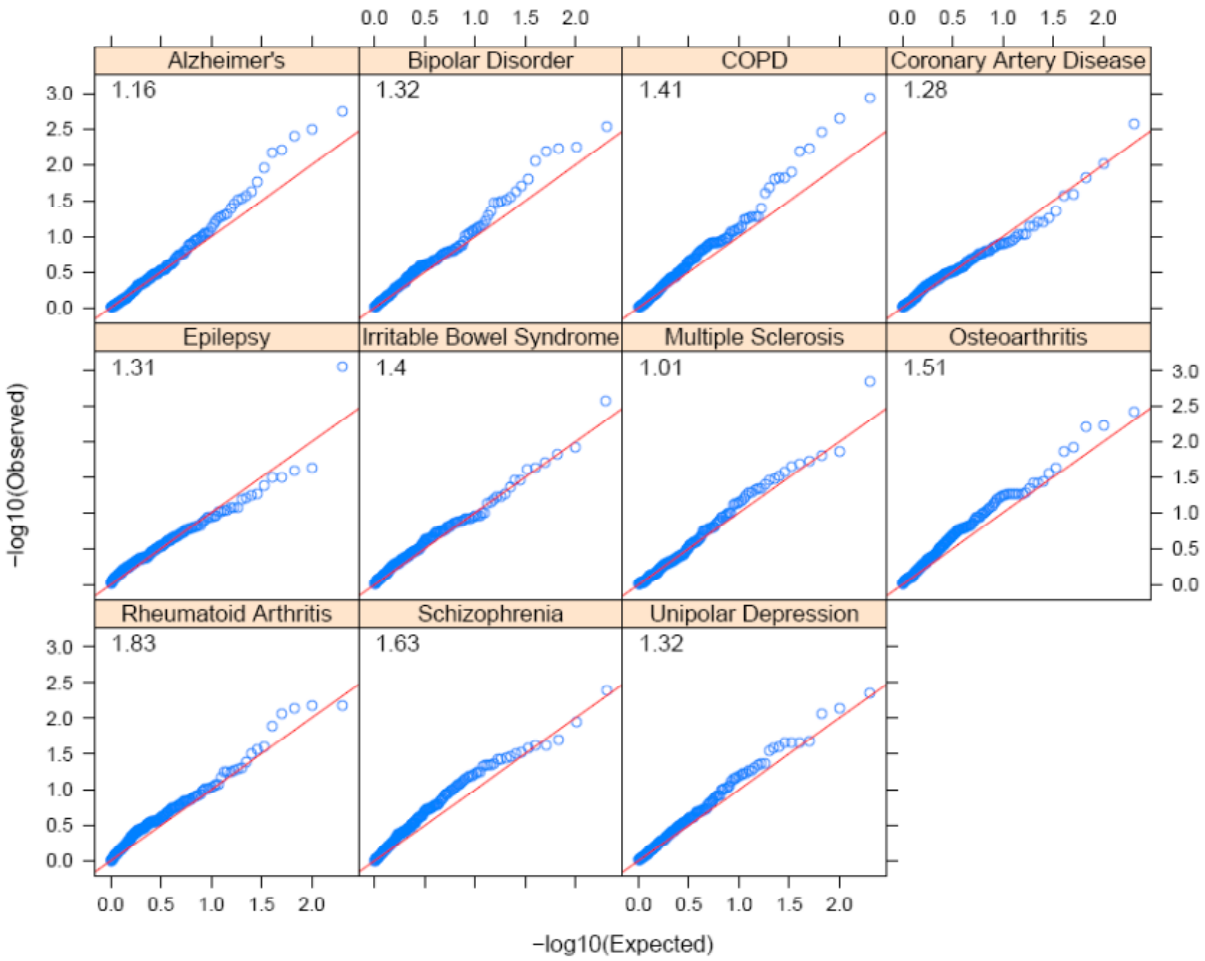

Fig. S6B.

Distribution of case control association p-values for rare amino acid-changing variants (MAF ≤ 0.5%). Genomic control values (λ) are shown in the upper left corner of each panel.

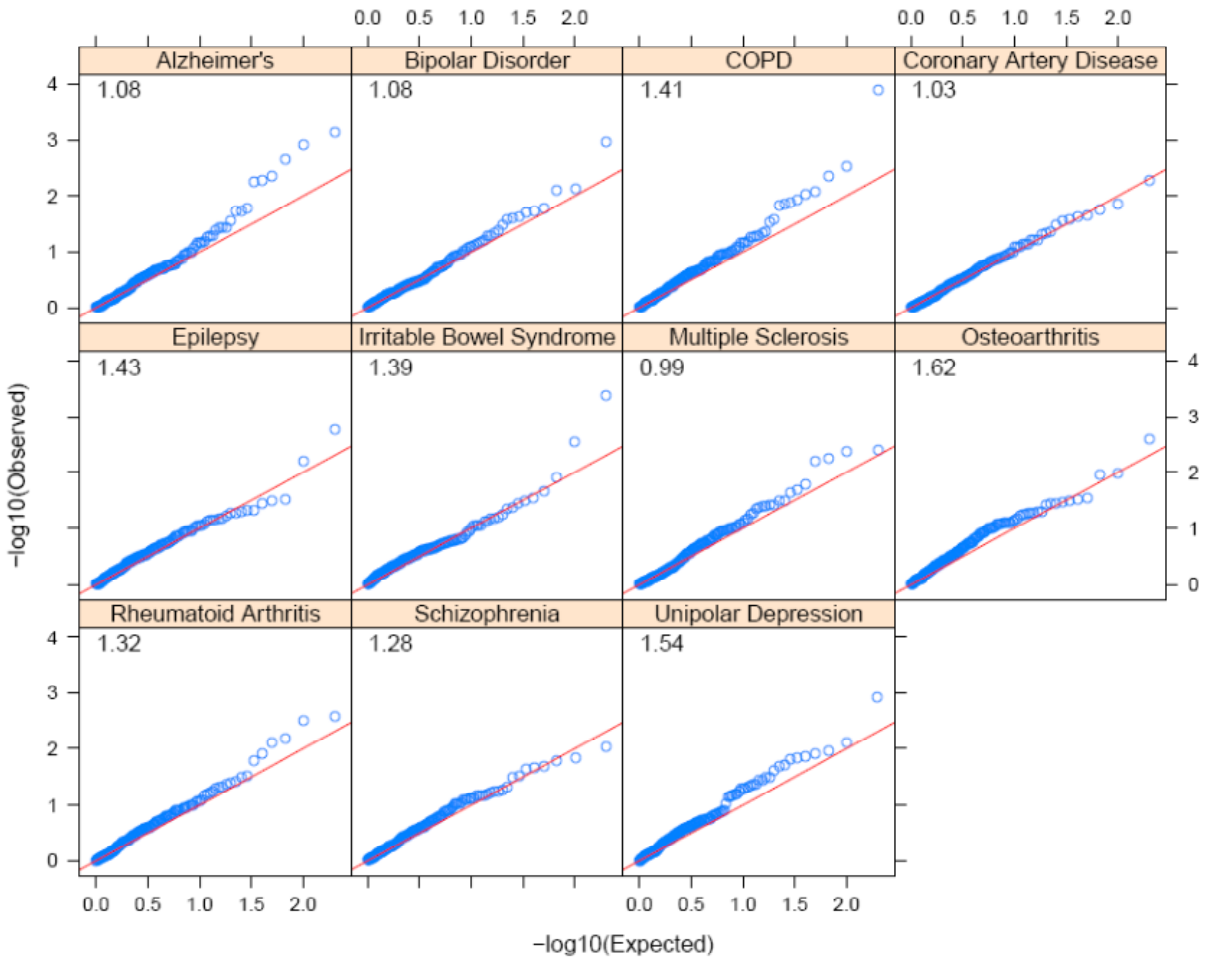

Fig. S6C.

Distribution of case control association p-values for rare amino acid-changing variants (MAF ≤ 0.5%) predicted to be functionally damaging. Genomic control values (λ) are shown in the upper left corner of each panel. Functionally damaging variants were defined as those predicted to be damaging by SIFT or PolyPhen, or occurring at evolutionary conserved base positions (phyloP ≥ 2).

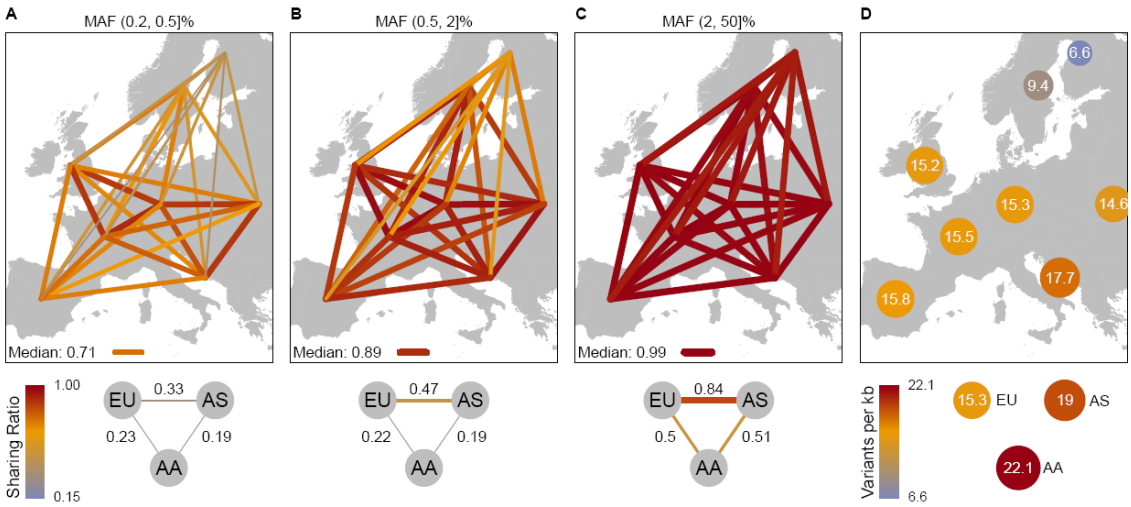

Fig. S7.

Allele sharing and variant abundance. (A-C) The average allele sharing between pairs of populations for variants in different minor allele frequency bins (MAF) computed as the frequency in the pooled population pair. The sharing between African Americans (AA) or Southern Asia (AS) with Europe (EU) is shown as the median value across the comparisons with each individual population in Europe. (D) The number of variants per kilobase found in population samples of 2,500 individuals.

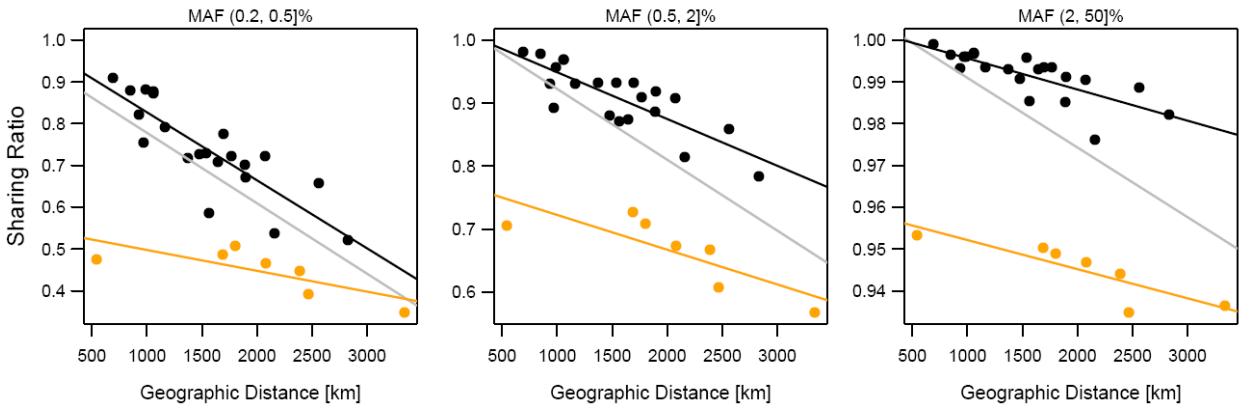

Fig. S8.

Sharing between pairs of European populations decreases with larger geographic distance in Europe. Each dot represents an edge in Figure 3. Sharing between the Finnish and other European population (orange dots) is generally lower than between other pairs of European populations (black dots). Allele sharing decreases with larger geographic distance, independent of MAF bin (gray lines, $p < 0.005$ in all cases), after excluding comparisons with the Finnish (black line, $p < 10^{-4}$) or among the comparisons with the Finnish (orange line, $p < 0.05$).

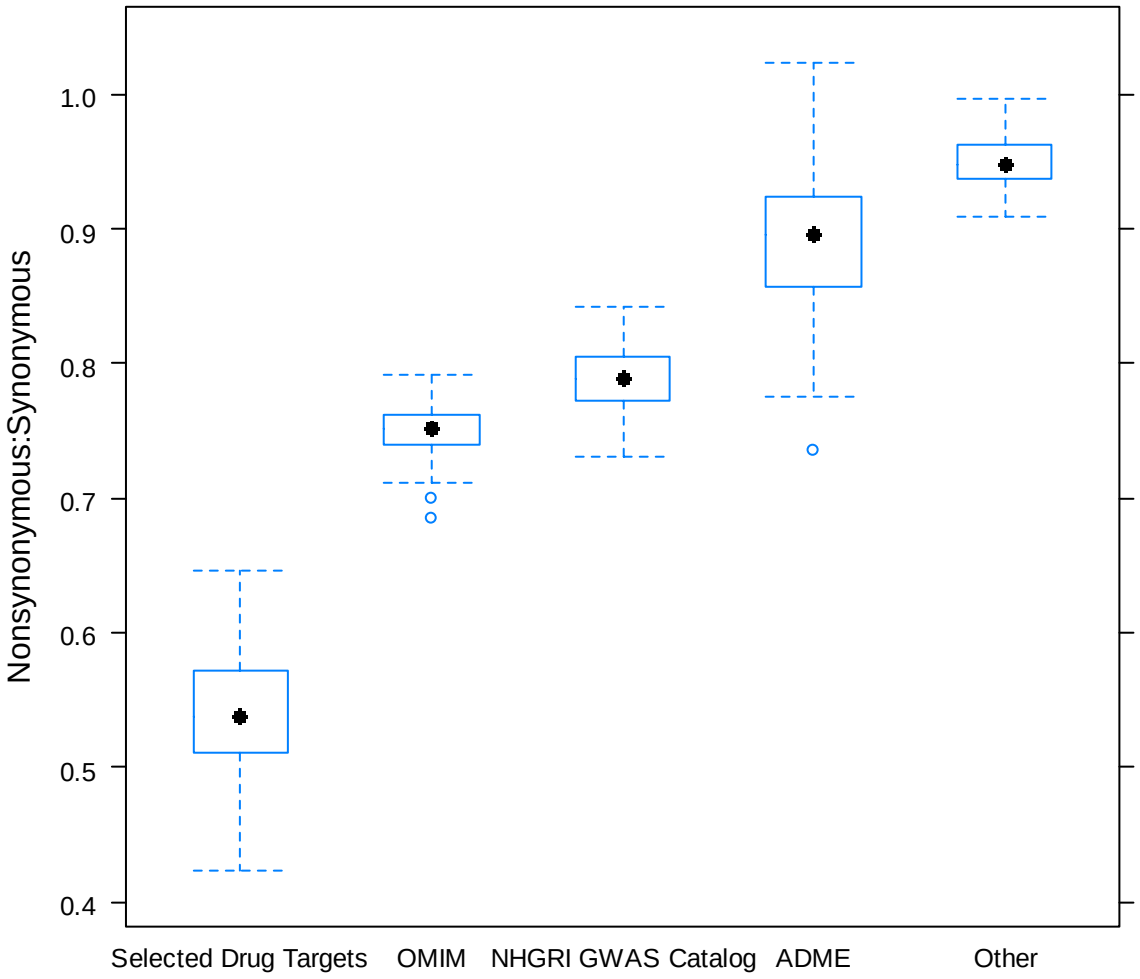

Fig. S9.

Ratio of nonsynonymous to synonymous variant allele carriage between the drug target genes in this study, other groups of genes and the rest of the protein-coding genome. The distribution of the ratio of non-reference NS and S alleles carried by each CEU subject sequenced at low depth by the 1000 Genomes Project. Ratios were computed for the drug target genes in this study versus OMIM, NHGRI GWAS Catalog, ADME (absorption, distribution, metabolism and excretion; see www.pharmaadme.org), and all other coding genes.

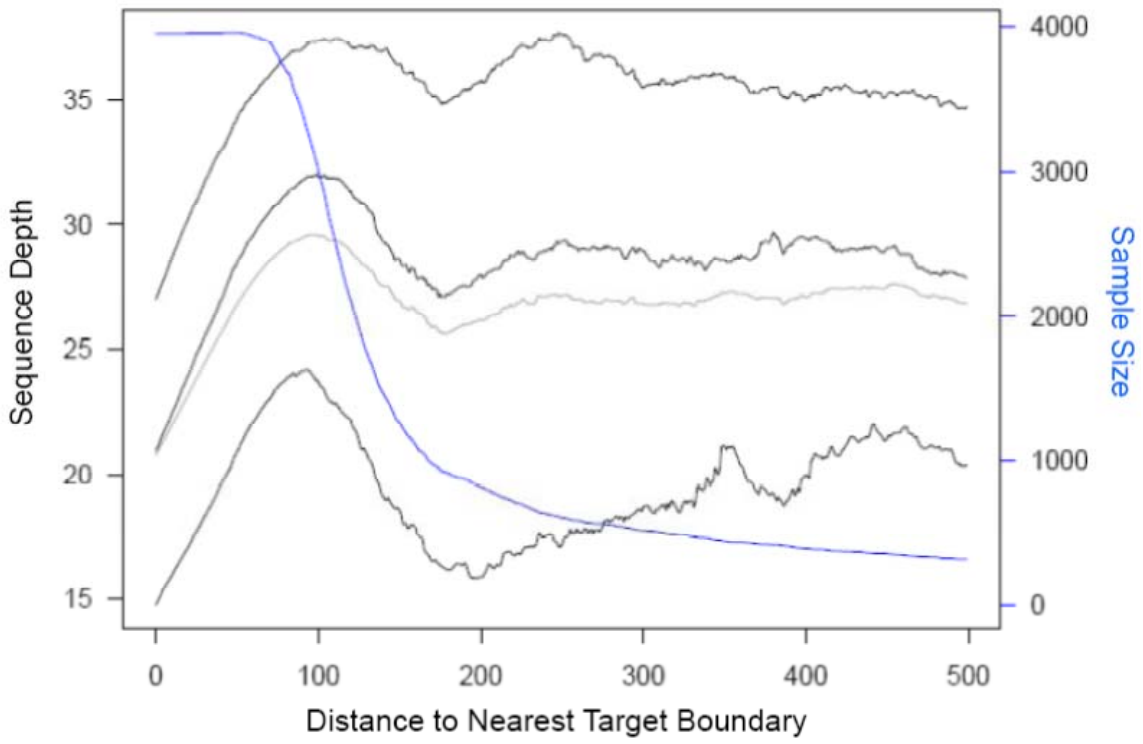

Fig. S10.

Impact of base position relative to target boundaries on sequence depth. Each sequenced base was rescaled by the number of bases from the nearest target boundary (x axis). Depth distribution summaries for each base position are shown on the y axis, with black lines corresponding to the 25th, 50th, and 75th percentiles and the gray line the mean. The blue line corresponds to the number of observations (i.e. sample size) at each base position. The sample size at a distance of one is two times the number of target regions.

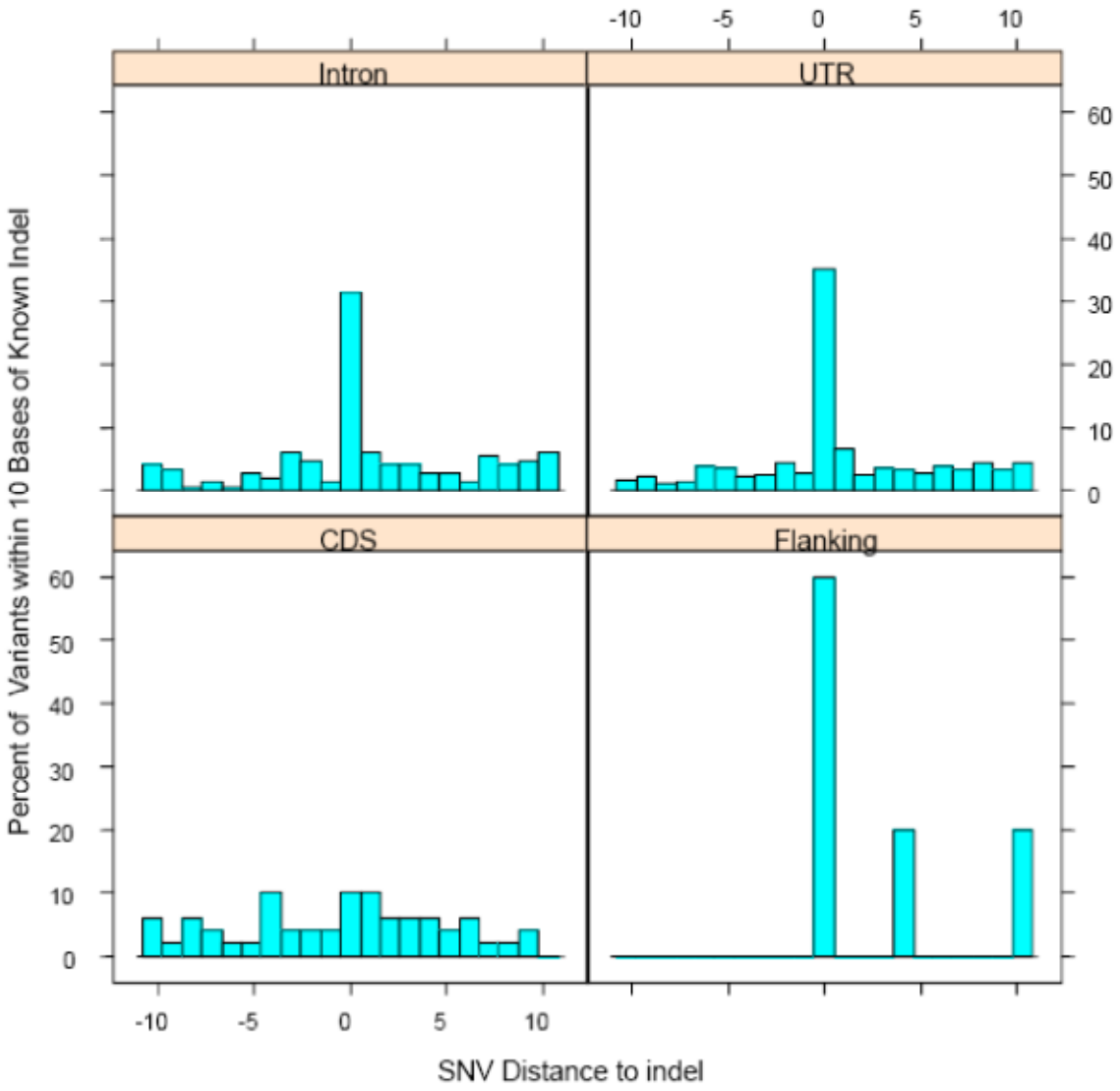

Fig. S11.

Distribution of SNVs within 10 bp of known indels. Variants are divided by location in introns, untranslated region exons (UTR), coding exons (CDS) and gene flanking regions.

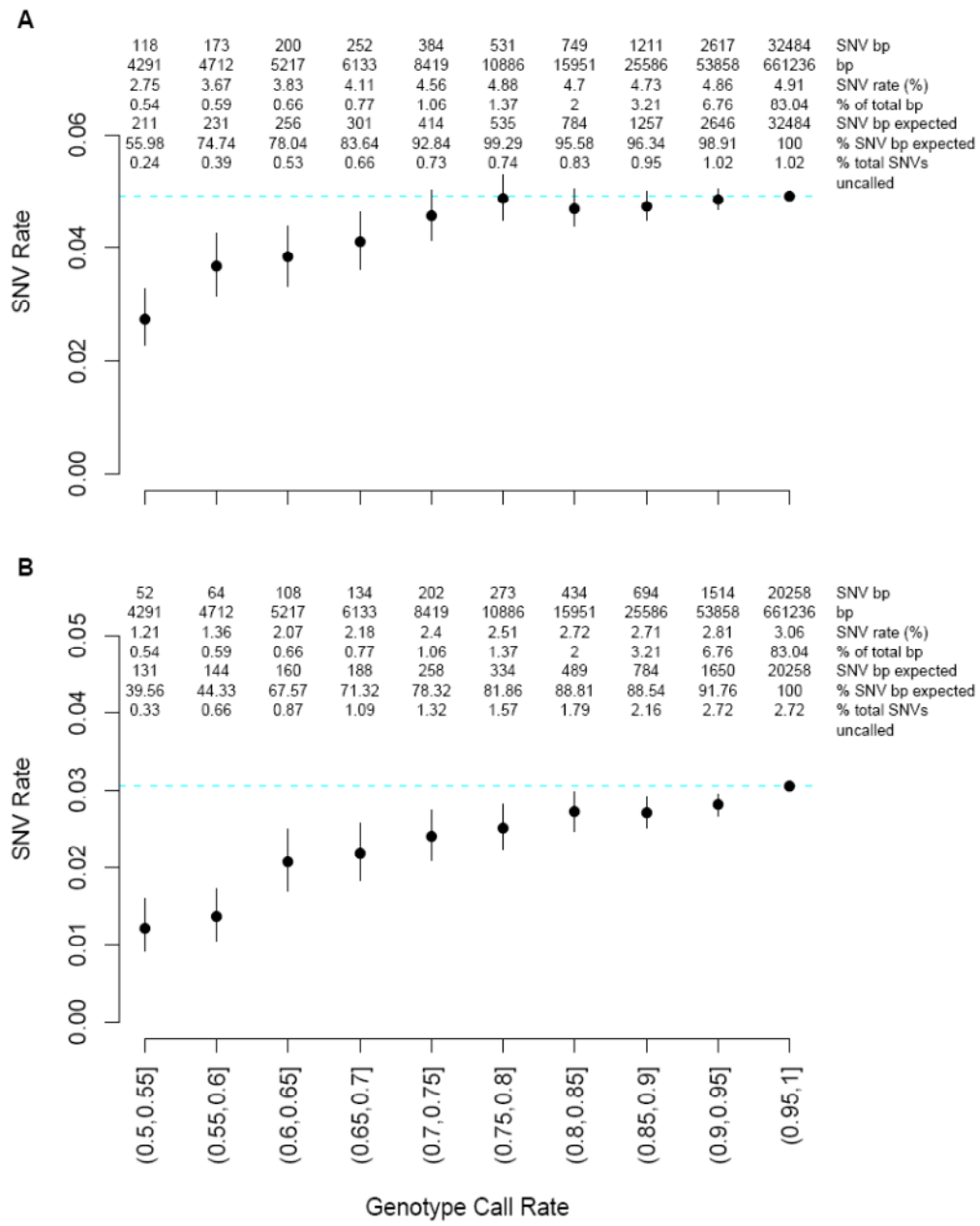

Fig. S12.

Influence of genotype call rates on SNVs discovery rates for (A) all variant and (B) singleton base positions. The mean SNV rate (dot) and 95% exact confidence interval are presented for bases that fall into each genotype call rate bin. Summary statistics within each call rate bin include, in order, the number of SNV positions, the number of sequenced bases (variable and non-variable), SNV rate, percent of sequenced bases (of total with call rate > 0.5), expected number of SNVs based on positions with >95% call rates, the percent of expected SNVs observed, and the cumulative percentage of uncalled SNVs.

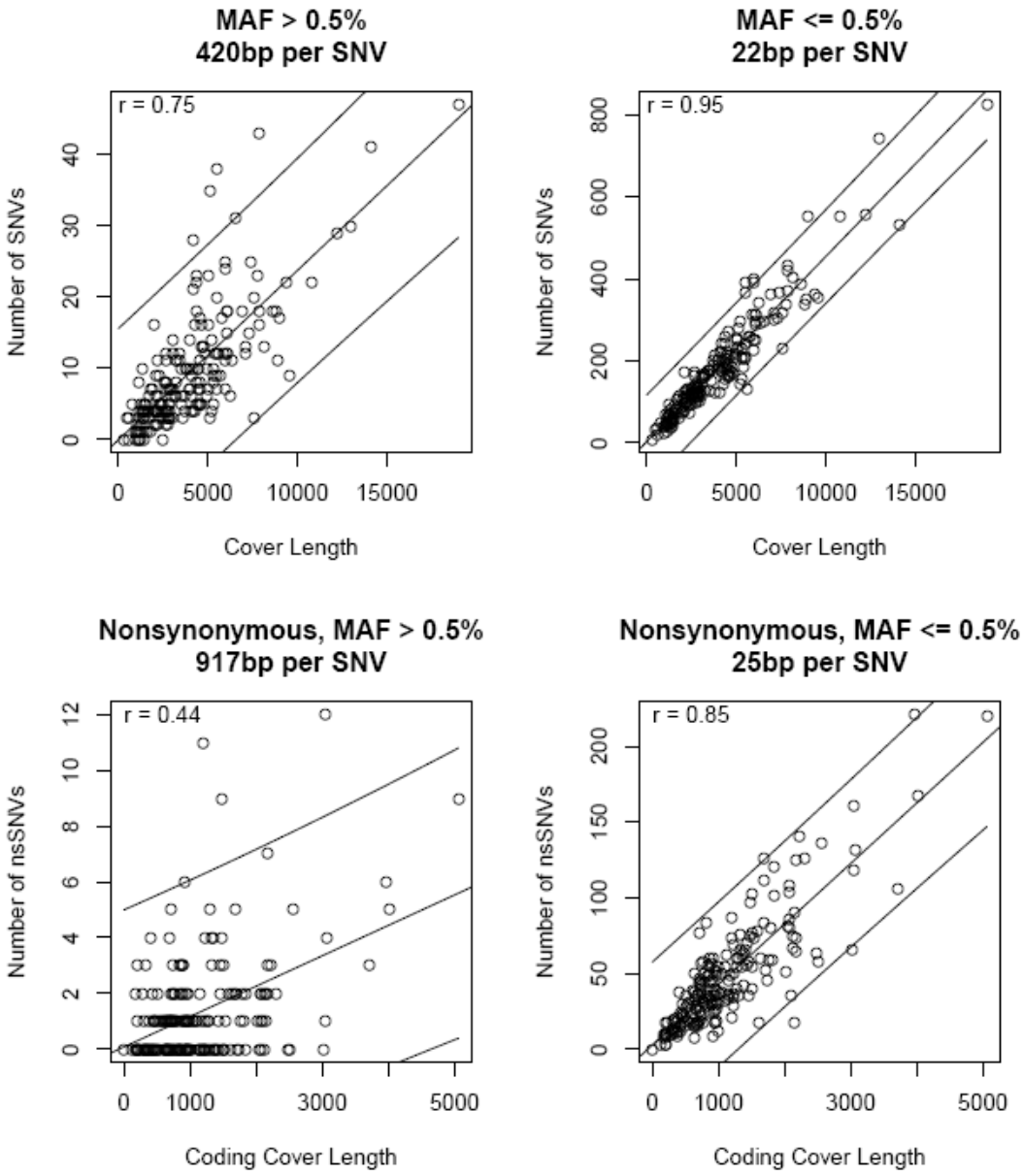

Fig. S13.

Relationship between the length of target sequenced in each gene and the number of SNVs observed in Europeans. The number of SNVs within successfully sequenced target regions is shown for each gene. The top row corresponds to SNVs observed across all target regions for each gene, and the bottom row to only NS SNVs and coding sequence. The slope of the regression line is given as the number of sequenced bases per SNV. Lines correspond to the ordinary least squares regression line and 99% prediction intervals used to identify outliers.

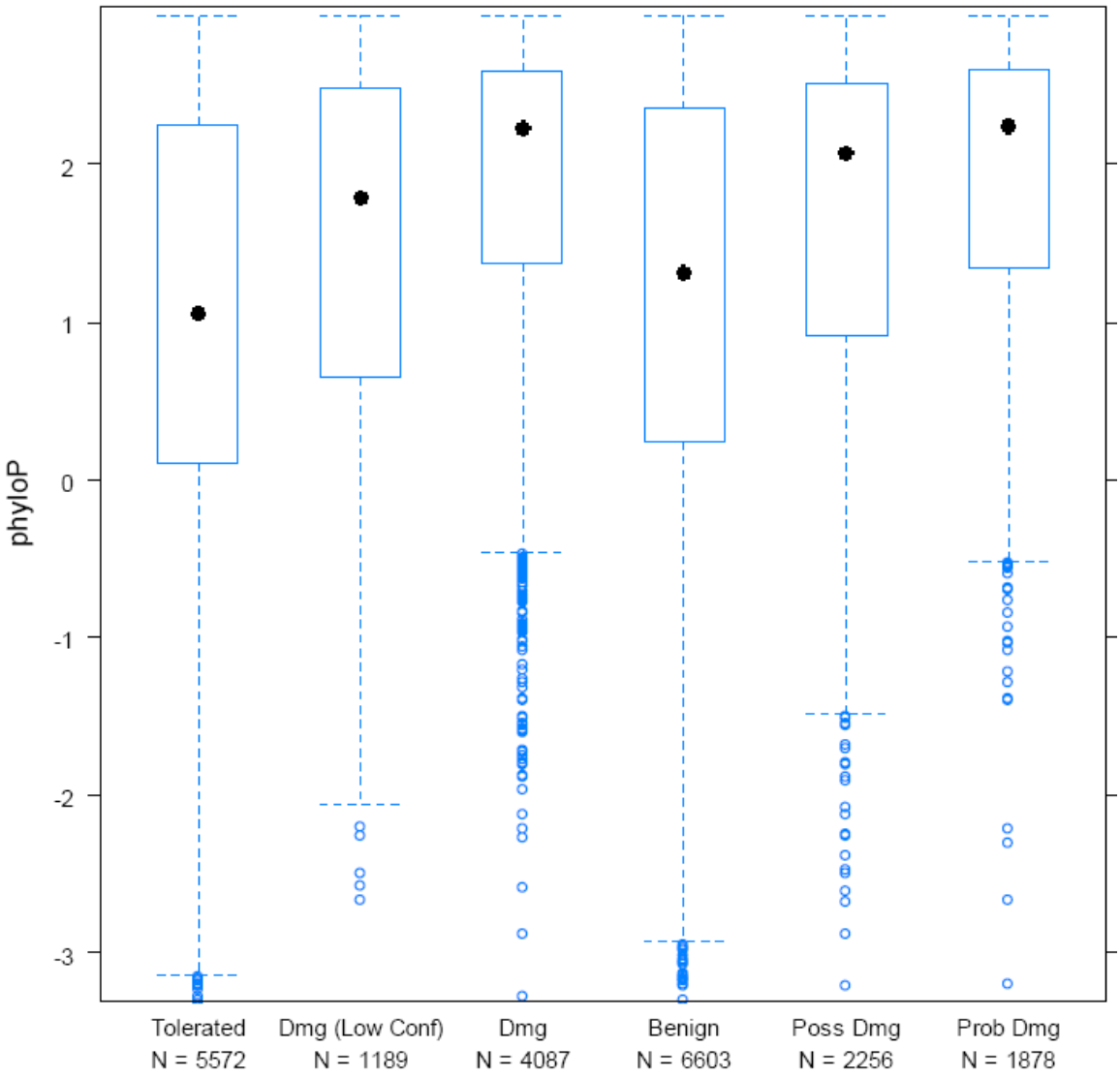

Fig. S14.

Relationship between phyloP conservation scores and SIFT and PolyPhen function predictions. Lower bars show minimum value that falls within 1.5 times interquartile range (IQR), points that fall below this value, boxes show IQR, black points correspond to the sample median, and upper bars extend to maximum value. Y-axis was truncated at -3 to emphasize differences around IQR; minimum values are as low as -9.5. Tolerated, Dmg (Low Conf) and Dmg correspond to SIFT predictions with scores >0.05, ≤0.05 with low confidence and scores ≤0.05, respectively. Benign, Poss Dmg and Prob Dmg correspond to PolyPhen predictions of benign, possibly damaging and probably damaging, respectively.

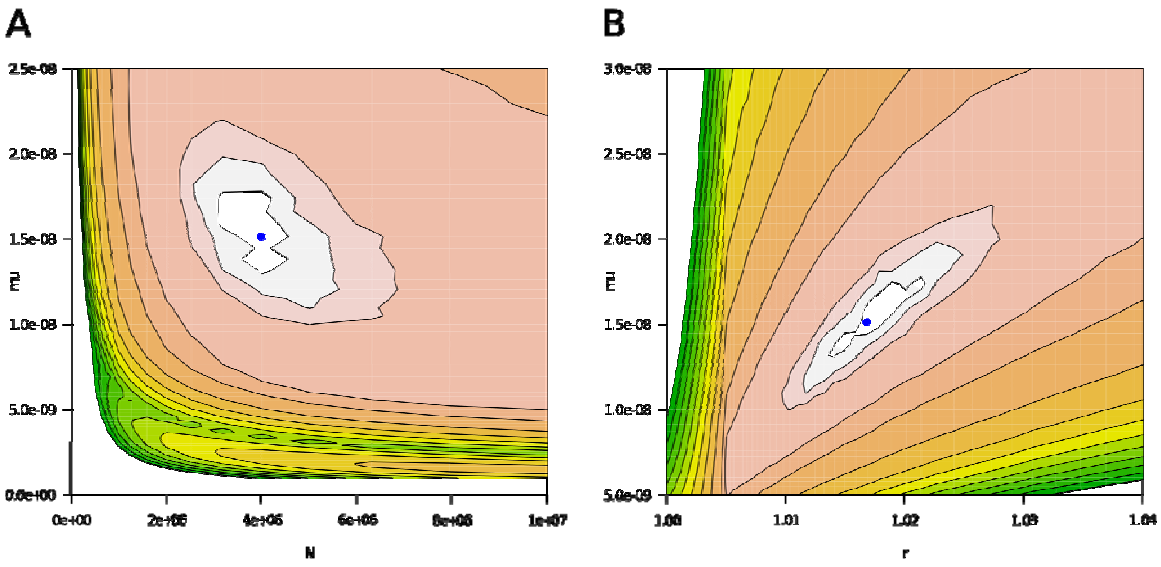

Fig. S15.

Profile likelihood surfaces of N_e and a single global (A) and of r and (B). These surfaces confirm that is an identifiable parameter in this inference scheme (i.e. the likelihood surface has a single point maximum as opposed to the ridge along fixed values of expected in traditional population genetic inference). The blue point depicts the joint MLE and the levels mark -2 10^0, -5 10^0, -10^1, -2 10^1, -5 10^1, -10^2-, ..., -10^5 log-likelihood units lower than the MLE estimate.

Table S1.

Sequenced genes.

Gene	Description	Chrom	Start	End	Length	Coding Length
ABCB1	ATP-binding Cassette, Sub-family B, Member 1	7	86970884	87180500	209616	3843
ADAM10	ADAM Metallopeptidase Domain 10	15	56675802	56829469	153667	2247
ADIPOQ	Adiponectin, C1Q And Collagen Domain Containing	3	188043157	188058946	15789	735
ADORA1	Adenosine A1 Receptor	1	201326405	201403156	76752	981
ADORA2A	Adenosine A2a Receptor	22	22996866	23168325	171460	4593
ADRB3	Adrenergic, Beta-3-, Receptor	8	37939673	37943341	3668	1227
ALOX5AP	Arachidonate 5-lipoxygenase-activating Protein	13	30207669	30236556	28887	486
APCS	Amyloid P Component, Serum	1	157824240	157825285	1045	672
APH1A	Anterior Pharynx Defective 1 Homolog A	1	148504423	148508156	3733	809
APH1B	Anterior Pharynx Defective 1 Homolog B	15	61356844	61385807	28964	774
APP	Amyloid Beta (A4) Precursor Protein	21	26174732	26465003	290271	2313
BDKRB2	Bradykinin Receptor B2	14	95740950	95780542	39592	1176
BICD1	Bicaudal D Homolog 1	12	32151448	32422408	270961	2928
BRD2	Bromodomain Containing 2	6	33044415	33057059	12645	2406
BRD3	Bromodomain Containing 3	9	135887784	135922913	35130	2181
BRD4	Bromodomain Containing 4	19	15209301	15252262	42962	4100
C5AR1	Complement Component 5a Receptor 1	19	52504944	52517167	12223	1053
CACNA1B	Calcium Channel, Voltage-dependent, N Type, Alpha 1B Subunit	9	139892062	140138897	246836	7020
CAMKK2	Calcium/calmodulin-dependent Protein Kinase Kinase 2, Beta	12	120159878	120220494	60616	1773
CASR	Calcium-sensing Receptor	3	123385220	123488032	102812	3237
CCKAR	Cholecystokinin A Receptor	4	26092116	26101140	9024	1287
CCKBR	Cholecystokinin B Receptor	11	6237542	6249932	12390	1344
CCL11	Chemokine (C-C Motif) Ligand 11	17	29636800	29639312	2512	294
CCL7	Chemokine (C-C Motif) Ligand 7	17	29621353	29623373	2021	300
CCL8	Chemokine (C-C Motif) Ligand 8	17	29670168	29672534	2367	300

CCR1	Chemokine (C-C Motif) Receptor 1	3	46218204	46224836	6632	1068
CCR3	Chemokine (C-C Motif) Receptor 3	3	46227186	46283166	55981	1068
CCR5	Chemokine (C-C Motif) Receptor 5	3	46386637	46392701	6064	1059
CCR9	Chemokine (C-C Motif) Receptor 9	3	45903023	45919671	16648	1110
CD28	CD28 Molecule	2	204279443	204310801	31358	663
CD3D	CD3d Molecule, Delta	11	117714999	117718669	3670	516
CD3E	CD3e Molecule, Epsilon	11	117680656	117692100	11445	624
CD3G	CD3g Molecule, Gamma	11	117720311	117729979	9669	549
CD4	CD4 Molecule	12	6768912	6800237	31325	1377
CDH2	Cadherin 2, Type 1, N-cadherin	18	23784933	24011189	226256	2721
CHRM3	Cholinergic Receptor, Muscarinic 3	1	237858996	238139343	280347	1773
CHRM4	Cholinergic Receptor, Muscarinic 4	11	46363216	46364734	1519	1440
CHRNA3	Cholinergic Receptor, Nicotinic, Alpha 3	15	76674706	76700377	25671	1518
CHRNA4	Cholinergic Receptor, Nicotinic, Alpha 4	20	61445109	61463192	18083	1884
CHRNA5	Cholinergic Receptor, Nicotinic, Alpha 5	15	76644961	76673515	28554	1407
CHRNA6	Cholinergic Receptor, Nicotinic, Alpha 6	8	42726920	42742776	15856	1485
CHRNA7	Cholinergic Receptor, Nicotinic, Alpha 7	15	30110018	30248541	138523	1509
CHRNB2	Cholinergic Receptor, Nicotinic, Beta 2	1	152806881	152818978	12097	1509
CLEC16A	C-type Lectin Domain Family 16, Member A	16	10945846	11183547	237702	3162
CNR2	Cannabinoid Receptor 2	1	24073047	24112404	39357	1083
CNTN5	Contactin 5	11	98397081	99732683	1335603	3303
CTSK	Cathepsin K	1	149035311	149047436	12125	990
CXCL1	Chemokine (C-X-C Motif) Ligand 1	4	74953973	74968249	14277	324
CXCL2	Chemokine (C-X-C Motif) Ligand 2	4	75181616	75183861	2245	324
CXCL3	Chemokine (C-X-C Motif) Ligand 3	4	75121170	75123354	2184	324
CXCL5	Chemokine (C-X-C Motif) Ligand 5	4	75080223	75083286	3064	345
CYSLTR1	Cysteinyl Leukotriene Receptor 1	X	77414786	77469743	54957	1014
CYSLTR2	Cysteinyl Leukotriene Receptor 2	13	48178952	48181499	2547	1041
DPP3	Dipeptidyl-peptidase 3	11	66004456	66033706	29250	2214
DPP4	Dipeptidyl-peptidase 4	2	162557001	162639298	82297	2301

DRD2	Dopamine Receptor D2	11	112785527	112851103	65577	1332
DRD3	Dopamine Receptor D3	3	115330247	115380589	50342	1203
DYRK3	Dual-specificity Tyrosine-(Y)-phosphorylation Regulated Kinase 3	1	204875504	204889165	13662	1784
EDG1	Sphingosine-1-phosphate Receptor 1	1	101475043	101479662	4619	1149
EDNRA	Endothelin Receptor Type A	4	148621575	148685555	63980	1284
EDNRB	Endothelin Receptor Type B	13	77367617	77447665	80049	1446
EGR1	Early Growth Response 1	5	137829080	137832903	3823	1632
ELA2	Elastase, Neutrophil Expressed	19	803291	807246	3955	804
EVI5	Ecotropic Viral Integration Site 5	1	92746841	93030549	283708	2433
FAAH	Fatty Acid Amide Hydrolase	1	46632575	46652104	19529	1740
FGF10	Fibroblast Growth Factor 10	5	44340831	44424623	83793	627
FH	Fumarate Hydratase	1	239727526	239749677	22151	1533
GABRA2	Gamma-aminobutyric Acid (GABA) A Receptor, Alpha 2	4	45946341	46086702	140362	1356
GABRA3	Gamma-aminobutyric Acid (GABA) A Receptor, Alpha 3	X	151086290	151370993	284704	1479
GHSR	Growth Hormone Secretagogue Receptor	3	173645617	173648940	3324	1175
GJD2	Gap Junction Protein, Delta 2, 36kDa	15	32831934	32834074	2141	966
GLP1R	Glucagon-like Peptide 1 Receptor	6	39124595	39163498	38903	1392
GPBAR1	G Protein-coupled Bile Acid Receptor 1	2	218833983	218836826	2843	992
GPR109A	G Protein-coupled Receptor 109A	12	121751793	121753857	2064	1092
GPR119	G Protein-coupled Receptor 119	X	129346095	129347102	1007	1008
GRIN1	Glutamate Receptor, Ionotropic, N-methyl D-aspartate 1	9	139152663	139183029	30367	2886
GRIN2B	Glutamate Receptor, Ionotropic, N-methyl D-aspartate 2B	12	13605411	14024319	418908	4455
GRM5	Glutamate Receptor, Metabotropic 5	11	87880626	88420838	540213	3543
GSK3B	Glycogen Synthase Kinase 3 Beta[Homo Sapiens]	3	121028233	121295954	267722	1302
HCRTR1	Hypocretin Receptor 1	1	31855888	31865508	9621	1278
HCRTR2	Hypocretin Receptor 2	6	55147025	55255377	108353	1335
HHIP	Hedgehog Interacting Protein	4	145786623	145879337	92715	2103
HRH1	Histamine Receptor H1	3	11153779	11280243	126465	1464
HRH3	Histamine Receptor H3	20	60223421	60228718	5297	1338
HTR1A	5-hydroxytryptamine (serotonin) Receptor 1A	5	63292034	63293302	1268	1269

HTR1B	5-hydroxytryptamine (serotonin) Receptor 1B	6	78228641	78229900	1260	1173
HTR2C	5-hydroxytryptamine (serotonin) Receptor 2C	X	113724807	114050880	326074	1377
HTR4	5-hydroxytryptamine (serotonin) Receptor 4	5	147810788	148013934	203146	1489
HTR6	5-hydroxytryptamine (serotonin) Receptor 6	1	19864367	19878642	14275	1323
IKBKB	Inhibitor Of Kappa Light Polypeptide Gene Enhancer In B-cells, Kinase Beta	8	42247986	42309122	61136	2271
IL13	Interleukin 13	5	132021764	132024700	2936	441
IL18	Interleukin 18	11	111519186	111540050	20865	582
IL1R1	Interleukin 1 Receptor, Type I	2	102125678	102162766	37089	1710
IL23A	Interleukin 23, Alpha Subunit P19	12	55018926	55020461	1536	570
IL28B	Interleukin 28B	19	44426033	44427609	1577	591
IL4	Interleukin 4	5	132037272	132046267	8995	462
IL5	Interleukin 5	5	131905035	131907113	2078	405
IL6	Interleukin 6	7	22732028	22738145	6118	639
IL7R	Interleukin 7 Receptor	5	35892748	35912681	19933	1380
IL8	Interleukin 8	4	74825139	74828297	3158	300
IL8RB	Chemokine (C-X-C Motif) Receptor 2	2	218698991	218710220	11229	1083
ITGA4	Integrin, Alpha 4	2	182029864	182110719	80855	3099
ITGAV	Integrin, Alpha V	2	187163045	187253873	90828	3147
ITGB1	Integrin, Beta 1	10	33229326	33287204	57878	2627
JAK3	Janus Kinase 3	19	17797961	17819800	21839	3375
KCNC2	Potassium Voltage-gated Channel, Shaw-related Subfamily, Member 2	12	73720163	73889778	169615	1979
KCNMA1	Potassium Large Conductance Calcium-activated Channel, Subfamily M, Alpha Member 1	10	78299366	79067757	768392	3574
KCNN4	Potassium Intermediate/small Conductance Calcium-activated Channel, Subfamily N, Member 4	19	48962525	48977249	14724	1284
KIAA1967	KIAA1967	8	22518202	22533929	15727	2772
L1CAM	L1 Cell Adhesion Molecule	X	152780163	152804802	24640	3774
LDHA	Lactate Dehydrogenase A	11	18372687	18385969	13282	999
LEP	Leptin	7	127668567	127684917	16350	504
LRRK2	Leucine-rich Repeat Kinase 2	12	38905081	39049354	144273	7584
MAG	Myelin Associated Glycoprotein	19	40474868	40496547	21680	1914
MAPK11	Mitogen-activated Protein Kinase 11	22	49044269	49050949	6681	1095

MAPK14	Mitogen-activated Protein Kinase 14	6	36103551	36186513	82962	1216
MCHR1	Melanin-concentrating Hormone Receptor 1	22	39405045	39408764	3720	1269
MCHR2	Melanin-concentrating Hormone Receptor 2	6	100474507	100548835	74328	1023
METAP2	Methionyl Aminopeptidase 2	12	94391953	94433746	41793	1437
MIF	Macrophage Migration Inhibitory Factor	22	22566565	22567409	844	348
MLNR	Motilin Receptor	13	48692475	48694514	2039	1239
MME	Membrane Metallo-endopeptidase	3	156280130	156384212	104082	2253
MMP12	Matrix Metallopeptidase 12	11	102238674	102250922	12248	1412
MMP9	Matrix Metallopeptidase 9	20	44070954	44078607	7653	2124
MS4A1	Membrane-spanning 4-domains, Subfamily A, Member 1	11	59979858	59994801	14944	894
NCSTN	Nicastrin	1	158579687	158595366	15679	2130
NFKBIL1	Nuclear Factor Of Kappa Light Polypeptide Gene Enhancer In B-cells Inhibitor-like 1	6	31622626	31634585	11960	1146
NLRP1	NLR Family, Pyrin Domain Containing 1	17	5345443	5428556	83113	4493
NLRP3	NLR Family, Pyrin Domain Containing 3	1	245646098	245679033	32935	3111
NMNAT2	Nicotinamide Nucleotide Adenylyltransferase 2	1	181484001	181654360	170360	994
NOS2A	Nitric Oxide Synthase 2, Inducible	17	23107919	23151682	43763	3462
NR1D1	Nuclear Receptor Subfamily 1, Group D, Member 1	17	35502567	35510499	7932	1845
NRXN1	Neurexin 1	2	50000992	51113178	1112187	4693
NTRK2	Neurotrophic Tyrosine Kinase, Receptor, Type 2	9	86473286	86828325	355039	2584
OPRK1	Opioid Receptor, Kappa 1	8	54300829	54326747	25918	1143
OPRM1	Opioid Receptor, Mu 1	6	154402136	154609693	207557	1488
OSM	Oncostatin M	22	28988818	28992840	4022	759
OXTR	Oxytocin Receptor	3	8767094	8786300	19206	1170
P2RX7	Purinergic Receptor P2X, Ligand-gated Ion Channel, 7	12	120055061	120108259	53198	1788
P4HA1	Prolyl 4-hydroxylase, Alpha Polypeptide I	10	74436981	74526630	89650	1676
P4HA2	Prolyl 4-hydroxylase, Alpha Polypeptide II	5	131555430	131591455	36026	1668
P4HB	Prolyl 4-hydroxylase, Beta Polypeptide	17	77394323	77411833	17510	1527
PDE4A	Phosphodiesterase 4A, CAMP-specific	19	10392333	10441306	48974	2969

PDE5A	Phosphodiesterase 5A, CGMP-specific	4	120634998	120769429	134431	2654
PGK1	Phosphoglycerate Kinase 1	X	77246425	77268980	22555	1254
PIK3CA	Phosphoinositide-3-kinase, Catalytic, Alpha Polypeptide	3	180349005	180435194	86189	3207
PLA2G7	Phospholipase A2, Group VII	6	46780068	46811069	31002	1326
PPARD	Peroxisome Proliferator-activated Receptor Delta	6	35418313	35503933	85621	1326
PRKAG1	Protein Kinase, AMP-activated, Gamma 1 Non-catalytic Subunit	12	47682322	47698863	16542	996
PSEN1	Presenilin 1	14	72672908	72756862	83955	1404
PSEN2	Presenilin 2	1	225124896	225150429	25534	1347
PSENEN	Presenilin Enhancer 2 Homolog (C. Elegans)	19	40928334	40929743	1409	306
PTGDR	Prostaglandin D2 Receptor	14	51804181	51813192	9011	1080
PTGER1	Prostaglandin E Receptor 1	19	14444278	14447174	2896	1209
PTGES	Prostaglandin E Synthase	9	131540433	131555165	14732	459
PTGIR	Prostaglandin I2 (prostacyclin) Receptor	19	51815565	51820194	4629	1161
PTGS1	Prostaglandin-endoperoxide Synthase 1	9	124173050	124197802	24752	1800
PTGS2	Prostaglandin-endoperoxide Synthase 2	1	184907592	184916179	8587	1815
PTHR1	Parathyroid Hormone 1 Receptor	3	46894240	46920293	26054	1782
PYGB	Phosphorylase, Glycogen; Brain	20	25176706	25226648	49942	2532
RIPK2	Receptor-interacting Serine-threonine Kinase 2	8	90839110	90872433	33323	1623
RORA	RAR-related Orphan Receptor A	15	58576755	59308794	732039	2032
RORC	RAR-related Orphan Receptor C	1	150045171	150070972	25802	1564
RTN4	Reticulon 4	2	55052829	55131468	78640	3613
SCD	Stearoyl-CoA Desaturase	10	102096762	102114578	17816	1080
SCN9A	Sodium Channel, Voltage-gated, Type IX, Alpha Subunit	2	166759941	166940749	180809	5934
SDHB	Succinate Dehydrogenase Complex, Subunit B, Iron Sulfur	1	17217804	17253252	35448	843
SDHD	Succinate Dehydrogenase Complex, Subunit D, Integral Membrane Protein	11	111462832	111471727	8895	480
SIRT1	Sirtuin 1	10	69314433	69348149	33716	2244
SIRT2	Sirtuin 2	19	44061037	44082342	21306	1171
SIRT3	Sirtuin 3	11	205030	226362	21332	1200
SIRT4	Sirtuin 4	12	119224546	119235430	10885	945
SIRT5	Sirtuin 5	6	13682812	13720500	37688	976
SIRT6	Sirtuin 6	19	4125106	4133596	8490	1068
SIRT7	Sirtuin 7	17	77463107	77469332	6225	1203

SLC10A1	Solute Carrier Family 10, Member 1	14	69312305	69333759	21455	1050
SLC10A2	Solute Carrier Family 10, Member 2	13	102494351	102517197	22846	1047
SLC5A1	Solute Carrier Family 5, Member 1	22	30769259	30836645	67386	1995
SLC6A4	Solute Carrier Family 6, Member 4	17	25545463	25586841	41379	1893
SLC6A9	Solute Carrier Family 6, Member 9	1	44234742	44269721	34979	2151
SP110	SP110 Nuclear Body Protein	2	230741896	230792932	51036	2202
STIM1	Stromal Interaction Molecule 1	11	3833509	4071015	237506	2058
STK39	Serine Threonine Kinase 39	2	168518776	168812365	293590	1638
SYK	Spleen Tyrosine Kinase	9	92603891	92700652	96762	1908
TACR1	Tachykinin Receptor 1	2	75129738	75280122	150385	1228
TACR2	Tachykinin Receptor 2	10	70833964	70846680	12716	1197
TACR3	Tachykinin Receptor 3	4	104730074	104860422	130348	1398
TBXA2R	Thromboxane A2 Receptor	19	3545504	3557658	12154	1160
TGFB1	Transforming Growth Factor, Beta 1	19	46528491	46551656	23165	1173
TGFBR1	Transforming Growth Factor, Beta Receptor 1	9	100907233	100956295	49062	1512
TLR4	Toll-like Receptor 4	9	119506431	119519589	13158	2520
TLR7	Toll-like Receptor 7	X	12795123	12818401	23278	3150
TLR9	Toll-like Receptor 9	3	52230138	52235219	5081	3099
TNFRSF1A	Tumor Necrosis Factor Receptor Superfamily, Member 1A	12	6308184	6321522	13338	1368
TNFSF11	Tumor Necrosis Factor (ligand) Superfamily, Member 11	13	42034872	42080148	45277	954
TNNI3K	TNNI3 Interacting Kinase	1	74436535	74782696	346162	2508
TRPC3	Transient Receptor Potential Cation Channel, Subfamily C, Member 3	4	123019633	123092285	72653	2547
TRPC6	Transient Receptor Potential Cation Channel, Subfamily C, Member 6	11	100827577	100959869	132293	2796
TRPM8	Transient Receptor Potential Cation Channel, Subfamily M, Member 8	2	234490782	234592905	102123	3315
TRPV1	Transient Receptor Potential Cation Channel, Subfamily V, Member 1	17	3415490	3459454	43964	2519
UTS2R	Urotensin 2 Receptor	17	77925490	77926659	1169	1170
ZAP70	Zeta-chain Associated Protein Kinase 70kDa	2	97696463	97722755	26292	1860

[a]Chromosome positions based on NCBI build 36.3.

Table S2A.

Differences in Gene Ontology terms between 202 study genes and the rest of the genome: molecular function.

	GENCODE[a]		Current Study				Odds
GO Term	Count	Percent	Count	Percent	Diff.	P-value	Ratio
G-protein coupled receptor activity	383	1.9	51	25.3	23.4	4.1E-113	17.4
ion channel activity	241	1.2	19	9.4	8.2	1.3E-23	8.6
receptor activity	1291	6.4	72	35.6	29.3	2.2E-60	8.1
cytokine activity	165	0.8	11	5.5	4.6	2.4E-11	7.0
protein heterodimerization activity	245	1.2	13	6.4	5.2	3.4E-10	5.6
receptor binding	254	1.3	11	5.5	4.2	9.3E-7	4.5
protein homodimerization activity	409	2.0	15	7.4	5.4	3.5E-7	3.9
protein binding	7127	35.0	131	64.9	29.8	6.5E-18	3.4
protein serine/threonine kinase activity	515	2.5	16	7.9	5.4	5.8E-6	3.3
peptidase activity	483	2.4	14	6.9	4.6	8.8E-5	3.0
protein kinase activity	536	2.6	15	7.4	4.8	8.4E-5	2.9
nucleic acid binding	1159	5.7	1	0.5	-5.2	2.3E-3	0.08

[a]Gene ontology description for all protein coding genes annotated by the GENCODE project.

Table S2B.

Differences in Gene Ontology terms between 202 study genes and the rest of the genome: cellular component.

	GENCODE		Current Study				Odds
GO Term	Count	Percent	Count	Percent	Diff.	P-value	Ratio
external side of plasma membrane	144	0.7	17	8.4	7.7	1.3E-32	12.8
membrane raft	121	0.6	14	6.9	6.3	3.4E-26	12.3
integral to plasma membrane	997	4.9	74	36.6	31.7	3.8E-88	11.1
postsynaptic membrane	155	0.8	15	7.4	6.7	2.5E-23	10.3
dendrite	168	0.8	15	7.4	6.6	2.2E-21	9.5
cell surface	269	1.3	22	10.9	9.6	1.5E-28	9.0
neuronal cell body	190	0.9	15	7.4	6.5	1.1E-18	8.4
plasma membrane	2939	14.5	113	55.9	41.5	1.5E-59	7.4
membrane fraction	522	2.6	26	12.9	10.3	2.0E-18	5.6
synapse	271	1.3	12	5.9	4.6	1.6E-7	4.6
extracellular space	748	3.7	26	12.9	9.2	4.6E-11	3.8
cell junction	403	2.0	14	6.9	5.0	3.0E-6	3.6
integral to membrane	4471	22.0	100	49.5	27.5	4.9E-20	3.4
endoplasmic reticulum	933	4.6	22	10.9	6.3	5.9E-5	2.5
endoplasmic reticulum membrane	553	2.7	13	6.4	3.7	3.1E-3	2.4
extracellular region	1867	9.2	34	16.8	7.7	3.8E-4	2.0

Table S2C.

Differences in Gene Ontology terms between 202 study genes and the rest of the genome: biological process.

GO Term	GENCODE Count	GENCODE Percent	Current Study Count	Current Study Percent	Diff.	P-value	Odds Ratio
positive regulation of peptidyl-tyrosine phosphorylation	47	0.2	11	5.5	5.2	1.5E-39	24.6
elevation of cytosolic calcium ion concentration	99	0.5	19	9.4	8.9	1.5E-58	21.0
cellular calcium ion homeostasis	69	0.3	13	6.4	6.1	7.2E-39	20.0
chemotaxis	130	0.6	18	8.9	8.3	1.4E-40	15.1
inflammatory response	245	1.2	31	15.4	14.1	1.5E-64	14.7
response to ethanol	80	0.4	11	5.5	5.1	2.8E-24	14.4
calcium ion transport	124	0.6	15	7.4	6.8	2.1E-29	12.9
cell surface receptor linked signaling pathway	233	1.2	26	12.9	11.7	2.2E-47	12.6
response to lipopolysaccharide	121	0.6	12	5.9	5.4	4.2E-19	10.4
synaptic transmission	178	0.9	17	8.4	7.5	4.2E-26	10.3
immune response	347	1.7	29	14.4	12.7	1.2E-38	9.6
response to hypoxia	166	0.8	14	6.9	6.1	9.3E-19	9.0
G-protein coupled receptor protein signaling pathway	905	4.5	56	27.7	23.3	5.9E-53	8.2
positive regulation of apoptosis	159	0.8	12	5.9	5.2	3.2E-14	7.9
response to drug	281	1.4	18	8.9	7.5	1.4E-17	6.9
positive regulation of cell proliferation	351	1.7	20	9.9	8.2	6.4E-17	6.2
cell-cell signaling	250	1.2	14	6.9	5.7	1.1E-11	5.9
protein amino acid phosphorylation	576	2.8	24	11.9	9.1	2.3E-13	4.6
negative regulation of cell proliferation	324	1.6	14	6.9	5.3	2.0E-8	4.6
ion transport	532	2.6	21	10.4	7.8	6.8E-11	4.3
positive regulation of transcription from RNA polymerase II promoter	375	1.8	14	6.9	5.1	6.5E-7	3.9
signal transduction	1309	6.4	41	20.3	13.9	1.4E-14	3.7
cell proliferation	326	1.6	11	5.5	3.8	7.4E-5	3.5
apoptosis	518	2.6	16	7.9	5.4	6.5E-6	3.3
cell adhesion	553	2.7	15	7.4	4.7	1.4E-4	2.8
transport	752	3.7	20	9.9	6.2	1.2E-5	2.8
proteolysis	477	2.4	11	5.5	3.1	9.0E-3	2.4

Table S3.

Genotype data quality assessments.

Validation Experiment	Measure	All variants	Singletons
130 sample duplicates	Heterozygote discordance	0.92%	1.5%
	Heterozygote error rate	0.50%	-
Capillary sequence, 245 singletons	False discovery rate	-	2.0%
1000 Genomes high coverage trios	Heterozygote discordance	0.95%	0.0%
30 parent offspring trios	Mendelian error rate	0.06%	4.8%

Table S4.
Overview of sequenced sample collections.

					Passed Quality Control	
Collection	**Ethnicity**	**Country**[a]	**Plated**[b]	**Sequenced**	**Count**	**Percent**
CoLaus	European	Switzerland	2086	2064	2059	99%
LOLIPOP	European	United Kingdom	549	541	541	99%
	Indian Asian	United Kingdom	499	497	497	100%
	Other	United Kingdom	285	284	284	100%
Metabolic Syndrome (GEMS), Trio[c]	European	Canada	35	35	35	100%
	European	Finland	45	45	45	100%
Metabolic Syndrome (GEMS), Case	European	Australia	188	188	186	99%
	European	Canada	283	281	280	99%
	European	Finland	75	75	75	100%
	European	Switzerland	158	158	157	99%
	European	United States	84	84	84	100%
Metabolic Syndrome (GEMS), Control	European	Australia	192	191	190	99%
	European	Canada	253	250	250	99%
	European	Finland	80	80	80	100%
	European	Switzerland	177	177	176	99%
	European	United States	90	90	85	94%
Coronary Artery Disease (MedStar)	European	United States	609	608	604	99%
Osteoarthritis (GOGO)	European	United Kingdom	300	298	298	99%
	European	United States	536	534	534	100%
Irritable Bowel Syndrome	European	Canada	165	165	165	100%
	European	United States	152	152	152	100%
Rheumatoid Arthritis	European	United Kingdom	615	611	611	99%
Multiple Sclerosis (geneMSA)	European	Netherlands	158	158	158	100%
	European	Switzerland	176	176	175	99%
	European	United States	339	339	337	99%
Multiple Sclerosis, Case	African American	United States	340	339	339	100%
Multiple Sclerosis, Control	African American	United States	260	254	252	97%
Epilepsy (GenEpa)	European	Switzerland	125	125	111	89%
Epilepsy (HitDIP)	European	Finland	185	183	164	89%
Alzheimer's Disease	European	Canada	705	700	687	97%

(genADA)						
Unipolar depression	European	Germany	775	758	741	96%
Bipolar disorder	European	Canada	376	376	374	99%
	European	United Kingdom	81	81	80	99%
	European	England	329	329	323	98%
Schizophrenia	European	Canada	254	254	254	100%
	European	Germany	336	336	330	98%
	European	United Kingdom	221	221	219	99%
	European	United Kingdom	298	298	296	99%
Chronic Obstructive Pulmonary Disease (HitDIP)	European	Norway	782	781	780	100%
Chronic Obstructive Pulmonary Disease (ECLIPSE)	European	Bulgaria	52	52	52	100%
	European	Canada	96	96	95	99%
	European	Czech Republic	27	27	27	100%
	European	Denmark	44	44	43	98%
	European	Netherlands	74	72	71	96%
	European	Norway	150	150	148	99%
	European	Slovenia	74	72	72	97%
	European	Spain	32	32	32	100%
	European	United Kingdom	187	186	185	99%
	European	United States	266	264	263	99%
1000 Genomes Project	European	Nigeria	3	3	3	100%
	European	United States	3	3	3	100%
Total			14204	14117	14002	99%

[a]Country where subjects were recruited into their respective study.
[b]Count of subject DNA samples that were plated for sequencing.
[c]A total of 30 trios were sequenced, however some trio members are included as cases or controls and hence included in the counts above.

Table S5A.

Single nucleotide variants observed in 12,514 European subjects by frequency and class.

	Nonsense Readthrough	NS	S	Splice	UTR	Intron	Flank	Total
Singleton	200	5978	3316	181	8323	4870	376	23244
Doubleton	31	1293	874	30	1927	1097	87	5339
(0.0001,0.001][a]	24	1463	1089	59	2623	1538	93	6889
(0.001,0.005]	3	196	177	14	478	304	16	1188
(0.005,0.02]	0	101	92	4	215	131	14	557
(0.02,0.05]	0	36	46	1	134	60	3	280
(0.05,0.5]	2	105	209	10	414	280	18	1038
Total	260	9172	5803	299	14114	8280	607	38535
Unobserved[b]	34	1823	1422	65	3538	2106	152	9140

[a]Excludes doubletons that may have MAF up to 0.00016 if 50% of genotypes are missing.
[b]SNVs observed in the overall study but not in 12,514 Europeans.

Table S5B.

Single nucleotide variants observed in 594 African American subjects by frequency and class.

	Nonsense Readthrough	NS	S	Splice	UTR	Intron	Flank	Total
Singleton	17	905	683	30	1649	964	69	4317
Doubleton	1	173	148	9	420	237	24	1012
(0.002,0.005][a]	1	173	211	10	464	258	20	1137
(0.005,0.02]	1	198	249	11	648	344	25	1476
(0.02,0.05]	2	88	139	5	323	171	21	749
(0.05,0.5]	2	129	260	10	555	414	20	1390
Total	24	1666	1690	75	4059	2388	179	10081
Unobserved[b]	270	9329	5535	289	13593	7998	580	37594

[a]Excludes doubletons that may have MAF up to 0.0034 if 50% of genotypes are missing.
[b]SNVs observed in the overall study but not in 594 African Americans.

Table S5C.

Single nucleotide variants observed in 567 South Asian subjects by frequency and class.

	Nonsense Readthrough	NS	S	Splice	UTR	Intron	Flank	Total
Singleton	16	806	573	25	1392	844	53	3709
Doubleton	1	163	128	2	322	186	14	816
(0.002,0.005][a]	1	115	122	4	260	161	13	676
(0.005,0.02]	2	123	120	7	313	172	12	749
(0.02,0.05]	1	41	77	4	149	50	6	328
(0.05,0.5]	1	93	202	8	434	303	18	1059
Total	22	1341	1222	50	2870	1716	116	7337
Unobserved[b]	272	9654	6003	314	14782	8670	643	40338

[a]Excludes doubletons that may have MAF up to 0.0035 if 50% of genotypes are missing.
[b]SNVs observed in the overall study but not in 567 South Asians.

Table S6.

Overlap of variants from Online Mendelian Inheritance in Man (OMIM) with those observed in the current study.

					European		Southern Asian		African American	
Variant Number	**Evidence**	***GENE* Variant**	**Mode of Inherit**	**Disease**	**MAF**	**MAC**	**MAF**	**MAC**	**MAF**	**MAC**
		APP								
chr21_26191825	Low	GLU665ASP	Dominant	ALZHEIMER DISEASE	8.0E-5	2	0.0	0	0.0	0
chr21_26185979	Low	ALA713THR	Dominant		8.0E-5	2	0.0	0	0.0	0
		CASR								
chr3_123455763	High	LEU13PRO	Recessive	HYPOCALCIURIC HYPERCALCEMIA	4.0E-5	1	0.0	0	0.0	0
chr3_123485908	Medium	PHE806SER	Dominant	HYPOPARATHYROID	4.0E-5	1	0.0	0	0.0	0
		CD3G								
chr11_117720349	Medium	MET1VAL	Recessive	IMMUNODEFICIENCY	4.0E-5	1	0.0	0	0.0	0
		DRD2								
chr11_112792867	Low	VAL154ILE	Dominant	MYOCLONUS-DYSTONIA SYNDROME	4.0E-5	1	0.0	0	0.0	0
		EDNRB								
chr13_77390541	High	GLY57SER	Recessive	HIRSCHSPRUNG DISEASE - SUS.	7.8E-3	195	8.9E-4	1	8.4E-4	1
chr13_77373231	Medium	SER305ASN	Recessive		1.3E-2	321	0.0	0	5.9E-3	7
		GHSR								
chr3_173648189	High	ARG237TRP	Recessive	SHORT STATURE - IDIOPATHIC	2.0E-4	5	1.8E-3	2	0.0	0
		LRRK2								
chr12_38990503	High	ARG1441CYS	Dominant		4.0E-5	1	0.0	0	8.4E-4	1
chr12_39020469	High	GLY2019SER	Dominant	PARKINSON'S DISEASE	3.6E-4	9	0.0	0	0.0	0
chr12_39043595	High	GLY2385ARG	Dominant		4.0E-5	1	8.9E-4	1	0.0	0
		MMP9								
chr20_44070974	Medium	MET1LYS	Recessive	METAPHYSEAL ANADYSPLASIA	8.0E-5	2	6.2E-3	7	0.0	0
		PLA2G7								
chr6_46785057	High	VAL279PHE	Recessive	PLATELET-ACTIVATING FACTOR ACETYLHYDROLASE DEFICIENCY	3.4E-4	8	1.8E-3	2	0.0	0
		PSEN1								
chr14_72707406	High	ALA79VAL	Dominant		4.0E-5	1	0.0	0	0.0	0
chr14_72723321	High	HIS163ARG	Dominant	ALZHEIMER DISEASE	1.2E-4	3	0.0	0	0.0	0
chr14_72723321	High	HIS163TYR	Dominant		1.2E-4	3	0.0	0	0.0	0
chr14_72748272	High	ASP333GLY	Dominant	CARDIOMYOPATHY - DILATED	0.0	0	0.0	0	2.5E-3	3
		PSEN2								
chr1_225138141	Medium	ALA85VAL	Dominant	ALZHEIMER DISEASE	4.0E-5	1	0.0	0	0.0	0
chr1_225139894	High	SER130LEU	Dominant	CARDIOMYOPATHY - DILATED	1.4E-3	34	0.0	0	0.0	0
chr1_225149872	Medium	ASP439ALA	Dominant	ALZHEIMER DISEASE	1.2E-4	3	0.0	0	0.0	0
		SCN9A								
chr2_166876329	Low	ILE62VAL	Dominant	FEBRILE CONVULSIONS	8.1E-5	2	0.0	0	0.0	0
chr2_166849262	High	ASN641TYR	Dominant	GENERALIZED EPILEPSY W/ FEBRILE SEIZURES	4.0E-5	1	0.0	0	0.0	0
chr2_166846542	High	LYS655ARG	Dominant		2.4E-3	59	0.0	0	2.5E-3	3
chr2_166842007	High	LEU858HIS	Dominant	ERYTHERMALGIA	4.0E-5	1	0.0	0	0.0	0
chr2_166837487	High	ARG996 CYS	Codom	PAROXYSMAL EXTREME PAIN DISORDER	4.2E-5	1	0.0	0	0.0	0
		SDHB								
chr1_17253094	High	ALA3GLY	Dominant	COWDEN-LIKE SYNDROME	1.7E-4	4	0.0	0	4.2E-2	49
chr1_17227710	Medium	HIS132PRO	Dominant	PARAGANGLIOMAS	4.0E-5	1	0.0	0	0.0	0

chr1_17226884	High	SER163PRO	Dominant	COWDEN-LIKE SYNDROME	1.5E-2	364	1.5E-2	17	3.4E-3	4
chr1_17221730	High	ARG242HIS	Dominant	PARAGANGLIOMAS 4 PHEOCHROMOCYTOMA	0.0	0	0.0	0	8.4E-4	1
		SDHD								
chr11_111463887	High	HIS50ARG	Dominant	CARCINOID TUMORS	8.6E-3	215	2.7E-3	3	0.0	0
chr11_111464873	High	PRO81LEU	Dominant	PARAGANGLIOMAS 1 PHEOCHROMOCYTOMA	4.0E-5	1	0.0	0	0.0	0
		SLC10A2								
chr13_102499774_A	Medium	THR262 MET	Recessive	BILE ACID MALABSORPTION - PRIMARY	2.4E-4	6	8.8E-4	1	8.4E-4	1
		SLC6A4								
chr17_25562500	High	ILE425VAL		OBSESSIVE-COMPULSIVE DISORDER - SUS.	8.8E-4	22	0.0	0	1.7E-3	2
		TACR3								
chr4_104860004	Medium	GLY93ASP	Recessive	HYPOGONADOTROPIC HYPOGONADISM	0.0	0	0.0	0	0.0	0

[a]MAF: minor allele frequency.
[b]MAC: minor allele count.

Table S7.

Sample sizes and summary statistics for case-control analyses.

Study	Case: Control Ratio	Cases	Controls	Genetic Distance[a]	Inflation λ[b] Common	Amino Acid	Functional
Alzheimer's	1:10	649	6490	0.018	1.09	1.16	1.08
Bipolar Disorder	1:6	778	4667	0.018	1.01	1.32	1.08
COPD	1:6	947	5682	0.017	1.18	1.41	1.41
Coronary Artery Disease	1:8	604	4832	0.018	1.19	1.28	1.03
Dyslipidemia	1:1	769	769	N/A	1.07	1.50	1.20
Epilepsy	1:50	120	6000	0.017	1.09	1.31	1.43
Irritable Bowel Syndrome	1:12	314	3768	0.014	1.26	1.40	1.39
Multiple Sclerosis	1:10	642	6420	0.019	1.12	1.01	0.99
Osteoarthritis	1:6	798	4788	0.019	2.18	1.51	1.62
Rheumatoid Arthritis	1:6	608	3648	0.019	1.31	1.83	1.32
Schizophrenia	1:4	1066	4264	0.018	1.12	1.63	1.28
Unipolar Depression	1:6	718	4308	0.020	1.35	1.32	1.54

[a]Median Euclidean genetic distance between cases and controls

[b]Genomic control λ for common variant and aggregate rare variant tests, including all amino acid-changing variants and just those predicted to be functional by PolyPhen or SIFT, or occurring at evolutionarily conserved bases.

Table S8.

Overlap of genes and disease traits with NHGRI GWAS Catalog.

Study Trait	Gene	NHGRI Reported Trait	NHGRI Reported Gene	Reference
Bipolar Disorder	*CNTN5*	Bipolar disorder and schizophrenia	*CNTN5*	(*77*)
Schizophrenia	*CNTN5*	Bipolar disorder and schizophrenia	*CNTN5*	(*77*)
COPD	*HHIP*	Chronic obstructive pulmonary disease	*HHIP*	(*49, 78*)
Coronary Artery Disease	*OPRM1*	Coronary heart disease	*OPRM1*	(*79*)
Unipolar Depression	*RORA*	Depression--quantitative trait	*RORA*	(*80*)
Unipolar Depression	*ITGB1*	Depression--quantitative trait	*ITGB1*	(*80*)
Multiple Sclerosis	*TNFRSF1A*	Multiple sclerosis	*TNFRSF1A*	(*81*)
Multiple Sclerosis	*IL6*[a]	N/A	N/A	(*76*)
Multiple Sclerosis	*IL7R*	Multiple sclerosis	*IL7R*	(*76, 81, 82*)
Multiple Sclerosis	*EVI5*	Multiple sclerosis	*EVI5, RPL5*	(*82, 83*)
Multiple Sclerosis	*CLEC16A*	Multiple sclerosis	*CLEC16A*	(*81, 82*)
Schizophrenia	*PTGS2*	Schizophrenia	Intergenic	(*84*)
Coronary Artery Disease	*CHRNA5*	Sudden cardiac arrest	*CHRNB4*	(*85*)
Coronary Artery Disease	*CHRNA3*	Sudden cardiac arrest	*CHRNB4*	(*85*)

[a]The association between *IL6* and multiple sclerosis reported in reference (*76*) was not included in the NHGRI GWAS catalog due to stringent significance criteria. As this was one of the top-associated genes in that report ($p = 5.9\times10^{-8}$; more significant than *IL7R* and several other associations reported in this table) and it overlaps with the current study, we include it here.

Table S9.

Statistically significant rare variant associations resulting from GWAS candidate gene analysis.

Study	GENE	Test	MAF[a]	Carrier/Noncarrier Count		P-value[b]		Odds Ratio	(95% CI)
				Controls	Cases	Unadjusted	Adjusted		
Multiple Sclerosis	*IL6*	Functionally Damaging	0.0006	4/6393	4/635	0.0014	0.0071	12.3	(3.1,49.8)
Multiple Sclerosis	*IL6*	Amino Acid Changing	0.0015	15/6382	6/633	0.0043	0.0215	4.9	(1.9,12.9)
Multiple Sclerosis	*TNFRSF1A*	Amino Acid Changing	0.0023	25/6372	8/631	0.0056	0.0277	3.6	(1.6,8.2)
Unipolar Depression	*ITGB1*	Amino Acid Changing	0.0044	42/4251	2/716	0.0240	0.0481	0.3	(0.1,1.1)

[a]Cumulative minor allele frequency in cases and controls.

[b]P-value adjusted for the number of candidate genes for each disease (see Supplementary Table X).

Table S10.
Sample sizes for population genetic and geographic analyses.

	# individuals sampled	# chromosomes in frequency spectra[a]
African-American	594	1,168
Southern Asia	567	1,068
Europe	12,514	22,000
North-Western Europe	2,489	4,546
North European	963	1,838
Finland	261	474
Western Europe	1,625	3,006
Central Europe	946	1,734
Eastern Europe	263	482
South-Western Europe	289	532
South-Eastern Europe	370	688

[a]All samples were down-sampled to retain 80% of all targeted sites, except for the European continental sample, where the number was rounded down to an even number (84.6% of all sites retained).

Table S11.

Single-nucleotide variant transition:transversion ratios.

	All Variants	Singletons	Doubletons	MAF > 0.1% and Missing < 10%
Nonsynonymous	2.10	1.90	2.57	2.25
Synonymous	4.79	4.25	5.23	5.20
UTR	2.00	1.82	2.29	2.45
Intron	2.04	1.97	2.22	2.15

Table S12.

Discordant genotypes and rates observed in 130 sample duplicates.

Sample 1	Sample 2 Genotype[a]			Discordance Rate	
Genotype	0	1	2	Overall	Heterozygote
All variant positions (N = 44,230)					
0	5,169,859	313	0	7.05E-5	9.23E-3
1		39,625	56		
2			24,418		
Variants in dbSNP (N = 2,641)					
0	257,216	146	0	6.22E-4	5.38E-3
1		36,573	52		
2			24,308		
Variants not in dbSNP (N = 41,589)					
0	4,912,643	167	0	3.48E-5	5.31E-2
1		3,051	4		
2			110		
Singleton variants (N = 25477)					
0	2,949,160	3	0	1.02E-6	1.45E-2
1		204	0		
2			0		

[a]Sample genotypes are categorized as 0, 1 and 2 corresponding to reference homozygote, heterozygote or non-reference homozygote, respectively. As duplicate sample order is arbitrary, all discordant genotype counts are presented in the upper triangle.

Table S13.

Genotype error rate estimates based on duplicate discordance assuming a single-allele error model.

Observed	True Genotype[a]		
Genotype	0	1	2
All variant positions (N = 44,230)			
0	1.00E+00	3.94E-3	0.00E+00
1	2.36E-11	9.95E-1	8.34E-5
2	0.00E+00	6.48E-4	1.00E+00
Total Error Rate	2.36E-11	4.59E-3	8.34E-5
Variants in dbSNP (N = 2,641)			
0	1.00E+00	1.36E-3	0.00E+00
1	9.09E-5	9.98E-1	7.48E-5
2	0.00E+00	6.56E-4	1.00E+00
Total Error Rate	9.09E-5	2.02E-3	7.48E-5
Variants not in dbSNP (N = 41,589)			
0	1.00E+00	2.67E-2	0.00E+00
1	4.65E-21	9.73E-1	1.73E-6
2	0.00E+00	6.27E-4	1.00E+00
Total Error Rate	4.65E-21	2.73E-2	1.73E-6

[a]Sample genotypes are categorized as 0, 1 and 2 corresponding to reference homozygote, heterozygote or non-reference homozygote, respectively.

Table S14.

Discordant genotypes and rates observed comparing genotypes from current study to two 1000 Genomes Project deep sequenced trios.

	1000 Genomes Genotype[a]			Discordance Rate	
Genotype	0	1	2	Overall	Heterozygote
CEU and YRI combined					
0	1900	12	0	4.21E-3	9.52E-3
1	3	1769	2		
2	0	0	349		
CEU trio					
0	932	4	0	3.18E-3	7.44E-3
1	2	800	0		
2	0	0	146		
YRI trio					
0	968	8	0	5.11E-3	1.12E-2
1	1	969	2		
2	0	0	203		

[a]Sample genotypes are categorized as 0, 1 and 2 corresponding to reference homozygote, heterozygote or non-reference homozygote, respectively.

Table S15.

Influence of coding length, GC content and average phyloP on measures of nonsynonymous gene diversity and mutation among genes.

Response	Common NS		Rare NS		cMAF		Mutation Rate	
	β	p[b]	β	p	β	p	β	p
(intercept)	0.1918		3.557		4.81E-4		1.29E-8	
Coding Length	0.000062	3.2E-8	0.0246	<1e-15	3.26E-6	<1e-15	6.27E-13	0.11
r^2	0.15		0.71		0.53		0.013	
Coding Length Adjusted[a]								
(intercept)	0.00241		0.0211		2.06E-6		4.88E-9	
GC Content	-0.00021	0.65	0.0334	2.2E-5	-2.82E-7	0.63	1.82E-8	1.9E-3
phyloP	-0.00122	9.0E-7	-0.0095	9.8E-8	8.32E-7	0.01	-3.40E-10	0.75
r^2	0.12		0.2		0.03		0.05	

[a]Mutation rate was not adjusted for coding length, all other response variables were divided by length of successfully sequenced coding regions

[b]P values computed by likelihood ratio F test (complete versus reduced model)

Table S16.

Annotation nonsynonymous SNVs by PolyPhen and SIFT as a function of allele frequency in 12,514 Europeans.

	PolyPhen			SIFT[a]		
	Benign	Possibly Damaging	Probably Damaging	Tolerated	Damaging (Low Conf)	Damaging
Singleton	3532	1248	1089	2937	763	2371
Doubleton	807	253	212	680	160	469
(0,0.001]	857	311	258	779	202	510
(0.001,0.005]	127	39	21	121	18	56
(0.005,0.05]	96	23	13	86	12	41
(0.05,0.5]	64	21	5	85	10	11

[a]Predictions are based on SIFT score as tolerated (score > 0.05), damaging with low confidence warning (score ≤ 0.05), and damaging (score ≤ 0.05).

Table S17.

Correspondence between PolyPhen and SIFT predictions of nonsynonymous SNVs in Europeans.

		PolyPhen	
SIFT	Benign	Possibly Damaging	Probably Damaging
Tolerated	4593	642	218
Damaging (Low Conf)	457	391	291
Damaging	1483	1199	1345

Additional Data table S1 (separate file)

Target regions of sequenced genes. The column names and a brief description, where needed, are given below.

Gene – RefSeq build 36 gene name
Gene37 – RefSeq build 37 gene name
Chromosome
Exon.NCBI.36.Start – Exon start position
Exon.NCBI.36.Stop – Exon stop position
Exon.plus.50.bp.flanking.sequence.NCBI.36.Start – Target start position
Exon.plus.50.bp.flanking.sequence.NCBI.36.Stop – Target stop position
Entrez.Gene.ID
Transcript – Entrez transcript ID
Ensembl.Gene.ID
Ensembl.Transcript.ID
code – Number of coding bases in target region
utr– Number of UTR bases in target region
intron – Number of intronic bases in target region
upstream – Number of upstream bases in target region
downstream – Number of downstream bases in target region
code.cover – Corresponding bases with at least 50% genotypes called
utr.cover
intron.cover
upstream.cover
downstream.cover
TargetLength – Total target length
CoverLength – Total target length with at least 50% of genotypes called

Additional Data table S2 (separate file)

Variants and their annotations. The column names and a brief description, where needed, are given below.

VARIANT_ID – Variant ID: NCBI build 36 chromosome, position and minor allele (if multi-allelic)
RSID – RefSNP ID
CHROMOSOME
POSITION
GENE – RefSeq build 36 gene name
REF_ALLELE – Reference allele
REF_ALLELE_COUNT – Count of observed alleles
VARIANT – Non-reference allele
VARIANT_COUNT – Count of observed alleles
MISSING – Fraction of missing genotypes
FEATURE – Variant feature
UP_ID – UniProt ID
UNIPROT_POSITION
AA1 – Reference amino acid
AA2 – Non-reference amino acid
MINOR_ALLELE – Which allele has frequency less than 0.5 in full resequenced sample
MINOR_ALLELE_COUNT – Minor allele count
MULTI_ALLELE – Which allele this is observed at this base position in descending frequency
HOM_N – Number of major homozygote genotypes observed
HOM_DEPTH_AVG – Major homozygote average depth
HOM_Q_AVG – Major homozygote average consensus quality
HET_N – Number of heterozygote genotypes observed
HET_DEPTH_AVG – Heterozygote average depth
HET_Q_AVG – Heterozygote average consensus quality
Eur.MA_COUNT – Minor allele count in Europeans
Eur.NOBS – Number of genotypes observed in Europeans
Eur.FREQ – European minor allele frequency (minor allele defined in full sample)
Europe.W.FREQ – Western European
Europe.W.NOBS
Europe.C.FREQ – Central European
Europe.C.NOBS
Europe.SW.FREQ – Southwestern European
Europe.SW.NOBS
Europe.S.FREQ – Southern European
Europe.S.NOBS
Europe.SE.FREQ – Southeastern European
Europe.SE.NOBS
Europe.E.FREQ – Eastern European
Europe.E.NOBS
Europe.NW.FREQ – Northwestern European

Europe.NW.NOBS
Europe.N.FREQ – Northern European
Europe.N.NOBS
Finnish.FREQ – Finnish
Finnish.NOBS
African_American.MA_COUNT
African_American.NOBS
African_American.FREQ
UN_Southern_Asia.MA_COUNT
UN_Southern_Asia.NOBS
UN_Southern_Asia.FREQ
POLYPHEN_PREDICTION – PolyPhen prediction for nonsynonymous variants
PSIC – PolyPhen position-specific independent counts
SIFT_PREDICTION – SIFT prediction
SIFT_SCORE – SIFT score
PHYLOP – phyloP score based on 46-way placental alignment
CHROMOSOME

Additional Data table S3 (separate file)

Folded site frequency spectra of all four-fold degenerate sites for the 188 autosomal genes used for the demographic inference. The spectra were calculated for a sample of 11,000 Europeans as detailed. One column corresponds to each gene and one row for each minor allele bin count.

Reference List

1. J. K. Pritchard, *Am J Hum Genet* **69**, 124 (2001).
2. G. V. Kryukov, L. A. Pennacchio, S. R. Sunyaev, *Am. J. Hum. Genet.* **80**, 727 (2007).
3. G. T. Marth *et al.*, *Genome Biol.* **12**, R84 (2011).
4. T. A. Manolio *et al.*, *Nature* **461**, 747 (2009).
5. E. E. Eichler *et al.*, *Nat. Rev. Genet.* **11**, 446 (2010).
6. J. Asimit, E. Zeggini, *Annu. Rev. Genet.* **44**, 293 (2010).
7. S. Gravel *et al.*, *Proc. Natl. Acad. Sci. U. S. A* **108**, 11983 (2011).
8. A. Coventry *et al.*, *Nat. Commun.* **1**, 131 (2010).
9. T. Ohta, *Nature* **246**, 96 (1973).
10. S. H. Williamson *et al.*, *Proc. Natl. Acad. Sci. U. S. A* **102**, 7882 (2005).
11. H. J. Muller, *Am. J. Hum. Genet.* **2**, 111 (1950).
12. A. P. Russ, S. Lampel, *Drug Discov. Today* **10**, 1607 (2005).
13. See supporting material at Science online. (2012).
14. M. A. Jobling, M. Hurles, C. Tyler-Smith, *Human Evolutionary Genetics: Origins, Peoples and Disease* (Garland Science, 2003).
15. M. Livi-Bacci, *A concise history of world population* (Wiley-Blackwell, ed. 2, 2007), pp. 1-250.
16. J. Wakeley, T. Takahashi, *Mol. Biol. Evol.* **20**, 208 (2003).
17. P. Awadalla *et al.*, *Am. J. Hum. Genet.* **87**, 316 (2010).
18. D. F. Conrad *et al.*, *Nat. Genet.* **43**, 712 (2011).
19. P. W. Messer, *Genetics* **182**, 1219 (2009).
20. A. L. Price *et al.*, *Am. J. Hum. Genet.* **86**, 832 (2010).
21. I. K. Kotowski *et al.*, *Am. J. Hum. Genet.* **78**, 410 (2006).
22. L. A. Hindorff *et al.*, *Proc. Natl. Acad. Sci. U. S. A* **106**, 9362 (2009).
23. E. Salmela *et al.*, *PLoS. One.* **3**, e3519 (2008).
24. C. D. Bustamante, E. G. Burchard, F. M. De la Vega, *Nature* **475**, 163 (2011).
25. O. Lao *et al.*, *Curr. Biol.* **18**, 1241 (2008).
26. E. S. Lander, N. J. Schork, *Science* **265**, 2037 (1994).
27. J. M. Akey *et al.*, *PLoS. Biol.* **2**, e286 (2004).
28. SeattleSNPs. http://pga.gs.washington.edu. (2012).
29. N. Ahituv *et al.*, *Am. J. Hum. Genet.* **80**, 779 (2007).
30. R. M. Durbin *et al.*, *Nature* **467**, 1061 (2010).
31. M. Firmann *et al.*, *BMC. Cardiovasc. Disord.* **8**, 6 (2008).
32. M. Preisig *et al.*, *BMC. Psychiatry* **9**, 9 (2009).
33. J. S. Kooner *et al.*, *Nat. Genet.* **40**, 149 (2008).
34. H. Ling *et al.*, *Obesity. (Silver. Spring)* **17**, 737 (2009).
35. D. F. Wyszynski *et al.*, *Am. J. Cardiol.* **95**, 194 (2005).
36. T. L. Assimes *et al.*, *J. Am. Coll. Cardiol.* **56**, 1552 (2010).
37. V. B. Kraus *et al.*, *Osteoarthritis. Cartilage.* **15**, 120 (2007).
38. C. Vignal *et al.*, *Arthritis Rheum.* **60**, 53 (2009).

39. S. E. Baranzini *et al.*, *Hum. Mol. Genet.* **18**, 767 (2009).
40. J. R. Oksenberg *et al.*, *Am. J. Hum. Genet.* **74**, 160 (2004).
41. B. A. Cree *et al.*, *Arch. Neurol.* **66**, 226 (2009).
42. N. Patterson *et al.*, *Am. J. Hum. Genet.* **74**, 979 (2004).
43. E. L. Heinzen *et al.*, *Am. J. Hum. Genet.* **86**, 707 (2010).
44. D. Kasperaviciute *et al.*, *Brain* **133**, 2136 (2010).
45. H. Li *et al.*, *Arch. Neurol.* **65**, 45 (2008).
46. P. Muglia *et al.*, *Mol. Psychiatry* **15**, 589 (2010).
47. C. Francks *et al.*, *Mol. Psychiatry* **15**, 319 (2010).
48. J. Vestbo *et al.*, *Eur. Respir. J.* **31**, 869 (2008).
49. S. G. Pillai *et al.*, *PLoS. Genet.* **5**, e1000421 (2009).
50. T. T. Ashburn, K. B. Thor, *Nat. Rev. Drug Discov.* **3**, 673 (2004).
51. Y. A. Lussier, J. L. Chen, *Sci. Transl. Med.* **3**, 96ps35 (2011).
52. J. Harrow *et al.*, *Genome Biol.* **7 Suppl 1**, S4 (2006).
53. M. Ashburner *et al.*, *Nat. Genet.* **25**, 25 (2000).
54. R. Li, Y. Li, K. Kristiansen, J. Wang, *Bioinformatics.* **24**, 713 (2008).
55. R. Li *et al.*, *Genome Res.* **19**, 1124 (2009).
56. I. M. Heid *et al.*, *Am. J. Epidemiol.* **168**, 878 (2008).
57. I. W. Saunders, J. Brohede, G. N. Hannan, *Genomics* **90**, 291 (2007).
58. R. Ihaka, R. Gentleman, *Journal of Computational and Graphical Statistics* **5**, 299--314 (1996).
59. V. Ramensky, P. Bork, S. Sunyaev, *Nucleic Acids Res.* **30**, 3894 (2002).
60. P. C. Ng, S. Henikoff, *Genome Res.* **11**, 863 (2001).
61. A. Stabenau *et al.*, *Genome Res.* **14**, 929 (2004).
62. A. Siepel *et al.*, *Genome Res.* **15**, 1034 (2005).
63. R. Blekhman *et al.*, *Curr. Biol.* **18**, 883 (2008).
64. D. G. MacArthur, C. Tyler-Smith, *Hum. Mol. Genet.* **19**, R125 (2010).
65. Z. A. Szpiech, M. Jakobsson, N. A. Rosenberg, *Bioinformatics.* **24**, 2498 (2008).
66. G. A. Watterson, *Theor. Popul. Biol.* **7**, 256 (1975).
67. I. Ebersberger, D. Metzler, C. Schwarz, S. Paabo, *Am. J. Hum. Genet.* **70**, 1490 (2002).
68. M. W. Nachman, S. L. Crowell, *Genetics* **156**, 297 (2000).
69. S. F. Schaffner *et al.*, *Genome Res.* **15**, 1576 (2005).
70. L. Excoffier, M. Foll, *Bioinformatics.* **27**, 1332 (2011).
71. A. J. Coffey *et al.*, *Eur. J. Hum. Genet.* **19**, 827 (2011).
72. M. R. Nelson *et al.*, *Am. J. Hum. Genet.* (2008).
73. S. A. Bacanu, J. C. Whittaker, M. R. Nelson, *Pharmacogenomics. J.* **12**, 93 (2012).
74. S. Hunter *et al.*, *Nucleic Acids Res.* **37**, D211 (2009).
75. J. A. Blake, C. J. Bult, J. A. Kadin, J. E. Richardson, J. T. Eppig, *Nucleic Acids Res.* **39**, D842 (2011).
76. J. H. Wang *et al.*, *Genome Med.* **3**, 3 (2011).
77. K. S. Wang, X. F. Liu, N. Aragam, *Schizophr. Res.* **124**, 192 (2010).
78. M. H. Cho *et al.*, *Nat. Genet.* **42**, 200 (2010).
79. G. Lettre *et al.*, *PLoS. Genet.* **7**, e1001300 (2011).
80. A. Terracciano *et al.*, *Biol. Psychiatry* **68**, 811 (2010).
81. P. L. De Jager *et al.*, *Nat. Genet.* **41**, 776 (2009).
82. D. A. Hafler *et al.*, *N. Engl. J. Med.* **357**, 851 (2007).

83. *Nat. Genet.* **41**, 824 (2009).
84. P. F. Sullivan *et al.*, *Mol. Psychiatry* **13**, 570 (2008).
85. B. E. Aouizerat *et al.*, *BMC. Cardiovasc. Disord.* **11**, 29 (2011).