Sciencexpress Reports

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

Matthew R. Nelson,1*† Daniel Wegmann,2* Margaret G. Ehm,1 Darren Kessner,2 Pamela St. Jean,1 Claudio Verzilli,1 Judong Shen,1 Zhengzheng Tang,3 Silviu-Alin Bacanu,1 Dana Fraser,1 Liling Warren,1 Jennifer Aponte,1 Matthew Zawistowski,6 Xiao Liu,4, Hao Zhang,4 Yong Zhang,4 Jun Li,5 Yun Li,3 Li Li,1 Peter Woollard,1 Simon Topp,1 Matthew D. Hall,1 Keith Nangle,1 Jun Wang,4,6 Gonçalo Abecasis,7 Lon R. Cardon,1 Sebastian Zöllner,7,8 John C. Whittaker,1 Stephanie L. Chissoe,1 John Novembre,2†‡ Vincent Mooser1‡

[1]Quantitative Sciences, GlaxoSmithKline, RTP, NC, USA; Upper Merion, PA, USA; and Stevenage, UK.

[2]Ecology and Evolutionary Biology, University of California–Los Angeles, Los Angeles, CA, USA.

[3]Genetics, Biostatistics, University of North Carolina–Chapel Hill, Chapel Hill, NC, USA.

[4]BGI, Shenzhen, China.

[5]Human Genetics, University of Michigan–Ann Arbor, Ann Arbor, MI, USA.

[6]Biology, The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen.

[7]Biostatistics, University of Michigan–Ann Arbor, Ann Arbor, MI, USA.

[8]Psychiatry, University of Michigan–Ann Arbor, Ann Arbor, MI, USA.

*These authors contributed equally to this work.

†To whom correspondence should be addressed: E-mail: matthew.r.nelson@gsk.com (M.R.N); jnovembre@ucla.edu (J.N.)

‡These authors contributed equally to this work.

Rare genetic variants contribute to complex disease risk; however, the abundance of rare variants in human populations remains unknown. We explored this spectrum of variation by sequencing 202 genes encoding drug targets in 14,002 individuals. We find rare variants are abundant (one every 17 bases) and geographically localized, such that even with large sample sizes, rare variant catalogs will be largely incomplete. We used the observed patterns of variation to estimate population growth parameters, the proportion of variants in a given frequency class that are putatively deleterious, and mutation rates for each gene. Overall we conclude that, due to rapid population growth and weak purifying selection, human populations harbor an abundance of rare variants, many of which are deleterious and have relevance to understanding disease risk.

Understanding the genetic contribution to human disease requires knowledge of the abundance and distribution of functional genetic diversity within and among populations. The "common-disease rare-variant" hypothesis posits that variants affecting health are under purifying selection, and thus should be found only at low frequencies in human populations (*1*–*3*). This hypothesis has become increasingly credible, since very large genome-wide association studies of common variants have explained only a fraction of the known heritability of most traits (*4*, *5*). Investigating the role of rare variants for complex trait mapping has led to tests that aggregate rare variants (*6*), and determine the abundance, distribution, and phenotypic effects of rare variants in human populations (*7*, *8*).

Population genetic models predict that mutation rates, the strength of selection, and demography affect the abundance of rare variants, although the relative importance of each is a long-standing question (*9*–*11*). To understand rare variant diversity in humans, we sequenced 202 genes in a sample of 14,002 well-phenotyped individuals (table S1). These genes represent approximately 1% of the coding genome and approximately 7% of genes considered current or potential drug targets (*12*), enriched for cell signaling proteins and membrane-bound transporters (table S2). A total of 864 kb were targeted, including 351 kb of coding and 323 kb of untranslated (UTR) exon regions (database S1). Over 93% of target bases were successfully sequenced at a median depth of 27 reads per site (*13*). Because rare variant discovery can easily be confounded with sequencing errors, we performed numerous experiments to demonstrate high data quality (table S3) (*13*). The sequenced subjects include two population samples (n = 1,322 and 2,059) and 12 disease collections (n = 125–1,125 cases, table S4). The self-reported ancestry of the sample was predominantly European (12,514), African American (594) and South Asian (567). Some of the following analyses focus on the European subset, which is well-powered to investigate rare variants. Based on our sample size we expect that 94% of variant alleles with minor allele frequency (MAF) 0.01% in Europeans were sampled at least once.

Sequencing revealed an abundance of rare (MAF<0.5%) single nucleotide variants (SNVs), compared to common variants (Fig. 1, A and B). We observed on average one variant per 17 bp in the overall sample and one variant per 21 bp in the Europeans (table S5). Among all variants, more than 95% were rare (MAF≤0.5%), and more than 74% were observed in only one or two subjects. ~90% of rare variants were not previously reported, as opposed to ~5% of common variants (MAF>0.5%) (fig. S1). For the large European subset, Watterson's θ_W, a metric of genetic diversity (Table 1), was much larger (40.38×10^{-4}) than in previous smaller scale studies, and an order of magnitude larger than the pairwise metric θ_π (3.96×10^{-4}). We observed a third allele at 2.0% of variable sites, and among those, 1.6% had a fourth allele. We found between 1.2 and 1.9 non-diallelic SNVs per kb of sequence (fig. S2), which tended to occur at sites under lower evolutionary conservation (fig. S3, (*13*)). The rate of variant discovery remained nearly constant with increasing sample size (Fig. 2A). We expect 111–153 variants per kb in a sample of 100,000 Europeans and 337–452 variants per kb in a sample of 1 million (Fig. 2, A and B).

These patterns are at odds with notions that human genetic diversity can be summarized by use of an effective population size (N_e) of 10,000 individuals (*14*). An N_e of 10,000 individuals is predictive of the average pairwise differences between human sequences (θ_π, Table 1), and is reflective of our emergence from a small population in Africa (*15*). However, the excess of rare variants observed here (i.e., $\theta_W >> \theta_\pi$) is a signature of the rapid growth and large population sizes that typify more recent human demographic history (*8*). When we fit a demographic

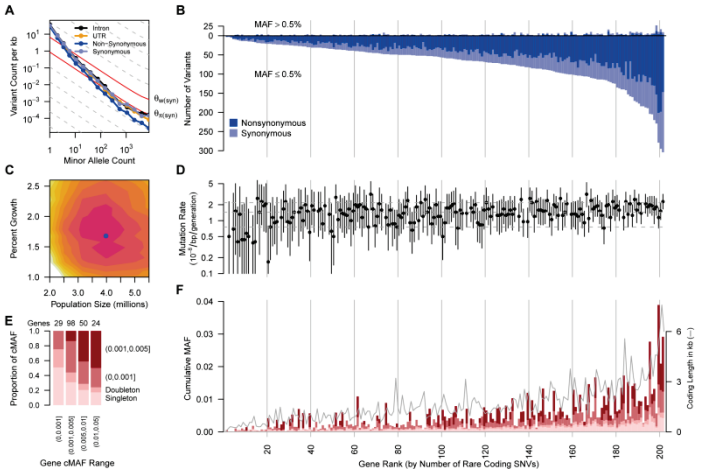

Fig. 1. (**A**) Frequency spectrum of variants relating the number of variants per kb within minor allele counts. Solid gray lines provide expectations from nucleotide diversity (θ_π) and the number of segregating sites (θ_W). (**B**) The number of common (MAF>0.5%, above the origin) and rare (MAF≤ 0.5%, below the origin) coding variants observed in each gene are shown as stacked bars of NS and S variants. (**C**) Log-likelihood surface of European population growth (*r*) and population size (N_e) in a demographic model. Colored contours correspond to 2 log-likelihood intervals. The blue point is the maximum likelihood estimate of *r* and N_e. (**D**) Per-gene mutation rates with 2 log-likelihood intervals. Horizontal lines are 10th, 50th and 90th mutation rate percentiles. Seven genes on the X chromosome and four genes with low target coverage or yielding too few common variants for inference (*ADRB3*, *CCR5*, *MIF* and *PTGER1*) were excluded. (**E**) Proportion of rare cumulative MAF (cMAF) accounted for by SNVs of increasing frequency. (**F**) Proportion of rare variants in four cMAF ranges falling within the MAF categories shown in (E). The successfully sequenced coding length of each gene (in kb) is overlaid as a gray line. cMAFs in (E) and (F) are for amino acid-changing variants in each gene predicted to be damaging or are evolutionarily conserved (phyloP≥2). Genes in (B), (D) and (F) are ordered by number of rare coding variants per gene and vertical lines correspond to rank deciles.

model to the four-fold degenerate synonymous (S) variants in Europeans, we obtained a maximum-likelihood estimate for a recent growth rate of 1.7**%** [95% confidence interval (CI) = 1.2%–2.3%], and a recent European effective population size of 4.0 million (95% CI = 2.5–5.0 million; Fig. 1C).

Taking advantage of the unprecedented size of this study for population genetics inference (*8*, *16*), we estimated mutation rates for each gene (Fig. 1D, (*13*)) and obtained a median estimate of 1.38×10^{-8} per bp per generation with 90% of estimates falling between 1.7×10^{-9} and 2.4×10^{-8}. Incorporating singleton discovery false negative rates from 2–8% resulted in median estimates no greater than 1.45×10^{-8}. These population-genetic-based rate estimates are similar to recent pedigree-based mutation rate estimates of 1.36×10^{-8} per bp per generation (*17*) and 1.17×10^{-8} per bp per generation (*13*, *18*). Further, these data reject a model of uniform mutation rates across genes ($P < 2 \times 10^{-8}$) and show synonymous mutation rates are correlated with the number of NS rare variants ($P = 0.04$) and GC content ($P < 2.4 \times 10^{-9}$) (*13*).

The excess of rare variants observed in coding regions is also due to an abundance of nonsynonymous (NS) variants segregating at low frequencies that are not seen at more common variant frequencies as a result of purifying selection. Summing across all frequencies of variant sites, S and intronic variants occurred more frequently (~70 variants per kb each) compared to UTR and nonsynonymous (NS) variant sites (~55 and ~45 per kb of UTR or NS sequence, respectively, Fig. 2A). Yet, examining the abundance of rare variants across functional categorizations of variant sites reveals little difference among classes when minor allele count is low (Fig. 1A). These patterns are likely due to an equal input of mutations for each category followed by purifying selection preventing deleterious NS and UTR variants from reaching higher frequencies (*13*, *19*). The ratio of NS:S in singletons is close to that expected amongst new mutations and then decreases with increasing frequency (Fig. 2C). Using the approach of (*2*) we estimate that while ~70% of all NS singletons in our sample are sufficiently deleterious that they will never reach frequencies >5%, only 13% of new NS mutations appear so deleterious that they would not be observed even as singletons in a sample of this size (*13*), putting an upper bound on the frequency of dominant lethal mutations (*15*). The output of functional prediction algorithms (Fig. 2, D and E) also suggest that rare variants are enriched for damaging variants.

On average, each subject carried a rare minor allele at 0.02% of all NS sites, of which ~56% are expected to be deleterious enough to never be fixed. Over 0.3% of sequenced subjects carried at least one mutation reported to be a dominant cause of disease (table S6) (*13*). We also identified variants at 0.5% < MAF ≤2 %, the so-called goldilocks variants (*20*), in that they would be common enough to be detected in large population samples and rare enough to be enriched for variants under purifying selection (Fig. 2, C to E). In the European sample, we observed 105 amino acid-changing variants in 73 genes falling within this frequency range. Half of these were predicted to be functionally damaging, relative to 31% of more common

Table 1. Comparison of classical population genetic measures of sequence diversity across studies.

Study	No. of genes*	N	Sample†	Length (Mb)	θ_π ($\times 10^{-4}$)	θ_W ($\times 10^{-4}$)
Akey (*27*)	132	23	EU	2.50	3.41	7.35
SeattleSNPs (*28*)	213	23	EU	7.26	6.81	6.36
Ahituv (*29*)	58	757	EU	0.13	4.32	10.11
Current study	202	500‡	EU	0.74	3.96	8.79
		11,000	EU	0.74	3.96	40.38
		500‡	SA	0.69	4.04	10.67
Akey *et al.*	132	24	AA	2.50	4.49	12.10
Seattle SNPs	213	24	AA	7.26	8.97	10.15
Current study	202	500‡	AA	0.70	4.89	13.78

*Studies differ in the relative proportion of coding and non-coding sequences.
†Ancestry: EU, European; AA, African-American; SA, South Asian.
‡Sampled to $n = 500$.

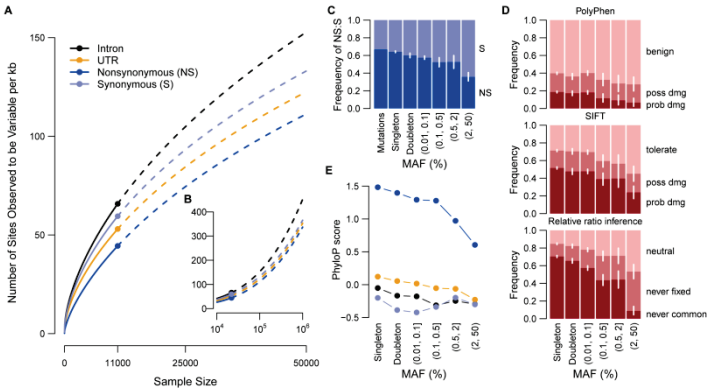

Fig. 2. (**A** and **B**) Number of variants per kilobase of intronic, UTR, nonsynonymous (NS) or synonymous (S) sequence with sample size increasing to 50,000 (A) and one million (B) Europeans. Observed numbers are given as a dot, solid and dashed lines indicate hyper-geometric expectations and jack-knife projections, respectively. (**C**) Expected ratios of NS to S variants in the absence of selection and observed ratios for different minor allele frequency (MAF) bins. (**D**) The proportion of NS variants predicted to be benign, possibly damaging or probably damaging by PolyPhen or SIFT and the proportion of NS variants that is neutral, deleterious such that they will never become common (MAF >5%) or never be fixed in Europeans as predicted by the relative ratios of NS:S variant abundances observed at different MAF (*2*). (C and D) 95% confidence intervals are represented by white lines. (**E**) phyloP score for intronic, UTR, NS and S variants for different MAF bins.

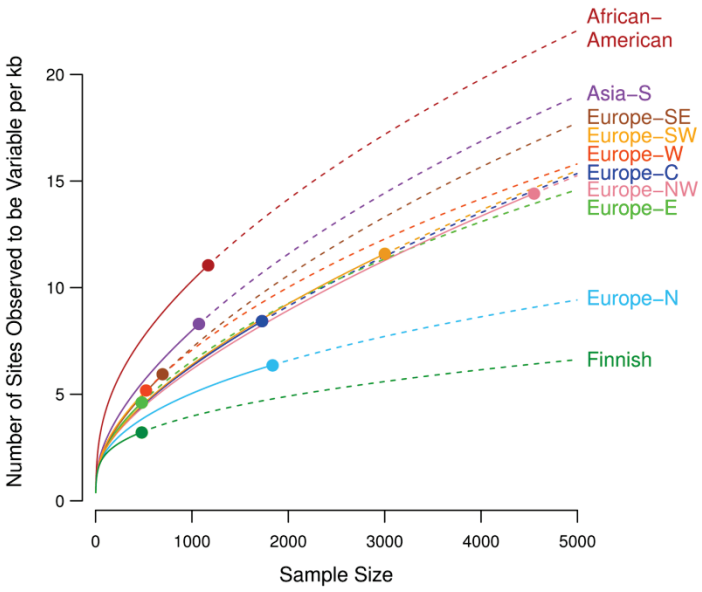

Fig. 3. Number of variants per kilobase of sequence with sample sizes increasing to 5,000 people for multiple populations. Observed numbers are given as a dot, solid and dashed lines indicate hyper-geometric expectations and jack-knife projections, respectively.

coding SNVs (>2%) and 65% of singletons. By comparison, we found 210 goldilocks variants in African Americans and 132 in South Asians, supporting the value of non-European samples for the genetic analysis of complex traits (*21*).

Rare variants can be tested in aggregate for an association with disease (*6*), where the power of the test is strongly correlated with the cumulative minor allele frequency (cMAF) of potentially deleterious SNVs within each gene (Fig. 1, E and F, and figs. S4 and S5). 37% of genes had cMAFs >0.5% of rare alleles predicted to be deleterious. We tested associations of common variants individually and rare coding variants in aggregate with the diseases represented in this study (*13*). When possible, we matched controls with cases using genome-wide genetic similarity. Nevertheless, type 1 error rate inflation consistent with effects of population stratification was observed (table S7 and fig. S6) and was worse for rare variant tests. There were no statistically significant rare variant associations, and thus no compelling evidence connecting any genes with the studied diseases. Of 13 more closely examined genes reported to be associated with six of the diseases investigated (table S8) (*22*), only the association of rare variants in *IL6* with multiple sclerosis was noteworthy (OR = 12, $P = 0.007$) (table S9).

Because rare variants are typically the result of recent mutations, they are expected to be geographically clustered or even private to specific populations. Using a measure of variant sharing between two samples (*7*), we found that for common variants, any two European populations appear to be panmictic, while for rare variants, European populations show lower levels of sharing (fig. S7). In general, the level of sharing depends on geographic distance, with the dependence increasing substantially with decreasing allele frequency (fig. S8). The Finnish population shows substantially lower levels of sharing with other European populations than predicted by geographic distance, consistent with hypotheses of a historical Finnish demographic bottleneck (*23*). Levels of rare variant sharing are even lower when comparing populations from distinct continents. Thus catalogs of rare variants will need to be generated locally across the globe (*7*, *24*).

We found substantial variation in the total abundance of variants across populations, even within Europe (Fig. 3 and fig. S7D), likely due to demographic history. In particular we observed a north-south gradient in the abundance of rare variants across Europe, with increased numbers of rare variants in Southern Europe and a very small number of variants among Finns, who had about one third as many variants as southern Europeans. The gradient is consistent with observed gradients in haplotype diversity (*25*) and a Finnish ancestral bottleneck (*23*). Association mapping approaches based on rare variant diversity levels will be more susceptible to subtle effects of population stratification (*26*) and more likely to result in false positive disease associations.

To evaluate our conclusions relative to the rest of the genome, we compared the NS:S variant ratios of the sequenced genes to the entire coding genome within the low coverage CEU 1000 Genomes Project data. The average per subject NS:S ratio from our 202 genes was 0.54, while all other genes had an average ratio of 0.94 ($P < 10^{-15}$) (fig. S9). By comparison, genes found in OMIM and the genome-wide association studies catalog (*22*) had average ratios of 0.75 and 0.78, respectively. This implies that the genes in this study are under stronger purifying selection, consistent with their choice as drug targets and importance to human health. Hence, our results cannot be simply extrapolated to the whole exome. Instead, it is likely that our results underestimate the average genetic diversity that will be found in more typical human gene coding regions, primarily regarding the amount of NS variation.

This large-scale resequencing study provides a unique description of variation for 202 drug target genes and insight into the very rare spectrum of variation. Although sequencing error might be a concern, we show that the error rates in this study are low (table S3). Another caveat

is that our inference of demographic parameters and mutation rates ignores the effects of background selection on synonymous variants. Despite these caveats, the results show there is an abundance of rare variation in human populations, and that surveys of common variants are only observing a small fraction of the genetic diversity in any gene. Further, as we observe, much of the rare variation in coding regions appears to be functional and may be crucial for yielding insights into the genetic basis of human disease. Because the genes studied are related to drug discovery, development or repositioning efforts this work has potential to help investigate target biology and drug response.

References and Notes

1. J. K. Pritchard, Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124 (2001). doi:10.1086/321272 Medline
2. G. V. Kryukov, L. A. Pennacchio, S. R. Sunyaev, Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**, 727 (2007). doi:10.1086/513473 Medline
3. G. T. Marth *et al.*; the 1000 Genomes Project, The functional spectrum of low-frequency coding variation. *Genome Biol.* **12**, R84 (2011). doi:10.1186/gb-2011-12-9-r84 Medline
4. T. A. Manolio *et al.*, Finding the missing heritability of complex diseases. *Nature* **461**, 747 (2009). doi:10.1038/nature08494 Medline
5. E. E. Eichler *et al.*, Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446 (2010). doi:10.1038/nrg2809 Medline
6. J. Asimit, E. Zeggini, Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.* **44**, 293 (2010). doi:10.1146/annurev-genet-102209-163421 Medline
7. S. Gravel *et al.*; 1000 Genomes Project, Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11983 (2011). doi:10.1073/pnas.1019276108 Medline
8. A. Coventry *et al.*, Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* **1**, 131 (2010). doi:10.1038/ncomms1130 Medline
9. T. Ohta, Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96 (1973). doi:10.1038/246096a0 Medline
10. S. H. Williamson *et al.*, Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7882 (2005). doi:10.1073/pnas.0502300102 Medline
11. H. J. Muller, Our load of mutations. *Am. J. Hum. Genet.* **2**, 111 (1950). Medline
12. A. P. Russ, S. Lampel, The druggable genome: an update. *Drug Discov. Today* **10**, 1607 (2005). doi:10.1016/S1359-6446(05)03666-4 Medline
13. Materials and methods are available as supplementary materials on *Science* Online.
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, Gene genealogies when the sample size exceeds the effective size of the population. *Mol. Biol. Evol.* **20**, 208 (2003). doi:10.1093/molbev/msg024 Medline
17. P. Awadalla *et al.*, Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. *Am. J. Hum. Genet.* **87**, 316 (2010). doi:10.1016/j.ajhg.2010.07.019 Medline
18. D. F. Conrad *et al.*; 1000 Genomes Project, Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712 (2011). doi:10.1038/ng.862 Medline
19. P. W. Messer, Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. *Genetics* **182**, 1219 (2009). doi:10.1534/genetics.109.105692 Medline
20. A. L. Price *et al.*, Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86**, 832 (2010). doi:10.1016/j.ajhg.2010.04.005 Medline
21. I. K. Kotowski *et al.*, A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. *Am. J. Hum. Genet.* **78**, 410 (2006). doi:10.1086/500615 Medline
22. L. A. Hindorff *et al.*, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9362 (2009). doi:10.1073/pnas.0903103106 Medline
23. E. Salmela *et al.*, Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. *PLoS ONE* **3**, e3519 (2008). doi:10.1371/journal.pone.0003519 Medline
24. C. D. Bustamante, E. G. Burchard, F. M. De la Vega, Genomics for the world. *Nature* **475**, 163 (2011). doi:10.1038/475163a Medline
25. O. Lao *et al.*, Correlation between genetic and geographic structure in Europe. *Curr. Biol.* **18**, 1241 (2008). doi:10.1016/j.cub.2008.07.049 Medline
26. E. S. Lander, N. J. Schork, Genetic dissection of complex traits. *Science* **265**, 2037 (1994). doi:10.1126/science.8091226 Medline
27. J. M. Akey *et al.*, Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* **2**, e286 (2004). doi:10.1371/journal.pbio.0020286 Medline
28. SeattleSNPs, http://pga.gs.washington.edu. 2012.
29. N. Ahituv *et al.*, Medical sequencing at the extremes of human body mass. *Am. J. Hum. Genet.* **80**, 779 (2007). doi:10.1086/513471 Medline
30. R. M. Durbin *et al.*, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061 (2010). doi:10.1038/nature09534 Medline
31. M. Firmann *et al.*, The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovasc. Disord.* **8**, 6 (2008). doi:10.1186/1471-2261-8-6 Medline
32. M. Preisig *et al.*, The PsyCoLaus study: methodology and characteristics of the sample of a population-based survey on psychiatric disorders and their association with genetic and cardiovascular risk factors. *BMC Psychiatry* **9**, 9 (2009). doi:10.1186/1471-244X-9-9 Medline
33. J. S. Kooner *et al.*, Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. *Nat. Genet.* **40**, 149 (2008). doi:10.1038/ng.2007.61 Medline
34. H. Ling *et al.*, Genome-wide linkage and association analyses to identify genes influencing adiponectin levels: the GEMS Study. *Obesity (Silver Spring)* **17**, 737 (2009). doi:10.1038/oby.2008.625 Medline
35. D. F. Wyszynski *et al.*, Relation between atherogenic dyslipidemia and the Adult Treatment Program-III definition of metabolic syndrome (Genetic Epidemiology of Metabolic Syndrome Project). *Am. J. Cardiol.* **95**, 194 (2005). doi:10.1016/j.amjcard.2004.08.091 Medline
36. T. L. Assimes *et al.*; Myocardial Infarction Genetics Consortium; Wellcome Trust Case Control Consortium; Cardiogenics, Lack of association between the Trp719Arg polymorphism in kinesin-like protein-6 and coronary artery disease in 19 case-control studies. *J. Am. Coll. Cardiol.* **56**, 1552 (2010). doi:10.1016/j.jacc.2010.06.022 Medline
37. V. B. Kraus *et al.*, The Genetics of Generalized Osteoarthritis (GOGO) study: study design and evaluation of osteoarthritis phenotypes. *Osteoarthritis Cartilage* **15**, 120 (2007). doi:10.1016/j.joca.2006.10.002 Medline
38. C. Vignal *et al.*, Genetic association of the major histocompatibility complex with rheumatoid arthritis implicates two non-DRB1 loci. *Arthritis Rheum.* **60**, 53 (2009). doi:10.1002/art.24138 Medline
39. S. E. Baranzini *et al.*, Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Hum. Mol. Genet.* **18**, 767 (2009). doi:10.1093/hmg/ddn388 Medline
40. J. R. Oksenberg *et al.*, Mapping multiple sclerosis susceptibility to the HLA-DR locus in African Americans. *Am. J. Hum. Genet.* **74**, 160 (2004). doi:10.1086/380997 Medline
41. B. A. Cree *et al.*, Modification of Multiple Sclerosis Phenotypes by African Ancestry at HLA. *Arch. Neurol.* **66**, 226 (2009). doi:10.1001/archneurol.2008.541 Medline
42. N. Patterson *et al.*, Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74**, 979 (2004). doi:10.1086/420871 Medline
43. E. L. Heinzen *et al.*, Rare deletions at 16p13.11 predispose to a diverse spectrum of sporadic epilepsy syndromes. *Am. J. Hum. Genet.* **86**, 707 (2010). doi:10.1016/j.ajhg.2010.03.018 Medline
44. D. Kasperaviciūte *et al.*, Common genetic variation and susceptibility to partial epilepsies: a genome-wide association study. *Brain* **133**, 2136 (2010). doi:10.1093/brain/awq130 Medline
45. H. Li *et al.*, Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch. Neurol.* **65**, 45 (2008). doi:10.1001/archneurol.2007.3 Medline
46. P. Muglia *et al.*, Genome-wide association study of recurrent major depressive

disorder in two European case-control cohorts. *Mol. Psychiatry* **15**, 589 (2010). doi:10.1038/mp.2008.131 Medline
47. C. Francks *et al*., Population-based linkage analysis of schizophrenia and bipolar case-control cohorts identifies a potential susceptibility locus on 19q13. *Mol. Psychiatry* **15**, 319 (2010). doi:10.1038/mp.2008.100 Medline
48. J. Vestbo *et al*.; ECLIPSE investigators, Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). *Eur. Respir. J.* **31**, 869 (2008). doi:10.1183/09031936.00111707 Medline
49. S. G. Pillai *et al*.; ICGN Investigators, A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet.* **5**, e1000421 (2009). doi:10.1371/journal.pgen.1000421 Medline
50. T. T. Ashburn, K. B. Thor, Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **3**, 673 (2004). doi:10.1038/nrd1468 Medline
51. Y. A. Lussier, J. L. Chen, The emergence of genome-based drug repositioning. *Sci. Transl. Med.* **3**, 96ps35 (2011). doi:10.1126/scitranslmed.3001512 Medline
52. J. Harrow *et al*., GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* **7**, (Suppl 1), S4, 1 (2006). doi:10.1186/gb-2006-7-s1-s4 Medline
53. M. Ashburner *et al*.The Gene Ontology Consortium, Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25 (2000). doi:10.1038/75556 Medline
54. R. Li, Y. Li, K. Kristiansen, J. Wang, SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713 (2008). doi:10.1093/bioinformatics/btn025 Medline
55. R. Li *et al*., SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124 (2009). doi:10.1101/gr.088013.108 Medline
56. I. M. Heid *et al*., Estimating the single nucleotide polymorphism genotype misclassification from routine double measurements in a large epidemiologic sample. *Am. J. Epidemiol.* **168**, 878 (2008). doi:10.1093/aje/kwn208 Medline
57. I. W. Saunders, J. Brohede, G. N. Hannan, Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference. *Genomics* **90**, 291 (2007). doi:10.1016/j.ygeno.2007.05.011 Medline
58. R. Ihaka, R. Gentleman, *J. Comput. Graph. Statist.* **5**, 299 (1996).
59. V. Ramensky, P. Bork, S. Sunyaev, Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30**, 3894 (2002). doi:10.1093/nar/gkf493 Medline
60. P. C. Ng, S. Henikoff, Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863 (2001). doi:10.1101/gr.176601 Medline
61. A. Stabenau *et al*., The Ensembl core software libraries. *Genome Res.* **14**, 929 (2004). doi:10.1101/gr.1857204 Medline
62. A. Siepel *et al*., Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034 (2005). doi:10.1101/gr.3715005 Medline
63. R. Blekhman *et al*., Natural selection on genes that underlie human disease susceptibility. *Curr. Biol.* **18**, 883 (2008). doi:10.1016/j.cub.2008.04.074 Medline
64. D. G. MacArthur, C. Tyler-Smith, Loss-of-function variants in the genomes of healthy humans. *Hum. Mol. Genet.* **19**, (R2), R125 (2010). doi:10.1093/hmg/ddq365 Medline
65. Z. A. Szpiech, M. Jakobsson, N. A. Rosenberg, ADZE: a rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics* **24**, 2498 (2008). doi:10.1093/bioinformatics/btn478 Medline
66. G. A. Watterson, On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256 (1975). doi:10.1016/0040-5809(75)90020-9 Medline
67. I. Ebersberger, D. Metzler, C. Schwarz, S. Pääbo, Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* **70**, 1490 (2002). doi:10.1086/340787 Medline
68. M. W. Nachman, S. L. Crowell, Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297 (2000). Medline
69. S. F. Schaffner *et al*., Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15**, 1576 (2005). doi:10.1101/gr.3709305 Medline
70. L. Excoffier, M. Foll, fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**, 1332 (2011). doi:10.1093/bioinformatics/btr124 Medline
71. A. J. Coffey *et al*., The GENCODE exome: sequencing the complete human exome. *Eur. J. Hum. Genet.* **19**, 827 (2011). doi:10.1038/ejhg.2011.28 Medline
72. M. R. Nelson *et al*., The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am. J. Hum. Genet.* **83**, 347 (2008). doi:10.1016/j.ajhg.2008.08.005 Medline
73. S. A. Bacanu, J. C. Whittaker, M. R. Nelson, How informative is a negative finding in a small pharmacogenetic study? *Pharmacogenomics J.* **12**, 93 (2012). doi:10.1038/tpj.2011.58 Medline
74. S. Hunter *et al*., InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, (Database issue), D211 (2009). doi:10.1093/nar/gkn785 Medline
75. J. A. Blake, C. J. Bult, J. A. Kadin, J. E. Richardson, J. T. Eppig; Mouse Genome Database Group, The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.* **39**, (Database issue), D842 (2011). doi:10.1093/nar/gkq1008 Medline
76. J. H. Wang *et al*.; Australian and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene), Modeling the cumulative genetic risk for multiple sclerosis from genome-wide association data. *Genome Med.* **3**, 3 (2011). doi:10.1186/gm217 Medline
77. K. S. Wang, X. F. Liu, N. Aragam, A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. *Schizophr. Res.* **124**, 192 (2010). doi:10.1016/j.schres.2010.09.002 Medline
78. M. H. Cho *et al*., Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat. Genet.* **42**, 200 (2010). doi:10.1038/ng.535 Medline
79. G. Lettre *et al*., Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project. *PLoS Genet.* **7**, e1001300 (2011). doi:10.1371/journal.pgen.1001300 Medline
80. A. Terracciano *et al*., Genome-wide association scan of trait depression. *Biol. Psychiatry* **68**, 811 (2010). doi:10.1016/j.biopsych.2010.06.030 Medline
81. P. L. De Jager *et al*.; International MS Genetics Consortium, Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat. Genet.* **41**, 776 (2009). doi:10.1038/ng.401 Medline
82. D. A. Hafler *et al*.International Multiple Sclerosis Genetics Consortium, Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.* **357**, 851 (2007). doi:10.1056/NEJMoa073493 Medline
83. M. Bahlo *et al*.; Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene), Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. *Nat. Genet.* **41**, 824 (2009). doi:10.1038/ng.396 Medline
84. P. F. Sullivan *et al*., Genomewide association for schizophrenia in the CATIE study: results of stage 1. *Mol. Psychiatry* **13**, 570 (2008). doi:10.1038/mp.2008.25 Medline
85. B. E. Aouizerat *et al*., GWAS for discovery and replication of genetic loci associated with sudden cardiac arrest in patients with coronary artery disease. *BMC Cardiovasc. Disord.* **11**, 29 (2011). doi:10.1186/1471-2261-11-29 Medline

Acknowledgments: We thank GSK colleagues who advised on the selection of genes and collections especially W. Anderson, L. Condreay, P. Agarwal, A. Hughes, J. Rubio, C. Spraggs and D. Waterworth, the sample preparation team especially J. Charnecki, M. E. Volk, D. Duran., D. Briley and K. King for data preparation, A. Slater for subject selection and preparation of genome-wide genotype data, E. Woldu for capillary sequencing, A. Nelsen, S. Buhta-Halburnt, L. Amos, J. Forte for consent review, M. Lawson for assistance in running the association analyses, J. Brown for discussions about gene feature analyses and S. Ghosh for providing reviews of the manuscript and G. Tian, H. Jiang, Z. Su, X. Sun, L. Yang and X. Zhang at BGI for sequencing. We acknowledge the work of collaborating clinicians and researchers who contributed to recruiting and characterizing subjects (*13*). J.N. and D.W. were supported by a Searle Scholars Program award to J.N.. D.K. is supported by an NIH Genome Analysis training grant. All variants described in this study have been submitted to dbSNP; accession numbers are included in database S2. Subject level sequence data for CoLaus and LOLIPOP studies are available in dbGaP. Additional subject level sequence data can be made available upon request from the authors under a Data Transfer Agreement for the purpose to understand, assess or extend the conclusions of this paper.

Supplementary Materials
www.sciencemag.org/cgi/content/full/science.1217876/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S15
Tables S1 to S17
References (*30*–*85*)
Databases S1 to S3

14 December 2011; accepted 3 May 2012
Published online 17 May 2012
10.1126/science.1217876