LETTER

Analysis of 6,515 exomes reveals the recent origin of most human protein–coding variants

Wenqing Fu[1], Timothy D. O'Connor[1], Goo Jun[2], Hyun Min Kang[2], Goncalo Abecasis[2], Suzanne M. Leal[3], Stacey Gabriel[4], David Altshuler[4], Jay Shendure[1], Deborah A. Nickerson[1], Michael J. Bamshad[1,5], NHLBI Exome Sequencing Project* & Joshua M. Akey[1]

Establishing the age of each mutation segregating in contemporary human populations is important to fully understand our evolutionary history[1,2] and will help to facilitate the development of new approaches for disease-gene discovery[3]. Large-scale surveys of human genetic variation have reported signatures of recent explosive population growth[4–6], notable for an excess of rare genetic variants, suggesting that many mutations arose recently. To more quantitatively assess the distribution of mutation ages, we resequenced 15,336 genes in 6,515 individuals of European American and African American ancestry and inferred the age of 1,146,401 autosomal single nucleotide variants (SNVs). We estimate that approximately 73% of all protein-coding SNVs and approximately 86% of SNVs predicted to be deleterious arose in the past 5,000–10,000 years. The average age of deleterious SNVs varied significantly across molecular pathways, and disease genes contained a significantly higher proportion of recently arisen deleterious SNVs than other genes. Furthermore, European Americans had an excess of deleterious variants in essential and Mendelian disease genes compared to African Americans, consistent with weaker purifying selection due to the Out-of-Africa dispersal. Our results better delimit the historical details of human protein-coding variation, show the profound effect of recent human history on the burden of deleterious SNVs segregating in contemporary populations, and provide important practical information that can be used to prioritize variants in disease-gene discovery.

As part of the US National Institutes of Health (NIH) Heart, Lung and Blood Institute (NHLBI)-sponsored Exome Sequencing Project (ESP), we sequenced the exomes of 6,515 individuals (Supplementary Table 1), including 4,298 European Americans and 2,217 African Americans. Exome data were subjected to standard quality control filters as previously described[6] (Supplementary Information), resulting in a data set of 1,146,401 autosomal protein-coding SNVs with a known ancestral state (709,816 and 643,128 in European Americans and African Americans, respectively) distributed across 15,336 protein-coding genes. To quantitatively estimate the age of each SNV (that is, the allele age), we developed a simulation approach to generate a series of coalescent trees for a specified demographic model, and estimated allele age based on a derivation described previously[7] (Supplementary Information). We verified the accuracy and robustness of this approach to factors including recombination rate heterogeneity, population growth, migration and purifying selection. Extensive coalescent simulations showed that we could accurately estimate the expected allele age from the simulated data, although the variance associated with any individual SNV can be large (Supplementary Figs 6 and 7).

We estimated the age of all 1,146,401 SNVs using 6 different demographic models[5,6,8–11], 3 of which considered recent explosive population growth[5,6,8] (Supplementary Table 2). Estimates of allele age were generally robust across different demographic models, with the largest discrepancies resulting in a twofold difference in average age across all SNVs (Supplementary Table 3 and Supplementary Fig. 8a). However, because most SNVs arose recently (see below), differences among demographic models were highly concordant (Supplementary Information). Accordingly, we report results based on a modified Out-of-Africa model[9] in which accelerated population growth began 5,115 years ago with a per-generation growth rate of 1.95% and 1.66% for European Americans and African Americans, respectively[6].

The site frequency spectrum (SFS) of protein-coding SNVs revealed an enormous excess of rare variants (Fig. 1a). Indeed, we observed an SNV approximately once every 52 base pairs (bp) and 57 bp in European Americans and African Americans, respectively, whereas in a population without recent explosive growth we would expect the SNVs to occur once every 257 bp and 152 bp in European Americans and African Americans, respectively (Supplementary Information). Thus, the European American and African American samples contain approximately fivefold and threefold increases in SNVs, respectively, attributable to explosive population growth, resulting in a large burden of rare SNVs predicted to have arisen very recently (Fig. 1b). For example, the expected age of derived singletons, which comprise 55.1% of all SNVs, is 1,244 and 2,107 years for the European American and African American samples, respectively. Overall, 73.2% of SNVs (81.4% and 58.7% in European Americans and African Americans, respectively) are predicted to have arisen in the past 5,000 years. SNVs that arose more than 50,000 years ago were observed more frequently in the African American samples (Fig. 1b), which probably reflects stronger genetic drift in European Americans associated with the Out-of-Africa dispersal.

The average age across all SNVs was 34,200 ± 900 years (± s.d.) in European Americans and 47,600 ± 1,500 years in African Americans, and these estimates were robust to sequencing errors (Supplementary Information and Supplementary Fig. 9). As expected, SNVs shared between European Americans and African Americans were significantly older (104,400 years and 115,800 years for European Americans and African Americans, respectively) than population-specific variants (5,400 years and 15,300 years in European Americans and African Americans, respectively; Fig. 1c) (t-test; $P < 0.00001$ by permutation). Furthermore, there were large and significant differences among the average allele ages of SNVs stratified by functional type (t-test; $P < 0.00001$ by permutation). For example, splice site, nonsense and non-synonymous SNVs were two to eight times younger than synonymous and non-coding variants (Fig. 1d). Moreover, we classified amino acids into four groups (non-polar and neutral, polar and neutral, acidic and polar, and basic and polar), and non-synonymous SNVs that resulted in changes between groups were significantly younger than

[1]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA. [2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA. [3]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA. [4]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [5]Department of Pediatrics, University of Washington, Seattle, Washington 98195, USA.
*Lists of participants and affiliations appear in the Supplementary Information.
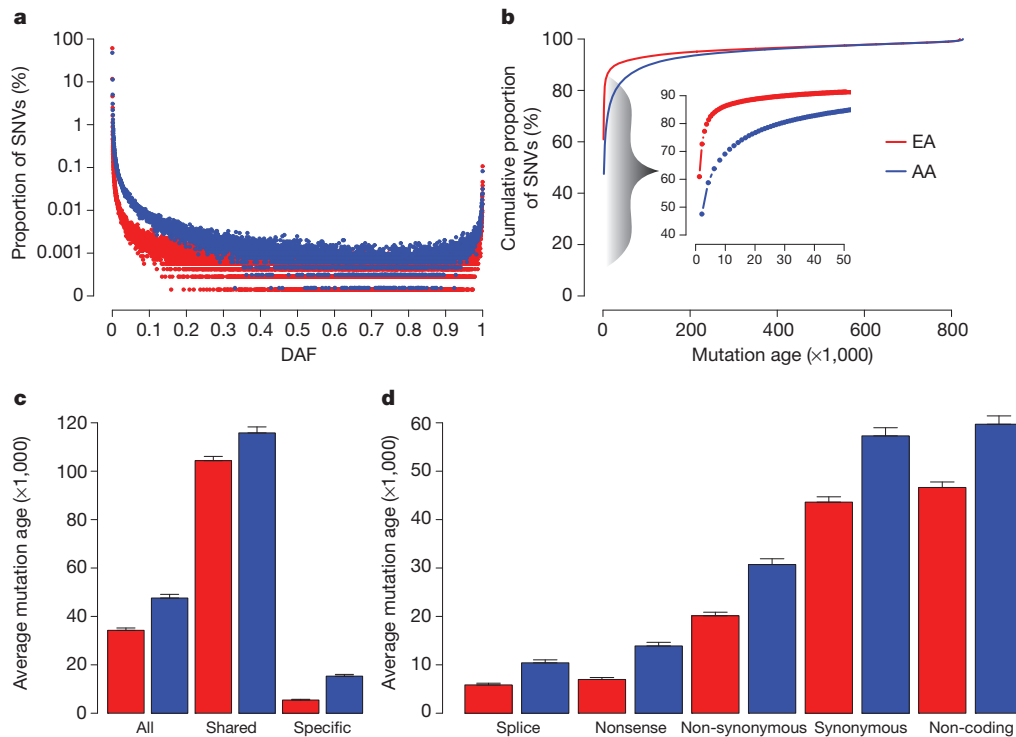
Figure 1 | The vast majority of protein-coding single-nucleotide variants arose recently. a, The site-frequency spectrum for European Americans (EAs, red) and African Americans (AAs, blue). DAF, derived allele frequency. **b**, Cumulative proportion of SNVs for a given allele age. The inset highlights the cumulative proportion of SNVs that are estimated to have arisen in the last 50,000 years. The x axis denotes allele age ($\times 1,000$) and the y axis indicates the the cumulative proportion of SNVs (%). **c**, Average age for all SNVs, SNVs found in both the European Americans and African Americans (shared), and SNVs found in only one population (specific). **d**, Average age for different types of variants. Error bars represent s.d.

those that resulted in changes within groups (t-test; $P < 0.00001$ by permutation; Supplementary Fig. 10a). These differences in average allele age are probably due to varying intensities of selective constraint among different classes of SNVs[12]. Consistent with this prediction, we observed significantly higher values of the neutrality index, a measure of the direction and degree of departure from neutral evolution, in genomic regions enriched for younger variants (Spearman's correlation; $P = 0.004$ and $P = 0.001$ for European Americans and African Americans, respectively; Supplementary Fig. 11), indicating a higher burden of deleterious SNVs.

To more directly identify putatively deleterious SNVs, we used four functional prediction methods (SIFT[13], PolyPhen2 (ref. 14), a likelihood ratio test[15], MutationTaster[16]) applicable to non-synonymous SNVs

and two conservation-based methods (GERP++[17] and PhyloP[18]) applicable to all SNVs (Supplementary Information). We found a strong inverse relationship between average SNV age and the number of methods that predicted a variant to be deleterious (Fig. 2a, b). Thus, SNVs predicted to be deleterious by multiple methods probably experience (on average) more intense purifying selection and may be of particular interest in disease mapping studies. The age of non-synonymous SNVs predicted to be deleterious by all 6 methods was 3,000 and 6,200 years in European Americans and African Americans, respectively, and 88.7% were less than 5,000 years old (92.9% and 80.6% in European Americans and African Americans, respectively).

The strengths and weaknesses of functional prediction methods vary substantially and as a result the accuracy of any single method
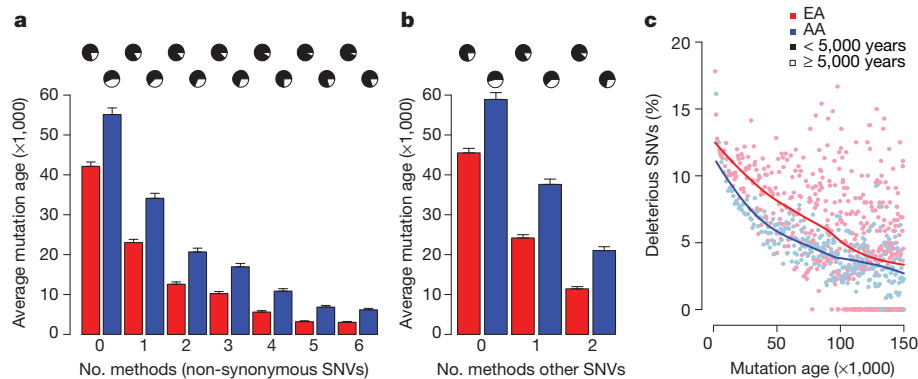

Figure 2 | Characteristics of allele age for deleterious single-nucleotide variants. a, b, Average age of non-synonymous and other SNVs as a function of the number of methods that predict the variant to be deleterious. Pie charts represent the proportion of SNVs that arose less than (black) or more than (white) 5,000 years ago. Error bars represent s.d. **c**, Relationship between the proportion of SNVs predicted to be deleterious and SNV age. Note, greater than 99% of deleterious SNVs are estimated to have arisen in the past 150,000 years. Solid lines represent a locally weighted scatterplot smoothing (LOESS) fit to the data.

is modest[15]. Therefore, we used a majority rule approach to identify a more conservative set of SNVs predicted to be deleterious[6]. Specifically, non-synonymous SNVs predicted to be functionally significant by at least four methods and all other SNVs (synonymous, splice and non-coding variants) predicted by two conservation-based methods were designated as deleterious. In total, 14.4% (164,688) of SNVs, including 152,633 nonsynonymous variants, met these criteria. We found that allele age was strongly related to the probability that a variant was predicted to be deleterious (Supplementary Fig. 12), with the fraction of SNVs predicted to be deleterious diminishing as allele age increased (Fig. 2c and Supplementary Fig. 13). The average age of conservatively defined deleterious variants was 5,200 ± 300 years for European Americans and 10,100 ± 600 years for African Americans. Moreover, 86.4% of these SNVs were predicted to have arisen in the past 5,000 years (91.2% and 77.0% for European Americans and African Americans, respectively), corresponding to the onset of accelerated population growth (Fig. 3a). In other demographic models, a similarly high proportion of deleterious SNVs were predicted to have arisen since the onset of accelerated growth rates, with the exact timing varying somewhat among models, but always in the timeframe of 5,000 to 10,000 years (Supplementary Table 3 and Supplementary Fig. 8b, c).

Moreover, 7,197 (57.4%) of the 12,533 genes in European Americans and 4,534 (37.5%) of the 11,607 genes in African Americans that harbour one or more deleterious variants only possess deleterious SNVs with an estimated age of less than 5,000 years (Fig. 3b). Thus, recent accelerated population growth has had a large influence on the number of genes harbouring deleterious variants in contemporary populations.

Notably, after correcting for exon length of each gene, 3 and 18 genes in European Americans and in African Americans, respectively, have a significant excess of deleterious variants that arose after the onset of recent accelerated growth ($P \leq 3 \times 10^{-6}$; Supplementary Table 4), including 12 genes that have been associated with human diseases[19] such as *LAMC1* (premature ovarian failure[20]), *LRP1* (Alzheimer's disease[21]), *CPE* (coronary artery atherosclerosis[22]) and *KIAA0196* (hereditary spastic paraplegia[23]).

Next, we investigated the distribution of ages for conservatively defined deleterious SNVs in 849 genes that cause Mendelian disorders[24], 2,663 genes associated with complex diseases[19], 1,226 genes considered 'essential' (that is, a mouse knockout associated with lethality or sterility)[25] and 11,711 genes classified as 'other' (Supplementary Information). The proportions of deleterious SNVs in genes for Mendelian disorders (15.9%), essential genes (15.2%) and genes associated with complex diseases (15.1%) were each significantly higher (Fisher's exact test, $P < 10^{-16}$) than the proportion in other genes (14.0%). In the European American samples, the proportion of deleterious SNVs did not decline monotonically as a function of age for Mendelian and essential genes. Instead, the proportion of deleterious variants with an estimated age of 50,000 to 100,000 years in Mendelian disease genes and 100,000 to 150,000 years in essential genes were elevated (Fig. 4a). This pattern was not observed in the African American samples (Fig. 4a). To explore this observation, we carried out simulations to estimate the probability that a deleterious SNV survives to the present day as a function of when the variant arose, the magnitude of selection and the presence or absence of an Out-of-Africa bottleneck (Supplementary
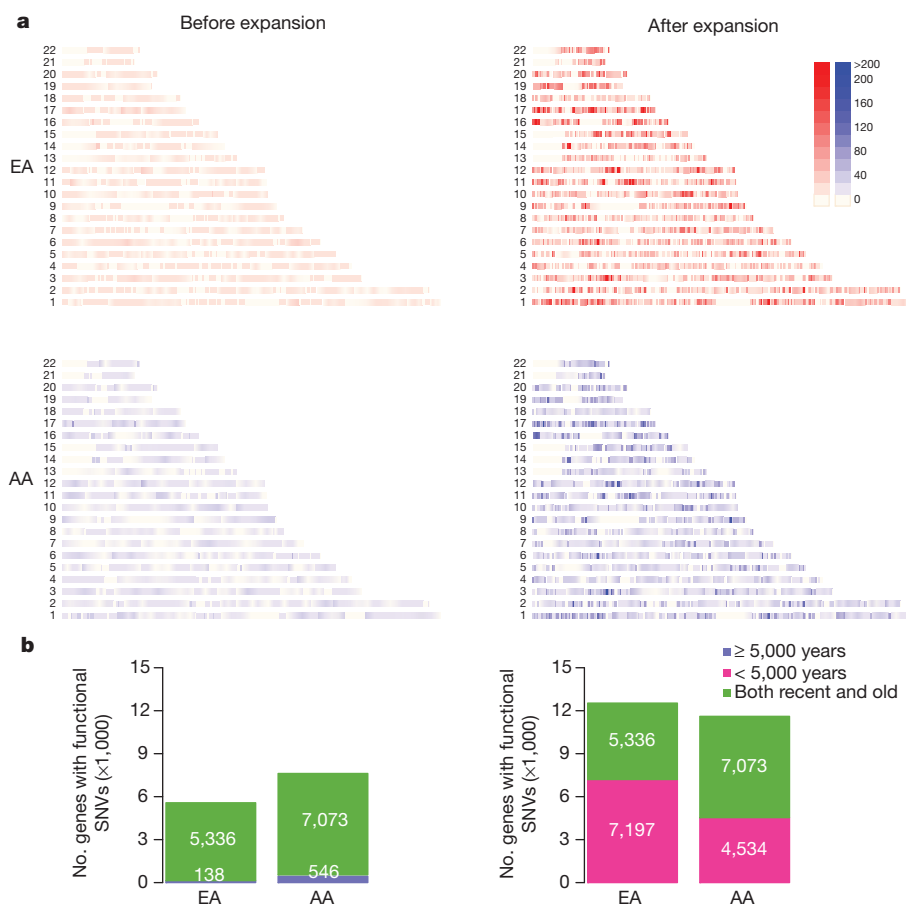

Figure 3 | Distribution of deleterious single-nucleotide variants across the exome before and after recent accelerated population growth. a, Heat map representation of deleterious SNV density (number of deleterious protein-coding variants per Mb) for European Americans (EAs, red) and African Americans (AAs, blue). The distributions of deleterious SNVs across the exome before and after recent accelerated population growth are shown in the left and right panels, respectively. Numbers on the *y* axis represent chromosome number. **b,** The bar plots summarize the number of genes segregating one or more deleterious SNVs that arose before (left) or after (right) recent accelerated population growth.
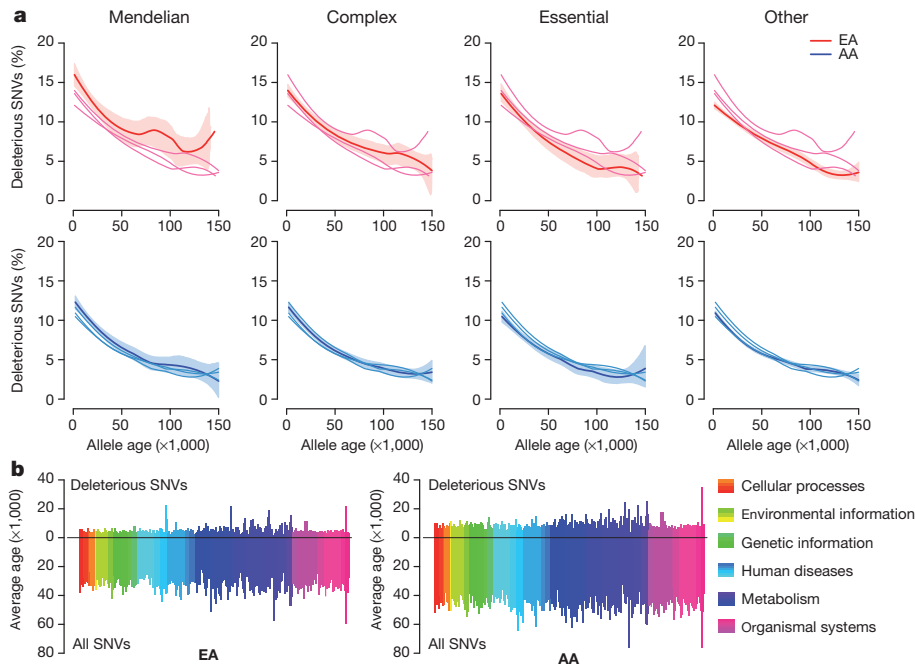
Figure 4 | Heterogeneity of allele age across genes and pathways.
a, Distribution of the proportion of deleterious SNVs for Mendelian, complex, essential and other genes in European Americans (EAs, top) and African Americans (AAs, bottom) versus age. Data for each of the four categories of genes is shown in each plot, with darker lines representing the specific gene class indicated by the column label. Shaded regions define 95% confidence intervals obtained by bootstrapping. **b**, Average ages for deleterious (projecting up) and all (projecting down) SNVs across 235 KEGG pathways that can be organized into six broad classes. Each of the six classes is comprised of multiple sub-classes, indicated by the different colour shadings (Supplementary Information).

Information). Simulations of deleterious alleles in the presence of a bottleneck recapitulated the patterns observed in European Americans (Supplementary Fig. 14). Specifically, in the presence of a bottleneck, weakly deleterious alleles (selection coefficient, $s \leq 0.001$) have an increased probability of survival precisely in the intervals 50,000 to 100,000 years and 100,000 to 150,000 years. Thus, our simulations suggest that genes underlying disease and essential genes are more functionally constrained relative to other genes, and the bottleneck associated with the Out-of-Africa dispersal led to less efficient purging of weakly deleterious alleles[26].

Finally, we found that the average age of deleterious variants (and the proportion of deleterious variants; Supplementary Fig. 15) was significantly different across 235 KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways (Kruskal–Wallis rank sum test; $P = 0.0025$ and $P < 0.00001$ for European Americans and African Americans, respectively; Fig. 4b and Supplementary Information). The average age across pathways did not vary significantly when all SNVs were considered (Kruskal–Wallis rank sum test; $P = 0.259$ and $P = 0.075$ for European Americans and African Americans, respectively), indicating that the differences observed for deleterious variants probably represent heterogeneity of functional constraint across pathways. In general, the average age of deleterious variants in metabolic pathways was older than that in other pathways (Mann–Whitney test, $P \leq 0.0001$ for European Americans and African Americans, respectively), suggesting they are subject to less functional constraint. Conversely, deleterious variants in human disease pathways (Mann–Whitney test, $P = 0.03$ for African Americans) and in pathways involved in organismal systems were significantly younger than other pathways (Mann–Whitney test, $P = 0.04$ and $P = 0.002$ for European Americans and African Americans, respectively).

In summary, the spectrum of protein-coding variation is considerably different today compared to what existed as recently as 200 to 400 generations ago. Of the putatively deleterious protein-coding SNVs, 86.4% arose in the last 5,000 to 10,000 years, and they are enriched for mutations of large effect (Supplementary Fig. 14) as selection has not had sufficient time to purge them from the population. Thus, it seems

likely that rare variants have an important role in heritable phenotypic variation, disease susceptibility and adverse drug responses. In principle, our results provide a framework for developing new methods to prioritize potential disease-causing variants in gene-mapping studies. More generally, the recent dramatic increase in human population size, resulting in a deluge of rare functionally important variation, has important implications for understanding and predicting current and future patterns of human disease and evolution. For example, the increased mutational capacity of recent human populations has led to a larger burden of Mendelian disorders, increased the allelic and genetic heterogeneity of traits, and may have created a new repository of recently arisen advantageous alleles that adaptive evolution will act upon in subsequent generations[27].

METHODS SUMMARY

Exome sequences were obtained for 6,823 individuals, who were sequenced to high coverage (median depth greater than 100×) on an Illumina GAII or HiSeq2000. Library construction, exome capture, sequencing, mapping, calling and filtering were carried out as described previously, with minor modifications[6] (see also Supplementary Information). After quality control and removal of related individuals, 6,515 individuals were retained. Ancestry of each individual was inferred by principal component analysis (PCA) performed on the sequence data. We developed a simulation approach based on coalescent theory to estimate allele age, which was applied to 1,146,401 autosomal SNVs with known ancestral states. A complete description of the materials and methods is provided in the Supplementary Information.

1. Kimura, M. & Ota, T. The age of a neutral mutant persisting in a finite population. *Genetics* **75**, 199–212 (1973).
2. Tishkoff, S. A. & Verrelli, B. C. Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu. Rev. Genomics Hum. Genet.* **4**, 293–340 (2003).
3. Slatkin, M. & Rannala, B. Estimating allele age. *Annu. Rev. Genomics Hum. Genet.* **1**, 225–249 (2000).
4. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740–743 (2012).

5. Nelson, M. R. et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337, 100–104 (2012).
6. Tennessen, J. A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64–69 (2012).
7. Griffiths, R. C. & Tavaré, S. The age of a mutation in a general coalescent tree. Commun. Stat. Stoch. Models 14, 273–295 (1998).
8. Coventry, A. et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nature Commun. 1, 131 (2010).
9. Gravel, S. et al. Demographic history and rare allele sharing among human populations. Proc. Natl Acad. Sci. USA 108, 11983–11988 (2011).
10. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5, e1000695 (2009).
11. Schaffner, S. F. et al. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15, 1576–1583 (2005).
12. Gibson, G. Rare and common variants: twenty arguments. Nature Rev. Genet. 13, 135–145 (2012).
13. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols 4, 1073–1081 (2009).
14. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. Nature Methods 7, 248–249 (2010).
15. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. Genome Res. 19, 1553–1561 (2009).
16. Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. Nature Methods 7, 575–576 (2010).
17. Davydov, E. V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLOS Comput. Biol. 6, e1001025 (2010).
18. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20, 110–121 (2010).
19. Becker, K. G., Barnes, K. C., Bright, T. J. & Wang, S. A. The genetic association database. Nature Genet. 36, 431–432 (2004).
20. Pyun, J. A., Cha, D. H. & Kwack, K. LAMC1 gene is associated with premature ovarian failure. Maturitas 71, 402–406 (2012).
21. Liu, Q. et al. Amyloid precursor protein regulates brain apolipoprotein E and cholesterol metabolism through lipoprotein receptor LRP1. Neuron 56, 66–78 (2007).
22. Jia, E. Z. et al. Association of the mutation for the human carboxypeptidase E gene exon 4 with the severity of coronary artery atherosclerosis. Mol. Biol. Rep. 36, 245–254 (2009).
23. Valdmanis, P. N. et al. Mutations in the KIAA0196 gene at the SPG8 locus cause hereditary spastic paraplegia. Am. J. Hum. Genet. 80, 152–161 (2007).
24. Blekhman, R. et al. Natural selection on genes that underlie human disease susceptibility. Curr. Biol. 18, 883–889 (2008).
25. Liao, B. Y., Scott, N. M. & Zhang, J. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol. Biol. Evol. 23, 2072–2080 (2006).
26. Lohmueller, K. E. et al. Proportionally more deleterious genetic variation in European than in African populations. Nature 451, 994–997 (2008).
27. Hawks, J., Wang, E. T., Cochran, G. M., Harpending, H. C. & Moyzis, R. K. Recent acceleration of human adaptive evolution. Proc. Natl Acad. Sci. USA 104, 20753–20758 (2007).

Author Contributions W.F. and J.M.A. conceived the analyses. D.A.N., S.G. and D.A. oversaw data generation and quality control. G.J., H.M.K. and G.A. developed algorithms and identified SNVs from the sequencing data. W.F. carried out the majority of analyses with contributions from T.D.O. W.F., M.J.B., J.S. and J.M.A. analysed the data and wrote the manuscript with contributions from all authors. W.F., T.D.O., S.M.L., J.S., D.A.N., M.J.B. and J.M.A. are members of the Seattle Grand Opportunity (GO) group and G.J., H.M.K., G.A., S.G. and D.A. are members of the Broad GO group, which are both sub-groups of the NHLBI Exome Sequencing Project (ESP).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to W.F. (wqfu@u.washington.edu) or J.M.A. (akeyj@u.washington.edu).

Filtered sets of annotated variants and their allele frequencies are available at http://evs.gs.washington.edu/EVS/ and genotypes and phenotypes from a large subset of individuals are also available through dbGaP (http://www.ncbi.nlm.nih.gov/gap) using the following accession information: NHLBI GO-ESP: Women's Health Initiative Exome Sequencing Project (WHI) – WHISP, WHISP_Subject_Phenotypes, pht002246.v2.p2, phs000281.v2.p2; NHLBI GO-ESP: Heart Cohorts Exome Sequencing Project (JHS), ESP_HeartGO_JHS_LDLandEOMI_Subject_Phenotypes, pht002539.v1.p1, phs000402.v1.p1; NHLBI GO-ESP: Heart Cohorts Exome Sequencing Project (FHS), HeartGO_FHS_LDLandEOMI_PhenotypeDataFile, pht002476.v1.p1, phs000401.v1.p1; NHLBI GO-ESP: Heart Cohorts Exome Sequencing Project (CHS), HeartGO_CHS_LDL_PhenotypeDataFile, pht002536.v1.p1, phs000400.v1.p1; NHLBI GO-ESP: Heart Cohorts Exome Sequencing Project (ARIC), ESP_ARIC_LDLandEOMI_Sample, pht002466.v1.p1, phs000398.v1.p1; NHLBI GO-ESP: Lung Cohorts Exome Sequencing Project (Cystic Fibrosis), ESP_LungGO_CF_PA_Culture_Data, pht002227.v1.p1, phs000254.v1.p1; NHLBI GO-ESP: Early-Onset Myocardial Infarction (Broad EOMI), ESP_Broad_EOMI_Subject_Phenotypes, pht001437.v1.p1, phs000279.v1.p1; NHLBI GO-ESP: Lung Cohorts Exome Sequencing Project (Pulmonary Arterial Hypertension), PAH_Subject_Phenotypes_Baseline_Measures, pht002277.v1.p1, phs000290.v1.p1; NHLBI GO-ESP: Lung Cohorts Exome Sequencing Project (Lung Health Study of Chronic Obstructive Pulmonary Disease), LHS_COPD_Subject_Phenotypes_Baseline_Measures, pht002272.v1.p1, phs000291.v1.p1.