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

Jacob A. Tennesen,1* Abigail W. Bigham,2† Timothy D. O’Connor,1* Wenging Fu,1 Eimear E. Kenny,3 Simon Gravel,3 Sean McGee,1 Ron Do,4,5 Xiaoming Liu,6 Goo Jun,7 Hyun Min Kang,7 Daniel Jordan,6 Suzanne M. Leal,9 Stacey Gabriel,4 Mark J. Rieder,1 Goncalo Abecasis,7 David Altshuler,4 Deborah A. Nickerson,1 Eric Boerwinkle,6,10 Shamil Sunyaev,4,8 Carlos D. Bustamante,3 Michael J. Bamshad,1,2‡ Joshua M. Akey,1‡ Broad GO, Seattle GO, on behalf of the NHLBI Exome Sequencing Project

1Department of Genome Sciences, University of Washington, Seattle, Washington, USA. 2Department of Pediatrics, University of Washington, Seattle, Washington, USA. 3Department of Genetics, Stanford University, Stanford, California, USA. 4Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. 5The Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. 6Human Genetics Center, University of Texas Health Sciences Center at Houston, Houston, Texas, USA. 7Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA. 8Division of Genetics, Brigham & Women’s Hospital, Harvard Medical School, Boston, MA, USA. 9Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. 10Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA.

As a first step toward understanding how rare variants contribute to risk for complex diseases, we sequenced 15,585 human protein-coding genes to an average median depth of 111x in 2,440 individuals of European (n=1,351) and African (n=1,088) ancestry. We identified >500,000 single nucleotide variants (SNVs), the majority of which were rare (86% with a minor allele frequency < 0.5%), novel (82%), and population-specific (82%). On average, 2.3% of the 13,595 SNVs each person carries were predicted to impact protein function of ~313 genes per genome, and ~95.7% of SNVs predicted to be functionally important were rare. This excess of rare functional variants is due to the combined effects of explosive, recent accelerated population growth and weak purifying selection. Furthermore, we show that large sample sizes will be required to associate rare variants with complex traits.
in either splice sites or targeted noncoding regions. The majority of SNVs (411,084; 82%) were novel with more novel SNVs observed in AA (240,341) than in EA (204,415); although the proportion of SNVs that were novel was higher in EA compared to AA (76.0% vs. 73.8%; $\chi^2 = 398.3$, df = 1, p-value < 10$^{-16}$). Approximately 98% (402,813) of novel SNVs were rare, and 48.9% of all novel, rare SNVs were nonsynonymous.

The AA and EA sample sizes provided ~90% power to detect variants with a MAF $\geq 0.1\%$, and nearly 100% power to detect common variants (MAF $\geq 5\%$) (18; Tables S1, S2, Fig. S10). With our large sample size, the proportion of singletons identified rapidly decreased to a nearly constant rate of new singleton discovery such that each additional exome contributed ~200 novel SNVs (18; Fig. S11). The number of SNVs per gene rapidly stabilized for common variants in small sample sizes (~100 individuals), whereas the number of rare variants continued to increase linearly (18; Fig. S12).

**Fig. 1. Characteristics of protein-coding variation in humans.** A) Number of nonsynonymous SNVs predicted to be functionally important as a function of seven different methods (18). B) Distribution of nucleotide diversity ($\pi$) across the genome in AA (blue) and EA (red). The value of $\pi$ for each gene is shown as a vertical line. The middle panel shows the difference in diversity between EA and AA ($\Delta\pi = \pi_{EA} - \pi_{AA}$), scaled between 0 and 1. C) Distribution of the proportion of total diversity, $\pi$, attributable to SNVs with different MAFs in the EA and AA samples. The x-axis is binned in increments of 0.5%.
Of the total SNVs, 57% (285,857) were singletons, and SNVs with ≤ 3 minor alleles accounted for 72% of all variants (18; Fig. S9D). The proportion of singletons observed in AA (49.3%, n=140,818) was lower than that observed in EA (50.7%, n=144,821), but the overall site frequency spectra (SFS) and the SFS for both AA and EA are highly skewed, exhibiting a large excess of rare variants relative to the standard neutral model (18, 19; Fig. S9D). The skew of the SFS was greater for EA than AA and at equal sample sizes, the odds that a SNV was a singleton was 1.7 times greater for EA than AA. Consistent with these observations, Tajima’s D was highly negative for both EA (-2.12) and AA (-2.10) and dropped precipitously as sample size increased (18; Fig. S9E), highlighting that sequencing a large number of individuals pro-
SNVs (74% of nonsynonymous and 6% of synonymous variants) are with the SFS that we designate SFS-Del; (likelihood ratio test, and Mutation Taster) and three conservation based methods applicable to nonsynonymous variants (PolyPhen2, SIFT, a)

Patterns of coding variation by gene and pathway. The median number of SNVs per gene was 24, ranged between 0 and 761, and was significantly different (Wilcoxon-Rank Sum Test; p-value < 10^-15) between AA (median 16, range 0-566) and EA (median 13, range 0-410). Mutational target size plays an important role in governing differences in polymorphism across loci, as gene length accounts for 76.6% of variation in the number of SNVs across genes (95% bootstrap CI = 73.9%-79.1%; p < 10^-15).

The proportion of rare variants per base pair (bp) in each gene was higher (mean = 2.015%; 95% range = 0.621% - 4.057%) than that of common variants (0.334%; 95% range = 0.000% - 1.143%), and the average ratio of rare to common alleles per bp was ~6:1. We identified 110 genes that showed an unusually high proportion of rare variants, including six histone genes that are likely subject to greater selective constraint (18; Table S3). The number of putatively functional variants also varied widely across genes (Fig. S14B) ranging from 0 to >100, with a median of two in both the EA and AA samples

Nucleotide diversity (π), varied considerably among genes, ranging from ~0 to 1.319% per bp (mean = 0.042%; Fig. 1B). Mean π in AA (0.047%) was significantly higher (p-value < 10^-15, paired t test) than π in EA (0.035%), and π per gene was modestly correlated (r² = 63%; p-value < 10^-15) between AA and EA (18; Fig. S15). Rare variants account for 4% of total diversity, more than any other MAF bin (of width 0.5%) in both EA and AA (Fig. 1C). Rare and low frequency SNVs comprise ~13% and 20% of total diversity in the EA and AA sample, respectively (Fig. 1C). In both samples, estimates of π were highest for HLA loci and other genes related to immune function, such as DEFBI08B, and olfactory receptors (Fig. 1B). When genes were grouped into functional categories by KEGG pathway, estimates of π were highest for pathways related to immune function and olfaction and lowest for pathways involved in basic cellular processes (18; Fig. S16).

Abundance of rare variation explained by human demographic history. The excess of rare variation across the exome is consistent with explosive human population growth (22). To investigate this further, we used an Out-of-Africa (OOA) demographic model (23) to predict the expected joint distribution of allele frequencies between the EA and AA samples via a diffusion approximation (18). The OOA model, modified to account for admixture, captures prominent features of the joint frequency distribution. However, both populations contain more rare variants than predicted by this model (18; Fig. 2A,B), most likely due to rapid population growth in the last few thousand years that is undetectable with smaller sample sizes (18; Fig. S9E). We revisited the demographic model from Gravel et al. (23), allowing for a reduced initial European expansion that is compensated for by accelerated growth starting after the split of European and Asian populations. Similarly, we introduced a phase of exponential growth in the African population starting at the same time. The resulting demographic model is an improved fit to the synonymous site-frequency spectrum (18; Fig. 2B), and strongly supports a recent, dramatic acceleration of population growth. The maximum-likelihood time for accelerated growth was 5,115 years ago (Fig. 2B).

The EA population growth, previously estimated at 0.5% per generation, is now modeled at the first step as 0.307% (sd ± 0.003%), followed by explosive growth of 1.95% (sd ± 0.03%) over the last 5,115 years. The growth in the AA sample during this same period is estimated to be 1.66% (sd ± 0.03%). The estimated standard deviations (18) are quite small, and for data sets of this scale, it is likely that other sources of uncertainty (e.g., mutation rate or model specification) play a more important role compared to finite genome fluctuations. The final population sizes in this model are lower than current census sizes, and we speculate that larger sample sizes will be necessary to fully capture the signature recent growth rate expansion imparted on patterns of DNA sequence variation.

Impact of natural selection on rare coding variation. To investigate the effect of purifying selection on nonsynonymous variants, we examined the relationship between MAF of nonsynonymous SNVs and functional prediction scores from SIFT, Polyphen2, a likelihood ratio test statistic, and Mutation Taster (18). Each prediction score showed a significant (p < 10^-16) negative correlation with MAF in the combined sample (Fig. 3A) as well as in each sample separately (18). Moreover, the proportion of predicted deleterious changes was negatively correlated with MAF (Fig. 3B). We next mapped 31,115 nonsynonymous SNVs to known protein structures and classified them into different structural categories (Fig. 3C; 18). Nonsynonymous SNVs in all categories, except sites that contact other protein chains, showed a significant excess of rare variants compared to synonymous sites (Fig. 3C), as expected if subjected to purifying selection. The relative effect sizes show that categories of variants with direct functional importance (e.g., active sites, enrichment 2.8%; ligand-binding residues, enrichment 1.7%) are under stronger constraint than categories important for structural stability. The exception is residues that form side-chain hydrogen bonds, which show a 2.8% enrichment of rare variants, suggesting that hydrogen bonds make

Abundance of rare variation explained by human demographic history. The excess of rare variation across the exome is consistent with explosive human population growth (22). To investigate this further, we used an Out-of-Africa (OOA) demographic model (23) to predict the expected joint distribution of allele frequencies between the EA and AA samples via a diffusion approximation (18). The OOA model, modified to account for admixture, captures prominent features of the joint frequency distribution. However, both populations contain more rare variants than predicted by this model (18; Fig. 2A,B), most likely due to rapid population growth in the last few thousand years that is undetectable with smaller sample sizes (18; Fig. S9E). We revisited the demographic model from Gravel et al. (23), allowing for a reduced initial European expansion that is compensated for by accelerated growth starting after the split of European and Asian populations. Similarly, we introduced a phase of exponential growth in the African population starting at the same time. The resulting demographic model is an improved fit to the synonymous site-frequency spectrum (18; Fig. 2B), and strongly supports a recent, dramatic acceleration of population growth. The maximum-likelihood time for accelerated growth was 5,115 years ago (Fig. 2B).

The EA population growth, previously estimated at 0.5% per generation, is now modeled at the first step as 0.307% (sd ± 0.003%), followed by explosive growth of 1.95% (sd ± 0.03%) over the last 5,115 years. The growth in the AA sample during this same period is estimated to be 1.66% (sd ± 0.03%). The estimated standard deviations (18) are quite small, and for data sets of this scale, it is likely that other sources of uncertainty (e.g., mutation rate or model specification) play a more important role compared to finite genome fluctuations. The final population sizes in this model are lower than current census sizes, and we speculate that larger sample sizes will be necessary to fully capture the signature recent growth rate expansion imparted on patterns of DNA sequence variation.

Impact of natural selection on rare coding variation. To investigate the effect of purifying selection on nonsynonymous variants, we examined the relationship between MAF of nonsynonymous SNVs and functional prediction scores from SIFT, Polyphen2, a likelihood ratio test statistic, and Mutation Taster (18). Each prediction score showed a significant (p < 10^-16) negative correlation with MAF in the combined sample (Fig. 3A) as well as in each sample separately (18). Moreover, the proportion of predicted deleterious changes was negatively correlated with MAF (Fig. 3B). We next mapped 31,115 nonsynonymous SNVs to known protein structures and classified them into different structural categories (Fig. 3C; 18). Nonsynonymous SNVs in all categories, except sites that contact other protein chains, showed a significant excess of rare variants compared to synonymous sites (Fig. 3C), as expected if subjected to purifying selection. The relative effect sizes show that categories of variants with direct functional importance (e.g., active sites, enrichment 2.8%; ligand-binding residues, enrichment 1.7%) are under stronger constraint than categories important for structural stability. The exception is residues that form side-chain hydrogen bonds, which show a 2.8% enrichment of rare variants, suggesting that hydrogen bonds make
a large contribution to protein folding and stability.

To investigate selective constraint acting on synonymous variants, we calculated the correlation between the derived allele frequency (DAF) of synonymous variants and their corresponding change in the relative adaptiveness value, or w score (24). The w score summarizes information about selective constraints on the efficiency of codon-anticodon coupling and the number of tRNA gene copies in the genome. Negative values indicate synonymous variants that may decrease translational efficiency or accuracy. We found a weak but significant positive correlation between DAF and change in w score (r = 0.03; p-value < 10^{-16}), consistent with the action of purifying selection (Fig. 3D).

We examined selective sweeps by identifying genes with high ratios of divergence (human-specific lineage substitutions relative to chimp and macaque) compared to polymorphism within humans, which are predicted to increase between-species divergence and decrease within-population diversity. We identified genes in which the ratio of nonsynonymous to synonymous divergence was high relative to the ratio of nonsynonymous to synonymous SNVs (25). We also identified genes with either a high or low ratio of π in AA relative to π in EA and genes with diversity estimates in the bottom 20th percentile in which at least one SNV had an F_{ST} ≥ 0.3. In total, 114 genes met one or more of these criteria (18; Table S4). Approximately 25% of these genes have been implicated as targets of positive selection (26).

Implications for disease and personal genomics. We evaluated gene-specific power of rare variant association studies in the EA and AA samples. We used Fisher’s Exact Test (FET), a robust approach for aggregate testing of rare variation at a locus (27), to determine the power to detect an association for each gene harboring rare causal variants with...
odds ratios (OR) of 1.5 or 5 in 400 cases and 400 controls (18). In both the EA and AA samples, cases and controls were sampled from 1,000 individuals selected to minimize any confounding effects of population stratification (18; Fig. S17), with power calculations assuming a Type I error rate of \( \alpha = 0.001 \). In each sample, power varies widely across loci, and <5% of genes (beige) achieve 80% power even when relatively strong effects (OR = 5) are modeled (Fig. 4A); when causal variants are assumed to have an OR of 1.5, no genes achieve 80% power (18; Fig. S18). Furthermore, although the AA sample has uniformly higher power per gene relative to the EA sample (Fig. 4A), caution is warranted because this is largely a function of our modeling assumptions (18).

The mean number of SNVs per exome (homozygous non-reference and heterozygous genotypes) was 13,595, and ~66% (8,893) of these sites were heterozygous. As expected, AAs had significantly more SNVs per exome than EAs (15,073 versus 12,406, Mann-Whitney test, \( p < 10^{-15} \)), which is true for all classes of sites (Fig. 4B). Moreover, on average, each individual possessed 35 nonsense variants and was homozygous for at least one non-reference nonsense variant; 318 individuals (181 AA and 137 EA) were compound heterozygotes for nonsense SNVs. The mean number of novel SNVs per individual was 549 overall, but AA had more than twice the number of novel SNVs compared to EA’s (762 versus 362, respectively; \( p = 1.9 \times 10^{-7} \) correcting for differences in the mean number of SNVs between populations). The fraction of overall variation that was novel in AA was higher than in EA (5% and 3%, respectively; \( p < 10^{-15} \)). Finally, while most protein coding variants were rare in the full AA and EA population samples, the majority of SNVs found in an average individual were common (Fig. 4C).

We next examined the distribution of functionally important variations...
tion, functionally important singletons, and the proportion of functionally important SNVs per individual (Fig. 4D) using both conservative and liberal criteria (18). On average, individuals possess between 318 and 580 predicted functional protein-coding SNVs depending on how functional variants are defined, with slightly more in AA than in EA individuals (Fig. 4D). The average number of predicted functional singletons per individual was more robust to the definition of functional variants, and ranged from 25 to 31, and also was slightly higher in AA compared to EA individuals (Fig. 4D). In both cases, however, there was more variation among individuals than between populations.

Finally, the average proportion of predicted functional SNVs per individual varied between 2.3% and 4.2% (Fig. 4D). When the more liberal definition of functional SNVs was used, EA individuals have a significantly higher proportion of predicted functional SNVs compared to AA individuals (Fig. 4D; Wilcoxon-Rank Sum Test; p-value < 10^-15), consistent with empirical estimates and theoretical expectations due to the lower EA effective population size (28, 29). However, when the more conservative definition was used, this pattern was reversed and AA individuals have a significantly higher proportion of predicted functional SNVs compared to EA individuals (Fig. 4D; Wilcoxon-Rank Sum Test; p-value < 10^-15). These results highlight how the definition of functional variants can influence inferences and underscore the importance of continued methodological development to robustly identify functionally important variation. Nonetheless, there was considerable rare genetic variation among individuals that is predicted to be functional, which could explain variability in disease risk and adverse drug response.

Conclusion. Our results have several important implications for human disease gene mapping and personal genomics. In particular, the vast majority of protein-coding variation is evolutionarily recent, rare, and enriched for deleterious alleles. Thus, rare variation likely makes an important contribution to human phenotypic variation and disease susceptibility. However, detecting the effects of rare variants requires very large sample sizes because the power to detect an association is low for most human genes. Accounting for the SFS on a gene-by-gene basis should facilitate the development of more powerful association tests. Additionally, because most rare SNVs are population-specific, replication of disease associations across populations may be challenging. Finally, as whole-genome sequencing at high coverage becomes increasingly feasible, statistical and experimental methods that accurately identify functionally important protein-coding and regulatory variation are needed to empower association studies, identify variants causally related to disease, and provide clinically actionable information.

References and Notes
18. Supporting material is available on Science Online.
Acknowledgments

47. Y. Y. Waldman, T. Tuller, T. Shlomi, R. Sharan, E. Ruppin, Translation et al
46. S. Gravel.; 1000 Genomes Project, Demographic history and rare allele
44. M. E. Zwick, D. J. Cutler, A. Chakravarti, Patterns of genetic variation in
et al. NISC Comparative Sequencing Program, Distribution and
42. G. M. Cooper et al
41. J. M. Schwarz, C. Rödelsperger, M. Schuelke, D. Seelow, MutationTaster
40. S. Chun, J. C. Fay, Identification of deleterious mutations within three human
39. P. Kumar, S. Henikoff, P. C. Ng, Predicting the effects of coding non-
38. I. A. Adzhubei, Robust relationship inference in genome-wide
37. A. Manichaikul et al
36. R. M. Durbin PAH_Subject_Phenotypes_Baseline_Measures, pht002277.v1.p1,
35. CDB is on the scientific advisory board for Personalis, Inc, Mubadala Medical Holding Company, 23andme “Roots into the future” project,
34. Acknowledgments: We acknowledge the support of the National Heart, Lung,
and Blood Institute (NHLBI) and the contributions of the research institutions,
study investigators, field staff and study participants in creating this resource for
biomedical research; and the Population Genetics Project Team. Funding for GO
ESP was provided by NHLBI grants RC2 HL-103010 (HeartGO),
RC2 HL-102925 (SeattleGO) and RC2 HL-102926 (BroadGO).
33. We acknowledge the support of the National Heart, Lung, and Blood Institute (NHLBI) and the contributions of the research institutions, study investigators, field staff and study participants in creating this resource for biomedical research; and the Population Genetics Project Team. Funding for GO ESP was provided by NHLBI grants RC2 HL-103010 (HeartGO), RC2 HL-102925 (SeattleGO) and RC2 HL-102926 (BroadGO). The exome sequencing was performed through NHLBI grants RC2 HL-102925 (BroadGO) and RC2 HL-102926 (SeattleGO). Filtered sets of annotated variants and their genotypes and phenotypes from a large subset of individuals are also available via 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).
32. Acknowledgments: We acknowledge the support of the National Heart, Lung, and Blood Institute (NHLBI) and the contributions of the research institutions, study investigators, field staff and study participants in creating this resource for biomedical research; and the Population Genetics Project Team. Funding for GO ESP was provided by NHLBI grants RC2 HL-103010 (HeartGO), RC2 HL-102925 (BroadGO) and RC2 HL-102926 (SeattleGO). Filtered sets of annotated variants and their genotypes and phenotypes from a large subset of individuals are also available via 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),