Analysis of protein-coding genetic variation in 60,706 humans

Exome Aggregation Consortium[#], Monkol Lek[1,2,3,4], Konrad J Karczewski[1,2*], Eric V Minikel[1,2,5*], Kaitlin E Samocha[1,2,6,5*], Eric Banks[2], Timothy Fennell[2], Anne H O'Donnell-Luria[1,2,7], James S Ware[2,8,9,10,11], Andrew J Hill[1,2,12], Beryl B Cummings[1,2,5], Taru Tukiainen[1,2], Daniel P Birnbaum[2], Jack A Kosmicki[1,2,6,13], Laramie Duncan[1,2], Karol Estrada[1,2], Fengmei Zhao[1,2], James Zou[2], Emma Pierce-Hoffman[1,2], David N Cooper[14], Mark DePristo[15], Ron Do[16,17,18,19], Jason Flannick[2,20], Menachem Fromer[1,6,21,16,17], Laura Gauthier[15], Jackie Goldstein[1,2,6], Namrata Gupta[2], Daniel Howrigan[1,2,6], Adam Kiezun[15], Mitja I Kurki[2,22], Ami Levy Moonshine[15], Pradeep Natarajan[2,23,24,25], Lorena Orozco[26], Gina M Peloso[2,24,25], Ryan Poplin[15], Manuel A Rivas[2], Valentin Ruano-Rubio[15], Douglas M Ruderfer[21,16,17], Khalid Shakir[15], Peter D Stenson[14], Christine Stevens[2], Brett P Thomas[1,2], Grace Tiao[15], Maria T Tusie-Luna[27], Ben Weisburd[2], Hong-Hee Won[2,23,24,25], Dongmei Yu[22,28], David M Altshuler[2,29], Diego Ardissino[30], Michael Boehnke[31], John Danesh[32], Roberto Elosua[33], Jose C Florez[2,23,24], Stacey B Gabriel[2], Gad Getz[15,23,34], Christina M Hultman[35], Sekar Kathiresan[2,23,24,25], Markku Laakso[36], Steven McCarroll[6,8], Mark I McCarthy[37,38,39], Dermot McGovern[40], Ruth McPherson[41], Benjamin M Neale[1,2,6], Aarno Palotie[42], Shaun M Purcell[21,16,17], Danish Saleheen[43,44,45], Jeremiah Scharf[22,28], Pamela Sklar[21,16,17,46,47], Patrick F Sullivan[48,49], Jaakko Tuomilehto[50], Hugh C Watkins[51], James G Wilson[52], Mark J Daly[1,2,6], Daniel G MacArthur[1,2†]

[1]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
[2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[3]School of Paediatrics and Child Health, University of Sydney, Sydney, NSW, Australia
[4]Institute for Neuroscience and Muscle Research, Childrens Hospital at Westmead, Sydney, NSW, Australia
[5]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA
[6]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[7]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA
[8]Department of Genetics, Harvard Medical School, Boston, MA, USA
[9]National Heart and Lung Institute, Imperial College London, London, UK
[10]NIHR Royal Brompton Cardiovascular Biomedical Research Unit, Royal Brompton Hospital, London, UK
[11]MRC Clinical Sciences Centre, Imperial College London, London, UK
[12]Genome Sciences, University of Washington, Seattle, WA, USA
[13]Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA, USA
[14]Institute of Medical Genetics, Cardiff University, Cardiff, UK
[15]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[16]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[17]Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[18]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[19]The Center for Statistical Genetics, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[20]Department of Molecular Biology, Massachusetts General Hospital, Boston, MA, USA
[21]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[22]Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
[23]Harvard Medical School, Boston, MA, USA
[24]Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA
[25]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA
[26]Immunogenomics and Metabolic Disease Laboratory, Instituto Nacional de Medicina Genomica, Mexico City, Mexico
[27]Department of Endocrinology and Metabolism, Instituto Nacional de Ciencias Medicas y Nutricion, Mexico City, Mexico
[28]Department of Neurology, Brigham and Women's Hospital, Boston, MA, USA
[29]Vertex Pharmaceuticals, Boston, MA, USA
[30]Department of Cardiology, University Hospital, Parma, Italy
[31]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA
[32]Department of Public Health and Primary Care, Strangeways Research Laboratory, Cambridge, UK
[33]Cardiovascular Epidemiology and Genetics, Hospital del Mar Medical Research Institute, Barcelona, Spain
[34]Department of Pathology and Cancer Center, Massachusetts General Hospital, Boston, MA, USA
[35]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden
[36]Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland
[37]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
[38]Oxford Centre for Diabetes,Endocrinology and Metabolism, University of Oxford, Oxford, UK
[39]Oxford NIHR Biomedical Research Centre, Oxford University Hospitals Foundation Trust, Oxford, UK
[40]Inflammatory Bowel Disease and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA
[41]Atherogenomics Laboratory, University of Ottawa Heart Institute, Ottawa, ON, Canada
[42]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
[43]Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA
[44]Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA
[45]Center for Non-Communicable Diseases, Karachi, , Pakistan
[46]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[47]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[48]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA
[49]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[50]Department of Public Health, University of Helsinki, Helsinki, Finland
[51]Radcliffe Department of Medicine, University of Oxford, Oxford, UK

[52]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA

[*] These authors contributed equally to this work and names appear in alphabetical order
[†] Corresponding author
[#] List of collaborators to appear in Supplementary

Summary

Large-scale reference data sets of human genetic variation are critical for the medical and functional interpretation of DNA sequence changes. Here we describe the aggregation and analysis of high-quality exome (protein-coding region) sequence data for 60,706 individuals of diverse ethnicities. The resulting catalogue of human genetic diversity has unprecedented resolution, with an average of one variant every eight bases of coding sequence and the presence of widespread mutational recurrence. The deep catalogue of variation provided by the Exome Aggregation Consortium (ExAC) can be used to calculate objective metrics of pathogenicity for sequence variants, and to identify genes subject to strong selection against various classes of mutation; we identify 3,230 genes with near-complete depletion of truncating variants, 79% of which have no currently established human disease phenotype. Finally, we show that these data can be used for the efficient filtering of candidate disease-causing variants, and for the discovery of human "knockout" variants in protein-coding genes.

Background

Over the last five years, the widespread availability of high-throughput DNA sequencing technologies has permitted the sequencing of the whole genomes or exomes (the protein-coding regions of genomes) of over half a million humans. In theory, these data represent a powerful source of information about the global patterns of human genetic variation, but in practice, are difficult to access for practical, logistical, and ethical reasons; in addition, the inconsistent processing complicates variant-calling pipelines used by different groups. Current publicly available datasets of human DNA sequence variation contain only a small fraction of all sequenced samples: the Exome Variant Server, created as part of the NHLBI Exome Sequencing Project (ESP)[1], contains frequency information spanning 6,503 exomes; and the 1000 Genomes (1000G) Project, which includes individual-level genotype data from whole-genome and exome sequence data for 2,504 individuals[2].

Databases of genetic variation are important for our understanding of human population history and biology[1–5], but also provide critical resources for the clinical interpretation of variants observed in patients suffering from rare Mendelian diseases[6,7]. The filtering of candidate variants by frequency in unselected individuals is a key step in any pipeline for the discovery of causal variants in Mendelian disease patients, and the efficacy of such

filtering depends on both the size and the ancestral diversity of the available reference data.

Here, we describe the joint variant calling and analysis of high-quality variant calls across 60,706 human exomes, assembled by the Exome Aggregation Consortium (ExAC; exac.broadinstitute.org). This call set exceeds previously available exome-wide variant databases by nearly an order of magnitude, providing unprecedented resolution for the analysis of very low-frequency genetic variants. We demonstrate the application of this data set to the analysis of patterns of genetic variation including the discovery of widespread mutational recurrence, the inference of gene-level constraint against truncating variation, the clinical interpretation of variation in Mendelian disease genes, and the discovery of human "knockout" variants in protein-coding genes.

Variant discovery and quality control

Details of the variant calling process are provided in Supplementary Information. Briefly, we assembled approximately 1 petabyte of raw sequencing data (FASTQ files) from 91,796 individual exomes drawn from a wide range of primarily disease-focused consortia (Supplementary Information Table 2). We processed these exomes through a single informatic pipeline and performed joint variant calling of single nucleotide variants (SNVs) and short insertions and deletions (indels) across all samples using a new version of the Genome Analysis Toolkit (GATK) HaplotypeCaller pipeline [Supplementary Information Section 1.3; Banks *et al.*, in preparation]. At each site, sequence information from all individuals was used to assess the evidence for the presence of a variant in each individual. We also performed systematic analysis of copy number variation across these individuals [Ruderfer et al., to be co-submitted].

We leveraged a variety of sources of internal and external validation data to calibrate filters and evaluate the quality of filtered variants (Supplementary Information Table 6). We adjusted the standard GATK variant site filtering[8] to increase the number of singleton variants that pass this filter, while maintaining a singleton transmission rate of 50.49%, very near the expected 50%, within sequenced trios. We then used the remaining passing variants to assess depth and genotype quality filters compared to >10,000 samples that had been directly genotyped using SNP arrays (Illumina HumanExome) and achieved 97-99% heterozygous concordance, consistent with known error rates for

rare variants in chip-based genotyping[9]. Relative to a "platinum standard" genome sequenced using five different technologies[10], we achieved sensitivity of 99.8% and false discovery rates of 0.13% for single nucleotide variants (SNVs), and corresponding rates of 95.1% and 1.95% for insertions and deletions (indels).

In order to generate allele frequencies based on independent observations without enrichment of Mendelian disease alleles, we restricted the final release data set to unrelated adults with high-quality sequence data and without severe pediatric disease; full details of the filtering process are described in the Supplementary Information. After filtering, the final ExAC data set comprises 60,706 individuals (Figure 1a). To identify the ancestry of each ExAC sample, we performed principal component analysis (PCA) on 5,400 common SNVs with high coverage across all of the exome capture technologies[11] represented in ExAC. This PCA allows us to distinguish the major axes of geographic ancestry within the ExAC sample, and to identify population clusters corresponding to individuals of European, African, South Asian, East Asian, and admixed American (hereafter Latino) ancestry (Figure 1b; Supplementary Information Table 3). We further separated Europeans into individuals of Finnish and non-Finnish ancestry given the enrichment of this bottlenecked population; the term "European" hereafter refers to non-Finnish European individuals.

We identified 10,195,872 candidate sequence variants in ExAC. We further applied stringent depth and site/genotype quality filters to define a subset of 7,404,909 high quality (HQ) variants, including 317,381 indels (Supplementary Information Table 6), corresponding to one variant for every 8 bp within the exome calling intervals. The majority of these are very low-frequency variants absent from previous smaller call sets (Figure 1c): of the HQ variants, 99% have a frequency of <1%, 54% are singletons (variants seen only once in the data set), and 72% are absent from both 1000G and ESP.

However, the density of variation in ExAC is not uniform across the genome, and the observation of variants depends on factors such as mutational properties and selective pressures. In the ~45M well covered (80% of individuals with a minimum of 10X coverage) positions in ExAC, there are ~18M possible synonymous variants, of which we observe 1.4M (7.5%). However, we observe 313K out of 499K (62.8%) of possible

CpG transitions (C to T variants, where the following base is G), while only observing 294K out of 9.3M transversions (3.1%) and 802K out of 8.9M other transitions (9%). A similar pattern is observed for missense and nonsense variants, with lower proportions due to selective pressures (Figure 1D). Of 123,629 HQ indels called in coding exons, 117,242 (95%) have length -6 to +6, with shorter deletions being the most common (Figure 1E). Frameshifts are found in smaller numbers and are more likely to be singletons than in-frame indels (Figure 1F), reflecting the influence of purifying selection.

Patterns of protein-coding variation revealed by large samples

The unprecedented density of protein-coding sequence variation in ExAC reveals a number of properties of human genetic variation undetectable in smaller data sets. For instance, 7.9% of HQ sites in ExAC are multiallelic (multiple different sequence variants observed at the same site), close to the Poisson expectation of 8.3% given the observed density of variation, and far higher than observed in previous data sets - 0.48% in 1000 Genomes (ExAC calling intervals) and 0.43% in ESP.

The size of ExAC also makes it possible to directly observe mutational recurrence: instances in which the same mutation has occurred multiple times independently throughout the history of the sequenced populations. For instance, among synonymous variants, a class of variation expected to have undergone minimal selection, 43% of validated *de novo* events identified in external datasets of 1,756 parent-offspring trios are also observed independently in our dataset (Figure 2a), indicating a separate origin for the same variant within the demographic history of the two samples. This proportion is much higher for transition variants at CpG sites, well established to be the most highly mutable sites in the human genome[12]: 87% of previously reported *de novo* CpG transitions at synonymous sites are observed in ExAC, indicating that our sample sizes are beginning to approach saturation of this class of variation. This saturation is detectable by a change in the discovery rate at subsets of the ExAC data set, beginning at around 20,000 individuals (Figure 2b), indicating that ExAC is the first human dataset large enough for this effect to be directly observed.

Mutational recurrence has a marked effect on the frequency spectrum in the ExAC data, resulting in a depletion of singletons at sites with high mutation rates (Figure 2c). We

observe a correlation between singleton rates (the proportion of variants seen only once in ExAC) and site mutability inferred from sequence context[13] ($r = -0.98$; $p < 10^{-50}$; Extended Data Figure 4d): sites with low predicted mutability have a singleton rate of 60%, compared to 20% for sites with the highest predicted rate (CpG transitions; Figure 2C). Conversely, for synonymous variants, CpG variants are approximately twice as likely to rise to intermediate frequencies: 16% of CpG variants are found in at least 20 copies in ExAC, compared to 8% of transversions and non-CpG transitions, suggesting that synonymous CpG transitions have on average two independent mutational origins in the ExAC sample. Recurrence at highly mutable sites can further be observed by examining the population sharing of doubleton synonymous variants (variants occurring in only two individuals in ExAC). For low-mutability mutations (especially transversions), these variants are more likely to be observed in a single population (representing a single mutational origin), while CpG transitions are more likely to be found in two separate populations (representing independent mutational events); as such, site mutability and probability of observation in two populations is significantly correlated ($r = 0.884$; Figure 2d).

We also explored the prevalence and functional impact of multinucleotide polymorphisms (MNPs), clusters of base substitutions on the same haplotype. We used a read-based phasing pipeline to assess all sites where multiple substitutions were observed within the same codon in at least one individual. We found 5,945 MNPs in ExAC (Extended Data Figure 3a), with an average of 23 per sample, where analysis of the underlying SNPs without correct haplotype phasing would result in altered interpretation. These include 647 instances where the effect of a protein-truncating variant (PTV) variant is eliminated by an adjacent SNP (rescued PTV) and 131 instances where underlying synonymous or missense variants result in PTV MNPs (gained PTV). We identified rescued PTV MNPs in *MLH1* (where PTV variants are associated with autosomal dominant Lynch syndrome), and *FANCA* (autosomal recessive Fanconi anemia)[14,15]. We also identified an ExAC sample as a carrier for a gained PTV MNP in *MUTYH*, where homozygous PTV variants are known to cause *MUTYH*-associated polyposis[16]. Additionally our analysis revealed 10 MNPs that have previously been reported as disease causing mutations in HGMD (Supplementary Information Table 9), including a stop-gained MNP in *COH1* which has previously been identified as a

recessive cause of Cohen syndrome[17]. We note that these variants would be missed by virtually all currently available variant calling and annotation pipelines.

Inferring variant deleteriousness and gene constraint

Deleterious variants are expected to have lower allele frequencies than neutral ones, due to negative selection. This theoretical property has been demonstrated previously in human population sequencing data[18,19] and here (Figure 1d, Figure 1e). This allows inference of the degree of natural selection against specific functional classes of variation: however, mutational recurrence as described above indicates that allele frequencies observed in ExAC-scale samples are also skewed by mutation rate, with more mutable sites less likely to be singletons (Figure 2c and Extended Data Figure 4d). Mutation rate is in turn non-uniformly distributed across functional classes - for instance, stop lost mutations can never occur at CpG dinucleotides (Extended Data Figure 4e). We corrected for mutation rates (Supplementary Information) by creating a mutability-adjusted proportion singleton (MAPS) metric. This metric reflects (as expected) strong selection against predicted PTVs, as well as missense variants predicted by conservation-based methods to be deleterious (Figure 2e).

The deep ascertainment of rare variation in ExAC also allows us to infer the extent of selection against variant categories on a per-gene basis by examining the proportion of variation that is missing compared to expectations under random mutation. Conceptually similar approaches have been applied to smaller exome datasets[13,20] but have been underpowered, particularly for the analysis of depletion of PTVs. We compared the observed number of rare (MAF <0.1%) variants per gene to an expected number derived from a selection neutral, sequence-context based mutational model[13]. The model performs extremely well in predicting the number of synonymous variants, which should be under minimal purifying selection, per gene ($r = 0.98$; Extended Data Figure 5).

We quantified deviation from expectation with a Z score[13], which for synonymous variants is centered at zero, but has distributions that are significantly shifted towards higher values (greater constraint) for both missense and PTVs (Wilcoxon $p < 10^{-50}$ for both). To reduce confounding by coding sequence length for PTVs, we also developed an expectation-maximization algorithm (Supplementary Information Section 4.4) using the observed and expected PTV counts within each gene to separate genes into three

general categories: null (observed ≈ expected), recessive (observed ≤50% of expected), and haploinsufficient (observed <10% of expected). This metric – the probability of being loss-of-function (LoF) intolerant (pLI) – separates genes of sufficient length into LoF intolerant (pLI ≥0.9, n=3,230) or LoF tolerant (pLI ≤0.1, n=10,374) categories. pLI is less correlated with coding sequence length (r = 0.17 as compared to 0.57 for the PTV Z score), outperforms the PTV Z score as an intolerance metric (Supplementary Information Table 11), and reveals the expected contrast between gene lists (Figure 3b). pLI is also positively correlated with a gene product's number of physical interaction partners ($p < 10^{-41}$). The most constrained pathways (highest median pLI for the genes in the pathway) are core biological processes (spliceosome, ribosome, and proteasome components; KS test $p < 10^{-6}$ for all) while olfactory receptors are among the least constrained pathways (KS test $p < 10^{-16}$), demonstrated in Figure 3b and consistent with previous work[5,21–23].

Critically, we note that LoF-intolerant genes include virtually all known severe haploinsufficient human disease genes (Figure 3b), but that 79% of LoF-intolerant genes have not yet been assigned a human disease phenotype despite the clear evidence for extreme selective constraint (Supplementary Information 4.11). These likely represent either undiscovered severe dominant disease genes, or genes in which loss of a single copy results in embryonic lethality.

The most highly constrained missense (top 25% missense Z scores) and PTV (pLI ≥0.9) genes show higher expression levels and broader tissue expression than the least constrained genes[24] (Figure 3c). These most highly constrained genes are also depleted for eQTLs ($p < 10^{-9}$ for missense and PTV; Figure 3d), yet are enriched within genome-wide significant trait-associated loci (χ^2 $p < 10^{-14}$, Figure 3e). Intuitively, genes intolerant of PTV variation are dosage sensitive: natural selection does not tolerate a 50% deficit in expression due to the loss of single allele. It is therefore unsurprising that these genes are also depleted of common genetic variants that have a large enough effect on expression to be detected as eQTLs with current limited sample sizes. However, smaller changes in the expression of these genes, through weaker eQTLs or functional variants, are more likely to contribute to medically relevant phenotypes. Therefore, highly constrained genes are dosage-sensitive, expressed more broadly across tissues (as expected for core cellular processes), and are enriched for medically relevant variation.

Finally, we investigated how these constraint metrics would stratify mutational classes according to their frequency spectrum, corrected for mutability as in the previous section (Figure 3f). The effect was most dramatic when considering stop-gained variants in the LoF-intolerant set of genes. For missense variants, the missense Z score offers information additional to Polyphen2 and CADD classifications, indicating that gene-level measures of constraint offer additional information to variant-level metrics in assessing potential pathogenicity.

ExAC improves variant interpretation in Mendelian disease

We assessed the value of ExAC as a reference dataset for clinical sequencing approaches, which typically prioritize or filter potentially deleterious variants based on functional consequence and allele frequency[6]. To simulate a Mendelian variant analysis, we filtered variants in 100 ExAC exomes per continental population against ESP (the previous default reference data set for clinical analysis) or the remainder of ExAC, removing variants present at ≥0.1% allele frequency, a filter recommended for dominant disease variant discovery[6]. Filtering on ExAC reduced the number of candidate protein-altering variants by 7-fold compared to ESP, and was most powerful when the highest allele frequency in any one population ("popmax") was used rather than average ("global") allele frequency (Figure 4a). ESP is not well-powered to filter at 0.1% AF without removing many genuinely rare variants, as AF estimates based on low allele counts are both upward-biased and imprecise (Figure 4b). We thus expect that ExAC will provide a very substantial boost in the power and accuracy of variant filtering in Mendelian disease projects.

Previous large-scale sequencing studies have repeatedly shown that some purported Mendelian disease-causing genetic variants are implausibly common in the population[25–27]. The average ExAC participant harbors ~53 variants reported as disease-causing in two widely-used databases of disease-causing variants (Supplementary Information Section 5.2). Most (~41) of these are high-quality genotypes but with implausibly high (>1%) AF in at least one population. We therefore hypothesized that most of the supposed burden of Mendelian disease alleles per person is due not to genotyping error, but rather to misclassification in databases.

We manually curated the evidence of pathogenicity for 192 previously reported pathogenic variants with allele frequency >1% either globally or in South Asian or Latino individuals, populations that are underrepresented in previous reference databases. Nine variants had sufficient data to support disease association, typically with either mild or incompletely penetrant disease effects; the remainder either had insufficient evidence for pathogenicity, no claim of pathogenicity, or were benign traits (Supplementary Information Section 5.3). 163 were reclassified as benign or likely benign following American College of Medical Genetics guidelines[28]. Supporting functional data were reported for 18 of these variants, highlighting the need to review cautiously even variants with experimental support.

We also sought phenotypic data for a subset of ExAC participants homozygous for reported severe recessive disease variants, again enabling reclassification of some variants as benign. North American Indian Childhood Cirrhosis is a recessive disease of cirrhotic liver failure during childhood requiring liver transplant for survival to adulthood, previously reported to be caused by *CIRH1A* p.R565W[29]. ExAC contains 222 heterozygous and 4 homozygous Latin American individuals, with a population allele frequency of 1.92%. The 4 homozygotes had no history of liver disease according to available phenotype data, and recontact with additional phenotyping in two individuals revealed normal liver function (Supplementary Information Table 15). Thus, despite the rigorous linkage and Sanger sequencing efforts that led to the original report of pathogenicity, the ExAC data demonstrate that this variant is not a fully penetrant cause of severe disease, a reminder of the importance of well-matched reference populations.

The above curation efforts confirm the importance of allele frequency filtering in analysis of candidate disease variants. However, literature and database errors are prevalent even at lower allele frequencies: the average ExAC exome contains 0.89 reportedly Mendelian variants in well-characterized dominant disease genes[30] at <1% popmax AF and 0.20 at <0.1% popmax AF. This inflation likely results from a combination of false reports of pathogenicity and incomplete penetrance, as we show for *PRNP* in the accompanying work [Minikel et al, submitted]. The abundance of rare functional variation in many disease genes in ExAC is a reminder that such variants should not be assumed to be causal or highly penetrant without careful segregation or case-control analysis[28,7].

Impact of rare protein-truncating variants

We investigated the distribution of PTVs, variants predicted to disrupt protein-coding genes through the introduction of a stop codon or frameshift or the disruption of an essential splice site; such variants are expected to be enriched for complete loss-of-function of the impacted genes. Naturally-occurring PTVs in humans provide a model for the functional impact of gene inactivation, and have been used to identify many genes in which LoF causes severe disease[31], as well as rare cases where LoF is protective against disease[32].

Among the 7,404,909 HQ variants in ExAC, we found 179,774 high-confidence PTVs (as defined in Supplementary Information Section 6), 121,309 of which are singletons. This corresponds to an average of 85 heterozygous and 35 homozygous PTVs per individual (Figure 5a). The diverse nature of the cohort enables the discovery of substantial numbers of novel PTVs: out of 58,435 PTVs with an allele count greater than one, 33,625 occur in only one population. However, while PTVs as a category are extremely rare, the majority of the PTVs found in any one person are common, and each individual has only ~2 singleton PTVs, of which 0.14 are found in PTV-constrained genes (pLI >0.9). The site frequency spectrum of these variants across the populations represented in ExAC recapitulates known aspects of demographic models, including an increase in intermediate-frequency (1%-5%) PTVs in Finland[33] and relatively common (>0.1%) PTVs in Africans (Figure 5b).

Using a sub-sampling approach, we show that the discovery of both heterozygous (Figure 5c) and homozygous (Figure 5d) PTVs scales very differently across human populations, with implications for the design of large-scale sequencing studies for the ascertainment of human “knockouts” described below.

Discussion

Here we describe the generation and analysis of the most comprehensive catalogue of human protein-coding genetic variation to date, incorporating high-quality exome sequencing data from 60,706 individuals of diverse geographic ancestry. The resulting call set provides unprecedented resolution for the analysis of very low-frequency protein-coding variants in human populations, as well as a powerful resource for the clinical interpretation of genetic variants observed in disease patients. The complete frequency

and annotation data from this call-set has been made freely available through a public website [exac.broadinstitute.org].

The very large sample size of ExAC also provides opportunities for a high-resolution analysis of the sensitivity of human genes to functional variation. While previous sample sizes have been adequately powered for the assessment of gene-level intolerance to missense variation[13,20], ExAC provides for the first time sufficient power to investigate genic intolerance to PTVs, highlighting 2,557 LoF-intolerant genes for which human disease phenotypes have not yet been identified. These unassociated genes are likely to fall into two main categories: those that cause severe haploinsufficient disease that has not yet been genetically characterized, and those where heterozygous inactivation results in embryonic lethality. In the accompanying work [Ruderfer et al., to be co-submitted] we show that ExAC similarly provides power to identify genes intolerant of copy number variation. Quantification of genic intolerance to both classes of variation will provided added power to disease studies.

The ExAC resource provides the largest database to date for the estimation of allele frequency for protein-coding genetic variants, providing a powerful filter for analysis of candidate pathogenic variants in severe Mendelian diseases. Frequency data from ESP[1] have been widely used for this purpose, but those data are limited by population diversity and by resolution at allele frequencies ≤0.1%. ExAC therefore provides substantially improved power for Mendelian analyses, although it is still limited in power at lower allele frequencies, emphasizing the need for more sophisticated pathogenic variant filtering strategies alongside on-going data aggregation efforts. ExAC also highlights an unexpected tolerance of many disease genes to functional variation, and reveals that the literature and public databases contain an inflated number of reportedly pathogenic variants across the frequency spectrum, indicating a need for stringent criteria for assertions of pathogenicity.

Finally, we show that different populations confer different advantages in the discovery of gene-disrupting PTVs, providing guidance for projects seeking to identify human “knockouts” to understand gene function. Individuals of African ancestry have more PTVs (140 on average), with this enrichment most pronounced at allele frequencies above 1% (Figure 5b). Finnish individuals, as a result of a population bottleneck, are

depleted at the lowest (<0.1%) allele frequencies but have a peak in frequency at 1-5% (Figure 5b). However, these differences are diminished when considering only LoF-constrained (pLI > 0.9) genes (Extended Data Figure 10). Sampling multiple populations would likely be a fruitful strategy for a researcher investigating common PTV variation. However, discovery of homozygous PTVs is markedly enhanced in the South Asian samples, which come primarily from a Pakistani cohort with 38.3% of individuals self-reporting as having closely related parents, emphasizing the extreme value of consanguineous cohorts for "human knockout" discovery (Figure 5d) [Saleheen *et al.*, to be co-submitted].

Even with this unprecedented collection of jointly processed exomes, many limitations remain. First, most ExAC individuals were ascertained for the presence or absence of a biomedically important disease and thus are not a random sampling of the population. To minimize this bias, we have made every effort to exclude severe pediatric diseases, but nevertheless, the inclusion of both cases and controls for several polygenic disorders means that ExAC may contain higher counts of disease-associated variants for certain phenotypes[34]. Second, future reference databases would benefit from including a broader sampling of human diversity, particularly from the Middle East, a population not represented in the present dataset. Third, some protein-coding exons lack coverage in ExAC entirely or have coverage levels confounded with exome capture technology, which in turn is confounded with cohort and continental population. Fourth, protein-coding exons are only one source of functionally important variation, and the future inclusion of whole genomes will also be critical to enable ascertainment of additional classes of variation and the identification of constrained regions outside of protein-coding sequence.

While the ExAC dataset dramatically exceeds the scale of previously available frequency reference datasets, much remains to be gained by further increases in sample size. Indeed, the fact that even the rarest transversions have mutational rates[13] on the order of 1×10^{-9} implies that almost all possible non-lethal SNVs likely exist in some person on Earth. ExAC already includes >70% of all possible protein-coding CpG transitions at well-covered sites; order of magnitude increases in sample size will eventually lead to saturation of other classes of variation.

ExAC was made possible by the willingness of multiple large disease-focused consortia to share their raw data, and by the availability of the software and computational resources required to create a harmonized variant call set on the scale of tens of thousands of samples. The creation of yet larger reference variant databases will require continued emphasis on the value of public data sharing.

References

1. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493,** 216–20 (2013).

2. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

3. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–496 (2011).

4. Stoneking, M. & Krause, J. Learning about human population history from ancient and modern genomes. *Nat. Rev. Genet.* **12,** 603–614 (2011).

5. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335,** 823–8 (2012).

6. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12,** 745–55 (2011).

7. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508,** 469–476 (2014).

8. DePristo, M. a *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

9. Voight, B. F. *et al.* The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLoS Genet.* **8,** e1002793 (2012).

10. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32,** 246–251 (2014).

11. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506,** 185–90 (2014).

12. Cooper, D. N. & Youssoufian, H. The CpG dinucleotide and human genetic disease. *Hum. Genet.* **78,** 151–155 (1988).

13. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* (2014). doi:10.1038/ng.3050

14. Bronner, C. E. *et al.* Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. *Nature* **368,** 258–261 (1994).

15. Levran, O. *et al.* Spectrum of sequence variations in the FANCA gene: An International Fanconi Anemia Registry (IFAR) study. *Hum. Mutat.* **25,** 142–149 (2005).

16. Tenesa, a *et al.* Association of MUTYH and colorectal cancer. *Br. J. Cancer* **95,** 239–42 (2006).

17. El Chehadeh-Djebbar, S. *et al.* Changing facial phenotype in Cohen syndrome: towards clues for an earlier diagnosis. *Eur. J. Hum. Genet.* **21,** 736–42 (2013).

18. Tennessen, J. a *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–9 (2012).

19. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47,** 435–444 (2015).

20. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9,** e1003709 (2013).

21. Jeong, H., Mason, S. P., Barabási, a L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411,** 41–42 (2001).

22. Goh, K.-I. *et al.* The human disease network. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 8685–8690 (2007).

23. Rolland, T. *et al.* Resource A Proteome-Scale Map of the Human Interactome Network. *Cell* **159,** 1212–1226 (2014).

24. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–60 (2015).

25. Bell, C. J. *et al.* Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci. Transl. Med.* **3,** 65ra4 (2011).

26. Xue, Y. *et al.* Deleterious- and disease-allele prevalence in healthy individuals: Insights from current predictions, mutation databases, and population-scale resequencing. *Am. J. Hum. Genet.* **91,** 1022–1032 (2012).

27. Piton, A., Redin, C. & Mandel, J.-L. XLID-Causing Mutations and Associated Genes Challenged in Light of Data From Large-Scale Human Exome Sequencing. *Am. J. Hum. Genet.* **93,** 368–383 (2013).

28. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17,** 405–423 (2015).

29. Chagnon, P. *et al.* A missense mutation (R565W) in cirhin (FLJ14728) in North American Indian childhood cirrhosis. *Am. J. Hum. Genet.* **71,** 1443–9 (2002).

30. Blekhman, R. *et al.* Natural Selection on Genes that Underlie Human Disease Susceptibility. *Curr. Biol.* **18,** 883–889 (2008).

31. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* 1–17 (2015). doi:10.1016/j.ajhg.2015.06.009

32. Kathiresan, S. Developing Medicines That Mimic the Natural Successes of the Human Genome. *J. Am. Coll. Cardiol.* **65,** 1562–1566 (2015).

33. Lim, E. T. *et al.* Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. *PLoS Genet.* **10,** e1004494 (2014).

34. Freischmidt, A. *et al.* Haploinsufficiency of TBK1 causes familial ALS and fronto-temporal dementia. *Nat. Neurosci.* **18,** (2015).

Acknowledgements

M.Lek is supported by the Australian National Health and Medical Research Council CJ Martin Fellowship and the Australian American Association Sir Keith Murdoch Fellowship. A.H.O is supported by Pfizer/ACMGF Clinical Genetics Fellowship. J.SW. is supported by Fondation Leducq and Wellcome Trust. A.J.H. is supported by NSF Graduate Research Fellowship. M.I.K is supported by Instrumentarium Science Foundation, Finland; Finnish Foundation for Cardiovascular Research; Orion Research Foundation and the University of Eastern Finland, Saastamoinen Foundation. P.N. is supported by John S. LaDue Memorial Fellowship in Cardiology, Harvard Medical School. G.M.P. is supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number K01HL125751. H.W. is supported by postdoctoral award from the American Heart Association (15POST23280019). R.E. is supported by Instituto Salud Carlos III-FIS-FEDER-ERDF: RD12/0042/0013, PI12/00232; Agència de Gestió Ajuts Universitaris de Recerca: 2014 SGR 240. M.I.M is supported by Wellcome Trust Senior Investigator, NIHR Senior Investigator;; EU Framework VII HEALTH-F4-2007-201413; Medical Research Council G0601261; Wellcome Trust 090532, 098381, 090367; NIH RC2-DK088389, U01-DK085545. J.M.S is supported by NINDS grants NS40024-09S1 and NS085048. P.S. is supported by NIMH grant MH095034 and MH089905. P.F.S is supported by Swedish Research Council award D0886501; NIMH grants MH077139 and MH094421; Yeargen Family; Stanley Center. D.G.M is supported by NIGMS R01 GM104371 and NIDDK U54 DK105566.

Author Contributions

M.Lek,K.J.K.,E.V.M.,K.E.S.,E.B.,T.F.,A.H.O.,J.S.W.,A.J.H.,B.B.C.,T.T.,D.P.B.,J.A.K.,L.D. ,K.E.,F.Z.,J.Z. ,M.J.D.,D.G.M. contributed to the analysis and writing of the manuscript. M.Lek, E.B.,T.F.,K.J.K.,E.V.M.,F.Z.,D.P.B.,D.N.C.,M.D.,R.D.,J.F.,M.F.,L.G.,J.G.,N.G., D.H., A.K.,M.I.K.,A.L.M.,P.N.,L.O.,G.M.P.,R.P.,M.A.R.,V.R.,D.M.R.,K.S.,P.D.S.,C.S., B.P.T.,G.T.,M.T.T.,B.W.,H.W.,D.Y. ,S.B.G.,M.J.D.,D.G.M. contributed to the production of the ExAC data set. D.M.A.,D.A.,M.B.,J.D.,R.E.,J.C.F.,S.B.G.,G.G.,C.M.H.,S.K.,M. Laakso,S.M.,M.I.M.,D.M., R.M.,B.M.N.,A.P.,S.M.P.,D.S.,J.S.,P.S.,P.F.S.,J.T.,H.C.W., J.G.W.,M.J.D.,D.G.M. contributed to the design and conduct of the various exome sequencing studies and critical review of manuscript.

Author Information

P.F.S is a scientific advisor to Pfizer.

ExAC data set is publicly available at http://www.exac.broadinstitute.org

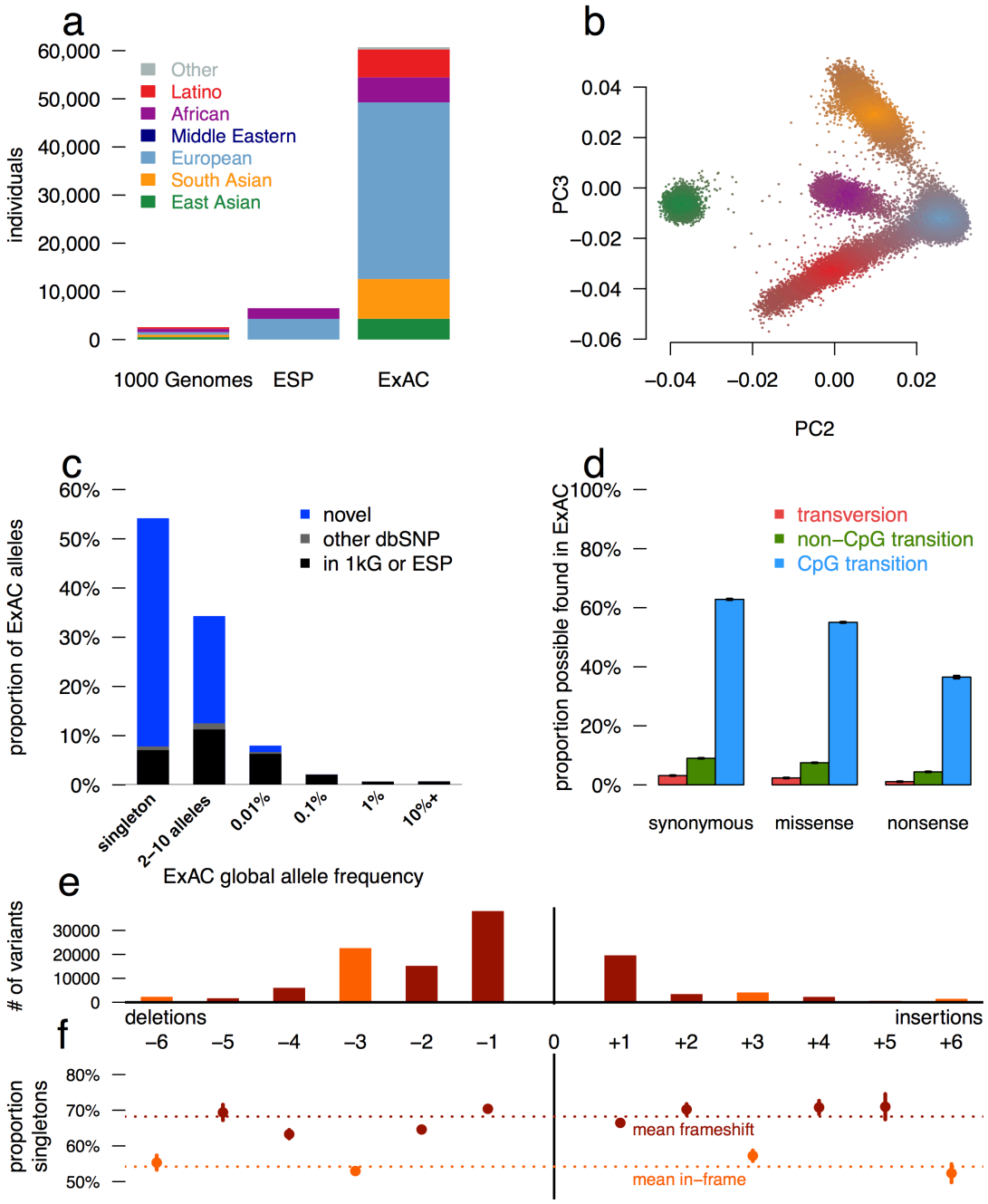

Figure 1. Patterns of genetic variation in 60,706 humans. a) The size and diversity of public reference exome datasets. ExAC exceeds previous datasets in size for all studied populations. b) Principal component analysis (PCA) dividing ExAC individuals into five continental populations. PC2 and PC3 are shown; additional PCs are in Extended Data Figure 2a. c) The allele frequency spectrum of ExAC highlights that the majority of genetic variants are rare and novel. d) The proportion of possible variation observed by mutational context and functional class. Over half of all possible CpG transitions are observed. e-f) The number (e) and frequency distribution (proportion singleton; f) of indels, by size. Compared to in-frame indels, frameshift variants are greater in number and are more common (have a lower proportion of singletons, a proxy for predicted deleteriousness on gene product).

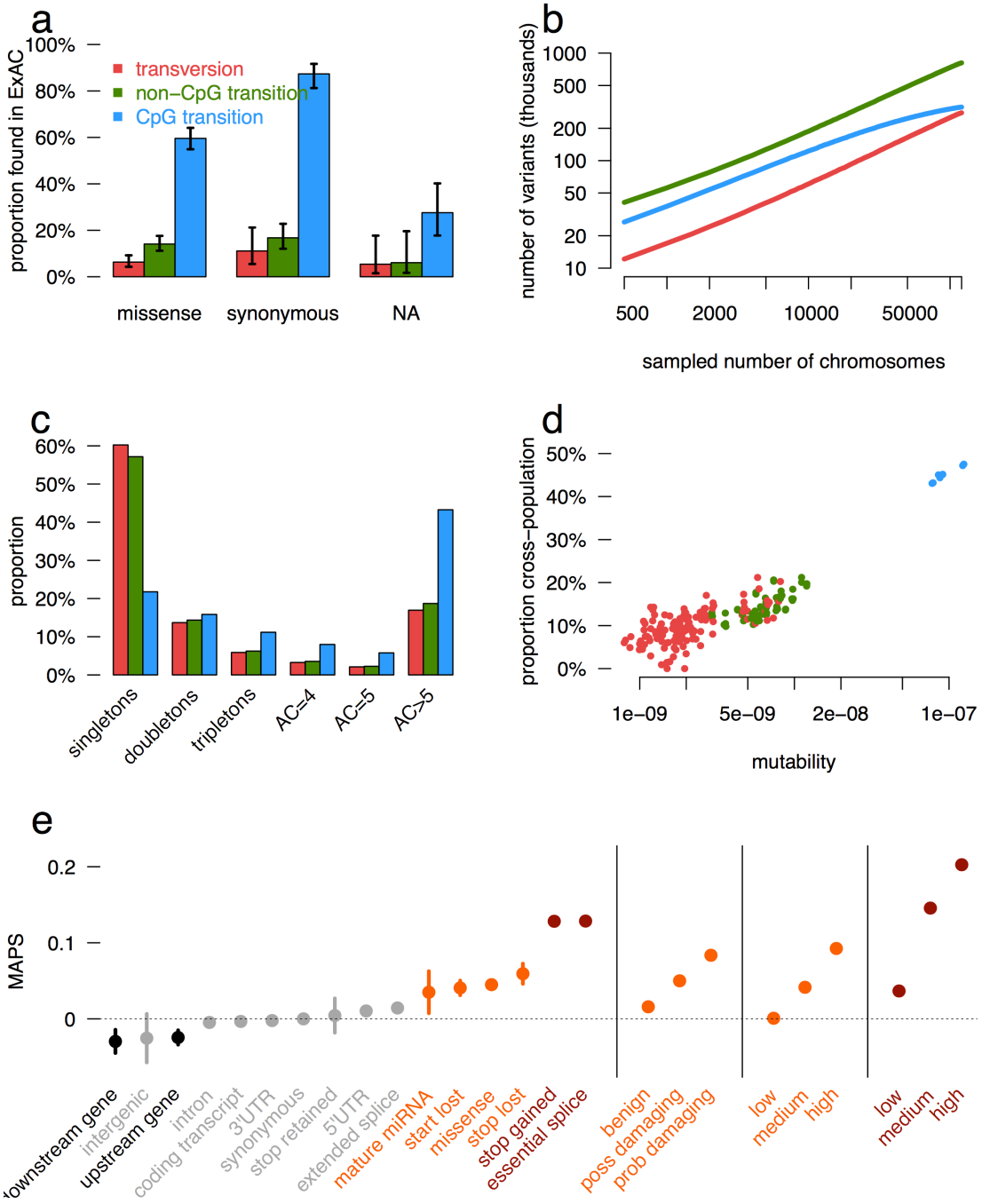

Figure 2. Mutational recurrence at large sample sizes. a) Number of unique variants observed, by mutational context, as a function of number of individuals (down-sampled from ExAC). CpG transitions, the most likely mutational event, begin reaching saturation at ~20,000 individuals. B) Proportion of validated *de novo* variants from two external datasets that are independently found in ExAC, separated by functional class and mutational context. Error bars represent standard error of the mean. Colors are consistent in a-d. c) The site frequency spectrum is shown for each mutational context. d) For doubletons (variants with an allele count of 2), mutation rate is positively correlated with the likelihood of being found in two individuals of different continental populations. e) The mutability-adjusted proportion of singletons (MAPS) is shown across functional classes. Error bars represent standard error of the mean of the proportion of singletons.

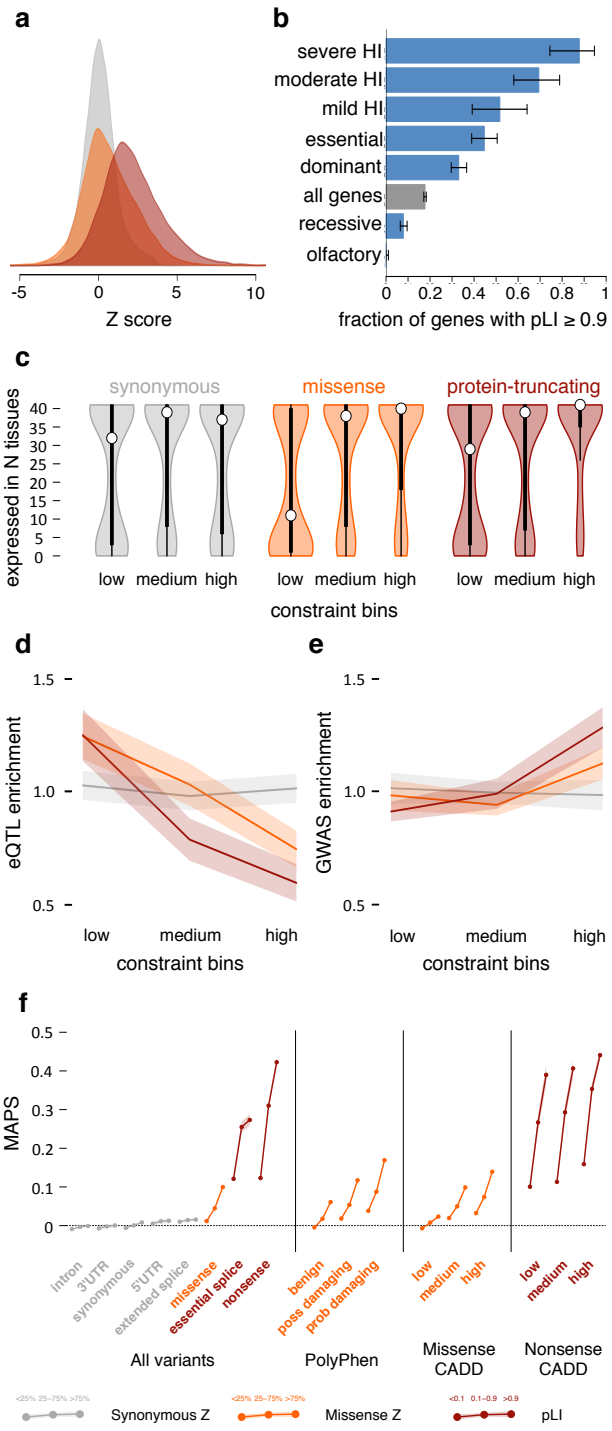

Figure 3. Quantifying intolerance to functional variation in genes and gene sets. a) Histograms of constraint Z scores [Samocha 2014] for 18,225 genes. This measure of departure of number of variants from expectation is normally distributed for synonymous variants, but right-shifted (higher constraint) for missense and protein-truncating variants (PTVs), indicating that more genes are intolerant to these classes of variation. b) The proportion of genes that are very likely intolerant of loss-of-function variation (pLI ≥ 0.9) is highest for ClinGen haploinsufficient genes, and stratifies by the severity and age of onset of the haploinsufficient phenotype. Genes essential in cell culture and dominant disease genes are likewise enriched for intolerant genes, while recessive disease genes and olfactory receptors have fewer intolerant genes. Black error bars indicate 95% confidence intervals (CI). c) Synonymous Z scores show no correlation with the number of tissues in which a gene is expressed, but the least missense- and PTV-constrained genes tend to be expressed in fewer tissues. Thick black bars indicate the first to third quartiles, with the white circle marking the median. d) Highly missense- and PTV-constrained genes are less likely to have eQTLs discovered in GTEx as the average gene. Shaded regions around the lines indicate 95% CI. e) Highly missense- and PTV-constrained genes are more likely to be adjacent to GWAS signals than the average gene. Shaded regions around the lines indicate 95% CI. f) MAPS (Figure 2d) is shown for each functional category, broken down by constraint score bins as shown. Missense and PTV constraint score bins provide information about natural selection at least partially orthogonal to MAPS, PolyPhen, and CADD scores, indicating that this metric should be useful in identifying variants associated with deleterious phenotypes. Shaded regions around the lines indicate 95% CI. For panels a,c-f: synonymous shown in gray, missense in orange, and protein-truncating in maroon.

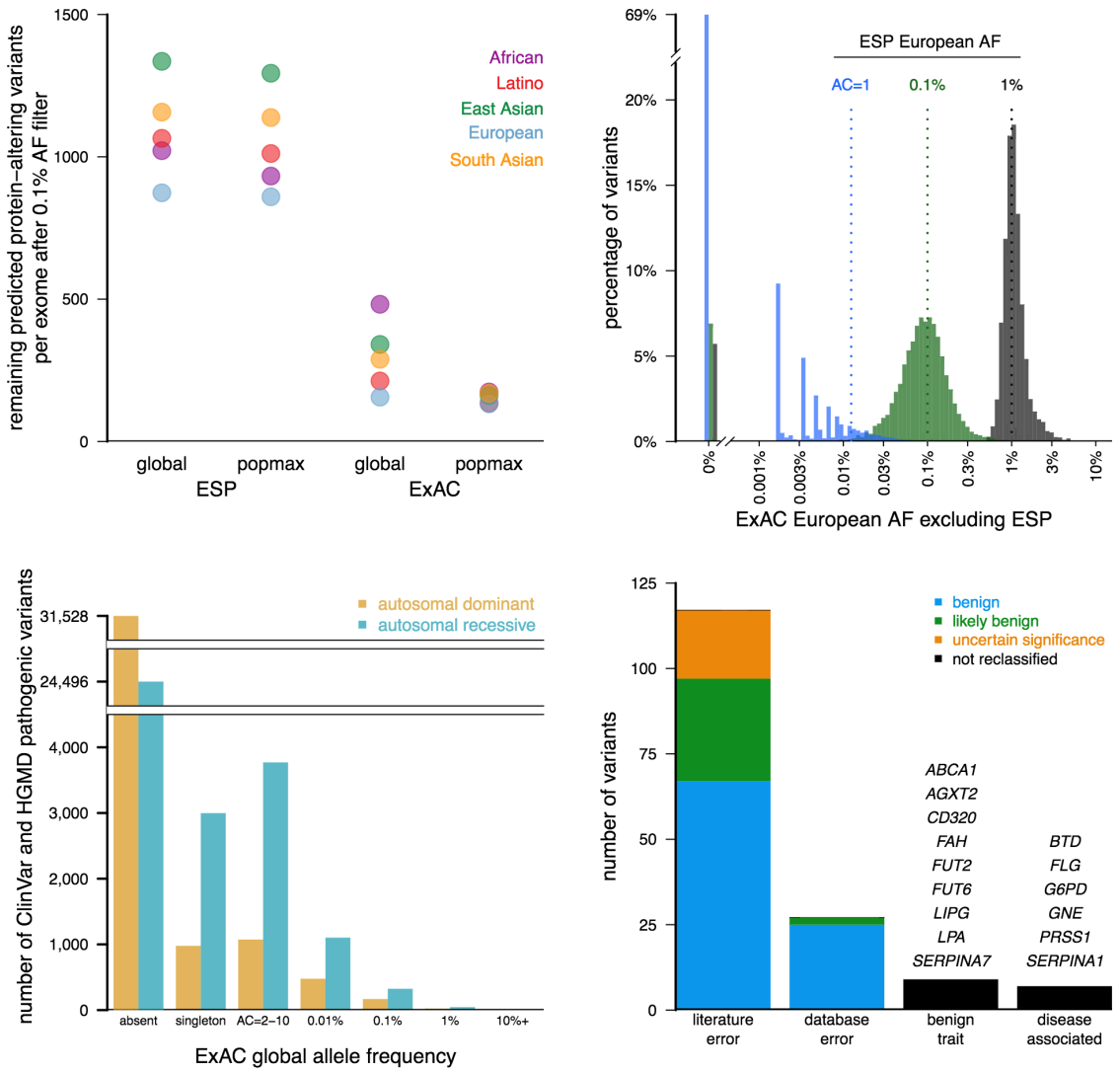

Figure 4. Filtering for Mendelian variant discovery. a) Predicted missense and protein-truncating variants in 500 randomly chosen ExAC individuals were filtered based on allele frequency information from ESP, or from the remaining ExAC individuals. At a 0.1% allele frequency (AF) filter, ExAC provides greater power to remove candidate variants, leaving an average of 154 variants for analysis, compared to 1090 after filtering against ESP. Popmax AF also provides greater power than global AF, particularly when populations are unequally sampled. b) Estimates of allele frequency in Europeans based on ESP are more precise at higher allele frequencies. Sampling variance and ascertainment bias make AF estimates unreliable, posing problems for Mendelian variant filtration. 69% of ESP European singletons are not seen a second time in ExAC (tall bar at left), illustrating the dangers of filtering on very low allele counts. c) Allele frequency spectrum of disease-causing variants in the Human Gene Mutation Database (HGMD) and/or pathogenic or likely pathogenic variants in ClinVar for well characterized autosomal dominant and autosomal recessive disease genes[30]. Most are not found in ExAC; however, many of the pathogenic variants found in ExAC are at too high a frequency to be consistent with disease prevalence and penetrance. d) Literature review of variants with >1% global allele frequency or >1% Latin American and South Asian population allele frequency confirmed there is insufficient evidence for pathogenicity for the majority of these variants. Variants were reclassified by ACMG guidelines[28].

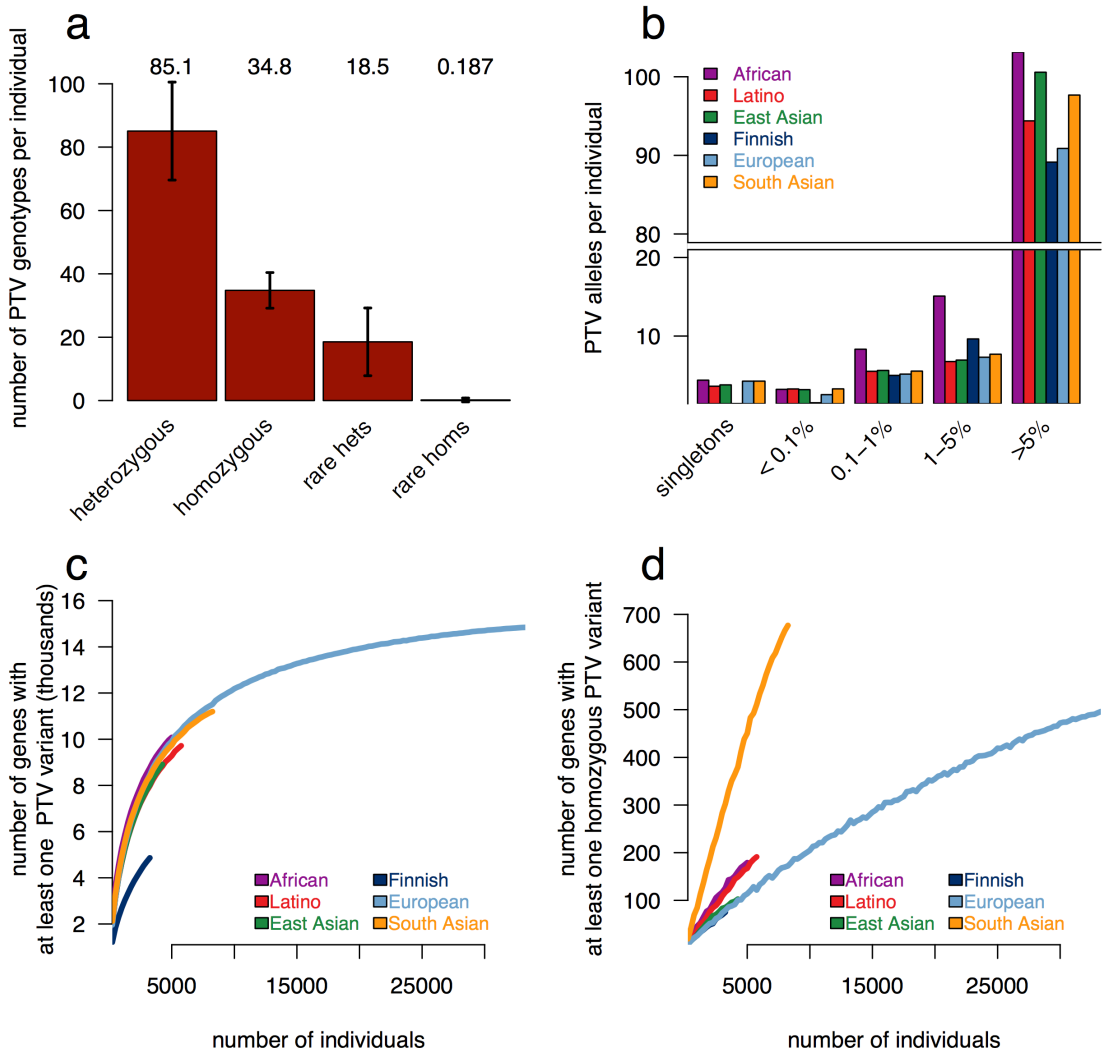

Figure 5. Protein-truncating variation in ExAC. a) The average ExAC individual has 85 heterozygous and 35 homozygous protein-truncating variants (PTVs), of which 18 and 0.19 are rare (<0.1% popmax AF), respectively. Error bars represent standard deviation. b) Breakdown of PTVs per individual (a) by popmax AF bin. Across all populations, most PTVs found in a given individual are common (>5% popmax AF). c-d) Number of genes with at least one PTV (c) or homozygous PTV (d) as a function of number of individuals, downsampled from ExAC.