

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

2 Exome Aggregation Consortium[#], Monkol Lek^{1,2,3,4}, Konrad J Karczewski^{1,2*}, Eric V
3 Minikel^{1,2,5*}, Kaitlin E Samocha^{1,2,6,5*}, Eric Banks², Timothy Fennell², Anne H O'Donnell-
4 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
5 Tukiainen^{1,2}, Daniel P Birnbaum², Jack A Kosmicki^{1,2,6,13}, Laramie E Duncan^{1,2,6}, Karol
6 Estrada^{1,2}, Fengmei Zhao^{1,2}, James Zou², Emma Pierce-Hoffman^{1,2}, Joanne
7 Berghout^{14,15}, David N Cooper¹⁶, Nicole Deflaux¹⁷, Mark DePristo¹⁸, Ron Do^{19,20,21,22},
8 Jason Flannick^{2,23}, Menachem Fromer^{1,6,24,19,20}, Laura Gauthier¹⁸, Jackie Goldstein^{1,2,6},
9 Namrata Gupta², Daniel Howrigan^{1,2,6}, Adam Kiezun¹⁸, Mitja I Kurki^{2,25}, Ami Levy
10 Moonshine¹⁸, Pradeep Natarajan^{2,26,27,28}, Lorena Orozco²⁹, Gina M Peloso^{2,27,28}, Ryan
11 Poplin¹⁸, Manuel A Rivas², Valentin Ruano-Rubio¹⁸, Samuel A Rose⁶, Douglas M
12 Ruderfer^{24,19,20}, Khalid Shakir¹⁸, Peter D Stenson¹⁶, Christine Stevens², Brett P
13 Thomas^{1,2}, Grace Tiao¹⁸, Maria T Tusie-Luna³⁰, Ben Weisburd², Hong-Hee Won³¹,
14 Dongmei Yu^{6,27,25,32}, David M Altshuler^{2,33}, Diego Ardissino³⁴, Michael Boehnke³⁵, John
15 Danesh³⁶, Stacey Donnelly², Roberto Elosua³⁷, Jose C Florez^{2,26,27}, Stacey B Gabriel²,
16 Gad Getz^{18,26,38}, Stephen J Glatt^{39,40,41}, Christina M Hultman⁴², Sekar Kathiresan^{2,26,27,28},
17 Markku Laakso⁴³, Steven McCarroll^{6,8}, Mark I McCarthy^{44,45,46}, Dermot McGovern⁴⁷, Ruth
18 McPherson⁴⁸, Benjamin M Neale^{1,2,6}, Aarno Palotie^{1,2,5,49}, Shaun M Purcell^{24,19,20}, Danish
19 Saleheen^{50,51,52}, Jeremiah M Scharf^{2,6,27,25,32}, Pamela Sklar^{24,19,20,53,54}, Patrick F
20 Sullivan^{55,56}, Jaakko Tuomilehto⁵⁷, Ming T Tsuang⁵⁸, Hugh C Watkins^{59,44}, James G
21 Wilson⁶⁰, Mark J Daly^{1,2,6}, Daniel G MacArthur^{1,2†}

22

23 ¹Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA,
24 USA

25 ²Program in Medical and Population Genetics, Broad Institute of MIT and Harvard,
26 Cambridge, MA, USA

27 ³School of Paediatrics and Child Health, University of Sydney, Sydney, NSW, Australia

28 ⁴Institute for Neuroscience and Muscle Research, Childrens Hospital at Westmead,
29 Sydney, NSW, Australia

30 ⁵Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA,
31 USA

32 ⁶Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard,
33 Cambridge, MA, USA

34 ⁷Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA

- 1 ⁸Department of Genetics, Harvard Medical School, Boston, MA, USA
- 2 ⁹National Heart and Lung Institute, Imperial College London, London, UK
- 3 ¹⁰NIHR Royal Brompton Cardiovascular Biomedical Research Unit, Royal Brompton
- 4 Hospital, London, UK
- 5 ¹¹MRC Clinical Sciences Centre, Imperial College London, London, UK
- 6 ¹²Genome Sciences, University of Washington, Seattle, WA, USA
- 7 ¹³Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston,
- 8 MA, USA
- 9 ¹⁴Mouse Genome Informatics, Jackson Laboratory, Bar Harbor, ME, USA
- 10 ¹⁵Center for Biomedical Informatics and Biostatistics, University of Arizona, Tucson, AZ,
- 11 USA
- 12 ¹⁶Institute of Medical Genetics, Cardiff University, Cardiff, UK
- 13 ¹⁷Google Inc, Mountain View, CA, USA
- 14 ¹⁸Broad Institute of MIT and Harvard, Cambridge, MA, USA
- 15 ¹⁹Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount
- 16 Sinai, New York, NY, USA
- 17 ²⁰Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount
- 18 Sinai, New York, NY, USA
- 19 ²¹The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at
- 20 Mount Sinai, New York, NY, USA
- 21 ²²The Center for Statistical Genetics, Icahn School of Medicine at Mount Sinai, New
- 22 York, NY, USA
- 23 ²³Department of Molecular Biology, Massachusetts General Hospital, Boston, MA, USA
- 24 ²⁴Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY,
- 25 USA
- 26 ²⁵Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital,
- 27 Boston, MA, USA
- 28 ²⁶Harvard Medical School, Boston, MA, USA
- 29 ²⁷Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA,
- 30 USA
- 31 ²⁸Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA
- 32 ²⁹Immunogenomics and Metabolic Disease Laboratory, Instituto Nacional de Medicina
- 33 Gen—mica, Mexico City, Mexico

1 ³⁰Molecular Biology and Genomic Medicine Unit, Instituto Nacional de Ciencias Médicas
2 y Nutrición, Mexico City, Mexico

3 ³¹Samsung Advanced Institute for Health Sciences and Technology (SAIHST),
4 Sungkyunkwan University, Samsung Medical Center, Seoul, Republic of Korea

5 ³²Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

6 ³³Vertex Pharmaceuticals, Boston, MA, USA

7 ³⁴Department of Cardiology, University Hospital, Parma, Italy

8 ³⁵Department of Biostatistics and Center for Statistical Genetics, University of Michigan,
9 Ann Arbor, MI, USA

10 ³⁶Department of Public Health and Primary Care, Strangeways Research Laboratory,
11 Cambridge, UK

12 ³⁷Cardiovascular Epidemiology and Genetics, Hospital del Mar Medical Research
13 Institute, Barcelona, Spain

14 ³⁸Department of Pathology and Cancer Center, Massachusetts General Hospital,
15 Boston, MA, USA

16 ³⁹Psychiatric Genetic Epidemiology & Neurobiology Laboratory, State University of New
17 York, Upstate Medical University, Syracuse, NY, USA

18 ⁴⁰Department of Psychiatry and Behavioral Sciences, State University of New
19 York, Upstate Medical University, Syracuse, NY, USA

20 ⁴¹Department of Neuroscience and Physiology, State University of New York, Upstate
21 Medical University, Syracuse, NY, USA

22 ⁴²Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm,
23 Sweden

24 ⁴³Department of Medicine, University of Eastern Finland and Kuopio University Hospital,
25 Kuopio, Finland

26 ⁴⁴Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

27 ⁴⁵Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford,
28 Oxford, UK

29 ⁴⁶Oxford NIHR Biomedical Research Centre, Oxford University Hospitals Foundation
30 Trust, Oxford, UK

31 ⁴⁷Inflammatory Bowel Disease and Immunobiology Research Institute, Cedars-Sinai
32 Medical Center, Los Angeles, CA, USA

33 ⁴⁸Atherogenomics Laboratory, University of Ottawa Heart Institute, Ottawa, ON, Canada

⁴⁹Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

⁵⁰Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA

⁵¹Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA

⁵²Center for Non-Communicable Diseases, Karachi, , Pakistan

⁵³Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

⁵⁴Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA

⁵⁵Department of Genetics, University of North Carolina, Chapel Hill, NC, USA

⁵⁶Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

⁵⁷Department of Public Health, University of Helsinki, Helsinki, Finland

⁵⁸Department of Psychiatry, University of California, San Diego, CA, USA

⁵⁹Radcliffe Department of Medicine, University of Oxford, Oxford, UK

⁶⁰Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA

19

* These authors contributed equally to this work and names appear in alphabetical order

† Corresponding author

List of collaborators to appear at the end of manuscript

23

1 **Summary**

2 Large-scale reference data sets of human genetic variation are critical for the medical
3 and functional interpretation of DNA sequence changes. Here we describe the
4 aggregation and analysis of high-quality exome (protein-coding region) sequence data
5 for 60,706 individuals of diverse ethnicities generated as part of the Exome Aggregation
6 Consortium (ExAC). The resulting catalogue of human genetic diversity contains an
7 average of one variant every eight bases of the exome, and provides direct evidence for
8 the presence of widespread mutational recurrence. We show that this catalogue can be
9 used to calculate objective metrics of pathogenicity for sequence variants, and to identify
10 genes subject to strong selection against various classes of mutation; we identify 3,230
11 genes with near-complete depletion of truncating variants, 72% of which have no
12 currently established human disease phenotype. Finally, we demonstrate that these data
13 can be used for the efficient filtering of candidate disease-causing variants, and for the
14 discovery of human “knockout” variants in protein-coding genes.

15

16 **Background**

17 Over the last five years, the widespread availability of high-throughput DNA sequencing
18 technologies has permitted the sequencing of the whole genomes or exomes (the
19 protein-coding regions of genomes) of hundreds of thousands of humans. In theory,
20 these data represent a powerful source of information about the global patterns of
21 human genetic variation, but in practice, are difficult to access for practical, logistical,
22 and ethical reasons; in addition, their utility is complicated by the heterogeneity in the
23 experimental methodologies and variant calling pipelines used to generate them. Current
24 publicly available datasets of human DNA sequence variation contain only a small
25 fraction of all sequenced samples: the Exome Variant Server, created as part of the
26 NHLBI Exome Sequencing Project (ESP)¹, contains frequency information spanning

1 6,503 exomes; and the 1000 Genomes (1000G) Project, which includes individual-level
2 genotype data from whole-genome and exome sequence data for 2,504 individuals².

3
4 Databases of genetic variation are important for our understanding of human population
5 history and biology¹⁻⁵, but also provide critical resources for the clinical interpretation of
6 variants observed in patients suffering from rare Mendelian diseases^{6,7}. The filtering of
7 candidate variants by frequency in unselected individuals is a key step in any pipeline for
8 the discovery of causal variants in Mendelian disease patients, and the efficacy of such
9 filtering depends on both the size and the ancestral diversity of the available reference
10 data.

11
12 Here, we describe the joint variant calling and analysis of high-quality variant calls
13 across 60,706 human exomes, assembled by the Exome Aggregation Consortium
14 (ExAC; exac.broadinstitute.org). This call set exceeds previously available exome-wide
15 variant databases by nearly an order of magnitude, providing substantially increased
16 resolution for the analysis of very low-frequency genetic variants. We demonstrate the
17 application of this data set to the analysis of patterns of genetic variation including the
18 discovery of widespread mutational recurrence, the inference of gene-level constraint
19 against truncating variation, the clinical interpretation of variation in Mendelian disease
20 genes, and the discovery of human “knockout” variants in protein-coding genes.

21

22 **The ExAC Data set**

23 Sequencing data processing, variant calling, quality control and filtering was performed
24 on over 91,000 exomes (see Online Methods), and sample filtering was performed to
25 produce a final data set spanning 60,706 individuals (Figure 1a). To identify the ancestry
26 of each ExAC individual, we performed principal component analysis (PCA) to

1 distinguish the major axes of geographic ancestry and to identify population clusters
 2 corresponding to individuals of European, African, South Asian, East Asian, and
 3 admixed American (hereafter Latino) ancestry (Figure 1b; Supplementary Information
 4 Table 3); we note that the apparent separation between East Asian and other samples
 5 reflects a deficiency of Middle Eastern and Central Asian samples in the data set. We
 6 further separated Europeans into individuals of Finnish and non-Finnish ancestry given
 7 the enrichment of this bottlenecked population; the term “European” hereafter refers to
 8 non-Finnish European individuals.

9

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

17

18 The density of variation in ExAC is not uniform across the genome, and the observation
 19 of variants depends on factors such as mutational properties and selective pressures. In
 20 the ~45M well covered (80% of individuals with a minimum of 10X coverage) positions in
 21 ExAC, there are ~18M possible synonymous variants, of which we observe 1.4M (7.5%).
 22 However, we observe 63.1% of possible CpG transitions (C to T variants, where
 23 the adjacent base is G), while only observing 3% of possible transversions and 9.2% of
 24 other possible transitions (Supplementary Information Table 9). A similar pattern is
 25 observed for missense and nonsense variants, with lower proportions due to selective
 26 pressures (Figure 1D). Of 123,629 HQ insertion/deletions (indels) called in coding

1 exons, 117,242 (95%) have length <6 bases, with shorter deletions being the most
2 common (Figure 1E). Frameshifts are found in smaller numbers and are more likely to
3 be singletons than in-frame indels (Figure 1F), reflecting the influence of purifying
4 selection.

5

6 **Patterns of protein-coding variation revealed by large samples**

7 The density of protein-coding sequence variation in ExAC reveals a number of
8 properties of human genetic variation undetectable in smaller data sets. For instance,
9 7.9% of HQ sites in ExAC are multiallelic (multiple different sequence variants observed
10 at the same site), close to the Poisson expectation of 8.3% given the observed density of
11 variation, and far higher than observed in previous data sets - 0.48% in 1000 Genomes
12 (exome intervals) and 0.43% in ESP.

13

14 The size of ExAC also makes it possible to directly observe mutational recurrence:
15 instances in which the same mutation has occurred multiple times independently
16 throughout the history of the sequenced populations. For instance, among synonymous
17 variants, a class of variation expected to have undergone minimal selection, 43% of
18 validated *de novo* events identified in external datasets of 1,756 parent-offspring trios^{8,9}
19 are also observed independently in our dataset (Figure 2a), indicating a separate origin
20 for the same variant within the demographic history of the two samples. This proportion
21 is much higher for transition variants at CpG sites, well established to be the most highly
22 mutable sites in the human genome¹⁰: 87% of previously reported *de novo* CpG
23 transitions at synonymous sites are observed in ExAC, indicating that our sample sizes
24 are beginning to approach saturation of this class of variation. This saturation is
25 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 exome-wide 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 context¹¹ ($r = -0.98$; $p < 10^{-50}$; 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). Low-mutability mutations (especially transversions), 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 (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), in cases where multiple substitutions were observed within the same codon in at least one individual. We found 5,945 MNPs (mean: 23 per sample) in ExAC (Extended Data Figure 3a) where analysis of the underlying SNPs without correct haplotype phasing would result in altered interpretation. These include 647 instances

1 where the effect of a protein-truncating variant (PTV) variant is eliminated by an adjacent
2 SNP (rescued PTV) and 131 instances where underlying synonymous or missense
3 variants result in PTV MNPs (gained PTV). Additionally our analysis revealed 8 MNPs in
4 disease-associated genes, resulting in either a rescued or gained PTV, and 10 MNPs
5 that have previously been reported as disease causing mutations (Supplementary
6 Information Table 10 and 11). We note that these variants would be missed by virtually
7 all currently available variant calling and annotation pipelines.

8

9 **Inferring variant deleteriousness and gene constraint**

10 Deleterious variants are expected to have lower allele frequencies than neutral ones,
11 due to negative selection. This theoretical property has been demonstrated previously in
12 human population sequencing data^{12,13} and here (Figure 1d, Figure 1e). This allows
13 inference of the degree of selection against specific functional classes of variation:
14 however, mutational recurrence as described above indicates that allele frequencies
15 observed in ExAC-scale samples are also skewed by mutation rate, with more mutable
16 sites less likely to be singletons (Figure 2c and Extended Data Figure 4d). Mutation rate
17 is in turn non-uniformly distributed across functional classes - for instance, stop lost
18 mutations can never occur at CpG dinucleotides (Extended Data Figure 4e). We
19 corrected for mutation rates (Supplementary Information Section 3.2) by creating a
20 mutability-adjusted proportion singleton (MAPS) metric. This metric reflects (as
21 expected) strong selection against predicted PTVs, as well as missense variants
22 predicted by conservation-based methods to be deleterious (Figure 2e).

23

24 The deep ascertainment of rare variation in ExAC also allows us to infer the extent of
25 selection against variant categories on a per-gene basis by examining the proportion of
26 variation that is missing compared to expectations under random mutation. Conceptually

1 similar approaches have been applied to smaller exome datasets^{11,14} but have been
2 underpowered, particularly when analyzing the depletion of PTVs. We compared the
3 observed number of rare (MAF <0.1%) variants per gene to an expected number derived
4 from a selection neutral, sequence-context based mutational model¹¹. The model
5 performs well in predicting the number of synonymous variants, which should be under
6 minimal selection, per gene ($r = 0.98$; Extended Data Figure 5b).

7

8 We quantified deviation from expectation with a Z score¹¹, which for synonymous
9 variants is centered at zero, but is significantly shifted towards higher values (greater
10 constraint) for both missense and PTV (Wilcoxon $p < 10^{-50}$ for both; Figure 3a). The
11 genes on the X chromosome are significantly more constrained than those on the
12 autosomes for missense ($p < 10^{-7}$) and loss-of-function ($p < 10^{-50}$). The high correlation
13 between the observed and expected number of synonymous variants on the X
14 chromosome ($r = 0.97$ vs 0.98 for autosomes) indicates that this difference in constraint
15 is not due to a calibration issue. To reduce confounding by coding sequence length for
16 PTVs, we developed an expectation-maximization algorithm (Supplementary Information
17 Section 4.4) using the observed and expected PTV counts within each gene to separate
18 genes into three categories: null (observed $\approx$ expected), recessive (observed $\leq 50\%$ of
19 expected), and haploinsufficient (observed $< 10\%$ of expected). This metric – the
20 probability of being loss-of-function (LoF) intolerant (pLI) – separates genes of sufficient
21 length into LoF intolerant ($pLI \geq 0.9$, $n=3,230$) or LoF tolerant ($pLI \leq 0.1$, $n=10,374$)
22 categories. pLI is less correlated with coding sequence length ($r = 0.17$ as compared to
23 0.57 for the PTV Z score), outperforms the PTV Z score as an intolerance metric
24 (Supplementary Information Table 15), and reveals the expected contrast between gene
25 lists (Figure 3b). pLI is positively correlated with a gene product's number of physical
26 interaction partners ($p < 10^{-41}$). The most constrained pathways (highest median pLI for

1 the genes in the pathway) are core biological processes (spliceosome, ribosome, and
2 proteasome components; KS test $p < 10^{-6}$ for all) while olfactory receptors are among
3 the least constrained pathways (KS test $p < 10^{-16}$), demonstrated in Figure 3b and
4 consistent with previous work^{5,15–18}.

5

6 Critically, we note that LoF-intolerant genes include virtually all known severe
7 haploinsufficient human disease genes (Figure 3b), but that 72% of LoF-intolerant genes
8 have not yet been assigned a human disease phenotype despite clear evidence for
9 extreme selective constraint (Supplementary Information Table 13). We note that this
10 extreme constraint does not necessarily reflect a lethal disease, but is likely to point to
11 genes where heterozygous loss of function confers some non-trivial survival or
12 reproductive disadvantage.

13

14 The most highly constrained missense (top 25% missense Z scores) and PTV ($pLI \geq 0.9$)
15 genes show higher expression levels and broader tissue expression than the least
16 constrained genes¹⁹ (Figure 3c). These most highly constrained genes are also depleted
17 for eQTLs ($p < 10^{-9}$ for missense and PTV; Figure 3d), yet are enriched within genome-
18 wide significant trait-associated loci ($\chi^2 p < 10^{-14}$, Figure 3e). Intuitively, genes intolerant
19 of PTV variation are dosage sensitive: natural selection does not tolerate a 50% deficit in
20 expression due to the loss of single allele. Unsurprisingly, these genes are also depleted
21 of common genetic variants that have a large enough effect on expression to be
22 detected as eQTLs with current limited sample sizes. However, smaller changes in the
23 expression of these genes, through weaker eQTLs or functional variants, are more likely
24 to contribute to medically relevant phenotypes.

25

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

8

9 **ExAC improves variant interpretation in Mendelian disease**

10 We assessed the value of ExAC as a reference dataset for clinical sequencing
11 approaches, which typically prioritize or filter potentially deleterious variants based on
12 functional consequence and allele frequency (AF)⁶. Filtering on ExAC reduced the
13 number of candidate protein-altering variants by 7-fold compared to ESP, and was most
14 powerful when the highest AF in any one population (“popmax”) was used rather than
15 average (“global”) AF (Figure 4a). ESP is not well-powered to filter at 0.1% AF without
16 removing many genuinely rare variants, as AF estimates based on low allele counts are
17 both upward-biased and imprecise (Figure 4b). We thus expect that ExAC will provide a
18 very substantial boost in the power and accuracy of variant filtering in Mendelian disease
19 projects.

20

21 Previous large-scale sequencing studies have repeatedly shown that some purported
22 Mendelian disease-causing genetic variants are implausibly common in the population^{20–}
23 ²² (Figure 4c). The average ExAC participant harbors ~54 variants reported as disease-
24 causing in two widely-used databases of disease-causing variants (Supplementary
25 Information Section 5.2). Most (~41) of these are high-quality genotypes but with
26 implausibly high (>1%) popmax AF. We therefore hypothesized that most of the

1 supposed burden of Mendelian disease alleles per person is due not to genotyping error,
2 but rather to misclassification in the literature and/or in databases.

3

4 We manually curated the evidence of pathogenicity for 192 previously reported
5 pathogenic variants with AF >1% either globally or in South Asian or Latino individuals,
6 populations that are underrepresented in previous reference databases. Nine variants
7 had sufficient data to support disease association, typically with either mild or
8 incompletely penetrant disease effects; the remainder either had insufficient evidence for
9 pathogenicity, no claim of pathogenicity, or were benign traits (Supplementary
10 Information Section 5.3). It is difficult to prove the absence of any disease association,
11 and incomplete penetrance or genetic modifiers may contribute in some cases.
12 Nonetheless, the high cumulative AF of these variants combined with their limited
13 original evidence for pathogenicity suggest little contribution to disease, and 163 variants
14 met American College of Medical Genetics criteria²³ for reclassification as benign or
15 likely benign (Figure 4d). 126 of these 163 have been reclassified in source databases
16 as of December 2015 (Supplementary Information Table 20). Supporting functional data
17 were reported for 18 of these variants, highlighting the need to review cautiously even
18 variants with experimental support.

19

20 We also sought phenotypic data for a subset of ExAC participants homozygous for
21 reported severe recessive disease variants, again enabling reclassification of some
22 variants as benign. North American Indian Childhood Cirrhosis is a recessive disease of
23 cirrhotic liver failure during childhood requiring liver transplant for survival to adulthood,
24 previously reported to be caused by *CIRH1A* p.R565W²⁴. ExAC contains 222
25 heterozygous and 4 homozygous Latino individuals, with a population AF of 1.92%. The
26 4 homozygotes had no history of liver disease and recontact in two individuals revealed

normal liver function (Supplementary Information Table 22). 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 either benign or insufficient to cause disease, highlighting the importance of matched reference populations.

The above curation efforts confirm the importance of AF filtering in analysis of candidate disease variants^{6,25,26}. However, literature and database errors are prevalent even at lower AFs: the average ExAC individual contains 0.89 (<1% popmax AF) reportedly Mendelian variants in well-characterized dominant disease genes²⁷ and 0.21 at <0.1% popmax AF. This inflation likely results from a combination of false reports of pathogenicity and incomplete penetrance, as we have recently shown for *PRNP*²⁸. 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^{7,23}.

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 disease²⁹, as well as rare cases where LoF is protective against disease³⁰.

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 ($pLI > 0.9$). ExAC recapitulates known aspects of population demographic models, including an increase in intermediate-frequency (1-5%) PTVs in Finland³¹ and relatively common (>1%) PTVs in Africans (Figure 5b). However, these differences are diminished when considering only LoF-constrained ($pLI > 0.9$) genes (Extended Data Figure 10).

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 low-frequency protein-coding variants in human populations, as well as a public resource [exac.broadinstitute.org] for the clinical interpretation of genetic variants observed in disease patients.

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^{11,14}, ExAC provides for the first time sufficient power to investigate genic intolerance to PTVs, highlighting 3,230 highly LoF-intolerant genes, 72% of which have no established human disease phenotype in OMIM or ClinVar. We note that this extreme constraint does not necessarily reflect a lethal disease, but is likely to point to genes where heterozygous loss of function confers some non-trivial survival or reproductive disadvantage. In independent work [Ruderfer et al., manuscript 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 provide 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 ESP¹ have been widely used for this purpose, but those data are limited by population diversity and by resolution at allele frequencies $\leq 0.1\%$. 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.

Finally, we show that different populations confer different advantages in the discovery of gene-disrupting PTVs, providing guidance for the identification of human “knockouts” to understand gene function. Sampling multiple populations would likely be a fruitful strategy for a researcher investigating common PTV variation. However, discovery of

1 homozygous PTVs is markedly enhanced in the South Asian samples, which come
2 primarily from a Pakistani cohort with 38.3% of individuals self-reporting as having
3 closely related parents, emphasizing the extreme value of consanguineous cohorts for
4 “human knockout” discovery^{32–34} (Figure 5d). Other approaches to enriching for
5 homozygosity of rare PTVs, such as focusing on bottlenecked populations, have already
6 proved fruitful^{31,32}.

7

8 Even with this large collection of jointly processed exomes, many limitations remain.
9 Firstly, most ExAC individuals were ascertained for biomedically important
10 disease; while we have attempted to exclude severe pediatric diseases, the inclusion of
11 both cases and controls for several polygenic disorders means that ExAC certainly
12 contains disease-associated variants³⁵. Secondly, future reference databases would
13 benefit from including a broader sampling of human diversity, especially from under-
14 represented Middle Eastern and African populations. Thirdly, the inclusion of whole
15 genomes will also be critical to investigate additional classes of functional variation and
16 identify non-coding constrained regions. Finally, and most critically, detailed phenotype
17 data are unavailable for the vast majority of ExAC samples; future initiatives that
18 assemble sequence and clinical data from very large-scale cohorts will be required to
19 fully translate human genetic findings into biological and clinical understanding.

20

21 While the ExAC dataset exceeds the scale of previously available frequency reference
22 datasets, much remains to be gained by further increases in sample size. Indeed, the
23 fact that even the rarest transversions have mutational rates¹¹ on the order of 1×10^{-9}
24 implies that the vast majority of possible non-lethal SNVs likely exist in some living
25 human. ExAC already includes >63% of all possible protein-coding CpG transitions at

1 well-covered synonymous sites; orders-of-magnitude increases in sample size will
2 eventually lead to saturation of other classes of variation.

3

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

9

10

1 **Online Methods**

2 **Variant discovery**

3 We assembled approximately 1 petabyte of raw sequencing data (FASTQ files) from
 4 91,796 individual exomes drawn from a wide range of primarily disease-focused
 5 consortia (Supplementary Information Table 2). We processed these exomes through a
 6 single informatic pipeline and performed joint variant calling of single nucleotide variants
 7 (SNVs) and short insertions and deletions (indels) across all samples using a new
 8 version of the Genome Analysis Toolkit (GATK) HaplotypeCaller pipeline. Variant
 9 discovery was performed within a defined exome region that includes Gencode v19
 10 coding regions and flanking 50 bases. At each site, sequence information from all
 11 individuals was used to assess the evidence for the presence of a variant in each
 12 individual. Full details of data processing, variant calling and resources are described in
 13 the Supplementary Information Sections 1.1-1.4.

14

15 **Quality assessment**

16 We leveraged a variety of sources of internal and external validation data to calibrate
 17 filters and evaluate the quality of filtered variants (Supplementary Information Table 7).
 18 We adjusted the standard GATK variant site filtering³⁶ to increase the number of
 19 singleton variants that pass this filter, while maintaining a singleton transmission rate of
 20 50.1%, very near the expected 50%, within sequenced trios. We then used the
 21 remaining passing variants to assess depth and genotype quality filters compared to
 22 >10,000 samples that had been directly genotyped using SNP arrays (Illumina
 23 HumanExome) and achieved 97-99% heterozygous concordance, consistent with known
 24 error rates for rare variants in chip-based genotyping³⁷. Relative to a “platinum standard”
 25 genome sequenced using five different technologies³⁸, we achieved sensitivity of 99.8%
 26 and false discovery rates (FDR) of 0.056% for single nucleotide variants (SNVs), and

1 corresponding rates of 95.1% and 2.17% for insertions and deletions (indels). Lastly, we
 2 compared 13 representative Non-Finnish European exomes included in the call set with
 3 their corresponding 30x PCR-Free genome. The overall SNV and indel FDR was 0.14%
 4 and 4.71%, while for SNV singletons was 0.389%. The overall FDR by annotation
 5 classes missense, synonymous and protein truncating variants (including indels) were
 6 0.076%, 0.055% and 0.471% respectively (Supplementary Information Table 5 and 6).
 7 Full details of quality assessments are described in the Supplementary Information
 8 Section 1.6.

9

10 **Sample filtering**

11 The 91,796 samples were filtered based on two criteria. First, samples that were outliers
 12 for key metrics were removed (Extended Data Figure 2b). Second, in order to generate
 13 allele frequencies based on independent observations without enrichment of Mendelian
 14 disease alleles, we restricted the final release data set to unrelated adults with high-
 15 quality sequence data and without severe pediatric disease. After filtering, only 60,706
 16 samples remained, consisting of ~77% of Agilent (33 Mb target) and ~12% of Illumina
 17 (37.7 Mb target) exome captures. Full details of the filtering process are described in the
 18 Supplementary Information Section 1.7.

19

20 **ExAC data release**

21 For each variant, summary data for genotype quality, allele depth and population specific
 22 allele counts were calculated before removing all genotype data. This variant summary
 23 file was then functionally annotated using variant effect predictor (VEP) with the LOFTEE
 24 plugin. This data set can be accessed via the ExAC Browser
 25 (<http://exac.broadinstitute.org>) or downloaded from
 26 ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/ExAC.r0.3.sites.vcf.gz. Full

1 details regarding the annotation of the ExAC data set are described in the
2 Supplementary Information Sections 1.9-1.10.

3

4

5

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. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–8 (2015).
9. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–84 (2014).
10. Cooper, D. N. & Youssoufian, H. The CpG dinucleotide and human genetic disease. *Hum. Genet.* **78**, 151–155 (1988).
11. 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
12. Tennessen, J. a *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–9 (2012).
13. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).

- 1 14. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B.
2 Genic intolerance to functional variation and the interpretation of personal
3 genomes. *PLoS Genet.* **9**, e1003709 (2013).
- 4 15. Jeong, H., Mason, S. P., Barabási, a L. & Oltvai, Z. N. Lethality and
5 centrality in protein networks. *Nature* **411**, 41–42 (2001).
- 6 16. Goh, K.-I. *et al.* The human disease network. *Proc. Natl. Acad. Sci. U. S. A.*
7 **104**, 8685–8690 (2007).
- 8 17. Rolland, T. *et al.* Resource A Proteome-Scale Map of the Human
9 Interactome Network. *Cell* **159**, 1212–1226 (2014).
- 10 18. Itan, Y. *et al.* The human gene damage index as a gene-level approach to
11 prioritizing exome variants. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13615–20
12 (2015).
- 13 19. GTEx Consortium. Human genomics. The Genotype-Tissue Expression
14 (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**,
15 648–60 (2015).
- 16 20. Bell, C. J. *et al.* Carrier testing for severe childhood recessive diseases by
17 next-generation sequencing. *Sci. Transl. Med.* **3**, 65ra4 (2011).
- 18 21. Xue, Y. *et al.* Deleterious- and disease-allele prevalence in healthy
19 individuals: Insights from current predictions, mutation databases, and
20 population-scale resequencing. *Am. J. Hum. Genet.* **91**, 1022–1032 (2012).
- 21 22. Piton, A., Redin, C. & Mandel, J.-L. XLID-Causing Mutations and
22 Associated Genes Challenged in Light of Data From Large-Scale Human
23 Exome Sequencing. *Am. J. Hum. Genet.* **93**, 368–383 (2013).
- 24 23. Richards, S. *et al.* Standards and guidelines for the interpretation of
25 sequence variants: a joint consensus recommendation of the American
26 College of Medical Genetics and Genomics and the Association for
27 Molecular Pathology. *Genet. Med.* **17**, 405–423 (2015).
- 28 24. Chagnon, P. *et al.* A missense mutation (R565W) in cirhin (FLJ14728) in
29 North American Indian childhood cirrhosis. *Am. J. Hum. Genet.* **71**, 1443–9
30 (2002).
- 31 25. Stenson, P. D. *et al.* The Human Gene Mutation Database: Building a
32 comprehensive mutation repository for clinical and molecular genetics,
33 diagnostic testing and personalized genomic medicine. *Hum. Genet.* **133**,
34 1–9 (2014).

- 1 26. Dewey, F. E. *et al.* Sequence to Medical Phenotypes: A Framework for
2 Interpretation of Human Whole Genome DNA Sequence Data. *PLOS*
3 *Genet.* **11**, e1005496 (2015).
- 4 27. Blekhman, R. *et al.* Natural Selection on Genes that Underlie Human
5 Disease Susceptibility. *Curr. Biol.* **18**, 883–889 (2008).
- 6 28. Minikel, E. V. *et al.* Quantifying prion disease penetrance using large
7 population control cohorts. *Sci. Transl. Med.* **8**, 322ra9–322ra9 (2016).
- 8 29. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes:
9 Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* 1–17
10 (2015). doi:10.1016/j.ajhg.2015.06.009
- 11 30. Kathiresan, S. Developing Medicines That Mimic the Natural Successes of
12 the Human Genome. *J. Am. Coll. Cardiol.* **65**, 1562–1566 (2015).
- 13 31. Lim, E. T. *et al.* Distribution and Medical Impact of Loss-of-Function
14 Variants in the Finnish Founder Population. *PLoS Genet.* **10**, e1004494
15 (2014).
- 16 32. Sulem, P. *et al.* Identification of a large set of rare complete human
17 knockouts. *Nat. Genet.* **47**, 448–452 (2015).
- 18 33. Narasimhan, V. M. *et al.* Health and population effects of rare gene
19 knockouts in adult humans with related parents. *Science (80-.).* **8624**, 1–8
20 (2016).
- 21 34. Saleheen, D. *et al.* *Human knockouts in a cohort with a high rate of*
22 *consanguinity.* *bioRxiv* (2015). doi:10.1101/031518
- 23 35. Freischmidt, A. *et al.* Haploinsufficiency of TBK1 causes familial ALS and
24 fronto-temporal dementia. *Nat. Neurosci.* **18**, (2015).
- 25 36. DePristo, M. a *et al.* A framework for variation discovery and genotyping
26 using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498
27 (2011).
- 28 37. Voight, B. F. *et al.* The MetaboChip, a Custom Genotyping Array for
29 Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits.
30 *PLoS Genet.* **8**, e1002793 (2012).
- 31 38. Zook, J. M. *et al.* Integrating human sequence data sets provides a
32 resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**,
33 246–251 (2014).

1

2 **Acknowledgements**

3 We would like to thank the reviewers and editor for their time, valuable comments and
4 suggestions. The scientific community for their support and comments on biorxiv, twitter
5 and other public forums. Brendan Bulik-Sullivan and Jon Bloom for their help with
6 mathematical notation.

7

8 M.Lek is supported by the Australian National Health and Medical Research Council CJ
9 Martin Fellowship, Australian American Association Sir Keith Murdoch Fellowship and
10 the MDA/AANEM Development Grant. K.J.K. is supported by NIGMS Fellowship
11 (F32GM115208). A.H.O. is supported by Pfizer/ACMG Foundation Translational
12 Genomic Fellowship. J.S.W. is supported by Fondation Leducq and Wellcome Trust.
13 A.J.H. is supported by NSF Graduate Research Fellowship. M.I.K is supported by
14 Instrumentarium Science Foundation, Finland; Finnish Foundation for Cardiovascular
15 Research; Orion Research Foundation and the University of Eastern Finland,
16 Saastamoinen Foundation. P.N. is supported by John S. LaDue Memorial Fellowship in
17 Cardiology, Harvard Medical School. G.M.P. is supported by the National Heart, Lung,
18 and Blood Institute of the National Institutes of Health under Award Number
19 K01HL125751. M.T.Tusie-Luna is supported by CONACyT grant 128877. H.W. is
20 supported by postdoctoral award from the American Heart Association
21 (15POST23280019). R.E. is supported by Instituto Salud Carlos III-FIS-FEDER-ERDF:
22 RD12/0042/0013, PI12/00232; Agència de Gestió Ajuts Universitaris de Recerca: 2014
23 SGR 240. S.K. is supported by grants from the National Institutes of Health
24 (R01HL107816), the Donovan Family Foundation and Fondation Leducq. S.J.G. is
25 supported by NIH/NIMH grant R01MH085521 and NARSAD: The Brain and Behavior
26 Research Foundation and the Sidney R. Baer, Jr. Foundation. M.I.M is supported by
27 Wellcome Trust Senior Investigator, NIHR Senior Investigator;; EU Framework VII
28 HEALTH-F4-2007-201413; Medical Research Council G0601261; Wellcome Trust
29 090532, 098381, 090367; NIH RC2-DK088389, U01-DK085545. R.M is supported by
30 Canadian Institutes of Health Research MOP136936; MOP82810, MOP77682,
31 Canadian Foundation for Innovation 11966, Heart & Stroke Foundation of Canada T-
32 7268. J.M.S is supported by NINDS grants NS40024-09S1 and NS085048. P.S. is
33 supported by NIMH grant MH095034 and MH089905. P.F.S is supported by Swedish
34 Research Council award D0886501; NIMH grants MH077139 and MH094421; Yeargen

Family; Stanley Center. H.C.W. is supported by BHF Centre of Research Excellence, NIHR Senior Investigator. M.T.Tsuang is supported by NIH/NIMH grant R01MH085560. D.G.M is supported by NIGMS R01 GM104371 and NIDDK U54 DK105566.

ATVB & Precocious Coronary Artery Disease Study (PROCARDIS): Exome sequencing was supported by a grant from the NHGRI (5U54HG003067-11) to Drs. Gabriel and Lander. **Bulgarian Trios:** Medical Research Council (MRC) Centre (G0800509) and Program Grants (G0801418), the European Community's Seventh Framework Programme (HEALTH-F2-2010-241909 (Project EU-GEI)), and NIMH(2P50MH066392-05A1). **GoT2D & T2DGENES:** NHGRI ("Large Scale Sequencing and Analysis of Genomes" U54HG003067), NIDDK ("Multiethnic Study of Type 2 Diabetes Genes" U01DK085526), NIH ("LowF Pass Sequencing and High Density SNP Genotyping in Type 2 Diabetes" 1RC2DK088389), National Institutes of Health ("Multiethnic Study of Type 2 Diabetes Genes" U01s DK085526, DK085501, DK085524, DK085545, DK085584; "LowF Pass Sequencing and HighF Density SNP Genotyping for Type 2 Diabetes" DK088389). The German Center for Diabetes Research (DZD). National Institutes of Health (RC2F DK088389, DK085545, DK098032). Wellcome Trust (090532, 098381). National Institutes of Health (R01DK062370, R01DK098032, RC2DK088389). **METSIM:** Academy of Finland and the Finnish Cardiovascular Research Foundation. **Inflammatory Bowel Disease:** The Helmsley Trust Foundation, #2015PG-IBD001, Large Scale Sequencing and Analysis of Genomes Grant (NHGRI), 5 U54 HG003067-13. **Jackson Heart Study:** We thank the Jackson Heart Study (JHS) participants and staff for their contributions to this work. The JHS is supported by contracts HHSN268201300046C, HHSN268201300047C, HHSN268201300048C, HHSN268201300049C, HHSN268201300050C from the National Heart, Lung, and Blood Institute and the National Institute on Minority Health and Health Disparities. **Ottawa Genomics Heart Study:** Canadian Institutes of Health Research MOP136936; MOP82810, MOP77682, Canadian Foundation for Innovation 11966, Heart & Stroke Foundation of Canada T-7268. Exome sequencing was supported by a grant from the NHGRI (5U54HG003067-11) to Drs. Gabriel and Lander. **Pakistan Risk of Myocardial Infarction Study (PROMIS):** Exome sequencing was supported by a grant from the NHGRI (5U54HG003067-11) to Drs. Gabriel and Lander. Fieldwork in the study has been supported through funds available to investigators at the Center for Non-Communicable Diseases, Pakistan and the University of Cambridge, UK.

Registre Gironi del COR (REGICOR): Spanish Ministry of Economy and Innovation through the Carlos III Health Institute [Red HERACLES RD12/0042, CIBER Epidemiología y Salud Pública, PI12/00232, PI09/90506, PI08/1327, PI05/1251, PI05/1297], European Funds for Development (ERDF-FEDER), and by the Catalan Research and Technology Innovation Interdepartmental Commission [SGR 1195].

Swedish Schizophrenia & Bipolar Studies: National Institutes of Health (NIH)/National Institute of Mental Health (NIMH) ARRA Grand Opportunity grant NIMHRC2MH089905, the Sylvan Herman Foundation, the Stanley Center for Psychiatric Research, the Stanley Medical Research Institute, NIH/National Human Genome Research Institute (NHGRI) grant U54HG003067. **SIGMA-T2D:** The work was conducted as part of the Slim Initiative for Genomic Medicine, a project funded by the Carlos Slim Health Institute in Mexico. The UNAM/INCMNSZ Diabetes Study was supported by Consejo Nacional de Ciencia y Tecnología grants 138826, 128877, CONACT- SALUD 2009-01-115250, and a grant from Dirección General de Asuntos del Personal Académico, UNAM, IT 214711. The Diabetes in Mexico Study was supported by Consejo Nacional de Ciencia y Tecnología grant 86867 and by Instituto Carlos Slim de la Salud, A.C. The Mexico City Diabetes Study was supported by National Institutes of Health (NIH) grant R01HL24799 and by the Consejo Nacional de Ciencia y Tecnología grants 2092, M9303, F677-M9407, 251M, and 2005-C01-14502, SALUD 2010-2-151165. **Schizophrenia Trios from Taiwan:** NIH/NIMH grant R01MH085560. **Tourette Syndrome Association International Consortium for Genomics (TSAICG):** NIH/NINDS U01 NS40024-09S1. **Exome Aggregation Consortium (ExAC):** 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., E.P., 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., J.B., D.N.C., N.D., 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., S.A.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., S.D., R.E., J.C.F., S.B.G., G.G., S.J.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., M.T.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://exac.broadinstitute.org>

Collaborators (alphabetical order)

Hanna E Abboud⁶¹, Goncalo Abecasis³⁵, Carlos A Aguilar-Salinas⁶², Olimpia Arellano-Campos⁶², Gil Atzmon^{63,64}, Ingvald Aukrust^{65,66,67}, Cathy L Barr^{68,69}, Graeme I Bell⁷⁰, Graeme I Bell^{70,71}, Sarah Bergen⁴², Lise Bjørkhaug^{66,67}, John Blangero^{72,73}, Donald W Bowden^{74,75,76}, Cathy L Budman⁷⁷, Noël P Burt², Federico Centeno-Cruz⁷⁸, John C Chambers^{79,80,81}, Kimberly Chambert⁶, Robert Clarke⁸², Rory Collins⁸², Giovanni Coppola⁸³, Emilio J Córdova⁷⁸, Maria L Cortes¹⁸, Nancy J Cox⁸⁴, Ravindranath Duggirala⁸⁵, Martin Farrall^{59,44}, Juan C Fernandez-Lopez⁷⁸, Pierre Fontanillas², Timothy M Frayling⁸⁶, Nelson B Freimer⁸³, Christian Fuchsberger³⁵, Humberto García-Ortiz⁷⁸, Anuj Goel^{59,44}, María J Gómez-Vázquez⁶², María E González-Villalpando⁸⁷, Clicerio González-Villalpando⁸⁷, Marco A Grados⁸⁸, Leif Groop⁸⁹, Christopher A Haiman⁹⁰, Craig L Hanis⁹¹, Craig L Hanis⁹¹, Andrew T Hattersley⁸⁶, Brian E Henderson⁹², Jemma C Hopewell⁸², Alicia Huerta-Chagoya⁹³, Sergio Islas-Andrade⁹⁴, Suzanne BR Jacobs², Shapour Jalilzadeh^{59,44}, Christopher P Jenkinson⁶¹, Jennifer Moran², Silvia Jiménez-Morale⁷⁸, Anna Kähler⁴², Robert A King⁹⁵, George Kirov⁹⁶, Jaspal S Kooner^{80,9,81}, Theodosios Kyriakou^{59,44}, Jong-Young Lee⁹⁷, Donna M Lehman⁶¹, Gholson Lyon⁹⁸, William MacMahon⁹⁹, Patrik KE Magnusson⁴², Anubha Mahajan¹⁰⁰, Jaume Marrugat³⁷, Angélica Martínez-Hernández⁷⁸, Carol A Mathews¹⁰¹, Gilean McVean¹⁰⁰, James B Meigs^{102,26}, Thomas Meitinger^{103,104}, Elvia Mendoza-Caamal⁷⁸, Josep M Mercader^{2,105,106}, Karen L Mohlke⁵⁵, Hortensia Moreno-Macías¹⁰⁷, Andrew P Morris^{108,100,109}, Laeya A Najmi^{65,110}, Pål R Njølstad^{65,66}, Michael C O'Donovan⁹⁶, Maria L Ordóñez-Sánchez⁶², Michael J Owen⁹⁶, Taesung Park^{111,112}, David L Pauls²⁵, Danielle Posthuma^{113,114,115}, Cristina Revilla-Monsalve⁹⁴, Laura Riba⁹³, Stephan Ripke⁶, Rosario Rodríguez-Guillén⁶², Maribel Rodríguez-Torres⁶², Paul Sandor^{116,68}, Mark Seielstad^{117,118}, Rob Sladek^{119,120,121}, Xavier Soberón⁷⁸, Timothy D Spector¹²², Shyong E Tai^{123,124,125}, Tanya M Teslovich³⁵, Geoffrey Walford^{105,26}, Lynne R Wilkens⁹², Amy L Williams^{2,126}

⁶¹Department of Medicine, University of Texas Health Science Center, San Antonio, TX, USA

- 1 ⁶²Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City,
- 2 Mexico
- 3 ⁶³Departments of Medicine and Genetics, Albert Einstein College of Medicine, New York
- 4 City, NY, USA
- 5 ⁶⁴Department of Natural Science, University of Haifa, Haifa, Israel
- 6 ⁶⁵Department of Clinical Science, University of Bergen, Bergen, Norway
- 7 ⁶⁶Department of Pediatrics, Haukeland University Hospital, Bergen, Norway
- 8 ⁶⁷Department of Biomedicine, University of Bergen, Bergen, Norway
- 9 ⁶⁸The Toronto Western Research Institute, University Health Network, Toronto, Canada
- 10 ⁶⁹The Hospital for Sick Children, Toronto, Canada
- 11 ⁷⁰Departments of Medicine and Human Genetics, University of Chicago, Chicago, IL,
- 12 USA
- 13 ⁷¹Department of Medicine, University of Chicago, Chicago, IL, USA
- 14 ⁷²South Texas Diabetes and Obesity Institute, University of Texas Health Science
- 15 Center, San Antonio, TX, USA
- 16 ⁷³University of Texas Rio Grande Valley, Brownsville, TX, USA
- 17 ⁷⁴Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC,
- 18 USA
- 19 ⁷⁵Center for Genomics and Personalized Medicine Research, Wake Forest School of
- 20 Medicine, Winston-Salem, NC, USA
- 21 ⁷⁶Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, NC,
- 22 USA
- 23 ⁷⁷North Shore-Long Island Jewish Health System, Manhasset, NY, USA
- 24 ⁷⁸Instituto Nacional de Medicina Genómica, Mexico City, Mexico
- 25 ⁷⁹Department of Epidemiology and Biostatistics, Imperial College London, London, UK
- 26 ⁸⁰Department of Cardiology, Ealing Hospital NHS Trust, Southall, UK
- 27 ⁸¹Imperial College Healthcare NHS Trust, Imperial College London, London, UK
- 28 ⁸²Nuffield Department of Population Health, University of Oxford, Oxford, UK
- 29 ⁸³Center for Neurobehavioral Genetics, University of California, Los Angeles, CA, USA
- 30 ⁸⁴Vanderbilt Genetics Institute, Vanderbilt University School of Medicine, Nashville, TN,
- 31 USA
- 32 ⁸⁵Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA
- 33 ⁸⁶University of Exeter Medical School, University of Exeter, Exeter, UK
- 34 ⁸⁷Instituto Nacional de Salud Publica, Mexico City, Mexico

- 1 ⁸⁸Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School
2 of Medicine, Baltimore, MD, USA
- 3 ⁸⁹Department of Clinical Sciences, Lund University Diabetes Centre, Malm_, Sweden
- 4 ⁹⁰Department of Preventive Medicine, University of Southern California, Los Angeles,
5 CA, USA
- 6 ⁹¹Human Genetics Center, The University of Texas Health Science Center, Houston, TX,
7 USA
- 8 ⁹²Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA
- 9 ⁹³Instituto de Investigaciones Biomédicas, Mexico City, Mexico
- 10 ⁹⁴Instituto Mexicano del Seguro Social, Mexico City, Mexico
- 11 ⁹⁵Department of Genetics, Yale University School of Medicine, New Haven, CT, USA
- 12 ⁹⁶MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff,
13 UK
- 14 ⁹⁷Center for Genome Science, Korea National Institute of Health, Chungcheongbuk-do,
15 Republic of Korea
- 16 ⁹⁸Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, Woodbury,
17 NY, USA
- 18 ⁹⁹Department of Psychiatry, University of Utah, Salt Lake City, UT, USA
- 19 ¹⁰⁰Nuffield Department of Medicine, University of Oxford, Oxford, UK
- 20 ¹⁰¹Department of Psychiatry, University of Florida, Gainesville, FL, USA
- 21 ¹⁰²General Medicine Division, Massachusetts General Hospital, Boston, MA, USA
- 22 ¹⁰³Institute of Human Genetics, Technische Universität München, Munich, Germany
- 23 ¹⁰⁴Institute of Human Genetics, German Research Center for Environmental Health,
24 Neuherberg, Germany
- 25 ¹⁰⁵Diabetes Research Center (Diabetes Unit), Massachusetts General Hospital, Boston,
26 MA, USA
- 27 ¹⁰⁶Research Program in Computational Biology, Barcelona Supercomputing Center,
28 Barcelona, Spain
- 29 ¹⁰⁷Universidad Autónoma Metropolitana, Mexico City, Mexico
- 30 ¹⁰⁸Estonian Genome Centre, University of Tartu, Tartu, Estonia, University of Tartu, Tartu,
31 Estonia
- 32 ¹⁰⁹Department of Biostatistics, University of Liverpool, Liverpool, UK
- 33 ¹¹⁰Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital,
34 Bergen, Norway

1 ¹¹¹Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic
2 of Korea

3 ¹¹²Department of Statistics, Seoul National University, Seoul, Republic of Korea

4 ¹¹³Department of Functional Genomics, University of Amsterdam, Amsterdam, The
5 Netherlands

6 ¹¹⁴Department of Clinical Genetics, VU Medical Centre, Amsterdam, The Netherlands

7 ¹¹⁵Department of Child and Adolescent Psychiatry, Erasmus University Medical Centre,
8 Rotterdam, The Netherlands

9 ¹¹⁶Department of Psychiatry, University of Toronto, Toronto, Canada

10 ¹¹⁷Department of Laboratory Medicine, University of California, San Francisco, CA, USA

11 ¹¹⁸Blood Systems Research Institute, San Francisco, CA, USA

12 ¹¹⁹Department of Human Genetics, McGill University, Montreal, Canada

13 ¹²⁰Department of Medicine, McGill University, Montreal, Canada

14 ¹²¹McGill University and G_ome Qu_bec Innovation Centre, Montreal, Canada

15 ¹²²Department of Twin Research and Genetic Epidemiology, King's College London,
16 London, UK

17 ¹²³Saw Swee Hock School of Public Health, National University of Singapore, Singapore,
18 Singapore

19 ¹²⁴Department of Medicine, National University of Singapore, Singapore, Singapore

20 ¹²⁵Cardiovascular & Metabolic Disorders Program, Duke-NUS Graduate Medical School
21 Singapore, Singapore, Singapore

22 ¹²⁶Department of Biological Sciences, Columbia University, New York, NY, USA

23

24

25

Figures

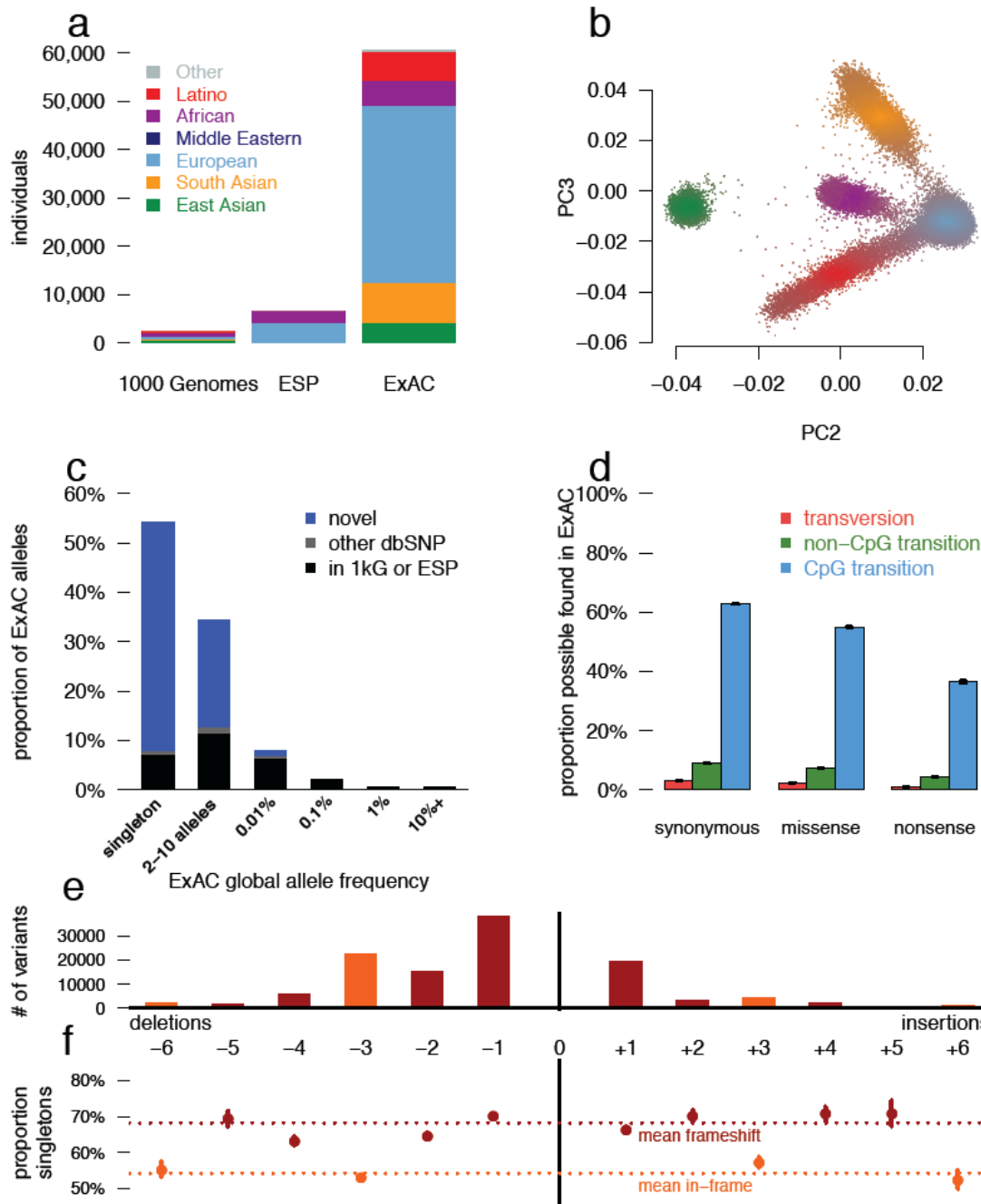
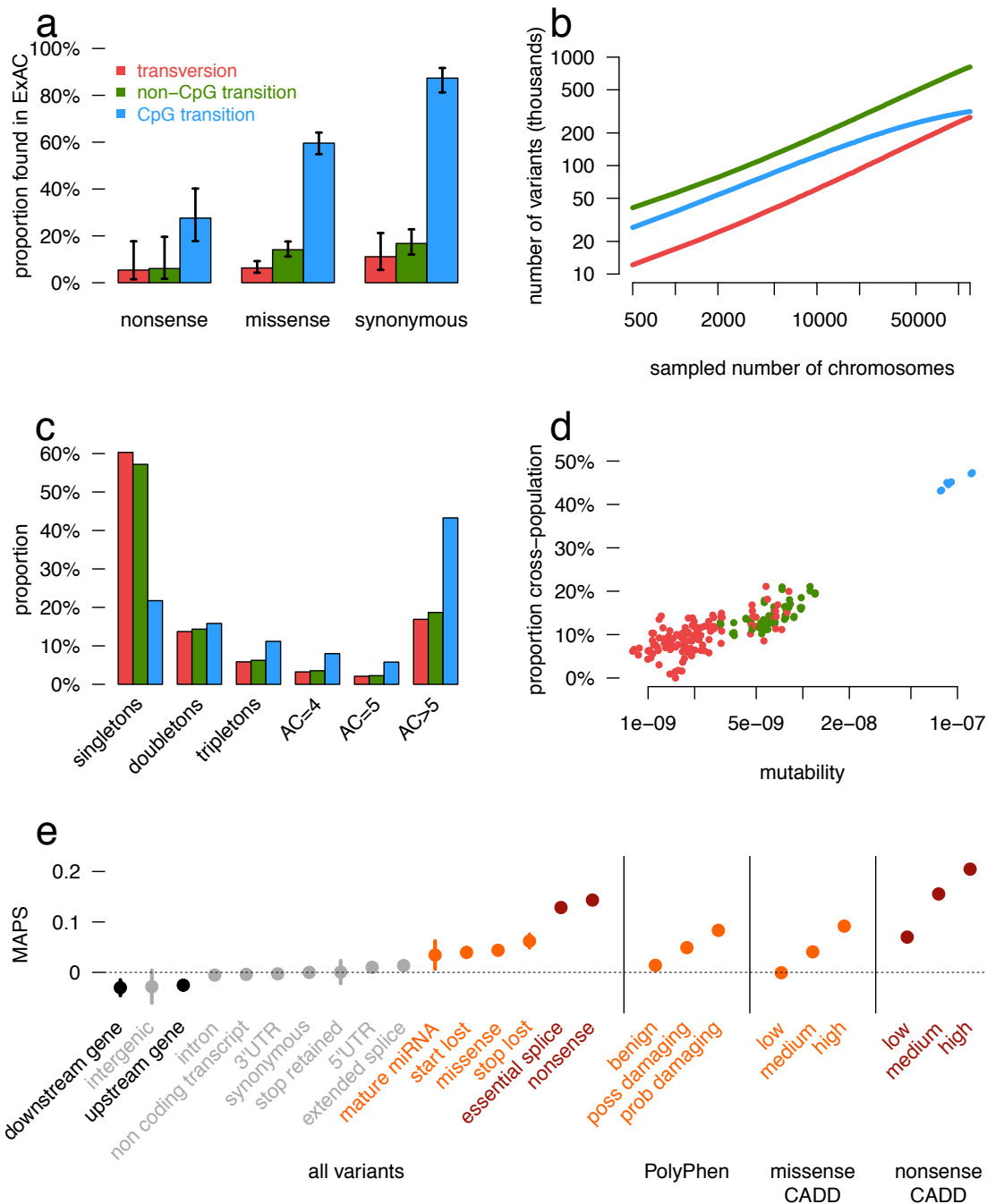


Figure 1. Patterns of genetic variation in 60,706 humans.

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

1 variants are rare and novel. d) The proportion of possible variation observed by mutational
2 context and functional class. Over half of all possible CpG transitions are observed. Error bars
3 represent standard error of the mean. e-f) The number (e) and frequency distribution (proportion
4 singleton; f) of indels, by size. Compared to in-frame indels, frameshift variants are less common
5 (have a higher proportion of singletons, a proxy for predicted deleteriousness on gene product).
6 Error bars indicate 95% confidence intervals.
7
8

1



2

3 Figure 2. Mutational recurrence at large sample sizes.

4 a) Proportion of validated *de novo* variants from two external datasets that are independently
5 found in ExAC, separated by functional class and mutational context. Error bars represent
6 standard error of the mean. Colors are consistent in a-d. b) Number of unique variants observed,
7 by mutational context, as a function of number of individuals (down-sampled from ExAC). CpG
8 transitions, the most likely mutational event, begin reaching saturation at ~20,000 individuals. c)
9 The site frequency spectrum is shown for each mutational context. d) For doubletons (variants

1 with an allele count of 2), mutation rate is positively correlated with the likelihood of being found in
2 two individuals of different continental populations. e) The mutability-adjusted proportion of
3 singletons (MAPS) is shown across functional classes. Error bars represent standard error of the
4 mean of the proportion of singletons.

5

6

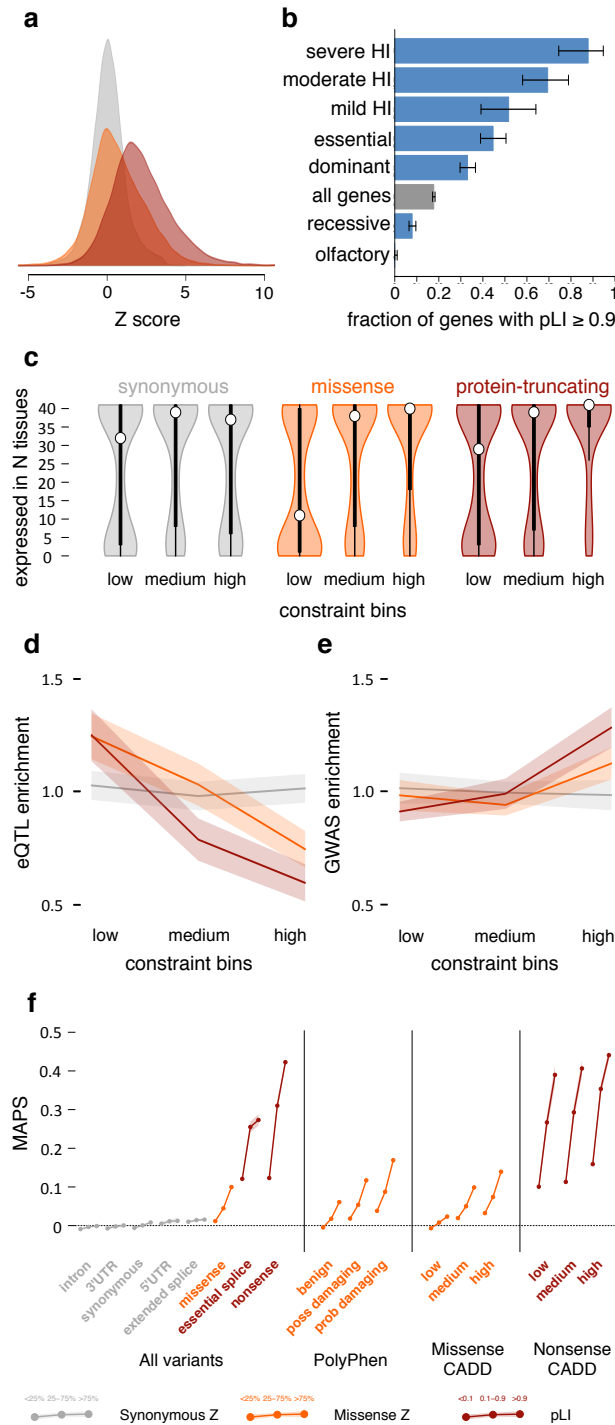
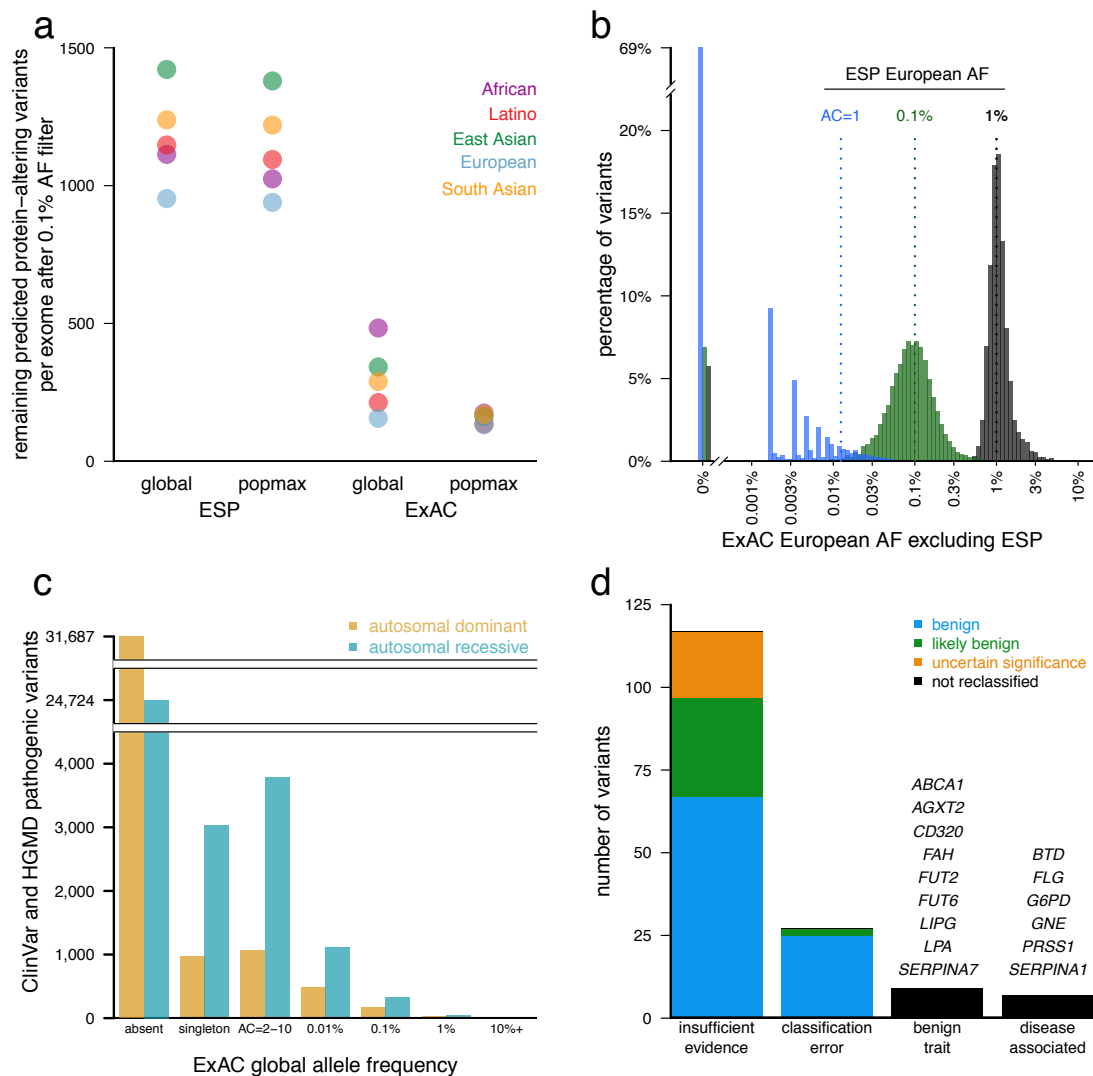


Figure 3. Quantifying intolerance to functional variation in genes and gene sets.

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 ($pLI \geq 0.9$) 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.

1



2

3 **Figure 4. Filtering for Mendelian variant discovery.**

4 a) Predicted missense and protein-truncating variants in 500 randomly chosen ExAC individuals
5 were filtered based on allele frequency information from ESP, or from the remaining ExAC
6 individuals. At a 0.1% allele frequency (AF) filter, ExAC provides greater power to remove
7 candidate variants, leaving an average of 154 variants for analysis, compared to 1090 after
8 filtering against ESP. Popmax AF also provides greater power than global AF, particularly when
9 populations are unequally sampled. b) Estimates of allele frequency in Europeans based on ESP
10 are more precise at higher allele frequencies. Sampling variance and ascertainment bias make
11 AF estimates unreliable, posing problems for Mendelian variant filtration. 69% of ESP European
12 singletons are not seen a second time in ExAC (tall bar at left), illustrating the dangers of filtering
13 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 genes²⁷. Most are not found in ExAC; however, many of the reportedly 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 or South Asian population allele frequency confirmed there is insufficient evidence for pathogenicity for the majority of these variants. Variants were reclassified by ACMG guidelines²³.

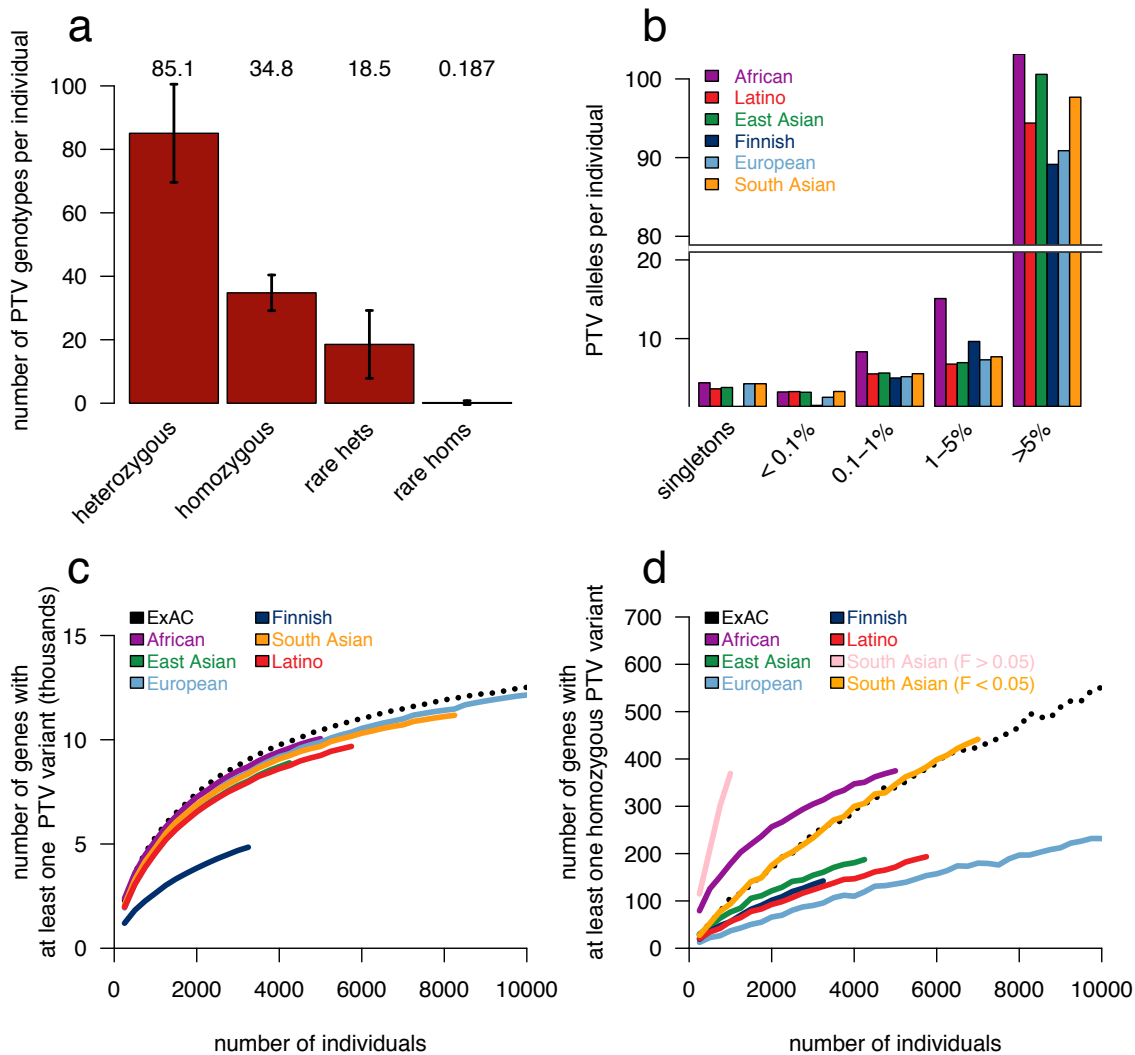
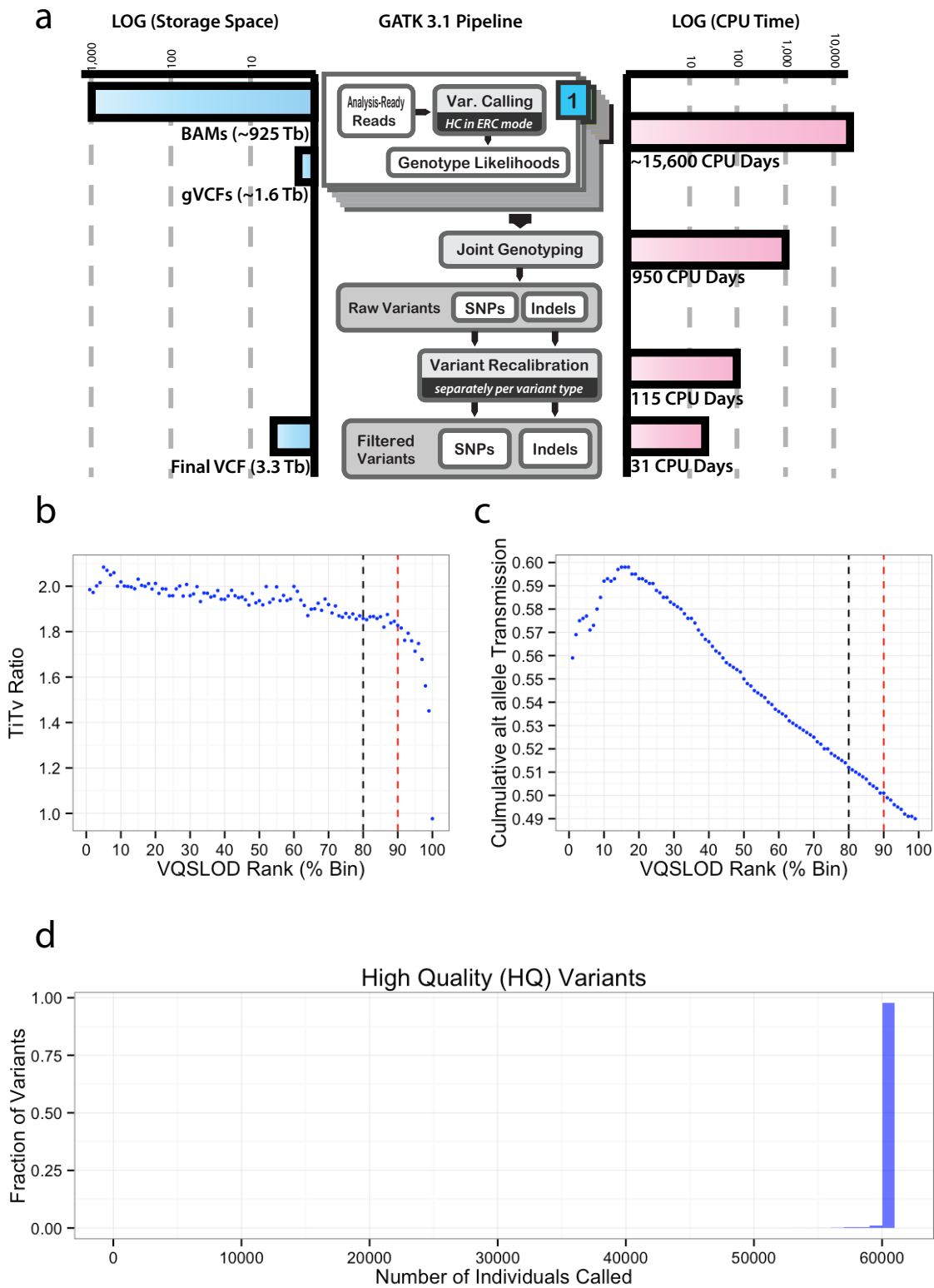


Figure 5. Protein-truncating variation in ExAC.

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. South Asian population is broken down by consanguinity (Inbreeding coefficient, F).

1



2

3 Extended Data Figure 1 The GATK 3.1 pipeline used for the joint calling of 91,796 exomes.

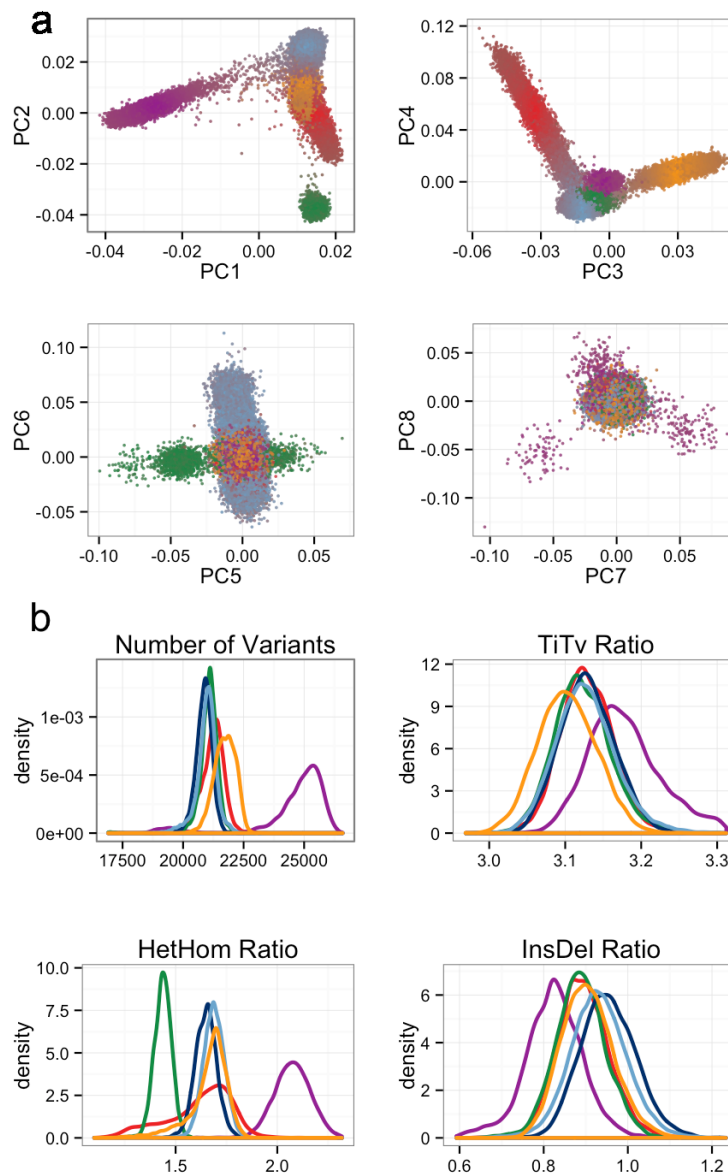
1 a) The resources used for the variant calling in terms of CPU days and storage (terabytes). b)
2 The impact of VQSLOD on singleton TiTv. c) The impact of VQSLOD on singleton transmission in
3 trios. Note: In b) and c) singleton variants discovered in joint called set was ordered by VQSLOD
4 in descending order (i.e. higher confident variants first) and then binned into percentiles. The
5 black dotted line indicates the current VQSR cut off and the red dotted line is where the less
6 stringent threshold was moved. d) The number of individuals called at each variant site as a
7 fraction of the total number of High Quality (HQ) variants.

8

9

10

1



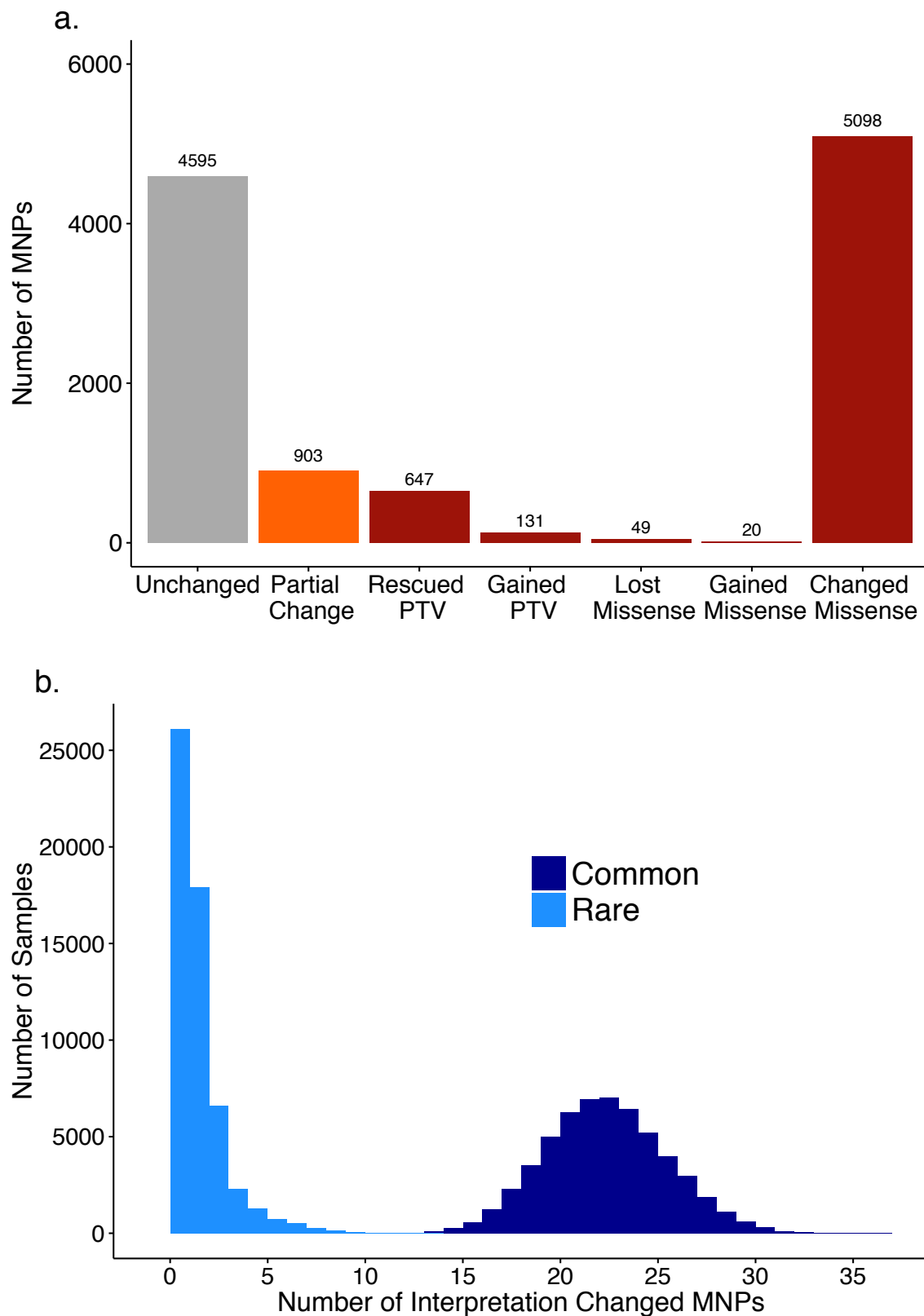
2

3 **Extended Data Figure 2. Principal component analysis (PCA) and key metrics used to filter**
 4 **samples.**

5 a) Principal component analysis using a set of 5,400 common exome SNPs. Individuals are
 6 colored by their distance from each of the population cluster centers using the first 4 principal

1 components. b) The metrics number of variants, TiTv, alternate heterozygous/homozygous
2 (HetHom) ratio and Insertion/Deletion (InsDel) ratio. Populations are their respective colors:
3 Latino (red), African (purple), European (blue), South Asian (yellow) and East Asian (green).
4
5
6

1



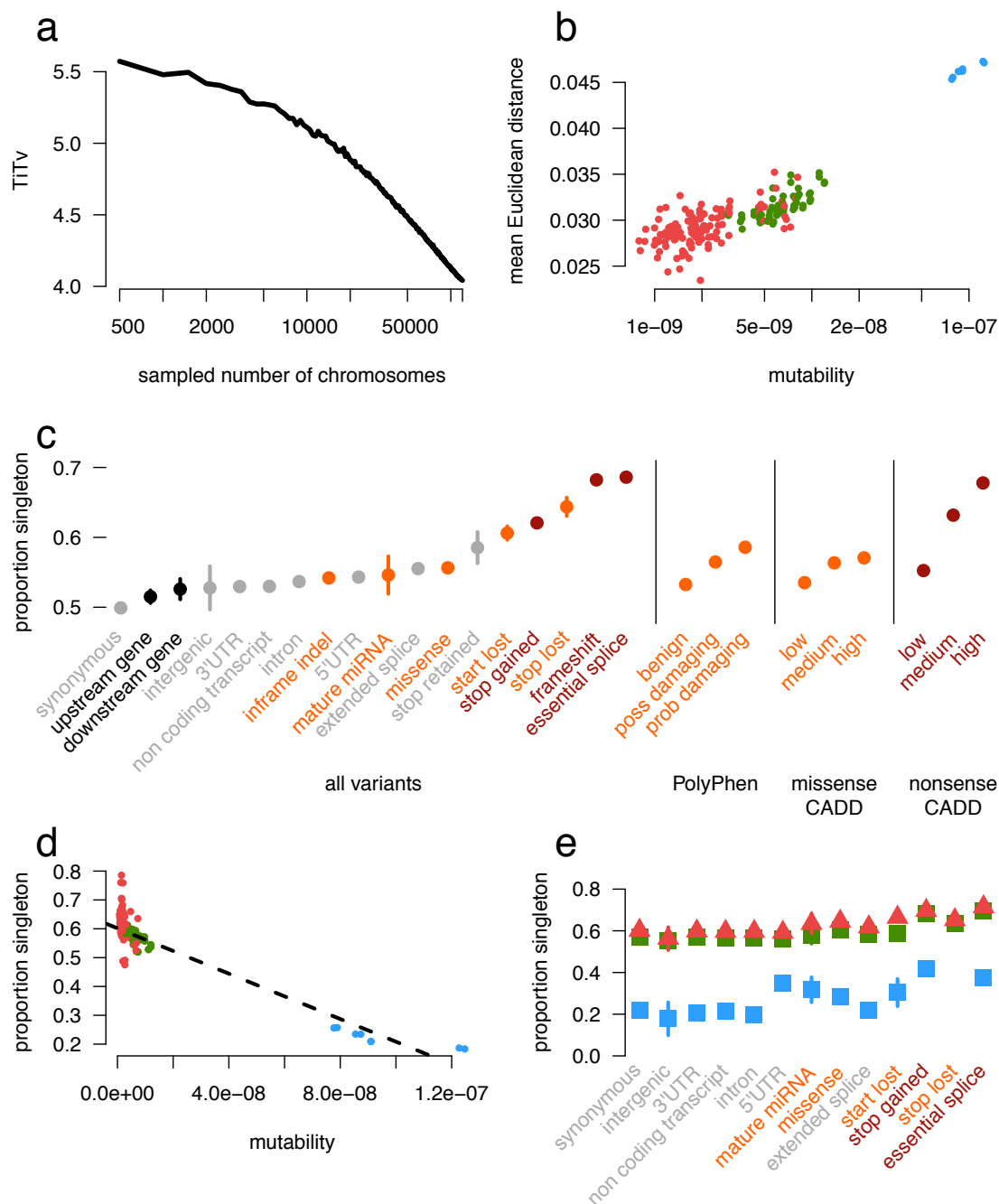
2

3 Extended Data Figure 3. Multi-nucleotide variants discovered in the ExAC data set.

- 1 a) Number of MNPs per impact on the variant interpretation. b) Distribution of the number of
- 2 MNPs per sample where phasing changes interpretation, separated by allele frequency. Common
- 3 > 1%, Rare < 1%. MNPs comprised of a rare and common allele are considered rare as this
- 4 defines the frequency of the MNP.

5
6
7

1



2

3 **Extended Data Figure 4. The impact of recurrence across different mutation and functional**
4 **classes.**

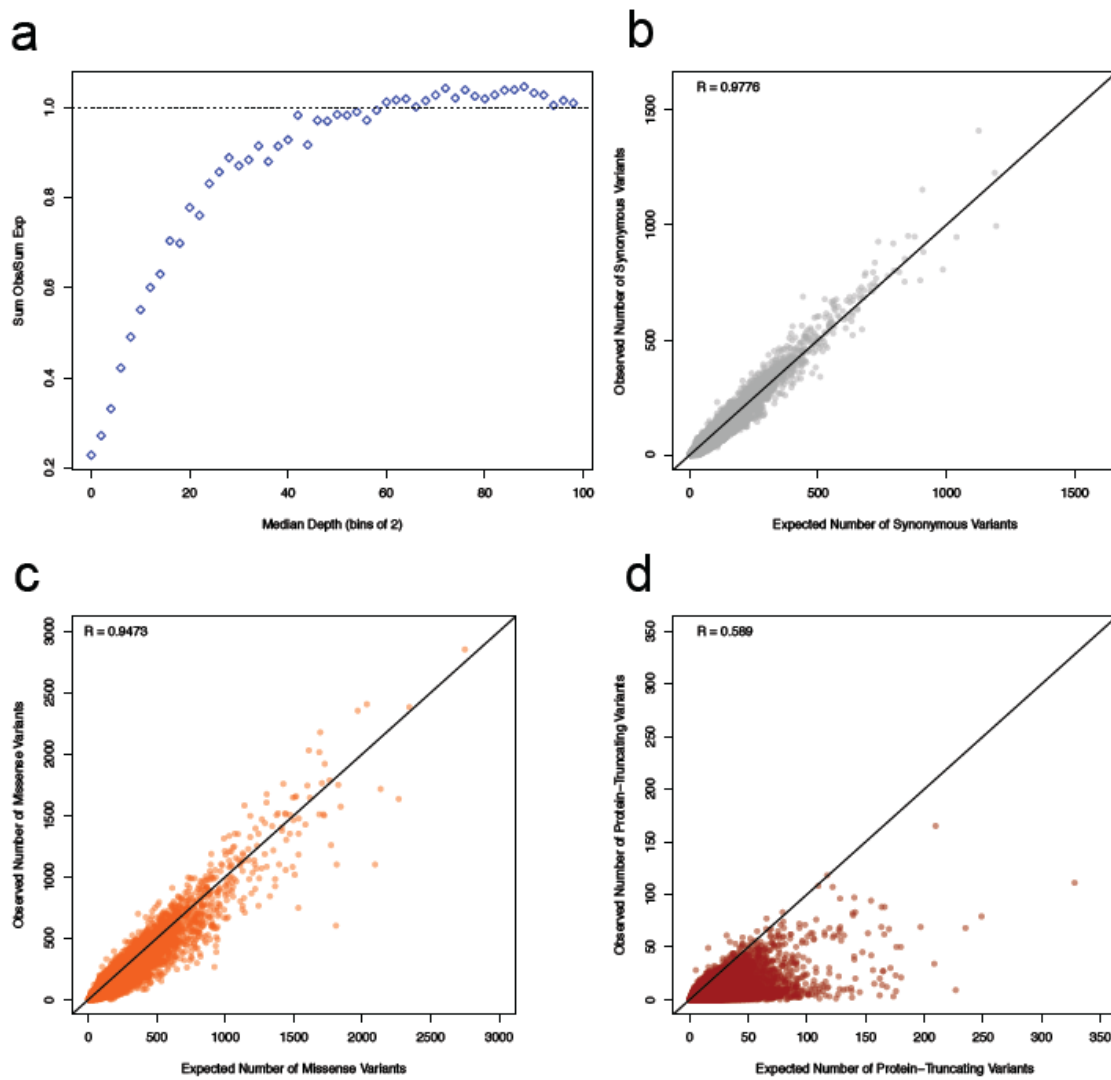
5 a) TiTv (Transition to transversion) ratio of synonymous variants at downsampled intervals of
6 ExAC. The TiTv is relatively stable at previous sample sizes (<5000) but changes drastically at
7 larger sample sizes. b) For synonymous doubleton variants, mutability of each trinucleotide
8 context is correlated with mean Euclidean distance of individuals that share the doubleton.
9 Transversion (red) and non-CpG transition (green) doubletons are more likely to be found in

1 closer PCA space (i.e. more similar ethnicities) than CpG transitions (blue) c) The proportion
2 singleton among various functional categories. The functional category stop lost has a higher
3 singleton rate than nonsense. Error bars represent standard error of the mean. d) Among
4 synonymous variants, mutability of each trinucleotide context is correlated with proportion
5 singleton, suggesting CpG transitions (blue) are more likely to have multiple independent origins
6 driving their allele frequency up. e) The proportion singleton metric from c) broken down by
7 transversions, non-CpG transitions, and CpG variants. Notably, there is a wide variation in
8 singleton rates among mutational contexts in functional classes, and there are no stop-lost CpG
9 transitions. Error bars represent standard error of the mean.

10

11

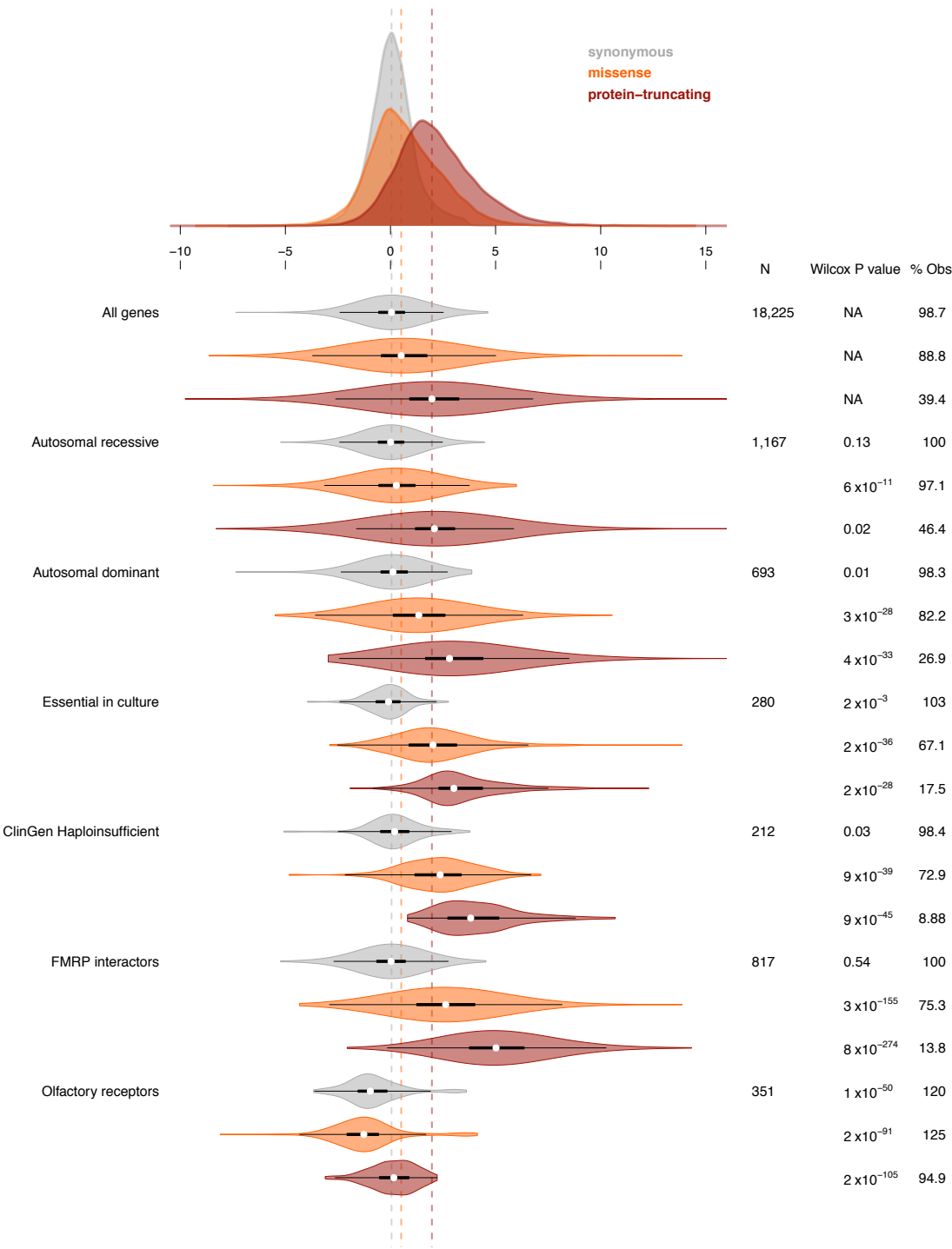
1



Extended Data Figure 5. Relationships between depth and observed vs expected variants as well as correlations between observed and expected variant counts for synonymous, missense, and protein-truncating..

a) The relationship between the median depth of exons (bins of 2) and the sum of all observed synonymous variants in those exons divided by the sum of all expected synonymous variants. The curve was used to determine the appropriate depth adjustment for expected variant counts. For the rest of the panels, the correlation between the depth-adjusted expected variants counts and observed are depicted for synonymous (b), missense (c), and protein-truncating (d). The black line indicates a perfect correlation (slope = 1). Axes have been trimmed to remove *TTN*.

1

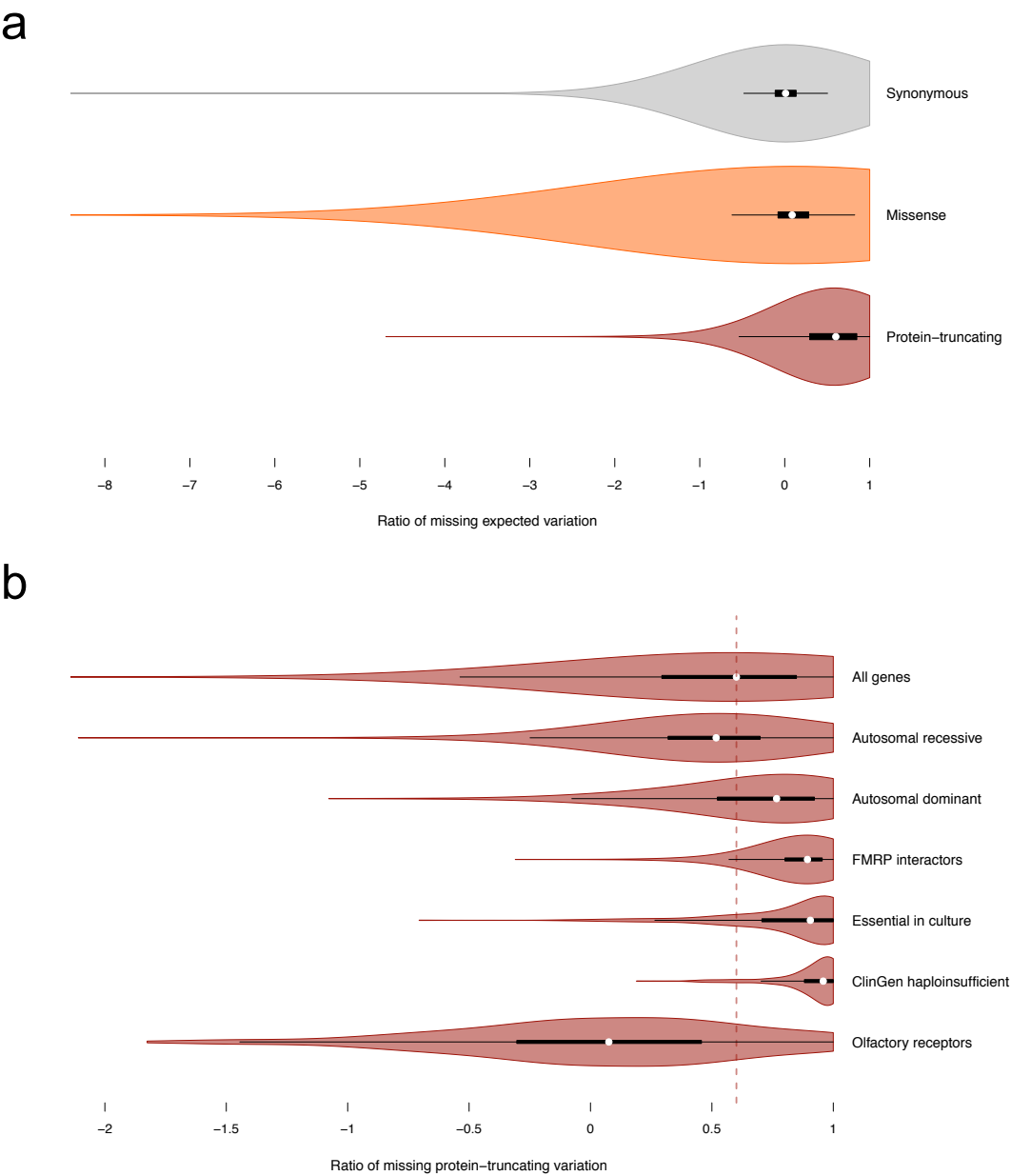


2

3 **Extended Data Figure 6. Distribution of synonymous, missense, and protein-truncating Z**
4 **scores for gene sets.**

- 1 The number of genes in the set, the Wilcoxon p-value for the difference from the full distribution,
- 2 and the percentage of expected variation observed are reported on the right. Thick black bars
- 3 indicate the first to third quartiles, with the white circle marking the median.
- 4
- 5

1



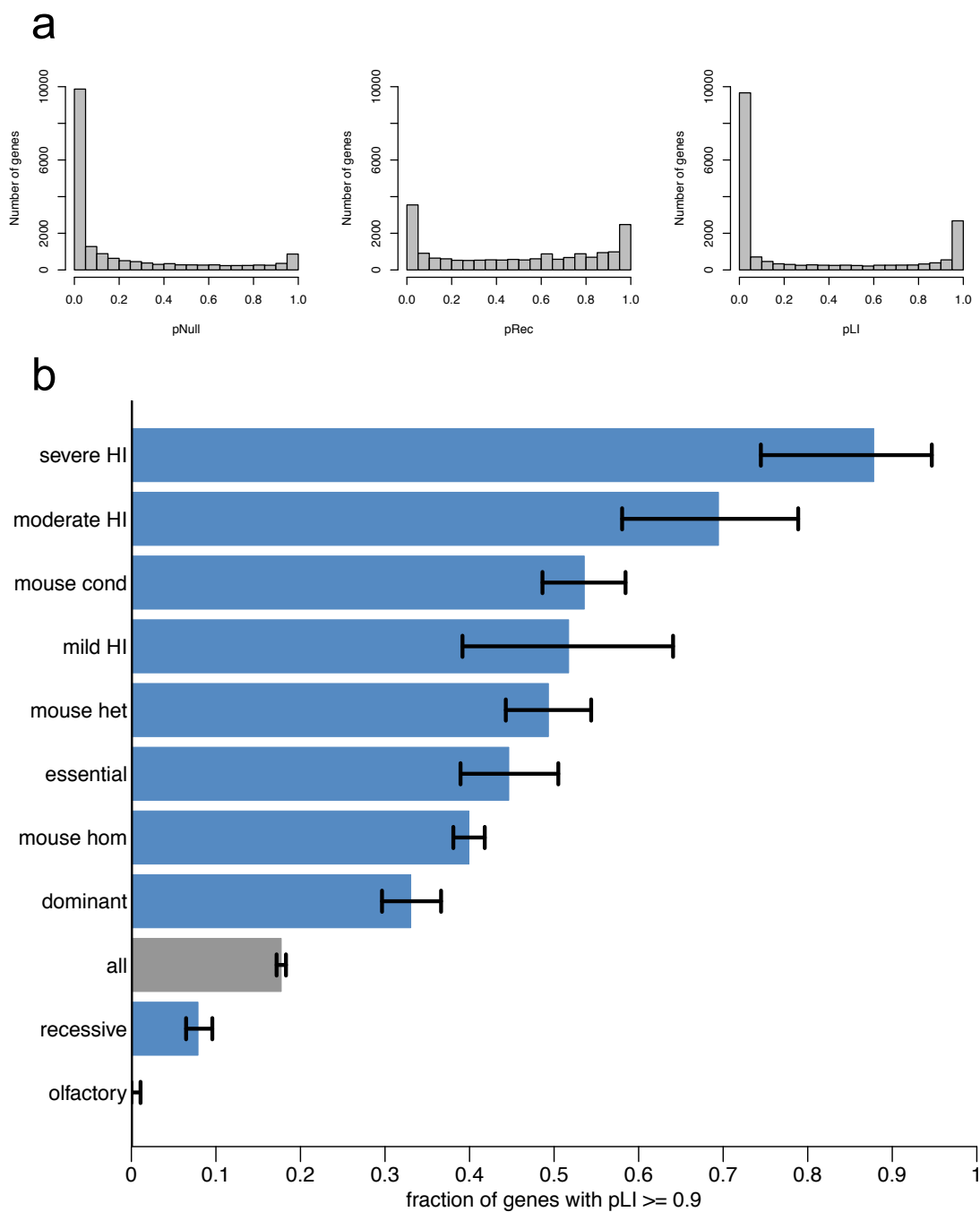
2

3 **Extended Data Figure 7. Ratio of missing synonymous, missense and protein truncating**
4 **variation.**

1 a) The distribution of the ratio of missing expected variation for synonymous, missense, and
2 protein-truncating as well as for gene sets of interest. Note that 1 means there were no variants
3 observed and negative values indicate more variation observed than expected. b) The median
4 ratio of missing protein-truncating variation for all transcripts is indicated by the dashed maroon
5 line. For a, the x-axis has been trimmed at -8 (out of -18) to highlight the patterns of the data.
6 Similarly, x-axis in the bottom panel has been trimmed at -2 (out of -5) to highlight the patterns of
7 the data. Thick black bars indicate the first to third quartiles, with the white circle marking the
8 median.

9
10

1

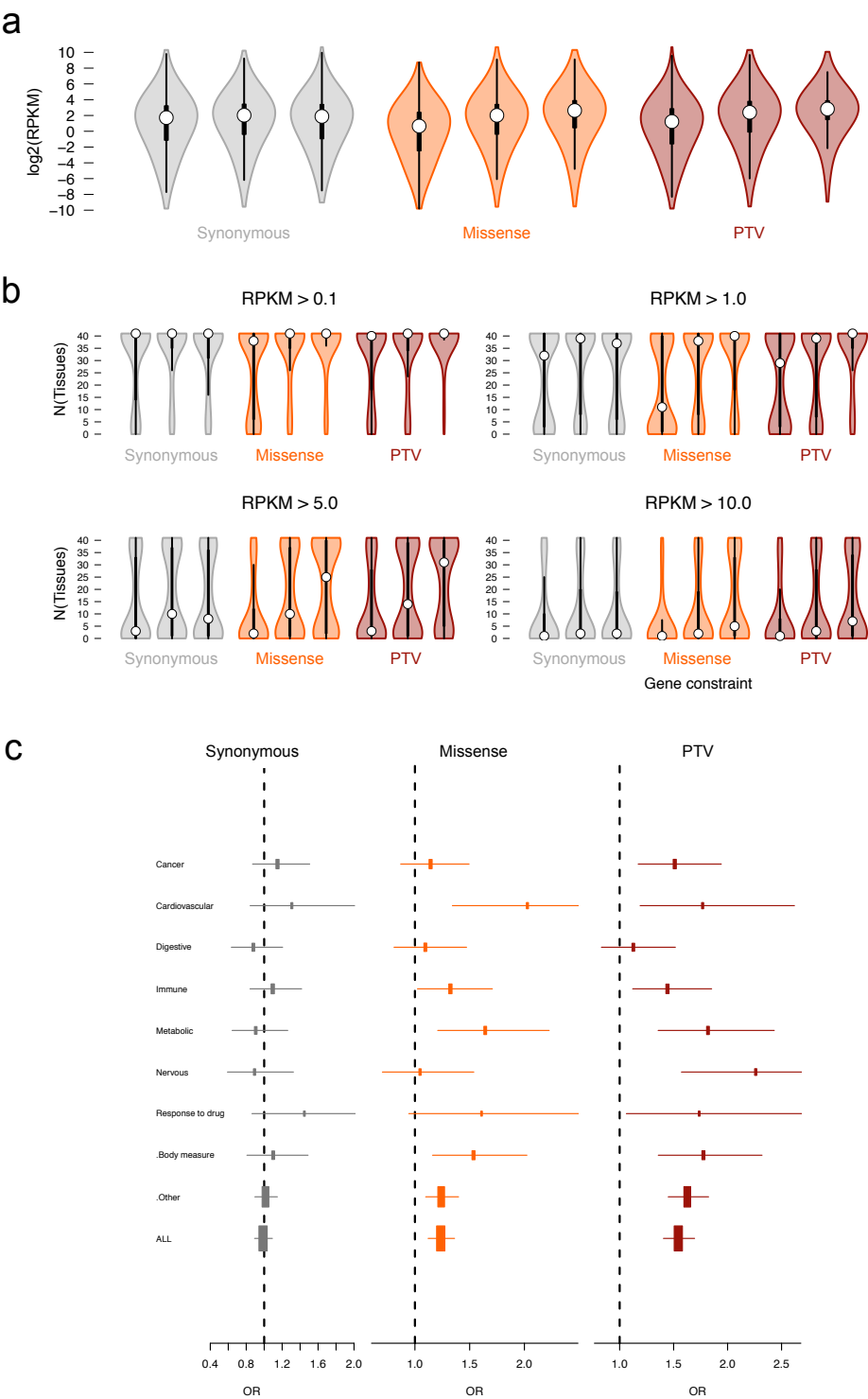


2

3 **Extended Data Figure 8. The distribution of pNull, pRec, and pLI across all transcripts and**
 4 **the fraction of genes in a gene set with pLI ≥ 0.9.**

a) The distributions of pNull, pRec, and pLI for all canonical transcripts. The distribution is roughly bimodal for each. pLI close to one indicates extreme intolerance to loss-of-function variation; we therefore take $pLI \geq 0.9$ as the cut-off for extreme loss-of-function intolerance and depict, in b, the fraction of genes from gene sets of interest that have $pLI \geq 0.9$. The black error bars indicate a 95% confidence interval. olfactory = olfactory receptor genes (n=371); recessive = recessive disease genes from Blekhman and Berg (n=1,183); all (n=18,225); dominant = dominant disease genes from Blekhman and Berg (n=709); mouse hom = genes that are lethal in mice when both copies are knocked out (n=2,760); essential = genes that are essential in cell culture as curated by Hart et al 2014 (n=285); mouse het = genes that are lethal in mice when one copy is knocked out (n=387); mild HI = haploinsufficient genes that cause a mild disease (n=59); mouse cond = genes that are lethal in mice when conditionally knocked out in adult mice (n=402); moderate HI = haploinsufficient genes that cause moderately severe disease (n=77); severe HI = haploinsufficient genes that cause severe disease (n=44). Please refer to Supplementary Table 10 for more details on gene lists.

1



2

3

Extended Data Figure 9. Application of pLI on RNA-Seq data and GWAS hits.

4

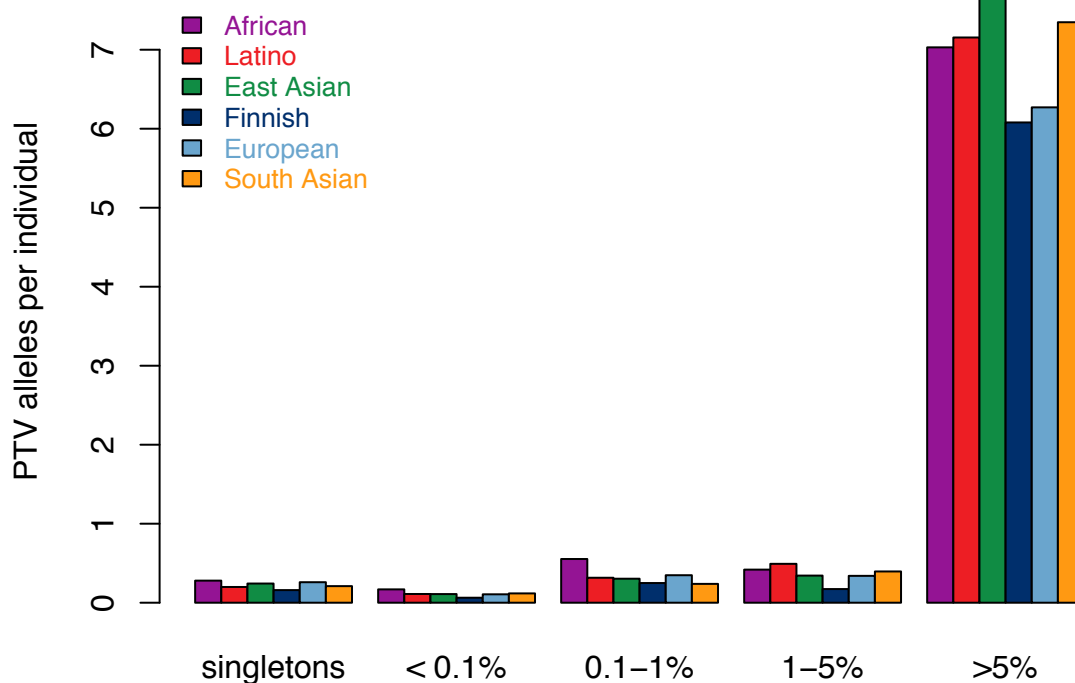
a) The relationship between constraint and median gene expression across all tissues. b) The

5

relationship between constraint and tissue expression at different RPKM cutoffs. Thick black bars

- 1 indicate the first to third quartiles, with the white circle marking the median. c) The odds ratio of
- 2 being a GWAS hit for each Experimental Factor Ontology trait for the most constrained genes vs
- 3 the middle bin. The error bars indicate a 95% confidence interval.
- 4
- 5

1



2

3

Extended Data Figure 10. Number of protein-truncating variants in constrained genes per individual by allele frequency bin.

4

5

Equivalent to Figure 5b limited to constrained ($pLI \geq 0.9$) genes.

6

7

8

9