Supplementary Information

1 Data Generation

Authors: Monkol Lek, Eric Banks, Tim Fennell, Konrad J. Karczewski, Fengmei Zhao, Eric V. Minikel, Mark J. Daly and Daniel G. MacArthur

1.1 Alignment and BAM processing

Alignment

The sequencing reads (i.e. fastq files) from exomes were aligned to the human genome reference (hg19) using bwa (v0.5.9) on a per lane basis.

```
bwa aln Homo_sapiens_assembly19.fasta -q 5 -l 32 -k 2 -t $NSLOTS -o 1 \
-f $output.1.sai $input.1.fastq.gz

bwa aln Homo_sapiens_assembly19.fasta -q 5 -l 32 -k 2 -t $NSLOTS -o 1 \
-f $output.2.sai $input.2.fastq.gz

bwa aln Homo_sapiens_assembly19.fasta -q 5 -l 32 -k 2 -t $NSLOTS -o 1 \
-f $output.unpaired.sai $output.unpaired.fastq.gz

bwa sampe -t $NSLOTS -T -P -f $output.aligned_bwa.sam Homo_sapiens_assembly19.fasta \
$output.1.sai $output.2.sai $input.1.fastq.gz $input.2.fastq.gz
```

Read duplicate marking

PCR duplicate reads were then marked for each lane level BAM using Picard MarkDuplicates. The lane level BAMs were then aggregated into a single BAM with lane level data represented as separate read groups (@RG) and Picard MarkDuplicates was applied again.

Insertion/Deletion Realignment

The re-alignment intervals for each BAM was determined using GATK RealignerTargetCreator and a list of known indel sites. Using this interval list, local realignment was then performed by GATK IndelRealigner.

Base Quality Recalibration

The base quality scores were then recalibrated using GATK BaseRecalibrator and a list of known variant sites. The new base quality scores were then applied using GATK PrintReads but retaining the original base quality scores within the BAM.

1.2 BAM quality control

The following Picard tools were used to generate various sample metrics:
CollectOxoGMetrics
CollectSequencingArtifactMetrics
CalculateHsMetrics
CollectAlignmentSummaryMetrics
CollectInsertSizeMetrics

The VerifyBAMID (v1.0.0) tool was used as an estimate of sample contamination using a set of 93,102 variants.

```
verifyBamID --verbose --ignoreRG --vcf ExomeContam.vcf --out $outfile --bam $input.bam
```

Samples that were outliers for key metrics or had contamination estimate (i.e. FREEMIX) > 0.075 were excluded from variant calling. In total **91,796 sample BAMs** were used for variant calling.

1.3 Variant calling

Single sample variant discovery

The Genome Analysis Toolkit (GATK) v3.1 (v3.1-144) HaplotypeCaller algorithm was used to generate gVCFs for all 91,796 BAMs across a defined exome interval set totaling 59,880,539 bp and known sites were annotated with dbSNP135 (Extended Data Figure 1a).

```
java -jar GenomeAnalysisTK.jar \
-T HaplotypeCaller --disable_auto_index_creation_and_locking_when_reading_rods \
-R Homo_sapiens_assembly19.fasta \
-o $output.vcf.gz \
-I $input.bam \
-L $input.intervals \
--minPruning 3 --maxNumHaplotypesInPopulation 200 -ERC GVCF \
--max_alternate_alleles 3 -variant_index_parameter 128000 \
-variant_index_type LINEAR -contamination 0.0
```

Joint Genotyping

The sample gvcfs were combined into 298 groups with sqrt(n) samples in each group.

```
java -jar GenomeAnalysisTK.jar \
```

```
-T CombineGVCFs --disable_auto_index_creation_and_locking_when_reading_rods \
-R Homo_sapiens_assembly19.fasta \
-o $output.vcf.gz \
-V gvcf.list \
--sample_rename_mapping_file rename_alias_file.txt
```

Joint genotyping was then performed using the 298 grouped gvcf files as input.

```
java -jar GenomeAnalysisTK.jar \
-T GenotypeGVCFs --disable_auto_index_creation_and_locking_when_reading_rods \
-R Homo_sapiens_assembly19.fasta \
-o $output.unfiltered.vcf.gz \
-D Homo_sapiens_assembly19.dbsnp.vcf \
-L Input.intervals \
-V all_combined_gvcfs.list
```

1.4 Variant recalibration and filtering

GATK Variant Quality Score Recalibration (VQSR) was used to filter variants. The SNP VQSR model was trained using HapMap3.3 and 1KG Omni 2.5 SNP sites and a 99.6% sensitivity threshold was applied to filter variants, while Mills et. al. 1KG gold standard and Axiom Exome Plus sites were used for insertions/deletion sites (Supplementary Information Table 1) and a 95.0% sensitivity threshold was used.

```
#Strip to sites only
java -jar MakeSitesOnlyVcf.jar \
INPUT=$input.unfiltered.vcf.gz \
OUTPUT=$output.sites_only.unfiltered.vcf.gz
```

```
#Perform SNP VQSR on sites only file
java -jar GenomeAnalysisTK.jar \
-T VariantRecalibrator --disable_auto_index_creation_and_locking_when_reading_rods \
-R Homo_sapiens_assembly19.fasta \
-input $input.sites_only.unfiltered.vcf.gz \
--num_threads 2 -recalFile $output.snps.recal \
-tranchesFile $output.snps.tranches \
-allPoly -tranche 100.0 -tranche 99.95 -tranche 99.9 -tranche 99.8 -tranche 99.6 \
-tranche 99.5 -tranche 99.4 -tranche 99.3 -tranche 99.0 -tranche 98.0 -tranche 97.0 \
-tranche 90.0 \
-an QD -an MQRankSum -an ReadPosRankSum -an FS -an MQ -an InbreedingCoeff \
-resource:hapmap,known=false,training=true,truth=true,prior=15 hapmap_3.3.b37.vcf.gz \
-resource:omni,known=false,training=true,truth=true,prior=12 1000G_omni2.5.b37.vcf.gz \
-resource:1000G,known=false,training=true,truth=false,prior=10
1000G_phase1.snps.high_confidence.b37.vcf.gz \
```

```
-resource:dbsnp137,known=false,training=false,truth=false,prior=7 dbsnp_138.b37.vcf.gz \
-resource:dbsnp129,known=true,training=false,truth=false,prior=3 \
dbsnp_138.b37.excluding_sites_after_129.vcf.gz \
--maxGaussians 6 -mode SNP \
-rscriptFile $output.snps.recalibration_plots.rscript

#Perform INDEL VQSR on sites only file
java -jar GenomeAnalysisTK.jar \
-T VariantRecalibrator --disable_auto_index_creation_and_locking_when_reading_rods \
-R Homo_sapiens_assembly19.fasta \
-input $input.sites_only.unfiltered.vcf.gz \
--num_threads 2 -recalFile $output.indels.recal \
-tranchesFile $output.indels.tranches \
-allPoly -tranche 100.0 -tranche 99.95 -tranche 99.9 -tranche 99.5 -tranche 99.0 -tranche
97.0
-tranche 96.0 -tranche 95.0 -tranche 94.0 -tranche 93.5 -tranche 93.0 -tranche 92.0 -
tranche 91.0 -tranche 90.0 \
-an FS -an ReadPosRankSum -an InbreedingCoeff -an MQRankSum -an QD \ -
resource:mills,known=false,training=true,truth=true,prior=12
Mills_and_1000G_gold_standard.indels.b37.vcf.gz \
-resource:axiomPoly,known=false,training=true,truth=false,prior=10
Axiom_Exome_Plus.genotypes.all_populations.poly.vcf.gz \
-resource:dbsnp137,known=true,training=false,truth=false,prior=2 dbsnp_138.b37.vcf.gz \ -
-maxGaussians 6 -mode INDEL \
-rscriptFile $output.indels.recalibration_plots.rscript

#Apply SNP VQSR on genotypes file (SNP sensitivity 99.6)
java -jar GenomeAnalysisTK.jar \
-T ApplyRecalibration --disable_auto_index_creation_and_locking_when_reading_rods \
-R Homo_sapiens_assembly19.fasta \
-o $onput.filtered.vcf.gz \
-input $input.filtered.vcf.gz \
-L $input.intervals \
-recalFile $input.exomes.snps.recal \
-tranchesFile $input.snps.tranches -ts_filter_level 99.6 -mode SNP

#Apply Indel VQSR on genotypes file (INDEL sensitivity 95.0)
java -jar GenomeAnalysisTK.jar \
-T ApplyRecalibration --disable_auto_index_creation_and_locking_when_reading_rods \
-R Homo_sapiens_assembly19.fasta \
-o $output.indels.filtered.vcf.gz \
-input $input.unfiltered.vcf.gz \
-L $input.intervals \
-recalFile $input.indels.recal \
-tranchesFile $input.indels.tranches -ts_filter_level 95.0 -mode INDEL
```

1.5 Assessment of variant filtering

To assess the relationship between Variant Quality Score Log Odds (VQSLOD) and various metrics, we rank ordered all variants in descending order and then binned into percentiles such that the 1% VQSLOD bin corresponds to the top 1% VQSLOD scoring variants. The VQSLOD score is known to have a bias towards common and well behaved variants (i.e. training data), thus the 99.6% sensitivity threshold used for filtering is achieved at much lower rank in larger exome data sets, when all variants are rank ordered by VQSLOD in descending order. The standard VQSR filtering results in ~20% of bi-allelic singleton SNPs to be filtered and ~60% of bi-allelic singleton Indels.

In order to reduce the number of singleton SNPs filtered, we reduced the stringency moving the threshold from VQSLOD=-1.8251 to -2.632 with minimal impact on singleton transition to transversion (TiTv) ratio and reducing the number of singleton variants filtered (Extended Data Figure 1b).

The adjustment from standard VQSR filtering was assessed using singleton transmission. We used 490 trios to assess the transmission of singletons from parents, where a detected transmission to the child is observed as a doubleton. Relaxing the variant filters resulted in an improved average transmission rate from 51.2% to 50.1% (17,867 aggregated transmissions, 17,825 aggregated non-transmissions) (Extended Data Figure 1c). In contrast, the singleton Indel PASS cut off was not adjusted after reviewing singleton insertion/deletion ratio and transmission.

One additional round of site filtering was performed to filter sites with inbreeding coefficient (InbreedingCoeff) < -0.2 to remove sites with excess heterozygous individuals and sites with AC (allele count) = 0, which occurs when remaining genotype calls in the release subset does not meet minimum quality threshold of DP >= 10 and GQ >= 20.

1.6 Data quality evaluation

NA12878

The NA12878 exome included in the release call set was evaluated against the NIST v2.18 consensus call set[1] to estimate per sample sensitivity and specificity. The analysis was performed within the exome target interval set and excludes regions known to contain segmental duplications, copy number variations and short tandem repeats. In total there were 13 SNP variants discovered that were absent in NIST v2.18. The 5 SNP variants with allele frequency (AF) < 0.1% (including 2 singletons) had read support in high coverage PCR-Free WGS of NA12878 and were not considered false positive sites. The SNP and Indel false discovery rate (FDR) for NA12878 was 0.056% and 2.17%, respectively. This evaluation interval set was also used to determine the overall SNP and Indel sensitivity of 99.8% and 95.1%, respectively.

PCR-Free Whole Genome Sequencing

An analysis of false discovery per sample was performed on 13 exomes included in the joint called set and that also have high coverage PCR-Free WGS (dbGAP Accession phs000655.v1.p1). The 13 exomes selected for evaluation are all Non-Finnish European and representative in terms of coverage and number of singleton variants discovered (Suplementary Information Table 6). Among the 13 exomes, 8 were from unrelated parents in trios and singleton counts needed to be adjusted. There are three samples (with no related sample in the larger call set) that have higher than expected number of singletons and are likely due to lower representation in European sub-populations. The WGS call set and underlying reads were used to determine the overall SNP and Indel FDR of 0.14% and 4.71%, respectively. The overall FDR across functional annotation classes missense, synonymous and protein truncating variants (including indels) were 0.076%, 0.055% and 0.471% (Supplementary Information Table 5). The per sample FDR by allele frequency and functional annotation are summarized in Supplementary Information Table 6. An additional evaluation on a set of 16 exomes from the 1000 Genomes project with PCR-Free WGS[2] resulted in an overall SNP and Indel FDR of 0.19% and 3.14%, respectively.

Illumina Human Exome Bead Chip Concordance

An overlapping set of 10,650 samples with exome chip genotyping were used to assess overall genotyping accuracy of heterozygous calls on 100,285 overlapping PASS sites. The overall heterozygous concordance was 99.6%. The singleton calls had a

concordance of 97.1% (104/107). The three discordant singleton calls had low mapping quality and/or poor allele balance.

Validated De-novo variants

A set of 699 MiSeq validated de-novo variants detected in samples included in the joint called set was used to assess sensitivity of filters applied. Prior to filtering 698/699 (99.9%) variants were re-discovered, while ~90% of variants remained after filtering was applied.

1.7 Sample Quality Control and selection

A set of 5,400 common SNPs that did not overlap with indels, were selected from 5,800 SNPs[3] and samples with outlier heterozygosity were removed before principal component analysis (PCA) was performed. Samples were clustered into population groups based on principal components 1-4 (PC1-4) and colored by their distance from each of the major population clusters (Extended Data Figure 2a). PCA was performed again on the samples within the European cluster to determine Finnish and non-Finnish Europeans.

Ancestry groups for the world (Figure 1a) are approximate and are based on continental populations, with adjustments for distinguishing South Asian from East Asian ancestry. South America and Mexico are specified as 'Latino.' North America and Europe are 'European' (data from Wikipedia 'World Populations' page, accessed 2/10/15).

The GATK VariantEval tool was used to calculate the per sample number of variants, transition/transversion (TiTv) ratio, alternate allele heterozygous/homozygous (Het/Hom) ratio and insertion/deletion (indel) ratio (Extended Data Figure 2b) and outlier samples within each population were removed. The minor population differences observed for TiTv is largely due to an increased exome capture heterogenity/differences within South Asian and African samples. In contrast, the Het/Hom population differences are due to the decreased heterozygosity observed, correlated to the distance from Africa and consistent with progressive rounds of migration of small sub-populations into a region[4]. These differences have been observed in previous population studies[5]. Lastly, the

differences in indel ratios are due to alignment to a European human draft reference instead of an ancestral reference, differences in insertion and deletion mutation rates and population demography[6].

The relatedness of samples were inferred using KING[7] and samples with degree of relatedness of 1 or 2 were excluded to produce a list of unrelated samples using the --unrelated option. The sex of each sample were inferred using heterozygosity of common variants on chromosome X and chromosome Y coverage (normalized to chromosome 20 coverage) and outliers excluded. Lastly, samples with a high fraction of genotype quality of zero (GQ=0) sites were excluded. Cohorts and sequencing projects with a large number of sample outliers were removed entirely.

After sample quality control, samples from cohorts with severe pediatric disease were excluded and samples absent of consent or data usage permissions required for public release of frequency data. The remaining samples from the cohorts/consortium are listed in Supplementary Information Table 2 and Supplementary Information Table 3. The resulting samples (n=60,706) were extracted from the larger set (n=91,796) using GATK SelectVariants. The subsetting performed by SelectVariants excludes non-variant sites and trims alleles.

1.8 Exome definition and coverage

The exome or exome calling intervals used by variant calling (Section 1.3) contains the Gencode v19 coding intervals and flanking 50 bases (~60 Mb). The 60,706 samples consists of ~77% of Agilent (33 Mb target) and ~12% of Illumina (37.7 Mb target) exome captures. We used a strict definition of coverage, where paired reads both need to have mapping quality (MQ) ≥ 20 and base quality (BQ) ≥ 10 and coverage capped at 250x. The target coverage for the exomes was 80% of targeted bases > 20x, which corresponds to ~65x mean coverage. A per base coverage (capped at a maximum of 100x) was determined for 10% of samples across exome calling intervals using a modified version of samtools coverage (Supplementary Information Table 8).

1.9 Variant site summary annotation

A custom GATK Annotation walker was developed to provide summary metrics in the INFO field of the VCF call set (Supplementary Information Table 4). The sex chromosome metrics were adjusted for the inferred sex, allowing for hemizygous counts for males on chromosome X and Y and variant calls ignored for females on chromosome Y. A custom Python script was developed to summarize genotype quality (GQ) and depth (DP) distributions at each variant site (Supplementary Information Table 8).

1.10 Functional annotation

Variant annotation was performed using Variant Effect Predictor (VEP)[8] version 81 was used with Gencode v19 on GRCh37. Protein-truncating variant (PTV) annotation was performed using LOFTEE (version 0.2; Loss-of-function Transcript Effect Estimator,Supplementary Information Table 8), a plugin to VEP. LOFTEE considers all stop-gained, splice-disrupting, and frameshift variants, and filters out many known false-positive modes, such as variants near the end of transcripts and in non-canonical splice sites, as described in the code documentation. VEP is used to determine Ensembl Gene ID and gene symbol, as well as Ensembl Transcript ID (and whether the transcript is canonical), for each functional consequence of the variant, and the PolyPhen2 and SIFT scores. CADD is annotated courtesy of Martin Kircher and Jay Shendure (Supplementary Information Table 8).

1.11 Universe of possible SNVs

We created a synthetic VCF with every possible single nucleotide substitution in the ExAC calling intervals using a custom Python script (Supplementary Information Table 8). These variants were annotated in the identical fashion as the ExAC VCF as in Section 1.10, and merged with coverage data as described in Section 1.8. Both VCFs were subsetted to well-covered sites, or sites where at least 80% of individuals with at least 10X coverage. These files were used to compute the breakdown of observed (ExAC) and possible (synthetic) variants by functional class and mutational class (Supplementary Information Table 9).

1.12 High Quality (HQ) variants

Variant sites were considered high-quality if they met the following criteria: (1) they were given a PASS filter status by VQSR (see above), (2) at least 80% of the individuals in the dataset had at least depth (DP) >= 10 and genotype quality (GQ) >= 20 (i.e. AN_Adj >= 60706*0.8*2 or 97130), (3) there was at least one individual harboring the alternate allele with depth >= 10 and GQ >= 20, and (4) the variant was not located in the 10 1-kb regions of the genome with the highest levels of multi-allelic (quad-allelic or higher) variation. The application of this criteria, subsequently excludes all variants in chromosome X and Y except for the pseudoautosomal regions.

1.13 Data sets and summary of various filters

The subsequent analysis in Supplementary Sections 2-6 use high quality variants unless explicitly specified. In contrast, the quality control analysis was performed on various intermediate and unreleased data sets to assess thresholds and their consequences. This has been summarized in Supplementary Information Table 7.

1.14 Data practices and statistical analysis

For all subsequent analyses, allele frequencies (global and population-specific) are calculated based on adjusted allele counts (i.e. only including individuals with depth (DP) >= 10 and genotype quality (GQ) >= 20). The “popmax” frequency is calculated as the highest frequency with this adjusted frequency across all populations (max(AC_Pop/AN_Pop) for all “Pop”s, as above). Sequence context for all SNVs is determined using a custom VEP plugin (Supplementary Information Table 8) to determine CpG transitions and mutational contexts.

All statistical analyses and plots were done using R 3.2.2. Correlations were performed using the Pearson correlation coefficient (r) unless otherwise specified.

2 Multi-nucleotide Polymorphism (MNPs)

Authors: Andrew Hill, Beryl Cummings, Konrad J. Karczewski, Monkol Lek and Daniel G. MacArthur

In order to locate phased MNPs across the ExAC dataset without phasing information included as part of the original variant calling, we implemented a post-hoc phasing approach. Phasing post-hoc is computationally expensive over a large number of samples, so we employed the following three pass approach to find a small set of candidate MNPs for which phase information was needed as the final step of calling.

First Pass

We used the ExAC sites VCF and associated Variant Effect Predictor (VEP) annotations to carry out a first round of MNP calling that identifies coding variants that fall within the same annotated codon on at least one transcript. Only combinations of two or three SNPs were considered. Combinations of SNPs matching these criteria were identified as candidate MNPs. For each candidate MNP identified, we also extracted several relevant annotations such as allele frequency information, VEP annotations, and the codon change produced by the MNP as determined by combining codon changes of the individual mutations.

Second Pass

We extracted sample genotypes for SNPs identified in first step using tabix and combined the sample genotypes with the annotation provided in the ExAC sites VCF. We performed a second pass of MNP calling using the same criteria as the first, but with an additional check to find samples that have the corresponding alternative genotype for all SNPs composing the candidate MNP.

Third Pass

Finally, we extracted only sites that passed the second round of calling, again using tabix, and combined them into a single file. This file was then split into subsets of 10 samples each using GATK SelectVariants in order to reduce phasing runtime. Each of these subsets was used as input to GATK ReadBackedPhasing, along with the corresponding BAM files for the samples contained in each subset. Use of ReadBackedPhasing, which uses read-based support for variant phasing is appropriate

in this case because variants in the same codon will almost always be spanned by the same read. ReadBackedPhasing reports phasing blocks for variants and places a PhasingInconsistent flag in the INFO field of the VCF if phasing support is poor for a given site. Homozygous sites are not phased by ReadBackedPhasing as phasing is guaranteed for homozygous variants.

A third and final round of calling was performed that was the same as the second, but with an additional check for matching phasing blocks for all the SNPs composing each candidate MNP (in the case of heterozygous variants) for each sample. MNPs marked with PhasingInconsistent by ReadBackedPhasing were flagged, allowing them to be filtered from the final set of MNPs. Results of calling on all sample subsets were aggregated to form a final set of MNPs for all samples.

2.1 MNP filtering and QC

Each unique genomic coordinate with a multi-nucleotide polymorphism was filtered according to average coverage in the ExAC callset (≥20), mapping quality (>55), VQSR filtering (all variants within MNP with PASS) and phasing consistency (PhasingInconsistent Flag via ReadBackedPhasing). MNPs that result in a gain or loss of a nonsense variant were further filtered to those designated high confidence via LOFTEE with no PhyloCSF flags. This left 11,443 MNP sites with an average of 45 MNPs per sample.

To ensure MNP discovery was not inflated by a particular population or cohort, we analyzed the distribution of MNPs within each subgroup. The number of MNPs discovered were correlated ($r > 0.9$) with the number of samples available for each population and cohort, indicating MNP discovery was not confounded by these factors. Similarly, the number of MNPs discovered per chromosome was correlated with the number of targeted intervals per chromosome in ExAC.

2.2 Random Inspection of Phasing

A large subset of MNPs (~1,000) for several samples were examined manually in IGV to look for support of annotation and phasing. Each MNP examined showed appropriate

read support for phasing and annotation within the same codon, indicating that our method was performing well.

2.3 Categorization of MNPs

In order to break down the overall set of MNPs into categories of their effect on variant interpretation, we defined the following two sets of categories.

The first set of categories is intended to define the functional outcome of MNPs:
Gained protein-truncating variant (PTV): Neither of the individual SNPs is a nonsense/stop-gained mutation, but the MNP is.
Rescued PTV: at least one of the individual SNPs is a nonsense/stop-gained mutation, but the MNP is not.
Gained Missense: the individual SNPs are synonymous, but the MNP results in a missense variant.
Lost Missense: at least one of the individual SNPs is a missense variant, but the MNP is synonymous.
Changed Missense: at least one of the individual SNPs is a missense variant and the MNP is a new missense variant with a different resulting amino acid.
Partially Changed Missense: The MNP is composed of two different missense variants that when considered together have the same amino acid outcome as only one of the variants (e.g. Missense A + Missense B = Missense A).
Unchanged: Either the outcome of the MNP is identical to that of the individual SNPs or one of the SNPs is a synonymous variant that does not change the outcome of an adjacent non-synonymous variant.

The second set of categories groups the first set based their broader impact on variant interpretation.
Interpretation Incorrect: contains *Gained PTV*, *Rescued PTV*, *Gained Missense*, *Lost Missense*, and *Changed Missense*. Defines cases where phasing is critical for functional interpretation as analysis of the SNPs in isolation will be incorrect without the context provided by identifying MNPs.

Interpretation Incomplete: contains *Partially Changed Missense*. Defines MNPs that may have a different interpretation depending on the nature of the analysis and which variants are prioritized during analysis.
I**nterpretation Unchanged:** contains *Unchanged*. Defines cases where local phasing has no immediate impact on functional interpretation.

Using the above classification we were able to determine the number of variants in which the functional interpretation changes (Extended Data Figure 3a) and the impact on a per-individual level (Extended Data Figure 3b). We also analyzed the consequences of misinterpreting MNPs in known disease-associated genes (Supplementary Information Table 10) and at reported pathogenic variants (Supplementary Information Table 11).

3 Recurrence

Authors: Konrad J. Karczewski, Jack A. Kosmicki, Eric V. Minikel, Kaitlin E. Samocha, Daniel P. Birnbaum, Daniel G. MacArthur and Mark J. Daly

Many population genetics models based on the infinite sites model[9] assume that if a variant is observed twice, the two observations are a result of identity by descent. However, as the number of sequenced individuals grows, the probability of observing two or more independent mutational events occurring at the same site in the genome increases. Indeed, recurrent mutations have been associated with many (primarily mendelian) diseases, including multiple endocrine neoplasia type 2B[10], achondroplasia[11], Apert syndrome[12], and progressive myoclonus epilepsy[13]. Additionally, recurrent mutations have been described in dominant diseases such as Huntington's disease[14] and Creutzfeldt-Jakob disease[15], and as risk factors for psychiatric diseases[16]. However, to our knowledge, this phenomenon has not been systematically quantitated at a global scale. Here, we describe an analysis of widespread mutational recurrence observed in exome sequence data from 60,706 individuals from the Exome Aggregation Consortium (ExAC). This effect is most pronounced among highly mutable CpG transitions, and in this dataset, we begin to saturate all possible synonymous CpG mutations.

The transition/transversion ratio (TiTv) is a metric commonly used in large-scale genomics datasets to assess quality of a variant callset. This metric has an expected value based on known mutational contexts of the genomic region sequenced: in a single human individual, a whole genome sequence (WGS) callset is expected to have a TiTv of ~2.1, while variants from whole exome sequence (WES) are expected to have TiTv of ~3, depending on the specific regions targeted. However, in the limit of sequencing all individuals in the world (~7 billion), each with a set of ~100 *de novo* variants in addition to inherited variants, we expect to saturate all possible single nucleotide variants in the genome (~9 billion), except those incompatible with life. In this scenario, the TiTv ratio for this callset will be ~0.5 (1 transition and 2 transversions for every site). Thus far, most TiTv metrics have been determined for relatively small (<5000 individual) datasets, where changes in TiTv are relatively small across sample sizes (Extended Data Figure 4a). In this section, we discuss multiple lines of evidence that demonstrate the presence of mutational recurrence, and attempt to quantify this phenomenon.

3.1 Recurrence among validated *de novo* variants

In order to directly observe mutational recurrence, we explored whether validated *de novo* variants from external datasets are also found in ExAC, which must indicate two separate mutational origins. Here, we downloaded a collection of 2156 validated *de novo* variants in 1750 parent-offspring trios from two studies: Deciphering Developmental Delay (DDD)[17] (1531 *de novo* variants, of which 1516 are validated, in 1133 probands) and schizophrenia[18] (640 *de novo* in 617 probands). We then calculated the rate of recurrence between the set of *de novo* variants and ExAC, only considering a variant recurrent if the chromosome, position, reference allele, and alternate allele were identical in the *de novo* dataset and ExAC. Both *de novo* studies were chosen based on the lack of ascertainment with respect to which *de novo* variants were validated (i.e. other studies validate a subset of identified *de novo* variants, usually likely protein truncating variants).

Across all 2156 *de novo* variants, 582 (26.99%) variants were present in ExAC. However, across various functional and mutational categories, there is a wide variation in *de novo* recurrence rate, from 5.4% of all stop-gained transversions to 87.3% of all synonymous CpG transitions (Figure 2a). This indicates that mutational contexts as well as selective pressures are crucial for the understanding of mutational recurrence.

3.2 Evidence for recurrence within ExAC

With the 1,798,481 synonymous variants in ExAC, we begin to observe saturation of CpG transitions, but not transversions or other transitions, as the same sites are observed with the same mutations from various sources. Additionally, we explored the frequency at which a synonymous doubleton variant is observed in one population (more likely to be identity-by-descent) or two populations (more likely to be independent mutational origins). Here, we observe that approximately 45% of CpG transitions occur in individuals from two separate multiple populations (Figure 2d), compared to 11% of transversions and 15% of non-CpG transitions. A similar trend is observed when considering the mean Euclidean distance in PCA space between the pair of individuals

that share the variant: the mutability of each context is correlated with the mean distance between individuals ($r = 0.88$; $p < 10^{-50}$; Extended Data Figure 4b).

As selection acts on particular variant classes, alleles will be purged from the population and so, their site frequency spectrum (SFS) will be shifted towards rare variation. The SFS can thus be used to provide information as to the deleteriousness of variant classes (Figure 1d and 1f). To compare multiple variant classes, we initially explored a "proportion singleton" metric (Extended Data Figure 4c), similar to those adopted in previous work[19].

However, at this rare ($AF < 10^{-4}$) frequency range, recent mutational events (and thus, the mutation rate) will have a stronger influence on the SFS (Figure 2b). Indeed, we observe a strong negative correlation between the proportion singleton and mutation rate for each mutational context among synonymous variants ($r = -0.9$; $p < 10^{-50}$; Extended Data Figure 4d).

As mutation rates vary widely among mutational classes (transitions, transversions, CpGs), the distribution of these variant types among functional classes (synonymous, missense, etc) will bias the proportion singleton metric. For instance, as a stop lost variant cannot be the result of a CpG transition, its proportion singleton rate will be inflated, as non-CpG variants are more likely to be singletons. In Extended Data Figure 4e, we show that various functional categories show differential proportion singleton rates for various mutation types.

We thus propose a correction to the percent singleton metric: we regressed out the expected proportion singleton for each functional class, using a linear model trained on synonymous variation weighted by number of observations in each mutational context. We term this metric MAPS (mutability-adjusted proportion of singletons), which correlates with deleteriousness based on biological expectations (Figure 2e).

4 Constraint

Authors: Kaitlin E. Samocha, Taru Tukiainen, Konrad J. Karczewski, Karol Estrada, James Zou, Eric V. Minikel, Daniel G. MacArthur and Mark J. Daly

4.1 Establishing the expected number of variants per gene

Probabilities of a mutation

Our metrics to evaluate a gene's intolerance to variation—their level of constraint—rely on comparing the observed variant counts to an expectation. In order to determine the expected number of variants per gene, we modified a previous method described in detail in Samocha et al (2014)[20]. This method begins with the creation of a mutation rate table, which provides the probability of each trinucleotide (XY_1Z) mutating to all other possible trinucleotides (XY_2Z). The mutation rate table is then used to determine the probability of all possible coding mutations as well as mutations in the conserved splice site bases. The result of the method is a per-gene (or per-exon) probability of mutation split by mutational class (synonymous, missense, nonsense, and splice site).

As in Samocha et al (2014)[20], we adjust the probabilities of mutation for regional divergence between humans and macaques. To do so, we create a divergence score, which is the number of divergent sites between the two species divided by the number of screened sites, for the gene as well as 1 MB upstream and downstream. A linear model is used to define the equation used to adjust the probabilities of mutation by the divergence score. More details can be found in the supplement of Samocha et al (2014)[20].

Two major changes were made between the previous version of the method and the one used in this paper: (1) we use GENCODE v19 annotations for transcripts instead of Refseq and (2) the expected number of variants, and not the probability of mutation, is adjusted for depth of sequencing coverage (see below). For this paper, we focus on the canonical transcript as defined by Ensembl v75 for each protein-coding gene and drop all transcripts that do not begin with a methionine, end with a stop codon, or whose length are not divisible by three. After all of these filters, there are 19,620 canonical transcripts that are used in all following analyses.

Determining the depth of coverage correction

We extract the number of rare (minor allele frequency < 0.1%) single nucleotide variants for every exon of the canonical transcripts and assign functional classes (synonymous, missense, nonsense, and splice site) based on the amino acid change or position in the splice site. We then need a way to account for the depth of sequencing coverage since regions that are poorly sequenced will, by definition, have fewer variants than expected. To do this, we determine the median depth of coverage for each exon. Given that synonymous variants are most likely to be free of extreme negative selection, we focus on those variants. Using only those exons with a median depth ≥ 50, which we consider to be well sequenced, we regress the number of rare synonymous variants on the probability of a synonymous mutation to determine the appropriate formula to predict the number of expected synonymous variants. This formula is applied to all exons (regardless of depth). To find the appropriate way to correct for sequencing coverage, we group exons by depth (bins of 2) and determine the sums of all observed and expected synonymous variants in these exons. The sum of observed synonymous variants divided by the sum of expected variants has a logarithmic relationship between depth bins of 0 and 50, where it then plateaus at ~1 (Extended Data Figure 5a). We fit the curve to determine the appropriate depth of coverage correction for exons with a median depth between 1 and 50.

$$depth\ adjusted\ count = \begin{cases} expected\ count, median\ depth\ \geq 50 \\ expected\ count * (0.089 + 0.217 * \ln(median\ depth), 1 \leq\ median\ depth < 50 \\ 0.089 * expected\ count, median\ depth < 1 \end{cases}$$

Expected number of variants

To determine the depth-corrected expected number of variants per exon, we use those exons with a median depth ≥ 50 and regress the number of rare synonymous variants on the probability of a synonymous mutation. These regressions are done separately for the autosomes with the pseudo-autosomal regions (PAR) of the X chromosome, the non-PAR regions of the X chromosome, and the Y chromosome. The resulting formulas are used to predict the depth-uncorrected expected number of synonymous, missense, and protein-truncating variants (PTVs; nonsense and splice site) variants for all exons. The correlation between the observed and depth-uncorrected expected number of synonymous variants per exon is 0.8360. We then correct these expected numbers by the above equation and observe an increased correlation between observed and depth-

corrected expected synonymous variants (r = 0.9283). Note that from this point forward, the expected number of variants always refers to the depth-corrected counts.

4.2 Creation of the constraint metric

Determining Z scores of the deviation of observation from expectation

We create a signed Z score to establish the significance of the deviation of observed variant counts per gene from expectation as in Samocha et al 2014[20]. To start, we sum all exon level variant counts across canonical transcripts. Here, the observed count is the number of unique variants with a VQSLOD ≥ -2.632 and 123 or fewer alternative alleles (minor allele frequency cut off of ~0.1%). If an exon has a median depth < 1, the variant counts for that exon are not included in the total for the transcript. We then remove all transcripts where no variants are observed. For the remaining 18,466 transcripts, we calculate the chi-squared value for the deviation of observation from expectation for each mutational class (synonymous, missense, and protein-truncating). The square root of these values is multiplied by -1 if the number of observed variants is greater than expectation (or 1 if observed counts are smaller than expected) to create the Z score.

A critical next step is to correct the Z scores so that the synonymous Z scores followed an approximately normal distribution. For the synonymous Z scores, we use a subset of transcripts whose synonymous Z scores fall in between -5 and 5. All synonymous Z scores are divided by the standard deviation of this outlier-removed subset to create the corrected Z scores.

A slightly different approach is used for missense and PTV Z scores. We take all transcripts with a missense Z score between -5 and 0 and combine them with those same Z scores multiplied by -1 (to create a mirrored distribution). All missense Z scores are divided by the standard deviation of the mirrored distribution to create the corrected missense Z scores. The same procedure is applied to the PTV Z scores.

Removing outliers

We then use these corrected Z scores to define outlier transcripts—specifically those with significantly elevated synonymous and missense counts or significantly depleted

synonymous and missense counts. These outliers are defined as transcripts with a synonymous Z < -3.71 and a missense Z < -3.09 or transcripts with a synonymous Z > 3.71 and a missense Z > 3.09. These filters remove a total of 241 transcripts, leaving 18,225 for all further analyses. The distribution of the synonymous, missense, and PTV Z scores are depicted in Figure 3a in the main text. Note that a Z score of ~3.09 is equivalent to a p-value of 10^{-3} and is considered the significance threshold when splitting transcripts into constrained and unconstrained classes.

Correlation of observed and expected counts

For the set of 18,225 cleaned transcripts, the correlation between the number of observed rare (minor allele frequency < 0.1%) synonymous variants and the expected number of variants given the above model is 0.9776. This correlation is higher than simply regressing the observed synonymous variants against number of coding bases in the gene (r = 0.9201), or the probability of a synonymous mutation (r =0.9349). This relationship between observed and expected mutation counts can be seen for synonymous, missense, and protein-truncating variants in Extended Data Figure 5b-d.

Power of the Z score analyses

To achieve a Z score of 3.09 (a p-value equivalent of 10^{-3}), the number of expected variants would need to be a minimum of 10. Following this criterion, 99.5% of transcripts could be evaluated for missense constraint. However, only 11,437 transcripts (62.8%) were mutable enough to have 10 or more expected PTVs in the ExAC dataset (see below).

4.3 Z score distributions for gene lists

We next investigate the synonymous, missense, and PTV Z score distributions for the following gene lists: autosomal recessive, autosomal dominant, essential in cell culture, ClinGen haploinsufficient, FMRP interactors, and olfactory receptors (Supplementary Information Table 12). For the synonymous Z scores (Extended Data Figure 6), most gene lists match the distribution of the full set of canonical transcripts (median Z = 0.05). The only notable exception is the list of olfactory receptors, which show 118% of the expected synonymous variation (Wilcoxon $p < 10^{-46}$).

Across all canonical transcripts, ~89% of all missense variation is observed and the median missense Z score is 0.51. As a note, higher (more positive) Z scores indicate increased selective constraint, while negative Z scores are given for transcripts where more variation was seen than expected. All of the gene sets tested significantly differ from the overall distribution (Extended Data Figure 6) with the recessive genes and olfactory receptors showing slightly lower missense Z scores. The rest of the gene sets have significantly higher missense Z scores (Wilcoxon $p < 10^{-28}$).

The PTV Z scores have the most skewed distributions (Extended Data Figure 6). Overall, only 39% of the expected protein-truncating variation is observed, giving the full set of canonical transcripts a median PTV Z score of 1.97. The Z scores for the autosomal recessive genes match the overall distribution fairly closely (Wilcoxon p = 0.02, median = 2.09). The olfactory receptors, as before, have significantly lower PTV Z scores (Wilcoxon $p < 10^{-50}$, median = 0.16), but unlike with synonymous and missense do not have more protein-truncating variation than expected (95% observed).

4.4 Creation of a new protein-truncating constraint score

The PTV Z score is correlated with gene length

The Z scores were created to evaluate the significance of the deviation of observed counts from expectation. Given this, it is sensitive to differences in power. For example, a gene with 0 observed variants would require ~10-11 expected variants to pass a significance threshold of 10^{-3} (Z score of 3.09). The expected number of variants per gene is based on the length and mutability of the transcript. Since the probability of having a protein-truncating mutation is small (roughly an order of magnitude less than the probability of a missense mutation), only 63% of the canonical transcripts are expected to have 10 or more PTVs in the ExAC dataset (59% if expecting 11 PTVs).

Due to this reliance on mutability, it is unsurprising that the PTV Z score is correlated with the coding length of the transcript (r = 0.5697). This correlation is not seen for the missense Z score (r = 0.0566). Therefore, larger transcripts will have more significant PTV deviations (and Z scores) than smaller transcripts and some transcripts that are truly intolerant of loss-of-function variation will be too small to achieve statistical

significance. These results motivated the search for a better metric to capture PTV constraint (discussed below).

Evaluating the ratio of missing protein-truncating variation

A natural metric to evaluate intolerance to protein-truncating variation is the amount of expected variation that was not observed. Truly intolerant transcripts should be missing most of the expected variation, which is independent of the length of the transcript. We defined the ratio of missing variation as one minus the quotient of the observed counts divided by the expected counts.

The correlation between the length of the transcript and the ratio of missing protein-truncating variation is 0.1561. The distributions of the ratio of missing synonymous, missense, and protein-truncating variation are depicted in Extended Data Figure 7a. The majority of transcripts fall between 0 and 1 for the ratio of missing protein-truncating variation, where 1 means the transcript is completely devoid of protein-truncating variation. Both the synonymous and missense distributions shift towards transcripts having more of their expected variation.

The ratio of missing protein-truncating variation is depicted for the gene lists used above in Extended Data Figure 7b. All gene sets are significantly different from the set of all canonical transcripts (referred to as "All genes" in the figure; Wilcoxon $p < 10^{-10}$ for all). Autosomal recessive genes and olfactory receptors have slightly more of their expected protein-truncating variation than the set of all transcripts. The rest of the gene sets are significantly more depleted for the expected protein-truncating variation than the full set of transcripts. The most striking signal comes from the haploinsufficient genes, none of which have more protein-truncating variation than expected.

Creation of pLI

One of the main goals of this work was to identify genes that are intolerant of loss-of-function variation. Given the continuous nature of the ratio of missing protein-truncating variation, it is slightly challenging to do this. To address this challenge, we estimated the probability of being loss-of-function intolerant (pLI) using the expectation-maximization (EM) algorithm.

The underlying premise of this analysis is to assign genes to one of three natural categories with respect to sensitivity to loss-of-function variation: null (where protein-truncating variation – heterozygous or homozygous - is completely tolerated by natural selection), recessive (where heterozygous PTVs are tolerated but homozygous PTVs are not), and haploinsufficient (where heterozygous PTVs are not tolerated). We assume tolerant (null) genes would have the expected amount of truncating variation and then took the empirical mean observed/expected rate of truncating variation for recessive disease genes (0.463) and severe haploinsufficient genes (0.089) to represent the average outcome of the homozygous and heterozygous intolerant scenarios respectively. These values (1.0, 0.463, 0.089) are then used as a three-state model to which we fit the observed/expected truncating variant rate of each gene via the following analysis.

Let $\pi := (\pi_{Null}, \pi_{Rec}, \pi_{HI})$ represent the proportion of all genes that fall into each of the three proposed categories: null, recessive, and haploinsufficient.

Let λ_{Null}, λ_{Rec}, and λ_{HI} denote the expected amount of loss-of-function depletion in each of the three categories. Based on the observed depletion of protein-truncating variation in the Blekhman autosomal recessive and ClinGen dosage sensitivity gene sets (Supplementary Information Table 12), we use:

$\lambda_{Null} = 1$
$\lambda_{Rec} = 0.463$
$\lambda_{HI} = 0.089$

For each gene *i*, We model the observed data (PTV counts) as a function of the unobserved class labels (Z_i) as follows:

$$Z_i \mid \pi \sim Cat(\pi_{Null}, \pi_{Rec}, \pi_{HI})$$

$$PTV_i \mid Z_i \sim Pois(N\lambda_{Z_i})$$

Here, PTV_i represents the observed number of PTVs in gene *i* and *N* is sample size, such that $N\lambda_{Z_i}$ is the expected number of loss-of-function variants in a gene belonging to

class Z_i in the ExAC data. Our goal is to find the maximum-likelihood estimate (MLE) for π (the mixing weights of the three gene classes), and to use this estimate to obtain an Empirical Bayes maximum a posteriori (MAP) estimate for Z_i – the probability of gene assignment to each category – for all genes *i=1…M*.

We use an expectation-maximization (EM) algorithm to find the MLE for π and Z_i, treating π as the parameters and the Z_i as the latent variables. We initialize the EM algorithm by setting $\pi^0 = (1/3\,, 1/3\,, 1/3)$.

In the E-step, we evaluate the distribution of the latent variables (Z_i) given the values of the parameters (π) from the previous iteration. The E-step is

$$p(Z_i \mid \pi_i, PTV_i) = \frac{Pois(PTV_i \mid N\lambda_{Z_i})\pi_i}{\sum_i Pois(PTV_i \mid N\lambda_{Z_i})\pi_i},$$

where $Pois$ denotes the Poisson likelihood. In the M-step, we update the parameters π with a new expectation taken under the distribution of the latent variables (Z_i) computed in the M step. The update is

$$\pi^{new} := \sum_Z p\left(Z_i \mid PTV_{i,}\pi^{old}\right)/Ngenes$$

We repeat these steps until the convergence criteria are met (π_{HI} changes by less than 0.001 from one iteration to the next).

When the EM has converged, the final mixing weights are used to determine each gene's probability of belonging to each of the categories (null, recessive, haploinsufficient).

$$Z_{i,Null} = Pois(PTV_i \mid N\lambda_{Null})$$
$$Z_{i,Rec} = Pois(PTV_i \mid N\lambda_{Rec})$$
$$Z_{i,HI} = Pois(PTV_i \mid N\lambda_{HI})$$

The final metric, pLI (the probability of being loss-of-function intolerant):

$$pLI = \frac{Z_{i,HI}}{\sum Z_i}$$

The closer pLI is to 1, the more likely the transcript is loss-of-function (LoF) intolerant. The overall distribution of pLI is fairly bimodal, with most genes looking either tolerant or intolerant of protein-truncating variation (Extended Data Figure 8a). Additionally, pLI is only modestly correlated with transcript length (r = 0.1668). However, we find that the most highly LoF-intolerant genes (pLI ≥ 0.9) are significantly longer than all genes (Wilcox $p < 10^{-50}$). The least intolerant genes are also significantly—but to a lesser extent—larger than all genes (Wilcox $p < 10^{-3}$).

In order to additionally confirm that the pLI metric was free of confounding with gene length, we compare the gene size distribution of genes with a pLI ≥ 0.99 versus genes that had the pLI equivalent for falling into the recessive category (pRec) ≥ 0.99. pRec is determined by the equation below:

$$pRec = \frac{Z_{i,Rec}}{\sum Z_i}$$

We find no significant difference in the distribution of gene length between genes with pLI ≥ 0.99 (n=1,803) and genes with pRec ≥ 0.9 (n=1,145; p = 0.3032).

We also show that longer genes are, in general, more depleted of protein-truncating variation (observed/expected), which can explain the enrichment of long genes in the set of genes with pLI ≥ 0.9. There is a relationship between deciles of gene length (bins of increasing gene length) and the observed depletion of PTVs in that bin: longer genes (deciles closer to 1) have a significantly lower rate of observed/expected ($p < 10^{-50}$).

Given that the X chromosome is hemizygous in males, we expect that genes on the X would be more constrained than those on autosomes. As expected, we find the genes on the X chromosomes are significantly more constrained than those genes on the autosomes for missense and loss-of-function (synonymous p = 0.0223; missense p = 4.43×10^{-8}; loss-of-function p = 2.50×10^{-75}). The high correlation between the observed and expected number of synonymous variants on the X chromosome (r = 0.9677 vs

0.9777 for autosomes) indicates that this difference in constraint is not due to a calibration issue.

We find that 3,230 (17.7%) of genes are confidently considered extremely loss-of-function intolerant since their pLI is 0.9 or greater. Similarly, there are 3,463 (19.0%) and 1,226 (6.7%) genes with pRec or pNull ≥ 0.9, respectively. pRec and pNull also show fairly bimodal distributions (Extended Data Figure 8a, Supplementary Information Table 13). As a warning, while we consider pLI to be a valuable metric to identify genes that appear haploinsufficient, we caution against using pRec as a similar metric for recessive disease genes. An appropriate recessive disease gene metric would benefit from including information about the site frequency spectrum of variants observed in the gene, among other properties.

Comparison to a previous haploinsufficiency metric: p(HI)

Our metric to evaluate loss-of-function intolerance was designed to identify genes that are intolerant of heterozygous loss-of-function variants, which would mean that these genes are likely acting via haploinsufficiency. Previously, Huang et al (2010) designed p(HI)—the probability of being haploinsufficient[21]—to determine how likely each gene was to be haploinsufficient. Huang and colleagues made this metric by using properties of established haploinsufficient and haplosufficient genes to train a predictive model. The properties included in the final model were "dN/dS between human and macaque, promoter sequence, embryonic expression and network proximity to known HI [haploinsufficient] genes"[21].

In order to compare pLI and Huang's p(HI), we took the 18,064 genes that had values for both metrics. Since p(HI) was trained on a set of haploinsufficient genes, we removed 64 genes that were part of their training data set and considered to be haploinsufficient by ClinGen's Dosage Sensitivity Map, which left 18,000 genes for analysis. While there are 3,175 genes in this set with pLI ≥ 0.9, there are only 613 with p(HI) ≥ 0.9. For this reason, we dropped the cut-off to 0.8, giving 3,878 genes for pLI and 1,061 for p(HI).

Within the 18,000 genes, 148 are considered haploinsufficient by ClinGen, 109 of which have a pLI ≥ 0.8. By contrast, only 51 of the 148 haploinsufficient genes have a p(HI) ≥ 0.8, and 80% of those (n = 41) also have pLI ≥ 0.8. Our metric identifies twice as many

genes at the same cut off, but a larger proportion of the genes in the high p(HI) tail are considered likely haploinsufficient by both metrics.

Supplementary Information Table 14a and b depict the breakdown of all genes and ClinGen haploinsufficient genes, respectively, by their pLI and p(HI) values. We took those data and found the enrichment of ClinGen haploinsufficient genes in the high pLI and p(HI) tails by setting as baseline the fraction of ClinGen haploinsufficient genes with pLI and p(HI) < 0.8 compared to all genes in that category (n = 29 and 13,681, respectively). The fraction of each other category was compared to this baseline to determine the enrichment of genes that fall into each of the other categories (pLI < 0.8 and p(HI) ≥ 0.8, etc) and is shown in Supplementary Information Table 14c. Genes uniquely flagged by both metrics have similar enrichments (10 for pLI vs 10.8 for p(HI)). The real enrichment, however, is found in the subset of genes that are considered likely haploinsufficient (≥ 0.8) by both metrics.

4.5 Evaluating protein-truncating constraint metrics

To determine which of the three protein-truncating constraint metrics (PTV Z, ratio of missing protein-truncating variation, and pLI) is the most useful to use as a general LoF intolerance measure, we perform two tests: (1) the ability to predict known haploinsufficient genes and (2) enrichment of *de novo* PTVs found in autism spectrum disorder cases.

We perform a logistic regression using the three protein-truncating constraint metrics to predict inclusion in the ClinGen haploinsufficient gene list. For all regressions, transcript length is included as a covariate. pLI has the highest Z-value (14.314), reflecting a more significant ability to predict haploinsufficient genes. The Z-value for PTV Z is 11.307 and is 12.164 for the ratio of missing protein-truncating variation.

For the enrichment of *de novo* PTVs, we use the published *de novo* variants from 3,982 cases with autism and 2,078 controls[22,23] and a previously described method that controls for the mutability of each gene[20]. In brief, the probability of mutation (for a specific mutation type) is summed across all genes in a gene set and compared to the total probability of mutation (of the same type) for all genes. That fraction becomes the

expected fraction of genes in the gene set that should harbor a *de novo* variant of the same type. We evaluate the observed overlap between the *de novo* list and the gene set of interest by invoking the binomial.

Since this method requires an established gene set, we took genes with pLI ≥ 0.9 (n = 3,230) and matched the set size using the genes with the highest PTV Z scores and ratio of missing protein-truncating variation. While the fold enrichment is greatest for the ratio of missing protein-truncating variation (enrichment = 1.9, $p < 10^{-21}$), pLI still outperforms the PTV Z score (Supplementary Information Table 15). No significant enrichments are seen when using the control *de novo* PTVs (fold enrichments between 0.81 and 0.91).

4.6 Applications of pLI

Given pLI's superior performance in predicting haploinsufficient genes and clearer interpretability than the ratio of missing protein-truncating variation, we chose to use pLI as our main metric of LoF intolerance.

As shown in Figure 3b in the main text, established haploinsufficient genes are enriched in the high pLI tail (pLI ≥ 0.9, χ^2 $p < 10^{-50}$). Of note, the enrichment in pLI stratifies with the severity of the disease caused by the haploinsufficient genes with increasingly severe phenotypes showing increased enrichment in the highly LoF-intolerant genes (manually curated from the ClinGen dosage sensitivity list by AODL). Critically, we note that LoF-intolerant genes include virtually all known severe haploinsufficient human disease genes (Figure 3b), but that 79% of these genes have not yet been assigned a human disease phenotype despite the clear evidence for extreme selective constraint.

The targets of FMRP are also strongly enriched in the high pLI tail (pLI ≥ 0.9, χ^2 $p < 10^{-50}$; Extended Data Figure 8b). Dominant disease genes and those essential in cell culture, however, are more evenly split between the two categories, but still enriched for pLI ≥ 0.9 (χ^2 $p < 10^{-30}$ and $p < 10^{-23}$, respectively). Olfactory receptors and recessive disease genes have low pLI scores overall, indicating that these sets are not likely haploinsufficient. These results do not mean that recessive genes are not important to disease, but that they can on average tolerate a heterozygous PTV.

We also study three gene lists that correspond to genes found in mice: those genes that are lethal as homozygous knock outs, genes that are lethal as heterozygous knock outs, and genes that are lethal when conditionally knocked out in adult mice (also described in Supplementary Information Table 12). As depicted in Extended Data Figure 8b, the conditional lethal genes are the most enriched in the most LoF-intolerant genes, followed by the heterozygous lethal, and then the homozygous lethal genes.

4.7 Gene expression and eQTLs

To further understand the characteristics of constrained genes we investigate the association of the synonymous Z score, missense Z score, and pLI with various gene expression and regulation metrics utilizing the multi-tissue gene expression data from the Genotype-Tissue Expression (GTEx) project[24] (GTEx Analysis V4, dbGaP Accession phs000424.v4.p1) spanning 53 tissue types sampled from 212 post-mortem donors downloaded from the GTEx portal (http://www.gtexportal.org) on July 29, 2015.

The medians of log2-transformed RPKM values for each tissue are correlated with the constraint scores after excluding sex chromosomal transcripts and transcripts not expressed in the given tissue (i.e. median RPKM = 0). Given the high correlation in gene expression between the various brain regions sampled in GTEx, a composite measure for brain expression is created by taking the median expression values for each gene across these eleven brain tissue types (only one of the duplicate measurements for each cerebellum and cortex was included). This composite brain expression measure is used instead of the individual brain regions when the per-gene median and maximum expression values across all tissues are calculated and similarly when the total number of tissues a given gene is expressed in is determined, therefore giving 41 as the maximum number of tissues in which a gene can be detected.

Consistently in each tissue, gene expression level is strongly and positively correlated with missense Z score and pLI, a result that is further strengthened after accounting for gene coding sequence length. The association with synonymous Z score, however, is non-significant or considerably subtler. Similar patterns of association are observed for the median and maximum gene expression across tissues (median gene expression is

depicted in Extended Data Figure 9a). Also, the total number of tissues a gene is expressed in is positively correlated with missense Z score and pLI at different RPKM cutoffs (Figure 3c; Extended Data Figure 9b).

The relationship between the constraint scores and gene regulatory variation detected in the GTEx data set is investigated in the 13 tissues with the largest sample sizes (expression and genotype data available for >60 individuals) that were included in the GTEx V4 eQTL analyses (Adipose – Subcutaneous, Artery - Aorta, Artery – Tibial, Esophagus - Mucosa, Esophagus - Muscularis, Heart - Left Ventricle, Lung, Muscle – Skeletal, Nerve – Tibial, Skin - Sun Exposed (Lower leg), Stomach, Thyroid and Whole Blood). The eQTL analysis follows the steps described in detail in the GTEx pilot phase manuscript[24].

Dividing the analyzed transcripts into three subsets based on their constraint scores (for Z: bottom quartile (<25%), the two middle quartiles grouped, top quartile (>75%); for pLI: pLI ≤ 0.1, 0.1 < pLI < 0.9, pLI ≥0.9), we calculate the proportion of eGenes, i.e. a gene with a significant eQTL (FDR 5%), out of all genes included in the eQTL analysis (expressed in at least ten individuals at >0.1 RPKM) in each of the constraint subsets for each of the 13 tissues and for synonymous, missense and LoF constraint scores separately. The power for eQTL discovery varies widely from tissue to tissue given the sample sizes per tissue, which range from 74 (Artery - Aorta) to 168 (Whole Blood). Independent of the total number of eGenes discovered, in each tissue, the most missense and loss-of-function constrained group of genes are significantly depleted of eGenes compared to the least constrained group (e.g. in skeletal muscle, $p < 10^{-24}$ for pLI). Such pattern is not seen when grouping the genes based on their constraint for synonymous variation. To have a metric comparable between tissues, we further normalize these eGene proportions by the total number of eGenes discovered in each tissue. Figure 3d shows the average proportion of eGenes in whole blood clearly demonstrating both the depletion (59.57% of the average for pLI) of eGenes among the most and enrichment (125.11% of the average for pLI) among the least missense and loss-of-function constrained genes.

4.8 Enrichment of GWAS signals

Next we investigate the same synonymous Z score, missense Z score, and pLI in the Genome-wide Association Studies (GWAS) Catalog[25] for the closest gene to signal; see Supplementary Information Table 12 [Hindorff et al, Accessed 02/04/2015]. We filter results to include only those GWAS signals that had been reported with a $p < 5.0 \times 10^{-8}$. In order to categorize GWAS results by ontologies, we only include those signals that have been mapped in the "Experimental Factor Ontology" (EFO, http://www.ebi.ac.uk/efo). We find 2,792 unique genes that have been listed in the Catalog and for which we have Z scores and pLI.

As performed in previous analyses, we divide variants by functional categories: synonymous, missense and loss-of-function, and each category was further divided in three constraint groups: Lowest (0 - 25% quantile for Z; pLI ≤ 0.1), Middle (25 – 75% quantile for Z; 0.1 < pLI < 0.9) and Highest (75 – 100% quantile for Z; pLI ≥ 0.9). Then we estimate the enrichment of genes in the GWAS catalogue as:

$$E_q = P_q * S$$
$$P_q = {}^{GWAS_q}/_{GWAS}$$
$$S = {}^{N}/_{GWAS}$$

where:
P_q is the proportion of GWAS genes in the quantile q
and S is a scaling factor (number of evaluated genes divided by number of GWAS hits)
The standard error for the proportions are similarly scaled:

$$SE = \sqrt{\left(\frac{P_q(1 - P_q)}{N_q}\right)} * S$$

We estimate the significance of the difference in the number of GWAS loci of highest vs the lowest constraint scores using a χ^2 test.

While only the loss-of-function category shows a clear and significant difference between the highest and the lowest constraint scores, we note a pattern in the missense category

where the less constrained genes have higher, albeit not significant, proportion of GWAS hits than the middle category (Figure 3e).

To better characterize this pattern we divide the GWAS hits by major EFO categories: Cancer, Cardiovascular, Digestive, Immune, Metabolic, Nervous, Response to drug, Body measure and Others, and compare the least constrained genes vs the middle category as well as the most constrained genes vs the middle category (Extended Data Figure 9c). Again, we see that on average, GWAS hits are enriched in the most LoF constrained genes and depleted in the least constrained. In this sub-analysis we also identify an enrichment of Cardiovascular, Metabolic, and body measurement GWAS hits in the most missense constrained genes, while these categories with enrichments were non-significant in least missense constraint genes.

4.9 Networks and pathway analysis

To better understand the set of genes considered intolerant of loss-of-function variation, we use the STRING database[26] to obtain a network of experimentally supported protein-protein physical interactions. The network consists of 14,160 genes (nodes) and 712,137 physical interactions (edges). For each gene, we compute the number of neighbors it has in the network (degree of the node), which corresponds to the number of interaction partners its encoded protein has. We run a linear regression between the pLI score of a gene and its number of interaction partners and find that genes with more partners are more likely to have high pLI scores (t-test $p < 10^{-41}$). A weaker positive correlation is found between the number of interaction partners and the missense Z score of a gene (t-test $p < 10^{-8}$). A weak negative correlation is observed between the number of partners and the synonymous Z score (t-test $p < 10^{-6}$).

The list of 186 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were obtained from Broad Institute GSEA. Each pathway is represented by the list of pLI scores for each of the genes in the pathway. For each pathway, we compute the Kolmogorov-Smirnov (KS) statistic between its list of pLI scores and the pLI scores of all the genes to quantify the enrichment or depletion of pLI for this pathway. Fifty-eight pathways show significant deviations in pLI from the rest of the genes at multiple-testing adjusted p-value of 10^{-7} (Supplementary Information Table 16).

For each pathway, we quantify the degree of its redundancy by computing the fraction of its genes with a duplication in the human genome[27]. Among the highly 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 $p < 10^{-16}$). More surprisingly, we identify multiple metabolic pathways, such as starch and sucrose metabolism (KS test $p < 10^{-9}$), as being highly unconstrained. Members of these pathways are also likely to have paralogous genes in the human genome.

4.10 Stratifying variants by pLI, Z scores, and MAPS

Finally, we repeat the frequency spectral analysis of functional categories as in Figure 2e (see Section 3). For each functional category, we separate the variants into low, medium, and high constraint for each overarching category type (Figure 3f). We use the synonymous Z score for benign mutational classes (gray), the missense Z score for the missense classes (orange), and pLI for the protein-truncating variants (maroon). For all categories, but particularly for missense and protein-truncating variants, the more constrained categories harbor a higher proportion of singleton variants (corrected for mutability; MAPS). For instance, missense variants in genes with the highest missense Z scores have a higher MAPS, indicating that a greater percentage of these variants are singletons than missense variants in low missense Z score genes. This trend is consistent for those missense variants predicted to be “probably damaging” by PolyPhen and missense and protein-truncating variants in the highest CADD tertile, indicating that the constraint score is complementary to these approaches and together, could be used to prioritize variants in pathogenic scans of clinical genomes.

5 Mendelian Analysis

Authors: Anne H. O'Donnell-Luria, Eric V. Minikel, James S. Ware and Daniel G. MacArthur

5.1 Comparison with 1000 Genomes and ESP

For comparisons to other datasets we used the 1000 Genomes Phase 3 and the NHLBI-GO Exome Sequencing Project (ESP) VCF files. The 1000 Genomes files were subsetted to the ExAC calling intervals using GATK SelectVariants. Allele frequency information for filtering was extracted from each dataset using custom Python scripts. Predicted protein-altering variants were defined as those with worst VEP annotations that were either "missense-like" (missense, in-frame indel, stop lost, start lost, and mature miRNA), or protein-truncating (frameshift, essential splice, and stop gained). We also created VCF files of 500 individuals for a simulated Mendelian analysis (for Figure 4a, Supplementary Information Table 18) and of ExAC lacking ESP individuals (for Figure 4b), using GATK SelectVariants. Details of the data sets and scripts used for this analysis is summarized in Supplementary Information Table 17.

5.2 Number of reportedly mendelian variants per person

We utilized a February 2014 data freeze of the Human Gene Mutation Database (HGMD) and the July 9, 2015 release of ClinVar. Briefly, ClinVar variants from XML and TXT releases were harmonized into a single tab-delimited file (Supplementary Information Table 17) and run through a Python implementation of vt normalize[28]. Membership of HGMD and ClinVar variants in ExAC was determined by matching a pos_id defined as chromosome, position padded to 9 digits, reference, and alternate alleles in normalized representation. For ExAC analyses, we ignored HGMD and ClinVar variants where the reference allele is the reportedly pathogenic mutation. Lists of autosomal dominant and autosomal recessive variants were the same as those used for constraint analysis detailed in Supplementary Information Table 12.

33,783 ClinVar[29] variants were annotated as pathogenic and non-conflicted, 7,987 (23.6%) of which were present in ExAC. An additional 14,778 HGMD variants were present in ExAC. This yielded a total of 22,765 reportedly pathogenic (ClinVar and/or

HGMD) variants in ExAC, of which 844 have AF>1% in at least one continental population. On average, ExAC individuals harbor 53.7 alleles previously reported as Mendelian disease-causing in HGMD or ClinVar, but 47.2 of these are variants with AF >1% in at least one continental population (Supplementary Information Table 21). Such high allele frequency is incompatible even with recessive inheritance, except in rare cases of balancing selection[30,31], and the well-known examples *CFTR* p.F508del (cystic fibrosis) and *HBB* p.E7V (sickle cell) contribute only 0.02 to the total number of such variants per person, suggesting that genuinely common true disease variants are not a major contributor to the total. Meanwhile, restricting to high-confidence variants only reduces the overall figure slightly, from 53.7 to 46.8 variants per person, suggesting that genotyping error is not a major contributor either. Indeed, as noted in the main text, most (41.0 variants per person) of the burden of reportedly Mendelian variation consists of HQ genotypes with >1% AF.

Our curation efforts confirm the importance of appropriate allele frequency filtering in analysis of candidate Mendelian disease variants. However, literature and database errors are also prevalent at lower allele frequencies, because even when excluding all variants with AF >1% in any population, there remain 5.8 reportedly pathogenic alleles per ExAC individual. While we were not able to assign an inheritance mode for all such genes, we analyzed the data for genes that do fall within our autosomal dominant or autosomal recessive gene lists. As noted in the main text, 0.89 reportedly pathogenic alleles per person fall in known autosomal dominant disease genes[32], implausibly suggesting that almost all individuals have a dominant genetic disease. In addition, 2.0 fall in autosomal recessive disease genes[32]. That figure is higher than the range of estimates (0.6 to 1.4) based on consanguineous families[33,34], and is particularly suspect as our total includes only *known* disease variants. Thus, a large fraction of the cumulative burden of reported disease variants in ExAC exomes must arise from some combination of false positives and incompletely penetrant variants.

5.3 Mendelian variant review

192 variants that were reported to be disease mutations (DM) in HGMD were selected for literature review and curation by AODL and EVM. 134 of these 192 variants have since been retired or reclassified in HGMD as of December 2015. The variants included:

(1) 75 variants with a global allele frequency of >1% (Supplementary Information Table 19) and (2) 117 variants global allele frequency of <1% but a population allele frequency of >1% in South Asians (SAS) or Latin American (AMR) ancestry samples (Supplementary Information Table 20). It was also required that the variants PASS all filters and have an adjusted AN of >60,000. A more abbreviated review was performed on the 75 variants with >1% global allele frequency, as this was in most cases much too high to be consistent with the disease prevalence. For the 117 variants with the AMR or SAS population allele frequency of >1%, all HGMD annotated literature support for each variant was reviewed for both for the level of evidence supporting disease pathogenicity and further curated according to ACMG criteria[35]. Specifically, author's confidence of pathogenicity, number of cases identified (heterozygous, compound heterozygous, or homozygous), functional data, and ancestry of probands was evaluated, in combination with the number of homozygotes and hemizygotes in ExAC for the variants. Variants were classified as disease associated (DA), benign traits (BT), insufficient evidence of pathogenicity (IE), and classification errors (CE). While disease associated and benign trait variants were excluded from the second round of analysis, the remaining 173 variants were evaluated according to ACMG criteria. Of these, 118 were reclassified as benign (B), 45 as likely benign (LB) and 10 as variants of uncertain significance (VUS). Of the 26 IE variants with functional data, there were 18 cases where it was described as positive in support of pathogenicity and 8 cases where the functional data was negative.

We hypothesized that false positive disease associations might be particularly prevalent for variants first identified in Latino or South Asian probands, as reference databases of genetic variation in these populations have been especially limited until recently. We therefore reviewed reported ancestry information for probands in the variants we had analyzed. For 27 (55%) of the 49 variants for which ancestry information was available, at least one (or all) of the original probands were of South Asian or Latino ancestry. In some of the remaining cases, the proband was from a related population.

6 Protein-truncating variation

Authors: Konrad J. Karczewski, Daniel P. Birnbaum and Daniel G. MacArthur

6.1 Generating a high-confidence set of protein-truncating variants (PTVs)

Of the 7.4M high-quality variants in ExAC, we detect 221,860 putative protein-truncating (frameshift, splice donor, splice acceptor, and stop-gained) variants. We apply stringent filters using LOFTEE (Supplementary Information Table 8), removing 42,186 variants as in Supplementary Information Table 23 and Supplementary Information Table 24, which results in 179,774 remaining variants, which we term high-confidence (HC) PTVs.

6.2 PTV burden across populations

On average, each individual harbors 85.1 heterozygous and 34.2 homozygous protein-truncating variants (PTVs; Figure 5a). The number varies across populations as shown in Supplementary Information Table 25.

To properly assess differential PTV burden across populations, we downsampled the ExAC dataset to 3000 individuals per population. Here, we observe different frequency spectra of protein-truncating variants across populations (Figure 5b). For instance, the Finnish population is depleted for low frequency variation, but modestly enriched for PTVs in the 1-5% frequency range, a result consistent with previous work[36]. Africans have a higher proportion of common PTVs (greater than 0.1% frequency).

However, when considering only PTVs in LoF-constrained genes (pLI > 0.9; see above), there is minimal difference between populations across allele frequency categories (Extended Data Figure 10).

6.3 PTVs in known disease genes

Of the 179,774 PTVs discovered, 19,403 occur in autosomal recessive disease genes (union of gene lists from Blekhman and Berg, Supplementary Information Table 12), 13,462 of which are singletons. 7,343 occur in autosomal dominant disease genes, 5,236 of which are singletons.

7 Data availability

The annotated VCF sites and coverage files are available to download directly via FTP (Supplementary Information Table 8) or for viewing on the ExAC Browser (http://exac.broadinstitute.org).

8 References

1. 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).

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

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

4. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science (80-.).* **319,** 1100–4 (2008).

5. Wang, J., Raskin, L., Samuels, D. C., Shyr, Y. & Guo, Y. Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics* **31,** 318–323 (2014).

6. Montgomery, S. B. *et al.* The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* **23,** 749–61 (2013).

7. Manichaikul, a. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010).

8. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).

9. Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61,** 893–903 (1969).

10. Carlson, K. M. *et al.* Parent-of-origin effects in multiple endocrine neoplasia type 2B. *Am. J. Hum. Genet.* **55,** 1076–1082 (1994).

11. Bellus, G. a *et al.* Achondroplasia is defined by recurrent G380R mutations of FGFR3. *Am. J. Hum. Genet.* **56,** 368–373 (1995).

12. Moloney, D. M. *et al.* Exclusive paternal origin of new mutations in Apert syndrome. *Nat. Genet.* **13,** 48–53 (1996).

13. Muona, M. *et al.* A recurrent de novo mutation in KCNC1 causes progressive myoclonus epilepsy. doi:10.1038/ng.3144

14. MacDonald, M. E. *et al.* The Huntington's disease candidate region exhibits many different haplotypes. *Nat. Genet.* **1,** 99–103 (1992).

15. Lee, H. S. *et al.* Ancestral Origins and Worldwide Distribution of the PRNP 200K Mutation Causing Familial Creutzfeldt-Jakob Disease. *Am. J. Hum. Genet.* **64,** 1063–1070 (1999).

16. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455,** 232–236 (2008).

17. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519,** 223–8 (2015).

18. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506,** 179–84 (2014).

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

20. 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

21. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* **6,** 1–11 (2010).

22. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515,** 209–215 (2014).

23. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515,** 216–221 (2014).

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

25. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001–D1006 (2014).

26. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39,** D561–D568 (2011).

27. Ouedraogo, M. *et al.* The Duplicated Genes Database: Identification and Functional Annotation of Co-Localised Duplicated Genes across Genomes. *PLoS One* **7,** e50653 (2012).

28. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31,** 2202–2204 (2015).

29. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980–5 (2014).

30. Andres, a. M. *et al.* Targets of Balancing Selection in the Human Genome. *Mol. Biol. Evol.* **26,** 2755–2764 (2009).

31. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E455–64 (2014).

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

33. Bittles, a H. & Neel, J. V. The costs of human inbreeding and their implications for variations at the DNA level. *Nat. Genet.* **8,** 117–121 (1994).

34. Gao, Z., Waggoner, D., Stephens, M., Ober, C. & Przeworski, M. An Estimate of the Average Number of Recessive Lethal Mutations Carried by Humans. *Genetics* **199,** 1243–1254 (2015).

35. 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).

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

37. Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42,** D1091–D1097 (2014).

38. Knox, C. *et al.* DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Res.* **39,** D1035–D1041 (2011).

39. Wishart, D. S. *et al.* DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **36,** 901–906 (2008).

40. Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34,** D668–D672 (2006).

41. Berg, J. S. *et al.* An informatics approach to analyzing the incidentalome. *Genet. Med.* **15,** 36–44 (2013).

42. Hart, T., Brown, K. R., Sircoulomb, F., Rottapel, R. & Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol. Syst. Biol.* **10,** 733 (2014).

43. Mainland, J. D., Li, Y. R., Zhou, T., Liu, W. L. L. & Matsunami, H. Human olfactory receptor responses to odorants. *Sci. Data* **2,** 150002 (2015).

44. Blake, J. a, Bult, C. J., Kadin, J. a, Richardson, J. E. & Eppig, J. T. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.* **39,** D842–8 (2011).

45. Georgi, B., Voight, B. F. & Bućan, M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet.* **9,** e1003484 (2013).

46. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: A database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.* **34,** 2393–2402 (2013).

47. Darnell, J. C. *et al.* FMRP Stalls Ribosomal Translocation on mRNAs Linked to Synaptic Function and Autism. *Cell* **146,** 247–261 (2011).

Resource	Link
Human Genome Reference (GRCh37/hg19)	http://www.broadinstitute.org/ftp/pub/seq/references/Homo_sapiens_assembly19.fasta
Exome calling intervals	ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/exome_calling_regions.v1.interval_list
Hapmap 3.3 (VQSR)	ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/2.8/b37/hapmap_3.3.b37.vcf.gz
1KG Omni 2.5 (VQSR)	ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/2.8/b37/1000G_omni2.5.b37.vcf.gz
Gold standard Indels (VQSR)	ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/2.8/b37/Mills_and_1000G_gold_standard.indels.b37.vcf.gz
dbSNP137 (VQSR)	ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/2.8/b37/dbsnp_138.b37.vcf.gz
Axiom Exome Plus (VQSR)	ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/Axiom_Exome_Plus.genotypes.all_populations.poly.vcf.gz

Supplementary Information Table 1. Publicly available resource files used for variant calling.

Consortium/Cohort	Samples
1000 Genomes	1,851
Bulgarian Trios	461
GoT2D	2,502
Inflammatory Bowel Disease	1,675
Myocardial Infarction Genetics Consortium	14,622
NHLBI-GO Exome Sequencing Project (ESP)	3,936
National Institute of Mental Health (NIMH) Controls	364
SIGMA-T2D	3,845
Sequencing in Suomi (SISu)	948
Swedish Schizophrenia & Bipolar Studies	12,119
T2D-GENES	8,980
Schizophrenia Trios from Taiwan	1,505
The Cancer Genome Atlas (TCGA)	7,601
Tourette Syndrome Association International Consortium for Genomics (TSAICG)	297
Total	60,706

Supplementary Information Table 2. The Exome Aggregation Consortium (ExAC) sample numbers from each Consortia/cohort.

Population	Male Samples	Female Samples	Total
African/African American (AFR)	1,888	3,315	5,203
Latino (AMR)	2,254	3,535	5,789
East Asian (EAS)	2,016	2,311	4,327
Finnish (FIN)	2,084	1,223	3,307
Non-Finnish European (NFE)	18,740	14,630	33,370
South Asian (SAS)	6,387	1,869	8,256
Other (OTH)	275	179	454
Total	33,644	27,062	60,706

Supplementary Information Table 3. ExAC samples summarized by population and sex.

VCF INFO Tag	Description
AC_Adj	Adjusted allele counts
AC_Pop	Population specific allele counts
AN_Adj	Number of chromosomes adjusted by sex
AN_Pop	Population specific number of chromosomes adjusted by sex
AC_Het	Number of heterozygous individuals
Het_Pop	Population specific heterozygous individuals
AC_Hom	Number of homozygous individuals
Hom_Pop	Population specific homozygous individuals
AC_Hemi	Number of hemizygous males on the chromosome X and Y
Hemi_Pop	Population specific hemizygous males on the chromosome X and Y
DP_HIST	Depth (DP) histogram in 20 equal intervals between 0-100
GQ_HIST	Genotype Quality (GQ) histogram in 20 equal intervals between 0-100

Supplementary Information Table 4. Additional site summary information included in the VCF.

Note: Pop indicates one of the population abbreviations summarized in Supplementary Information Table 3

a)

Singleton	<0.1%	0.1-1%	1-10%	>10%
0.389%	0.696%	0.143%	0.315%	0.029%

b)

missense	synonymous	ptv
0.076%	0.055%	0.471%

Supplementary Information Table 5. SNP false discovery rate compared to PCR-Free WGS.
a) Overall SNP false discovery rate across various allele frequencies. b) Overall false discovery rate across function annotation classes: missense, synonymouse and protein truncating variants (ptv, including indels).

Supplementary Information Table 6. Per sample SNP and Indel false discovery rate by allele frequency and functional annotation. Note: Provided Excel file.

Call set	Samples	Filters	Alleles	Analysis
All samples	91,796	None	NA	
All samples - Variant site filtering	91,796	adjusted VQSR PASS	NA	Exome chip concordance Singleton transmission Validated de-novo sensitivity False discovery using PCR-Free WGS
ExAC r0.3 public release	60,706	None	10,195,872	
ExAC r0.3 public release - Variant site filtering	60,706	adjusted VQSR PASS	9,042,960	NA12878 sensitivity/specificity
ExAC r0.3 public release - Variant site and Genotype filtering	60,706	adjusted VQSR PASS High quality variants (HQ)	7,404,909	All analyses except Section 4 or unless specified

Supplementary Information Table 7. Call sets used for QC evaluation and analysis.

Resource	Link
ExAC r0.3 public release	ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/ExAC.r0.3.sites.vep.vcf.gz
Exome Calling Intervals	ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/exome_calling_regions.v1.interval_list
Summary coverage metrics	ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/coverage
LOFTEE (VEP plugin)	https://github.com/konradjk/loftee
Context Annotation (VEP plugin)	https://github.com/konradjk/loftee/blob/master/context.pm
CADD Anotation	http://krishna.gs.washington.edu/download/CADD/v1.3/ExAC_r0.3.tsv.gz
Analysis code for figures generated	https://github.com/macarthur-lab/exac_2015
Python script for calculating addtion metrics contained in INFO	https://github.com/macarthur-lab/exac_2015/blob/master/src/prepare_exac_sites_vcf.py

Supplementary Information Table 8. Publicly available ExAC resources.

Supplementary Information Table 9. The breakdown of observed (ExAC) and possible (synthetic) variants by functional class and mutational class. Note: Provided as Excel file.

Gene	Disease	Inheritance	MNP Class	Genomic Positions
RFXANK	Bare lymphocyte syndrome, complementation group B	Autosomal Recessive	Rescued PTV	19:19308027, 19:19308028
FBP1	Fructose-1,6-bisphospatase deficiency	Autosomal Recessive	Rescued PTV	9:97365720, 9:97365722
MLH1	Lynch syndrome	Risk Autosomal Dominant	Rescued PTV	3:37090485, 3:37090486
FANCA	Fanconi Anemia	Autosomal Recessive	Rescued PTV	16:89865604, 16:89865605
MSH6	Lynch Syndrome	Risk Autosomal Dominant	Changed Missense	2:48027683, 2:48027684
DNAH5	Primary ciliary dyskinesia	Autosomal Recessive	Changed Missense	5:13811775, 5:13811776
MUTYH	MUTYH-associated polyposis	Autosomal Recessive	Gained PTV	1:45795077, 1:45795078
ABCA1	Tangier disease	Autosomal Recessive	Gained PTV	9:107578478, 9:107578480
TJP2	Progressive cholestatic liver disease	Autosomal Recessive	Gained PTV	9:71843021, 9:71843023
VWF	Von Willebrand Disease	Autosomal Recessive/ Dominant*	Gained PTV	12:6138595, 12:6138597

Supplementary Information Table 10. Additional Examples of MNPs with functional impact on disease-associated genes. PTV = protein-truncating variant

Gene	Disease	Inheritance	MNP Class	PubMed ID	Genomic Positions
COH1	Cohen Syndrome	Autosomal Recessive	Gained PTV	23188044	8:100791048, 8:100791049
ANGPTL3	Familial combined hypolipidemia	Autosomal Recessive	Gained PTV	20942659, 22659251	1:63063287, 1:6306328
CLCN1	Myotonia congenita	Autosomal Recessive/ Dominant*	Gained PTV	7874130	7:143027909, 7:143027910
TNFRSF13B	Combined variable immunodeficiency	Autosomal Recessive/ Dominant*	Gained PTV	16007087	17:16843689, 17:16843690
GLB1	Morquio Disease B	Autosomal Recessive	Partial Change	1928092 21497194	3:33093471, 3:33093472
ALPL	Hypophosphatasia	Autosomal Recessive	Changed Missense	11855933	1:21889705, 1:21889706
COL7A1	Epidermolysis bullosa dystrophica	Autosomal Recessive	Changed Missense	16971478 21448560	3:48613970, 3:48613971
*UNC5C**	Colorectal cancer risk	Risk Autosomal Dominant	Changed Missense	21893118	4:96127798, 4:96127799
UROD	Hepatoerythropoietic porphyria	Autosomal Recessive	Changed Missense	8176248	1:45479389, 1:45479390

Supplementary Information Table 11. MNPs identified in our study that are also in HGMD.

Note: * Variants in the gene are associated with both the autosomal recessive and dominant form of the disease. **Variant is annotated as disease-causing in HGMD and as VUS in ClinVar. PTV = protein-truncating variant.

List name	Reference	Comments
Drug targets	Drugbank (Law 2014[37], Knox 2011[38], Wishart 2008[39], Wishart 2006[40])	Drugbank was accessed Feb 9, 2015. For each FDA-approved drug, only the #1 "mechanistic" target was included.
Dominant disease genes	Blekhman 2008[32] and Berg 2013[41]	
Recessive disease genes	Blekhman 2008[32] and Berg 2013[41]	
ClinGen haploinsufficient	ClinGen (http://www.ncbi.nlm.nih.gov/projects/dbvar/clingen/)	Genes with sufficient evidence for dosage pathogenicity (level 3) as determined by the ClinGen Dosage Sensitivity Map accessed 02/27/2015.
Essential in culture	Hart 2014[42]	
GWAS hits	Welter 2014[25]	Closest gene 3' and 5' of GWAS hits in the NHGRI GWAS catalog (genome.gov/gwastudies/) accessed 02/09/2015.
Olfactory receptors	Mainland 2015[43]	
Essential in mice	Blake 2011[44], Georgi 2013[45], Liu 2013[46]	
FMRP interacting genes	Darnell 2011[47]	
OMIM	https://github.com/macarthur-lab/gene_lists/blob/master/other_data/omim.use.tsv	Cleaned OMIM gene list mapping to HGNC gene names.

Supplementary Information Table 12. Sources of gene lists used in analyses.

Note: Gene list is also available at https://github.com/macarthur-lab/gene_lists

Supplementary Information Table 13. Z scores, pLI, pRec and pNull for all genes. Note: Provided as Excel file.

a. Breakdown of all genes (n = 18,000) by their pLI and p(HI) values

	p(HI) < 0.8	p(HI) ≥ 0.8
pLI < 0.8	13681	441
pLI ≥ 0.8	3258	620

b. Breakdown of ClinGen haploinsufficient genes (n = 148) by their pLI and p(HI) values

	p(HI) < 0.8	p(HI) ≥ 0.8
pLI < 0.8	29	10
pLI ≥ 0.8	68	41

c. Enrichment of ClinGen haploinsufficient genes in each pLI and p(HI) category

	p(HI) < 0.8	p(HI) ≥ 0.8
pLI < 0.8	1.0	10.8
pLI ≥ 0.8	10.0	31.6

Supplementary Information Table 14. Probability of Loss of Function (pLI) and Probability of Haploinsufficient) p(HI)) counts for all genes and ClinGen.

The breakdown of all genes (a) and ClinGen haploinsufficient genes (b) by their pLI and p(HI) values. (c) The enrichment of ClinGen haploinsufficient genes that fall into the high pLI and p(HI) tails when taking the fraction of ClinGen genes with pLI and p(HI) < 0.8 comapred to all genes.

	PTV Z > 3.891 (n=3,230)	Ratio missing PTVs > 0.9061 (n=3,230)	pLI ≥ 0.9 (n=3,230)
PTV fold enrichment	1.3656	1.9224	1.6290
p-value	5.07×10^{-12}	5.12×10^{-22}	8.31×10^{-20}

Supplementary Information Table 15. The enrichment of *de novo* protein-truncating variants (PTVs) from autism cases with the top loss-of-function-intolerant genes as defined by PTV Z, the ratio of missing protein-truncating variation, and pLI.

Supplementary Information Table 16. 58 pathways that show significant deviations in pLI. Note: Provided as Excel file.

Resource	Links
1000 Genomes Phase3	ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/
ESP	http://evs.gs.washington.edu/EVS/
Allele frequency filtering scripts	https://github.com/macarthur-lab/exac_papers/blob/master/src/1kg_af.py https://github.com/macarthur-lab/exac_papers/blob/master/src/esp_af.py
Clinvar processing scripts	https://github.com/macarthur-lab/clinvar
Disease gene lists	https://github.com/macarthur-lab/gene_lists

Supplementary Information Table 17. Resources and scripts used for mendelian analysis.

		Filter Applied			
Population	**Total**	**ESP global**	**ESP popmax**	**ExAC global**	**ExAC popmax**
AFR	13586.4	1021.6	932.4	482	173.9
AMR	12211.1	1065	1011.7	212.6	135
EAS	12426.3	1335.5	1293.7	340.9	161.4
NFE	11927.8	873.7	859.7	155.7	132.3
SAS	12203.1	1157.1	1138	288.7	169.4

Supplementary Information Table 18. Number of missense and protein-truncating variants per individual.

Mean number of missense, protein-truncating and equivalent variants per exome in 100 exomes from each of the populations in the leftmost column, total and after applying any of four filters for <0.1% allele frequency. The plot of filtered variants appears in Figure 4a.

Supplementary Information Table 19. HGMD variants with >1% global allele frequency variant curation. Note: Provided Excel file.

Supplementary Information Table 20. HGMD variants with >1% population (AMR and SAS) allele frequency variant curation. Note: Provided Excel file. Over time, some of the DM HGMD variants have been updated in the HGMD database. The classifications as of 8/15/2015 are noted in the table.

	absent	AC 1	AC 2-10	0.01%	0.10%	1%	10%+
Total	106887	8016	9360	3470	1346	231	61
HGMD	99211	7543	8792	3258	1224	147	27
ClinVar	25890	2649	3479	1237	394	107	37
… of which literature-only (OMIM and GeneReviews)	11243	1319	1779	625	229	77	27
Homozygotes present in ExAC	N/A	0	41	335	909	213	57
Autosomal dominant genes	31345	1167	1255	537	206	27	8
Autosomal recessive genes	24488	3527	4162	1205	380	70	12
X linked genes (dominant or recessive)	14508	151	104	70	43	3	0
Inheritance mode not annotated	36546	3171	3839	1658	717	131	41

Supplementary Information Table 21. ExAC frequencies of reportedly pathogenic variants.

Note: AC 1 = allele count of 1 (singleton), AC 2-10 = allele count of 2-10. All gene lists are based on Blekhman detailed in Supplementary Information Table 12.

#	Demographics	History/Exam	Family History	Liver Testing
1	56 yo Female	Type II diabetes diagnosed at age 40, complicated by neuropathy and hypertension since age 50. She did not self-report liver disease. On physical exam at age 56, no signs of liver disease, no jaundice.	7 siblings, 6 with type II diabetes mellitus, 3 deceased from diabetic complications. None of the siblings were reported to have liver disease.	Normal liver function testing. Test value (normal range): • AST 28 U/L (0-32) • ALT 12 U/L (0-33) • Total bilirubin 0.3 mg/dL (0-1) • Direct bilirubin 0.07 mg/dL (0.2-0.6) • Total protein 7.3 g/dL (6-8.3) • Albumin 4.3 g/dL (3.5-5.2)
2	53 yo Female	Generally healthy. Control in study.	9 siblings, 3 with hypertension, otherwise healthy. None of the siblings were reported to have liver disease.	Normal liver function testing • AST 20 mg/dL • ALT 17 mg/dL • GGT 9 mg/dL Abdominal US with normal liver and biliary tract.
3	62 yo Male	Type II diabetes. Deceased. Myocardial infarction.	Unknown	Not done
4	81 yo Female	Type II diabetes.	Unknown	Not done

Supplementary Information Table 22. ExAC Cases homozygous for CIRH1A p.R565W (16:69199289 C / T) from SIGMA T2D cohort.

Supporting read data for the 4 homozygous individuals appear http://exac.broadinstitute.org/variant/16-69199289-C-T#all and https://github.com/macarthur-lab/exac_2015/blob/master/data/cirh1a_r565w_hom_screenshots.pdf

AST = asparate aminotransferase, ALT = alanine aminotransferase, GGT = Gamma-glutamyl transferase

Filter	Description	Number of variants
ANC_ALLELE	Alternate allele is ancestral state	52
END_TRUNC	Variant falls in last 5% of transcript	10900
EXON_INTRON_UNDEF	Exon or intron boundaries undefined for this transcript	235
NON_CAN_SPLICE	Variant falls in non-canonical splice site	920
NON_CAN_SPLICE_SURR	Variant falls in exon with non-canonical splice site	1134
Multiple filters	More than one of the above filters is applied	166

Supplementary Information Table 23. Number of variant filtered by LOFTEE.

Flag	Description	Number of variants
NAGNAG_SITE	Splice acceptor has in-frame AG acceptor site one codon away	1570
PHYLOCSF_UNLIKELY_ORF	Variant in transcript with a more likely reading frame based on its conservation pattern	1565
PHYLOCSF_WEAK	Variant in transcript with a dubious protein-coding status based on its conservation pattern	16001
SINGLE_EXON	Variant falls in a single exon transcript	9532
Multiple flags	More than one of the above flags is applied	111

Supplementary Information Table 24. Number of variant flagged by LOFTEE.

	African (AFR)	Latino (AMR)	East Asian (EAS)	Finnish (FIN)	European (NFE)	South Asian (SAS)
Heterozygous	102.22	75.58	81.07	74.16	76.67	81.62
Homozygous	38.53	43.2	42.88	39.67	39.8	40.15

Supplementary Information Table 25. Number of protein-truncating variants per individual by population.