Figure Legends

Figure 1 - Schematic workflow.
ALoFT uses a VCF file as input and annotates premature Stop, frameshift-causing indel and canonical splice-site mutations with functional, conservation, network, mismapping and annotation issue features. Then it predicts the pathogenicity (as either benign, recessive or dominant disease-causing) of premature stop mutations using a model trained on known data. ALoFT can also take a 5-column tab-delimited file containing chromosome, position, variant ID, Ref allele and Alt allele columns as input.

Figure 2 - ALoFT classification of 1000 Genomes, ESP6500 and HGMD variants.
a) Benign LoF score for premature stop variants in 1000 Genomes Phase1 data set (1KG) and HGMD. For this plot, we randomly selected one variant per gene. b) HGMD and 1KG premature stop variants on the dominant disease-causing gene - NF2. The benign 1KG LoF variant truncates 2 isoforms, whereas HGMD LoF variants truncate 7 to 12 isoforms. The red triangle denotes the disease-causing LoF variants that affect 7 isoforms and are different from the isoforms affected by 1KG LoF variant. c) Relative positive of premature stop variants on coding transcripts and their allele frequencies. Compared to HGMD variants, 1KG and ESP6500 LoF variants are enriched in the last 5% of the coding sequence. d) Predicted benign LoF scores for premature stop variants in the last coding exon.

Figure 3 – AloFT classification of premature stop variants from Mendelian disease, autism and cancer studies.
a) ALoFT dominant LoF score, GERP and CADD score for Mendelian disease mutations obtained from the Center for Mendelian Genomics studies. b) The top two panels show the dominant LoF scores of *de novo* nonsense mutations in autism patients and siblings; mutations in patients are further separated by gender, as shown in yellow background in the bottom two panels. c) For cancer somatic mutations with predicted diseasing-causing score higher or equal to the threshold of 0.33 (X-axis), we calculated the fraction of mutations occurring in various gene categories. We calculated the fraction of somatic premature stop mutations in 504 known cancer driver genes and 504 randomly selected genes. To ensure that the cancer driver genes and the selected random genes have similar length distributions, the 504 random genes were selected from genes with matched length. We also made sure that the randomly picked genes are sampled from genes known to have premature stop polymorphisms from the 1000 Genomes cohort. Similarly, we compared the fraction of somatic premature stop mutations in 397 LoF-tolerant genes and 397 randomly selected genes with similar length distribution. LoF-tolerant genes are genes that have at least one homozygous LoF variant in at least one individual in the 1000 Genomes cohort.

Supplementary Materials and Methods

1. Description of ALoFT annotation pipeline

ALoFT provides extensive annotation for SNPs that introduce a premature-stop codon, SNPs affecting the splice sites and indels that lead to frameshifts. Initial sequence-based annotation of the coding variants is performed by Variant Annotation Tool[1]. The output of VAT is augmented with various features specific to LoF variants. The input files can be in VCF format or a tab-delimited 5-column file that includes

chromosome, variant position, variant ID, reference allele and alternate allele. LoF variants annotated with various features are output as three separate files.

a. A VCF-formatted file containing summarized annotations.

b. Tab-delimited file containing extensive annotations for premature Stop variants and indels leading to frameshift.

c. Tab-delimited file containing annotations for variants that affect the canonical splice sites.

The output of ALoFT annotation pipeline is discussed below and the overview of the pipeline is shown in Figure S1.

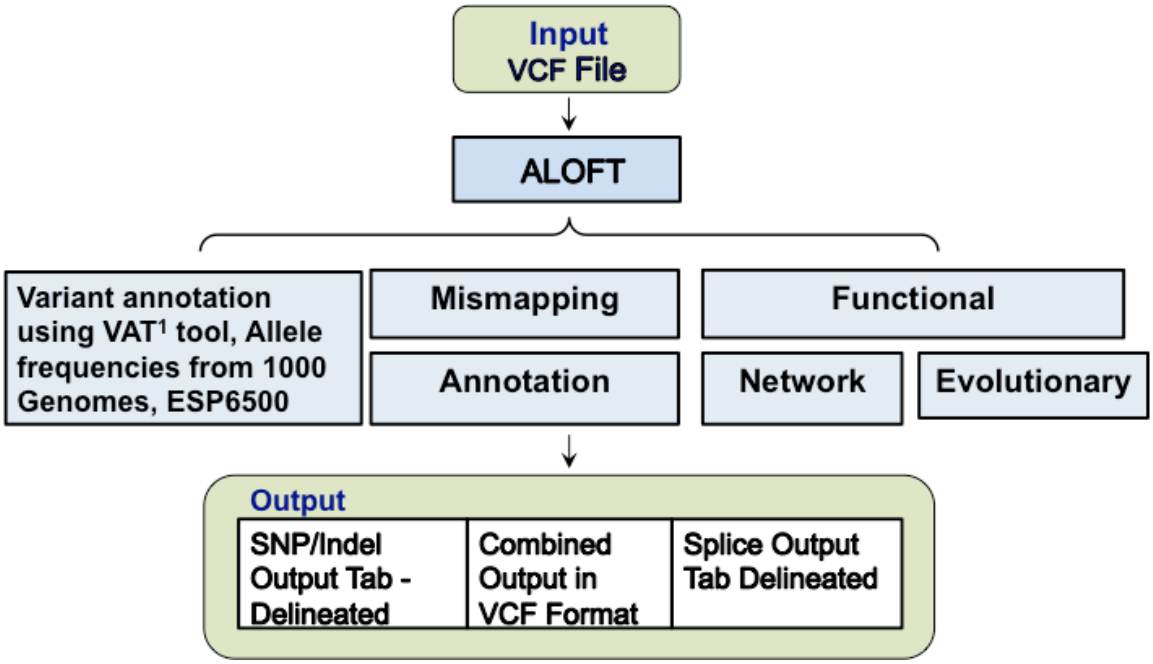

Figure S1 - ALoFT annotation pipeline

1.1 Functional features

We annotated domains affected by the LoF variants with PFAM and SMART domain information. The 3D structure of the protein is essential for proper folding and function of proteins. Therefore, we incorporated structure-based features: SCOP domains and disordered residues into our pipeline. In addition, we annotated signal peptide and trans-membrane domains. PFAM, SCOP, signal peptide and trans-membrane domain annotations were obtained by querying Ensembl Release 73 using the Ensembl PERL API[2]. Post-translationally modified residues (phosphorylated, acetylated, and ubiquitinated sites) are annotated based on data from PhosphositePlus [3]. Disordered residues have been known to be important in protein-protein interaction surfaces and have been implicated in disease-causing mechanisms[4,5]. We obtained disordered residues in proteins using DISOPRED[6]. For all functional features, we assessed if the premature stop variant affected a functional feature and if the region lost due to the premature Stop led to loss of functional domains/features. We also identified

transcripts containing a premature Stop as candidates for nonsense-mediated decay (NMD) if the distance of the premature Stop from the last exon-exon junction was greater than 50 base pairs.

1.2 Network features

We calculated proximity parameters for each LoF-affected gene that correspond to the number of disease genes directly connected to it in a protein-protein interaction network. Human protein-protein interaction networks were downloaded from BioGrid[7] (the version used is BIOGRID-ORGANISM-Homo_sapiens-3.2.95). The list of dominant and recessive disease genes were obtained from the list curated from OMIM[8,9]. Shortest path from a gene to the nearest disease gene are also included in the ALoFT output.

1.3 Evolutionary features

ALoFT includes GERP score of the LoF variant position. In case of indels, the mean GERP score is provided. In addition, ALoFT evaluates the evolutionary conservation of the region that is lost due to the truncation. This is calculated as the percentage of region lost that occurs in GERP-constrained elements. dN/dS values for human-macaque and human-mouse orthologs were obtained from Ensembl using Biomart[10].

1.4 Mismapping errors

ALoFT flags potential false positive variant calls by identifying homologous regions in the genome where the potential for mismapping is high. The following features are annotated:

a. Variants in segmentally duplicated regions
b. Variants in genes that have paralogs
c. Variants in genes that have pseudogenes

Paralogs of human genes were obtained from Ensembl. Pseudogene information was derived from the GENCODE pseudogene resource[11].

1.5 Annotation errors

Variants that lead to a premature Stop codon, indels that lead to frameshift and variants in splice sites are annotated as LoF variants based on sequence annotation and are under assumed to lead to LoF. However, this assumption is not always valid. The various ways where the inferred LoF annotation might not be correct is captured under the following flags:

a. lof_anc: Indicates that the LoF variant allele is the same as the ancestral allele and is likely to be a functional allele.
b. near_start: The variant is in the first 5% of the coding sequence.
c. near_end: The variant is in the last 5% of the coding sequence.
d. alt_canonical_site: SNPs in splice sites are flagged as potentially not LoF when the alternate allele represents the canonical splice site i.e when the alternate allele is GT at the donor or AG at the acceptor site.
e. noncanonical_splice_flank: Variants in exons that are flanked by noncanonical splice sites. Some of these exons could be due to spurious exon annotations in the gene models.
f. Small_intron: Variants in introns less than 15 bp long

1.6 Other features

Aloft includes all the annotation features derived from VAT. This includes transcript-specific annotation of the coding SNP. In addition, ALoFT provides allele frequency information for the variants based on reference population studies, specifically, ALoFT

output includes allele frequency information for LoF variants from the Phase1 1000 Genomes as well as ESP6500 datasets. ESP6500 dataset was downloaded from Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) [November 8, 2013]. An overview of all the features output by ALOFT is shown in Table S1.

MISMAPPING	Number of paralogs to genes containing LoF variants. LoF variant in segmental duplication Number of pseudogenes of genes containing LoF variant
ANNOTATION ISSUES	Alternative allele is ancestral allele NAGNAG pattern indicating alternative splice sites Alternative allele is the consensus splice site LoF containing exon flanked by non-canonical splice sites Splice LoF in short introns (<15bp) LoF within first or last 5% of coding sequence
NETWORK	Shortest path to disease-causing gene Proximity parameter
EVOLUTIONARY	GERP score GERP element dN/dS (macaque) dN/dS (mouse) Percentage of conserved exons removed due to truncation. Calculated as the fraction of removed exons covered by GERP-constraint elements
FUNCTIONAL INTERPRETATION	NMD prediction LoF in PFAM, SMART domains PFAM, SMART domain lost due to truncation LoF in trans-membrane, signal peptides Transmembrane domain, signal peptides lost due to truncation LoF in SCOP domain, disordered region SCOP domain, disordered region lost due to truncation LoF in post-translational modified sites (PTM) PTM lost due to truncation
OTHER	1000 Genomes, ESP6500 allele frequency Partial/full LoF (LoF affecting some isoforms of a gene/ all isoforms) Coding variant annotations using VAT[1] tool

Table S1: Features output by ALoFT for LoF variants.

2. Pathogenicity prediction for LoF mutations

To predict pathogenicity of LoF variants, we trained a Random Forest model to differentiate between benign, heterozygous and homozygous disease-causing LoF

variants. For the training data, we only used premature Stop variants because indel calling methods are not yet robust. Benign variants were derived from 1000 Genomes Phase1 dataset, comprising of 1,092 individuals. Premature Stop mutations leading to disease were obtained from HGMD. We used the variation-specific and gene features that are output by ALoFT to build the classifier. Background mutation rates vary amongst genes. Therefore, we also included the following gene/transcript-specific features, which take into account the effects of length and the background mutation rate for each gene. The following gene/transcript-specific features were included:

a. Conservation: We calculated synonymous and non-synonymous SNP density based on variation data from 1000 Genomes Phase1, average GERP scores of synonymous and non-synonymous SNPs, percentage of synonymous and non-synonymous SNPs in GERP-constrained elements, percentage of coding transcript overlapping with constrained GERP elements and average heterozygosity.

b. Network: We obtained gene centrality scores of various networks from Khurana et al.[12]

c. Transcript expression levels in 25 tissues from GTex[13]. For each transcript, we calculated the average expression values across individuals for particular tissue. Tissue specificity is calculated using entropy-based method.

d. Number of validated miRNA binding sites per gene.

In total, we used 101 features to train our model.

2.1 Training data

Benign premature stop varaints are SNPs homozygous in at least one individual in the Phase1 1000 Genomes. Nonsense SNPs from HGMD are classified as those causing recessive or dominant disease based on 'recessive' and 'dominant' genes curated from the Online Mendelian Inheritance in Man database, OMIM[9,14]. Mutations that lead to dominant inheritance of diseases can do so both via loss of function as well as gain of function mechanisms. However, it is reasonable to assume that LoF variants in dominant disease genes are most likely to result in LoF. Nonetheless, we only included dominant genes predicted to be haplo-insufficient[15] in the training data to make sure that we are predominantly probing loss-of-function effects. The final training dataset includes 404 (in 387 genes) benign variants, 2,365 (in 117 genes) dominant and 4,837 (in 665 genes) recessive premature Stop mutations.

2.2 Three-class classification

Descriptive features are transformed into binary values - "-1" and "1", e.g. whether truncating PFAM domain. Missing values are replaced with weighted average of three classes. We then use random forest algorithm to train our model and evaluate the performance with 10-fold cross-validation. To reduce bias, we included only one variant per gene in the training data (except for dominant genes, we randomly selected three variants per gene. The average number of dominant mutations per gene is 20). The variant is picked randomly from the list of mutations and the longest affected transcript is used. Thus each training model was based on 387 benign premature stop variants, xx# of dominant mutations and 665 recessive mutations. We repeated this process 40 times. We calculated multi-class AUC using the methodology developed by Hand and Till [16]. We assigned the class with largest probability as the predicted outcome. Figure S2 shows the precision calculations for 5 training models. Precision is calculated as the fraction of true positives among predictions.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

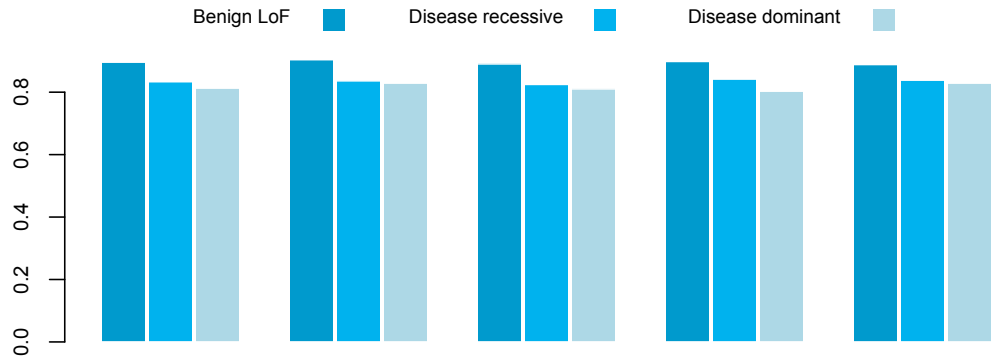

Figure S2 – Precision plot for 5 models.

Figures S3a and S3b are feature importance plots. In Figure S3a, the feature importance is calculated by randomly permuting the values of a feature, training the model with the permuted values for the feature, and calculating the change in mean accuracy of classification as each feature is probed. Figure S3b is an analogous importance plot where the mean change in Gini coefficient is calculated to assess the importance of a feature. The importance plot is not directly interpretable because many variables are correlated. Therefore, evaluating each feature individually is not meaningful.

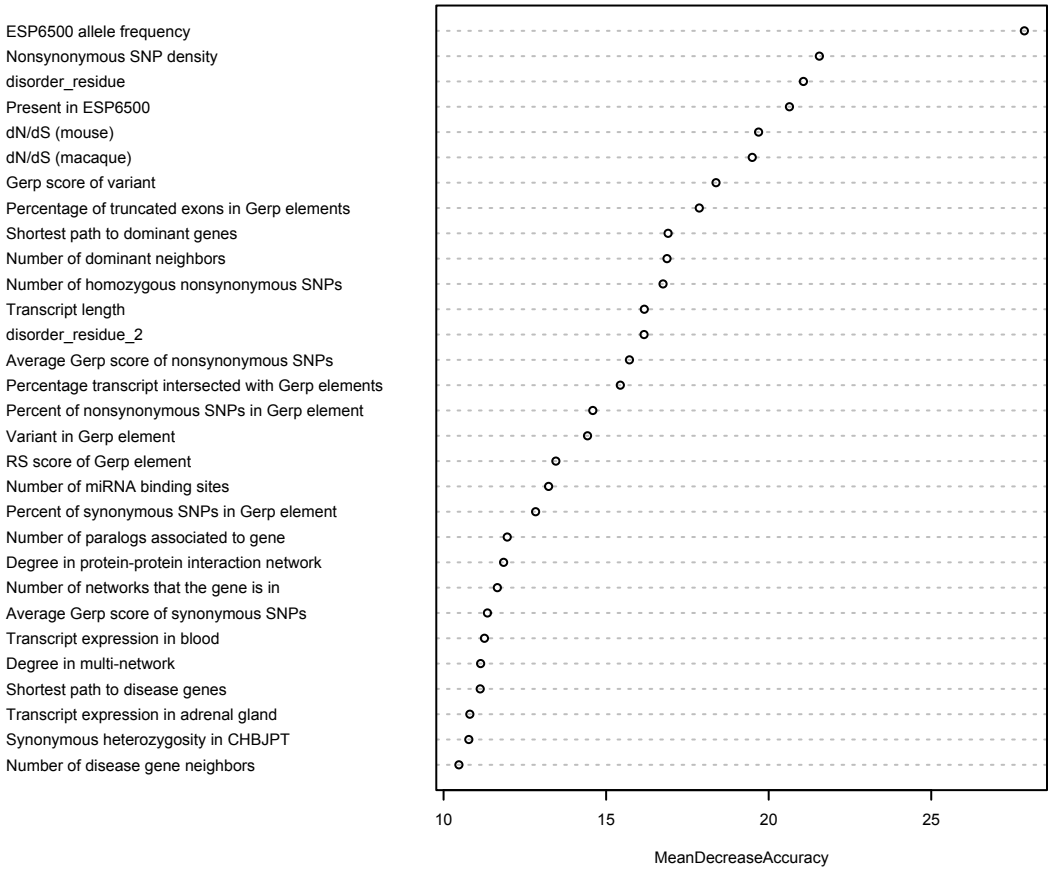

Figure S3a - Importance plot

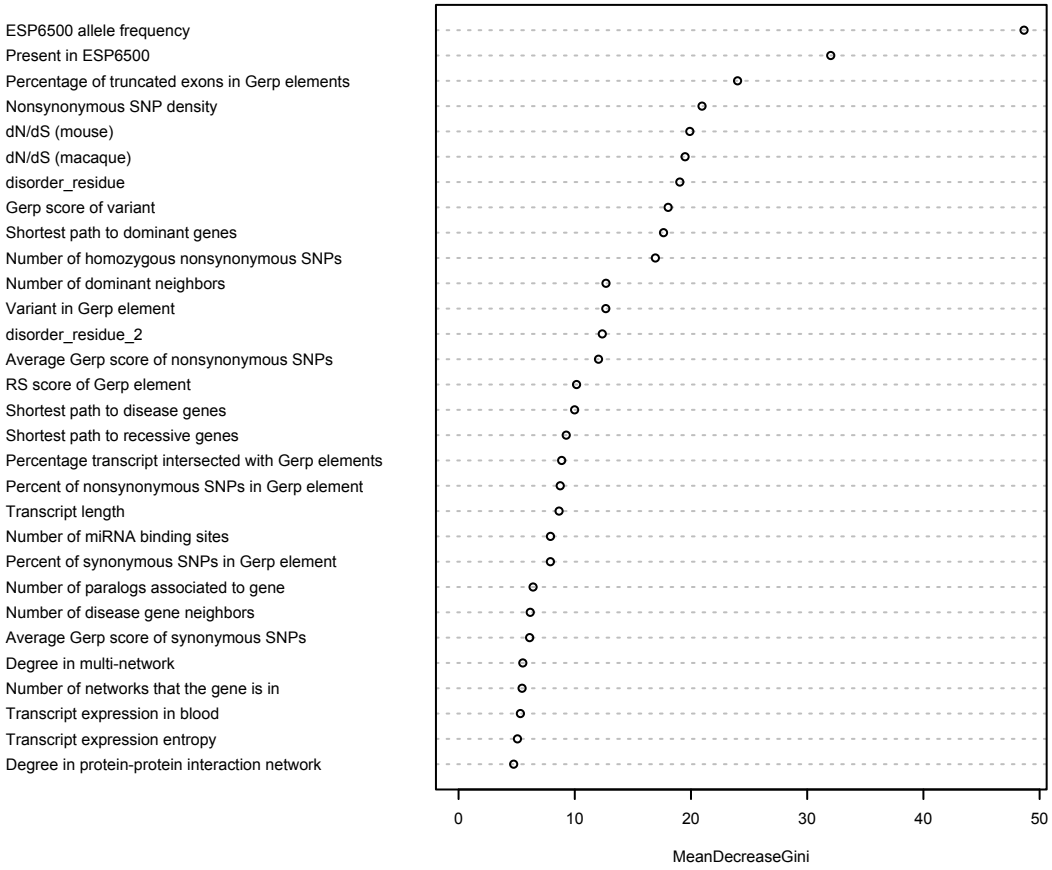

Figure S3b - Importance plot

We also ran the classifier on several different training datasets. As showed in Table S2, the method is pretty robust. Here are the details of the various runs.

No.	Training datasets	Multi-class AUC
1	Dominant genes (includes only haploinsufficient genes); Longest transcript	0.955
2	Same as 1, except removed all Olfactory receptor genes	0.954
3	Same as 1, except randomly picked transcript	0.952
4	Same as 1, except used all dominant genes (without haploinsufficiency filter)	0.921

Table S2 - Robustness of method with respect to training data

2.3 Application of prediction model

2.3.1 Applied to known disease-causing mutations from CMG

We applied our method on known pathogenic mutations from published Center For Mendelian Genomics studies (http://data.mendelian.org/CMG/), which contain 4

dominant and 9 recessive stop-gained mutations. We also obtained GERP and CADD[17] score for these variants.

2.3.2 Applied to 1000 Genomes Phase1 data

We applied our method to the healthy cohort of 1,092 individuals from the Phase1 1000 Genomes data. Among the 6,069 stop-gained mutations, 107, 2639 and 3323 mutations are predicted as dominant, recessive and tolerant respectively (Table S3). Detailed results are available at Table S6.

Predictions	Number of premature-stop mutations (total 6,069)
Dominant	253 - 4.17%
Recessive	2,823 - 46.5%
Benign	2,993 - 49.3%

Table S3 - Pathogenicity prediction for premature-stop mutations from1000 Genomes.

We also calculated per individual statistics for predicted dominant, recessive and benign premature stop mutations (Table S4). We counted the number of alternative alleles for each category.

Predictions	Average alternative allele counts per individual (percentage)
Dominant	0.89 (0.64%)
Recessive	8.3 (5.99%)
Benign	127.9 (93.4%)

Table S4 - Average per individual statistics for 1000 Genomes.

2.3.3 Applied to *de novo* mutations from autism patient

We collected *de novo* stop-gained mutations from four autism studies[18-21]. There are 19 and 53 mutations in siblings and probands respectively. Most individuals have one de-novo premature stop mutation (Table S5). The prediction results are included in Table S7.

	Number of premature-stop mutations
Siblings	19 samples (1 mutation)
Autism males	33 samples (1 mutation); 2 samples (2 mutations)
Autism females	14 samples (1 mutation); 1 sample (2 mutations)

Table S5 - number of de novo premature-stop mutations per individual

We obtained the 33 confident autism genes (FDR<0.1) from Rubeis et al.,[22]. Premature-stop mutations in these genes show significantly higher scores than premature-stop mutations in other genes (Only *de novo* LoFs in probands are used; p-value:0.008; Wilcoxon rank-sum test).

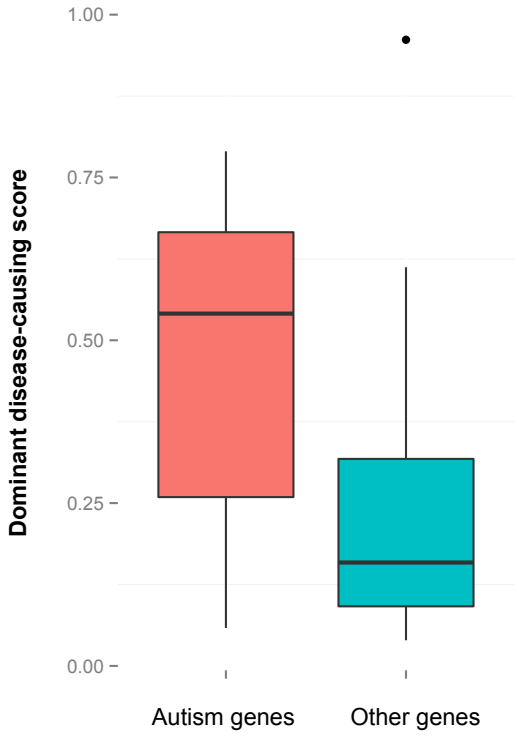

Figure S5 - Prediction scores for autism *de novo* LoFs in confident risk genes.

2.3.4 Applied to somatic mutations from cancer genome sequencing

We obtained somatic premature-stop mutations from Alexandrov et al[23]. This includes ~6,000 exomes in 30 different cancer types. Cancer genes are from COSMIC cancer gene consensus[24].

References

1 Habegger, L. *et al.* VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics* **28**, 2267-2269, doi:10.1093/bioinformatics/bts368 (2012).

2 Flicek, P. *et al.* Ensembl 2013. *Nucleic acids research* **41**, D48-55, doi:10.1093/nar/gks1236 (2013).

3 Hornbeck, P. V. *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic acids research* **40**, D261-270, doi:10.1093/nar/gkr1122 (2012).

4 Vacic, V. *et al.* Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLoS computational biology* **8**, e1002709, doi:10.1371/journal.pcbi.1002709 (2012).

5 Dunker, A. K. & Obradovic, Z. The protein trinity--linking function and disorder. *Nature biotechnology* **19**, 805-806, doi:10.1038/nbt0901-805 (2001).

6 Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F. & Jones, D. T. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**, 2138-2139, doi:10.1093/bioinformatics/bth195 (2004).
7 Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic acids research* **34**, D535-539, doi:10.1093/nar/gkj109 (2006).
8 Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research* **33**, D514-517, doi:10.1093/nar/gki033 (2005).
9 Blekhman, R. *et al.* Natural selection on genes that underlie human disease susceptibility. *Current biology : CB* **18**, 883-889, doi:10.1016/j.cub.2008.04.074 (2008).
10 Flicek, P. *et al.* Ensembl 2014. *Nucleic acids research* **42**, D749-755, doi:10.1093/nar/gkt1196 (2014).
11 Pei, B. *et al.* The GENCODE pseudogene resource. *Genome biology* **13**, R51, doi:10.1186/gb-2012-13-9-r51 (2012).
12 Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587, doi:10.1126/science.1235587 (2013).
13 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nature genetics* **45**, 580-585, doi:10.1038/ng.2653 (2013).
14 Boone, P. M. *et al.* Deletions of recessive disease genes: CNV contribution to carrier states and disease-causing alleles. *Genome research* **23**, 1383-1394, doi:10.1101/gr.156075.113 (2013).
15 Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS genetics* **6**, e1001154, doi:10.1371/journal.pgen.1001154 (2010).
16 Hand, D. J. & Till, R. J. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach Learn* **45**, 171-186, doi:Doi 10.1023/A:1010920819831 (2001).
17 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310-315, doi:10.1038/ng.2892 (2014).
18 Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-299, doi:10.1016/j.neuron.2012.04.009 (2012).
19 Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241, doi:10.1038/nature10945 (2012).
20 Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-245, doi:10.1038/nature11011 (2012).
21 O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-250, doi:10.1038/nature10989 (2012).
22 De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215, doi:10.1038/nature13772 (2014).

23 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).
24 Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research*, doi:10.1093/nar/gku1075 (2014).