**Methods**

**Abbreviations**

SNP: Single Nucleotide Polymorphism

pLoF: Putative Loss of Function

VAT: Variant Annotation Tool

PFAM: the Protein Families Database

SMART: Simple Modular Architecture Research Tool

SCOP: Structural Classification of Proteins

NMD: Nonsense-Mediated Decay

1KG: 1000 Genomes Phase 1

### AUC: Area Under the Curve

CMG: the Center for Mendelian Genomics

**1. Description of ALoFT annotation pipeline**

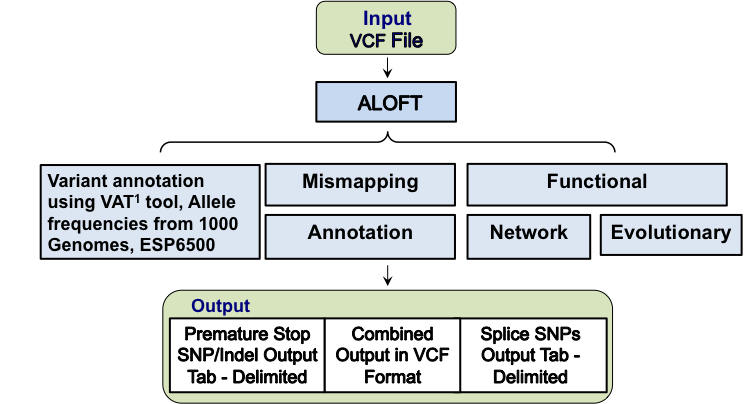
ALoFT provides extensive annotation for SNPs that introduce a premature stop codon, SNPs affecting splice sites and indels that lead to frameshift. Initial sequence-based annotation of the coding variants is performed by the Variant Annotation Tool1 (VAT). The output of VAT is augmented with various features specific to pLoF 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 Supplementary Figure 1.



**Supplementary Figure 1 - ALoFT annotation pipeline**

1.1 Functional features

We annotated domains affected by the pLoF variants with PFAM and SMART domain information. The 3D structure of a protein is essential for proper folding and function of proteins. Therefore, we incorporated two 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 API2. Post-translationally modified residues (phosphorylated, acetylated, and ubiquitinated sites) are annotated based on data from PhosphositePlus3. Disordered residues have been known to be important in protein-protein interaction surfaces and have been implicated in disease-causing mechanisms4,5. We obtained disordered residues in proteins using DISOPRED6. For all functional features, we addressed the following questions: 1) does the premature stop variant affect a functional feature? and 2) are functional, structural or other domains removed due to truncation? 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 pairs7-9.

1.2 Network features

We calculated proximity parameters for each pLoF-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 BioGrid10 (the version used is BIOGRID-ORGANISM-Homo\_sapiens-3.2.95). Dominant and recessive disease genes were obtained from lists curated from OMIM11-13. Shortest path from a gene to the nearest disease gene in the protein-protein interaction network is also included in the ALoFT output.

1.3 Evolutionary features

ALoFT includes GERP score of the pLoF 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 coding region lost that occurs in GERP-constrained elements. dN/dS values for human-macaque and human-mouse orthologs were obtained from Ensembl using Biomart.

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:

1. Variants in segmentally duplicated regions
2. Variants in genes that have paralogs
3. Variants in genes that have pseudogenes

Paralogs of human genes were obtained from Ensembl. Pseudogene information was derived from the GENCODE pseudogene resource14.

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 pLoF variants based on sequence annotation and are under assumed to lead to loss-of-function. 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:

1. lof\_anc: Indicates that the pLoF variant allele is the same as the ancestral allele and is likely to be a functional allele.
2. near\_start: The variant is in the first 5% of the coding sequence.
3. near\_end: The variant is in the last 5% of the coding sequence.
4. 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).
5. 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.
6. 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 of 1000 Genomes Project (1KG), ESP6500, as well as ExAC datasets. 1KG includes genetic variation data obtained from whole genome and exome sequencing of 1,092 healthy individuals. ESP6500 consists of genetic variants from exome sequencing of a cohort of 2,203 African-American and 4,300 European-American unrelated individuals enrolled in the National Heart, Lung, and Blood Institute Exome Sequencing Project15. 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]. Version 0.3 ExAC dataset was downloaded from <http://exac.broadinstitute.org/>, containing 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. An overview of all the features output by ALoFT is shown in Supplementary Table 1.

|  |  |
| --- | --- |
| **MISMAPPING** | Number of paralogs to genes containing pLoF  pLoF variant in segmental duplication  Number of pseudogenes of genes containing pLoF |
| **ANNOTATION ISSUES** | Alternative allele is ancestral allele  NAGNAG pattern indicating alternative splice sites  Alternative allele is the consensus splice site  pLoF containing exon flanked by non-canonical splice sites  Splice variants in short introns (<15bp)  pLoF 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  In PFAM, SMART domains  PFAM, SMART domain lost due to truncation  In trans-membrane, signal peptides  Transmembrane domain, signal peptides lost due to truncation  In SCOP domain, disordered region  SCOP domain, disordered region lost due to truncation  In post-translational modified sites (PTM)  PTM lost due to truncation |
| **OTHER** | 1KG, ESP6500, ExAC allele frequency  Partial/full pLoF (pLoF affecting some isoforms of a gene/ all isoforms)  Coding variant annotations using VAT1 tool |

**Supplementary Table 1: Features output by ALoFT for pLoF variants**

**2. Pathogenicity prediction for pLoF mutations**

To predict pathogenicity of pLoF variants, we trained a Random Forest model to differentiate between benign, heterozygous and homozygous disease-causing premature stop variants. For the training data, we only used premature stop variants caused by single nucleotide polymorphisms because indel calling methods are not yet robust. Benign set includes homozygous variants from 1KG. Premature stop mutations leading to diseases were obtained from HGMD. We used the variation and gene specific features that are output by ALoFT to build the classifier. 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 1KG, 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 for synonymous and non-synonymous SNPs in 1KG.

b. Network: We obtained gene centrality scores of various networks from Khurana et al.16

c. Transcript expression levels in 25 tissues from GTex17. For each transcript, we calculated the average expression values across individuals for particular tissue. Tissue specificity is calculated using Shannon entropy-based method18,19.

d. Number of validated miRNA binding sites per gene obtained from miRWalk20.

e. Average heterozygosity is calculated as , where p is minor allele frequency, q is the reference allele frequency, l is the length of coding transcript.

In total, we used 108 features to train our model (Supplementary Table 2 – attached as additional file).

2.1 Training data

Benign premature stop variants are SNPs homozygous in at least one individual in 1KG. Premature stop mutations 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, OMIM12,13. The training data consists of variants from autosomes only. 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 most pLoF variants in dominant disease genes cause loss-of-function. Nonetheless, we only included dominant genes predicted to be haplo-insufficient21 in the training data to make sure that we are predominantly probing loss-of-function effects. The final training dataset was derived from 397 (in 380 genes) benign variants, 3,300 (in 136 genes) dominant and 5,342 (in 796 genes) recessive premature stop mutations (Supplementary Table 3).

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Used as | # Genes | # Variants |
| Benign homozygous variants from 1KG | Training | 380 | 397 |
| Heterozygous1KG variants | Test | 3,997 | 5,170 |
| HGMD dominant | Training | 136 | 3,300 |
| HGMD recessive | Training | 796 | 5,342 |
| Other HGMD variants (excluded training variants) | Test | 868 | 6,145 |

**Supplementary Table 3 – HGMD and 1KG premature stop variants used for training and testing**

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-validations. To reduce bias, we included only one variant per gene in the training data for the benign and recessive classes. The average number of dominant mutations per gene is 24 (Supplementary Table 3). Therefore, we randomly selected three variants per gene for the dominant class in order to obtain a reasonably balanced training dataset. The variant is picked randomly from the list of mutations and the longest affected transcript is used. Thus, each training model was based on 380 benign premature stop variants, ~341 dominant mutations and 796 recessive mutations. Stratified sampling is used in random forest to achieve balanced three-class training.

We repeated this process 40 times. We calculated multi-class AUC for the test set using the methodology developed by Hand and Till22. We assigned the class with highest probability as the predicted outcome. Supplementary Figure 2 shows the precision calculations for 5 out of the 40 training models. Precision is calculated as the fraction of true positives among predictions. As an example, for recessive predictions, we counted the number of correct predictions as true positives, the rest of the recessive predictions as false positives.

**Precision**



**Supplementary Figure 2 – Precision plot for 5 models**

We also used several different training datasets for the prediction (Supplementary Table 4). The different training settings were used to test whether the classifier is robust to different training datasets.

|  |  |  |
| --- | --- | --- |
| No. | Training datasets | Multi-class AUC |
| 1 | Dominant genes (includes only haplo-insufficient genes); Longest transcript | 0.970 |
| 2 | Same as 1, except removed all Olfactory receptor genes | 0.969 |
| 3 | Same as 1, except randomly picked transcript | 0.968 |
| 4 | Same as 1,except used all dominant genes (without haplo-insufficiency filter) | 0.934 |

**Supplementary Table 4 - Robustness of method with respect to training data**

In Supplementary Figure 3, the importance of a feature is calculated by evaluating the decrease in mean accuracy of the test set when the value of the feature is randomly permuted. The importance plot is not directly interpretable because some of the prediction variables are correlated. The description of the features can be found at Supplementary Table 2.

To further evaluate the features important for the classification, we built several prediction models using different sets of features for the training. Supplementary Table 5 shows the features used for prediction and their corresponding multi-class AUC of the test set.

**Supplementary Figure 3 - Importance plot**

|  |  |  |
| --- | --- | --- |
| No. | Features | Multi-class AUC |
| 1 | ESP6500 and ExAC features (Allele frequency, presence/absence of allele) | 0.787 |
| 2 | Network features | 0.807 |
| 3 | Functional features | 0.852 |
| 4 | Evolutionary features | 0.864 |
| 7 | All features | 0.970 |

**Supplementary Table 5 – Classifier performance evaluation**

2.3 Application of prediction model

2.3.1 Applied to known disease-causing mutations from the Center for Mendelian Genomics studies (<http://data.mendelian.org/CMG/>)

We applied our method to classify Mendelian pathogenic mutations discovered in the Center For Mendelian Genomics studies. After excluding training variants, there are 3 dominant and 5 recessive premature stop mutations. We also obtained GERP and CADD23 scores for these variants.



**Supplementary Figure 4 - CMG prediction**

2.3.2 Applied to known disease-causing mutations from ClinVar

ClinVar24 variants were obtained from <https://github.com/macarthur-lab/clinvar> . In order to validate ALoFT predictions, we first excluded all ClinVar variants in genes that were used in the training set. We then labeled the remaining ClinVar variants as those leading to disease via the dominant or recessive mode of inheritance using an orthogonal list of dominant/recessive genes obtained from Berg et al.25 To avoid potential bias that might arise due to enrichment of disease variants in particular genes, we randomly picked one variant per gene for the analysis shown in Figure 2a. The final set used to validate ALoFT contains 197 variants in genes known to cause disease through the dominant mode of inheritance and 111 variants in recessive genes.

2.3.3 Applied to 1KG data

We applied our method to classify premature stop variants in the healthy cohort of 1,092 individuals from the 1KG data. Among the 5,495 premature stop variants (excluding chrX), 148, 3070 and 2277 variants are predicted as dominant, recessive and tolerant, respectively (Supplementary Table 6 - as additional file and Supplementary Table 7).

We used high confidence variants for the calculation of per individual statistics as described below. 1) While ALoFT provides several flags that identifies likely false positive variant calls arising due to mismapping and annotation errors, we conservatively excluded only those pLoF variants that correspond to the ancestral allele as they are unlikely to result in loss-of-function. 2) Variants present at > 1% frequency in either the European or African American population of the 1KG cohort, but absent in the ESP6500 cohort were also removed as likely erroneous calls. 3) For the 1KG set, only variants called from exome sequencing were included in order to make a fair comparison with the ESP6500 data that is also based on exome capture. We calculated per individual statistics for predicted dominant, recessive and benign premature stop mutations and is shown in Supplementary Table 7 and Supplementary Figure 5. Per individual calculations are based on 246 individuals of African ancestry and 379 individuals of European ancestry. [[[to add ExAC calculation …] ]]

|  |  |  |
| --- | --- | --- |
| Predictions | Number of premature stop variants | Number of premature stop alleles per individual  Mean (Median) |
| Dominant | 148 | 0.11 (0) |
| Recessive | 3,070 | 2.77 (2) |
| Benign | 2,277 | 50.2 (50) |

**Supplementary Table 7 - Pathogenicity predictions of premature stop variants and average per individual statistics for 1KG cohort.**



**Supplementary Figure 5 - Distribution of predicted dominant and recessive premature stop alleles in the 1KG individuals.**

2.3.3 Applied to *de novo* mutations from autism patient

We collected *de novo* premature stop mutations from four autism studies26-29. There are 19 and 53 mutations in siblings and probands respectively. Most individuals have one *de novo* premature stop mutation (Supplementary Table 8). The prediction results are included in Supplementary Table 9 (as additional file). (2 out of 53 proband mutations overlap our training data, excluded in Figure 2b).

|  |  |
| --- | --- |
|  | Number of *de novo* premature stop mutations  (Number of samples) |
| Siblings | 1 (19) |
| Autism males | 1 (33); 2 (2) |
| Autism females | 1 (14); 2 (1) |

**Supplementary Table 8 - number of *de novo* premature stop mutations per individual**

We obtained the list of 33 confident autism genes (FDR < 0.1) from Rubeis *et al*.30 and observed that dominant disease-causing score for premature stop variants in these genes are significantly higher than those in other genes (Only *de novo* pLoFs in probands are used; p-value: 5e-3; Wilcoxon rank-sum test).



**Supplementary Figure 6 - Prediction scores for autism *de novo* pLoFs in confident risk genes**

2.3.4 Applied to somatic mutations from cancer genome sequencing

We obtained somatic premature stop mutations from Alexandrov et al31.

This includes 6,535 exomes in 30 different cancer types. Cancer genes are from COSMIC cancer gene consensus32.

**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 Cheng, J., Belgrader, P., Zhou, X. & Maquat, L. E. Introns are cis effectors of the nonsense-codon-mediated reduction in nuclear mRNA abundance. *Molecular and cellular biology* **14**, 6317-6325 (1994).

8 Belgrader, P. & Maquat, L. E. Nonsense but not missense mutations can decrease the abundance of nuclear mRNA for the mouse major urinary protein, while both types of mutations can facilitate exon skipping. *Molecular and cellular biology* **14**, 6326-6336 (1994).

9 Thermann, R. *et al.* Binary specification of nonsense codons by splicing and cytoplasmic translation. *The EMBO journal* **17**, 3484-3494, doi:10.1093/emboj/17.12.3484 (1998).

10 Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic acids research* **34**, D535-539, doi:10.1093/nar/gkj109 (2006).

11 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).

12 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).

13 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).

14 Pei, B. *et al.* The GENCODE pseudogene resource. *Genome biology* **13**, R51, doi:10.1186/gb-2012-13-9-r51 (2012).

15 1000 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

16 Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587, doi:10.1126/science.1235587 (2013).

17 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nature genetics* **45**, 580-585, doi:10.1038/ng.2653 (2013).

18 Schug, J. *et al.* Promoter features related to tissue specificity as measured by Shannon entropy. *Genome biology* **6**, R33, doi:10.1186/gb-2005-6-4-r33 (2005).

19 Shannon, C. E. The mathematical theory of communication. 1963. *M.D. computing : computers in medical practice* **14**, 306-317 (1997).

20 Dweep, H., Sticht, C., Pandey, P. & Gretz, N. miRWalk--database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *Journal of biomedical informatics* **44**, 839-847, doi:10.1016/j.jbi.2011.05.002 (2011).

21 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).

22 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).

23 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).

24 Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research* **42**, D980-985, doi:10.1093/nar/gkt1113 (2014).

25 Berg, J. S. *et al.* An informatics approach to analyzing the incidentalome. *Genetics in medicine : official journal of the American College of Medical Genetics* **15**, 36-44, doi:10.1038/gim.2012.112 (2013).

26 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).

27 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).

28 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).

29 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).

30 De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215, doi:10.1038/nature13772 (2014).

31 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

32 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).