Predicting the impact of putative loss-of-function variants in protein-coding genes

Suganthi Balasubramanian[1,2*], Yao Fu[1*], Mayur Pawashe[2], Mike Jin[2], Jeremy Liu[2], Daniel G. MacArthur[3,4], Mark Gerstein[1,2,5]
[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520 USA, [2]Molecular Biophysics and Biochemistry Department, Yale University, New Haven 06520, CT, USA, [3]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA, [4]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA, [5]Department of Computer Science, Yale University, New Haven 06520, CT, USA

Variants predicted to result in the loss of function (LoF) of human genes have recently attracted considerable interest both because of their established clinical impact as well as their surprising prevalence in seemingly healthy humans. To better understand the impact of putative LoF variants, we developed ALoFT (**A**nnotation of **L**oss-**o**f-**F**unction **T**ranscripts), to annotate and predict the disease-causing potential of LoF variants. Our method is able to distinguish between dominant and recessive LoF variants discovered by Mendelian studies. Investigation of premature stop variants discovered in a sample of over 1,000 whole genome-sequenced individuals suggests that each individual carries about two heterozygous LOF alleles that can potentially lead to disease if present in the homozygous state. When applied to *de novo* LoF variants in autism-affected families, ALoFT predicts that the variants are more disruptive in autism patients than in unaffected siblings. Finally, we show that premature stop variants predicted to be pathogenic by ALoFT are enriched in, THE SOM. MUT FOUND IN KNOWN known cancer driver genes.

One of the most notable findings from personal genomics studies is that all individuals harbor LoF variants in some of their genes[1]. A systematic study of LoF variants from 180 individuals revealed that there are over 100 putative LoF variants in each individual[2]. Thus, several genes are knocked out either completely or in an isoform-specific manner in apparently healthy individuals. Remarkably, recent studies have led to the discovery of protective LoF variants associated with beneficial traits. The potential of LoF variants to identify valuable drug targets has fueled an increased interest in a more thorough understanding of putative LoF variants. For example, nonsense variants in PCSK9 are associated with low LDL levels[3,4] which has prompted the active pursuit of the inhibition of PCSK9 as a potential therapeutic for hypercholesterolemia[5-7]. Other examples include nonsense and splice mutations in APOC3 associated with low levels of circulating triglycerides, a nonsense mutation in SLC30A8 resulting in about 65% reduction in risk for Type II diabetes and two splice variants in the Finnish population in LPA that protect from coronary heart disease[8-11].

About 12% of known disease-causing mutations in the Human Gene Mutation Database (HGMD) are due to nonsense mutations[12]. Even though premature stop variants often lead to loss of function and are thus deleterious, predicting the functional impact of premature stop codons is not straightforward. Aberrant transcripts containing premature stop codons are typically removed by nonsense-mediated decay (NMD), an mRNA surveillance mechanism[13]. However, a recent large-scale expression analysis demonstrated that 68% of predicted NMD events due to premature stop variants are unsupported by RNASeq analyses[14]. A study aimed at understanding disease mutations using a 3D structure-based interaction network suggests that truncating mutations can give rise to functional protein products[15]. Moreover, premature stop codons in the last exon are not subject to NMD. Further, when a variant affects only some isoforms of a gene, it is difficult to infer its impact on gene function without the knowledge of the isoforms that are expressed in the tissue of interest and how their levels of expression affect gene function. Finally, loss-of-function of a gene might not have any impact on the fitness of the organism.

We have developed a pipeline called ALoFT (**A**nnotation of **L**oss-**o**f-**F**unction **T**ranscripts), to provide extensive annotation of putative LoF variants. In this study, we include premature stop-causing SNPs, frameshift-causing indels and variants affecting canonical splice sites as putative LoF variants, also referred to as premature truncating variants. An overview of the pipeline is shown in Supplementary Figure 1. The main features of ALoFT include (1) function-based annotations; (2) evolutionary conservation;

and (3) biological network data. For comprehensive functional annotation, we integrated several annotation resources such as PFAM and SMART functional domains[16,17], signal peptide and transmembrane annotations, post-translational modification sites, NMD prediction[18,19], and structure-based features such as SCOP domains and disordered residues. Evolutionary conservation can be used as a proxy for identifying functionally important regions. Therefore, ALoFT provides variant position-specific GERP scores, which is a measure of evolutionary conservation[20]. In addition, we evaluate if the region removed due to the truncation of the coding sequence is evolutionarily conserved based on GERP constraint elements[21]. ALoFT also outputs dN/dS values (ratio of missense to synonymous substitution rates) for macaque and mouse that are computed from human-macaque and human-mouse orthologous alignments, respectively. ALoFT includes two network features shown to be important in disease prediction algorithms: a proximity parameter that gives the number of disease genes connected to a gene in a protein-protein interaction network and the shortest path to the nearest disease gene[2,22]. The pipeline also includes features to help identify erroneous LoF calls, potential mismapping, and annotation errors, because LoF variant calls have been shown to be enriched for annotation and sequencing artifacts[2]. A detailed description of all the annotations provided by ALoFT is included in the Online Methods (Supplementary Table 1). Detailed documentation and github link to source code can be found at aloft.gersteinlab.org.

uses

To understand the impact of putative LoF variants on gene function we developed a prediction method to differentiate between disease-causing and benign variants. While there are several algorithms to predict the effect of missense coding variants on protein function, there is a paucity of methods that are applicable to nonsense variants[23-26]. Additionally, current prediction methods that infer the pathogenicity of variants do not take into account the zygosity of the variant[27,28]. The majority of LoF variants in healthy population cohorts are heterozygous. It is likely that a subset of these variants will cause disease in the recessive state. Therefore, we developed a prediction model to classify premature stop variants into those that are benign, those that lead to recessive disease and those that lead to dominant disease using the annotations output by ALoFT as predictive features (Fig. 1, Online Methods).

To build the ALoFT classifier, we used three classes of variants as training data sets: premature stop variants that are homozygous in at least one individual in the Phase1 1000 Genomes data (1KGP1) that represent benign stop variants, homozygous premature stop mutations from HGMD that lead to recessive disease and heterozygous premature stop variants in haplo-insufficient genes that lead to dominant disease[22,29]. We built the ALoFT classifier to distinguish among the three classes using a random forest algorithm[30]. ALoFT provides class probability estimates for each mutation. We obtain good discrimination between the three classes. The average multiclass test AUC (area under the curve) with 10-fold cross-validation is 0.96. The precision for the three classes are as follows: Dominant=0.85, Recessive=0.84, Benign=0.89. The classifier is robust to the choice of training data sets and performs well with different training data sets (Supplementary Table 2, Supplementary Fig. 2).

We analyzed the feature importance to understand the contribution of different features to the classification (Supplementary Fig. 3). The presence/absence of an allele in 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 Project, ESP6500 cohort, and its frequency appear to be the most important features for the classification. We retrained the random forest model excluding the ESP6500-related features. The classifier still performs well with an average multiclass test AUC =0.93 and the precision for the three classes are as follows: Dominant=0.83, Recessive=0.79, Benign=0.80. We also systematically evaluated the classifier using models trained on specific sets of features. We find that integrating all the features improves prediction accuracy of the classifier and is not dominated by any single feature (Supplementary Table 3).

AF2.

In order to estimate the number of premature stop disease alleles in a healthy individual, we applied ALoFT to 5,495 premature stop variants from the 1KGP1 dataset (Online Methods). The predicted benign LoF score for premature stop variants in this population cohort have a wide range of values (Fig. 2a, Supplementary Table 4). On average, each individual is a carrier of 2.2 rare heterozygous premature stop alleles that are predicted to be disease-causing in the homozygous state (Supplementary Table 5). This is likely to be an upper estimate because the prediction accuracy can be affected by a number of confounding factors that include incomplete penetrance of disease alleles, variable expressivity, genetic modifiers, compensatory mutations, marginal variant calls and imperfect training datasets. Current estimates of the genetic burden of disease alleles in an individual vary widely, ranging from 1.1 recessive alleles per individual to 31 deleterious alleles[31-34].

not Low.

Next, we looked at premature stop variants in the 1KGP1 cohort in known disease-causing genes. We find that variants in 1KGP1 are more likely to be benign compared to known disease-causing mutations in the same genes (Fig. 2a, p-value: 9e-3). The occurrence of LoF variants in known disease genes in healthy individuals can be rationalized as follows: most variants are rare and heterozygous and are predicted to lead to disease only in the homozygous state, some variants affect only some isoforms of a gene that has other unaffected functional isoforms, some variant affects isoform that is not expressed in tissue of interest and some variants give rise to truncated proteins that are functional (for example, variants in the last exon). 12.6% of premature stop variants in the presumed healthy 1KGP1 individuals and the disease-causing variants are on different isoforms. For example, the premature stop variant in NF2 in 1KGP1 affects 2 isoforms, whereas the premature stop mutations in HGMD affect the other 7 isoforms (Fig. 2b). ALoFT predicts that the heterozygous 1KGP1 variant is a benign LoF variant. Heterozygous truncating mutations in NF2 are known to cause the most severe disease, while missense mutations cause milder phenotypes[35]. Therefore, we do not expect to observe any LoF variant in NF2 in the presumed healthy individuals. This suggests that premature stop variants in some NF2 isoforms are not disease-causing.

We next applied ALoFT to predict the effect of premature stop variants in the final exons. It is often assumed that premature stop variants in the last coding exon are likely to be benign because they escape NMD; as a result, in many cases the effect will be the expression of a truncated protein rather than a complete loss of function. However, examples of disease-causing mutations in the last exon are also known[36]. Therefore, we applied ALoFT to see if we could distinguish between benign and disease-causing LoF variants in the last coding exon. To this end, we expanded our analysis to include the ESP6500 and HGMD datasets. A higher proportion of rare variants are observed in ESP6500 cohort due to its larger sample size and higher sequencing depth (Fig. 3a). Nonetheless, a large number of both common and rare premature stop variants are seen at the end of the coding genes in both the 1KGP1 and ESP6500 datasets. In contrast, fewer disease-causing HGMD variants are seen at the ends of coding genes (Fig. 3a).

ALoFT predicts that both common and rare premature stop variants in the last coding exon in the 1KGP1 and ESP6500 cohort are likely to be benign, whereas HGMD mutations tend to be disease-causing (Fig. 3b). Thus, ALoFT is able to differentiate between rare but benign premature stop variants seen in healthy individuals and the rare disease-causing HGMD alleles.

GREAT !

We further evaluated ALoFT by predicting the effect of nonsense mutations in several recently published disease studies. We classified premature stop mutations from the Center For Mendelian Genomics studies and predicted the mode of inheritance and pathogenicity of all of the truncating variants (Fig. 4a). Our method showed that heterozygous disease-causing variants have significantly higher dominant disease-causing scores than the homozygous disease-causing variants (p-value: 5.6e-3; Wilcoxon rank-sum test). We used two other measures, GERP score, which is a measure of evolutionary conservation, and CADD score, which gives a measure of pathogenicity, to classify recessive versus dominant LoF variants[37]. Both CADD and GERP scores are not able to discriminate between recessive and dominant disease-causing mutations (Fig. 4a).

De novo LoF SNPs have been implicated in autism based on analysis of sporadic or simplex families (families with no prior history of autism)[38-41]. We applied our method to *de novo* LoF mutations discovered in these studies. Our method shows that the proportion of dominant disease-causing *de novo* LoF events is significantly higher in autism patients versus siblings (Fig. 4b; p-value: 5.3e-3; Wilcoxon rank-sum test). Previous studies suggest that there is a higher mutational burden in female patients[42]. We observe a similar pattern for LoF mutations – female probands have a higher portion of predicted deleterious *de novo* LoF variants than male probands (p-value: 0.039). A recent study based on exome sequencing of 3,871 autism cases delineated 33 risk genes at FDR < 0.1[43]. We observe that the *de novo* LoF mutations in the autism patients in the 33 risk genes have higher dominant disease causing LoF score than the *de novo* LoF variants in other genes (Supplementary Fig. 5; p-value: 2.8e-3). Supplementary Table 7 includes the ALoFT predictions for *de novo* LoF variants.

BUT LESS MUT

Lastly, we applied our prediction method to infer the effect of somatic premature stop variants from a compilation of ~6,000 cancer exome sequencing studies[44]. As shown in Figure 4c, somatic LoF mutations tend to occur in known cancer driver genes compared to randomly sampled genes whose length distribution matches that of the known driver genes. Moreover, deleterious somatic LoF variants are enriched in driver genes, whereas they are depleted in LoF-tolerant genes (genes that contain at least one homozygous LoF variant in the 1KGP1 population). To classify driver genes as tumor suppressors, Vogelstein proposed a "20/20" rule where a gene is classified as a tumor suppressor if > 20% of the observed mutations in that gene are LoF mutations[45]. Therefore, somatic LoF variants in driver genes are expected to be mostly in tumor suppressors rather than oncogenes. Thus, ALoFT can be used to identify deleterious LoFs in potential tumor suppressor genes.

FURTH. ER

BE ENRICHED IN

In summary, we describe a tool for predicting the impact of premature Stop variants in the context of a diploid model, i.e. discriminating whether premature stop variants are likely to lead to recessive or dominant disease. Better identification and characterization of LoF variants has both diagnostic and therapeutic implications. ALoFT allows for the identification and prioritization of high impact putative disease-causing LoF variants in individual genomes. Integrating benign LoF variants with phenotypic

DISC

information will help us to identify protective LoF variants which are valuable drug targets[46,47]. Lastly, diseases caused by LoF variants provide opportunities for targeted therapy using drugs that either enable read-through of the premature stop, thus restoring the function of the mutant protein, or NMD inhibitors that prevents degradation of the LoF-containing transcript by NMD[48-54]. This is especially useful in the context of rare diseases where targeting the same molecular phenotype leading to different diseases alleviates the need to design a new drug for each individual disease. Further work will be needed both to correlate the predictions of ALoFT with experimental assays of protein loss of function, and to study the phenotypic impact of heterozygous and homozygous LoF variants in large clinical cohorts.

Ab/Nb

References

1. Balasubramanian, S. *et al.* Gene inactivation and its implications for annotation in the era of personal genomics. *Genes Dev* **25**, 1-10 (2011).
2. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-8 (2012).
3. Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat Genet* **37**, 161-5 (2005).
4. Cohen, J.C., Boerwinkle, E., Mosley, T.H., Jr. & Hobbs, H.H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med* **354**, 1264-72 (2006).
5. Banerjee, Y., Shah, K. & Al-Rasadi, K. Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *N Engl J Med* **366**, 2425-6; author reply 2426 (2012).
6. Milazzo, L. & Antinori, S. Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *N Engl J Med* **366**, 2425; author reply 2426 (2012).
7. Stein, E.A. *et al.* Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *N Engl J Med* **366**, 1108-18 (2012).
8. Flannick, J. *et al.* Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat Genet* **46**, 357-63 (2014).
9. Lim, E.T. *et al.* Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet* **10**, e1004494 (2014).
10. Tachmazidou, I. *et al.* A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nat Commun* **4**, 2872 (2013).
11. Timpson, N.J. *et al.* A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. *Nat Commun* **5**, 4871 (2014).
12. Stenson, P.D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* **133**, 1-9 (2014).
13. Isken, O. & Maquat, L.E. Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes Dev* **21**, 1833-56 (2007).
14. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-11 (2013).

15. Guo, Y. *et al.* Dissecting disease inheritance modes in a three-dimensional protein network challenges the "guilt-by-association" principle. *Am J Hum Genet* **93**, 78-89 (2013).
16. Letunic, I., Doerks, T. & Bork, P. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res* (2014).
17. Finn, R.D. *et al.* Pfam: the protein families database. *Nucleic Acids Res* **42**, D222-30 (2014).
18. 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**, 2138-9 (2004).
19. 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 Res* **40**, D261-70 (2012).
20. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-13 (2005).
21. Davydov, E.V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025 (2010).
22. Huang, N., Lee, I., Marcotte, E.M. & Hurles, M.E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* **6**, e1001154 (2010).
23. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20 (2013).
24. Cooper, G.M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* **12**, 628-40 (2011).
25. Karchin, R. Next generation tools for the annotation of human SNPs. *Brief Bioinform* **10**, 35-52 (2009).
26. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-81 (2009).
27. Hu, J. & Ng, P.C. Predicting the effects of frameshifting indels. *Genome Biol* **13**, R9 (2012).
28. Rausell, A. *et al.* Analysis of stop-gain and frameshift variants in human innate immunity genes. *PLoS Comput Biol* **10**, e1003757 (2014).
29. 1000 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
30. Breiman, L. Random Forests. *Machine Learning* **45**, 5-32 (2001).
31. Bell, C.J. *et al.* Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med* **3**, 65ra4 (2011).
32. Chong, J.X., Ouwenga, R., Anderson, R.L., Waggoner, D.J. & Ober, C. A population-based study of autosomal-recessive disease-causing mutations in a founder population. *Am J Hum Genet* **91**, 608-20 (2012).
33. Cooper, D.N. *et al.* Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. *Hum Mutat* **31**, 631-55 (2010).

34. Xue, Y. *et al.* Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am J Hum Genet* **91**, 1022-32 (2012).
35. Evans, D.G. Neurofibromatosis type 2 (NF2): a clinical and molecular review. *Orphanet J Rare Dis* **4**, 16 (2009).
36. Inoue, K. *et al.* Molecular mechanism for distinct neurological phenotypes conveyed by allelic truncating mutations. *Nat Genet* **36**, 361-9 (2004).
37. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
38. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-99 (2012).
39. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-5 (2012).
40. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-41 (2012).
41. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-50 (2012).
42. Jacquemont, S. *et al.* A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. *Am J Hum Genet* **94**, 415-25 (2014).
43. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-15 (2014).
44. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-21 (2013).
45. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-58 (2013).
46. Kaiser, J. The hunt for missing genes. *Science* **344**, 687-9 (2014).
47. Alkuraya, F.S. Human knockout research: new horizons and opportunities. *Trends Genet* (2014).
48. Bhuvanagiri, M. *et al.* 5-azacytidine inhibits nonsense-mediated decay in a MYC-dependent fashion. *EMBO Mol Med* **6**, 1593-609 (2014).
49. Bhuvanagiri, M., Schlitter, A.M., Hentze, M.W. & Kulozik, A.E. NMD: RNA biology meets human genetic medicine. *Biochem J* **430**, 365-77 (2010).
50. Du, M. *et al.* PTC124 is an orally bioavailable compound that promotes suppression of the human CFTR-G542X nonsense allele in a CF mouse model. *Proc Natl Acad Sci U S A* **105**, 2064-9 (2008).
51. Hirawat, S. *et al.* Safety, tolerability, and pharmacokinetics of PTC124, a nonaminoglycoside nonsense mutation suppressor, following single- and multiple-dose administration to healthy male and female adult volunteers. *J Clin Pharmacol* **47**, 430-44 (2007).
52. Kerem, E. *et al.* Ataluren for the treatment of nonsense-mutation cystic fibrosis: a randomised, double-blind, placebo-controlled phase 3 trial. *Lancet Respir Med* **2**, 539-47 (2014).
53. Peltz, S.W., Morsy, M., Welch, E.M. & Jacobson, A. Ataluren as an agent for therapeutic nonsense suppression. *Annu Rev Med* **64**, 407-25 (2013).
54. Welch, E.M. *et al.* PTC124 targets genetic disorders caused by nonsense mutations. *Nature* **447**, 87-91 (2007).

Acknowledgments

We thank Patrick McGillivray and Daniel Spakowicz for comments on the manuscript. This work was supported by grants 5R01GM104371 (US National Institutes of Health/National Institute of General Medical Sciences) to S.B. and D.G.M., and U54HG006504 (Yale Center for Mendelian Genomics) to M.G.

Figures

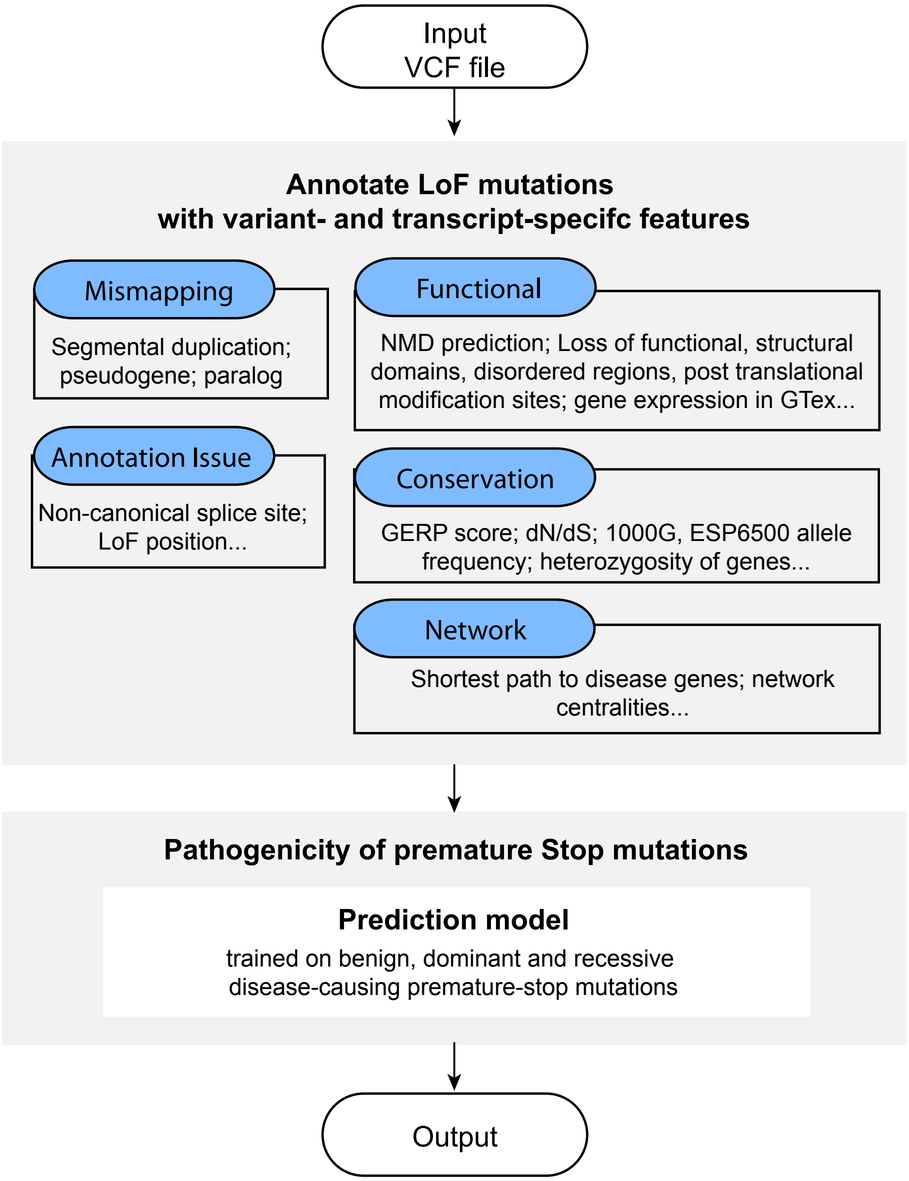
Input
VCF file
Annotate LoF mutations
with variant- and transcript-specifc features
Mismapping
Segmental duplication; pseudogene; paralog
Functional
NMD prediction; Loss of functional, structural domains, disordered regions, post translational modification sites; gene expression in GTex...
Annotation Issue
Non-canonical splice site; LoF position...
Conservation
GERP score; dN/dS; 1000G, ESP6500 allele frequency; heterozygosity of genes...
Network
Shortest path to disease genes; network centralities...
Pathogenicity of premature Stop mutations
Prediction model
trained on benign, dominant and recessive disease-causing premature-stop mutations
Output

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 features. ALoFT also flags potential mismapping and annotation errors. Using the annotation features, ALoFT predicts the pathogenicity (as either benign, recessive or dominant disease-causing) of premature stop mutations based on a model trained on known data. ALoFT can also take as input a 5-column tab-delimited file containing chromosome, position, variant ID, reference allele and alternate allele as its columns.

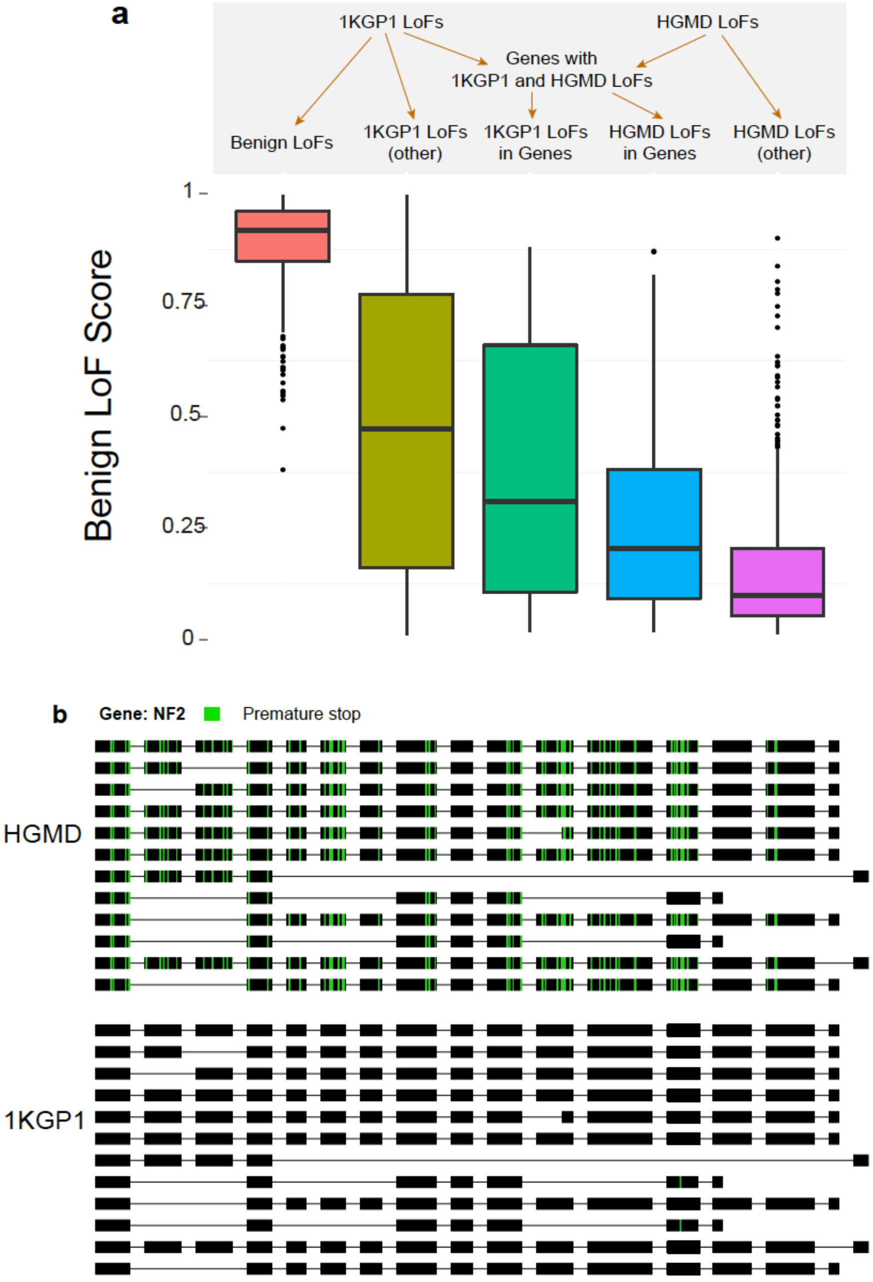

Figure 2 - ALoFT classification of 1000 Genomes and HGMD variants

a) Benign LoF score for premature stop variants in 1KGP1 and HGMD. For this plot, we randomly selected one variant per gene. The third (dark green) and fourth (blue) box plot pertains to premature stop variants in healthy 1KGP1 individuals occuring in disease-causing genes obtained from HGMD. “1KGP1 LoFs in other” include variants in 1KGP1 in genes not in HGMD i.e. non-disease genes. “HGMD LoFs other” include variants in only those disease genes where 1KGP1 LoF variants are not seen.

b) HGMD and 1KGP1 premature stop variants on the dominant disease-causing gene NF2. The benign 1KGP1 LoF variant truncates 2 isoforms, whereas HGMD LoF variants truncate 7 to 12 isoforms.

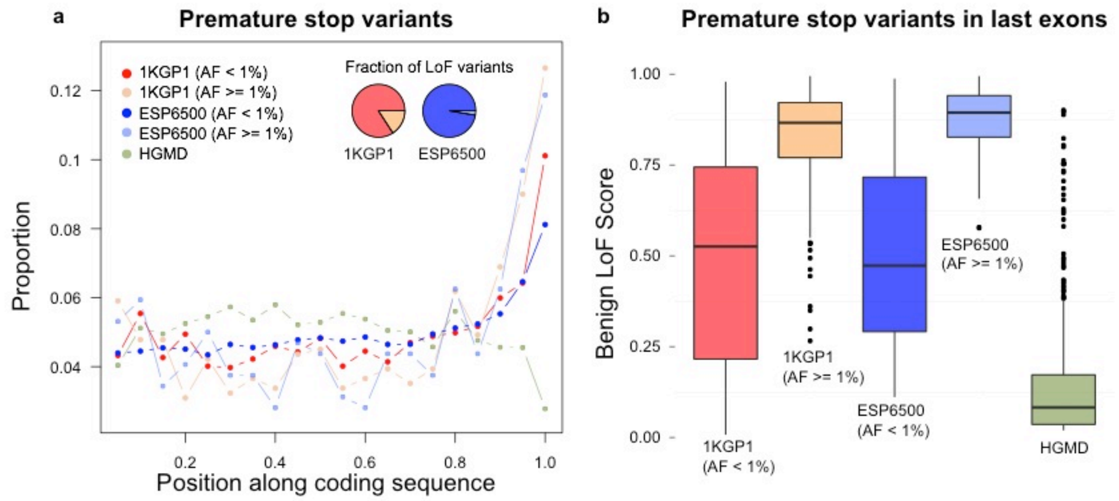

Figure 3: a) Position of premature stop variants in coding transcripts. Compared to HGMD variants, both common and rare 1KGP1 and ESP6500 variants are enriched in the last 5% of the coding sequence. “AF” stands for allele frequency. Variants at allele frequency less than 1% are considered to be rare variants. Variants with at least 1% allele frequency are considered as common.

b) Predicted benign LoF scores for premature stop variants in the last coding exon. Training variants are excluded in this plot.

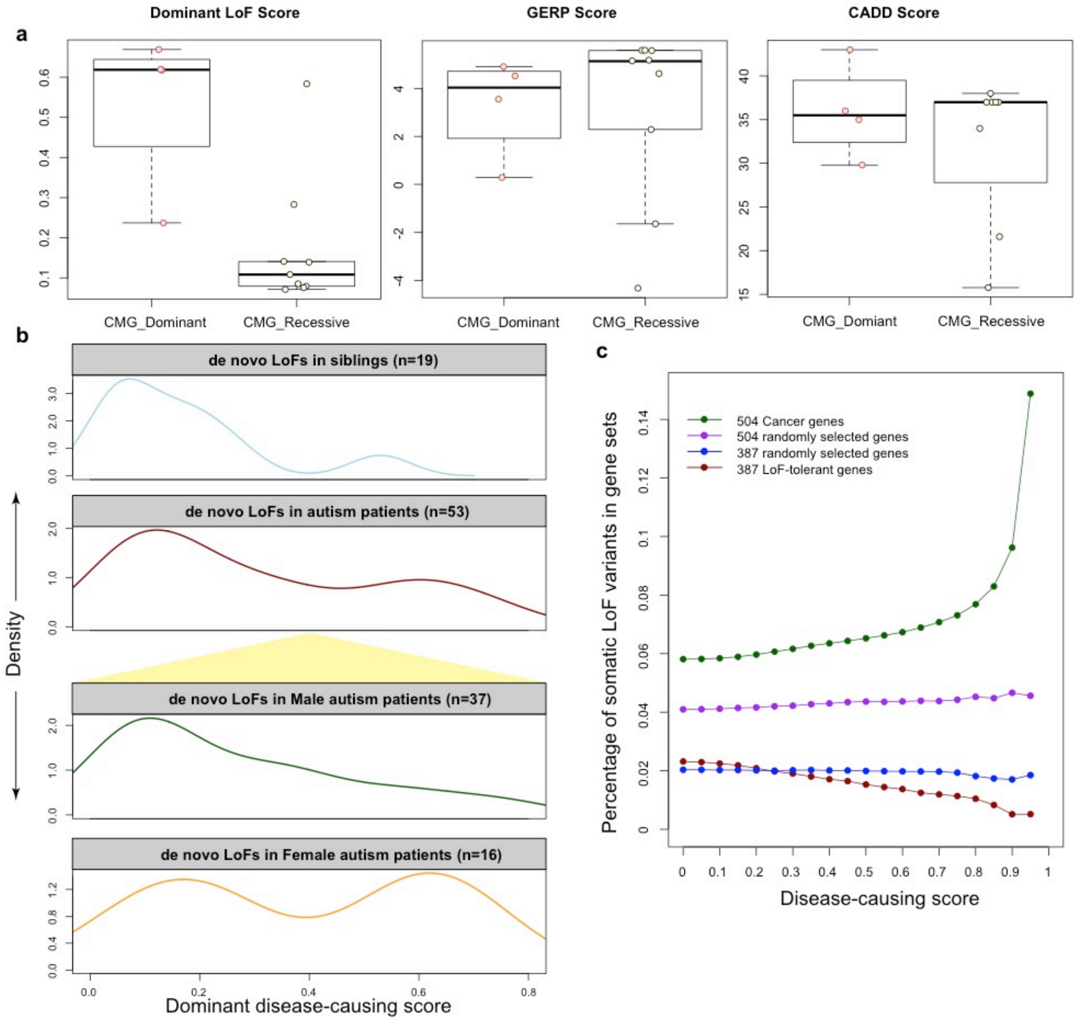

Figure 4 – ALoFT classification of premature stop variants from Mendelian disease, autism and cancer studies

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* premature stop 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) The fraction of mutations occurring in various gene categories (Y-axis) as a function of predicted diseasing-causing LoF score for cancer somatic premature stop variants (X-axis). Disease-causing score is calculated as (1- predicted benign LoF score).

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. 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 1KGP1 cohort.