Authors:
Suganthi Balasubramanian*, Yao Fu*, Mayur Pawashe, Mike Jin, Jeremy Lu, Daniel MacArthur, Mark Gerstein

Variants predicted to result in the loss of function (LoF) of human genes have attracted considerable recent 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 on average each individual carries three heterozygous 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 variants are more disruptive in patients than in unaffected siblings. Finally, we show that premature stop variants predicted to be pathogenic by ALoFT are enriched in 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 in identifying 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] and hence the active pursuit in 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. 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], 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 (Online Methods). ALoFT includes two network features previously shown to be important in disease prediction algorithms: proximity parameter that gives the number of disease genes that are 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 Supplementary Material and Methods section (Supplementary Table 1). Detailed documentation, input data files and source code linked to github can be found at aloft.gersteinlab.org.

To understand the impact of putative LoF variants on gene function we developed a prediction method to differentiate disease-causing and benign variants. While several algorithms to predict the effect of missense coding variants on protein function have been published, there is a paucity of methods that are applicable to nonsense variants[23,24]. Additionally, current prediction methods that infer the pathogenicity of variants do not take into account the zygosity of the variant[25,26]. 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. In addition to the features output by ALoFT, we also used gene-specific features for classification as shown in Figure 1a (details included in the Supplementary Material and Methods section).

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[22,27,28]. We built the ALoFT classifier to distinguish among the three classes using a random forest algorithm[29]. ALoFT provides class probability estimates for each mutation. We obtain good discrimination between the three classes. The average multiclass test AUC with 10-fold cross-validation is 0.962 (Check with Yao if we are going with the autosomes-only prediction model. Yao: did you exclude both chr X and chr Y). The precision for the three classes are as follows: Dominant=0.85, Recessive=0.85, 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 Figure 3a shows the important features in the classification. It is seen that presence/absence of an allele in ESP6500 and allele frequency seem to be

important features for the classification. However importance plot is not directly interpretable when the features used for deriving the prediction model are correlated. We removed the two features pertaining to ESP6500 allele frequency: presence or absence of allele in ESP6500 cohort and allele frequency of variant in ESP6500 and reevaluated the random forest model. The classifier still performs well with an average multiclass AUC =0.925. The precision for the models obtained by excluding allele frequency in ESP6500 are as follows: Dominant=0.80, Recessive=0.77, Benign=0.79 (Check with Yao if needs to be redone for autosomes only). As expected, allele frequency improves the classification accuracy. We evaluated the model by running the classifier using only two features: presence/absence of a variant in ESP6500 cohort and allele frequency. The average multiclass test AUC drops to 0.682. Using only the ESP6500 related features, the precision of the model for the three classes are: Dominant = 0.28, Recessive = 0.81, Benign= 0.96. Thus, allele frequency from ESP6500 cohort is a good feature for predicting benign mutations, but performs poorly in predicting dominant mutations. Therefore, only allele frequency in ESP6500 cohort as predictive features is inadequate for classification.

We applied ALoFT to 5,567 premature stop variants (5,164 heterozygous) from the 1KGP1 dataset. The predicted benign LoF score for premature stop variants in this population cohort have a wide range of values (Figure 2a). 2880 premature stop variants in the 1KGP1 dataset are predicted to be benign, 2505 variants are predicted to cause recessive disease, and a further 182 variants are predicted to lead to dominant disease (Supplementary Tables 2 and 6). On average, each individual is a carrier of about eight rare heterozygous premature stop alleles that are predicted to be disease-causing in the homozygous state (Supplementary Table 3) and less than one dominant variant. These are likely inflated estimates due to a number of confounding factors that include false positive variant calls, low penetrance of disease alleles, variable expressivity, genetic modifiers, compensatory mutations and limitations of the prediction model. Current estimates of the genetic burden of disease alleles in an individual vary widely. Estimates range from 1.1 recessive alleles per individual to 31 deleterious alleles.

To get a more accurate estimate of the number of premature Stop disease alleles in a healthy individual, we removed variants that are unlikely to cause LoF and variants that are highly likely to be false positive calls using the annotations provided by ALoFT. While ALoFT provides several flags that identifies likely false positive variant calls arising due to mapping and annotation errors, we conservatively only removed likely erroneous LoF annotations where the premature Stop variant is the ancestral allele. In addition, we removed potential false positive calls, identified as variants that are present at > 1% frequency in either the European or African American population of the 1KGP1 cohort but are absent in the ESP6500 cohort. We applied ALoFT to the filtered 1KGP1 variants and estimate that each individual has about 3 rare heterozygous premature stop alleles that are predicted to be disease-causing in the homozygous state, 0.25 dominant alleles and 57 alleles that are predicted to be benign (Supplementary Tables 3 and 4).

Next, we looked at premature stop variants in the 1KGP1 cohort in known disease-causing genes. Of the 476 variants, 146 variants are predicted to be benign, and 255 are recessive variants occurring in the healthy population only in the heterozygous state. Interestingly, in some cases (27.2% of variants), the variants in the presumed healthy 1KGP1 individuals and the disease-causing variants are in the same gene, but on different isoforms (Figure 2b). 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 (Figure 2b). Considering that mutations in NF2 are well-characterized dominant mutations[28], we should not observe any LoF variant in NF2 in the presumed healthy individuals. Truncating mutations in NF2 are known to cause the most severe disease, while missense mutations cause milder phenotypes[28]. Therefore, this suggests that isoform-specific premature stop variants are responsible for disease and are not seen in the individuals presumed to be healthy in the 1KGP1 cohort.

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 dominant negative mutations in the last exon are also known[30]. 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. Given the larger sample size in the ESP6500 cohort (6503 individuals) relative to the 1KGP1(1092 individuals), it is not surprising that we find rarer alleles in ESP6500 cohort and thus the average allele frequency of premature Stop variants is lower in the ESP6500 cohort. Nonetheless, a large number of premature stop variants are seen at the end of the coding genes in both the 1KGP1 and ESP6500 datasets (Figure 2c). Premature stop variants in the last coding exon in the 1KGP1 and ESP6500 cohort are likely to be benign, whereas HGMD mutations in the last coding exon tend to be disease-causing (Figure 2d, median benign LoF scores for 1KGP1, ESP6500 and HGMD are: 0.60, 0.50, and 0.05 respectively).

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 3a). Our method showed that heterozygous disease-causing variants have significantly higher dominant disease-causing scores than the homozygous disease-causing variants (p-value: 0.017; Wilcoxon rank-sum test). We used two other measures, GERP score which is a measure of evolutionary conservation and CADD score that gives a measure of pathogenicity, to classify recessive versus dominant LoF variants[31]. Both CADD and GERP scores are not able to discriminate between recessive and dominant disease-causing mutations (Fig 3a).

De novo LoF SNPs have been implicated in autism based on analysis of sporadic or simplex families (families with no prior history of autism)[32-35]. We applied our method to *de novo* LoF mutations discovered in these autism studies. Our method shows that the proportion of dominant disease-causing *de novo* LoF events is significantly higher in autism patients versus siblings (Fig 3b; p-value: 0.005; Wilcoxon rank-sum test). Previous studies suggest that there is a higher mutational burden in female patients[36]. 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: 0.038). A recent study based on exome sequencing of 3871 autism cases delineated 33 risk genes at FDR < 0.1[37]. *De novo* LoF mutations culled from the above mentioned published research 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: 0.003). Supplementary Table 7 includes the ALoFT predictions for *de novo* LOF variants.

Lastly, we also examined somatic stop-causing mutations in several cancers. 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[38]. Therefore, we expect to see a higher proportion of deleterious somatic LoF variants in driver genes than the rest of the genes. We applied our prediction method to infer the effect of somatic premature stop variants from a compilation of ~6,000 cancer exome sequencing studies[39]. As shown in Figure 3c, 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 and depleted in LoF-tolerant genes (genes that contain at least one homozygous LoF variant in the 1KGP1 population).

In summary, we describe a tool for predicting the impact of nonsense SNPs in the context of a diploid model, i.e. discriminating whether nonsense SNPs are more 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 information will help us to identify protective LoF variants which are valuable drug targets[40,41]. 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[42-48]. 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.

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. Castellana, S. & Mazza, T. Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. *Brief Bioinform* **14**, 448-59 (2013).
24. Karchin, R. Next generation tools for the annotation of human SNPs. *Brief Bioinform* **10**, 35-52 (2009).
25. Hu, J. & Ng, P.C. SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. *PLoS One* **8**, e77940 (2013).

26. Rausell, A. *et al.* Analysis of stop-gain and frameshift variants in human innate immunity genes. *PLoS Comput Biol* **10**, e1003757 (2014).
27. 1000 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
28. Evans, D.G. Neurofibromatosis type 2 (NF2): a clinical and molecular review. *Orphanet J Rare Dis* **4**, 16 (2009).
29. Breiman, L. Random Forests. *Machine Learning* **45**, 5-32 (2001).
30. Inoue, K. *et al.* Molecular mechanism for distinct neurological phenotypes conveyed by allelic truncating mutations. *Nat Genet* **36**, 361-9 (2004).
31. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
32. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-99 (2012).
33. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-5 (2012).
34. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-41 (2012).
35. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-50 (2012).
36. 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).
37. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-15 (2014).
38. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-58 (2013).
39. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-21 (2013).
40. Kaiser, J. The hunt for missing genes. *Science* **344**, 687-9 (2014).
41. Alkuraya, F.S. Human knockout research: new horizons and opportunities. *Trends Genet* (2014).
42. Bhuvanagiri, M. *et al.* 5-azacytidine inhibits nonsense-mediated decay in a MYC-dependent fashion. *EMBO Mol Med* **6**, 1593-609 (2014).
43. Bhuvanagiri, M., Schlitter, A.M., Hentze, M.W. & Kulozik, A.E. NMD: RNA biology meets human genetic medicine. *Biochem J* **430**, 365-77 (2010).
44. 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).
45. 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).
46. 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).
47. 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).

48. Welch, E.M. *et al.* PTC124 targets genetic disorders caused by nonsense mutations. *Nature* **447**, 87-91 (2007).