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

Suganthi Balasubramanian[1,2,6$,*,#], Yao Fu[1,7$,*], Mayur Pawashe[2], Mike Jin[2], Jeremy Liu[2], Patrick McGillivray[2], Konrad J. Karczewski[3,4], 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
[6]Regeneron Genetics Center, Tarrytown, NY 10591, USA
[7]Bina Technologies, Part of Roche Sequencing, Belmont 94002, CA, USA

[#] Corresponding authors: suganthi.bala@regeneron.com, pi@gersteinlab.org
* Contributed equally to this work
[$] Current address

Variants predicted to result in the loss of function (LoF) of human genes have recently attracted considerable interest both because of their clinical impact as well as their surprising prevalence in seemingly healthy humans. To better understand the impact of putative LoF variants (pLoF), we developed ALoFT (Annotation of Loss-of-Function Transcripts), to annotate and predict the disease-causing potential of LoF variants. Using data from Mendelian disease gene discovery projects, we show that ALoFT is able to distinguish between pLoF variants that are deleterious as heterozygotes and variants that cause disease only when they are homozygous. Investigation of variants discovered in healthy human populations suggests that each individual carries at least two heterozygous premature stop alleles that could potentially lead to disease if present as homozygotes. When applied to *de novo* pLoF variants in autism-affected families, ALoFT predicts that the variants are more deleterious in patients than in unaffected siblings. Finally, analysis of somatic variants in > 6,500 cancer exomes shows that pLoF variants predicted to be deleterious by ALoFT are enriched in known cancer driver genes.

One of the most notable findings from personal genomics studies is that all individuals harbor loss-of-function variants in some of their genes[1]. A systematic study of LoF variants from the 1000 Genomes revealed that there are over 100 putative LoF (pLoF) variants in each individual[2-4]. Recently, a larger study aimed at elucidating rare LoF events in 2,636 Icelanders generated a catalog of 1,171 genes that contain either homozygous or compound heterozygous LoF variants with a minor allele frequency less than 2%[5]. 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 pLoF 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[6,7] which prompted the active pursuit of the inhibition of *PCSK9* as a potential therapeutic for hypercholesterolemia[8] and led to the development of two drugs which have been recently approved by the FDA. 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, two splice variants in the Finnish population in *LPA* that protect against coronary heart disease, and two LoF-producing splice variants and a nonsense mutation in *HAL* associated with increased blood histidine levels and reduced risk of coronary heart disease[9-13].

About 12% of known disease-causing mutations in the Human Gene Mutation Database (HGMD) are due to nonsense mutations[14]. pLoF variants are also prioritized in cancer studies where various filtration schemes are used to narrow down causal mutations [15,16]. 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[17]. However, a recent large-scale expression analysis demonstrated that 68% of predicted NMD events due to premature stop variants are unsupported by RNA-Seq analyses[18]. 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[19]. Moreover, premature stop codons in the last exon are generally 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 (nonsense) 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) functional domain annotations; (2) evolutionary conservation; and (3) biological networks. For comprehensive functional annotation, we integrated several annotation resources such as PFAM and SMART functional domains[20,21], signal peptide and transmembrane annotations, post-translational modification sites, NMD prediction[22,23], and structure-based features such as SCOP domains and disordered residues. For evolutionary conservation, ALoFT outputs variant position-specific GERP scores, which is a measure of evolutionary

conservation[24] and 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. In addition, we evaluate if the region removed due to the truncation of the coding sequence is evolutionarily conserved based on constrained elements[25]. ALoFT includes 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[2,26]. 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 Supplementary Table 1. Documentation and github link to source code can be found at aloft.gersteinlab.org.

To understand the impact of pLoF 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[27-30]. Additionally, current prediction methods that infer the pathogenicity of variants do not take into account the zygosity of the variant[31,32]. The majority of pLoF variants in healthy population cohorts are heterozygous. It is likely that a subset of these variants will cause disease as homozygotes. Therefore, we developed a prediction model to classify premature stop causing variants into three classes: those that are benign, that lead to recessive disease (disease-causing only when homozygous) and that lead to dominant disease (disease-causing as heterozygotes) using the annotations output by ALoFT as predictive features (Fig. 1, Supplementary Information).

To build the ALoFT classifier, we used three classes of premature stop variants as training data: benign variants, dominant disease-causing and recessive disease-causing variants. The benign set includes homozygous premature stop variants discovered in a cohort of 1,092 healthy people, Phase1 1000 Genomes data (1KG). 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 represent the two disease classes[3,26]. In addition to loss-of-function effects, truncating mutations can also lead to gain of function. However, gain of function mutations are difficult to model systematically as the effect of variant can only be understood in the context of the biology of the gene and can vary widely for different genes and gene classes. In order to minimize errors that might arise due to inadequate modeling of gain-of-function effects and focus only on LoF, we only use predicted haploinsufficient genes as the training data for dominant model. We built the ALoFT classifier to distinguish among the three classes using a random forest algorithm[33]. For each mutation, ALoFT provides three class probability estimates, and we obtain good discrimination between each class. The average multiclass test AUC (area under the curve) with 10-fold cross-validation is 0.97. The precision for the three classes are as follows: Dominant=0.86, Recessive=0.86, Benign=0.96. The classifier is robust to the choice of training data sets and performs well with different training data sets (Supplementary Table 4, Supplementary Fig. 2). The prediction output provides three scores for each pLoF variant that correspond to the probability of the pLoF being benign, dominant or recessive disease-causing allele. In addition, ALoFT also provides the predicted pathogenicity. The pathogenic effect of pLoF variant is assigned to the class that corresponds to the maximum score. Though trained with premature stop SNVs, our

method is also applicable to frameshift indels. 99.4% of HGMD disease-causing ('DM') frameshift indels are predicted as pathogenic based on the maximum ALoFT score.

We analyzed the importance of the various features to the classification (Supplementary Fig. 3). The global allele frequency of variants in the Exome Aggregation Consortium, a dataset comprised of sequence variations obtained from an analysis of 60,706 unrelated individuals of diverse ethnicities (ExAC[34], http://exac.broadinstitute.org), appears to be the most important feature for the classification. When we removed this feature and other features related to allele frequency (i.e. both ExAC and ESP) and retrained the random forest model, the classifier still performs well with an average multiclass test AUC of 0.93. (The precision for the three classes are as follows: Dominant=0.84, Recessive=0.80, and Benign=0.75). We also systematically evaluated the classifier using models trained on various specific sets of features (Supplementary Table 5). Overall, we find that classifier is not driven by any single feature and integrating many features improves prediction accuracy.

Validation of classifier

We applied ALoFT to elucidate the pathogenicity of pLoF variants in various disease scenarios. Using case studies, we show that ALoFT provides robust predictions for the effect of pLoFs.

Case study 1: Application of ALoFT to understand pLOFs in Mendelian disease

We evaluated ALoFT by predicting the effect of known disease-causing premature stop mutations from ClinVar[35] (details in Supplementary Information) and predicted the mode of inheritance and pathogenicity of all of the truncating variants (Fig. 2a). ALoFT is clearly able to distinguish between pLoFs that possibly lead to disease in a heterozygous state versus those that do so only in a homozygous state. Our method shows that heterozygous disease-causing variants have significantly higher dominant disease-causing scores than the homozygous disease-causing variants (p-value: 1.3e-13; 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 pLoF variants[36]. Both CADD (p-value: 0.13) and GERP (p-value: 0.49) scores are not able to discriminate between recessive and dominant disease-causing mutations (Fig. 2a). We also tested our method on a smaller dataset from the Center For Mendelian Genomics studies[37] and was able to correctly recapitulate the pathogenic effect of pLoF variants and their inheritance pattern (Supplementary Fig. 4).

Case study 2: Application of ALoFT to understand *de novo* pLoFs implicated in autism

De novo pLoF 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* pLoF 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. 2b; p-value: 8.4e-4; Wilcoxon rank-sum test).

Autism spectrum disorder is known to be four times more prevalent in males than females suggesting a protective effect in females. Previous studies show that there is a higher mutational burden for non-synonymous mutations in females ascertained for autism spectrum disorder[42]. Therefore, we investigated the differences in the impact of *de novo* pLoF variants in male versus female autism patients. We also observe a similar pattern for pLoF mutations as has been found for missense variants – female probands have a higher proportion of predicted deleterious *de novo* pLoF variants than male probands (Fig. 2b; p-value: 0.03). 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* pLoF mutations in the autism patients in the 33 risk genes have higher dominant disease causing score than the *de novo* pLoF variants in other genes (Supplementary Fig. 5; p-value: 5e-3). Supplementary Table 6 includes the ALoFT predictions for *de novo* pLoF variants. Thus, ALoFT predictions corroborate the role of *de novo* pLoF variants in autism as shown by others using entirely different approaches.

Case Study 3: Identification of pathogenic somatic pLoF variants in cancer

We applied our prediction method to infer the effect of somatic premature stop variants (somatic pLoFs) from a compilation of 6,535 cancer exomes[44]. As shown in Figure 2c, somatic pLoFs are enriched in known cancer driver genes compared to randomly sampled genes of matched lengths. Moreover, deleterious somatic LoFs are strongly enriched in driver genes and depleted in LoF-tolerant genes (genes that contain at least one homozygous LoF variant in the 1KG population). Due to aneuploidy and clonal heterogeneity of cancer cells, we define an overall measure of deleteriousness as (1-Benign ALoFT acore) as shown in the X-axis of Figure 2c. 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]. From 505 genes with pLOF mutations identified by ALoFT, 317 met the 20/20 rule (62.7%), while 188 genes do not. Considering the fact that loss-of-function of tumor suppressor genes can lead to oncogenesis, ALoFT can be used to identify cancer-related pLoFs, especially those in tumor suppressors.

Case study 4: Distinguishing between benign and pathogenic pLoFs

Finally, we applied ALoFT to predict the effect of premature stop variants in the final exons of protein-coding genes. It is often assumed that premature stop variants in the last coding exon are likely to be benign because they could 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[46]. 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 applied ALoFT to understand the effect of pLoF variants in ESP6500, ExAC and HGMD datasets. A higher proportion of rare variants are observed in ESP6500 and ExAC cohort due to its larger sample size and higher sequencing depth (Fig. 3a). A large number of both common and rare premature stop variants are seen at the end of the coding genes in the 1KG, ESP6500 and ExAC 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 1KG, ESP6500 and ExAC cohort are likely to be benign, whereas HGMD mutations in the last coding exon 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.

Application to personal genomes: Estimating the number of pathogenic pLoFs in a healthy genome or understanding pLoFs in an individual genome

The above case studies clearly illustrate the validity of the ALoFT score in elucidating the effect of pLoF variants. In order to estimate the number of pLoF disease alleles in a healthy individual, we applied ALoFT to premature stop variants from the 1KG and ExAC datasets (Supplementary Information). The predicted benign score for pLoFs in these cohorts has a wide range of values (Fig. 4, Supplementary Tables 8,9). Furthermore, due to differences in sequencing coverage and variant calling approaches, the number of potential disease pLoFs per individual varies among datasets. In general, the number increases with higher coverage and larger cohorts where joint variant calling methods result in improved sensitivity in the identification of rare variants. To conservatively estimate a lower bound for per individual statistics (Supplementary Information), we applied a stringent filtering strategy to restrict to high confidence pLoFs. On average, each individual is a carrier of at least two rare heterozygous premature stop alleles that are predicted to be disease-causing in the homozygous state (Supplementary Table 9). Current estimates of the genetic burden of disease alleles (all types of variations, including LoFs) in an individual vary widely, ranging from 1.1 recessive alleles per individual to 31 deleterious alleles[47-51]. In connection with this, it should be noted that the referenced studies are based on diverse methods of identifying variants ranging from targeted panel-based candidate gene studies to whole genome sequencing and disease databases include incorrect disease annotations and common variants and about 27% of variants were excluded by Bell *et al*. in their estimate of carrier burden for severe recessive diseases[47]. ALoFT classifies 3.7% of HGMD mutations as tolerant mutations. Some notable examples of HGMD LoF variants predicted to be tolerant occur in genes such as *FLG, C4orf26 and APOA2*. Filaggrin LoF mutations are linked to susceptibility to atopic dermatitis, a skin condition leading to eczema (PMID: 27659773). Mutations in C4orf26 lead to *Amelogenesis Imperfecta*, a disorder of tooth development. While these mutations are pathogenic, they are not lethal and are also known to be genetically heterogeneous (PMID: 20878018).The estimation of the number of deleterious pLoF alleles can be affected by a number of confounding factors that include incomplete penetrance of disease alleles, variable expressivity, compensatory mutations, marginal variant calls and imperfect training datasets (Supplementary methods).

Next, we looked at premature stop variants in the 1KG cohort in known disease-causing genes. We find that variants in 1KG are more likely to be benign compared to known disease-causing mutations in the same genes (Fig. 4; green vs. blue boxes, p-value: 6.9e-9). Our results provide a possible rationale for this observation. Firstly, variants predicted to be benign in 1KG often affect isoforms that are different from the isoforms containing the disease-causing HGMD variant. This suggests that LoFs in healthy individuals may affect minor isoforms (Supplementary Fig. 7). About 12.4% of premature stop variants in the presumed healthy 1KG individuals in known disease genes and the disease-causing variants in the same genes are on different isoforms. Secondly, some variants predicted to be benign in 1KG occur in the last exon or later in the protein-coding transcript relative to the disease-causing variant in the same

transcript. The effect of such variants is possibly the production of truncated proteins that are sufficiently functional. Lastly, a majority of 1KG variants seen in the disease genes are predicted to be disease-causing only if they are homozygous. However, they occur as rare heterozygous variants in the 1KG cohort.

A study on British Pakistanis with related parents identified 781 genes containing rare LoF homozygous variants (PMID: 26940866). They found homozygous LoF variants in recessive Mendelian disease genes in 42 people, however 33 of them did not have the disease phenotype. We applied ALoFT to classify these homozygous LoF variants and ALoFT indeed predicts that 19 of them would cause disease (Supplementary table xx). However, the lack of a discernible phenotype could be due to incomplete penetrance of the mutations or due to modifier effects. The penetrance of some disease mutations are known to be age and sex-dependent (PMID: 19785764). It is well established that there is widespread occurrence of disease variants with reduced penetrance in the general population (PMID: 23820649). While studies in consanguineous populations have been used to identify recessive disease genes (PMID: 25558065, 27435318), absence of disease provides an opportunity to look for modifiers in their genetic background.

In summary, we describe a tool for predicting the impact of pLoF variants in the context of a diploid model, i.e. discriminating whether pLoF variants are likely to lead to recessive or dominant disease. Better identification and characterization of pLoF variants has both diagnostic and therapeutic implications. ALoFT allows for the identification and prioritization of high impact putative disease-causing pLoF variants in individual genomes. Integrating benign LoF variants with phenotypic information will help us to identify protective variants which are valuable drug targets[52,53]. Gene functions important for species propagation might actually be deleterious as one ages; thus, LoF variants in such genes provide an intriguing avenue to discover targets for aging-related diseases[54]. 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[55-61]. 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.

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. 1000 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
4. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
5. Sulem, P. *et al.* Identification of a large set of rare complete human knockouts. *Nat Genet* (2015).
6. 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).
7. 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).
8. Stein, E.A. *et al.* Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *N Engl J Med* **366**, 1108-18 (2012).
9. Flannick, J. *et al.* Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat Genet* **46**, 357-63 (2014).
10. Lim, E.T. *et al.* Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet* **10**, e1004494 (2014).
11. Tachmazidou, I. *et al.* A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nat Commun* **4**, 2872 (2013).
12. 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).
13. Yu, B. *et al.* Association of Rare Loss-Of-Function Alleles in HAL, Serum Histidine Levels and Incident Coronary Heart Disease. *Circ Cardiovasc Genet* (2015).
14. 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).**
15. **Snape, K. *et al.* Mutations in CEP57 cause mosaic variegated aneuploidy syndrome. *Nat Genet* 43, 527-9 (2011).**
16. **Snape, K. *et al.*** Predisposition gene identification in common cancers by exome sequencing: insights from familial breast cancer. *Breast Cancer Res Treat* **134**, 429-33 (2012).
17. Isken, O. & Maquat, L.E. Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes Dev* **21**, 1833-56 (2007).
18. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-11 (2013).

19. 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).
20. Letunic, I., Doerks, T. & Bork, P. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res* (2014).
21. Finn, R.D. *et al.* Pfam: the protein families database. *Nucleic Acids Res* **42**, D222-30 (2014).
22. 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).
23. 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).
24. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-13 (2005).
25. 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).
26. Huang, N., Lee, I., Marcotte, E.M. & Hurles, M.E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* **6**, e1001154 (2010).
27. 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).
28. 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).
29. Karchin, R. Next generation tools for the annotation of human SNPs. *Brief Bioinform* **10**, 35-52 (2009).
30. 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).
31. Hu, J. & Ng, P.C. Predicting the effects of frameshifting indels. *Genome Biol* **13**, R9 (2012).
32. Rausell, A. *et al.* Analysis of stop-gain and frameshift variants in human innate immunity genes. *PLoS Comput Biol* **10**, e1003757 (2014).
33. Breiman, L. Random Forests. *Machine Learning* **45**, 5-32 (2001).
34. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv* (2015).
35. Landrum, M.J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980-5 (2014).

36. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
37. Chong, J.X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet* **97**, 199-215 (2015).
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. Inoue, K. *et al.* Molecular mechanism for distinct neurological phenotypes conveyed by allelic truncating mutations. *Nat Genet* **36**, 361-9 (2004).
47. Bell, C.J. *et al.* Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med* **3**, 65ra4 (2011).
48. 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).
49. 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).
50. 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).
51. Tabor, H.K. *et al.* Pathogenic variants for Mendelian and complex traits in exomes of 6,517 European and African Americans: implications for the return of incidental results. *Am J Hum Genet* **95**, 183-93 (2014).

52. Kaiser, J. The hunt for missing genes. *Science* **344**, 687-9 (2014).
53. Alkuraya, F.S. Human knockout research: new horizons and opportunities. *Trends Genet* (2014).
54. Yizhak, K., Gabay, O., Cohen, H. & Ruppin, E. Model-based identification of drug targets that revert disrupted metabolism and its application to ageing. *Nat Commun* **4**, 2632 (2013).
55. Bhuvanagiri, M. *et al.* 5-azacytidine inhibits nonsense-mediated decay in a MYC-dependent fashion. *EMBO Mol Med* **6**, 1593-609 (2014).
56. Bhuvanagiri, M., Schlitter, A.M., Hentze, M.W. & Kulozik, A.E. NMD: RNA biology meets human genetic medicine. *Biochem J* **430**, 365-77 (2010).
57. 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).**
58. 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).
59. 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).
60. 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).
61. 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

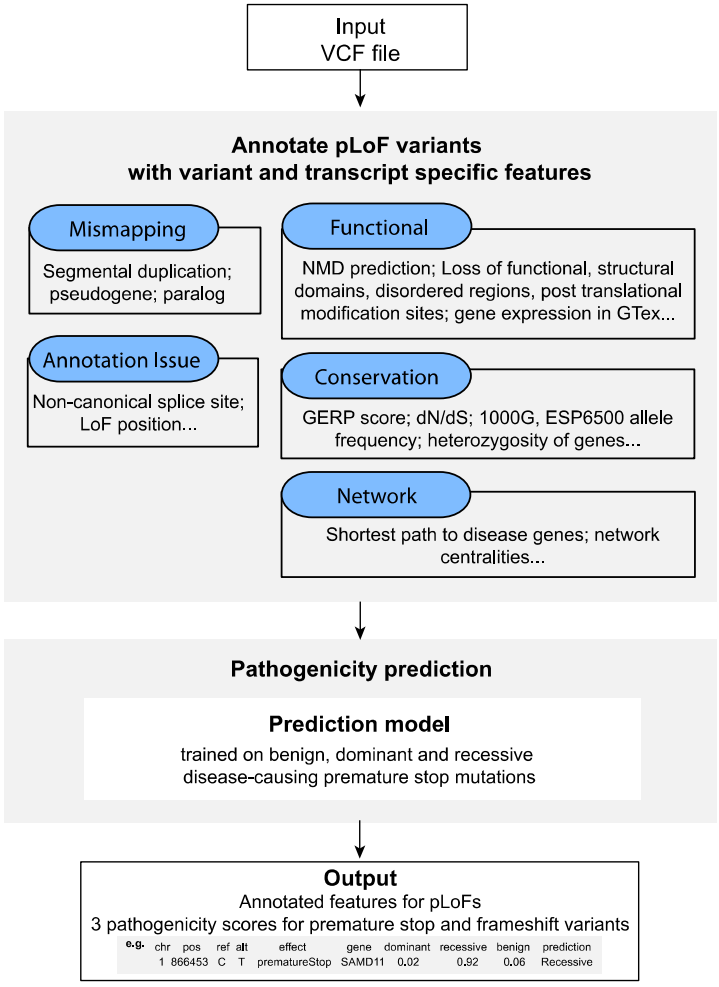

Figure 1 - Schematic workflow.
ALoFT uses a VCF file as input and annotates premature stop, frameshift-causing indel and canonical splice-site mutations with functional, conservation, network features. ALoFT also flags potential mismapping and annotation errors. Using the annotation features, ALoFT predicts the

pathogenicity (as either benign, recessive or dominant disease-causing) of premature stop and frameshift 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.

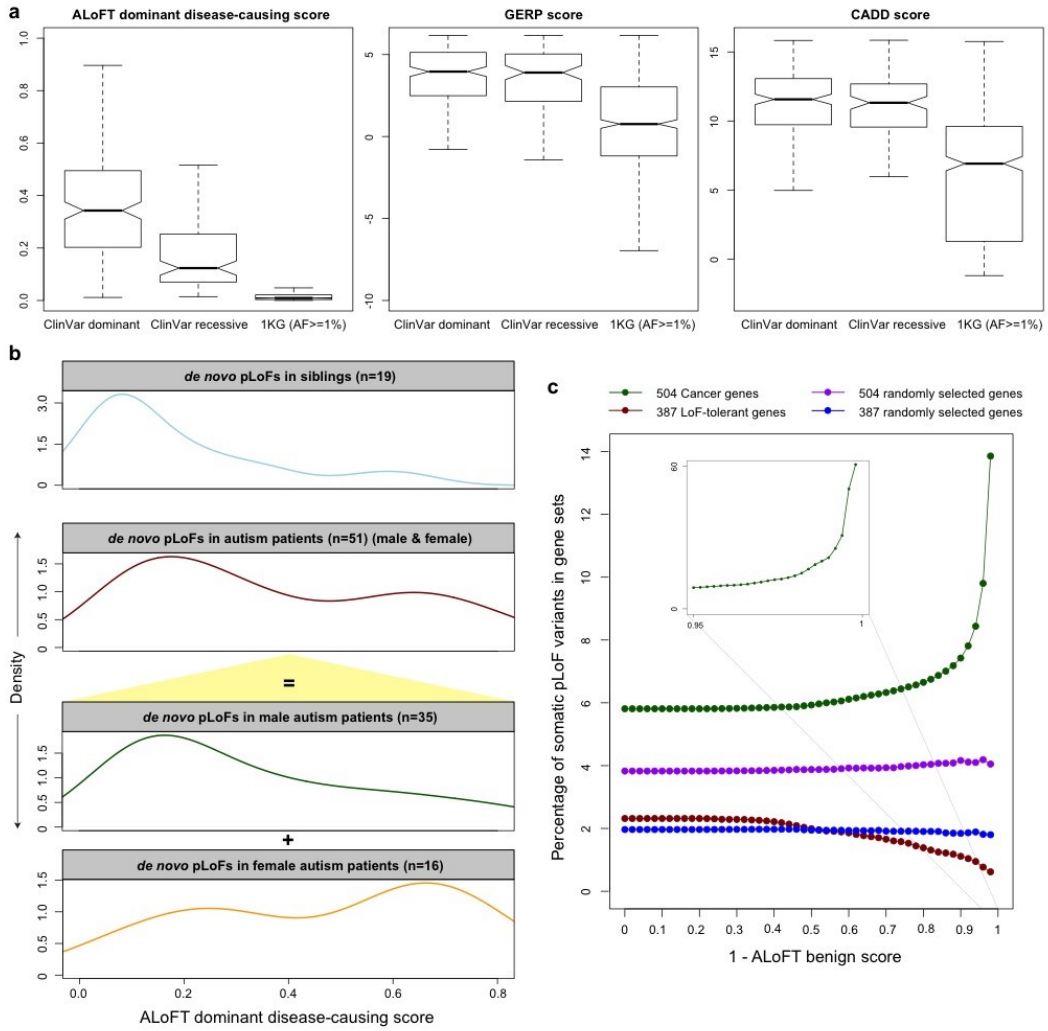

Figure 2 – ALoFT classification of premature stop variants from Mendelian disease, autism and cancer studies

a) ALoFT dominant disease-causing score, GERP and CADD score for ClinVar and 1KG common (AF>=1%) variants. All training variants are excluded. Average tolerance scores are 0.097 and 0.115 respectively for ClinVar dominant and recessive datasets.

b) The top two panels show the dominant disease-causing scores of *de novo* premature stop mutations in autism patients and siblings; mutations in patients are further separated by gender, as shown in the bottom two panels.
c) The fraction of mutations occurring in various gene categories (Y-axis) as a function of predicted diseasing-causing score for cancer somatic premature stop variants (X-axis). Disease-causing score is calculated as (1- predicted benign 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 1KG cohort.

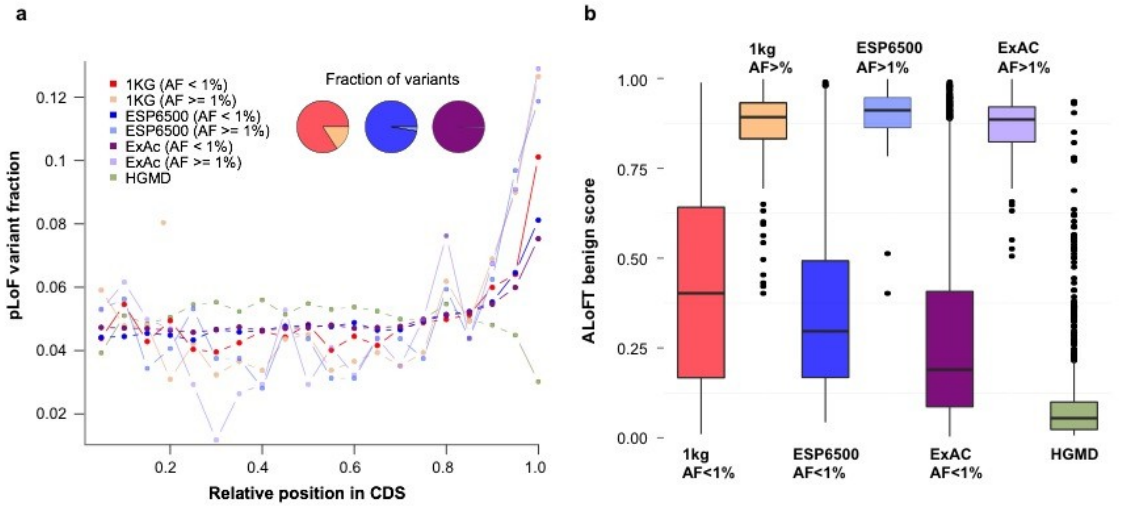

Figure 3 – pLoFs in last exons

a) Position of premature stop variants in coding transcripts. Compared to HGMD variants, both common and rare 1KG, ESP6500 and ExAC 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 scores for premature stop variants in the last coding exons.
Training variants are excluded in this plot.

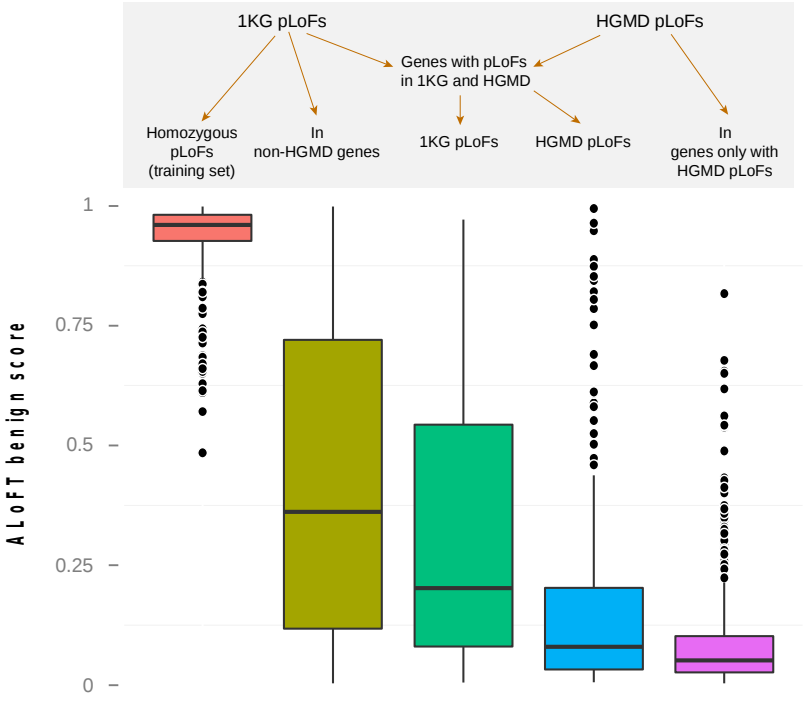

Figure 4 - ALoFT classification of 1000 Genomes and HGMD variants

Benign score for premature stop variants in 1KG and HGMD. For this plot, we randomly selected one variant per gene. “Benign pLoFs” set includes homozygous premature stop variants discovered in 1KG. The third (dark green) box plot pertains to premature stop variants in healthy 1KG individuals occuring in disease-causing genes obtained from HGMD. The fourth (blue) box plot pertains to pLoF variants in the subset of HGMD genes where 1KG pLoFs are also seen. “1KG pLoFs in non-HGMD genes” include 1KG variants not in HGMD genes, i.e. non-disease genes. “In genes only with HGMD pLoFs” include HGMD variants in only those disease genes where 1KG pLoFs are not seen.