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

Suganthi Balasubramanian1,2\*,#, Yao Fu1\*, Mayur Pawashe2, Mike Jin2, Jeremy Liu2, Daniel G. MacArthur3,4, Mark Gerstein1,2,5,#

1Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520 USA, 2Molecular Biophysics and Biochemistry Department, Yale University, New Haven 06520, CT, USA, 3Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA, 4Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA, 5Department of Computer Science, Yale University, New Haven 06520, CT, USA

# Corresponding authors

[suganthibala@gmail.com](mailto:suganthibala@gmail.com)

[pi@gersteinlab.org](mailto:pi@gersteinlab.org)

\* Contributed equally to this work

**Variants predicted to result in the loss of function 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. Our method is able to distinguish between dominant and recessive pLoF variants discovered by Mendelian studies. Investigation of variants discovered in over 1,000 human genome sequences suggests that each individual carries about two heterozygous premature stop alleles that could potentially lead to disease if present in the homozygous state. When applied to *de novo* pLoF variants in autism-affected families, ALoFT predicts that the variants are more disruptive in autism patients than in unaffected siblings. Finally, analysis of 6,535 cancer exomes shows that somatic 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 genes1. A systematic study of pLoF variants from 180 individuals revealed that there are over 100 putative LoF variants in each individual2. 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%3. 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 levels4,5 which has prompted the active pursuit of the inhibition of PCSK9 as a potential therapeutic for hypercholesterolemia6-8. 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 splice variants and a nonsense mutation in HAL associated with increased blood histidine levels and reduced risk of coronary heart disease9-13.

About 12% of known disease-causing mutations in the Human Gene Mutation Database (HGMD) are due to nonsense mutations14. 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 mechanism17. However, a recent large-scale expression analysis demonstrated that 68% of predicted NMD events due to premature stop variants are unsupported by RNASeq analyses18. A study aimed at understanding disease mutations using a 3D structure-based interaction network suggests that truncating mutations can give rise to functional protein products19. 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 domains20,21, signal peptide and transmembrane annotations, post-translational modification sites, NMD prediction22,23, 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 conservation24 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 GERP constrained elements25. 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 gene2,26. 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 artifacts2. 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.

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 variants27-30. Additionally, current prediction methods that infer the pathogenicity of variants do not take into account the zygosity of the variant31,32. 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 causing 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: benign homozygous premature stop variants, dominant heterozygous and recessive homozygous disease premature stop mutations. The benign set includes homozygous premature stop variants discovered in a cohort of 1092 healthy people, Phase1 1000 Genomes data (1KGP1). 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 classes26,33. In addition to LoF 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 is very context dependent depending on the biology o 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 algorithm34. 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 0.97. 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 3, 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 LoF score.

We analyzed the feature importance to understand the contribution of different features to the classification (Supplementary Fig. 3). The global allele frequency of variants in the Exome Aggregation Consortium (ExAC), a dataset comprised of sequence variations obtained from analysis of unrelated 60,706 individuals of diverse ethnicities (<http://exac.broadinstitute.org>) appears to be the most important feature for the classification. We removed features related to allele frequency and retrained the random forest model. The classifier still performs well with an average multiclass test AUC =0.93 and the precision for the three classes are as follows: Dominant=0.84, Recessive=0.80, Benign=0.75. We also systematically evaluated the classifier using models trained on specific sets of features. Thus, even though allele frequency is an important feature, integrating all the features improves prediction accuracy of the classifier and the overall accuracy is not driven by any single feature (Supplementary Table 4). It should also be noted that the importance plot is not directly interpretable because some of the prediction variables are correlated.

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 (is robust the appropriate word to use) 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 nonsense mutations from ClinVar35 (details in Methods) and predicted the mode of inheritance and pathogenicity of all of the truncating variants (Fig. 4a). ALoFT is clearly able to distinguish between pLoFs that 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 variants36. 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. 4a). We also tested our method on a smaller dataset from the Center For Mendelian Genomics studies (<http://data.mendelian.org/CMG/>) and was able to correctly recapitulate the pathogenic effect of pLoF variants and their inheritance pattern (Supplementary figure 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)37-40. We applied our method to *de novo* LoF mutations discovered in these studies. Our method also shows that the proportion of dominant disease-causing *de novo* LoF events is significantly higher in autism patients versus siblings (Fig. 4b; 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 in females ascertained for autism spectrum disorder41. Therefore, we investigated the differences in *de novo* LoF variants in male versus female autism patients. We also observe a similar pattern for pLoF mutations – female probands have a higher proportion of predicted deleterious *de novo* pLoF variants than male probands (Fig. 4b; p-value: 0.03). A recent study based on exome sequencing of 3,871 autism cases delineated 33 risk genes at FDR < 0.142. 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: 5e-3). Supplementary Table 7 includes the ALoFT predictions for *de novo* LoF 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 exomes43. As shown in Figure 4c, 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 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 mutations44. Considering the fact that loss-of-function of tumor suppressor genes can lead to oncogenesis, ALoFT can be used to identify disease-causing pLoFs in cancer genes, especially those in tumor suppressors.

Case study 4: Distingusihing 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 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 known45. 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). Nonetheless, a large number of both common and rare premature stop variants are seen at the end of the coding genes in the 1KGP1, 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 1KGP1, ESP6500 and ExAC 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.

Application to 1KGP1: 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 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 5,495 premature stop variants from the 1KGP1 dataset (Online Methods). The predicted benign LoF score for pLoFs in this population cohort have a wide range of values (Fig. 2a, Supplementary Table 5). On average, each individual is a carrier of two rare heterozygous premature stop alleles that are predicted to be disease-causing in the homozygous state (Supplementary Table 6). 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 alleles46-50. The estimation and disease-causing potential of deleterious 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. It should be noted that the HGMD data includes 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 diseases46.

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: 6.9e-9). Our results provide a possible rationale for this observation. Firstly, variants predicted to be benign in 1KGP1 often affect isoforms that are different from the isoforms containing the disease-causing HGMD variant. This suggests that LoFs in healthy individuals affect minor isoforms.12.4% of premature stop variants in the presumed healthy 1KGP1 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 1KGP1 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 functional. Lastly, a majority of 1KGP1 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 1KGP1 cohort.

In summary, we describe a tool for predicting the impact of pLoF 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 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 targets51,52. 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 diseases53. 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 NMD54-60. 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. Sulem, P. *et al.* Identification of a large set of rare complete human knockouts. *Nat Genet* (2015).

4. 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).

5. 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).

6. 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).

7. Milazzo, L. & Antinori, S. Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *N Engl J Med* **366**, 2425; author reply 2426 (2012).

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

34. Breiman, L. Random Forests. *Machine Learning* **45**, 5-32 (2001).

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. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-99 (2012).

38. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-5 (2012).

39. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-41 (2012).

40. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-50 (2012).

41. 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).

42. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-15 (2014).

43. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-21 (2013).

44. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-58 (2013).

45. Inoue, K. *et al.* Molecular mechanism for distinct neurological phenotypes conveyed by allelic truncating mutations. *Nat Genet* **36**, 361-9 (2004).

46. Bell, C.J. *et al.* Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med* **3**, 65ra4 (2011).

47. 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).

48. 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).

49. 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).

50. 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).

51. Kaiser, J. The hunt for missing genes. *Science* **344**, 687-9 (2014).

52. Alkuraya, F.S. Human knockout research: new horizons and opportunities. *Trends Genet* (2014).

53. 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).

54. Bhuvanagiri, M. *et al.* 5-azacytidine inhibits nonsense-mediated decay in a MYC-dependent fashion. *EMBO Mol Med* **6**, 1593-609 (2014).

55. Bhuvanagiri, M., Schlitter, A.M., Hentze, M.W. & Kulozik, A.E. NMD: RNA biology meets human genetic medicine. *Biochem J* **430**, 365-77 (2010).

56. 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).

57. 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).

58. 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).

59. 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).

60. 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**

Macintosh HD:Users:suganthi.bala:Documents:Data:aloft_submitted:dec2015_edition:Figure 1_feb13_2016.pdf

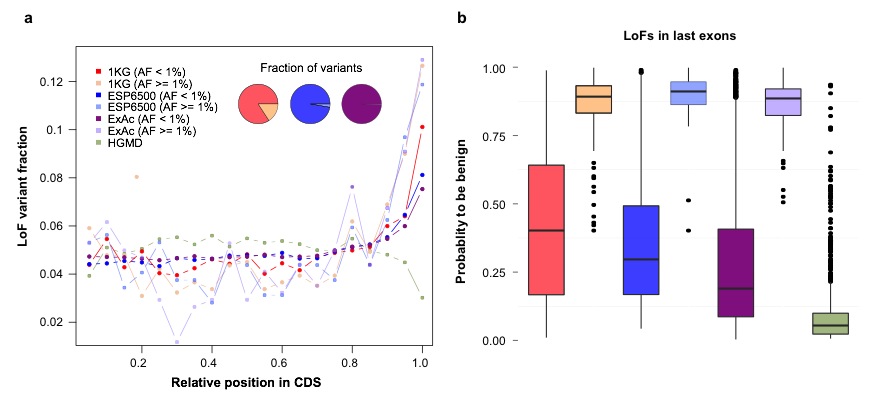
**Figure 1 - Schematic workflow.**

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**

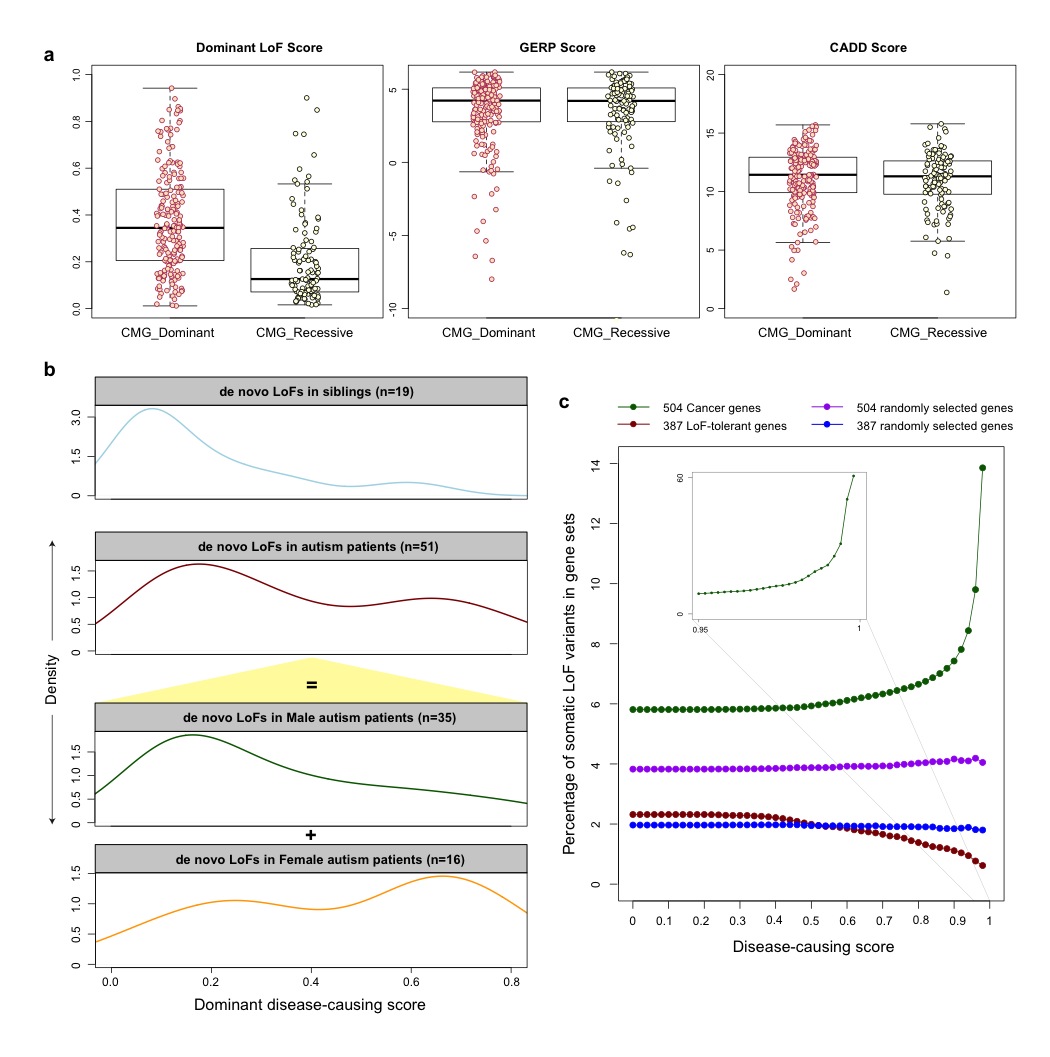
Benign LoF score for premature stop variants in 1KGP1 and HGMD. For this plot, we randomly selected one variant per gene. The third (dark green) box plot pertains to premature stop variants in healthy 1KGP1 individuals occuring in disease-causing genes obtained from HGMD. The fourth (blue) box plot pertains to LoF variants in the subset of HGMD genes where 1KGP1 LoFs are also seen. “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.



**Figure 3:** a) Position of premature stop variants in coding transcripts. Compared to HGMD variants, both common and rare 1KGP1, 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 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 ClinVar variants. All training varinats are excluded. Average tolerance scores are 0.097 and 0.115 respectively for dominant and recessive datasets.

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.