Loss-of-function genetic variants (LoF) are of great interest because they are rare variants often associated with diseases. However, recent sequencing efforts indicate the wider prevalence of LoF variants in seemingly healthy humans. To better understand putative LoF variants, we developed a classifier that uses ALoFT (**A**nnotation of **L**oss-**o**f-**F**unction **T**ranscripts), a pipeline that annotates LoF variants with functional, evolutionary and network features, to distinguish among benign, recessive and dominant disease-causing premature stop variants. We applied ALoFT to several datasets of healthy and diseased cohorts, including the 1000 Genomes Phase1 dataset and predict that approximately 47% of premature stop variants occurring as heterozygous alleles in healthy individuals can potentially lead to disease if present as homozygous, while 49% of the variants are predicted to be benign. ALoFT is able to distinguish between disease-causing dominant heterozygous and recessive homozygous premature stop variants in accord with recently published Mendelian studies. Using this method we also show that a higher proportion of *de-novo* premature stop variants are deleterious in autism probands than in controls and that cancer driver genes contain a higher proportion of predicted pathogenic premature stop variants than other genes, which is in agreement with published results.

About 12% of known disease-causing mutations in the Human Gene Mutation Database (HGMD) are due to nonsense mutations[1]. 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[2]. However, a recent large-scale expression analysis demonstrated that 68% of predicted NMD events due to premature stop variants are unsupported by RNASeq analyses[3]. 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[4]. 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.

One of the most notable findings from personal genomics studies is that all individuals harbor LoF variants in some of their genes[5]. A systematic study of LoF variants from 180 individuals revealed that there are hundreds of putative LoF variants in an individual[6]. 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[7,8] and hence the active pursuit in the inhibition of PCSK9 as a potential therapeutic for hypercholesterolemia[9-11]. 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[12-15].

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. The main features of ALoFT include 1. Function – based annotations 2. Conservation 3. Network. 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[6]. An overview of the pipeline is shown in Supplementary Figure 1. For comprehensive functional annotation, we integrated several functional 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 for macaque and mouse (ratio of missense to synonymous substitution rates) that are computed from human-macaque and human-mouse orthologous alignments respectively. 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[6,22]. A detailed description of all the annotations provided by ALoFT is included in the Supplementary Material and Methods section (Table S1). Detailed documentation, input data files and source code linked to github can be found at http://aloft.gersteinlab.org.

To understand the impact of putative LoF variants on gene function, we developed a prediction method to differentiate disease-causing variants from benign variants. Here, we focus on premature stop variants that arise either due to a SNP or an indel where a frameshift leads to a premature stop. 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 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 AUC with 10-fold cross-validation is 0.955. The precision for the three classes are as follows: Dominant=0.82, Recessive=0.83, Benign=0.89. The classifier is very robust to the choice of the training data sets and performs well with different training data sets (Table S2).

We applied ALoFT to 5665 heterozygous premature stop variants from Phase1 1000Genomes. The predicted benign LoF score for the premature stop variants in seemingly healthy people have a wide range of values (Figure 2a). 2993 premature stop variants in 1000 Genomes dataset are predicted to be benign, 2823 variants can lead to recessive disease and 253 variants can lead to disease via a dominant mode of inheritance (Table S2 and S6). On average, each individual is a carrier of about eight rare heterozygous premature stop alleles that can be disease-causing in the homozygous state (Table S3).

Next, we looked at premature stop variants in the 1000 genomes cohort in known disease-causing genes. This subset of 474 variants indicate that seemingly healthy people carry premature stop variants in disease-causing genes. As expected, ALoFT predicts that most of these mutations are benign or will cause disease only in the recessive state but are seen in the healthy population as heterozygous variants. Interestingly, in some cases, the variant in the presumed healthy 1000 genome 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 the 1000 genomes cohort 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. Therefore, this suggests that isoform-specific premature stop variants are responsible for disease and are not seen in the presumed healthy 1000 Genomes individuals.

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 and therefore the truncated protein will be expressed and will not lead to 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. A large number of premature stop variants are seen at the end of the coding genes in both the 1000 Genomes and ESP6500 datasets (Figure 2c). Variants in the last coding exon in the 1000 Genomes and ESP6500 cohort are more likely to be benign, whereas HGMD mutations in the last coding exon tend to be disease-causing (Figure 2d, median benign LoF scores for 1000 Genomes, 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 also 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). We applied our method to de-novo LoF mutations discovered in autism[32-35]. 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]. Mutations in these 33 genes have higher dominant disease causing LoF score than others (Figure S5; p-value: 0.003). Table S7 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 the gene had greater than 20% of the mutations that 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 the Figure 3c, we observe that somatic LoF mutations tend to occur in known cancer driver genes compared to randomly sampled genes whose length distribution matched 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 1000 genomes population.

To our knowledge, ALoFT is the first tool that predicts the impact of nonsense SNPs in the context of a diploid model, i.e. whether nonsense SNP will lead to recessive or dominant disease. This method is applicable to premature stop variants and frameshift-causing indels. ALoFT allows for the identification and prioritization of high impact putative disease-causing LoF variants in a personal genome from amongst benign LoF variants. 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 provides an unique opportunity for targeted therapy of a wide variety of diseases using drugs that either enable read-through of the premature stop restoring the function of the mutant protein or an NMD inhibitor that prevents degradation of the LoF-containing transcript by NMD. 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.

1. 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).

2. Isken, O. & Maquat, L.E. Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes Dev* **21**, 1833-56 (2007).
3. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-11 (2013).
4. 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).
5. Balasubramanian, S. *et al.* Gene inactivation and its implications for annotation in the era of personal genomics. *Genes Dev* **25**, 1-10 (2011).
6. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-8 (2012).
7. 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).
8. 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).
9. 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).
10. Milazzo, L. & Antinori, S. Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *N Engl J Med* **366**, 2425; author reply 2426 (2012).
11. Stein, E.A. *et al.* Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *N Engl J Med* **366**, 1108-18 (2012).
12. Flannick, J. *et al.* Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat Genet* **46**, 357-63 (2014).
13. Lim, E.T. *et al.* Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet* **10**, e1004494 (2014).
14. Tachmazidou, I. *et al.* A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nat Commun* **4**, 2872 (2013).
15. 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).
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. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-41 (2012).
34. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-5 (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).