Loss-of-function variants (LOF) attract great clinical interest, as it is believed that most of them are pathogenic. About 20% of known disease-causing mutations in the human gene mutation database, HGMD, are due to nonsense mutations. However, one of the most notable findings from personal genomics studies is that all individuals harbor LOF variants in some of their genes. A systematic study of LOF variants from large-scale sequencing studies revealed that there are about 100 putative LOF variants in each individual. Thus, several genes are knocked out either completely or in a transcript-specific manner in apparently healthy individuals. Remarkably, recent studies have led to the discovery of LOF variants that are beneficial. For example, nonsense variants in PCSK9 are associated with low LDL levels. Therefore, several pharmaceutical companies are actively pursuing the inhibition of PCSK9 as a potential therapeutic for hypercholesterolemia. 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. Therefore, there is great interest and benefit in understanding putative LOF variants.

It is often assumed that premature Stop-causing variants are deleterious as they are predicted to lead to loss-of-function and therefore are likely to cause disease. Truncated proteins may be harmful to the cell and such transcripts are typically removed by nonsense-mediated decay (NMD), an mRNA surveillance mechanism. However, understanding the functional impact of premature Stop codon is not straightforward. Transcripts containing premature Stop codons do not always undergo NMD. Moreover, transcript-specific LOF variants may or may not affect the function of the gene. In addition, loss-of-function of a gene might not have any overall impact on the fitness of the organism. Here, we present a method to infer the effect of a class of LOF variants, SNPs that lead to premature Stop codons. We have developed a pipeline called ALOFT (Annotation of Loss-Of-Function Transcripts), to provide extensive functional annotation of LOF variants. Using the features from ALOFT, we developed a prediction model in order to distinguish high impact LOF variants from low impact variants. We show that the model performs well on various datasets that include healthy cohorts as well as disease data, separating LOF events with high impact from those that are not.

ALOFT provides extensive functional annotation of LOF variants. The main modules of ALOFT include 1. Function – based annotations 2. Evolutionary features 3. Network features. In addition, the pipeline has two modules to help identify erroneous LOF calls: mismapping and annotation error modules that provide information that alerts the user about potential mismapping and annotation errors. An overview of the pipeline is shown in Figure 1a.

We integrated several functional annotation resources to get the most comprehensive functional annotation. They include annotations such as PFAM and SMART functional domains, signal peptide and transmembrane annotations, post-translational modification sites, structure-based features such as SCOP domains and disordered residues. For all functional features, we assessed if the Stop-causing variant affected a functional feature and if the region truncated due to the premature Stop led to loss of functional domains/features. We also identified transcripts containing a premature Stop as candidates for nonsense-mediated decay (NMD) if the distance of the premature Stop from the last exon-exon junction was greater than 50 base pairs.

Evolutionary conservation can be used as a proxy for identifying functionally important regions. ALOFT provides variant position-specific GERP scores. In addition, we evaluate if the truncated region is conserved based on GERP constraint elements and the percentage of truncated exons that are within GERP constrained elements. ALOFT also outputs dn/ds values for macaque and mouse.

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. Detailed description of all the annotations provided by ALOFT is included in the Supplementary Material and Methods section.

To understand the effect of putative LOF variants on gene function, we developed a prediction method to differentiate high impact LOF variants from low impat benign variants. Current prediction methods that infer the pathogenicity of variants do not take into account the ploidy of the variant. The majority of LOF variants in healthy population cohorts are heterozygous. It is likely that a subset of these variants will be pathogenic in the recessive state. Therefore, we developed a prediction model to classify nonsense variants into those that are benign, 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 included the following gene-based features in our prediction method: missense SNP density in a gene derived from Phase1 1000 Genomes variation data, GERP scores as a measure of conservation for missense variants, presence of missense SNPs in constrained GERP elements, number of miRNA binding sites and average heterozygosity of each gene. We also used the allele frequency of variants from the ESP6500 project and gene expression data from xx tissues obtained from GTex consortium.

To build the classifier, we used three training datasets: premature Stop-causing variants that are homozygous in at least one individual in the Phase1 1000 Genomes data that represent benign stop-causing variants, nonsense mutations from HGMD that lead to recessive disease and those where the mode of inheritance is dominant. Using the functional, evolutionary and other features described above, we built a classifier that distinguishes the three classes using a random forest algorithm. We obtain very good discrimination between the three classes (Fig 1b). The accuracy of the predictions are Dominant=0.8601767, Recessive=0.8012719, Benign=0.9198012. Thus, this method can be used to identify high impact premature Stop variants that are predicted to be disease causing from a list of LOF variants in a personal genome.

We tested the classifier on a dataset of premature Stop-causing variants from Phase1 1000Genomes that represents a healthy cohort. In Figure 2c, we see that the predicted benign LOF score for the premature Stop variants in seemingly healthy people have intermediate values ranging between benign and disease-causing scores. Based on the results of the classification, we predict that 3242 premature Stop-causing variants in 1000 Genomes dataset are benign, 2793 variants can lead to recessive disease and 104 variants can lead to disease via a dominant mode of inheritance. Thus, 48% of premature Stop variants in apparently healthy individuals from the 1000 genomes population harbor loss of function mutations that are high impact variants. A majority of these high impact variants are heterozygous and will cause disease only in the recessive state. Next, we looked at the subset of premature Stop mutations in known disease-causing genes that are also present in the 1000 genomes population. This subset of

mutations indicates that seemingly healthy people carry Stop-causing mutations in disease-causing genes. As expected, our classifier predicts that most of these mutations will cause disease only in the recessive state but are seen in the healthy population as heterozygous variants. However, 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 transcripts. This is illustrated in Fig 2d. Thus, transcript-specific premature Stop-causing variants are responsible for disease and are not seen in the presumed healthy 1000 Genomes individuals. In other cases, the 1000G LOF variant and the disease-causing HGMD variant are on the same transcript. However, the 1000G LOF variant truncates the protein at a position where there function of the protein is not affected whereas the disease-causing LOF affects the function (Supplementary figure).

We next applied our classifier to predict the effect of premature Stop-causing variants in the last exon. It is often assumed that premature Stop-variants in the last 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, it is possible that such mutations might still affect function if functional residues are lost due to truncation. We applied our classifier to see if we could distinguish between benign LOF variants in the last exon from disease-causing variants in the last exon. Specifically, we used Stop-causing mutations in the last exon from 1000 genomes, ESP6500 and HGMD to differentiate benign LOF variants from disease-causing ones. It has been observed that there are more number of Stop-causing variants at the end of the coding genes in both the 1000 Genomes and ESP6500 datasets (Fig 2a). The classifier correctly predicts that most variants in the last exon in the 1000 Genomes and ESP6500 cohort are benign, whereas the disease-causing HGMD mutations in the last exon are not (Fig. 2b).

Finally, we applied this method to infer the effect of nonsense mutations in several recently published disease studies. We classified premature Stop mutations from the Center For Mendelian Genomics studies and correctly predict the mode of inheritance and pathogenicity of all of the truncating variants (Fig 3a). We used GERP score of the LOF variant position and CADD score that scores all positions in the genome for deleteriousness to classify recessive versus dominant LOF variants. However, both CADD and GERP scores are not good discriminators of recessive versus dominant disease.

We also validate our method by applying our classifier to four different autism studies. De-novo LOF SNPs have been implicated in autism. Our method shows that dominant disease-causing de-novo LOF events are significantly higher in autism cases versus controls (Fig 3b). Moreover, female autism probands have a higher proportion of deleterious de-novo LOF variants than male probands. This is in agreement with studies that show that while autism is more prevalent amongst males than females, the severity of the disease is much higher in females.

Lastly, we also examined somatic Stop-causing mutations in several cancers. To classify driver genes as tumor suppressors or oncogenes, 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. Therefore, we expect to see deleterious LOF variants in driver genes. As shown in the Fig 3c and 3d, somatic LOF mutations in cancer driver genes are predicted to be deleterious whereas somatic mutations in LOF-tolerant genes are predicted to be benign. Thus, our prediction method can be used to identify pathogenic LOF variants and driver genes that are tumor suppressors.

To our knowledge, this is the first method 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 classification will allow us to identify and prioritize high impact putative disease-causing LOF variants from a list of LOF variants in a personal genome. Moreover, this method allows us to identify benign LOF variants. Integrating benign LOF variants with phenotypic information will help us to identify protective/beneficial LOF variants. Lastly, diseases caused by LOF variants provides a 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 a 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 as elegantly opined by Brooks et al.