Introduction

Loss-of-function variants (LOF) attract great clinical interest, as it is believed that most of them are potentially pathogenic. About 20% of known disease-causing mutations in the human gene mutation database, HGMD, are due to nonsense mutations. However, recent sequencing efforts demonstrate the presence of null variants in seemingly healthy people. Moreover, some LOF variants are also known to be beneficial. For example, LOF variants in PCSK9 are associated with low LDL levels and several pharmaceutical companies are actively pursuing the inhibition of PCSK9 as a potential therapeutic for hypercholesterolemia. Therefore, there is great interest in understanding putative LOF variants. Here, we present a method to infer the effect of LOF variants, primarily those due to premature Stop codons.

Coding variants have been the subject of extensive research. Several algorithms have been published to infer the effect of missense coding variants on protein function. However, there is a paucity of methods that are applicable to nonsense variants. (mention of SIFT indel). In addition, current prediction methods that infer the pathogenicity of variants do not take into account the ploidy of the variant. It is clear from recent exome studies that rare heterozygous variants abound in the human genome. It is likely that a subset of these variants will be pathogenic in the recessive state. To address these issues, we have developed ALOFT (Annotation of Loss-Of-Function Transcripts), a pipeline to extensively annotate putative LOF variants with a variety of functional and evolutionary features. Using these features, we developed a predictive model to classify nonsense variants into those that are benign, that lead to recessive disease and those that lead to dominant disease. We have focused on SNPs that lead to a premature Stop codon. While this method can be extended to frame-shift causing indels, we did not include indels in the training model as indel calling methods are not robust and have a high error rate.

ALOFT pipeline

We developed a pipeline to provide 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 which outputs information that alerts the user about potential mismapping and annotation errors. An overview of the pipeline is shown in Figure 1a. Detailed description of all the fannotations provided by ALOFTis included in the Supplementary Material and Methods section.

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, and information on post-translational modification sites. The 3D structure of the protein is essential for proper folding and function of proteins. Disordered residues have been known to be important in protein-protein interaction surfaces and have been implicated in disease-causing mechanisms. Therefore, we included structure-based features such as SCOP domains and information on predicted 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 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.

Previous studies have shown that neighbors of disease genes in a protein-protein interaction network often cause the same or related diseases. The network module includes two features: 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.

Classifier

We used the features from ALOFT to build a classifier that distinguishes benign premature Stop variants from those that lead to recessive disease and those that lead to dominant disease. In addition to the features output by ALOFT, we included the following gene-based features based on Phase1 1000 Genomes variation data in our prediction method : nonsynonymous SNP density in a gene, conservation of nonsynonymous variants as GERP scores, presence of nonsynymous 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.

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 SNPs from HGMD that lead to recessive disease and those where the mode of inheritance is dominant. Mutations that lead to dominant inheritance of diseases can do so both via loss of function as well as gain of function mechanisms To make sure that we are predominantly modeling loss-of-function effects, we only used mutations in predicted haploinsufficient genes as the model for dominant diseases. 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.

Applying the classifier

We applied this method to elucidate the effect of premature Stop-causing variants from Phase1 1000Genomes as a test dataset. Nonsense mutations catalogued in HGMD are disease-causing and therefore have lower probability of being benign. In Figure 2c, we see that the benign LOF score for the premature Stop variants in seemingly healthy people have intermediate values ranging between benign and disease-causing scores. Based on our classifier, 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. Next, we looked at the subset of premature Stop mutations that are present in disease-causing genes in the 1000 genomes population. This subset of mutations indicate that seemingly healthy people carry Stop-causing mutations in disease-causing genes. Our classifier predicts that most of these mutations will cause disease in the recessive state but are seen in the healthy population as heterozygous variants. 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 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 clearly 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). However, 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). While it is known that autism is more prevalent amongst males, it has been seen that the severity of the disease is much higher in females. This seems to be corroborated by the fact that female autism probands overall have a higher proportion of deleterious de-novo LOF variants than male probands. We also examined somatic Stop-causing mutations in several cancers. 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.

Discussion/Conclusion

To our knowledge, this is the first method that predicts the impact of nonsense SNPs on protein function (might have to modify this in the light of SIFT indels and the latest immune genes paper). Moreover, this method predicts the pathogenicity of a nonsense variant in the context of a diploid model, i.e. whether nonsense SNP will lead to recessive or dominant disease. Predicting the effect of nonsense SNPs enables one to identify putative disease-causing LOF variants. Disease-causing premature Stop variants provide 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 targetting 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. Additionally identifying and integrating benign LOF variants with phenotypic information will allow us to identify protective/beneficial LOF variants similar to those seen PCSK9, LPA, SLC30A8 etc.

Online Methods