Loss-of-function variants (LoF) attract great clinical interest, as it is believed that most cause disease. In this study, we include (1) premature Stop-causing SNPs, (2) frameshift-causing indels and (3) variants affecting canonical splice sites as putative LoF variants. About 12% of known disease-causing mutations in the human gene mutation database (HGMD) are due to nonsense mutations[1]. It is often assumed that premature Stop variants are deleterious as they are predicted to lead to loss-of-function. However, understanding 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 shows that 68% of predicted NMD events due to premature Stop variants were not supported by RNASeq analyses[3]. In fact, 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. In addition, understanding isoform-specific LoF variants is complex. 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 level 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 about 100 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 LoF variants that are beneficial. For example, nonsense variants in PCSK9 are associated with low LDL levels[7,8]. Therefore, several pharmaceutical companies are actively pursuing 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]. Therefore, there is great interest in a more thorough understanding of putative LoF variants.

We have developed a pipeline called ALoFT (**A**nnotation of **L**oss-**O**f-**F**unction **T**ranscripts), to provide extensive functional annotation of putative LoF variants. The main features of ALoFT include 1. Function – based annotations 2. Conservation features 3. Network features. In addition, the pipeline has 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, structure-based features such as SCOP domains, disordered residues and prediction of NMD[18,19]. Evolutionary conservation can be used as a proxy for identifying functionally important regions. ALoFT provides variant position-specific GERP scores, which is a measure of evolutionary conservation[20]. In addition, we evaluate if the region lost due to the truncation is conserved based on GERP constraint elements and the percentage of

exons lost that are within GERP constrained 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. The source code is available from https://github.com/gersteinlab/ALoFT. Detailed documentation and input data files can be found at http://aloft.gersteinlab.org

To understand the effect of putative LoF variants on gene function, we developed a prediction method to differentiate high impact disease-causing variants from low impact benign variants. Here, we focus on premature Stop-causing variants that arise either due to a SNP or an indel where a frameshift leads to a premature Stop. Current prediction methods that infer the pathogenicity of variants do not take into account the zygosity of the variant[23,24]. 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, 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 some gene-specific features for classification as shown in Figure 1a (details included in the Supplementary Material and Methods section).

Suganthi Balasubra…, 12/15/14 3:11 PM

Comment [1]: SB to MG: I want to remove the terms "high impact" and "low impact". Benign variants that are protective are also high impact variants and cannot be classified as low impact variants.

To build the classifier, we used three training datasets: premature Stop variants that are homozygous in at least one individual in the Phase1 1000 Genomes data that represent benign stop-causing variants, homozygous premature Stop mutations from HGMD that lead to recessive disease and heterozygous premature Stop variants in haploinsufficient genes that lead to dominant disease[25]. All the mutations from HGMD are not used for training. The training dataset is restricted to mutations that lead to dominant or recessive disease based on a curated gene list [26]. Using the functional, conservation and other features described above, we built a classifier that distinguishes the three classes using a random forest algorithm (ref for random forest). The classifier provides class probability estimates for the model. We obtain very good discrimination between the three classes (Fig 1b). The accuracies for the three classes with 5-fold cross-validation are as follows: Dominant=0.88, Recessive=0.82, Benign=0.93. The multiclass AUC is 0.953. The classifier is very robust to the choice of the training data sets and performs well when trained with different training data sets (Table S1)

We tested the classifier on a dataset of XX# of premature Stop variants from Phase1 1000Genomes, (excluding the YY# of homozygous LoF variants used as the training dataset) that represents a healthy cohort. We found that the predicted benign LoF score for the premature Stop variants in seemingly healthy people have values ranging between benign and disease-causing scores (Figure 2a). We predict that 3323 premature Stop variants in 1000 Genomes dataset are benign, 2639 variants can lead to recessive disease and 107 variants can lead to disease via a dominant mode of inheritance (Table S2 and S5). Thus, 43.5% of heterozygous premature Stop variants in apparently healthy individuals from the 1000 genomes population are predicted to cause disease in the recessive state. On average, each individual is a carrier of about eight rare recessive premature Stop alleles (Table S3). As expected, individuals from African

population possess the highest number of premature Stop variants (Figure S4) and the difference is caused by excess of benign mutations (Figure S5). ()

Suganthi Balasubra…, 12/15/14 5:08 PM
Comment [2]: wondering if this will open a can of worms in terms of LoF numbers per individual. I am inclined to remove this bit.

Next, we looked at premature Stop variants in the 1000 genomes cohort in known disease-causing genes, of which none were expected, as these are healthy individuals. However, ALoFT predicted that several variants cause disease in the recessive state but are present in this cohort only as heterozygous variants. Interestingly, in some cases (%XX), 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). Thus, isoform-specific premature Stop-causing variants are responsible for disease and are not seen in the presumed healthy 1000 Genomes individuals. In other cases, the LoF variant in 1000 genomes and the disease-causing HGMD variant are on the same transcript. However, the LoF variant in the 1000 genomes samples truncates the protein at a position much later than the disease-causing variant. Presumably, the former doesn't affect function significantly whereas the latter does.

We next applied our classifier to predict the effect of premature Stop variants in the last exon. 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, it is known that some disease-causing premature Stop mutations are present in the last coding exon. Therefore, we applied our classifier 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 (Fig 2c). The classifier correctly predicts that most variants in the last coding exon in the 1000 Genomes and ESP6500 cohort are benign, whereas HGMD mutations in the last coding exon are not (Fig. 2d).

We further evaluated the classifier 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 dominant variants have significant dominant disease-causing score than recessive ones (p-value: 0.003; Wilcox 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[27]. 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[28-31]. 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: 0.006; Wilcox rank-sum test). Autism is more prevalent amongst males than females. However, the severity of the disease is known to be much higher in females. Previous studies suggest that there is a higher mutational burden in female patients [32]. We observe a similar pattern for LoF mutations – female probands have a higher portion of deleterious de-novo LoF variants than male probands (p-value: 0.017). The published studies identified candidate genes based on

statistical enrichment of de novo LoF variants in probands over the unaffected individuals. Our classifier identified yy# disease-causing de novo LoF variants in zz# genes, and is agnostic to unaffected and proband status. A recent study based on exome sequencing of 3871 autism cases delineated xxx # of risk genes[33]. Mutations in this set of autism genes[34] have higher dominant disease causing score than others (Figure S6; p-value: 0.008).

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[35]. Therefore, we expect to see a higher proportion of deleterious somatic LoF variants in driver genes than the rest of the genes. We validated our prediction method by inferring the effect of somatic premature Stop variants from a compilation of ~6,000 cancer exome sequencing studies[36]. As shown in the Fig 3c, a higher proportion of somatic LoF mutations in known cancer driver genes are predicted to be deleterious than somatic mutations in LoF-tolerant genes and randomly sampled genes whose length distribution matched that of the known driver genes.

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/beneficial LoF variants which are valuable drug targets[37]. 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. Hu, J. & Ng, P.C. SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. *PLoS One* **8**, e77940 (2013).
24. Rausell, A. *et al.* Analysis of stop-gain and frameshift variants in human innate immunity genes. *PLoS Comput Biol* **10**, e1003757 (2014).

25. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
26. Blekhman, R. *et al.* Natural selection on genes that underlie human disease susceptibility. *Curr Biol* **18**, 883-9 (2008).
27. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
28. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-99 (2012).
29. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-41 (2012).
30. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-5 (2012).
31. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-50 (2012).
32. 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).
33. Poultney, C.S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-15 (2014).
34. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-15 (2014).
35. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-58 (2013).
36. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-21 (2013).
37. Kaiser, J. The hunt for missing genes. *Science* **344**, 687-9 (2014).