# Response Letter

### Reviewer 1:

|  |  |
| --- | --- |
| **Reviewer comment 1** | The authors need to reason more clearly or state more explicitly in the main text which ALoFT values they consider to be robust indicators for pathogenicity. This would truly help geneticists discriminate between benign and damaging LoFs. Figures 2-4 address that to some degree, but solely presenting these graphs is not informative enough in my opinion for the general genetics reader. It is essential to show the power of LoF values in concrete numbers as the selling point this article; namely that ALoFT can be used by the genetics community to distinguish LoF that are disruptive from benign variation. It is difficult to understand Figure 2 at first glance. |
| **Authors’ response** | We agree with the reviewer that graphs alone are not enough for the general genetics reader. However, we would like to bring to the reviewer’s attention that we have included tables of predicted ALoFT scores as Supplementary Tables 5 and 7 and uploaded them as additional files. Given that they are big files, we did not include it in the Supplementary pdf but did include it as Supplemental additional files in our original submission. ***However, in this submission, we will also include it as part of the Supplementary “pdf” file so as to avoid any confusion.***  ***We have also clarified this further in the text clearly indicating that ALoFT provides three scores corresponding to the three classes: benign, dominant and recessive. The dominant and recessive scores are the scores relevant to pathogenicity. Moreover, the effect of the variant is clearly output in the prediction file in the last column of the file. For example, in Supplementary Table 5, the last column indicates if the pLoF variant will be benign or have a dominant or recessive effect.*** |
| **Changes in text** | 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.  Changes made on Page 4, first paragraph |

|  |  |
| --- | --- |
| **Reviewer comment 2** | The sections on cancer genomics and autism seem somewhat separate entities in the manuscript. It is not clear how these, especially the section and figure on cancer genomics, integrate and relate to the preceding material. |
| **Authors response** | The section on cancer genomics and autism are included to illustrate the utility of ALoFT in prioritizing pLOF variants. We used known case studies and show that the ALoFT scores of pLOF variants corroborates previously published research, thus validating the method. In the case of autism, previous studies have shown that *de novo* pLoF variants are associated with autism spectrum disorder based on differences in mutation rates in genes in unaffected individuals versus affected probands. We have been able to show the same results from an entirely different approach. In the case of cancer, ALoFT is able to discriminate between pLoF variants in known driver genes versus other genes. ***We have made these points clearer in the text.*** |
| **Changes in text** | We have reorganized the text so that the motivation for the different sections is clear. We now lay out the paper in the following way:  1. Description of the ALoFT classifier, the training datasets used and the evaluation of the classifier using metrics such as AUC, PPV etc.  2. Validating ALoFT predictions by applying them to known disease variants and corroborating published studies.  3. Application of ALoFT to understand pLoFs in personal genomes. |

|  |  |
| --- | --- |
| **Reviewer comment 3a** | 3) Source databases. It seems somewhat surprising that the authors relied so heavily on 1000 genomes data for this paper. This seems somewhat outdated now that there are other public resources available that comprise data of thousands of exomes (i.e. ExAC with >60,000 exomes). Because the authors specifically mention that allele frequency is the most important feature for ALoFT classification, it would seem that up-to-date integration of allele frequency from as large and well annotated a database as possible is essential to optimize the use of ALoFT as a bioinformatic variant analysis tool. |
| **Authors’ response** | Of the 5,495 SNPs that introduce a premature stop codon in the 1KGP1 dataset, we only used 397 homozygous SNPs for the training model. Thus, only a very small fraction of the 1KGP1 was used as the training dataset. However, we used allele frequencies from both 1KGP1 and ESP6500 as prediction features.  There were two reasons for not using data from ExAC consortium: 1. This work was planned and most of the work was executed before the ExAC frequencies were released. 2. We did not revise the model with ExAC frequency before submission as the ExAC data was released under Fort Lauderdale agreement and it was unclear if we could use the entire data set for training our model. Based on the reviewer’s critique, we requested Daniel MacArthur, the lead member of ExAC consortium and a co-author on this paper if we could use the allele frequency data and homozygous SNP data for training our model. ***We have revised the manuscript by including ExAC allele frequency as a feature***. However, we are not allowed to use the homozygous variant information from ExAC for training the model as this constitutes global use of this data.  We would also like to bring to the attention of the reviewer that we specifically show that ***while allele frequency improves the accuracy of the prediction, the classifier performs well even in the absence of allele frequency.*** In fact, prediction accuracy was lower when allele frequency features were the **only features** used for training the model (this was included in Supplementary table 4, multiclass AUC =0.79 for a prediction model trained based on only allele frequency from ESP6500 and ExAC. Network (AUC=0.81), functional (AUC=0.85) and evolutionary (AUC=0.86) features outperform allele frequency features when used separately to train the classifier). Thus, integration of all features improves prediction accuracy (AUC = 0.97)**.**  It should also be noted that it is not possible to estimate the relative importance of each feature because some features are correlated with each other. Therefore, the effect of permuting the values of one feature will be masked by other features that could compensate for this feature in random forest method.  ***We have revised the text by updating the model with allele frequency from ExAC. We have also included ExAC in Figure 3.*** |
| **Changes in text** | Changes made in Page 4, Paragraph 2. Figure 3 modified to include ExAC. |

|  |  |
| --- | --- |
| **Reviewer comment 3b** | 3) Does the authors' claim that per individual only 2 LoF variants (of  over 100 putative variants) could lead to disease if present in  homozygous state stand up when validated in larger database? This number seems rather low. |
| **Authors Response** | **ExAC does not provide individual level information and hence we calculated this number for 1KGP1. Also, please note that this number pertains only to pLoFs that result due to the introduction of Stop codon due to a SNP (splice and frameshift indels are not included in this calculation).**  **This response to this comment needs to made stronger** |
| **Change in text** | Yao: Is it possible to estimate a per individual number for ExAC based on the allele frequency? I realize we don’t have sample level information. The ExAC paper is on bioarxiv. We should check to see if they have any such discussion. Also, can you please estimate this number including indels in the 1KGP1? If including indels, then please change manuscript accordingly. Unsure if this will add anything as Phase1 only provided us with a small subset of mostly high frequency indels. |

|  |  |
| --- | --- |
| **Reviewer comment 4** | It would appear that the same population - same samples/variants were used for model training and then used again test the software  performance as shown in Fig 2 and Fig 3. Is this really the case? If  not please clarify. If so, how is this justified? A different an  independent set of data should be used to validate the trained model. |
| **Authors Response** | We apologize for the confusion due to the use of 1KGP1 and HGMD nomenclatures for both training and testing. ***Completely different variants were used for training and testing.*** For example, all training variants are removed in Fig 2 and Fig 3. ***We have added Supplementary table 2 to clarify this issue.*** The model was trained on ***homozygous*** LoF variants in 1KGP1. The rest of the analysis is done on ***heterozygous*** LoF variants. For the training sets comprising of disease mutations, only HGMD mutations in genes that could be assigned to either the dominant or the recessive category based on OMIM were used. Of the 1,800 (Yao: is this number correct for the updated HGMD dataset) disease genes in HGMD, only 932 genes could be assigned to dominant or recessive category. ***Thus, we only used a subset of HGMD variants for training and the remaining variants for other analysis.*** Please note that the legend to Figure 3 clearly stated that training variants are not included in the figure. |
| **Changes in text** | We have added Supplementary table 2 that includes the number of variants that were used for training from the 1KGP1 and HGMD datasets. Besides training the classifier, all other reported analyses exclude variants that were used to train the model. |

|  |  |
| --- | --- |
| **Reviewer comment 5** | 5) Some known dominant or recessive disease causing mutations from CMG were tested by ALoFT, as well as GERP and CADD shown in Fig 4. However, the number of tested variants is very small at the end of the day (4 dominant, 9 recessive); with even a few outliers, these variants could also overlap with HGMD variants in training dataset. In such small testing sample set, it seems 1 out of 4 dominant mutations was an outlier and low; while 2 out of 9 recessive mutations were outliers, and LoF scores were overlapping with dominant mutations, which were very high. A larger number of tested samples would increase overall confidence in the robustness of the approach. |
| **Authors response** | We realize that the number of testing variants is small. To address the reviewer’s point about the robustness of the approach, we used ALoFT to classify pathogenic variants from ClinVar that do not overlap with the training variants. This is now shown as new Figure 4a. CMG figure is moved to the supplement as supplementary figure 4 |
| **Change in text** | We have added a new Figure 4a which shows the performance of ALoFT on an orthogonal dataset of disease mutations obtained from ClinVar. |

|  |  |
| --- | --- |
| **Reviewer comment 6** | 6) A major shortcoming is that the software predictions were not  supported by any experimental studies of function or mechanism. This would provide ultimate validation and confidence in the ALoFT approach. Overall, to follow other examples of new bioinformatics approaches that produced findings that have tangible translation and utility, it might have been more compelling to have started from the outset with a dominant and/or recessive disease family with an unknown mutation, test a list of nonsense mutations, use dominant LoF score and inheritance pattern to evaluate the mutations, and use filtered gene (mutation) list and perform functional study to find causative mutation. This would provide the ultimate validation of the robustness and utility of the approach. |
| **Authors response** | Experimental validation is beyond the scope of this work. However, we understand the reviewers point about validating predictions. To show the robustness of the method, we applied the classifier to several known case studies and show that our prediction results agree with published results. |

|  |  |
| --- | --- |
| **Reviewer Comment 7** | 7) The authors discuss that while LoF variants are still observed in  healthy controls, they tend to affect minor isoforms. They then  demonstrate this by showing 12 isoforms of NF2 with premature stop mutations in all 12 isoforms in the HGMD cohort and only 2 premature stop mutations in healthy controls. In the HGMD cohort however, there are multiple premature stop mutations in 'minor isoforms' such as isoform 7, 8, and 10 yet they are still classified to be 'disease-causing'. Further, the authors make this point using only one gene as an exemplar. The authors are encouraged to use multiple genes in support of this isoform trend. |
| **Authors Response** | We provide three reasons that could explain the presence of pLoF variants in known disease genes in healthy controls. One of the three reasons is that some pLoF variants in known disease-causing genes in healthy controls affect isoforms that are different from the isoforms that carry disease-causing pLoF variants. Thus, disease-causing pLoFs and pLoFs in healthy controls occur in mutually exclusive isoforms. The major and minor classification is not related to the length of the isoforms. So we cannot conclude that isoforms 7, 8 and 10 are ‘minor’ isoforms. Perhaps the use of the term ‘minor’ and major isoforms is probably confusing. We have modified the text to make it clearer. We also include Supplementary figure xx that shows the occurrence of 1KGP1 pLoFs and disease-causing pLoFs in mutually exclusive transcripts. However, due to concerns raised by reviewer 2, we have removed the discussion pertaining to NF1. |
| **Change in text** | SB look into making a collage of some examples. |

|  |  |
| --- | --- |
| **Reviewer comment 8** | 8) Dominant gain of function mutations are not included in this study.  Perhaps a brief comment on GoF variants in the overall landscape might  be valuable. |
| **Authors response** | 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. It depends on the biological context of the gene that can vary widely. In order to minimize errors that might arise due to inadequate modeling of GoF effects and focus only on LoF, we chose to only use predicted haploinsufficient genes as the training data for dominant model. While this is clearly mentioned in the Supplementary text, we have moved this detail to the main text. We also revised the manuscript to elaborate a bit more on this point. |
| **Change in text** | Changes made on Pages 3 and 4 (last paragraph of Page 3 continuing into Page 4). |

|  |  |
| --- | --- |
| **Reviewer comment 9** | 9) The Supplemental section is poorly structured. Supplemental  information is merged into a Supplemental Methods file wherein it is  not always immediately clear what methods description and legends are.  Readability would probably improve when each Supplemental Figure/  Table is presented on a separate page. Please, add a list of all used  abbreviations to the Supplemental Data. Abbreviations are not always  explained. |
| **Authors Response** | We have made the requested modifications as per the reviewers suggestion. We have added a list of all the abbreviations as well as presented each Supplemental Figure and Table on separate pages.  [needs to be checked carefully] |

### Reviewer 2:

|  |  |
| --- | --- |
| **Reviewer 2 Comments (General Points)** | |
| **Reviewer Comment 1** | The authors have sought to provide 'real-life' examples of how  ALoFT could be useful. However, their limited expertise in these areas (e.g. cancer, NF2, clinical genomics) has led to some simplistic  assumptions and I would recommend that they seek input from experts in the relevant fields, of which there must be many in their local environment. |
| **Authors Response** | We understand the referee’s take on the complexities associated with relating mutations in NF2 to a phenotype. Our literature review of NF2 mutations indicate that premature truncating mutations in NF2 are associated with severe phenotypes whereas missense mutations are associated with milder phenotypes. However, we agree with the reviewer that such an approach might be rather simplistic. Therefore, we have removed Figure 2b, the NF2 example. Instead we now include Supplementary Figure xx which shows several examples where known disease-causing LoF mutations and pLoFs in the same genes are in mutually exclusive transcripts. |
| **Changes in text** | Removed Figure 2b. Added Supplementary Figure xx instead of Fig 2b to illustrate the point. |

|  |  |
| --- | --- |
| **Reviewer Comment 2** | The use of terminology in the manuscript is loose and sometimes  confusing. Please review this carefully throughout. In particular the  authors should reconsider their use of the term LoF, as they are  evaluating variants that cause premature truncation which can cause  reduction, loss or gain of function. At the very least 'putative  loss-of-function' should be consistently used throughout, including in  the name of the tool. |
| **Authors Response** | We agree with the reviewer that variants that cause premature truncation can do so by LoF, GoF or reduction of function. Therefore, we prefer to label such variants as putative loss-of-function variants (pLoFs) In this revision, we have defined and consistently labeled them as pLoFs. Biesecker et al. have also used this terminology in their paper titled “**Individualized iterative phenotyping for genome-wide analysis of loss-of-function mutations**“ (PMID: 26046366). |
| **Changes in text** | We have consistently used the term putative loss-of-function (pLoF) throughout the manuscript. |

|  |  |
| --- | --- |
| **Reviewer 2 Comments (Major Points)** | |
| **Reviewer comment 1** | The training sets are clearly present in some of the data analyzed  in the results, this should be noted and potential caveats addressed,  e.g. overfitting. As an example, the authors note that ALoFT performs well at distinguishing HGMD disease-causing variants in the last exons of genes, but it is unclear from the methods if any of these were used for training. Some explanation of the overlap between the sets is necessary to provide confidence in the robustness of these analyses. |
| **Authors Response** | We apologize for the confusion. Completely different variants are used for training and the following analysis***. We have added Supplementary table 2 to clarify this issue.*** The model was trained on ***homozygous*** LoF variants in 1KGP1. The analysis was done on ***heterozygous*** LoF variants. For the training sets comprising of disease mutations, only HGMD mutations in genes that could be assigned to either the dominant or the recessive category based on OMIM were used. Of the 1,800 disease genes in HGMD, only 932 genes could be assigned to dominant or recessive category. ***Thus, we only used a subset of HGMD variants for training and used the remaining variants for analysis.***  ***No training variants are included in the analysis pertaining to Figure 3 where we show that ALoFT performs well in distinguishing, disease-causing HGMD mutations. This had been included in the Figure 3 legend.*** |
| **Change in text** | We have added Supplementary table 2 that includes the number of variants that were used for training from the 1KGP1 and HGMD datasets. All analyses exclude variants that were used to train the model. |

|  |  |
| --- | --- |
| **Reviewer comment 2** | HGMD designations of pathogenicity are universally known to not be  robust, and this has been documented many times. Whilst the authors may not have any other datasets to hand, this fact should at least be highlighted, else the unfortunate use of these classifications, as truth sets will be propagated. |
| **Authors Response** | We have made a reference to this by describing confounders due to imperfect training datasets in our initial submission (Please see in Page 6 , the underlined statement). However, we have elaborated and made this clearer in the revised text. |
| **Change in text** | 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 polymorphisms and about 27% of HGMD variants were excluded by Bell et al. in their estimate of carrier burden for severe recessive diseases.  Changes made on Page 6 in the paragraph below the heading “Application to 1KGP1: Estimating the number of pathogenic pLoFs in a healthy genome or understanding pLoFs in an individual genome”. |

|  |  |
| --- | --- |
| **Reviewer comment 3** | The impact of strandedness of gene transcripts on annotation was  not explicitly addressed, but is essential to incorporate. For example  in the sentence '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' are annotation discrepancies due to strandedness included within this? |
| **Authors’ response** | Our annotation algorithm (VAT, vat.gersteinlab.org) takes into account the strandedness of gene transcripts and correctly annotates variants based on their strand direction. So this is not an issue in ALoFT. |

|  |  |
| --- | --- |
| **Reviewer comment 4** | 4. I was unconvinced about the assumptions drawn on isoforms. i.e.  that LoFs in healthy individuals affect minor isoforms. Although as a  trend it seems reasonable, much more data needs to be presented and evaluated before proposing that, for example, some truncating NF2 mutations are not disease-causing. It clearly cannot be the only  reason; e.g. many LoFs in disease genes in healthy individuals cause disease in other individuals, often within the same family. |
| **Authors’ response** | We agree with the reviewer that isoform-specific effect of pLOFs cannot be the only reason for seeing pLoFs in disease genes in both healthy controls and cases. However, in our previous submission, we have included other potential reasons such as incomplete penetrance, compensatory mutations etc. Nonetheless, we understand the reviewers critique in the context of NF2 and have restructured the manuscript. The reviewer also agrees, that “as a trend, it seems reasonable”. To this point, we also had included in the previous submission manuscript that ~12% of pLOFs in known disease genes fall in this category. In this revision, we have now included more examples (Supplementary Figure XX) to show this trend. |
| **Changes in text** | Supplementary Figure XX added to show this trend. |

|  |  |
| --- | --- |
|  | **Reviewer 2 Comments (Other Points)** |
| **Reviewer comment 1** | The focus in the introduction on specific truncating variants in  genes seems partial both in terms of mechanism and disease (all heart disease related). In addition to not being representative of the  overall knowledge base in this area it is at odds with the focus of  the paper, which includes autism and cancer, but not heart disease.  There are cancer examples that could be cited. |
| **Authors’ response** | The intention of the introductory paragraph was to showcase protective pLOFs reported in literature that happen to be cardiovascular-related. However, we have included examples of pLOFs and their relevance in cancer in the new revised version. We also would like to clarify that the focus of the paper is not on autism or cancer. Examples from autism and cancer were used to show that ALoFT predictions corroborate published results. Therefore, the method is robust and can be used in various disease contexts.  (I don’t know if she is referring to protective variants in cancer. Yao, if you know of any examples of LoFs in cancer with protective effects please include ). |
| **Changes in text** | In page 2, the following sentence along with citations to two papers have been included.  pLoF variants are also prioritized in cancer studies where various filtration schemes are used to narrow down causal mutations 15,16. |

|  |  |
| --- | --- |
| **Reviewer comment 2** | 2. Please don't use the term 'premature stop-causing SNPs'. I assume these are stop-gain (nonsense) variants. In fact I was confused as to whether 'premature stop' was being used to collectively describe all variant classes that could lead to premature truncation (i.e. including frames-shifting indels)' or just stop-gain variants (i.e.base substitutions that result in a stop codon) or both at different times in the paper. Please make this clear and define and then consistently use terms. The authors used: nonsense, premature stop variant, premature stop mutation, premature stop-causing SNPs,premature stop alleles and premature stop codon in the manuscript!!. |
| **Authors’ response** | SNPs that can introduce a premature Stop codon, indels that result in framshifts, and variants that affect splice sites have been collectively referred to as LoF variants. Some have correctly argued that the molecular functional effect of such variants need not always be LoF. Hence they are also referred to as premature truncating variants. To make it clear and consistent throughout the manuscript, we have modified the manuscript to refer to them as pLoFs (putative loss-of-function variants). |
| **Changes in text** |  |

|  |  |
| --- | --- |
| **Reviewer comment 3** | In turn this meant I was unclear whether some of the analyses were  restricted to stop-gain variants or also included the putative  loss-of-function variants due to indels. This needs to be clear. I  think it is essential that the tool is fully evaluated on indels as  well as stop-gain variants. |
| **Authors’ response** | The training model was built using SNPs that introduce a premature Stop codon as indel calling methods are not robust a. However, ALoFT works on indels and we have updated the text to make this clear. We have also included the performance of ALoFT on indels in this revised manuscript. |
| **Changes in text** | Yao: Since you have already run the modified ALoFT ( I think I sent this to you) on HGMD, please just add one or two sentences, saying that xx% of HGMD indels are predicted to be pathogenic. You can calculate a PPV for indels based on HGMD indel predictions. Is this clear? |

|  |  |
| --- | --- |
| **Reviewer comment 4** | I think the cancer driver analysis part is weak. A version of this  strategy is already incorporated as standard in cancer driver  discovery pipelines. I would advise omitting this, or at least  checking with some of the major cancer institutes whether they would use ALoFT in that context. |
| **Authors’ response** | ALoFT discriminates disease-causing somatic pLOFs from benign somatic LoFs. We applied ALoFT to a variety of diseases to illustrate the fact that the method works well and produces meaningful results. (Not sure what she means by a version of this is already incorporated as standard). |
| **Changes in text** | Yao: It might be useful to include how many somatic pLoFs are there in the Alexandrov set, what percentage of these are predicted to be pathogenic by ALoFT and to show that even in cancer there are many many benign pLoFs. Please modify this section to make it stronger. |

|  |  |
| --- | --- |
| **Reviewer comment 5a** | 5a. The authors have a tendency to be simplistic in places: For  example: the statement 'premature stop codons in the last exon are not subject to NMD', is an oversimplification. (???) |
| **Authors’ response** | I am unsure why the reviewer is stating that this is an oversimplification. |
| **Changes in text** |  |

|  |  |
| --- | --- |
| **Reviewer comment 5b** | 5b. And, "The effect of such variants is the production of truncated proteins that are functional." needs to be tempered as it seems to state unequivocally that all variants in the last exon or after the last reported disease-causing variant will produce a functional truncated product, which is not true. |
| **Authors’ response** | We agree with the reviewer on the simplistic verbiage and have modified the text appropriately. |
| **Changes in text** | The effect of such variants could be the production of truncated proteins that are likely sufficiently functional. |

|  |  |
| --- | --- |
| **Reviewer comment 5c** | 5c. Similarly, the authors mention that prediction at the clinical level  is confounded by issues such as penetrance, expressivity but then say. 'Therefore, we do not expect to observe any LoF variant in NF2 in the presumed healthy individuals'. In fact healthy individuals can have NF2 LoFs, particularly if sampled at younger ages, as many of the features are non-specific and have insidious onset, only about half get tumors, usually benign ones.  'While the occurrence of LoF variants in known disease genes in  healthy individuals might be surprising' - It has been known and  extensively documented for over 20 years that healthy individuals can carry LoFs in disease genes!!!. It may be unexplained but it should not be surprising! |
| **Authors’ response** | We agree with the reviewer about NF2 and have removed this example to illustrate the point that pLoFs in disease genes that occur in presumed healthy individuals are isoform-specific. Instead, we include many examples in Supplementary figure xx which shows this trend.  We observed healthy individuals can carry LoFs in known disease genes. Besides, issues like penetratrance, we show that isoform-specific truncation may be another reason for this observation. We understand this is not the only explanation, as there are many other factors, stated in the manuscript: incomplete penetrance of disease alleles, variable expressivity, compensatory mutations, marginal variant calls and etc.. |
| **Changes in text** |  |

|  |  |
| --- | --- |
|  | **Reviewer 2 Comments (Figures)** |
| **Reviewer comment 1** | 1. Figure 2b - the clinical used transcript of NF2 is defined (RefSeq  NM\_000268.3) and should be identified as such, as this most likely  corresponds to those reported in HGMD and displayed in the figure. |
| **Authors’ response** | We have removed this Figure now. Please see Response to Reviewer 2, comment 5c. |
| **Changes in text** |  |

|  |  |
| --- | --- |
| **Reviewer comment 2** | 2. Figure 3a - this figure and accompanying description in the text  essentially show that HGMD and exome datasets will differ in reporting variants at ends of genes; surely this is expected due to the very different nature of how these datasets are collated? |
| **Authors’ response** | Yao/Mark: I am not sure what the issue is? The reviewer’s comment is unclear to me. |
| **Changes in text** |  |

|  |  |
| --- | --- |
| **Reviewer comment 3** | 3. Figure 4a - the contrast between the dominant score and the GERP and CADD scores is striking. As a minor point, it would be good to note in either the legend or figure something about the  recessive/tolerated score. Did the recessive variants also have a low  tolerated and thus high recessive score, as one might automatically  assume? |
| **Authors’ response** | Yes, the recessive variants also have a low benign score. We have added in the legend that the recessive variants also have a low tolerated score. |
| **Changes in text** | Change made in the Figure 4a legend. |

|  |  |
| --- | --- |
| **Reviewer comment 4** | 4. Figure 4c - minor point, is the labelling of the y-axis correct?  I.e. 0.14% of somatic variants with high disease-causing score were in  known cancer genes? Or should the range be from 0 to 14? |
| **Authors’ response** | We thank the reviewer for pointing out this error. We have made the appropriate correction. |
| **Changes in text** | Correction made. |

|  |  |
| --- | --- |
|  | **Reviewer 2 Comments (Methods)** |
| **Reviewer comment 1** | 1. Return of near\_start, near\_end defined as within first/last 5% of  gene - is this appropriate? Does it lend undue weight to an arbitrary  value? If not an arbitrary value, please provide justification for  selection of 5%. |
| **Authors’ response** | The 5% value is chosen as it has been shown in MacArthur et al that there is an enrichment of pLoFs at either end of the genes. We have also seen this trend in 1KGP1 as well as ESP6500 and ExAC datasets. However, given that these are simplistic filters, we wish to bring to the attention of the reviewer that we do not use this as a feature in our prediction algorithm. This is part of the annotation pipeline and is only an annotation flag that provides the user information about the pLoF. |
| **Changes in text** |  |

### Reviewer 3:

|  |  |
| --- | --- |
|  | **Reviewer 3 Comments** |
| **Reviewer comment 1** | I have a number of technical concerns. The most important concern is that this works attempts to solve two problems at once: 1) predict whether LoF variants in a given gene are pathogenic, and 2) predict whether a particular putative LoF mutation is truly a LoF. These two problems require differentapproaches to development and validation of a classifier. For the combined problem, the proposed classifier likely results in overfitting due to simple labeling. |
| **Authors’ Response** | We understand the reviewer’s criticism about making the distinction between molecular loss of function from its effect on phenotype. However, ALoFT has been developed to address point 1: inferring the pathogenic effect of pLOF variants. While elucidating effects at a molecular level is feasible to some extent, as we have noted above, understanding the exact mechanism can be tricky (LoF, GoF, change of function etc). Moreover, in practical terms we are ultimately interested in understanding the effect of genotype on phenotype. Therefore, to date, all prediction programs (there are numerous such programs for missense variants) developed to understand genetic variations have been trained on benign versus disease models. **ALoFT elucidates the pathogenic potential of pLoF variants.** ALoFT is similar to missense variant prediction programs where variants are classified as benign or pathogenic, but not whether it leads to LoF, GoF etc. Compared to missense variants, the likelihood that pLOF variants will affect the phenotype by molecular loss-of-function is much higher. Moreover, we also maximized the probability of modeling LoF effects by choosing variants in haploinsufficient genes as a model for the dominant class. This will minimize errors that might arise due to inclusion of gain-of-function mutations in the training set. However, we don’t believe it is essential to model molecular LoF. Rather we are more interested in understanding the effect of pLoF variant on the phenotype. To this end, we also trained our model based on variants in genes known to cause disease due to dominant mode of inheritance. We show that ALoFT has good predictive power (AUC = 0.934) , albeit a bit lower than the classifier trained on haploinsufficient genes (please check Supplementary table 3). |

Minor Comments not addressed yet

Minor comments

1. Figure 1 shows the ALoFT pipeline, but does not specify what the

output data is. Please specify the "output" in more detail; i.e. the

authors could certainly mention the LoF scores in the Figure 1 output.

The authors should also discuss which LoF scores are indicators for

pathogenicity, as mentioned above

2. References 6 and 7 are letters to the editor that can be deleted.

Note that a treatment to knock-down APOC3 has been reported - see

PMID: 25470695, which could be cited to parallel the structure of the

discussion of PCSK9 LoF variants.

3. Please ensure that all gene annotations are in italics.

4. Supplementary Tables 4 and 7 seem absent from the Supplement.

5. Page 1 (space): ....Finnish population in LPA that protect against

coronary > ....Finnish population in LPA that protect against coronary

6. Please check for consistency in the manuscript and figures that you

refer to premature stop mutations (no capital S in stop).

7. Figure 1: transcript-specifc > transcript specific

8. Figure 1, legend: The authors mention that ALoFT can also be used

to annotate a 5-column tab delimited file. I wonder whether the

position needs to be according to a specific genome browser version?

9. "Current estimates of genetic burden of disease alleles in an

individual vary widely, ranging from 1.1. recessive alleles per

individual to 31 deleterious alleles33-37. It should be noted that the

prediction 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." Please, note that the whole genome/exome

has not been sequenced in all referenced studies, which further

impacts the predicted number of deleterious alleles.

10. Figure 2a: This figure and its legend are somewhat confusing at

first glance.

- Please, explain what Benign LoFs are in the legend.

- Please rename 1KGP1 LoFs (other) > 1KGP1 LoFs in non-HGMD genes

- Please rename 1KGP1 LoFs in Genes > 1KGP1 LoFs in HGMD genes

- Please rename HGMD LoFs in Genes > Genes with LoFs in 1KGP1 and HGMD

- Please rename HGMD LoFs (other) > Genes with HGMD LoFs only

11. Page 3: "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 the production of truncated proteins that

are functional" > suggest > 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 the production of truncated

proteins that are sufficiently functional.

12. Figure 4a: Please include common variants into the graphs as references.

13. Page 5: "When multiple cancer genomes are not available, somatic

LoFs that are predicted to be disease-causing by ALoFT can be used to

identify potential tumor suppressors." Can the authors indicate what

ALoFT score would indicate that a somatic mutation is (likely) a tumor

suppressor?

14. Supplement 1.1. Functional features: "The 3D structure of the

protein is essential for proper folding and function of proteins." >

The 3D structure of a protein is essential for proper folding and

function of proteins.

15. Supplement 1.1. "annotated based on data from PhosphositePlus 3" >

annotated based on data from PhosphositePlus3

16. "For all functional features, we assessed 1. Does the premature

stop variant affect a functional feature? 2. Are..." > For all

functional features, we addressed the following questions: 1) Does the

premature stop variant affect a functional feature? and 2) Are...

17. "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." Please, provide a reference to support the choice of 50

bp.

18. "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." This Supplemental 1.1. section is about "Functional

features" and then ends with this NMD sentence. Why is this sentence

incorporated in a section about functional features? The rest of the

section discusses disruptions of protein domains.

19. Supplement 1.4 Mismapping errors. Please, provide the reader with

insights into how many 1) segmentally duplicated regions there are in

the genome?; 2) how many genes have paralogs?; and 3) how many genes

have pseudogenes?

20. Supplement 1.5 Annotation errors. Generally the authors poorly

support based on what they annotate as parameters in a-f to

distinguish promising from false positive LoFs. Please, provide

explanations/references in this section to help the reader understand

why you have taken these cutoffs.

21. Supplement 1.5.a. "lof\_anc: Indicates that the LoF variant allele

is the same as the ancestral allele and is likely to be a functional

allele." I do not completely understand what you mean here. Is the LoF

in this case the stop codon that is normally used?! Please, explain

more clearly.

22. In Section 2. Pathogenicity prediction for LoF mutations, you

could refer to section 2.1. when you mention that "benign variants

were detected from 1KGP1". That said, I do wonder about the occurrence

of these variants in ExAC? What is the frequency of benign variants?

How sure can one be that these variants are truly benign?

23. In section 2 points d and e seem to have different spacing of sentences.

24. Section 2 (and Supplementary Table 3): "In total, we used 106

features to train our model.

(<http://aloft.gersteinlab.org/features/#prediction_features>)". Suggest

that the excel file of these 106 features be included as a

supplementary excel document. The reader should have access to this

information through Nature Genetics.

25. Section 2.2. Three-class classification. "The average number of

dominant mutations per gene is 20." What is this value of 20 based on?

Please, provide an explanation or reference. Also, I wonder whether it

is a statistically validated method to pick 3 variants per gene for

the dominant class?

26. Supplementary Figure 2: Please, add in Supplementary Figure 2 at

the Y-axis that this is the Precision score, and explain more clearly

what you define as a 'true positive' and a 'false positive'.

27. Supplementary Table 2: Please, explain the reasoning for removing

olfactory receptors, randomly picking transcripts and analysis of all

dominant genes in the Table legend.

28. Supplementary Table 3: 7. > 7

29. Section 2.3.1. Applied to known disease-causing mutations from the

Center for Mendelian Genomics studies

(<http://data.mendelian.org/CMG/>). Why did the authors use CMG data and

not much larger databases such as HGMD or LOVD to support their ALoFT

zygosity claims?

30. Supplementary Figure 4: Distribution of predicted dominant and

recessive premature stop alleles in the 1KGP1 individuals. The authors

say that they made their calculations based on 246 individuals of

African ancestry and 379 individuals of European ancestry. Together

these numbers represent only 625 genomes of the 1000. Why did the

authors not include more individuals?

31. Supplementary Table 6, legend: "de novo" is not consistently

displayed in italics.