Peer Review File

Article Information: http://dx.doi.org/10.21037/tlcr-20-1197

Reviewer comments:

This is a well-designed extreme phenotype study in heavy smoker individuals who develop lung adenocarcinoma at an early age (extreme cases, n=50) or did not present lung cancer (extreme controls, n=50).
The authors performed WES and identified 619 different variants between cases and controls and validated 107 in a second cohort (TCGA). They conclude that this strategy may help to identify high-risk subjects.
The methods and results section can be improved, and a gene burden analysis should be performed in genes with several rare variants to target the most relevant genes.

We thank the reviewer for the comments. We provide our point-by-point responses and the changes performed below, which we think have substantially improved our manuscript. The changes can be revised in the tracked changes version of the manuscript that we submit.

- *Comment #1:* ***Abstract****. Please, detail TCGA*

  Reply #1: TCGA has been detailed in the abstract, as well as the characteristics of the studied individuals (pg. 4, lines 101-102).

- ***Methods:***
- *Comment #2: The bioinformatic analyses describe the primary analysis, but the prioritization of variants is poorly described. This reviewer assumes that the authors are selecting rare variants (SNV and indels), but no MAF is indicated in the method section, and no pathogenicity criteria (ACMG, CADD score) are reported. This is considered a standard procedure to report rare variants.*

  Reply #2: We thank the reviewer for the comment. We selected nonsynonymous variants located in exonic regions and/or those in the intronic splice site flanking regions. Variant priorization was performed according to the comparison between MAF in cases and controls, not that of the general population. We did not restrict our analysis based on ACMG/ clinical variant classifiers since we do not expect the variants to be either pathogenic or benign (or VOUS) but rather variants associated with the selected phenotypes. We have revised and clarified this information in the methods (pages 9 and 10, lines 214-218). We have also included in Supplementary table 2 the MAF values corresponding to the European (Non-Finnish) ancestry, the ACMG classification and the CADD raw scores, because we agree that it is relevant information.

- *Comment #3: Please, detail if you have used reference datasets from gnomAD or CSVS to compare the allelic frequencies in your extreme cases or controls with Non-Finnish European or Spanish reference populations. The study will improve if you present a Table with MAF in reference datasets.*

Reply #3: for the selection of variants of interest we considered the p-value associated with the extreme cases and controls MAF comparison, rather than the general population MAF data. Nevertheless, as commented before, we have included the allele frequency data form ExAC_NFE in Supplementary Table 2, given that it is the database with higher number of sequenced alleles and therefore, the most representative for genotypes of European ancestry. We are willing to add additional data, if it is deemed necessary.

- ***Results***
- *Comment #4: In general, a better organization of the main findings is recommended. The authors present data for missense variants, but there is no description on LoF variants (nonsense, novel splice, startloss, stopgain). I also do not find the finding for indels. A comparison of LoF or misssense variants with the synonymous variants found in extreme cases should be also presented to support the pathogenic effect of missense/LoF.*

  Reply #4: we thank the reviewer for this comment. Among the selected variants, we identified 586 nonsynonymous SNV, 7 frameshift deletions, 7 frameshift insertions, 3 nonframeshift deletions, 10 nonframeshift insertions and 6 stop gains. We have included this information in a new column in Supplementary Table 2.

- *Comment #5:* ***Table 1*** *should be improved. No MAF data are shown in extreme cases and no reference data are shown for NFE or Spanish. The criteria for pathogenicity (ACMG) should be included, as the CADD score for each variant.*

  Reply #5: We have now included all this information in Supplementary table 2. In order to improve the clarity, we think that it is better not to include it again in Table 1 since it is does not add relevant information (all the variants represented in Table 1 are nonsynonymous and are benign or likely benign).

- *Comment #6:* ***Supplementary table 3*** *shows genes with several variants. There are 22 genes with at least 3 variants. A gene burden analysis could be performed for these genes by comparing with extreme controls, gnomAD and CSVS.*

  Reply #6: We thank the reviewer for this extremely interesting suggestion. We have used REBET: The subREgion-based BurdEn Test, which is implemented in the R package to calculate the p-value of the extreme case/control comparison at the gene level. We have included this new information in the Methods (pg. 10, lines 221-223), in the Results (pg. 12, lines 268-271) and in Supplementary Table 3.

- *Comment #7: Discussion is too large and it should be reduced to methodological aspects of the design and main findings.*

  Reply #7: We agree with the reviewer. The discussion has been shortened, following his/her suggestion (several paragraphs).

- *Comment #8: Figure 2 legend should be revised: "…variants differentially expressed between individuals presenting extreme phenotype…"There are no gene expression data. This is a list of rare variants found in extreme cases and controls*

  Reply #8: We thank the reviewer for the thoughtfully revision of the text. We fully agree and we have revised the legend accordingly (pg. 29, lines 753-754).

- *Comment #9:* ***Suggested References:***

  Karczewski, K.J., Francioli, L.C., Tiao, G. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020). https://doi.org/10.1038/s41586-020-2308-7

  Peña-Chilet M, Roldán G, Perez-Florido J, et al. CSVS, a crowdsourcing database of the Spanish population genetic variability. Nucleic Acids Res. 2020 Sep 29:gkaa794. doi: 10.1093/nar/gkaa794 Epub ahead of print. PMID: 32990755.

  Reply #9: We have read the interesting paper of *Karczewski et al,* but we prefer not to include it, since we did not classify the variants according to their functionality. Regarding the paper by Peña-Chilet et al, we´re very aware of the results, since part of them have been generated by our colleagues and close collaborators from Hospital of Navarra. We are also very aware of the need of local genetic databases and the additional value that they add to the interpretation of the genetic variants. Consequently, we have cited the manuscript in the discussion (pgs. 18-19, lines 486-502, reference 34).