Reviewer's report

Title: RNAseq analysis of bronchial epithelial cells to identify COPD associated genes and SNPs

Version: 0 Date: 02 Jan 2018

Reviewer: Ian Hall

Reviewer's report:

This paper describes an attempt to identify COPD risk genetic variants through expression based approaches, most notably by taking advantage of DAE to try and identify cis acting risk SNPs which control gene expression and hence could contribute to development of COPD. This is an interesting approach although there are a number of key issues which need to be considered before concluding that risk alleles have been identified. Ideally of course this kind of study requires replication in an independent dataset: some attempt to do this has been made here using publically available GTex datasets although the approach is different from that used in the discovery analyses. The paper is generally reasonably well written, but does not always flow logically, in that the reason for undertaking some of the analyses is not clearly explained and the way in which sample sizes are described also leads to some confusion. the literature on GWAS discovery for COPD/lung function risk variants is also not fully cited although to be fair this is not the main thrust of the paper.

The following points need consideration.

1 The initial hypothesis as I understand it is that one could identify novel genetic risk factors, but by preselecting genes for the RNA seq analyses this design does not allow a true discovery approach. In practice 35 genes were selected: the main paper does not say what criteria were used, and describes these as 'COPD associated' genes. This needs to be explained in the main paper....the detail is actually in the supplementary material but the actual selection appears to be more due the fact that some of these genes control key pathways which are believed to be important in COPD rather than anything to do with genetic association. The risk of this approach is of course the same as in the candidate gene studies done 15 years ago, that one is preselecting genes to follow up based on their potential role in disease and hence the ability to identify truly novel genetic markers is limited.

2 When assessing gene expression between COPD and controls, there are obvious potential differences other than disease status to be taken into account. The table shows that smoking status was reasonably matched (although not for pack years): it would also be helpful to know if any subjects were on medication at the time of bronchoscopy.
3 Some key methodological details are missing: for example, no actual data are given on RNA quality, which is critical for the RNA seq analysis (the paper just cites a reference saying that these criteria were used). RIN scores etc should be added.

4 I was not clear why the SLDA approach was used to select genes for further study.

5 The most confusing element of the study for me as presented at present is the key data shown in Table 3. The original concept was to use DAE to look for allele driven expression effects. There were 30 cases and 30 controls, although not all BEC samples were adequate for RNA seq. This of course means that for any relatively low frequency minor alleles there will only be a small number of informative individuals (ie heterozygotes). However the table contains 'n' values of 128-159. Where do these come from? They can't be numbers of individuals based on my understanding of what has been done. How were these numbers then used in the statistical approach used? As this Table contains the key findings of the study it really needs to be clear what has actually been done.

6 Further analyses were then carried out in populations where bronchitis and emphysema predominant phenotypes had been identified. It isn't clear if these were pre-specified analyses, and where results are reported (for example for ERCC5 on chronic bronchitis predominant COPD) it isn't clear what the comparison group is. If these were post-hoc analyses they should be reported as such and any conclusions drawn from these treated accordingly.

7 Going back to the original study design, it would be helpful for the reader to have a feel for the numbers of SNPs which were within the regions sequenced. Unless I misunderstood the approach, I presume the hypothesis here is that if a SNP is identified by RNA seq and shows allele specific association with expression and disease status, this effect might be driven by a regulatory region SNP with which the coding SNP is in tight linkage disequilibrium. Only single SNPs are reported for each gene in the Tables, but I suspect there may be other SNPs in these genes as well. Assuming that is the case, how many SNPs were assessed and how were multiple testing issues resolved?
Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

No

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

Yes

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

Unable to assess

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

Quality of written English
Please indicate the quality of language in the manuscript:

Acceptable

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?
If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

none direct

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal