Reviewer’s report

Title: Distinct signatures of lung cancer types: aberrant mucin O-glycosylation and compromised immune response

Version: 0 Date: 01 Feb 2019

Reviewer: Aleksandar Dakic

Reviewer’s report:

Comments on "Different molecular signatures in lung cancer types from integrative bioinformatic analyses of RNASeq data"

This manuscript reports on a gene-expression signature aimed at discriminating between two most common forms of non small-cell lung carcinoma - adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC). Along with this main aim authors attempt to identify key pathophysiological molecular processes differentiating the two diseases. They do this by using and comparing three standard bioinformatics pipelines, and TCGA + GTEx data integrated in three different, partially overlapping ways. As such there is some need to consolidate and refocus the manuscript according to the importance of these different aims (bioinformatics methods and data curation comparison versus biological discovery).

Here are my comments ordered by sections of the manuscript. In addition, writing often seems as if it is in a draft version, so it needs to be edited further. I list some errors or language ambiguities at the end:

Pages are not numbered so I will comment in terms of numbering of the pdf file I received.

Section 2 - Materials and Methods

Currently, the GitHub link leads to no page…

Pg. 5 ln 44 - it is possible to include experimental design factors in limma too

Pg. 5 ln 49 - should be reworded to something like "Other sources of batch effects … have not been corrected for … as they did not affect the final conclusion (data not shown or Supplementary data or GitHub)"

Pg. 6 ln 5 replace "estimated" with "summarised" or similar!

Pg. 6 ln 12 - It is not quite clear how authors constructed unified GTEx + TCGA dataset and what is meant by "batch-free Recount2 protocol" - Recount2 acknowledges the need for batch correction - to cite them: "Although all recount2 samples have been processed and summarized
with a single pipeline, so-called 'batch' effects could occur and should be considered in downstream analyses, particularly when comparing among studies."

Pg. 6 In 30 - "see 2.1" is meant? / Pg. 6 In 33 "see 2.2" is meant?

Section 3 - Results

Section 3.1 - Overall, on the first reading it is not quite clear at all what is compared with what - this needs more clarity. From page 8, lines 50-56 it seems as if there are three different tumor vs. normal tissue data sets for each cancer - but they are often subsets of each other! It needs to clarified at the beginning that samples in "paired" data sets are fully contained in "all" data sets when it comes to both tumor and normal tissue samples. It also needs to clarified at the beginning that normal tissue samples in "paired" datasets are fully contained in "unpaired" data sets; and that "unpaired" and "all" datasets re-use the same normal tissue samples. This raises a question if it is sensible to report on both "unpaired" and "all" analyses separately - they share all normal tissue samples and 90% of tumor samples in case of both LUAD and LUSC (calculated from Table S1) - nothing particularly different can be expected there. Aim stated on page 9, line 5 - "to remove artefacts due to partially paired datasets" is not pursued further in this study if I understand well, so "unpaired" analysis seems like an unnecessary complication sacrificing clarity. If biological discovery is of interest, then partially matched "all" dataset would be an obvious second choice (as preferred "paired" dataset is too small for signature discovery).

Pg. 8 In 39-40 - in addition to removing discordant and low tumor purity samples, you also removed "outlier" samples with correlation of less than 0.6 with… any other sample? (Methods, Section 2.1) - so this should be added in tis sentence.

A clear sentence or two are needed here and in section 3.3 to explain how first LUAD vs. normal and LUSC vs. normal DEA are performed separately - then genes that change in the same direction in both cancer types removed from the lists of DE genes to form a new, cancer-type-specific lists of DE genes. Sentence on page 10, line 58 to page 11, line 2 trying to do this job is confusing.

Section 3.2 - The section title is ambiguous and the first part of it potentially misleading. As mentioned above, "unpaired" and "all" datasets share all normal tissue samples and 90% of cancer samples; "paired" and "all" datasets share 45%-68% of normal tissue samples and 10% of cancer samples. Of course the results will be mildly different. The similar statement in the Conclusion on page 15, In 59 needs to be moderated or deleted.

I recommend to substantially trim this section to relevant information. Some of it could be mentioned, but it is best to keep these analyses of different statistical approaches and different datasets to dedicated method-comparison papers and to simulated or well-known data sets with ground truth - so that something substantial can be concluded. Also, comparing a method without batch correction and patient-specific effect correction ("edgeR-TCGAb") with the one that includes them ("edgeR") does not deserve that much attention - of course there is going to be difference with such a wide range of potential confounding factors.
Section 3.4 - The soft clustering plots for LUSC and LUAD should be placed in the same figure, not separately in a figure and a supplementary figure, because the purpose of the work is to compare the two cancers.

Pg. 11, Ln 46 - It would be more clear to say "We extracted the genes that show upregulation trajectory across stages in one cancer type and downregulation trajectory in the other …"

Sections 3.6 and Figures 4 and 5 - Co-expression signatures: It seems separate analyses have been run for LUSC and LUAD gene expression data - this is bound to identify somewhat different groups of genes, which are then difficult to compare and leads to increasingly convoluted and confusing explanation of results. Why not perform the joint analysis with sample group annotations included - this way the differential usage of the same set of modules could be compared directly, highlighting which modules are more or less active in the two cancer types and highlighting key differential genes within them. This functionality is implemented in CEMiTool (Ref. 15)

My understanding is that your analysis did not adjust for the tumor purity, apart from removing tumors of purity less than 60%. With this in mind it needs to be acknowledged in section 3.3 and 3.7, when interpreting the putative involvement of the innate immune and complement system, that it is difficult to determine whether different immune infiltration is due to intrinsic cancer characteristic or different cancer cell purity due to tissue extraction. Additional difficulty is of course comparing both cancers to normal lung tissue first. As a consequence, all immune involvements need to be interpreted with caution.

Section 3.7 - For clarity, the reference to Table 2 should be placed where the process of narrowing down a list of candidate genes is described, not at the end of the section - definitively before the mention of Table S3.

Pg. 13, Ln 43 - narrowing down the subsets of LUAD- and LUSC-specifically mis-regulated genes to those belonging to "modules with a functional signatures" limits the further analysis to only functionally annotated genes and potentially prevents a novel discovery - is this really intended, particularly when the discriminating signature is the end point?

Section 3.8 - The statement "ITAG6, HABP2, FABP5 and RND3 are predictive for overall survival based on either FDR and/or p-values" is vague and insufficient. Along with p-values (their level is not specified), attention should be on the size of the effect i.e. the interpretation of the "group_low" coefficient in terms of survival. In addition, the two excel sheets of Table S4 contain the same values, only different headings. The "group_low" coefficients for genes ITAG6, HABP2, FABP5 and RND3 do not have FDR or p-values close to 0.05 or any similar threshold.

Section 3.9 - Figure 7 (or a supplementary figure) should first show how the chosen set of candidate signature genes performs in classifying LUAD and LUSC on training (working) data sets. The first branch of the dendrogram should be separating them well. In Figure 7, panel A, clustering with a test (independent) data set and 4 core genes does not do such a good job (LUAD and LUSC samples are still intermixed at much lower dendrogram levels) so the statements in the text should be moderated. Can authors present the figure with the entire set of
candidate signature genes? Do panel A and B of the Figure 7 present two different test data sets, from Ref. 79 and Ref. 80, respectively? If so this should be specified in text and/or the figure legend. If not, how and why are 14 samples in the panel B preselected? I assume that authors do not want to go into formal assessment of the signature's classification performance (sensitivity, specificity, precision, accuracy), because the original aim was less ambitious. But then can we call this a signature?

Figures:

Figure 1. Perhaps more thought could be put into what to include in Figure 1 - why select only upregulated genes in LUAD and downregulated genes in LUSC and not show complete picture in more focused format? Again, nothing new can be gained by including "unpaired" analysis as it shares 100% normal tissue samples and 90% of tumour samples with the "all" analysis. Then for instance comparing limma, edgeR and edgeR-TCGAb could go to a supplementary figure - this is not the primary goal of this paper.

Figure 2. This is perfect material for a supplementary figure, should not be in the main text.

Figure 3. Include clustering plots for both LUAD and LUSC next to each other.

Figure 7. Show panel A with the full set of available signature genes like in the panel B - either in supplementary or original figure.

Table S1. The breakdown of numbers of samples per analysis should be placed in the main text or a small table there. Consider excluding "unpaired" analysis altogether.

Writing:

- Pg. 4 ln 26. "an increased availability of matching health tissue samples" is meant?

- Pg. 5 ln 14. replace "samples retained ... have been pruned by 19 'discordant LUSC' samples" with "pruned from" or reword.. "removed from the analysis" etc.; Similar issue on Pg. 10, ln 58.

- Pg. 9 ln 13 remove one "either"

- Pg. 10 ln 43 "to a total" is meant?

- Pg. 7 ln 43 - "recently developed"? / ln 46 - "in-built" or similar correction?

- Lot more corrections for Discussion and Conclusion sections, here are some:

- Pg. 15 ln 2 - "have already been…”; ln 9 - "contribute”; ln 11 - either "more often" or "show tendencies for" are sufficient; ln 20 - delete "also”; ln 33 - immuno-compromised";
In 35 - "recent scenario"?; In 39 - delete "which"; In 43 - "more specifically deregulated"?; In 56 - delete "allowed to"; delete "elusive" - they are common forms of lung cancer.

- Pg. 15 ln 59 - as above this conclusion is based on partially overlapping datasets, so it is potentially misleading.

- Pg. 16 ln 7 - meaning not clear; Overall, the last few paragraphs need rework.

**Are the methods appropriate and well described?**
If not, please specify what is required in your comments to the authors.

Yes

**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.

No

**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:

Needs some language corrections before being published

**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.

I declare that I have no competing interests.

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