Author’s response to reviews

Title: Integrative analysis of genetic and epigenetic profiling of lung squamous cell carcinoma (LSCC) patients to identify smoking level relevant biomarkers

Authors:

Bidong Ma (mabidong1234@126.com)
Zhiyou Huang (335416169@qq.com)
Qian Wang (wangq@tianjiagenomes.com)
Jizhou Zhang (380372496@qq.com)
Bin Zhou (13736332220@163.com)
Jiaohong Wu (mazheng0817@163.com)

Version: 1 Date: 24 Aug 2019

Author’s response to reviews:

Current date: 08/24, 2019
Manuscript ID: BIDM-D-19-00043R1
Title: Integrative analysis of genetic and epigenetic profiling of lung squamous cell carcinoma (LSCC) patients to identify smoking level relevant biomarkers

Dear editor:

Thank you so much for your help with our manuscript, entitled “Integrative analysis of genetic and epigenetic profiling of lung squamous cell carcinoma (LSCC) patients to identify smoking level relevant biomarkers” by Bidong Ma, Zhiyou Huang, Qian Wang, Jizhou Zhang, Bin Zhou, Jiaohong Wu. We greatly appreciate the reviewers’ comments and suggestions. We have revised the manuscript according to the comments of the reviewers. The comments are valuable and very helpful for revising and improving our manuscript.
Our point-by-point response to the reviewers’ comments is appended below. Revised portion are marked by “Track Changes” option in the manuscript.

We are hopeful that our manuscript is suitable for publication in BioData Mining.

Thank you so much for your consideration.

Sincerely yours,

Jiaohong Wu,
Wenzhou Medical University affiliated People's Hospital

Reviewer #1:

Comments to respond:

1. Add some recent review about correlation of smoke, air pollution and molecular profiles of the patients with the risk of lung cancer i.e "The health and social implications of household air pollution and respiratory diseases". Simkovich et al., 2019, "Tobacco biomarkers and genetic/epigenetic analysis to investigate ethnic/racial differences in lung cancer risk among smokers". Musphy et al., 2018

   Thank you for your suggestion. We have added these recent reviews about the correlation of smoke, air pollution and lung cancer patient’s molecular profiles into our revised manuscript background section, on line 59 of the manuscript.

2. In the results sections, explain what all the acronimous stand for i.e : DEGs=Differentially expressed genes?

   I am very appreciated your suggestions. Yes, DEGs is differentially expressed genes, and we have defined it on line 111, where it is the first time presented in this manuscript.

3. Remove the GOTERMS from the text to make the results section more easy to be read

   Thank you for your suggestion. We have followed your suggestion and removed all GOterms according to make the results easier to read.
4. It is difficult to understand the difference between figure 1 and figure 2. Do both of them show upregulated genes in high smoking patients? Explain better in the legend. In general, use bigger fonts in the figures or add tables to see in details the lists of genes. We greatly appreciated your suggestions. To answer your first question, there are several label errors in Fig 1 and 2. We have redrawn these figures, using bigger fonts. For your second question, we have added S1A, S1B, S2A, S2B Tables to show all listed genes.

5. Page 10, line 17, while describing the genes belonging to the Hippo pathway, add the recent review about the importance of YAP and TAZ in lung carcinogenesis "YAP and TAZ in lung cancer: oncogenic role and clinical targeting" Lo Sardo et al., 2018. This review supports also the finding showed in figure 5A, (the strong overexpression of TAZ in high smokers vs low smoking patients). In the discussion section TAZ is defined as tafazzin. Are the authors sure that it is tafazzin and not the hippo pathway transcriptional coactivator (TAZ/WWTR?) Thank you for your suggestion. We have added this review into our manuscript on line 233 to describe the important of YAP1 and WWTR1 in lung carcinogenesis. For your second question, yes the abbreviation of TAZ is “WW Domain Containing Transcription Regulator 1”, not tafazzin. We have corrected it in the manuscript, and change the alias to the HGNC gene symbol.

6. In page 12, it is not clear why AIRE, SLC6A3 and PENK genes were selected. Authors should explain clearly on their basis of what rationale they were selected. In the discussion section, add some references about the biological role of these genes in lung cancer. We are appreciated by reviewer’s concern and suggestion. For clarification, we added explain about why we chose AIRE, SLC6A3 and PENK genes and showed their DNA methylation pattern in the analysis. In addition, we have added several references about biological roles of these genes in human lung cancer according in the discussion section. The reasons why we chose AIRE, SLC6A3 and PENK are on line 283, and the references are on line 309 to 344.

7. Finally, in order to give stronger biological readout of this in descriptive analyses based on published datasets, authors should show at least one experiment that validates some of the genes that are differentially regulated in high and low smoking patients. For example, they can analyze the expression of the proteins encoded by those genes in patient. They can also use already deposited data of immunoistochemistry, if present. More appreciated would be some functional expriment in lung cancer cell lines where the expression of a pair (ore one) of those interesting genes is manipulated in order to see the phenotipic effect on cell proliferation, colony formation,
cell cycle profile, or some oncogenic mechanism. Otherwise the work appears a list of differentially regulated genes that are not placed in their biological context.

Thank you for pointing it out. Due to funding shortages, we were unable to conduct a corresponding validation experiment. Protein expression data for these three genes are also lacking in the public database. However, we have cited several tumor-related references to illustrate the biological significance of these genes.

8. English editing is required

Thank you for your suggestion. We have checked our English grammar carefully, however, if still not enough, we may consider English editing services or ask for help from English native speakers.

Reviewer #2:

1. The TCGA DNA methylation data of LSCC is not seq-based profiles. The samples were profiled using Illumina 450K beadarray. The author should methylation instead of methyl-seq.

Thank you for your suggestion. We have corrected methyl-seq to methylation accordingly in the whole manuscript.

2. There is a typo in the formula to calculate the p value of enrichment.

We are appreciated you pointing it out. We have corrected our mistake accordingly.

3. As shown by the PCA on expression profiles, the dichotomized smoking levels were not correlated with the top two PCs. Then, the question is what is driving the variation. The author could identify the source and include the source as covariate in the following analyses.

Thank you for your suggestion. There are several confounding factors, as mentioned in the following question. In addition, although the dichotomized smoking levels has nothing to do with the overall expression profile, there is still a correlation to some specific genes, as the conclusion of our article.
4. The authors could look into the smoking pack years more carefully. Some demographic summary should be provided. In this particular case, gender might be a confounding factor to smoking groups. Are genders balanced between smoking groups? If not, the differential analyses might give us differentially expressed genes or differentially methylated CpGs between genders.

We are very appreciated your concerns. We have summarized demographic information in S3 Table and submit it in supporting information. For your second concern, among 299 samples, there are 33 females and 117 males in the high smoking group, 46 females and 103 males in the low smoking group. So genders are unbalanced. There are other confounding factors such as ethnic and race, we have added several reference to discuss such factors in Discussion section, on line 365.

5. For the integrative analysis, how the author defined the gene and CpG site pair needs to be further illustrated. Was it based on PCC, R2, or p value of correlation test?

To respond reviewer’s question, first of all, top 50 hyper- and hypo-methylated CpG islands on 94 genes were selected from methylation data, with a p-value < 0.05. The second step of integrative analysis is to select gene with high transcriptional expression and low CpG island methylation status, or low transcriptional expression and high CpG island methylation status. We explain it on line 256 and 268.

6. English writing needs to be improved. There are many fragments and some sentences are not comprehensible. For example, in the first sentence of Background, what was ranked the second place? Line 62, it should be "leading to". ...

Figures should provide extra information which was not presented in the text. I found some of the figures not informative.

Thank you for your suggestion. We have checked our English grammar carefully, however, if still not enough, we may consider English editing services or ask for help from English native speakers.