Peer Review File

Article information: http://dx.doi.org/10.21037/tcr-22-245

Reviewer A

This paper is an interesting approach to considering hubs of differentially expressed genes in lung cancer. The authors choose to look at genes that are different rather than genes that are similar. While this can be an appropriate approach, the rationale for this approach is not explained in the paper, which creates an impression that only these genes are interesting. The authors need to modify the introduction to explain why this rationale was chosen and they need to modify the limitations section of their discussion to explain that important genes that are similarly expressed may be important and cannot be considered in this approach.
The discussion also needs some overhaul to be a better synthesis of the question and the findings. It is currently just a hodge podge of listing of very select genes that are differentially expressed apparently randomly chosen and not much discussion of the patterns of the hubs that emerge from the analysis.
Reply 1: Thanks for the comments. The purpose of our research is to find hub differentially expressed genes (DEGs) in lung cancer compared with normal tissue. The hub DEGs could serve as potential prognostic markers in NSCLC and therapeutic target in lung cancer for personalized oncology. We have added some findings and research results in the Discussion section.
Changes in the text: we have modified the Introduction section (see Page 4, lines 71-74, 84-87) and the Discussion section (see Page 14, lines 299-303) using red color sentences.

the authors are overusing abbreviations making reading difficult. The following abbreviations should not be used and instead the words simply spelled out to enhance readability: DEG, BP, CC, MF, PPI should all be removed as abbreviations.
Reply 2: Thanks for the comments. We have removed some abbreviations words in our text as advised. DEGs have been widely used in a large number of research papers, and were also applied many times in our study.
Changes in the text: We have modified “BP”, “CC”, “MF” and “ PPI” to “biological process”, “cellular component”, “molecular function” and “protein–protein interaction” throughout the manuscript, respectively.

There are some typos:
219 According to WHO criteria for lung tumors classification and diagnosis, the
220 Lung cancer - "the Lung" - there should be no "the" and "Lung" should not be capitalized.
Reply 3: Thanks for the comments. We have modified “the Lung cancer” to “lung cancer” (see Page 11, lines 224-225).

263 the Phosphorylation of CDC25A protein phosphatase to delay cell cycle progression - "Phosphorylation" should not be capitalized.

Reply 4: Thanks. We have revised “Phosphorylation of CDC25A” to “phosphorylation of CDC25A” (see Page 13, line 267).

Reviewer B

Overall Comments

At first glance, this manuscript requires extensive English language editing. There are places within the manuscript with obvious grammatic errors and vague expressions which may impede and mislead readers. Please proofread the manuscript more carefully. Here are a few of them just looking at the Abstract.

Reply 1: We gratefully appreciate for your comments. We will proofread the manuscript more carefully and strengthen professional English writing.

Lines 32-34, “The Database for Annotation, Visualization, and Integrated Discovery (DAVID) has been used to investigate..”, to “was used to investigate”.

Reply 2: Thanks. We have modified our text as advised (see Page 2, line 34-35).

Lines 37-38, “We identified 84 common DEGs maybe play important roles”, to “that may play an important role”.

Reply 3: Thanks for the comments. We have revised our text as advised (see Page 2, line39).

Lines 40-42, “There were 87 DEGs identified only in SCLC but not in NSCLC tissues compared with normal lung tissues. The 28 DEGs were identified only in NSCLC but not in SCLC tissues”, to “There were 87 DEGs unique to SCLC tissues and 28 DEGs unique to NSCLC ones”.

Reply 4: Thanks for the comments. We have revised our text as advised (see Page 2, lines 41-42).

Lines 45-46, “The 14 hub DEGs were highly correlated with the prognosis of NSCLC”, prognosis in which direction? Overall survival? Sub-stage survival?

Reply 5: Thank you for your rigorous comments. We have modified our text as advised. The 14 hub DEGs were highly correlated with the overall survival of NSCLC (see Page 2, line 46). We have also modified “prognosis” to “overall survival” (see Page 3, line 49, 52; Page 11, lines 242).

Comments on the Abstract

The abstract appears very superfluous lacking a lot of detail pertained to the purpose of the study, the methodology employed, the results retrieved, and the conclusions derived. In terms of Background, there is a lack of purpose going beyond simply that “lung cancer is the most and the highest mortality worldwide”. In terms of Methods, there is lack of: 1) the type of expression profiles used (i.e. microarray); 2) samples of NSCLC and SCLC tissues analysed; 3) the type of functional enrichment (i.e. biological process, molecular pathway, cellular component, KEGG pathway?), as opposed to “pathways and biological role of the DEGs”; 3)

the type of Kaplan-Meier analysis (i.e. survival analysis? Overall? By cancer stage?) 4) which DEGs were used for all of the above analysis (i.e. the overlap of DEGs between NSCLC and SCLC? DEGs unique to each subtype?); 5) analysis tool for deriving hub genes and how this was done. In terms of Results, there is lack of: 1) specific functions or categories enriched 2) magnitude of overall and sub-state prognosis. In terms of conclusion, there is lack of insight as to why these DEGs hold prognostic and therapeutic value.

Reply 6: Thank you for your rigorous comments. We strictly control the number of words in the abstract to meet the standards for submission. We did not show a lot of detail pertained to the purpose of the study, the methodology employed, the results retrieved, the conclusions derived and only describe the more important results and conclusions.

Changes in the text: We have revised the background section of the Abstract (see Page 2, line 28-30). In terms of Methods, 1) GSE43346, GSE40275 and GSE18842 are expression profiles based on the Platforms of Affymetrix Array. We have revised the Methods as advised using red color sentences (see Page 5, lines 93-95); 2) Samples of NSCLC and SCLC tissues analyzed were showed on the Methods (see Page 5, lines 91-93); 3) We have revised our text as advised (see Page 2, lines 34-35); 4) Overall survival analysis was performed using Kaplan-Meier plotter. We have modified our text in several places (see Page 3, line 49, 52; Page 11, lines 242); 5) We have not got overall survival data associated with SCLC patients using Kaplan-Meier plotter and GEPIA. We only performed an overall survival analysis for NSCLC patients; 6) STRING and Cytoscape were used to identify hub DEGs for follow-up research and publish numerous papers We have revised our text as advised (see Pages 2-3, lines 46-52). RRM2, CHEK1, TYMS and SERPINB5 were significantly associated with the overall survival of NSCLC patients and also play an important role in other cancers (see Pages 11-14 in terms of Discussion, lines 240-293). Our findings need more molecular biology studies to be confirmed. Next, we will carry out research on the biological molecular mechanism of hub DEGs.

Comments on the Introduction

The authors provide a thorough epidemiological-based background as to how lung cancer and its subtypes are portrayed at present, whether that is in terms of mortality or therapeutic advances. From Lines 69-72, it seems that SCLC survival has not cached up to that of NSLC. Reading at the Lines 72-74, I am lacking to understand which “differences of these subtypes” the authors are alluding to, and in what directions are they being referred to. In this way it is unclear why authors decided to explore markers in both NSCLC and SCLC. In Lines 77-78, the authors should focus on citing advances in terms of treatment and prognosis specifically in NSCLC and SCLC. References 7 and 10 are irrelevant to the topic explored. Here is a list of newer references on the topic authors need to acknowledge: 1) https://doi.org/10.3390/biology10111200, 2) PMCID: PMC6357337 3) https://doi.org/10.1097/md.0000000000020183 4) https://doi.org/10.7717/peerj.8779 . Lastly, the introduction is missing a rationale and a statement of novelty for this study, as opposed to the literature.

Reply 7: Thank you for your comments. We have added research results to express more clearly and to make easier for readers. We have updated the references 7 and 10 related to the prognosis of lung cancer.

Changes in the text: We have revised the Introduction section as advised using red color sentences (see Page 4, lines 69-74, 84-88).The cited references have been updated (see Page 16, lines 348-350).

Comments on the Methods

"Microarray Data and Identification of DEGs" The authors describe the basic characteristics of the three gene expression studies included along some report of the methodology ensued for differential gene expression analysis. The former should be presented rather as a supplementary table, including some information regarding the patient characteristics (if available) and other information pertained to the array platform employed. Further, the authors should state what adjustment was employed for differential gene expression, why these logFC values were used. Moreover, it is unclear why the authors selected these datasets as opposed to others, and what was the search strategy employed. This study (PMID: 30787969) appears to have analysed the datasets used by the authors, with others which the authors disregarded. Why was this done?

Reply 8: We really appreciate the good question raised by the reviewers which has helped to improve the quality of our manuscript. We choose gene expression profiles mainly based on the following aspects: 1) expression profiling platform; 2) number of specimens; 3) tissue specimen. GSE43346, GSE40275 and GSE18842 are expression profiles based on the Platforms of Affymetrix Array. We selected expression profiles of large sample size opposed to others, and to make research results more reliable. Moreover, our study does not require the basic characteristics of the patients.

Changes in the text: We have revised the Methods as advised using red color sentences (see Page 5, lines 94-95).

"Gene ontology and KEGG enrichment analysis"

Did the authors used an adjustment for deriving a significant P-value? If so, what adjustment was used.

Reply9: Thanks. We directly obtained *p*-values from the DAVID v6.8 database without an adjustment.

"PPI Network Construction and Module Analysis"

It is unclear which DEGs were used to create which networks. Was it the overlap of DEGs between the two lung cancer subtypes, or those unique to each one? The authors are missing information about the confidence score employed for the construction of the PPI network.

Reply 10: Thanks for your comments. The overlapping DEGs in SCLC&NSCLC, unique to SCLC or NSCLC were used to construct the PPI Network, respectively. The confidence score employed for the construction of the PPI network are detailed in references 11 and 12.

Changes in the text: We have revised as advised using red color sentences (see Page 8, line 176-177).

“Validation of hub DEGs and Survival analysis”
The authors are presenting here the term “hub DEGs”, although this has not been defined nor explored appropriately. The authors should conduct an interaction analysis as previously described using a recent high confidence methodology via the cytohubba pluggin of Cytoscape [1) https://doi.org/10.3389/fonc.2021.779042 2) https://doi.org/10.1097/meg.0000000000002349], to validate whether these genes are highly interacting in the PPI network.
Reply 11: Thanks for your comments. We will conduct an interaction analysis to validate the interaction of hub DEGs using a recent high confidence methodology of the study by Giannos, Panagiotis et al. (https://doi.org/10.1097/meg.0000000000002349)

Comments on the Results
“Differentially expressed genes in SCLC and NSCLC”
Please use the terms “unique to” either NSCLC or SCLC, or the “overlap” between the two through out Lines 131-139. Otherwise, it is difficult for readers to keep track.
Reply 12: Thanks for your comments.
Changes in the text: We have revised the Results and the other sections as advised using red color sentences (see Page 7, lines 137-141).

“Functional and pathway enrichment analysis”
The transition statement in Lines 147-148 “while the common DEGs only in SCLC not in NSCLC were mainly” sounds confusing. Please use the terms “unique to” either NSCLC or SCLC, or the “overlap” between the two. Can authors clarify what “anyone pathway” refers to (Lines 174)
Reply 13: Thanks for your comments. We have revised manuscript to make precise and clear for readers.
Changes in the text: We have revised the Results and the other sections as advised using red color sentences (see Pages 7-8, lines 146-177). The DEGs unique to NSCLC were not enriched in any signaling pathway (all *P* value > 0.05, see Page 8, lines 176-177).

“PPI Network and Module Analysis”
The authors refer to hub genes, without any clear definition of what is meant. Equally, interaction-based analysis was not ensued, and should followed to confirm the findings or at least characterise the interactome of these DEGs, so called “hubs”. As previously described, please refer to and acknowledge a recent validation methodology using cytohubba pluggin of Cytoscape [1) https://doi.org/10.3389/fonc.2021.779042 2) https://doi.org/10.1097/meg.0000000000002349].
Reply 14: Thanks for your comments. We will conduct an interaction analysis to validate the interaction of hub DEGs using a recent high confidence methodology of the study by Giannos, Panagiotis et al. (https://doi.org/10.1097/meg.0000000000002349)

“Survival analysis and Cross-validation of hub DEGs”
It is very confusing as which DEGs (i.e. those unique to NSCLC, or SCLC or overlap between the two) were used. Authors should simplify Lines 191-205, so that is more clear

what was queried.

Reply 15: Thanks. We have revised manuscript as advised using red color sentences (see Pages 9-10, lines 199-204).

Comments on the Discussion

The authors attempt to summarise the value of their study in Lines 223- 227, however with the plethora of studies on the literature, the novelty of the current manuscript is poorly defined. Authors need to reiterate how their analysis is different. Some additions to the analysis, such as interaction-based analysis, have been proposed previously.

Reply 16: Thanks for the comments. We have added some findings and research results in the Discussion section.

Changes in the text: we have modified the Discussion section using red color sentences (see Page 14, lines 299-303).

Comments on the Conclusion

Authors reiterate the purpose of their study in Lines 304-305 ("This study has been performed to profile the stable differences and similarities between SCLC and NSCLC"), however this appears at variance with what is written at the end of the introduction (Lines 84-86, "We applied bioinformatics methods to further investigate the relevant biomarkers and potential molecular mechanisms in lung cancer.") or at the Abstract. Authors should clarify what is the fundamental purpose of the study and its novelty. Lines 305-307 are quire superfluous and unnecessary, given that since these subtypes are different, it is likely that they share different patterns of differential expression. Lines 307-309 should be adapted such that to showcase what the major functional findings were.

Reply 17: Thanks for your comments. We have modified our manuscript as advised and deleted the sentence("Lines 304-307","This study has been performed to profile the stable differences and similarities between SCLC and NSCLC. The results showed some common DEGs maybe play important roles in both SCLC and NSCLC, while some genes were only significantly differential expressed in SCLC or NSCLC").

Changes in the text: We have updated the Conclusion using red color sentences (see Pages 14, lines 299-303).

Reviewer C

The manuscript presents a study on profiling the stable differences and similarities between SCLC and NSCLC. The study identifies some common DEGs for both SCLC and NSCLC, as well as DEGs that are unique in SCLC or NSCLC. The author further conducts functional analysis which indicates that the identified DEGs may play different biological functions and are significantly enriched in different pathways. Further survival analysis shows that several DEGs can be potentially informative of the prognosis of NSCLC patients. Although statistical analyses are solid and interesting findings have been presented, the manuscript is not very well written (the writing should be substantially improved before it can be accepted for

publication).

Reply 1: We gratefully appreciate for your comments. We will continue to strengthen professional English writing.

Comments

1. The authors should carefully check grammar errors and typos throughout the manuscript. Here I list several examples. In the abstract, "Lung cancer is the most frequently diagnosed malignant tumor and the highest mortality worldwide, and divided into two differential histologic subtypes, non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC)", "We identified 84 common DEGs maybe play important roles in both SCLC and NSCLC", "KEGG re-analysis of 14 hub DEGs showed that RRM2, CHEK1 and SERPINB5 enriched in p53 signaling pathway, 47 RRM2 and TYMS enriched in Pyrimidine metabolism pathway maybe play a key role in NSCLC and SCLC, and 49 significantly related with prognosis in patients with NSCLC". There are multiple grammar issues in each of these sentences. Page 4, line 71, "during 2009 through 2010" can be replaced by "in 2009-2010". Page 5, lines 92-93, "GSE43346 contained 42 normal tissue samples and 23 SCLC samples. GSE40275 includes 43 normal tissue samples". Please use consistent tense (either present tense or past tense). After a quick skimming I can already see grammar issues in almost all paragraphs.

Reply 2: Thanks for your comments. We will carefully check grammar errors and strengthen professional English writing. We have revised the manuscript to use consistent tense.

Changes in the text: We have updated the Abstract, the Introduction and the Methods using red color sentences (see pages 2-3, lines 27-52; Page 4, lines 71-74; Page 5, lines 94-95;) .

2. Page 4, lines 75-78, "Gene expression profile array and bioinformatics analysis have been applied to study potential clinical biomarkers and molecular mechanisms (of human diseases?). Some key differentially expressed genes (DEGs) have been identified by integrated analysis and are significantly associated with the treatments and prognosis in some cancers (6-11)." Could the authors explain what "integrated analysis" means here? Also, by "are significantly associated with treatments and prognosis in some cancers", do you mean these identified DEGs have been used to determine the type of treatment and the option of prognosis in clinical practice, or statistical analyses have shown evidence of significant association? It's a little bit unclear what do you mean by "some genes are significantly associated with treatments and prognosis". Also, references 7, 8, 9, and 11 are for relevant research on lung cancers, but this sentence is giving a literature review of research on cancers in general. I'm wondering if the authors can add a thorough literature review specifically for recent research on differentially expressed genes across different subtypes of lung cancer and their potential roles in cancer mechanisms.

Reply 3: Thanks for your comments. We have revised the manuscript as advised. Some studies have shown those hubs DEGs are statistically significantly associated with treatments and prognosis in cancers. More basic experiments and molecular mechanisms are needed to be confirmed in clinical practice. There are few relevant studies on DEGs across different subtypes of lung cancer, and more studies on a certain subtype of SCLC or NSCLC. We have updated the references 7 and 10 related to the prognosis of lung cancer.

Changes in the text: We have modified "integrated analysis" to "integrated bioinformatics analysis" using red color sentences (see page 4, lines 78-79). The cited references have been

updated (see Page 16, lines 348-350).

3. Page 4, lines 90-94, haven the data from GSE43346, GSE40275 and GSE18842 gone through the same data generating process and quality control steps, and if so can the authors provide some detailed information on this (or references for these steps)?
Reply 4: Thanks for your comments. We directly use the GEO2R online tool through the same data generating process and quality control steps.

4. Page 5, line 100 "Gene ontology and KEGG enrichment analysis". Have all the pathways from GO and KEGG databases been analyzed? Also have the authors considered adjusting for multiple comparisons?
Reply 5: Thanks. We have analyzed all signaling pathways from GO and KEGG databases.

5. In the analyses have the authors considered adjusting for the effect of potential confounders? For example, Page 9 Line 190, I'm curious if there are additional information on these lung cancer patients such as smoking status.
Reply 6: We gratefully appreciate for your comment. Online Kaplan-Meier plotter and GEPIA database were applied to validate the prognosis value of hub DEGs in lung cancer patients. We have not got more characteristics of lung cancer patients for further prognosis analysis via online Kaplan-Meier plotter and GEPIA database.

6. One general comment is that the authors can highlight the novelty and main contributions (e.g., new findings) of the analyses presented in the manuscript and their connection with previous findings in Introduction and/or Discussion. This can help distinguish their work from the previous work.
Reply 7: We really appreciate the suggestion raised by the reviewers. We have carefully revised the Abstract, the Discussion and the Conclusions as advised text using highlighted changes (red).

Reviewer D

• The authors have satisfactorily addressed most of my concerns and I would recommend the manuscript for publication after a final minor revision.

• However, authors are presenting here the term "hub DEGs", a term has not been defined nor explored appropriately. Classically, "hub DEGs" or "hub genes" are defined as those genes with a specific connectivity in an interaction network of encoded proteins. To address this, authors should expand their analysis using two recent high confidence methodologies (doi: 10.3389/fonc.2021.779042, 10.1097/meg.0000000000002349). Particularly, authors should use the Cytoscape plugin called Cytohubba, and report what the overall interaction value of the "proposed biomarkers" is among the 12 available interaction algorithms (of the tool) in the SCLC and the NSCLC protein networks, separately. In more detail, authors should report the raw values for the 12 algorithms: Degree, Closeness, Betweenness, Radiality, Stress, EcCentricity, BottleNeck, Edge Percolated Component (EPC), Maximum Neighborhood Component (MNC), Density of Maximum Neighborhood Component (DMNC) and Maximal Clique Centrality (MCC)). If possible, authors may wish to report how these values rank in each network.

• The figures would benefit from further editing such as font compatibility and at instances were text or shapes appear misaligned.

Reply: Thanks for your comments and detailed explanation. The "hub DEGs" or "hub genes" are defined as those specific connectivity genes in an interaction network of encoded proteins. The plugin MCODE (https://doi.org/10.3389/fimmu.2022.814303, doi: 10.1097/MD.0000000000025553, doi: 10.1002/cam4.3907. Epub 2021 May 2, doi: 10.3389/fgene.2020.613744. eCollection 2020 et al) of the Cytoscape have been widely used to study hub genes in protein-protein interaction networks.

We adopt the 12 topological analysis methods of Cytoscape plugin cytoHubba to validate the hub DEGs that we have screened. Top 47 hub DEGs are indentified according to 12 topological algorithms ranked in protein-protein interaction (PPI) network. The 7 of 12 topological analysis methods have identified 100% (47/47) hub DEGs that we have screened using plugin MCODE in the SCLC&NSCLC PPI networks. The remaining 5 methods have identified at least 25 of the 47 hub DEGs (Table S1). The 7 of 12 methods have identified at least 8 of the 11 hub DEGs we have screened using plugin MCODE unique to SCLC PPI networks (Table S2). The 11 of 12 methods also have identified at least 7 of the 8 hub DEGs unique to NSCLC PPI networks (Table S3). The 12 topological analysis methods of the plugin cytoHubba further support the results of our study.

Reviewer E

The authors have tried to address the majority of my questions. However, the manuscript still requires a lot of English writing editing, and I'm not sure if the authors can improve the writing within a short period of time. My suggestion would be that if the English writing is accepted for publication in the journal then the manuscript can be accepted. Otherwise it should be sent back to the authors to be

further revised before it can be considered for publication.

Reply: We gratefully appreciate for your comments. We have asked English-speaking experts to improve English writing editing of the manuscript and proofread the manuscript more carefully.