Peer Review File

Article information: http://dx.doi.org/10.21037/tcr-22-709

Reviewer Comments

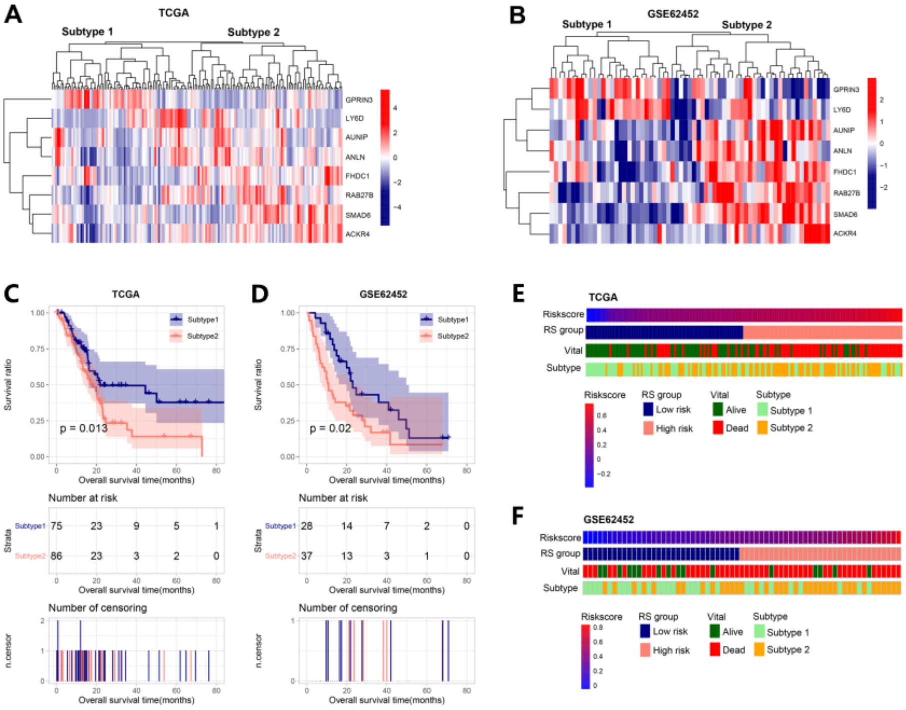
Main Findings

Qu et al. employ a bioinformatics analysis to identify mRNA transcripts in pancreatic cancer that predict the patient's survival. Overall, the methods and results of this paper are well done and straightforward to follow. The study is captivating and should be of interest and relevance to the broader bioinformatics and cancer research community. However, Qu et al. have not referred to the vast pool of recent works on pancreatic cancer, the subtypes thereof, and the link between disease aggressiveness and gene expression – more on this in the comments section – which should strengthen the paper and put it in context with the current literature.

Comment 1: The categorisation of pancr9eatic tumours into the two groups using gene expression data is, for all intent, a molecular subtyping approach. Gene expression studies have also identified and described subtypes of pancreatic cancer with prognostic and biological relevance, including the following: https://doi.org/10.1016/j.ccell.2017.07.007, https://doi.org/10.18632/oncotarget.25632, https://doi.org/10.1038/ng.3398, https://doi.org/10.1038/s41575-019-0109-y, https://doi.org/10.3390/cancers13020322, and https://doi.org/10.1371/journal.pone.0257084. Could the authors show the relationship (possible overlaps in patient samples) between the transcription subtypes that have been previously defined and the "survival associated" subtypes identified using their approach using a figure similar to Figure 1D of https://doi.org/10.1371/journal.pone.0257084? They could use only the mRNA molecular subtypes (by Bailey et al., Collisson et al., and Moffitt et al.) of the same TCGA samples analysed in the current article in the Supplemental Information of the manuscript found here: https://doi.org/10.1016/j.ccell.2017.07.007. Also, the above research articles should be cited, and a paragraph included in the introduction about these previous studies (some used the same TCGA datasets) to bring the current research into context with recent developments.

Reply 1: Thanks to the reviewers for the detailed suggestions and references provided. We have added Figure 11 based on references, and the specific analysis process is as follows. First: The approach in plos one is to compare the authors' proteome-based expression level classification results with the clustering results of three known documents (by Bailey et al., Collisson et al., and Moffitt et al), The clustering results of the three known documents are available, but cannot be used in the comparison with our clustering results, because our results are based on the transcriptome and plos one is based on the proteome; Second: We consulted many literatures (such as https://pubmed.ncbi.nlm.nih.gov/30018740/ etc.) and obtained some literatures that classified the TCGA PRAD transcriptome, and the samples they used It is the same as us, based on the TCGA PRAD transcriptome, but the published literature does not provide information on the specific classification of the samples in their results (such as sample 1-Classic), and we cannot obtain such information. Compare with samples included in our cluster.

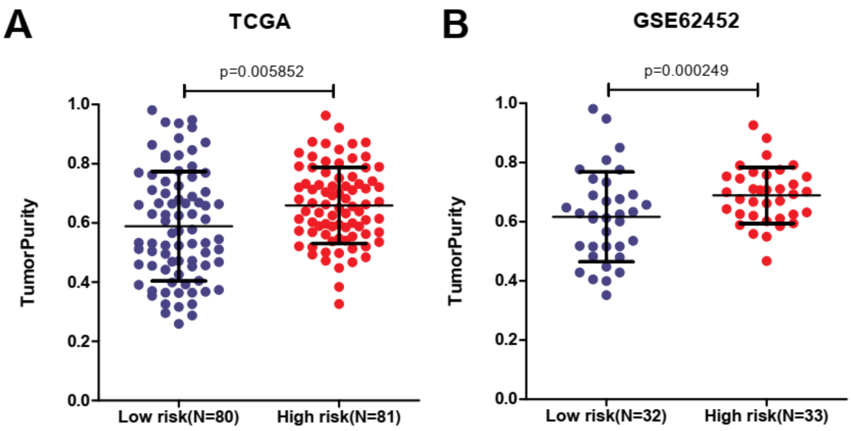
Although the results of the published literature cannot be obtained for comparison, we can learn from the algorithms in the literature. In the PAAD samples we included in the analysis, cluster analysis was performed based on 8 characteristic factors to obtain different subtype categories, and then examine different subtype categories. The relationship between subtype categories and risk groupings, in both the TCGA training set and the GSE62425 validation dataset, we performed based on 8 eigenfactors as described in the literature https://pubmed.ncbi.nlm.nih.gov/30018740/ Clustering, different subtypes, subtype1 and 2 were obtained, as shown in supplementary figure 3 A and B, it can be seen from the figure that the distribution of expression levels of the 8 genes is relatively consistent.
Afterwards, the correlation between different subtypes and survival prognosis was investigated in subtypes 1 and 2 obtained in the TCGA training set and the GSE62425 validation data set, respectively. The results are shown in supplementary figure 3 C and D. The results show that the results in the TCGA training set and the GSE62425 validation data set are consistent, subtype1 has a better prognosis, and then a heat map display is performed based on the information of the samples in the subtype, as shown in supplementary figure 3 E and F. The data information of this part can be found in the attached table 10.



Changes in the text: see page 9 line181-187, and page15 line 310-321

Comment 2: Is tumour purity a driving factor? A boxplot and/or t-test result showing the difference in the purity of the tumours between the two groups should suffice to answer this question. Tumour purity information can be found in the Supplemental Information of the manuscript here: https://doi.org/10.1016/j.ccell.2017.07.007.
Reply 2: Thanks to the reviewer's suggestion, we use the estimate package in R3.6.1 to evaluate the TumorPurity of the TCGA and GSE62452 dataset samples respectively. For the data results, please refer to the attached table 11, and then use t-test to group samples with

different risks. The TumorPurity in the difference comparison, the display figure and the comparison p-value are shown in supplementary figure 4. In both TCGA and GSE62452 datasets, the p-value is less than 0.05. We believe that there is a correlation between tumor purity and our risk grouping.



Changes in the text: see page 9 line189-192, and page15 line323-327

Comment 3: in many instances, e.g., in line 229, the author state that “The high-risk group showed a shorter survival time compared to the low-risk group for five genes (ANLN, LY6D, RAB27B, SMAD6, and AUNIP), and the low-risk group showed a shorter survival time compared to the high-risk group for three genes”. Could the author indicate the actual values of the median survival duration of each group? E.g., …shorter survival (median OS = 14 months) compared to the other groups (OS = 40 months)?
Reply 3: Thanks to the reviewers for their comments, we have revised the manuscript to add data of the median survival duration of each group in attached table 12.
Changes in the text: see page 12 line 252-253

Comment 4: The model provided by the author is highly accurate; however, they do not show any confidence intervals. Could the author also provide these confidence intervals, which can be generated, for example, by running the model multiple times using bootstrapping or re-sampling approaches?
Reply 4: Thanks to the reviewers for their comments, we added these confidence intervals in attached table 12
Changes in the text: see page 11 line 242

Comment 5: In many instances, the authors mention the “survival time” and “survival prognosis” without being specific. The TCGA has reported these survival measures, overall survival, progression-free survival, disease-specific survival, and disease-free survival. Could the author categorically state which measure they used and refer to in each instance?
Reply 5: Thanks to the reviewers for their suggestions, we have revised the details of the text. All survival time in this article refers to overall survival time.
Changes in the text: see page 12 line 257-258, page 12 line 247, 251, 257,page 13 line 270,

page 14 line 291,294, page 30 line 663

Comment 6: the discussion section is relatively lengthy because the authors list what is known about every gene in their list. The section can be more concise by stating only a few details about the genes. Some of the genes can be mentioned in one sentence.
Reply 6: We agree with the reviewers and have streamlined the Discussion section
Changes in the text: see page 16-18, line 349-380

Comment 7: On line 74: "Therefore, in-depth exploration of the underlying mechanisms of pancreatic ductal adenocarcinoma and identification of effective prognostic indicators are important for clinical treatment decisions and patient management." The research papers listed previously show molecular subtypes of clinical relevance and the associated mRNA, proteomics, genomics, and transcriptomics signatures. The current manuscript should refer to the efforts of these papers as they utilised the same datasets from the TCGA.
Reply 7: Thanks to the reviewer's suggestion, we refer to the efforts of these papers. And added relevant descriptions and citations in the text.
Changes in the text: see page 4, line 74-80

Comment 8: The fonts in Supplementary Figure 1 seem too small.
Reply 8: We have modified the fonts in Supplementary Figure 1.