Reviewer’s report

Title: A 25-gene classifier predicts overall survival in resectable pancreatic cancer

Version: 0 Date: 05 Jul 2017

Reviewer: Dung-Tsa Chen

Reviewer's report:

Summary: The manuscript is well organized and written. Several issues are needed to be addressed.

Strengths:

1. Excellent introduction of current clinical issue of pancreatic cancer and potential utility of genomic signature.
2. Significant effort to collect and synchronize the 9 valuable gene expression data of pancreatic cancer.
3. Comprehensive analysis of univariate and multivariate analysis with comparison to clinical variables and other molecular subtypes.

Weaknesses:

1. Training set is based on TCGA to test OS>36 months vs. 2-6 months OS. How are the results if other datasets are used as the training set? Or are these 1,400 genes using TCGA also significant in other datasets when comparing OS >36 months vs. 2-6 months OS? What are the distributions (sample size) of other datasets regarding OS: >36 months and 2-6 months? Any justification to use TCGA instead of other dataset as training set?
2. It is unclear how different datasets are pooled. It seems dataset is normalized individually. Since each dataset may use different platform of microarray/RNA-seq, special treatment is performed on probeset level to convert into gene level data for later data analysis. However, it is unclear if any cohort adjustment (or batch correction) is performed prior to data pooling. In other words, how the validation cohorts are pooled. Explanation is needed.
3. It has become a standard to provide function codes with output of data analysis to ensure reproducibility and transparency. Since R software is used in the study, authors should
consider Rmarkdown using knitr package to generate the report, such as from data processing from probeset to gene level, data pooling, gene signature development of 1400 and 25 genes in training set and the validation in test set, as well as the comparison to clinical variables and three molecular subtype classifiers in univariate and multivariate analyses.

4. Evaluation of the random noise gene seems biased. The 25-gene signature is not directly derived from the whole 15,261 genes. It is the result after tuning up the 1,400 genes, meaning it is selected from the 1,400 genes, not from the 15,261 genes. Thus, the evaluation should be based on the 1,400 genes.

5. The algorithm from 1400 genes to 25 genes is unclear. It seems LASSO is used to finalize the 25 genes using the glmnet package. However, how is the penalty regularization parameter $\lambda$ value finalized? Detailed explanation is needed.

6. The TCGA data as training set is for PDAC. However, some of the other datasets may not be PDAC. This may create heterogeneous issue. Please justify.

7. How many of the 25-gene signature overlap with other signatures and the three molecular subtypes.

**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.

Yes

**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.

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