Peer Review File

Article information: https://dx.doi.org/10.21037/atm-22-1931

Reviewer #1:

A good paper that I liked to read, and one question to address:

Major concerns

Comment 1: From the discussion "differential gene expression analyses were performed between TNBC and HR+/HER2-, HR+/HER2+, HR-/HER2+, and normal tissues". It would be interesting to establish some measures of reproducibility based on confidence.

Response: Thank you for the sincere comment.

Comment 2: "Co-expression network analysis was also conducted to identify TNBC-related gene modules. Finally, the genes in the intersection between the DEGs and the selected TNBC-related modules were named TNBC-related genes and were adopted for signature development". Now, such TNBC-related genes are interesting but tricky, as they would need validations from both computational and biological standpoints. The authors should add some considerations about these points.

Response: We are grateful for the suggestions. Cautious interpretations should be noted since no experimental validation was performed in our study. We have added some considerations into the discussion (**Manuscript Tracked Changes, page 17, lines 350-352**).

Reviewer #2:

This is a very interesting manuscript as triple negative breast tumor are very aggressive tumors that needs novel approaches to be treated.

Major concerns

Comment 1: Authors found a six mRNA signature with prognostic value that in combination with clinico-pathological data. This combination of molecular and clinico data could improve OS prediction.

Response: Thank you so much for your positive assessment. We hope that we can apply this model to an accurate prediction of overall survival in clinical practice and provide more supporting evidence for clinical decision-making.

Reviewer #3:

This is an interesting study attempting to establish a prognostic gene expression signature for TNBC patients. However, in order to be published I suggest the following edits.

Major concerns

Comment 1: Abstract includes many technical definitions, for example:'" (HR 7.49, 95% CI 3.57-15.70) or categorical (HR 15.70, 95% CI 3.67-67.22) predictors…" which makes the people less familiar with those statistical parameters to understand the study. Abstract should be more general and include the main conclusions, without including technical details.

Response: Thank you for your comment. According to your suggestion, we have revised some sentences in the abstract to be easily understood (**Manuscript Tracked Changes, page 2, lines 39-45**).

Comment 2: TNBC heterogeneity is a highly important issue and a central topic in this paper, and thus should be discussed in the introduction and discussion in more detail. For example, authors state: "TNBC can be further divided into four or six subtypes according to different classifications." However, there are more classifications that should be discussed. Look for example at Miquel Ensenyat-Mendez et al., 2021, "Current Triple-Negative Breast Cancer Subtypes: Dissecting the Most Aggressive Form of Breast Cancer", discussing different classifications based on based on gene expression, metabolic pathways, methylation etc. Or using proteomic analysis Vasudevan et al., 2021, ("Drug-Induced Resistance and Phenotypic Switch in Triple-Negative Breast Cancer Can Be Controlled via Resolution and Targeting of Individualized Signaling Signatures") found at least 17 different subtypes in TNBC subgroup. At least 10 subtypes were found by Sarah-Jane Dawson et al., "A new genome-driven integrated classification of breast cancer and its implications".

Response: Thank you so much for the suggestions. We have discussed the TNBC heterogeneity, such as diverse subtypes divided by gene expression, metabolic pathways, or proteomic data. For example, Miquel et al. reported that different classifications based on gene expression, metabolic pathways, methylation, etc. These contents had been added into the introduction (**Manuscript Tracked Changes, page 4, lines 63-67**) and discussion (**Manuscript Tracked Changes, page 14, lines 284-291**).

Comment 3: It is important to add at least one independent dataset to the paper in order to show whether the results are reproducible and similar parameters are obtained from this independent dataset, such as key modules, correlations between the relevant parameters as presented in the paper, gene annotations in the modules, 267 TRGs etc., the final gene signature. I see that the authors discuss this issue in the discussion, line 324-326, however it is not clear what does it mean" but not all genes involved in the signature were found". Not all the genes were TRG? Or differentially expressed? Or technically were not present in the datasets? Since this kind of validation is important the author should discuss these points and make more efforts to prove the validity of the conclusions. What will happen if the authors randomly divide the data set into two - mostly 50:50 for training and validation? The same conclusions will be obtained?

Response: Thank you for your good suggestion. We have performed external validation using our previously published FUSCC-TNBC dataset. Multivariate Cox regression was performed to establish a six-gene model and derive risk scores. Higher risk scores suggested a worse prognosis (HR 2.9, CI 95% 1.3-6.4; P = 0.007). The top quartile of risk score was set as the cut-off value and presented a good performance in prognostic stratification (see Supplementary Figure. S5). Notably, the model of the pathological stage combined with risk score also exhibited a good performance (C-index = 0.72) (**Manuscript Tracked Changes, page 13-14, lines 270-276**).

The small sample size of TCGA-TNBC (n= 158) leads to the fact that dividing the dataset into two (mostly 50:50 for training and validation) will influence the model training and statistical power. The bootstrap resampling method we adopted in our study is useful when the size of dataset is small and it is difficult effectively divide the training or validation sets.

Comment 4: Enriched, TNBC-related, modules in the section "GO and KEGG pathway enrichment analysis "should be discussed in more detail. For example, what is new? What corresponds to the previously known, TNBC pathways from the literature?

Response: Thank you for the suggestion. According to the constructive comments, we have discussed in detail which pathways are newly discovered and which are already reported in TNBC. For example, PI3K-Akt signaling pathway and MAPK signaling pathway play an important role in TNBC. Many studies have exhibited targeting PI3K-Akt signaling pathway or MAPK signaling pathway has great therapeutic potential. More detailed contents were added into discussion. More detailed contents were added into the discussion (**Manuscript Tracked Changes, page 15, lines 302-314**).

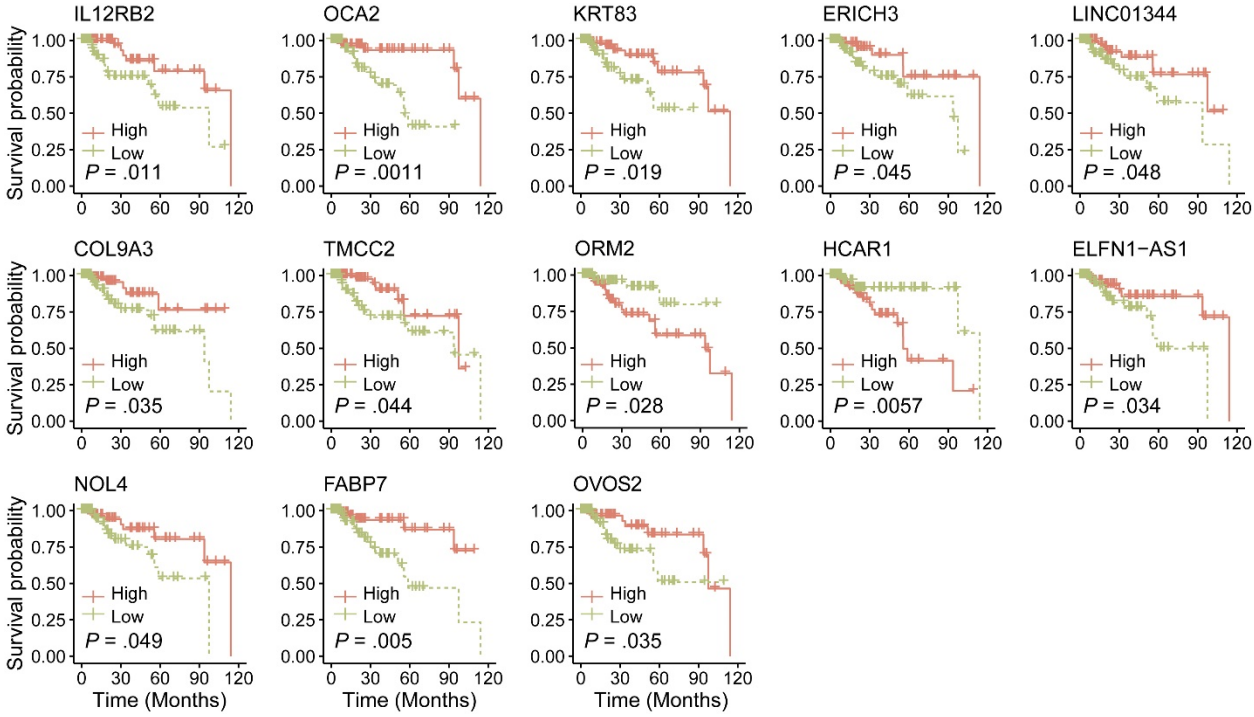
Comment 5: line 219 -why 6 genes were selected as final number to comprise the signature, this should be discussed in a more clear manner.

Response: Thanks for this comment. We used the LASSO Cox regression model to select these six prognostic TRGs, and we have elucidated why these six genes were selected as final number to constitute the signature in a more comprehensible manner (**Manuscript Tracked Changes, page 11, lines 224-226**).

Comment 6: In figure 4 - what is the time scale (months)? Should be indicated in the x axis. Which color indicates low levels? All figures should include the necessary info to understand the data easily. I suggest to add more details to the figure legends so one can understand the plots easily.

Response: Thanks for this good suggestion. We have added more necessary info into the figure 4 and figure legends for ease of understanding (**Manuscript Tracked Changes, page 25, lines 555-557**).

Figure 4 are shown as follows.



Comment 7: The section "Performance of the prognostic signature and nomogram." should be judged by the expert from the field.

Response: Thanks for this kind suggestion.

Page 5 Line 91-84:

We also included our previously published cohort of 465 TNBC patients treated at Fudan University Shanghai Cancer Center (FUSCCTNBC). The RNA-sequencing data of FUSCCTNBC are available in the Sequence Read Archive (RNA-seq: SRP157974).

Page 17 Line 357-361:

~~Second, external validation was not performed despite great necessity. We have tried several datasets containing high-throughput sequencing data of TNBC from the Gene Expression Omnibus, but not all genes involved in the signature were found. Other datasets did not provide detailed clinicopathological characteristics or follow-up data. Nevertheless, we performed validation using the bootstrap resampling method.~~ Second,

Page 18 Line 343:

All data in the current study were available in TCGA.

Page 18 Line 376:

All data in the current study were available in TCGA and FUSCCTNBC datasets.