Peer Review File

Article Information: https://dx.doi.org/10.21037/tcr-22-2444

This study used gene expression data on breast tumor and adjacent normal tissue samples from TCGA cohort. Authors first identified genes differentially expressed between tumor (by stage) and normal tissues, and based on these differentially expressed genes, they further identified two gene sets that are associated with breast cancer survival. Potential mechanisms were explored, with a focus on immune-related pathways. The study may have some value with a goal of identifying genes that may play a role in breast cancer development, and gene sets associated with breast cancer progression and mortality. There are some aspects, however, need to be clarified and improved.

1. In the introduction, where the breast cancer statistics are described based on the US population, but when describe the rationale of the study (lines 68-71): as, despite the advances in treatment modalities, mortality rate does not decrease,---this is not true, in the US women, the breast cancer mortality rate decreased,… here, authors may mean, the breast cancer morality rate in Chinese population, should be clear? Maybe discuss the rationale, like, are there any similar diagnosis/prognosis biomarker study using gene expression data in breast cancer? what is new (or research gap) for this study regarding bioinformatic analysis or other aspects of the study, compared to other previous studies in the area?

   Reply 1: Thank you for your comments. We have corrected the statement here, and complemented the rational discussion of breast cancer mortality.

2. **Describe rationale why DE genes are not only differentially expressed between tumor vs. normal tissue, also have to gradually upregulated or downregulated across tumor stage—to be defined as BC-developmental genes?**

   Reply 2: We have discussed the rational of DE genes selected to be upregulated or downregulated to be defined as BC-developmental genes in the section of Discussion.

3. **Describe STEM analysis—why and how, to be used in the study after the DE analysis by Limma package? Lines 102-105, not clear why this method should be used, and how.**

   Reply 3: Thank you for your comments here. We have complemented the introduction of STEM analysis in the section of Methods.

4. **Lasso was used to identify gene sets that are most associated with breast cancer survival, may need more details how Lasso will be used, two sets…, and the lasso step is based on DE genes defined as BC-developmental genes only, I am wondering if you use any DE genes (tumor vs. normal?), the gene sets could be different?**

   Reply 4: Thank you for your question. We have complemented more details of Lasso used in the analysis of gene sets for the prediction of BC. As discussed in the section of Discussion, the outcome of Lasso regression analysis should be different if we used DE genes instead of BC-developmental genes set. However, DE genes tend to contain numerous transcripts that did not participate in the development of BC due to selected bias and analytic error, resulting in the relatively low predictive efficacy of model.

5. **In the evaluation of the two models, can you do ROC analysis on clinical factors plus gene sets (two models), and ROC analysis on clinical factors only, and also ROC analysis on BC1 or BC2 gene sets only, so that we can see if adding the BC1 or BC2 will significantly add additional value for survival? If not better than the clinical factors only model, the gene expression panel may not be clinically important?**

   Reply 5: Thank you for your suggestion. In the original study, we have performed the ROC analysis on BC1 and BC2 gene sets, and the AUC of BC1 and BC2 model were 0.80 and 0.66. In this revision, we have performed the ROC analysis on clinical factors, and the AUC was 0.49 (see supplemental materials). These results indicated the clinical importance of BC gene sets in the prediction of survival, especially for BC1 gene set.

6. **The writing is ok, but should be read through carefully, where in many places, descriptions are not precise.**

   Reply 6: This manuscript has been extensively revised by a native English speaker.

7. **Figure 3, footnotes are labeled wrong? Like D: scatter plot of risk score from model BC1 (should be BC2?, B is BC1)? E-F is for BC1, and G-H also for BC1 (should be BC2)?**

   Reply 7: Sorry for our mistake here. We have corrected the footnote for Figure 3G and H from BC1 to BC2.

8. **For risk score calculation, at lines 134-139, need to describe more details, how the score was calculated based on gene expression, like for up-regulated or down-regulated genes, high score represents what? if favorable or unfavorable genes?**

   Reply 8: We have complemented the details of risk score calculation in the section of Methods.

9. **Figure 3F: BC1 is unfavorable gene set-, but HR is 1.041 (1.017-1.065) for risk score (high vs. low group?)—weakly associated with survival, but in H: BC2 is favorable gene set, but HR is 2.16 (1.66-2.70, high vs. low risk score group?), with stronger increased risk for survival (should be high score is better if gene set is favorable), the conclusion is that BC1 is significantly associated with survival, but not BC2?**

   Reply 9: To explore the prognostic efficacy, we have performed ROC analysis of BC1 and BC2 model, and the results of AUC of ROC analysis favored better prognostic efficacy of BC1 model. Moreover, to test these two models in cohorts from different populations, we assessed the testing power of two models of BC development-related gene sets in GSE4922. Results also indicated the testing power of BC1 model was higher than BC2 model. Therefore, we concluded that BC1 model was of high prognostic and predictive efficacy of BC patients.

10. **Figure 4, E and F, the difference on drug sensitivity seems very small between low vs. high risk group, but the P<0.05?**

    Reply 10: We have noticed the raw data in low and high risk groups from Figure 4E and 4F seems to be of relatively low heterogeneity in high risk group, while which seems to be of great heterogeneity, contributing to the significant P value in Figure 4E and 4F.