POINT-BY-POINT RESPONSE TO REVIEWER 1

****

Independent Review Report, Reviewer 1

EVALUATION

Sun et. al, perform transcriptomics analysis to build a model for the evaluation of tumour progression in NMIBC which consider the tumour immune microenvironment. The authors explain that this is a novel prognostic model for NMIBC. This study adds to the literature on NMIBC and has highlighted the need of considering various tumour micro-environment factors/interactions in the predictive modelling for prognostic. ***However, there are some concerns that include limitations in the methodology used such as re-normalization based on house-keeping gene, when input data is obtained from 2 completely different methodologies i.e. microarray and RNASeq.*** I recommend the following to take into consideration.

Response: We greatly appreciate the Reviewer for reading our manuscript and reviewing it, which helps us improve it to a better scientific level. After following the Reviewer's suggestions, we believe that we had overcome some limitations in our study when we first submitted it. Both major revision suggestions are essential for maintaining the statistical rigor of our conclusions. From each suggestion, we obtained an overview of the well-established methods and found the fully applicable method to our study. We would like to thank the Reviewer again for the valuable suggestions. Our point-by-point responses are as follows.

****

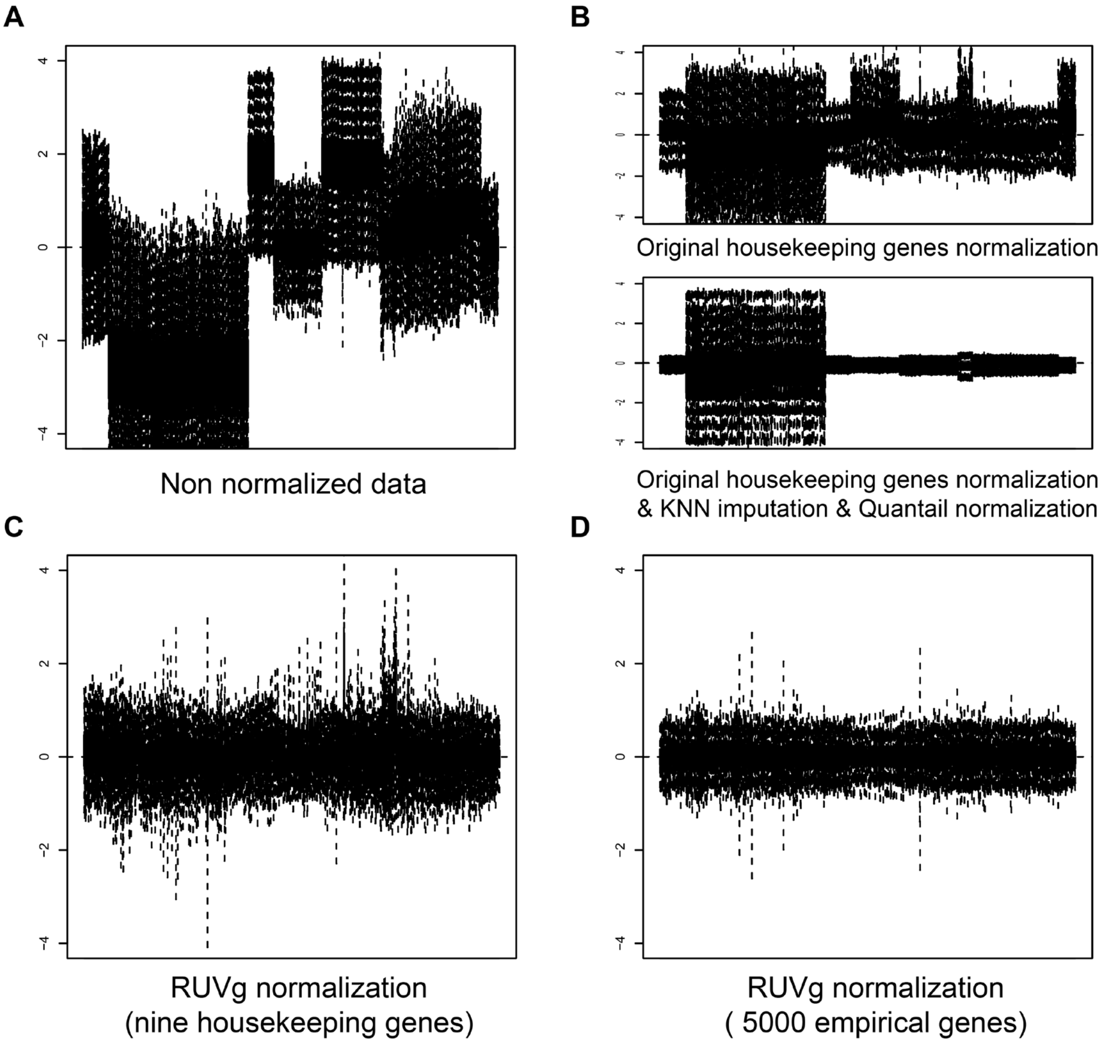
Major:

(1) As the authors highlighted in their discussion, re-normalization schema before model establishment has a great role to play in the accuracy of prediction model building. These are essential steps for Unbalanced Transcriptome Data (which could be created when data is prepared in a different lab setting, research groups, or due to different batches or using different sequencing platforms/methodology). In this study authors used the average expression of the 9 different housekeeping (HK) genes for re-normalization, they simply divide the raw gene expression with the averages. The origin of data used in this study is from 2 completely different platforms and it is very unlikely that even the HK genes expression among those datasets is comparable since it is obtained by 2 completely different methods (if HK expression is different, how come this can be considered for normalization). Integration of microarray and RNASeq data is still not well established. https://www.frontiersin.org/articles/10.3389/fbioe.2019.00358/full This review explains various algorithms that tried to solve these issues, but every algorithm has some assumptions to consider. Authors should also take a look at various batch effects related to normalization methods such as RUV (https://www.nature.com/articles/nbt.2931). I would suggest the authors re-consider this step using some well-established approach. Visualization plots such as PCA, RLE would be

very useful to look for the differences among datasets due to technology or study of origin.

Response: We greatly appreciate the Reviewer's insightful comment. After using RLE and PCA plots to evaluate our integrated data matrix, we discovered significant batch effects from both technology platforms and data sources. The recommended references helped a lot in improving this step. From Liu's review, we fully understood the general assumptions and commonly used methods for normalizing unbalanced transcriptomics data. In Risso's article, we acquired a helpful tool, RUVg, that can be implemented in our study.

We evaluated five different re-normalization conditions of their batch effect eliminating abilities by PCA and RLE plots. As we can see from Appendix Figure 1A and 2A, the non-normalized data are mainly distributed according to the origin of data sources. In Appendix Figure 1B, 2B and 2C, although the distribution of the integrated data matrix tends to be more homogerous, the PCA still isolates the RNA-Seq dataset from the microarray ones. Moreover, the RNA-Seq data were discovered with a more dispersed feature than microarray data after our original re-normalization procedure (Appendix Figure 1B, the graph below). This observation is coordinated with an SVA study's results in dealing with a combined RNA-Seq/microarray data set(Mooney et al., 2013).



Non normalized data

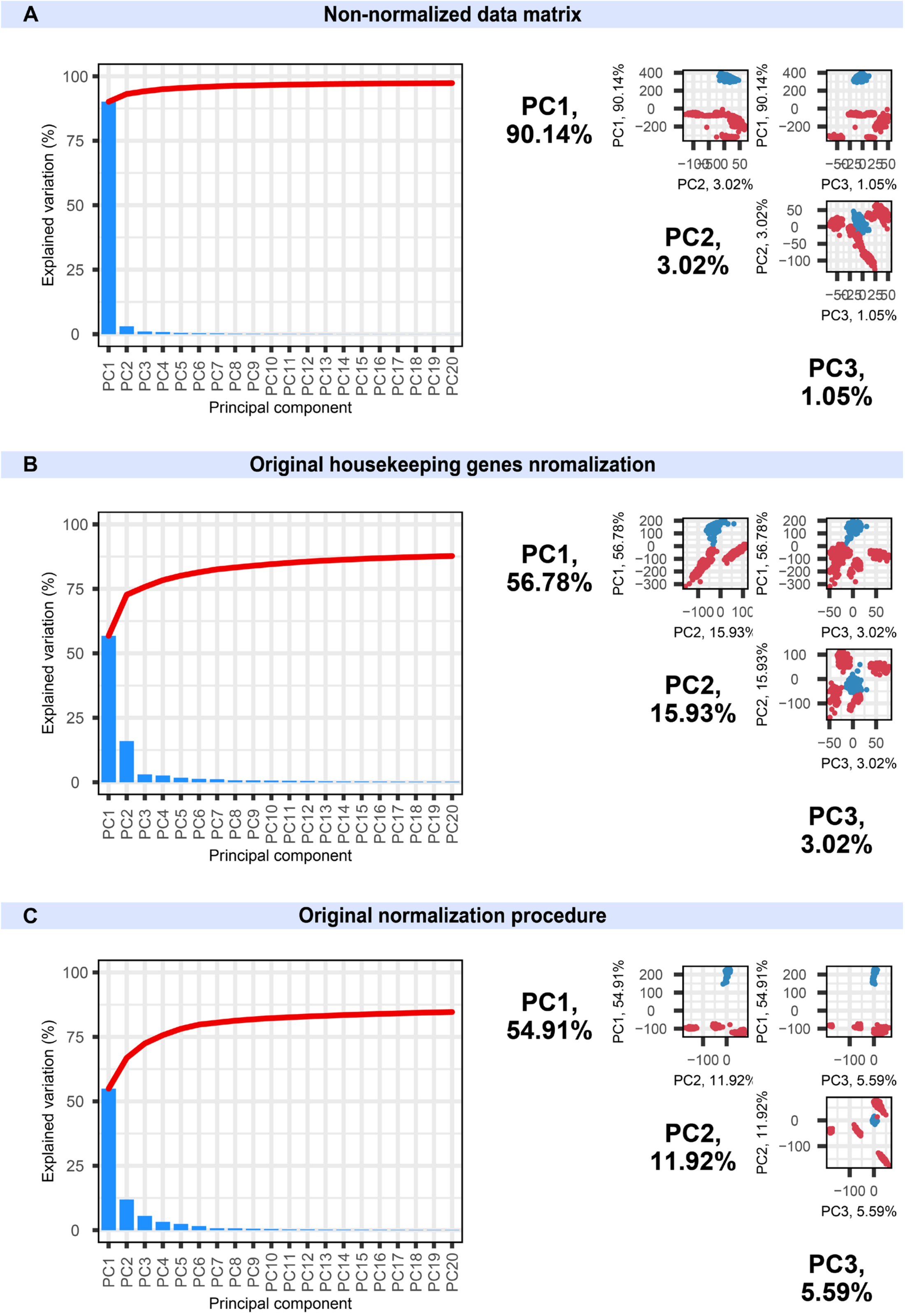
Original housekeeping genes normalization

Original housekeeping genes normalization & KNN imputation & Quantail normalization

RUVg normalization (nine housekeeping genes)

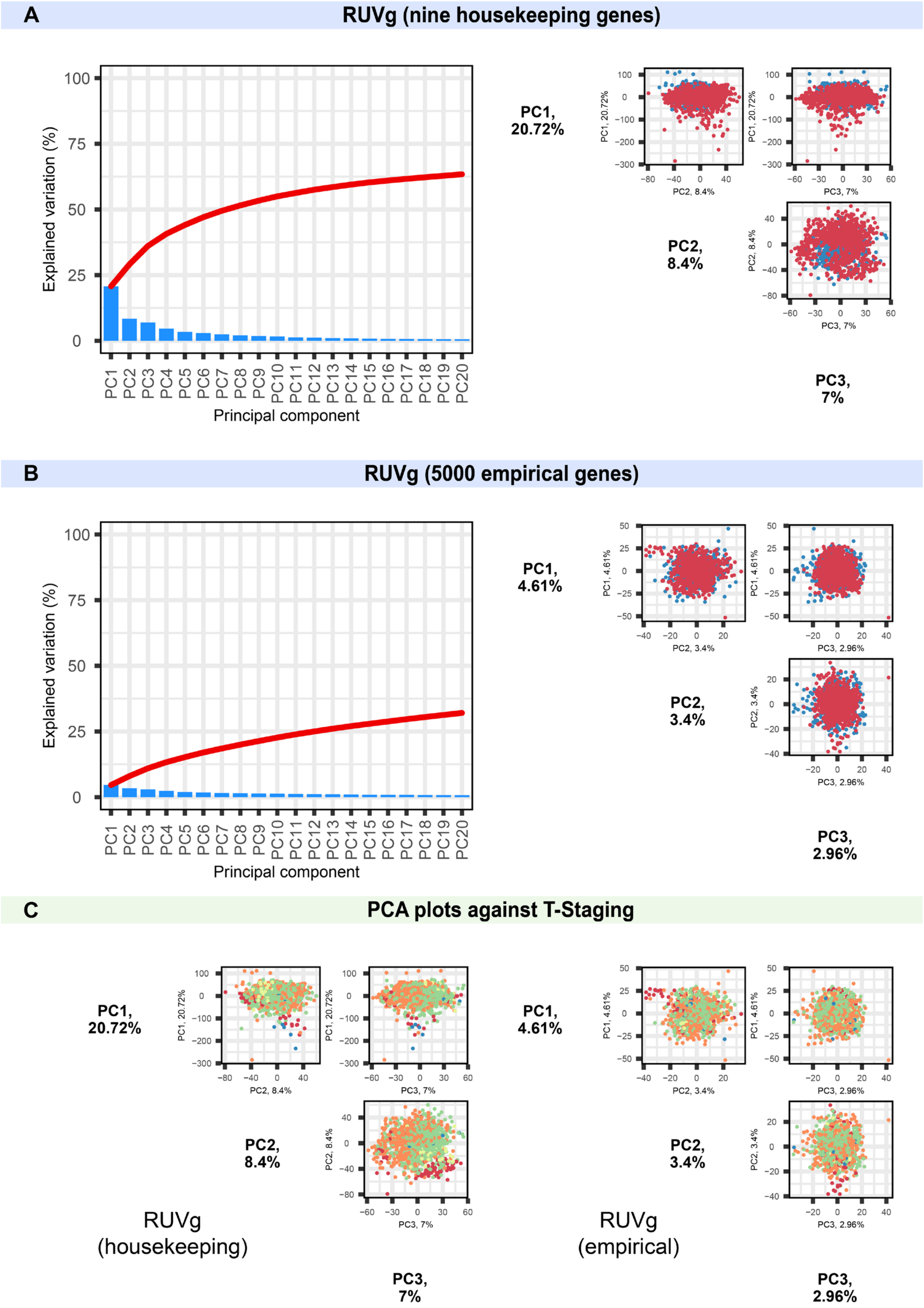
RUVg normalization (5000 empirical genes)

Appendix Figure 1. RLE plots for 1370 transcriptomics data after different normalization procedures.



Appendix Figure 2. SCREE and PCA plots for 1370 transcriptomics data from different steps in our original normalization procedure.

Then we tried the RUVg normalization method using two pre-defined negative control gene sets: our original nine housekeeping genes and 5000 *in silico* empirical control genes filtered by the embedded method in "RUVSeq" package. Both strategies can help correct the batch effect from technology platforms and dataset sources (Appendix Figure 3A and 3B). Furthermore, both strategies can also successfully isolate some of the T0 stage non-cancer samples from urothelial cancer samples. At last, we chose the RUVg normalization method using 5000 empirical negative control genes as our final strategy to re-do the normalization and related subsequent analyses. Because when we used nine housekeeping genes as negative control genes, the distribution of output data was less homogenized than the output data generated by using 5000 empirical genes (Appendix Figure 1C and 1D).



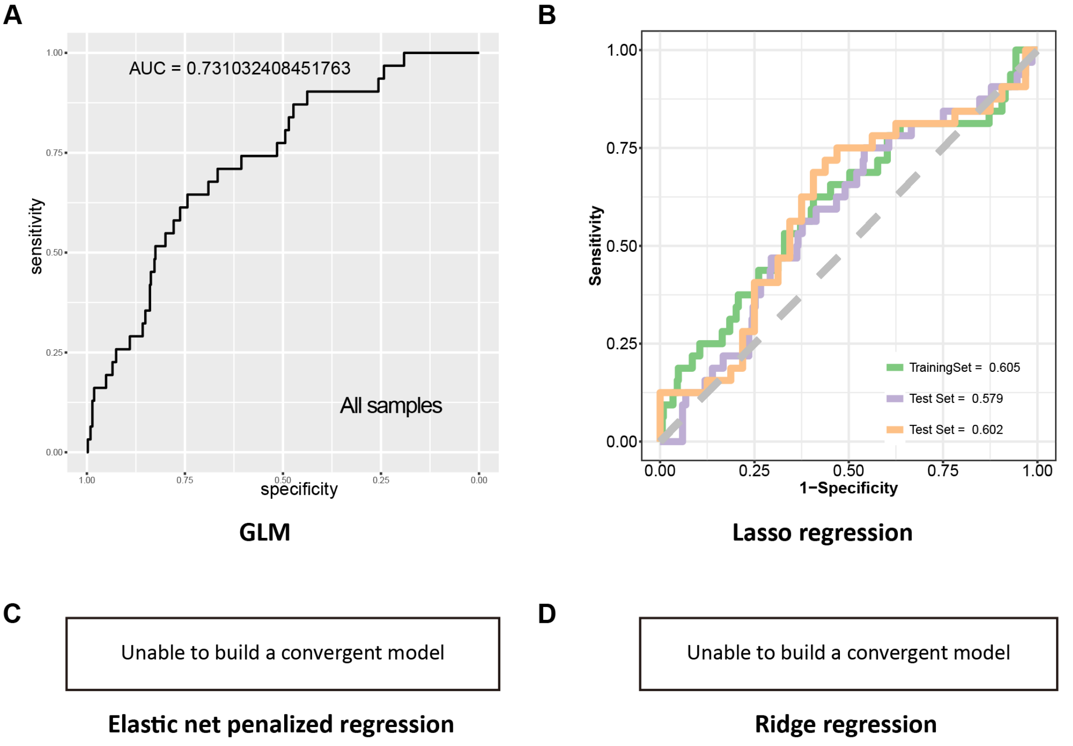
Appendix Figure 3. SCREE and PCA plots for 1370 transcriptomics data after RUVg normalization procedures used two different negative control gene sets.

In summary, we have tried RUVg with both nine housekeeping genes and 5000 empirical genes as negative control genes as our new re-normalization procedure. The empirical strategy was confirmed with the best re-normalization effect and, in the meantime, kept the possibility of distinguishing T0 stage cancers. The modified description of this step can be found in Line 169-172.

(2) Similarly when DEG analysis is performed; the author used the arithmetic means of the FDR values. There are other methods defined for combing the p-values from multiple analyses (https://www.nature.com/articles/s41598-021-86465-y). or here are some other suggestions for performing Meta-analysis (https://doi.org/10.1093/bib/bbaa019).
Response: We were lucky to have an expertise with deep understanding of biological statistics reviewing our manuscripts. From this question, we as well learned a lot about meta-analysis methods by combining evidences, such as effet sizes and p-values, obtained from multiple experiments. Toro-Domínguez and colleuges reviewed the well-established meta-analysis methods based on effect sizes, P-values combination, and ranks combination. Yoon proposed two variant methods (wFisher and ordmeta) based on the classical Fisher's method, all belonging to the P-values combination methods. We used the wFisher method to re-do our DEG analysis, because the ordmeta method is too time-consuming. All the corresponding modifications are highlighted in the manuscript (Line 142-147). Table 2, Figures 2-5, and Supplementary Tables 6-8 were also revised accordingly.

Minor

(1) Why elastic net penalized regression model was used? have you tried any other methods, if yes could you please provide the comparison with those?
Response: YES. We have tried four model-building methods: generalized linear model (GLM), lasso regression, elastic net penalized regression and ridge regression. The GLM model was conducted by the "glm" function, and the latter three models were conducted by the "glmnet" package of R language. We re-built all four models after revising the DEG and re-renormalization steps. In Appendix Figure 4, the ability to predict the PFS status of NMIBC patients by four methods was shown. Ridge regression, this time, delivered the best performance, so we changed our final modeling strategy to ridge regression instead. Description in the manuscript was also revised accordingly.
However, as we showed in Figure 5 and Appendix Figure 4, the performance of the renewed model declined a lot. We have tried many ways to select significantly correlated genes or try immune cell enrichment methods other than the ssGSEA and Z-score. None of them delivered a good result. Now that we have reached the maximum number of times allowed for self-extension for re-submitting manuscripts. We had to choose our best model at the moment and revise our manuscript according to it.
Nevertheless, we were confused about how the overall outcome is barely satisfactory when every step is better. Please kindly review our results again. We were willing to optimize the model building step if the Reviewer considers it necessary.



GLM

Lasso regression

Elastic net penalized regression

Ridge regression

Appendix Figure 4. ROCs of four model-building strategies in predicting the PFS status of NMIBC patients.

(2) Provide examples of similar models for other diseases, if exist?

Response: Actually, part of our inspiration beginning this work came from the successful use of 21-gene RT-PCR assay (Paik et al., 2004) in predicting breast cancer patients' reaction to chemotherapy and quantifying the likelihood of locoregional and distant recurrence under specific conditions. The 21-gene recurrence score (RS) is calculated by the expression of 16 genes, which is normalized relative to the expression of the five reference genes. The sixteen genes incorporated into the equation are grouped based on function, correlated expression, or both. Apart from GSTM1, CD68, and BAG1, other thirteen genes are classified into proliferation, invasion, HER2, and estrogen groups. Nowadays, as the most validated multigene assays, the clinical application of 21-gene assay is endorsed by the NCCN Guidelines of Breast Cancer.

The second example we want to mention is using an 18-gene T cell-inflamed gene expression profile (GEP) (Ayers et al., 2017) to predict response to pan-tumor drug pembrolizumab. With the development of immune checkpoint inhibitors, there is a need to identify predictive biomarkers for precision medicine implementation. Havel's review (Havel et al., 2019) has summarized various attempts on this task. Building the 18-gene GEP model (Ayers et al., 2017) also includes a normalization step with eleven housekeeping genes showing low variance across samples and a modeling step with eighteen functional genes classified into antigen-presenting cell abundance, T cell/NK cell abundance, IFN activity, and T cell exhaustion. The model was developed and widely evaluated in a series of pembrolizumab clinical trials including Keynote-001, Keynote-012 and Keynote-028. Predictive value was independently confirmed in a pan-tumor circumstances. In the manuscript, we added the GEP study to the Discussion section where we mentioned our limitation of failing to discover predictive values as the lack of therapeutic response data (Line 375-378).

(3) Please provide some comparison of this model with single dimension studies.

Response: Yes. We have added the discussion about the limitation of single dimension studies in Line 364-368.

(4) line 248, "The expression of 72 PFS", how this number is obtained?

Response: The methodology to filter survival-related genes is described in the "Identification of Immune-Cell-Specific DEGs Related to Survival" paragraph of the "Material and Methods" section (Line 148-161). The sheets documenting the results of this step are shown in Supplementary Table 7. 173 and 77 records are left after filtering the table with the green background format, indicating genes representing the candidate immune cells with significant association with PFS and OS, respectively. The corresponding unique genes numbers were 72 (PFS-related) and 34 (OS-related). However, after we re-do the DEGs and re-normalization analyses, the number of genes were changed to 110 (PFS-related) and 41 (OS-related) correspondingly. Please find the renewed results (Line 246) and tables (Supplementary Table 7) in our re-submitted version.

Reference

Ayers, M., Lunceford, J., Nebozhyn, M., Murphy, E., Loboda, A., Kaufman, D. R., et al. (2017). IFN-γ–related mRNA profile predicts clinical response to PD-1 blockade. *Journal of Clinical Investigation* 127, 2930–2940. doi:10.1172/JCI91190.

Havel, J. J., Chowell, D., and Chan, T. A. (2019). The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nat Rev Cancer* 19, 133–150. doi:10.1038/s41568-019-0116-x.

Mooney, M., Bond, J., Monks, N., Eugster, E., Cherba, D., Berlinski, P., et al. (2013). Comparative RNA-Seq and Microarray Analysis of Gene Expression Changes in B-Cell Lymphomas of Canis familiaris. *PLoS One* 8, e61088. doi:10.1371/journal.pone.0061088.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 351, 2817–2826. doi:10.1056/NEJMoa041588.