Dear Editors and Reviewers,

Thanks a lot for spending a substantial amount of time looking over our manuscript entitled "Integrated analysis of single-cell and bulk RNA-seq establishes a novel signature for prediction in gastric cancer". We have carefully read the valuable comments and suggestions from the Editors and Reviewers and now completed a revision of the manuscript. Below please see the point-by-point response to reviewers' comments. We are grateful to the Editors and Reviewers for their constructive suggestions that have improved both the quality and the clarity of the manuscript. The language editing has been performed by AJE Language Editing. The point-by-point responses to each comment are as below.

Thanks again for your consideration and hope you and the reviewers find our data are convincing and could address their concerns.

Looking forward to hearing from you.

Best wishes.

Sincerely yours,

Xiangjun Jiang, Fei Wen

Reviewer 1:

(1) Please, validate and confirm these findings by utilizing the cancer genomic atlas (TCGA) data.

Response: The STAD dataset in TCGA was used to further verify the prediction accuracy of the gastric cancer prediction model. Because there were few normal gastric tissue samples in the STAD data, normal gastric tissue samples from the GTEx dataset were included. Figure 5 shows that the LASSO method-based gastric cancer prediction model had a high prediction accuracy.

(2) Could the authors highlight the significance of TCGA data to study the complex interaction of immune cells in the tumor microenvironment of

cancer as well as cancer cells? reference: SnapShot: TP53 status and macrophages infiltration in TCGA-analyzed tumors. Int Immunopharmacol. 2020 Sep;86:106758. doi: 10.1016/j.intimp.2020.106758.

Response: Thank you for your valuable feedback. We have carefully read the article and have emphasized in our paper the significance of TCGA data for studying the complex interaction between immune cells in the tumor microenvironment of cancer and cancer cells.

(3) Could the authors discuss the possible mechanisms for these findings?

Response: Our study mainly developed a prediction model for gastric cancer. The model was derived from a set of characteristic genes identified by differential expression analysis between gastric cancer tissue and normal gastric tissue in epithelial cells. Our prediction model demonstrated good predictive performance, indicating the feasibility of constructing prediction models through differential expression analysis of cancer and normal tissue. Moreover, intersecting the differential expression analysis results from bulk RNA sequencing helped to eliminate genes with minor differences, which was beneficial to improve the predictive performance of the model.

(4) Please add a diagrammatic figure to summarize the findings.

Response: Thank you for your valuable guidance. We further validated the prediction accuracy of the gastric cancer prediction model using the STAD dataset in TCGA. The TCGA prediction results of the LASSO model have been added to Figure 5.

Reviewer 2:

(1) What is the purpose of including bulk RNA sequencing data?

Response: The purpose of including bulk RNA sequencing data is to provide cost-effective screening of differentially expressed genes. The cost of single cell sequencing is relatively high, while bulk RNA sequencing is a more cost-effective option. The combination of bulk RNA sequencing with single cell sequencing allows for cost-effective and accurate identification of differentially expressed genes that can be used for clinical diagnosis of gastric cancer.

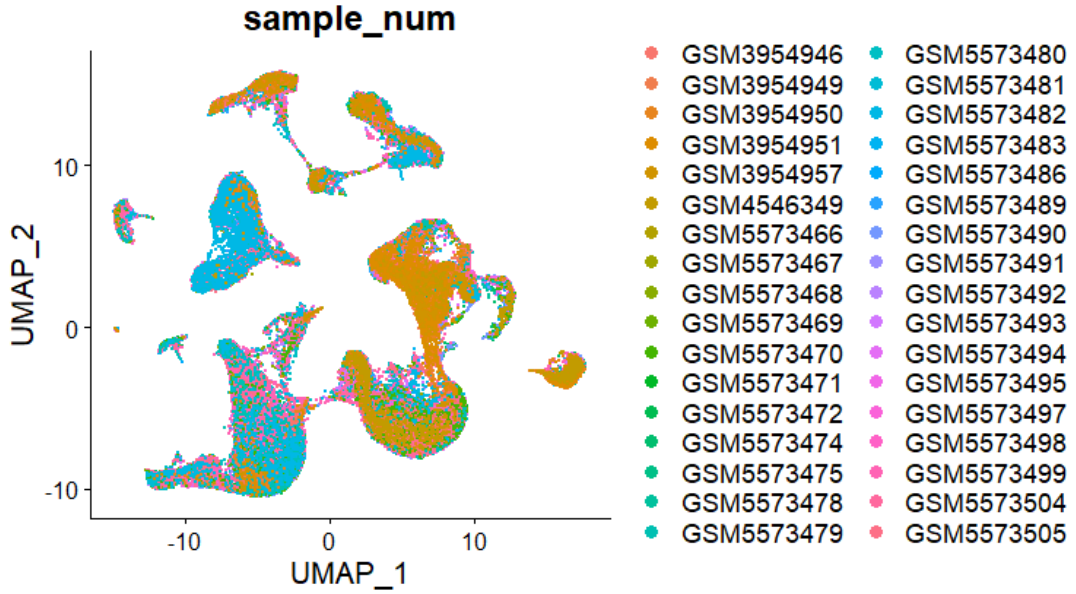
(2) Have you considered building a predictive model by comparing healthy with cancer epithelium from the single-cell RNA seq data? How much difference was there between the classifier derived from the combination of scRNA and bulk and the one derived from the single-cell only?

Response: Thank you for your valuable guidance. Your question is very meaningful. We chose to combine single-cell RNA sequencing and bulk RNA sequencing for feature gene selection mainly because bulk RNA sequencing is more accurate in detecting differentially expressed genes than single-cell sequencing, as the accuracy of single-cell sequencing decreases to some extent

as the sequencing depth increases. On the basis of increasing single-nuclear sequencing-related data, we will consider comparing the classifier derived from single-cell RNA sequencing data with the one derived from the combination of single-cell and bulk RNA sequencing.

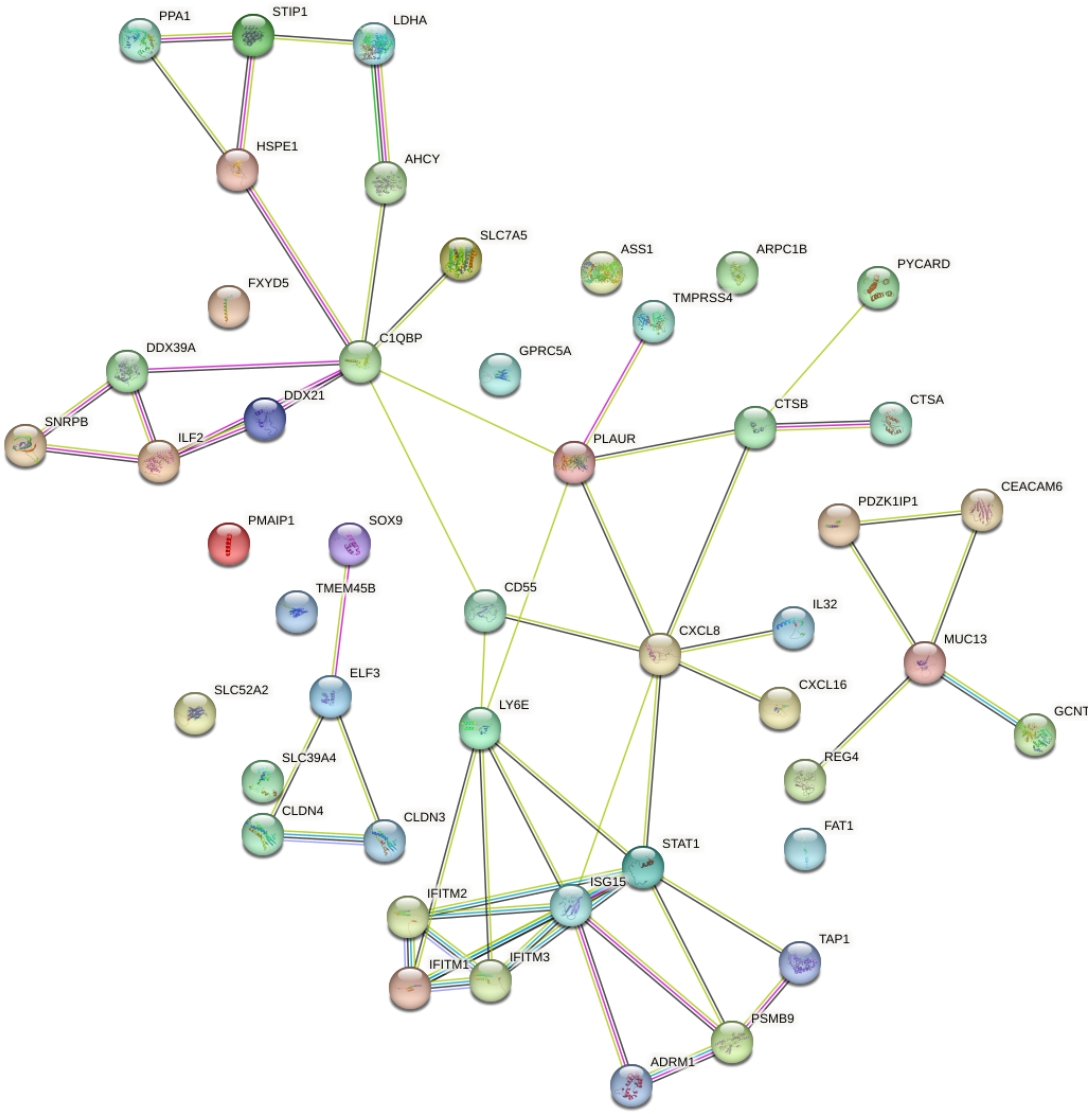
(3) It's known that patient heterogeneity may impact the feature identification, as shown in the bulk RNA sequencing data. Did you notice any batch effects in the single-cell RNA sequencing data?

Response: Batch effects can also occur in single-cell RNA sequencing and were addressed during the Seurat data processing step. For the purposes of this paper, UMAP plots related to batch effects after Seurat data processing are provided below. Due to space limitations, these results were not included in the main text.



(4) In addition, what are the regressed-out features and their biological function, which could be verified by GSEA analysis.

Response: Thank you for your suggestion. Using the LASSO model as an example, the following genes were not included in the model: ADRM1, AHCY, ARPC1B, ASS1, C1QBP, CD55, CEACAM6, CLDN3, CLDN4, CTSA, CTSB, CXCL16, CXCL8, DDX21, DDX39A, ELF3, FAT1, FXYD5, GCNT3, GPRC5A, HSPE1, IFITM1, IFITM2, IFITM3, IL32, ILF2, ISG15, LDHA, LY6E, MUC13, PDZK1IP1, PLAUR, PMAIP1, PPA1, PSMB9, PYCARD, REG4, SLC39A4, SLC52A2, SLC7A5, SNRPB, SOX9, STAT1, STIP1, TAP1, TMEM45B, and TMPRSS4. Further GSEA analysis was conducted on these genes using differential gene expression analysis results from bulk RNA sequencing, but no statistically significant conclusions were reached. Analysis of these genes in the STRING database revealed that they are still enriched in immune response and epithelial composition pathways.



(5) There is a lack of validation of the predicted model. An additional experiment is critical to verify if the prediction model is valid. The possible method includes immunohistochemistry staining, real-time PCR, or in situ hybridization. I would suggest adding some forms of evidence to support the prediction.

Response: We appreciate your valuable feedback and suggestion for including additional experimental validation to support the predictive model. While it is true that validation experiments are important, I believe that the integration of single-cell sequencing with bulk RNA sequencing provides reliable and accurate identification of differentially expressed genes, making the need for additional experimental validation unnecessary. Nonetheless, we will take your feedback into consideration and ensure that any future studies include comprehensive validation experiments to further support our findings.

(6) What are the differences across three scRNA-seq data? I recommend a more detailed discussion on the collected data and the biological differences

between the samples. For example: compiling a supplementary table that includes cancer stages, phenotypes, and prognosis.

Response: The relevant information of the single-cell sequencing data, including pathological type, gender, age, and the source datasets, is presented in the appendix. We included normal gastric tissue and atrophic gastritis samples from the GSE134520 and GSE150290 datasets, as well as gastric cancer samples from the GSE183904 dataset. Our study did not compare single-cell sequencing data across different tumor stages, and specific comparison results have been summarized in Kumar et al.'s article. Therefore, relevant information such as cancer stage, phenotype, and prognosis is not presented. reference: Kumar V, Ramnarayanan K, Sundar R, Padmanabhan N et al. Single-Cell Atlas of Lineage States, Tumor Microenvironment, and Subtype-Specific Expression Programs in Gastric Cancer. Cancer Discov 2022 Mar 1;12(3):670-691.

(7) The clustering and cell identity assignment based on the markers didn't show the cancer populations as the ones in the published articles. How are the labeling differ from the original articles included in the analysis?

Response: Thank you for your suggestion. We have attempted to differentiate cancer populations based on markers. Despite our efforts to ensure the accuracy and reliability of the analysis methods, tumor-related marker genes were expressed in almost all epithelial clusters, and it was not possible to differentiate tumor populations based on these markers. This may be due to factors such as sample sources, quality control standards, sequencing platforms, or data processing methods. We believe that building a predictive model by differentiating tumor cell subgroups is also a feasible method, and tumor subgroups can be differentiated using marker genes or infercnv.

(8) While building the classifier, was there hold-out data (normal cells v.s. cancer cells) to validate the prediction model classified the samples correctly?

Response: We divided GEO data into test sets and verification sets in a ratio of 6:4, and the latter was used to verify the accuracy of the model. Meanwhile, we added TCGA data for verification according to the reviewer's opinion.

(9) The text in the figures (Fig 2, 3, 4, 5) is small. Please ensure that the labels are legible.

Response: Thanks for your suggestion, we have modified the text size of the picture.

Reviewer 3:

(1) Are there controversies in this field? What are the most recent and important achievements in the field? In my opinion, answers to these questions should be emphasized. Perhaps, in some cases, novelty of the recent achievements should be highlighted by indicating the year of publication in the text of the manuscript.

Response: Thank you for your valuable feedback. We have made further revisions to the Introduction and Discussion sections of the article, in an effort to highlight the latest achievements in the field.

(2) The results and discussion section is very weak and no emphasis is given on the discussion of the results like why certain effects are coming in to existence and what could be the possible reason behind them?
Response: Thank you for your advice. We have revised the discussion section.

(3) Conclusion: not properly written.
Response: Thank you for your advice. We have revised the conclusions section.

(4) Results and conclusion: The section devoted to the explanation of the results suffers from the same problems revealed so far. Your storyline in the results section (and conclusion) is hard to follow. Moreover, the conclusions reached are really far from what one can infer from the empirical results.
Response: Thank you for your advice. We have revised the results and discussion section.

(5) The discussion should be rather organized around arguments avoiding simply describing details without providing much meaning. A real discussion should also link the findings of the study to theory and/or literature.
Response: Thank you for your advice. We have revised the discussion section.

(6) Spacing, punctuation marks, grammar, and spelling errors should be reviewed thoroughly. I found so many typos throughout the manuscript.
Response: Thank you for the suggestion. We further polished the language of the article.

(7) English is modest. Therefore, the authors need to improve their writing style. In addition, the whole manuscript needs to be checked by native English speakers.
Response: Thank you for the suggestion. We further polished the language of the article.

Reviewer 4:
(1) Please elaborate on the process, parameters, and results of the variance analysis for scRNA-seq. Setting the criteria of "logFC>0.5 & p<0.05", there are 934 genes left, why exclude lowly expressed genes?
Response: The purpose of our study is to construct a gastric cancer prediction model for clinical application. The limitation of gene expression differences is conducive to the clinical application of immunohistochemical/tissue batch RNA sequencing.

(2) Similarly, please illustrate why exclude lowly expressed genes in differential analysis of Bulk-seq. The volcano map (Figure 3D), on the other hand, marks them clearly.

Response: Thank you for the suggestion. We further polished the language of the article.

(3) The manuscript lacks an introduction to the screening process for so-called important genes.

Response: Thank you for your comments. We describe in more detail the screening process of characteristic genes.

(4) The flow chart seems to fail to show the data analyzing and processing flow.

Response: Thank you for your comments. We modified the flow chart to make it easier for readers to understand the whole research idea.

(5) The first part in the results section duplicates the content of Mast cells, but lacks the content of Chief Cells.

Response: Thanks for your comments, we have modified the recurring mast cells.

(6) The screening criteria of difference analysis for Bulk-seq are inconsistent in the manuscript (logFC＞1.5) and volcano map (logFC＞1). At the same time, the colors in the volcano map seem to be unprecise. Please double-check the volcano map (Figure 3D).

Response: Thank you for your comments. We confirmed the volcanic map.

(7) The supplementary materials are incorrectly marked in the third part of the results section. Table S3 corresponds to prob_1se, while prob_min corresponds to Table S4.

Response: Thank you for your comments. We have confirmed the corresponding order of the forms.

(8) Abbreviations need to be defined when they first appear in the text. The so-called first time here is also calculated separately in the abstract, the text (from the preface to the discussion), each illustration and each table annotation.

Response: Thanks for your comments, we have modified the abbreviation as required.

(9) The results of the two LASSO models were intersected to obtain 11 intersecting genes. Comma segmentation is missing between some genes. Please double-check the manuscript text (including symbols, singular and

plural of words, etc.) to maintain the rigor of the paper.

Response: Thank you for your comments. We have double-checked punctuation, spelling and so on in the manuscript.