Peer Review File

Article information: http://dx.doi.org/10.21037/tcr-22-703

Review comments:

Comment 1: The literature is missing the recent methods that study the prognostic genomic biomarkers for PCa including TNM staging or Gleason score classes. I suggest the authors to highlight PMID: 30890858 and/or PMID: 31835700.
Reply 1: We added the recent methods the two studies used to study the prognostic genomic biomarkers for PCa including TNM stage or Gleason score classes in the Introduction (see Page 5, Line 89-93).
Changes in the text: "To solve this problem, integrated bioinformatics methods such as Robust Rank Aggregation (RRA), ImaGEO, minimum Redundancy Maximum Relevance (mRMR), as support vector machine (SVM), and MetaDE, have been applied in various cancer studies, such as non-small cell lung cancer (NSCLC), cervical cancer, colorectal cancer, Esophageal Squamous cell carcinoma (ESCC) (16-21)."

Comment 2: The author uses many tools/measurements without highlighting why they use them for the non-technical readers. e.g. using ROC without highlighting why to use ROC and/or whats the clinical or computational measurement is here.
Reply 2: We added statement of the significance of the tools/measurements and the reason why we use them as advised.

(i) For the MetaQC method, we added Line 139-143 (Page 7-8) in 2.2 Data processing and quality control.
(ii) For GO and KEGG analysis, we added Line 171-174 (Page 9) in 2.6 GO annotation and KEGG pathway enrichment analysis.
(iii) For ROC analysis, we added Line 219-224 (Page11) in 2.10 The ROC and clinical attribute analysis of the hub genes.

Changes in the text: (number corresponding to reply 2)

(i) "The data quality control (QC) step is vital for bioinformatics analysis, in order to assess the quality and consistency of the datasets and improve the reliability and accuracy of the results. The MetaQC method provides systematic quality assessment of microarray data across studies to decide inclusion/exclusion criteria for genomic meta-analysis."
(ii) "GO annotation analysis provides explain and annotate of gene functions by three dimensions :cellular component (CC), molecular function (MF), and biological process (BP). Meanwhile, KEGG analysis provides the information of the biological pathway the genes participate in."
(iii) "A ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier as a function of its discrimination threshold. And ROC curve analysis has been well established in clinical diagnostic application for evaluating a marker's capability of discriminating between individuals who experience disease onset and individuals who do not (41)."

Comment 3: Optional, but I highly suggest to do the survival analysis using Kaplan-meier tool, it can be done online in databases that are hosted by usegalaxy.org.
Reply 3: In order to maintain the consistency of the analytical methods, we used Kaplan-Meier tool based on survminer package and survival package, which are widely used for survival analysis in R. Also, we added some statement in the explanatory text of Figure 16 to explain the information of the survival curve more clearly (see Page 39, Line872-875).
Changes in the text: "Figure 16. Association between the expression level of KRT5 and MYLK and disease-free survival time in the TCGA-PRAD dataset. The orange line indicates samples

with highly expressed genes, and the green line designates the samples with lowly expressed genes."

Comment 4: Overall, the manuscript requires minor English language editing with a few places where obvious grammatic errors and vague expressions are present
Reply 4: We fixed the gramma errors and vague expressions in the manuscript as advised.
Changes in the text: The changes are too trivial to list here, but can be seen under the "Track Changes" function.

Comment 5: At the beginning of the last paragraph in the Introduction (Lines 82-86), authors provide some vague information regarding the "massive application of the microarry screening" and the Gene Expression Omnibus as "one of the most widely used online gene expression profile databases". This information is not particularly insightful. Thereafter, authors correctly suggest that "The search for tumor-related genes and their related molecular mechanism has extensively involved the use of microarray analysis in pursuit of discovering tumor-specific biomarkers, drug therapeutic targets, and prognosis predictors". Authors should consider replacing Lines 82-86 and provide a statement suggesting that integrated bioinformatics analyses, such as the one embarked by authors, have been systematically applied thus far to derive potential clinical biomarkers and molecular mechanisms in some cancers. This should be accompanied by adding recent literature in terms of advances, notably, doi: 10.3390/biology10111200, 10.1371/journal.pone.0251962, 10.3389/fonc.2021.779042, 10.1097/MEG.0000000000002349, 10.1038/s41598-021-96274-y, 10.3390/genes12091339 .

Reply 5: For Lines 82-86, we have modified our text by replacing a more concise statement as advised (see Page 4, Line 81-83). Then, we added statement suggesting that integrated bioinformatics analyses have been systematically applied to derive potential clinical biomarkers and molecular mechanisms in some cancers, and added recent literature in terms of advanced of integrated bioinformatics analysis as advised (see Page 5, Line 86-95).

Changes in the text:

(i) The beginning 2 sentences of the third paragraph was deleted and replaced by "With the rapid development of high-throughput screening technology, bioinfomatic analysis has become a powerful tool in biomedical field for predicting disease-associated genes, disease subtypes, and disease treatment (14)."

(ii) "However, due to the small sample sizes in individual studies and the use of different technological platforms, substantial inter-study variability and difficult statistical analyses have been generated (15). To solve this problem, integrated bioinformatics methods such as Robust Rank Aggregation (RRA), ImaGEO, minimum Redundancy Maximum Relevance (mRMR), and MetaDE, have been applied in various cancer studies, such as non-small cell lung cancer (NSCLC), cervical cancer, colorectal cancer, Esophageal Squamous cell carcinoma (ESCC) (16-21). These methods can integrate data from different independent studies and obtain more clinical samples for data mining, for ease of achieving more robust and accurate analysis."

Comment 6: At the end of the last paragraph in the Introduction (Lines 90-110), authors chose to summarise the findings of their study. This will appear trivial to readers since the methodology, or the purpose of the study has not yet been discussed. Authors should clearly outline their aim and rationale. To this end, since much of the literature has already explored candidate gene biomarkers in prostate cancer, author should clarify the focus of their study. A proposed aim here is to suggest that there is potentially a scarcity of studies on interaction-based analysis of DEGs in this type of cancer.
Reply 6:

(i) We add the statement that outline our aim and rationale and clarify the focus of our study as advised (see Page5, Line 93-101).

(ii) We rewrote the last paragraph in the introduction by deleting some trivial statement

about the finding of this study (see Page 6, Line103-112)

Changes in the text:

(i) "It's worth noting that although numerous studies have already explored candidate gene biomarkers in PCa, most of these studies merely analyze individual dataset or utilize Venn diagram to directly combine the screened differential expressed genes from different datasets (DEGs), which may overlook some crucial biological information due to the high heterogeneity in PCa (22-27). Thus, we aim to suggest and improve the potential scarcity of studies on interaction-based analysis of DEGs in PCa."

(ii) "In this study, 4 microarray datasets from Gene Expression Omnibus (GEO) database were analyzed. We innovatively combined 2 integrated bioinformatics method MetaQC/ MetaDE and Robust Rank Aggreg (RRA) method to improve the efficiency and accuracy of differential expressed genes (DEGs) screening. After 368 DEGs (120 upregulated and 248 downregulated) were detected, the gene ontology (GO) functional annotation and KEGG pathway enrichment analysis of these genes were performed, and the PPI network of the DEGs was constructed. 11 hub genes were detected from the PPI network and 4 of 11 hub genes CAV1, KRT5, SNAI2, MYLK show potential clinical diagnostic and prognostic value and could be used as novel candidate biomarkers and therapeutic targets for PCa after the survival and clinical attribute analysis."

Comment 7: To address this, authors should expand their analysis by doing an overlap across all 11 topological algorithms from Cytohubba and acknowledge recent advances where this high confidence methodology has been applied (doi: 10.3389/fonc.2021.779042). Instead, authors chose to pursue only 1 algorithm (Degree) and overlap this with 3 other similar ones (Betweenness, Closeness and Stress) and suggested that these have been "most widely used in previous study". This does not particularly augment the rationale behind the methodological selection ensued by the authors. Hence, hub genes here should be oriented based on the overlap of all algorithms, to address all different quantitative aspects of the interactions between the DEGs derived.

Reply 7: We overlapped the Top 25 hub genes (the former was Top 20 hub genes) detected by all 11 topological algorithms from Cytohubba as advised, and screened 11 hub genes which were identified by at least 8 in 11 algorithms (see Line 393-401, Page 18-19). And the results of 3.9 Expression level analysis of the hub genes, 3.10 Association between methylation and expression of hub genes and 3.11 ROC and Clinical attribute analysis of the hub genes, changed correspondingly.

Changes in the text:

(i) "The top 25 hub genes were screened by the Cytohuba plug-in tool in Cytoscape according to the 11 topological algorithms respectively. 11 common hub genes that identified by at least 8 among 11 methods were identified, utilizing online Venn diagram tool (http://bioinformatics.psb.ugent.be/webtools/Venn/) (Table S4). Among the 11 hub genes, VEGFA, VCL, CAV1, KRT5, PTGS2, GJA1, SNAI2, CCL2, CXCL12, and MYLK were down-regulated. However, contrastingly, TWIST1 were up-regulated in primary prostate cancer tissue (Table 4)."

(ii) The content in Table 4 changed correspondingly.

(iii) The results of 3.9 Expression level analysis of the hub genes, 3.10 Association between methylation and expression of hub genes and 3.11 ROC and Clinical attribute analysis of the hub genes, changed correspondingly.

Comment 8: In the Methods, authors provide information that may be quite superfluous and perhaps unnecessary at first sight to readers and an example of this is in the "2.2 Data processing and quality control". Nevertheless, this information is vital but a large portion of it (in this and across all sections of the Methods) could be moved a supplementary Methods document. Authors should be more concise and discuss their initial methodology in summation. Lastly, authors should state how significance was established for each part of the analysis (such as in "GO annotation and KEGG pathway enrichment analysis").

Reply 8:

(i) For the superfluous statement in the Methods "2.2 Data processing and quality control", we deleted the detailed statement of the MetaQC package and moved this information into the added supplementary Methods document as advised.

(ii) Meanwhile, we state the significance of MetaQC analysis as advised (the same as reply 2, see Page 7-8, Line 139-143).

Changes in the text:

(i) Microarray raw data of the 8 datasets was downloaded via txt format from the corresponding platform. The original data of GSE3325, GSE6956, and GSE55945 was gathered by employing log2 transformation using the Limma Package (version 3.40.6) in R (http://www.bioconductor.org/packages/release/bioc/html/limma.html). For the five datasets GSE17951, GSE32571, GSE46602, GSE69223, and GSE89194, the original data was used since the gene expression data has already undergone log2 transformation. Then interquartile range (IQR) method in the MetaDE Package (version 1.0.5) was used to summarize the multiple probes to one intensity (28). The data quality control (QC) step is vital for bioinformatics analysis, in order to assess the quality and consistency of the datasets and improve the reliability and accuracy of the results.

(ii) The MetaQC method provides systematic quality assessment of microarray data across studies to decide inclusion/exclusion criteria for genomic meta-analysis. The full method of data processing and quality control step are shown in the supplementary methods document.

Comment 9: Since authors selected to focus on hug genes and their interactions, they should consider expanding their analysis in terms of functional classification by performing GO and KEGG analysis on the network-based molecular clusters derived. This methodology has been described (https://www.frontiersin.org/articles/10.3389/fnins.2022.915907 and https://www.mdpi.com/2079-7737/10/11/1200).

Reply 9: We performed expanded GO and KEGG analysis on the network-based molecular clusters derived as advised (Page18, Line 311-321). We added Figure 9 and 10 showing the result of GO and KEGG analysis of DEGs in modules. And the detailed result of the GO and KEGG analysis was included in Table S3 and S4. The results in the previous Table 4 (removed) in the text was moved to Table S4.

Changes in the text:

(i) The previous Table 4 in the text was moved to a supplementary result (Table S4).

(ii) "The GO enrichment results (Figure 9 and Table S3) showed that the genes in Module 1 were most enriched with muscle contraction (ontology: BP), cytosol (ontology: CC) and structural constituent of muscle (ontology: MF); and genes in Module 2 were most enriched with glutathione metabolic process (ontology: BP), extracellular region (ontology: CC) and glutathione transferase activity (ontology: MF). Meanwhile, the pathway enrichment results (Figure 10 and Table S4) showed that the genes in Module 1 were principally enriched invascular smooth muscle contraction, focal adhesion, and regulation of actin cytoskeleton. The genes in Module 2 were principally enriched in chemical carcinogenesis, drug metabolism-cytochrome P450, and metabolism of xenobiotics by cytochrome P450."

(iii) "Figure 9. GO enrichment analysis of DEGs in the top 2 modules. (A) The top 14 enriched GO terms of DEGs in module 1. (B) The top 16 enriched GO terms of DEGs in module 2." (see Page 37, Line 832-834)

(iv) "Figure 10. KEGG pathway enrichment analysis of DEGs in the top 2 modules. (A) The enriched pathways of DEGs in module 1. (B) The enriched pathways of DEGs in module 2." (see Page 37, Line 836-838)

Comment 10: The results of the authors in the Discussion should be summarised to avoid redundancy. At present, the Discussion is too long which could impede the understanding of readers. However, the Discussion could be expanded by providing an insight as to how the

derived biomarkers may be implicated in current therapeutic approaches from mechanistic evidence. This would nicely fit into the conclusion, to suggest that prospective experimental validation is required.

Reply 10: We deleted detailed discussion of QC and the individual enriched pathway discussion but keep the summary discussion of enriched pathway to focus on the potentially clinical application of the hub genes. And we expanded the Discussion by providing how the derived biomarkers may be implicated in current therapeutic approaches from mechanistic evidence as advised (see Page 45-46, Line 644-668).

Changes in the text:

(i) Line 585-596 in Page 47 in previous manuscript was deleted.

(ii) Line 612-645 in Page 48-49 in previous manuscript was deleted.

(iii) Line 646-653 in Page 49-50 in previous manuscript was deleted.

(iv) "Caveolin 1 (CAV1) is a carcinogenic membrane protein associated with endocytosis, extracellular matrix tissue, cholesterol distribution, cell migration and signal transduction. Previous studies have found that CAV1 is involved in liver cancer, colon cancer, breast cancer, kidney cancer, lung cancer and skin cancer etc. , and acted as a promoter or inhibitor of cancer according to cancer type and progress (69-71). Multiple endogenous and exogenous agents, such as Chrysotobibenzyl, Cordycepin and Giantol, have been used to modulate CAV-1 expression to regulate lung cancer progression (72-74). KRT5 is one of the human keratin proteins, primarily expressed in epidermal basal keratinocytes (75). Cimpean AM et.al reported that the expression level of KRT5 is in correlation with the prognosis and TNM stage in head and neck squamous cell carcinomas (HNSCC) (76). And Ricciardelli C et.al founded that K5 overexpression in serous ovarian cancer is associated with recurrence and chemotherapy resistance (77). SNAI2 encodes a zinc-finger protein of the Snail family of transcription factors, and plays an important part in epithelial-mesenchymal transition (EMT). Tao F et al. and Tian X et al. reported that the miR-203/SNAI2 axis plays a role in regulating prostate tumor growth, migration, angiogenesis and stemness (78). Meanwhile, the dynamic expression of SNAI2 in prostate cancer can predicts tumor progression and drug sensitivity, and loss of SNAI2 in prostate cancer correlates with clinical response to androgen deprivation therapy (79, 80). Myosin light chain kinase (MYLK) catalyzes the phosphorylation of myosin light chain and regulates the invasion and metastasis of some malignant tumors including lung cancer, colorectal cancer and breast cancer (22). Lin et al. found that MYLK promotes the progression of hepatocellular carcinoma by altering the cytoskeleton to enhance EMT (81). However, the specific role of these genes in the current therapeutic approaches in prostatec cancer is still indistinct and prospective experimental validation is required."

Comment 11: Authors should consider placing the limitations at the end of their discussion and highlight an inherent limitation of their analysis, which underlies (among others) inconsistent TNM staging between the selected studies / datasets, as these were possibly ensued under different classification systems / years.

Reply 11: We added the limitations of our study, which highlighting the inherent limitation of our analysis, at the end of discussion as advised (see Page28, Line525-534).

Changes in the text: "The limitations of our study were as follows: First, our results were not validated at further biological experimental level. Second, the sample size of the involved datasets were comparatively small, and the clinical tumor staging such as TNM stage and gleason score of the selected samples was inconsistent, possibly ensured under different classification systems/ years, which can affect the gene expression in prostate cancer. Finally, our study only focused on the genes which were identified having significant expression level change between cancerous and non-cancer samples in multiple datasets. But we did not consider other characteristics like age, tumor classification and staging. Therefore, some underlying biological information may be neglected in our study."

Comment 12: Lastly, authors should note which TNM stage classification of prostate cancer

they are referring.

Reply 12: We added the clinical and histopathological of the patient cohorts in selected 5 datasets in Table S1 (the information of GSE55946 is not available) (see Page12, Line 245-247). Meanwhile, we only refer to TNM stage of the TCGA-PRAD dataset in 3.11 ROC and Clinical attribute analysis of the hub genes, which under the 7th edition of the American Joint Committee on Cancer TNM staging system. Thus, we added which TNM stage classification of prostate cancer the TCGA-PRAD dataset referring (see Page 12, Line 232-233).

Changes in the text:

(i) "The clinical and histopathological data of the patient cohorts in selected 5 datasets are listed in Table S1 (the information of GSE55945 is not available) (43-45)."

(ii) The TNM stage classification of TCGA-PRAD dataset refers to the 7th edition American Joint Committee on Cancer (AJCC) system (42).