Response to Reviewer 2 Comments

The authors are grateful for your careful reading and valuable suggestions. We worked to improve the points needed for the revisions, with a more comprehensive introduction background, and improved the presentation of the results. Also, we answer specifically each issue raised. We hope this revision meets your expectations and we are at your disposal for further clarifications.

Point 1: The introduction section is rather too small and lacks to explain some important knowledge in the field for the background readers. For example, why ensemble-based feature selection methods are useful, bring some literature using them and more examples of gene signatures in CCRC or related cancer types.

Response 1: We elaborated the explanations needed in the penultimate paragraph of the introduction section:

> Nowadays, the scientific community is still searching for new biomarkers for ccRCC, and feature selection methods using survival analysis provide a robust exploratory methodology before experimental validations. Survival analysis is a field of statistics that predicts the time until an event of interest happens in many domains [15]. The most commonly used method for survival analysis is the Cox Regression model [16]. The Cox model is semi-parametric, that is, the distribution of the event of interest is unknown. In addition, Cox models are widely used for censored data, i.e., when the event is not observed during the study period due to loss to follow-up, study termination, or the patient's death by other causes. Regularized Cox models provide suitable predictions for high-dimensional data using penalty functions with the main regularizers Lasso-Cox, Ridge-Cox, and Elastic net-Cox [15]. Ensemble learning methods are committees of machine learning models, in other words, they combine the majority of the votes for each model in an ensemble or they adjust the weighted vote of each model. Moreover, this approach results in a more robust, efficient, and stable model compared to singular models. In this work, we applied Cox models and ensemble methods using gene expression to predict the overall survival (OS) after diagnosis of ccRCC.
>
> Lasso-Cox regression generated most of the reviewed gene signatures for ccRCC [9,10,13,17–19]. All the studies reviewed in this work use TCGA-KIRC dataset to train and validate the results. Fewer studies validated their results with other datasets such as GEO database [10,13,2], ICGC-RECA [2,11], and data from Fudan University Shanghai Cancer Center (FUSCC). The most common methodologies used to discover and validate gene signatures were differentially expression analysis (DEA), and gene set enrichment analysis (GSEA). Only one study compared its methodology to 3 other biomarker signatures from our literature selection [9]. In addition, there was a lack of comparisons between the gene signatures. As far we know, our study presents a most comprehensive comparison between gene signatures, including ensemble methods, machine learning, and feature selection.

Point 2: In general Figures need to be of better quality and more explanation is needed in the figure legend. In Figure, the authors need to describe the Venn diagram and short description of what the readers should interpret. For example, the overall between all the sets are zero. What does this denote? In addition, there is no explanation about the circo plot? Can this plot be moved to the supplementary? Because Figure 2 has so many plots and seems a bit cluttered.

Response 2: We improve the explanations in the figure legends. Regarding the quality of the figures, we provide the vectorized figures in PDF with high resolution. We changed the disposition of Figure 2 to clarify the Venn diagram and the Volcano plots. We also updated Figure 2a, and followed your suggestion to move the Figure 2d of circo plot to the supplementary (Figure A7).

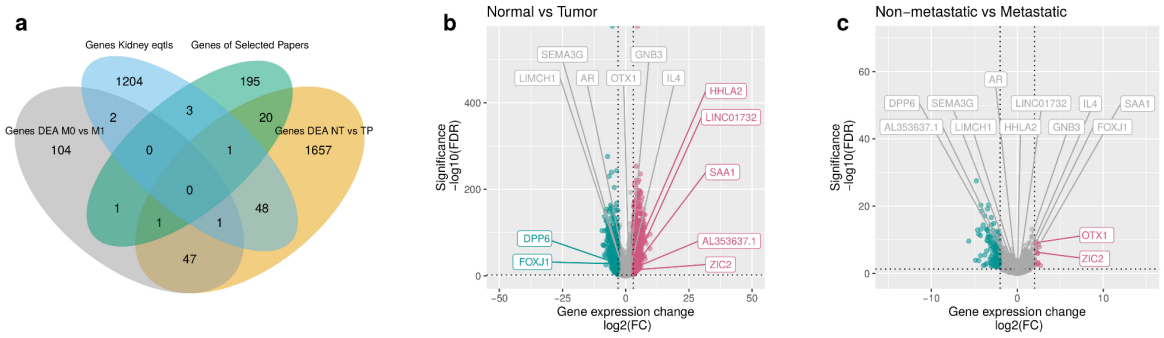

Figure 2. Selected genes through mRMR. (a) Venn diagram of prefiltered gene sets. A total of 3284 prefiltered genes is given by the sets of DEA between non-metastatic versus metastatic (156), normal tissues versus primary tumor (1775), genes from literature (221), significant eQTLs genes (1259), and 124 genes overlapping in two or three intersections of sets. (**b**) Volcano Plot of DEA comparing normal tissues versus primary tumor samples of TCGA-KIRC. In green, we see the downregulated genes of normal tissues versus primary tumors (DPP6 and FOXJ1). In red, we see the upregulated genes (HHLA2, LINC01732, SAA1, AL353637.1, and ZIC2). In gray, we see the not significant genes with low Fold-Change. (**c**) Volcano Plot of DEA comparing non-metastatic versus metastatic samples. In red, we see the upregulated genes (OTX1 and ZIC2).

Point 3: Regarding literature search for gene signatures, why the pubmed search was limited to period of 2015 to 2020? Is there any specific reason?

Response 3: The search on Pubmed was conducted in January 2021, and since this date, we performed the analyses, implemented the models, executed the feature selection methods, produced the figures, and wrote the manuscript. The majority of papers were published in the last five years since 2020, therefore we excluded the period from 2008 to 2014. We added a new plot in Supplementary Figures showing the number of publications by year, and showing that most publications are in the period between 2015 and 2020.

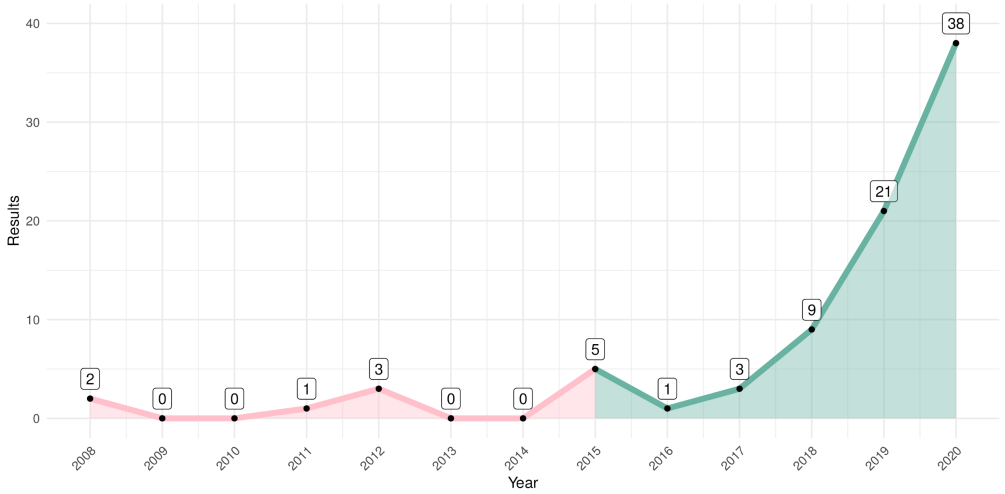

Figure A1. Number of papers published on PubMed by year on query performed in January 2021. Initially, in green, the gene signatures published in the period of 2015 to 2020 were selected to be compared. After the exclusion criteria, we obtained the 14 gene signatures.

Point 4: In the case of feature selection, Bioinformatics, and machine Learning analysis: The package information and methods were explained clearly. However, the count of preselected genes is stated as 3304. But the addition of differentially expressed gene + literature signatures+ eQTLs genes, the total number of genes don't correspond. That is, DEG between normal % tumor is 1775, and between M0 & M1 is 422. Genes from literature is 221 and genes with significant eQTLs are 1259. This total count is 3677. Therefore, authors should clarify how the 3304 genes were selected.

Response 4: We executed a double-check of the number of prefiltered genes, and the total of genes is 3284, obtained by the union of all sets, matching with Figure 2a. A total of 3284 prefiltered genes is given by the sets of DEA between non-metastatic versus metastatic (156), normal tissues versus primary tumor (1775), genes from literature (221), significant eQTLs genes (1259), and 124 genes overlapping in two or three intersections of sets. We include this information in the legend of Figure 2.

Figure 2. Selected genes through mRMR. (a) Venn diagram of prefiltered gene sets. A total of 3284 prefiltered genes is given by the sets of DEA between non-metastatic versus metastatic (156), normal tissues versus primary tumor (1775), genes from literature (221), significant eQTLs genes (1259), and 124 genes overlapping in two or three intersections of sets.

Point 5: There is a difference in the survival time for the two datasets. For TCGA data it is 10 years and for ICGC data it is 7 years. The authors need to clarify how they accounted for the difference in the time between the datasets, especially when considering survival analysis.

Response 5: We restrict the 10-years prediction for TCGA-KIRC to exclude outliers in the long tail of the density plot of the patient's overall survival. For the ICGC-RECA dataset, we decided to maintain a 7-years prediction in order to include all samples of this dataset for external validation. We included this information in the subsection "2.5. Model Evaluation and Statistical Analysis", and in the Supplementary Figures (Figure A6).

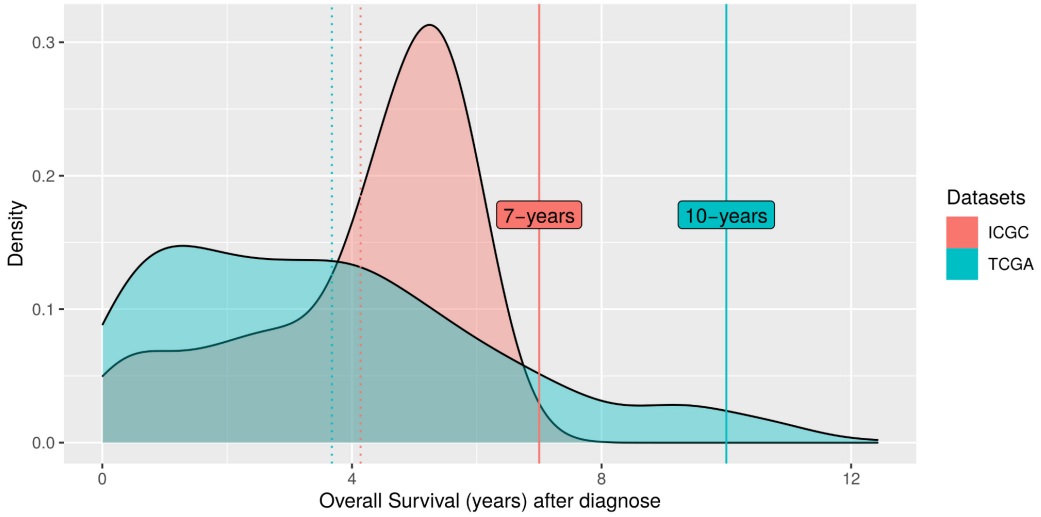

Figure A6. Density plot of the distribution of the patient's overall survival in TCGA-KIRC and ICGC-RECA. The dotted line indicates the mean of distributions, and the solid lines indicate the time prediction used for internal and external validations. We restrict the 10-years prediction for TCGA-KIRC to exclude outliers in the long tail of the density plot of the patient's overall survival. For the ICGC-RECA dataset, we decided to maintain a 7-years prediction in order to include all samples, and limit the time prediction to the range of distribution of this dataset for external validation.

Point 6: The authors should pay attention for grammatical errors and incomplete sentences. For instance, Clinical Characteristics of the ccRCC Cohorts: table of sample comparison shown. The wrong short form is written here, "The mRMR executed a supervised gene selection of 3304 genes with three clinical features: overall survival (OS) days, OS status, age and sex." In this sentence, there is mention of 3 clinical features and 4 features were listed. Authors should also make the citations in the right place. For example, in page 13, in the sentence "Guanine Nucleotide Binding Protein Beta Polypeptide 3 (GNB3) is involved in 462 various transmembra[104]ne signaling systems such as in GTPase activity" the citation is in the wrong place. Similary in the discussion section, the following sentence should be re-written or removed. "As next steps, we are applying this approach of machine learning and feature selection to 500 find potential cancer biomarkers in multiples levels of biological data available in 501 TCGA, such as long non-coding RNAs (lncRNAs), methylation, single-nucleotide 502 variants (SNV), and copy number variants (CNV)."

Response 6: We performed a systematic revision to search for grammatical errors and typos. We also corrected the mentioned items:

- We corrected the sentences in subsections "3.1. Clinical Characteristics of the ccRCC Cohorts" and "3.2. mMRM Gene Selection".
- We removed the misplaced citation.

- About the last sentence in the discussion section, we rewrite as follows:

> In future works, we are expanding the machine learning approach presented in this work to find potential cancer biomarkers using multiples levels of biological data available in TCGA, by analyzing and integrating data of long non-coding RNAs (lncRNAs), methylation, single-nucleotide variants (SNV), and copy number variants (CNV).

Point 7: As the article is not focusing and rather the readers do not know what are the results so far with this type of analysis. Like wise in the conclusion section “We obtained satisfactory results combining RNASeq data (ICGC and TCGA), survival information, and machine learning strategies.” This sentence is too general, authors should write what is satisfactory?

Response 7: We have modified the conclusion by adding the following clarifications:

> Our main goal was to compare distinct gene signatures from literature and generate new gene signatures using feature selection methods. We contributed by providing a list of new genes, some of them not previously reported as biomarkers for ccRCC. The gene signature created by the mRMR method achieved a score of 0.82 with AUC, being the best performer. We identified two clusters of genes with high expression (SAA1, OTX1, ZIC2, LINC01732, GNB3 and IL4) and low expression (AL353637.1, AR, HHLA2, LIMCH1, SEMA3G, DPP6, and FOXJ1) which are correlated with poor prognosis. We validated our 13-gene signature for ccRCC and confirmed our results with the literature, and by comparing each cancer stage of ccRCC with CPTAC and the survival effects of gene expression of individual genes in TCGA. We believe that further studies on the involvement of these genes in renal carcinogenic processes can improve the understanding of cancer biology. After experimental validations, new possible applications in clinical practices can benefit from the biomarker found with machine learning and feature selection.