Author’s response to reviews

Title: Immunogenomic Pathways Associated with Cytotoxic Lymphocyte Infiltration and Survival in Colorectal Cancer.

Authors:

Yuanyuan Shen (yshpf@mail.missouri.edu)
Yue Guan (guanyue@health.missouri.edu)
Justin J. Hummel (jjh42v@mail.missouri.edu)
Chi-Ren Shyu (shyuc@missouri.edu)
Jonathan B. Mitchem (mitchemj@health.missouri.edu)

Version: 1 Date: 30 Nov 2019

Author’s response to reviews:

27 November, 2019

Ozgur Kutuk
Editor
BMC Cancer

Dear Professor Kutuk:

We would like to thank you for the opportunity to revise our manuscript entitled “Immunogenomic Pathways Associated with Cytotoxic Lymphocyte Infiltration and Survival in Colorectal Cancer” by Yuanyuan Shen, Yue Guan, Justin J. Hummel, Chi-Ren Shyu, and Jonathan B. Mitchem. The reviewer comments are insightful and have helped us to improve our work. We appreciate their consideration. We have answered the comments and have outlined the changes below.

Reviewer 1:

Major concerns:
1. Paper needs to be revised to not have jargon and make more broadly understandable to the reader. Problem areas include:
   a) Non-standard abbreviations should be spelled out throughout. It is very difficult to read with so many. Like CRC and TCGA (standard); MCP, RSC, DEGs and many more not standard.
We apologize for any confusion from the use of abbreviations. In this revision, we have changed non-standard abbreviations to the long-form in the abstract and have continued to use some standard abbreviations (TCGA and CRC). In the text, we have used only the abbreviations utilized by Becht, et al in the paper describing the Microenvironment Cell Populations-Counter paper when it seems appropriate to match the figures included in the paper. In the figures, we have used the abbreviations but have been more specific in the description to avoid making the figures more difficult to understand.

b) What is Level 3 mRNA data?

The Cancer Genome Atlas provides data in 5 levels and level 3 data refers to “aggregated data” and is downloaded as .txt file as Reads Per Kilobase of transcript, per Million mapped reads (RPKM). This is the publicly available data (https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/data-levels). We have clarified this in the manuscript and also as a part of addressing “other suggestions” #2 below. We have also provided a new reference regarding TCGA workflow for further clarification (Han, et al. Journal of Neuro-Oncology. (2018) 136:463–474).

2. It is not clear if there is a circular argument and/or the input and output are not independent. Could you discuss if the genes that make up differential expression are the same or similar genes that are used to define CL using the MCP Counter method? In particular, "Immunoregulatory interactions between lymphoid and non-lymphoid cell”. Specifically, what genes defined the CL abundance and how do these compare to the genes differentially expressed in your analysis.

The cytotoxic lymphocyte gene enrichment set includes CD8A, EOMES, FGFBP2, GNLY, KLRC3, KLRC4, and KLRD1. The list of these genes is now included in the methods section. As suggested by the reviewer, we also were concerned regarding the ability of these genes to influence pathway enrichment analysis. Therefore, none of these genes were included in the pathway enrichment analysis. This has been clarified in the methods.

3. Related to #2, would be helpful to provide more detail about CL classification using MCP-Counter method. What goes into the classification, how different are tissues as far as CL presence? How powerful is this classification for the data set you are using - do tissues fall into clear groups. What is the median CL score and distribution for your data set?

The gene panel used for the delineation of the cytotoxic lymphocytes is now included in the methods to provide more detail regarding how the abundance score is created and we have added the paper reference here as well as outlined above. We have only included colorectal cancer tissues in this study, so there was no comparison to other tissues. In the original paper describing the MCP-Counter method, the authors use mixing experiments and immunohistochemistry to validate the association of this gene panel with cytotoxic lymphocytes. This included the use of different leukocyte mixtures, as well as mixture with fibroblast and cancer cell lines.

The median and quartiles are included below in table format and will be included in Supplemental Table 1 and the median included in the results. After separating the patients into High and Low groups based on the median score, a t-test was performed to confirm there was a significant difference in cytotoxic lymphocyte scores between CL-high and CL-low groups.
There was a significant difference in cytotoxic lymphocyte scores between CL-High and CL-Low groups (73.66± 64.2 v 14.07± 6.69, p<&lt;0.0001, Supplemental Figure 1) This is now included in the manuscript as Supplemental Figure 1. This analysis is also included in the results.

Supplemental Table 1. Distribution of cytotoxic lymphocyte abundance scores including median with 95% CI and quartiles for the entire patient cohort.

<table>
<thead>
<tr>
<th>Score (95% CI)</th>
<th>Minimum:</th>
<th>1.785</th>
</tr>
</thead>
<tbody>
<tr>
<td>25th Percentile:</td>
<td>13.95</td>
<td></td>
</tr>
<tr>
<td>Median:</td>
<td>26.72 (24.1-30.1)</td>
<td></td>
</tr>
<tr>
<td>75th Percentile:</td>
<td>50.61</td>
<td></td>
</tr>
<tr>
<td>Maximum:</td>
<td>546.3</td>
<td></td>
</tr>
</tbody>
</table>

4. Is the differential expression in right sided cancers (hi vs lo CL) due to the MSI tumors. Is the differential expression still present if the MSI tumors are excluded?

We agree with the reviewer that there is a possibility for differences dependent on whether MSI-H patients are included or not. Unfortunately, the limited number of patients do not allow for the ability to compare patients based on stage and location while including microsatellite status as a variable. However, MSI-H patients made up only 41% of the CL-high patients in the right sided cancer group (Table 1), the group with the vast majority of these patients. Recent studies have demonstrated that microsatellite status may not be the best predictor of patient outcome or response to immune based therapy (Mlecnik, et al, Immunity. 2016 Mar 15;44(3):698-711; Fabrizio, et al, Journal of gastrointestinal oncology, 2018. 9(4), p.610). Additionally, in the recent consensus molecular subtypes paper, there were MSI patients included in two subtypes suggesting heterogeneity among these patients (Guinney, et al, 2015 Nat Med 21(11):1350-63). Given that cytotoxic lymphocytes are the critical mediators of anti-tumor immunity and the limitations discussed above, we chose to include these patients together. To further clarify our reasons for doing the analysis in this way, we have provided further support for this rationale in the background and discussion sections.

5. Could you expand upon survival analysis? Could you graphically represent gene expression vs survival? What do you consider significantly different survival? Possibly report statistics supporting the genes reported in Table 2.

Survival analysis was done using the Kaplan-Meier method and assessed by Cox regression model. P values derived from Kaplan-Meier analysis are now included in Table 2. Representative survival curves are now included in Figure 3 and the remainder of survival curves and analysis have been uploaded as a supplementary data.

6. How many genes are in the Immunoregulatory pathway (Table2); what is the denominator?

Two hundred ninety-seven genes are in this pathway. We have included this in the results.
7. Table 2, Pathways column is not helpful. If you had a companion model of CL interacting with non-lymphoid cell, might make some sense.

We have eliminated that column from Table 2.

8. Shorter discussion - more to the point.

We have revised the discussion including removal of some elements as suggested below to make this more succinct.

9. The CD48 discussion is confusing. Since it is differentially expressed in both metastatic groups - but opposite relationship to survival.

Since the submission of our original manuscript, CD48 has been removed from this pathway by Reactome during their dataset update process. We have removed this from the discussion.

Other Suggestions
1. Need to better define early, localized, metastatic. There is overlap in staging. Are you using specific TNM and not stage? What is it?

In this study, we used the TNM classification system. This is provided by The Cancer Genome Atlas data. We did include patients with Stage I and II in the “early” group, patients with Stage I, II, and III in the “localized”, and Stage IV as “metastatic”. Patients with Stage III disease were combined with Stages I and II for analysis of “localized” disease to increase patient numbers. This does create some overlap with the “early” stage group as suggested. The Stage III group; however, did constitute 40-60\% of patients in the “localized” group depending on location, so we felt this remained an important analysis. The patient grouping is included in the methods and we have expounded on our reasoning in the discussion.

2. Looks like you used RSEM normalized RNA-seq data, but talk about download of data off of Illumina HiSeq and Illumina-GA (raw data). Not sure how and when this downloaded other data is used. If not used, no need to mention.

RSEM normalized data was used for MCP-counter correlation analysis and survival analysis. Illumina-HiSeq and Illumina-GA raw counts were used for the differentially expressed gene analysis as the raw count data from HiSeq and GA are needed for the edgeR analysis. This is updated in the methods.

3. Make sure there is sufficient detail in methods that someone could reproduce your analysis.

Our hope is to be transparent and allow our methods to be utilized by others. Through the above mentioned changes, we have updated the methods sections to more accurately reflect our process. Additionally, Figure 1 outlines the process utilized in this manuscript. We have also included further supplemental files including interval analyses based on the suggestions of Reviewer 2.
Reviewer 2:

Major concerns:
1. The methods section for RNA-seq DEG analysis would indicate that only DEGs with positive (?) fold change of $>$1 were identified, so that only those expressed higher in CL-H than CL-L (?) were included in the further analysis. If so, why were the genes that have higher expression in CL-L were ignored? This needs clarification. Also, the reference 12 here does not seem right.

Genes were considered to be differentially expressed with absolute fold change $>$1 and we apologize for not including this originally. We have changed the methods to reflect this. Additionally, the reference noted is incorrect. Thank you for bringing this to our attention. The references have been rechecked and made consistent with the manuscript.

2. Based on Figure 2, it is not clear how the authors conclude that "the p adjust value and the ratio of DEGs in each pathway, we found that "Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell" was the most significantly altered pathway at all three locations and at every stage"? Based on the Figure 2, it seems the "GPCR ligand binding" would be the strongest pathway for Early and Localized cancers and the "Signaling by Interleukins" the strongest for the Metastatic cancers? The 'Immuno-pathway' is among the 3 top ones for each, but not the top one? Further clarification is needed or better wording, such as noting only the enrichment in this pathway in all stages, as stated in the Conclusions.

We agree with the reviewer regarding the interpretation of the pathway enrichment analysis and a more accurate depiction of the data would be to state that the “Immunoregulatory interactions” pathway is the pathway that is most highly enriched and included at all sites and stages. We have changed the manuscript to reflect this more accurately as suggested.

3. The authors list many links in the "Availability of data and materials", but none of them seem to contain any processed raw data, such as gene lists with fold changes (or gene ratios) and p-values (all of which would clarify the points 1 and 2 above). The processed raw data/results should be provided as a supplement.

The results from the differentially expressed gene, MCP-Counter scores, pathway enrichment, and Kaplan-Meier analysis are now uploaded as supplementary data.

4. Since these patients were treated with various regimen (and not with immunotherapy), is the survival analysis and the associated DEG genes (as potential targets to improve immune based therapy?) in this context meaningful? The authors may want to elaborate on this a little more, the current narrative is very short.

We agree with the reviewer regarding the heterogeneity of treatment and the lack of use of immune based therapies in this patient group. Data generated from patients receiving immune based therapy pre- and post- therapy (possibly during) would be the most ideal as has been demonstrated in patients with melanoma and non-small cell lung cancer. Unfortunately, to the best of our knowledge, this data currently does not exist in colorectal cancer. In this case, we have tried our best to look at a surrogate marker for response to therapy, cytotoxic lymphocyte
infiltration. This has been demonstrated in studies in CRC to be predictive of response to both immune based therapy and conventional therapeutic methods. For this reason, we think this type of analysis remains meaningful as we do further studies. We have expounded on this in the discussion.

Minor edit suggestions:
- Please spell out CL-L (=CL-Low) on the abstract

This is completed. Also addressed as indicated by Reviewer 1.

- READ should be rectum adenocarcinoma, not just rectal cancer (=REC) on the Abbreviations and Figure 1 legend. Also, COAD is colon adenocarcinoma.

This is corrected.

- Typo: Illumine-ga in methods.

Corrected. Thank you for bringing this to our attention.

- Spell out TME in methods

This is completed.

- Add microsatellite analysis-related acronyms under the Table 1, and also explain how the p-value was derived.

Abbreviations are now included in Table 1. We have included the following regarding the generation of p values in patient demographic variable in the methods: “Patients' basic clinical features were summarized by descriptive statistics, including means and standard deviation, and an unpaired t-test was used for normally distributed continuous data. Categorical variables were compared using Fisher’s exact and chi-square tests. A p value < 0.05 was considered statistically significant.”

- Can you explain "p adjust value", results, page 8.

The “p adjust value” refers to the p value adjusted for false discovery rate. We have changed this in the text to read “p value adjusted for false discovery rate” to eliminate confusion.

- Check for a typo ("posit") on page 9, row 51.

We have removed this section from the discussion while addressing comments from Reviewer 1.

Again, we would like to thank you and the reviewers for your time and consideration. We look forward to your future response.

Sincerely,