Author's response to reviews

Title: Systematic genomic identification of colorectal cancer genes delineating advanced from early clinical stage and metastasis

Authors:

Hojoon Lee (hojoon.lee@stanford.edu)
Patrick Flaherty (pjflaherty@wpi.edu)
Hanlee P Ji (genomics_ji@stanford.edu)

Version: 3
Date: 15 November 2013

Author's response to reviews: see over
November 14, 2013

To the Editors of BMC Medical Genomics,

We are submitting our revised manuscript titled “Systematic genomic identification of colorectal cancer genes delineating clinical stage and metastasis” for consideration. With the advent of genomic medicine, there is increasing and substantial interest across the biomedical research community in identifying cancer genes relevant to both biology and clinical phenotype.

We appreciate the constructive and helpful feedback on our manuscript. We are extremely encouraged that both reviewers found the article to be meritorious and of interest. We have revised the manuscript to affect the changes suggested by each comment. The manuscript now is much improved as a result.

We have provided a point-by-point response to the reviewers comments and also noted the revisions based on their suggestions. Our responses are listed on the following pages.

Thank you for your considering our revised manuscript.

Sincerely,

Hanlee Ji, M.D.
Responses to the Reviewers

Dr. Ramon Salazar’s Comments

1) I find the integrative approach provided in this paper of great value and highly intuitive, it can represent a methodological break through in the validation of elastic net regularized regression gene centered methods, but its current clinical application is nule and influenced by the wrong definition of advanced disease where lymph node metastasis is made equivalent to distant metastasis, which is clinically a very inappropriate definition.

We appreciate the reviewer’s corrections of our clinical staging description and have corrected it accordingly. As the reviewer has noted, the distinction between lymph node metastasis and distant metastasis is important. To clarify this major point, we have carefully defined our terminology in both the title and throughout the main text. We group stage I and II together and call it early stage. We group stage III and IV together and call it advanced stage rather than metastasis throughout the manuscript. We have changed the title of the manuscript to reflect this distinction as well to “Systematic genomic identification of colorectal cancer genes delineating advanced from early clinical stage.”

2) On page 17 when the authors try to explain why there are no genes that delineate M status independently, the most important weakness of the whole paper, they claim that this indicates that lymph node metastasis is a sufficient, early indicator for metastatic behaviour. This is a wrong statement and, curiously, the authors correct themselves in the next sentence where they accept the smaller number of stage IV samples may have affected their sensitivity for identifying genes associated with distant metastasus in other organs.

We thank the reviewer for noting the ambiguity of this part of the manuscript. We have improved this section by tightening definitions and removing ambiguity in our results discussion. We deleted the sentence on page 17 that that lymph node metastasis is a sufficient, early indicator for metastatic behaviour. Instead, we added a sentence that lymph node metastasis is critical to advanced stage compared to T status as expected by AJCC staging method.

3) Further down on the las line on page 17 thay claim that dominant inverse association in T status and clinical stage suggest that loss of or lower expression of genes is a more frequent in advanced stage clinical disease, which counterdicts the previous statement describing the frequent direct association in N status and MSI implies that the gain or higher expression genetic aberrations occur more frequently in advanced clinical stage,

We agree with the reviewer that our discussion lacked clarity. Accordingly, we have revised this section of the discussion to improve clarity in the main text and briefly describe those changes here. We find that the set genomic predictors for each clinical parameter differ. While, initially, this observation is surprising to us, it makes sense that the clinical behavior and thus certain clinical parameter may be influenced by a different
set of pathways. The set of predictors reflects this. We added text on page 18 clarifying that all of the genomic features common in both clinical stage and N status showed the same association direction.

Dr. Iris Simon’s Comments

1) Our main critic of the analysis is that there is no independent validation. Public datasets could have been used to confirm loss/gain of expression of the top genes, copy number, methylation etc, even if only one of the genomic data forms is available.

We appreciate the reviewer’s criticism of this point of the paper. Throughout the analysis we made concerted effort to identify a comprehensive genomic/clinical data set with an adequate number of samples. Surprisingly, nearly all of the studies either were restricted to a single genomic microarray platform (e.g. gene expression) and lacked sufficient numbers with nearly all of them being under 100 samples. As a solution to the dearth of multiplatform genomic data with high quality clinical annotation, we provided bootstrap analysis for this purpose. To enhance our point, we added the publications that showed the loss/gain of copy number, methylation of top gene (WRN) at page 16.

In response to the reviewer, we identified an independent copy number data set from TCGA for independent validation involving an independent 354 samples. These samples lacked data from any other platform. From this independent data set, we confirmed that MALT1 was a hit seen in both analyses. We have augmented our discussion section (page 22) to describe this clearly. In the future, we will use additional TCGA data sets that have annotated clinical outcome data. The completion of the TCGA project is anticipated at the end of 2014.

2) page 1, paragraph 2: metastatic CRC usually refers to stage IV; therefore the description of stage I and II should be “early stage” CRC and stage III + IV should be called “advanced” or “late stage” cancer...

We appreciate the difficulty the reviewer finds with the current clinical delineation of colorectal cancer. We share in this struggle to accurately define phenotype – especially in the absence of clinical outcome data in the TCGA data set. While there are smaller data sets with clinical outcome data available, none have the number of samples and breadth in terms of assays (gene expression, copy number variation, etc) that the TCGA provides. **We have changed our terminology throughout the paper as the reviewer suggested to advanced disease rather than metastatic.** We expect to have more genomic data from metastatic samples by TCGA that enables us to identify the genes associated with distant metastasis.

3) The number of stage IV patients is unfortunately low (as in most public studies) and there is no analysis of primary tumor versus metastasis. That limits the conclusion for the analysis of metastatic behavior.

Indeed, the number of stage IV patients is low, as with most genomic studies with publically available genomic data sets. By grouping stage III and IV patients as advanced
disease, we improve the sample size at the risk of increasing within-group heterogeneity. **We state this important caveat in the discussion section at page 20.** For a future analysis, we will pursue additional metastatic samples for inclusion. The TCGA project will be completed at the end of 2014 and will provide more data with additional Stage IV samples.

4) **It is not obvious why this 30-gene signature was selected to represent genes for aggressiveness, considering how many signatures have been published in the last years. It would have been more appropriate to select signatures that had independent validations.**

We appreciate the reviewer’s astute commentary based on their gene expression literature. The majority of the previously published gene signatures were based on gene expression data and did not incorporate other cancer genetic features such as mutations or methylation. Our analysis considered these events as well as gene expression and so was likely to pick up different events. In addition, there are a limited number of gene expression analyses looking at metastasis. For example, Shibayama et al. published a review article focused on metastatic gene signatures and covering clinical significant gene sets derived from gene expression studies (Cancers 2011, 3, 2858-2869). Only two studies among the 15 they cited had samples from all clinical stages. For example, large study from Genomic Health looked at Stage II and III colorectal cancers and not Stage IV.

For our study we spent considerable time and effort comparing, contrasting and evaluating different gene sets for inclusion in the initial set that played a direct biological role in cancer and its pathogenesis. Ideally, the elastic-net regularized regression would handle all of the features and smoothly incorporate only those that have a significant improvement on prediction accuracy. But, we and others have found that inclusion of too many irrelevant features leads to performance degradations of that algorithm. Even so, elastic net is the state of the art for regularized regression and we find it effective if the set of genes is appropriately sized.

We chose the genes for inclusion in our initial set using two data sources: COSMIC and TCGA. We chose COSMIC because it has been both curated and validated for identifying genes of immediately relevance for cancer. However, we were concerned that by imposing such high standards on genes that make it into the initial set, we might miss out on genes that are important, but not yet validated. To address this issue, we added genes sets obtained from the TCGA group reasoning that not doing so would ignore the initial analysis that group performed and would risk leaving out important predictors that have not yet been validated. **This selection strategy yielded an initial genes set that is not too constrained as to miss many important predictors, yet not so large that it would degrade the performance of the elastic net.** We have augmented our methods section to clearly describe our selection strategy on page 4.
The authors describe the analysis of MSI-patients using the elastic-net analysis. The results confirm known genes as the main markers of tumors with MSI-status. It would be interesting to know if also some patients with MSS status show the same pattern. And, could the method be used to identify or characterize patients whose MSI-status is not known?

As the reviewer astutely noted, we can characterize the MSI-status (MSS, MSI-L, MSI-H) of patients with our regression model. When we made prediction of MSI-status of 198 samples, regression model correctly characterized MSI-status of 181 samples (93.0%). In this analysis, we misclassified one (out of 33 MSI-L) to MSS, one (out of 24 MSI-H) to MSI-L, and 12 (out of 141 MSS) to MSI-L.

Based on the mean squared error in table 3, using gene expression alone (0.799) has almost the same predictive power as the integrative genomics approach (0.782). Does this indicate the predictive power of integrative genomics approach is essentially driven by gene expression data?

We agree that the predictive power of integrative genomics approach is essentially driven by gene expression data. However, prediction is not main goal of our study. We aimed to identify critical genes that are supported by more than one genomic biomarker as well as understand the additional information added by heterogeneous data. Several top candidates such as GNAS (supported by methylation and mutation) and FHIT (supported by methylation and mutation) would not be detected if the analysis was solely based on gene expression data. We have modified our discussion to state again the main aim of our study and to point out this modest increase in predictive power over gene expression at page 19.

WRN might indeed be an interesting gene for further translational studies as the ReqQ helicases play a role not only in tumor progression but also in response to chemotherapy. Note: WRN is a marker of CIMP molecular subtypes and therefore described in several colorectal cancer studies. (e.g. Am J Pathol. 2010 Dec;177(6):2731-40).

We greatly appreciate that the reviewer has drawn our attention to this study that includes WRN as a component of a subtype profile. We have cited this study in the results (p. 16). It improves the link with previous research and ensures that we properly acknowledge important supporting evidence.