Reviewer's report

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

Version: 1 Date: 8 October 2013

Reviewer: Iris Simon

Reviewer's report:

Review by Iris Simon (Sr. Director R&D) and Sun Tian (Sr. Bioinformatician)

Systematic genomic identification of colorectal cancer genes delineating clinical stage and metastasis
Lee HJ, Flaherty P and Ji HP

The paper describes the integrative analysis of the publically available TCGA dataset to identify genes that are predictive for stage and aggressiveness of colorectal cancer tumors. The TGCA data consist of gene expression data (by microarray and RNAseq), micro RNA analysis, copy number alteration, mutation analysis and methylation data. Such diverse genomic datasets from the same samples are becoming more and more available, and therefore such integrative analysis is very important and relevant. It is interesting to see the extend (or sometimes lack of) interdependence of the different methods. We appreciate that the authors make many of the interim analysis and results available as supplementary information and on their web page (even though that makes reading the manuscript a bit fragmented).

The integrative genomics approach using elastic-net and combined ranking score is valid and well described.

Overall the manuscript describes a new method for integrative analysis but does not validate the clinical and biological findings in depths.

Major Compulsory Revisions

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.

Some additional comments:

Introduction

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.

The view is also a bit too static, as stage alone is not fully describing the aggressiveness of a tumor. There are stage II patients with a very aggressive
disease who will have metastasis within the next 3 years and there are stage III tumors that grow very slow and are curable.

Unfortunately, the TCGA patients have no outcome data and this is a major disadvantage of this study. Stage I-III patients should be separated in good outcome versus poor outcome, rather than by stage alone.

Methods

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.

Page 4, paragraph 1. The authors' initial input includes 1192 genes and cross validation is evaluated on TCGA dataset. These 1192 genes can be considered as 2 parts:

Part 1: 484 genes are from COSMIC + 20 genes are from previous knowledge. The pre-selection of these genes in part 1 is independent from statistical test performed on TCGA data.

Part 2: 32 mutated genes are from TCGA + 353 genes associated with copy number changes in TCGA dataset + a 30 genes expression signature associated with tumor aggressiveness in TCGA dataset + a 344 genes methylation signature in TCGA dataset. The pre-selection of these genes in part 2 is already dependent on the statistical test performed on TCGA data, the dataset on which cross validation will be performed. Even the statistical model used by author (elastic-net + combined ranking score) is different from the statistical model used in original TCGA dataset, a pre-selection of genes entering into cross validation procedure involve the dataset used for evaluating cross validation itself will introduce bias.

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.

Results:

Page 11/ 12: 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?

Page 14, paragraph 3 and table 3. 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?

Discussion

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).

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests