Author's response to reviews

Title: Integrated network analysis and logistic regression modeling identifies stage-specific genes in Oral Squamous Cell Carcinoma

Authors:

Vinay Randhawa Mr (vinay.plp@gmail.com)
Vishal Acharya Dr (vishal@ihbt.res.in)

Version: 3  Date: 27 May 2015

Author's response to reviews: see over
ANSWERS TO REVIEWERS' COMMENTS

EDITOR'S RESPONSE

(1) Since the main argument of the authors is that they have identified a prognostic cancer signature, this should ideally be evaluated in an independent dataset. If this is not possible the authors should amend their claim that this a prognostic gene signature, and revise the manuscript accordingly. For example, in this case, prognostic should be removed from the title as this is misleading.

Response: Considering the suggestions made by editor and both reviewers, we have validated the model on an external independent dataset. After careful search, we obtained an external dataset (GEO ID: GSE41613) to validate the classification model. This independent dataset is measured with Affymetrix array and containing expression and cancer stage related phenotypic data of HPV-negative OSCC stage samples (54 late- and 41 early-stage). The GEO dataset was assessed for outlier samples and initial pre-processing was performed as described for all earlier used datasets. The model constructed using merged dataset was evaluated on this preprocessed and annotated (independent) OSCC dataset. A reasonable ~61% accuracy (Q) was obtained using our predicted model in an independent dataset. However, we believe that with an increase in the sample size, there might be likely chance of obtaining higher accuracy with the selected 13 key hub genes.

The term “prognostic gene signature” and “prognostic” has been removed from the manuscript title and has been modified accordingly. Most importantly, the title of the article has also been modified accordingly. Also, the text has been rephrased at several sections of the revised manuscript.

(2) The level of english is not up to publication quality in places and should be proofread and the grammar corrected where necessary.

Response: As per suggestion, the revised manuscript has been carefully proof-read and corrected for grammatical mistakes.
(1) Answers to Reviewer- Mark Ibberson

Major Compulsory Revisions

1) In Figure 2 the effect of batch correction on the PCA is striking, but it is unclear if there still remain any batch effects since the points are too small and the batch differences can’t be seen very clearly. It would be clearer to make the points larger and color them according to batch since this is what you are trying to correct for.

Response: The impact of batch effect on genomics study might be minimized; it may not be completely controlled, defined or eliminated even with perfect design and best effort in all stages of experiment. In our analysis, COMBAT implementation have removed dataset-specific biases and eliminated batch effects to a greater extent, though not eliminated completely. We assume that all the variance associated to batch effect that could be eliminated has been taken care of by COMBAT. The same has been described in the revised manuscript accordingly.

Considering the reviewer’s suggestion, image (Figure 2) and corresponding text has been modified accordingly in main manuscript. Now, the data points in Figure 2 have been made larger and colored accordingly.

Relative Log Expression (RLE) boxplots assess global bias in samples are also proposed to assess batch effect removal methods with great accuracy [1]. Therefore in order to probe deeper, RLE boxplots (Additional file 2: Figure S4) were also assessed to explore the effect of batch removal. The RLE plot highlighted the existence of 7 clear batches in the simply combined datasets (without COMBAT process) (Figure S4-A). However, COMBAT implementation greatly improved the appearance of RLE boxplots (Figure S4-B). Additionally, compared to simply combined dataset, the mean of RLE plot for COMBAT processed dataset was distributed around zero and had almost similar spread for all samples which indicates a good batch effect removal using COMBAT.
2) Correlations of the modules to only a single measure (cancer stage) is shown (Figure 4A). Do the modules correlate to any other measures? For example, are there any correlations to technical measures such as publication date of the gene expression or other batch related parameters? This is important to check and would add to the confidence that these modules are really correlated to biological measures.

**Response:** We agree with the reviewer's suggestion that correlating modules with other phenotypes or technical measures would add up to the confidence level of modules. Since the phenotypic description of grade or type was not available for the samples, we were not able to consider the same for our analysis. However, since we have obtained significant gene ontology and pathway enrichment for the representative genes, we therefore considered the pink module to be of statistical and possibly biological significance and therefore used in our analysis. Despite the reviewer's suggestion, we were not able to correlate modules to technical measures as suggested to further enhance and boost our manuscript.

3) The classification accuracy of the gene signature was assessed by cross-validation using 20% of the data as the test and 80% as training set. The reported accuracy is therefore only applicable to the merged dataset used in the study. In order to be correct the classification accuracy needs to be tested on an independent dataset. Why did the authors not set aside one or 2 public datasets for this purpose from the start of the analysis?

**Response:** We thank reviewer for constructive suggestions regarding model construction and validation. Fortunately, an external dataset was obtained for validation of our model obtained. We feel that in the original manuscript, we were not able to clearly elaborate the method regarding “feature selection” and “classification model” and same has been modified in the revised manuscript.

For reviewer's concern, in the present study, we have considered elastic net implemented in “glmnet” which has been used for two purposes: (1) variable (feature) selection, and (2) classification model.
(1) Variable selection - To reduce the dimensionality in feature space (genes), ‘glmnet’ was applied to perform elastic net feature selection using linear regression modeling. This procedure predicted a best subset of hub genes having strongest correlations to phenotype. At this step feature space was reduced from 63 hub genes to 13 key hub genes.

(2) Classification - After performing feature selection, a classifier model was also built (classification problem) with ‘glmnet’ and tested using 5-fold cross-validation using above-mentioned 13 gene features (key hub genes). We randomly stratified original merged data into equal-sized and non-overlapping sets where 70% was used for training and remaining 30% for testing. For the first fold dataset, a model was constructed and tested on non-overlapping test set. Similarly, training and testing was carried out five times using one distinct set for testing and other four sets for training. The classification accuracy (AUC) of five generated models was found to be 0.88, 0.73, 0.85, 0.84 and 0.72, with an average of ~0.81.

So, a final classification model was build to determine whether identified 13 genes were able to discriminate between early- and late-stage OSCC samples. We believe that we have used standard procedure for feature selection and classifier model construction. However, we still agree with the reviewer’s concern about the features identified. Therefore, we have validated the model on an external independent dataset. After careful search, we have obtained an external dataset (GEO ID: GSE41613) for validation. This independent dataset is measured with Affymetrix array and containing expression and cancer stage related phenotypic data of HPV-negative OSCC stage samples (54 late- and 41 early-stage). The GEO dataset was assessed for outlier samples and initial pre-processing was performed as described for all earlier used datasets. The model constructed using whole merged dataset was evaluated on this preprocessed and annotated OSCC dataset. Using our model, a reasonable ~61% accuracy (Q) was obtained in an independent dataset. However, we believe that with an increase in the sample size, there might be likely chance of obtaining higher accuracy with selected 13 key hub genes.

The term “prognostic gene signature” and “prognostic” has been removed from the manuscript title and has been modified accordingly. Most importantly, the title of the article has also been modified accordingly. Also, the text has been rephrased at several sections of the revised manuscript.
4) The manuscript is poorly written in places and difficult to follow. It should be proof-read and corrected by a native English speaker
Response: As per reviewer’s suggestion, the entire manuscript has been proof-read carefully and also corrected for grammatical mistakes.

Minor essential revisions

1) Results (1st paragraph): There are 4 major steps outlined in the text, but only 3 in the corresponding Figure 1.
Response: We agree with the reviewer’s suggestion and text has been modified accordingly in the revised manuscript.

2) Line 300: ‘At a glance, clustering analysis of genes showed a distinct separation between two groups’. Figure S5 is unclear and difficult to see anything. Please comment and explain what you mean.
Response: We agree with the reviewer’s suggestion and image (Figure S6) and corresponding text in the main manuscript has been modified accordingly. We have added a horizontal side bar in which OSCC samples are color coded. Since we have incorporated one more supplementary figure, now Figure S5 has been changed to Figure S6. Appropriate changes have been mentioned in main manuscript.

3) It should be made explicit in the main text that only tumor expression data was used for the module identification.
Response: We agree with the reviewer’s suggestion and text has been modified accordingly.

4) Figure 6 needs more description and explanation
Response: We agree with the reviewer’s suggestion and text corresponding to explanation of Figure 6 has been modified accordingly.
(2) Answers to Reviewer (Eugenia Migliavacca)

Major Compulsory Revisions

1. Line 300: Because the genes clustered in Figure S5 showed a clear separation between the two groups, it does not imply that these genes are of biological significance. These genes were selected to separate the 2 groups by applying LIMMA and not surprisingly they separate the 2 groups. At this point of the analysis we know about their statistical significance based on the LIMMA results, but we do not know about their biological significance.

Response: We agree with the reviewer's suggestion that genes obtained after implementing t-test are statistically significant but may not be of biological significance. A student t-test may infer little if anything about the biology [2], however fold-change on the other hand lends itself to more biologically meaningful assessment of genes [3] [4] [5]. Overall, we assumed that simultaneous implementation of both p-value and fold change may result in more biologically meaningful sets of genes. As per suggestion, Line 300 and the associated text has been modified in the revised manuscript to suit accordingly.

2. Line 324: The subtitle 2.2 should be changed and should not contain “in independent datasets”. To my understanding, the authors randomly selected 100 samples from the original samples, therefore there are no independent datasets. The authors showed the robustness of the modules defined in 2.1 by resampling but not the preservation in independent datasets.

Response: We agree with the reviewer's suggestion that we have performed analysis by resampling procedure rather than using independent dataset for computing module preservation. This text has been rectified and subtitle 2.2 has been modified accordingly in the revised manuscript.

3. Line 494: There is a fundamental problem with the main statement in this paragraph. To my understanding, the 5-fold cross validation was performed only at the last stage of the
classifier (model building, C). The final classifier is based on the results of the 3 stages depicted in Figure 1, therefore these results are not reflecting a “true” cross validation process since the labels regarding all samples have been already used in the first part of the procedure (steps A and B). The authors should add an “external” loop of cross validation starting ideally before the merging of the data or at least before the non-specific filtering of the genes with low variability. Alternatively they should validate the proposed model in an external data set, if available.

**Response:** We thank reviewer for constructive suggestions regarding model construction and validation. Fortunately, an external dataset was obtained for validation of our model obtained. We feel that in the original manuscript, we were not able to clearly elaborate the method regarding “feature selection” and “classification model” and same has been modified in the revised manuscript.

For reviewer’s concern, in the present study, we have considered elastic net implemented in “glmnet” which has been used for two purposes: (1) variable (feature) selection, and (2) classification model.

1. Variable selection - To reduce the dimensionality in feature space (genes), ‘glmnet’ was applied to perform elastic net feature selection using linear regression modeling. This procedure predicted a best subset of hub genes having strongest correlations to phenotype. At this step feature space was reduced from 63 hub genes to 13 key hub genes.

2. Classification - After performing feature selection, a classifier model was also built (classification problem) with ‘glmnet’ and tested using 5-fold cross-validation using above-mentioned 13 gene features (key hub genes). We randomly stratified original merged data into equal-sized and non-overlapping sets where 70% was used for training and remaining 30% for testing. For the first fold dataset, a model was constructed and tested on non-overlapping test set. Similarly, training and testing was carried out five times using one distinct set for testing and other four sets for training. The classification accuracy (AUC) of five generated models was found to be 0.88, 0.73, 0.85, 0.84 and 0.72, with an average of ~0.81.

So, a final classification model was built to determine whether identified 13 genes were able to discriminate between early- and late-stage OSCC samples. We believe that we have used
standard procedure for feature selection and classifier model construction. However, we still agree with the reviewer’s concern about the features identified. Therefore, we have validated the model on an external independent dataset. After careful search, we have obtained an external dataset (GEO ID: GSE41613) for validation. This independent dataset is measured with Affymetrix array and containing expression and cancer stage related phenotypic data of HPV-negative OSCC stage samples (54 late- and 41 early-stage). The GEO dataset was assessed for outlier samples and initial pre-processing was performed as described for all earlier used datasets. The model constructed using whole merged dataset was evaluated on this preprocessed and annotated OSCC dataset. Using our model, a reasonable ~61% accuracy (Q) was obtained in an independent dataset. However, we believe that with an increase in the sample size, there might be likely chance of obtaining higher accuracy with selected 13 key hub genes. Most importantly, the title of the article has also been modified accordingly. Also, the text has been rephrased at several sections of the revised manuscript.

Minor Essential Revisions

4. Line 59: the number should be reported as 260,000 new cases and 124,000 Deaths

Response: We agree with the reviewer’s suggestion and text has been modified accordingly.

5. Line 85: “to only few orders of magnitudes” I would specify the order of Magnitude

Response: We agree with the reviewer’s suggestion and text has been modified accordingly.

6. Line 91: The sentence starting with “But these studies” is unclear

Response: We agree with the reviewer’s suggestion and text has been modified and rephrased accordingly.

7. Line 239: It is not clear what does it mean “initial number or replicated measurements”.

Response: Non-relevant text has been removed from the main manuscript and changes have been incorporated in Table 1.
8. Line 247: The number of discarded (or retained) chips should be reported for each study, maybe by adding a column in Table 1
Response: We have considered reviewer’s suggestion and same has been incorporated in Table 1.

9. Line 257: The authors should clearly specify if they are considering 355 late stage OSCC and 131 early stage OSCC, or if the 355 tumor samples there are both early and late stage carcinoma.
Response: We considered reviewer’s suggestion and text has been modified accordingly clearly mentioning the sample type and numbers.

10. Line 277: Figure S3. The labels are not readable. The authors could add a horizontal side bar in which each study is color coded.
Response: We agree with the reviewer’s suggestion and image (Figure S3) and text in the main manuscript has been modified accordingly.

11. Line 300: Figure S5. The labels are not readable. The authors could add a horizontal side bar in which different OSCC stages are color coded. The heatmap is very dark, the color scale should be changed from -20, 20 to a more appropriate interval.
Response: We agree with the reviewer’s suggestion and image (Figure S6) and text in the main manuscript has been modified accordingly. Since we have incorporated one more supplementary figure, Figure S5 has been now changed to Figure S6. Appropriate changes have been mentioned in main manuscript.

12. Line 346: Sign greater or equal is not correctly printed
Response: The text has been modified accordingly.

13. Line 346: Are the p-values adjusted for multiple testing? If yes, please specify by which method, if not, please report the adjusted p-values.
Response: p-values had not been adjusted for multiple testing initially. But considering the reviewer’s suggestion, for each correlation, p-values computed are now being adjusted for
multiple testing corrections using the Benjamini & Hochberg method for generating FDR adjusted p-value (q-value). In the revised manuscript, p-values are now replaced by adjusted p-values.

14. Line 358: "Pink module (114 genes) was therefore finally selected based on strongest correlation with stage and its apparent statistical importance." Pink and black modules have the same correlation and the same statistical significance, but the pink module has a slightly higher MS. The sentence should be rephrased.
Response: Accordingly, we have mentioned in the main manuscript that both modules have the opposite correlation with positive (0.33) for pink and negative (-0.33) for black with stage phenotype. And also the text has been rephrased and modified to highlight further importance of pink module.

15. Line 388: Sentence not clear.
Response: We agree with the reviewer’s suggestion and text has been modified accordingly.

16. Line 397: Sentence difficult to understand
Response: We agree with the reviewer’s suggestion and text has been modified accordingly.

17. Line 438: Sentence should be modified "... biological process, molecular function, and cellular component [40]."
Response: We agree with the reviewer’s suggestion and text has been modified accordingly.

18. Line 1158: Signs greater or equal are not correctly printed
Response: Accordingly the signs are “greater than (>)” only and not “greater or equal” in the main manuscript and carefully checked.

19. In the supplementary tables S3 and S4 p-values should be reported with not more than 3-4 significant digits
Response: We considered reviewer’s suggestion and text in “tables S3” and “table S4” has been modified accordingly.
References


