Reviewer's report

Title: Derivation of a Bronchial Genomic Classifier for Lung Cancer in a Prospective Study of Patients Undergoing Diagnostic Bronchoscopy

Version: 2  Date: 1 April 2015

Reviewer: Denis Bertramd

Reviewer’s report:

The paper presents a new classifier based on gene expression able to identify cancer patients after a bronchoscopy. 299 broncho epithelial cells were collected from 223 patients diagnosed with cancer and 76 patients diagnosed with a benign disease. Using the expression data, the author described how they select a small set of 23 genes to construct a classifier that could be used to predict patient with cancer. Then they test the classifier on an independent data-set of 123 patients.

Major Compulsory Revisions

* The authors did not give a convincing explanation on how the markers could be used to identify cancer patients. Indeed, the result on the test data-set shows a high sensitivity but a very low PPV (0.47) which mean that a large fraction of patients will be wrongly predicted to have cancer. I would suggest to discuss those results in more details and explain how the 23 markers and the associated classifier could be used in a clinical setting.

* The test(s) used for the p-value computations should be described on the text.

* I understand that using genes which expression is correlated with clinical variables, like smoking status and pack years, could give more information to detect cancer patients. But I am not convinced that a gene associated with patient gender could help to identify cancer patients. What is the effect of using the patient gender instead of the RPS4Y1 expression on the result of the classifier?

* The 3 sub-sections in the “CLASSIFIER DEVELOPMENT” section should be expended to provide more details about the methods used.

* Page 9: “First, the HG-U133A RMA expression values were adjusted by a gene-wise constant which shifted the mean of each gene’s expression levels in the test set to the mean observed in the training set”.

How the constant have been computed?

* Page 9: “Second, for the classifier genes where a corresponding HG-U133A probeset was not available (LYPD2 and RNF150), the gene’s mean expression value in the training set was used for all of the test set samples.”

As it is not possible to evaluate the contribution of the genes LYPD2 and
RNF150 to identify cancer patient on the test data-set, I would suggest to remove them from the classifier. The result on the test data-set should be similar, and this will lead to a shorter list of markers.

Minor Essential Revisions

*page 5 line 95: “Additional file 10”
The additional file number should be ordered in the order that they appear on the text.

*page 11 line 217: What is T score ?

*page 11 line 228: “entire training set..” -> “entire training set.”

*page 15 line 312: “However, the differences in the specific genes selected this may be due to differences in the feature selection process.”
This sentence is not clear and should be rewritten.

*page 19: the table 1 legend seems to be truncated.

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