Peer Review File

Article information: http://dx.doi.org/10.21037/atm-20-2047

Reviewer:

Comment 1: Wang et al. has proposed a method to identify blood biomarkers for the inference of tumors. The idea is interesting and of clinical significance.
Reply 1: Thank you for your comments.

Comment 2: However, generally speaking, the description of results was hard to follow. For example, what did the authors mean by "enriched the blood-tissue related genes" on line 134? What is the result of this step? A supplementary table may need to be provided.
Reply 2: We thank the reviewer for the comments and valuable suggestion.
The Genotype-Tissue Expression (GTEx) project collected RNAseq data of ~17,000 non-disease samples, involving ~50 tissues, including whole blood, lung, breast, etc. Firstly, using data from GTEx, we analyzed the Spearman correlation between whole blood samples and the other tissues respectively for each gene and each tissue. As a result, the blood-non blood tissue correlated genes for each tissue were calculated. Secondly, survival hazard genes and differential expression genes of each cancer type in The Cancer Genome Atlas (TCGA) were selected by Cox regression model and Wilcoxon rank sum test, respectively (Figure 1, Supplementary File 1). Only primary tumor tissues were used for Cox regression. And the differential expression genes were calculated between tumor and normal tissues (if exist). For example, we selected ovary cancer to test our workflow. We defined our inferred blood biomarkers as genes which show their significance in both correlation analysis and Cox regression analysis (false discovery rate, FDR<0.2), including positive hazard genes and negative hazard genes (Figure 2a). For instance, Kruppel like factor 13 (KLF13) gene expression shows a significant correlation between ovary and blood. Meanwhile, Cox regression reveals that KLF13 is a significant survival risk factor in ovarian cancer. Combining these two results, KLF13 is screened out to be a potential up-regulated prognostic blood biomarker for ovarian cancer. In other words, it is reasonable to think that high expression

of KLF13 in ovarian tumors means a reduced survival time, and that higher KLF13 expression may also be detected in the blood. Therefore, by measuring the expression of KLF13 in the blood, the prognosis of ovarian cancer may be evaluated.

Changes in the text: We have extended the 'Identification of potential positive and negative blood biomarkers' section of RESULTS and the 'Blood biomarker screening' section of MATERIALS AND METHODS. And a supplementary data (Supplementary File 1) was provided.

Comment 3: Another example is that the sentence at lines 159 to 163 had too much information to understand.

Reply 3: We defined a combined p value, by multiplying the p values of two steps, to assess the potential of a gene as a blood biomarker. Take the inference process of lung cancer prognostic blood biomarkers as an example, we firstly calculated the expression correlation between blood and lung for each gene using GTEx RNAseq data. For the second step, we collected the survival hazard genes of Lung squamous cell carcinoma (LUSC) by Cox regression using TCGA RNAseq data. The genes with $p \geq 0.05$ in any steps were excluded and the combined p values were calculated.

Besides the p values, the directions were also be considered. The genes with rho>0 in Spearman analyses and hazard ratio (HR)>1 in Cox regression, or rho<0 in Spearman analyses and HR<1 in Cox regression were considered as positive (or called: hazard, risk, up-regulated) blood biomarkers. Oppositely, the genes with rho>0 in Spearman analyses and HR<1 in Cox regression, or rho<0 in Spearman analyses and HR>1 in Cox regression were considered as negative (or called: risk-free, down-regulated) blood biomarkers.

Then we defined several gene sets. For example, the 'Lung LUSC Cox Positive 20' gene set means: The Spearman tests were used between blood and lung data of GTEx, and the Cox hazard genes were calculated using LUSC data of TCGA, and the positive blood biomarker genes with top 20 smallest combined p values were selected. For another example, the 'Breast BRCA Diff. Negative 50' gene set means: the Spearman tests were used between blood and breast data of GTEx, and the differential expression genes were calculated using breast invasive carcinoma (BRCA) data of TCGA, and the negative blood biomarker genes with top 50 smallest combined p values were selected. The gene sets used in this paper can be found in Supplementary File 2.

Changes in the text: We have extended the 'Differences of blood sample between solid tumor patients and healthy donors using BB Infer genes' section of RESULTS. And a supplementary data (Supplementary File 2) was provided.

Comment 4: When mentioning differential analysis, a description of the groups used should be clearly stated. For example, was it normal vs tumor, or was it one tumor vs the rest of the tumors. As far as I know, in PDAC dataset, there is no healthy samples, therefore I was curious how did the authors performed their analysis.

Reply 4: The differential analyses were calculated between tumor and normal samples. If a cancer type does not have a normal sample, this step (the right way in Figure 1) will be skipped. But we can still infer prognostic blood biomarkers by Cox regression (the left way in Figure 1).

Changes in the text: We have extended the 'Identification of potential positive and negative blood biomarkers' section of RESULTS and the 'Blood biomarker screening' section of MATERIALS AND METHODS.

Comment 5: The Methods part in Abstract might need heavy revision, so that it won't be including details like which package or versions used, but instead has an overall summary of the methodology logically presented.

Reply 5: We are sorry for this mistake. The Abstract have been revised. We hope it could meet the requirements.

Changes in the text: We have revised the Methods part in Abstract.