Author’s response to reviews

Title: Unsupervised gene expression analyses identify IPF-severity correlated signatures, associated genes and biomarkers

Authors:

Yunguan Wang (Yunguan.Wang@cchmc.org)
Jaswanth Yella (Jaswanth.Yella@cchmc.org)
Jing Chen (Jing.Chen2@cchmc.org)
Francis McCormack (mccormfx@ucmail.uc.edu)
Satish Madala (Satish.Madala@cchmc.org)
Anil Jegga (anil.jegga@cchmc.org)

Version: 1 Date: 02 Sep 2017

Author’s response to reviews:

Reviewer #1 (Remarks to the Author):

In "Abstract", it said "The differential expressed genes among all six subgroups of IPF ...". Is this accurate? It seems that most DEGs are between subgroups versus control, instead of among subgroups.

Answer: We thank the reviewer for pointing this out and this has been now fixed in the “Abstract”.

P7, Lines 26 -31 "Expression data containing only IPF samples were transformed using PCA to reduce the dimensionality, ..." However, the PCA plot (Figure S2) contains both IPF and Control, not "only IPF". Please clarify.

Answer: We thank the reviewer for pointing this out and apologize for the confusion. PCA for figure S2 were done in the following steps. The principal components (PCs) were calculated using only 131 IPF samples, and then the 12 control samples plus 131 IPF samples were projected on these PCs. Additional details on PCA can be found at Scikit-learn documentation for PCA at http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html#sklearn.decomposition.PCA.transform). In this way, we ensured the PCs used for IPF patient clustering and for figure S2 to be the same. In the revised version, we clarified this in the method section (under section, “Clustering, principle component analysis (PCA), and differential expression analysis).
P7: Lines 31-33 "The number of clusters was chosen as the smallest number that allowed maximal difference in average FEV1, FVC, and DLCO among clusters." Are the IPF clusters identified from PCA or hierarchical clustering? Further elaboration is required.

Answer: IPF patient clusters were learned using hierarchical clustering followed by PCA. A series of cluster numbers were tested. For each cluster number, average FEV1, FVC and DLCO of each cluster were calculated. The number of clusters was then chosen as the smallest number that allowed maximal difference in average FEV1, FVC, and DLCO among clusters. We have clarified this in the revised manuscript (see previous comment).

P9, Lines 41-43: "We performed Ward clustering followed by PCA on the gene expression profiles of 131 UIP/IPF patients, and identified 6 distinct patient clusters (C1 through C6) (Figure S2c)." It's unclear what criteria in PCA or hierarchical clustering are used to identify these 6 clusters in Figure S2C, since they are not completely separated in the PCA plot. Are they identified based on gene expression profiling, or based on clinical parameters (i.e. DLCO, FVC, FEV1…), or both?

Answer: We apologize for the confusion. IPF patient clustering was completely based on expression data. Clinical parameters (DLCO, FVC or FEV1) were only used for validation purpose. Regarding figure S2c, we agree that c6 was not separated from c3 on the first two principal components. However, this does not indicate that c3 and c6 are not separable if more principle components are used. In our study, we used all principle components for clustering analysis.

P10, line 16-19: "To examine transcriptomic differences between these patient clusters, we performed differential analysis using the R package 'limma' and identified 2968 DEGs (Table S2)." Are the 2968 DEGs identified based on comparison of each subgroup with control? It should be indicated here and in Table S2, if that's the case.

Answer: We thank the reviewer for pointing this. The 2968 DEGs identified were based on comparison with control. In the revised version, we have clarified this.

Figure 4:

1. What parameter is used to score each patient for ROC analysis? The Y-axis on the right panel is illegible due to the low resolution.

Answer: The Y-axis is true positive rate and in the revised version of the manuscript, we have corrected this (both Figure 4 and Figure S2).

2. The sample information of the validation dataset is not consistent with original data. For example, GSE10667 has 15 controls in original dataset. However, In Figure 4, it shows "12 Control". Please verify the datasets and redo the statistics if required.
Answer: We apologize for this typo. GSE10667 indeed has 15 control and we have corrected this in the revised manuscript.

3. The prediction accuracy is great between IPF and normal control (Figure 4), but poor between IPF and explant (Figure S5). Further analysis or explanation is required to explain.

This would be a limitation for the application of the gene signature, and thus need to be discussed.

Answer: The advanced IPF gene set was not able to differentiate normal IPF from acute exacerbate IPF, but achieved 0.71 and 0.83 accuracy in the other validation sets with IPF biopsies and IPF explants. We believe such performance is reasonable.

Reviewer #2 (Remarks to the Author):

Minor suggestions:

While the identified gene signatures did an exemplary job at identifying IPF from normal and severe from mild IPF, it will be very helpful both diagnostically and experimentally if these gene signatures can differentiate IPF from other ILDs. If possible, it will be great if such results can be added to the manuscript or at least discussed in the discussion section.

Answer: We thank the reviewer for this suggestion. As part of our ongoing studies, we do have preliminary results from using the core IPF geneset to differentiate IPF from other ILDs. The preliminary results were promising but not confirmative and need additional extensive analysis. Further, the primary focus of this study was to identify subgroups within IPF. Hence, we did not include the comparison of IPF with other ILDs.

Page 16, line 25: MMP7 is listed twice.

Answer: We apologize for the error (typo). We have corrected this (MMP3 instead of MMP7) in the revised version.