Reviewer 2 v.1

Comments to the Author

Therapeutic Advances in Respiratory Disease Review

“NF-kappa B signal transduction through the NFKB1-GSK3B

bridge and a series of inflammatory pathways that mediate sepsis-associated ARDS”

Zhang et al have studied sepsis-induced ARDS and lung injury using publicly-available microarray datasets from the GEO repository. To minimize batch effects, they used batch correction and used PCA analysis to evaluate data quality. The authors proceeded to perform differential expression and GSEA enrichment and constructed a weighting model using forest tree, STRING and functional enrichment methods to identify pathways. Finally, they used the "TRRUST" database to identify regulatory interactions. The authors found enrichment of immunological genes. The genes enriched for cell-substrate adherens junction and genes involved in the NF-kappa B signal transduction. Overall, this is an interesting paper that adds new knowledge to the field and to understanding genomic basis for ARDS due to sepsis. Issues that limit enthusiasm and need to be addressed are listed below.

1. The concept is appealing, however, the strategy and the methods are confusing without the clear rationale for the different technologies that were applied.

2. The authors used five different expression arrays, four from Affymetrix (GSE76293 HG-U133_Plus_2, GSE66890 HuGene-1_0-st, GSE10474 HG-U133A_2, GSE10361 HG-U133A) and one from Illumina (GSE32707 GPL10558 Illumina HumanHT-12). It is unclear if the arrays were combined or analyzed independently. This needs to be clarified. The section of the batch effect describes using “combit package” a package that does not exist in any of the R repositories. Furthermore, the details on how the data was analyzed were not described in detail or missing, like the case of the GSE32707 arrays.

3. There are variabilities between the datasets that can’t be overlooked for example, GSE76293 originated from polymorphonuclear neutrophils (PMNs) from bronchoalveolar lavage and blood, while the rest of the samples are from whole blood. The samples were from 2007 to 2016 at multiple centers, and the potential differences in diagnosis between centers need to be discussed.

4. What phenotypes were compared for each dataset and analysis? What phenotypes were used for the GSEA and differential expression? This requires clarification.

5. There is inconsistency with p-values described as P-value < 0.05, in others as p-value < 5% and in other instances they used FDR, but the FDR was not present in the tables (adj.P.Val is not necessarily FDR, need to see the commands used in limma). During differential expression analysis, a threshold

that combines FDR and fold change is usually utilized. For example, an absolute value of logFC of > 1 and FDR of < 0.01 is commonly used for gene expression. The tables values of logFC < 1.0 are listed as differentially expressed whereas they are not.

6. The random forest analysis general information about the method, but it is not clear what was the input for the algorithm. An input table is necessary.

7. The authors need to provide the software used. They indicate using the method “train_test_split” for the training and testing, did they used the python scikit-learn package or some other software? The same critique is for the protein-based network identification and the pathway-enriched method. Figure 1 helps to understand the general approach, but the method section should be rewritten to provide all the scripts and tables used at each step in a GitHub repository.