Identification of Sources of Platform Specific Bias in Single Cell RNA Sequencing

Rohan Verma, Nikila Joshi, Ziyu Ren, Paul Reyfman, Scott Budinger, Alexander Misharin

Northwestern University, Division of Pulmonary and Critical Care, Feinberg School of Medicine, Chicago, IL, USA;

ABSTRACT

Single cell transcriptomics is a powerful tool for unbiased marker-free discovery of the new cell types and their activation states. Here in addition to a quick overview of single cell RNA sequencing, we report identification of the systematic bias in detection of specific genes and, using computational and statistical approaches, demonstrate how this bias originates during the data acquisition, propagates through bioinformatics pipelines and affects estimation of the differentially expressed genes. Our findings are of high importance for the large scale integrative studies, such as Human Cell Atlas project. We also propose computational approaches for mitigating this bias.

METHODS

• Datasets: human cell datasets from 10x Genomics (SC030) and others from Northwestern Division of Pulmonary and Critical Care lung transplant data (SC02 NextSeq SC14/15 HiSeq).
• Programs: It was used for all data analysis and for single cell data analysis and visualization the package Seurat was used alongside the goatips package to produce plots.
• Libraries were prepared following standard protocol found on 10x website before sequencing on both NextSeq and HiSeq machines.
• The 10x cellranger pipeline was used in each case to perform alignment and initial filtering/quality checking to produce filtered matrices for further analysis.
• The established workflow from the Seurat package was for all samples analyzed using identical parameters for filtering the dataset before moving forward with analysis of genes.
• Data were log transformed and mitochondrial genes and number of unique molecular identifiers in each cell were regressed out.
• Variable genes were then identified and selected using a mean variability plot to examine dispersion for each gene (using log(variance)/mean).
• For clustering and dimensional reduction the first 9 principle components were used.
• Differential gene tests were done using negative binomial distribution to compare the two platforms.
• For each pair of cells a gene was detected in, we took the ratio of scaled counts to get a count on NextSeq to counts on HiSeq ratio which was then averaged for each gene and plotted.
• The distribution of genes along the distribution produced by mean NextSeq/HiSeq.

RESULTS

• Figure 1: Initial discrepancies detected based on choice of sequencing platform shown at bulk (left) and individual cell type level (alveolar macrophages, right). This was demonstrated in all libraries analyzed (SC02 NextSeq SC14/15 HiSeq).
• Figure 2: Identification and abundance of differentially expressed genes, particularly ribosomal genes, detected between platforms. The relative skew toward the HiSeq platform is evident in the middle region indicating the largest skewing.
• Figure 3: Density plots of genes used for building principle components used in clustering, genes differentially expressed between libraries, mitochondrial genes, and ribosomal genes along the curve of NextSeq to HiSeq ratio by gene abundance. DE genes appear at skewed towards HiSeq as they are skewed to the right tail of the distribution.
• Figure 4: Genes used to correct cluster skewing and Before(left) and After(right) scores were calculated in Seurat and demonstrate notable improvement.

CONCLUSIONS

• We report systematic gene detection bias between platforms Impacting Differential Gene Expression and Analysis and a list of 45 genes overlapping between the datasets shown to consistently have skewed detection towards the HiSeq platform.
• All of these genes were abundantly detected in samples and must had multiple isoforms.
• Almost all principle component genes are from the region with the smallest discrepancy between number of reads detected in families in HiSeq, thus initial clustering and cell type analysis is not affected by choice of platform.
• Many of these genes were ribosomal genes and a bias towards them could be seen splitting each cluster of cells in half based on heat maps, thus making it difficult analysis at the level of individual cell types.
• Thus we propose, as an improvement over regressing out effects from all genes or just ribosomal genes, the generation a custom gene list to be removed from analysis at the individual cell type level if doing integrative analysis as this led to a greater overlap between identical cells in TSNE plots and minimized differential genes detected between the two.

ACKNOWLEDGMENTS

We would like to thank 10x Genomics and Drs. Alexander Misharin and Scott Budinger for access to the datasets and resources used in this study.

REFERENCES


Northwestern University, Division of Pulmonary and Critical Care, Feinberg School of Medicine, Chicago, IL, USA;