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

Rohan Verma, Nikita Joshi, Ziyou Ren, Paul Reyfman, Scott Budinger, Alexander Misharin
[1]Northwestern University, [2]Division of Pulmonary and Critical Care, Feinberg School of Medicine, Chicago, IL, USA;

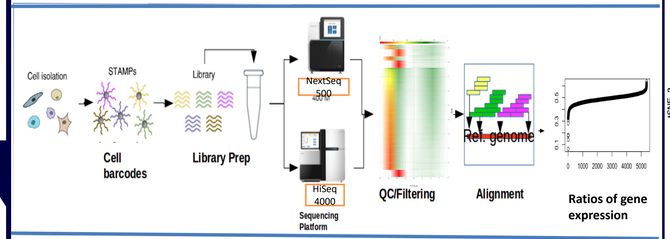
HYPOTHESIS

If there is no bias in the gene detection between the two platforms (NexSeq vs HiSeq), then we expect that there will be no differentially expressed genes within the same sample.

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.

OVERVIEW



METHODS

- Datasets: human cell datasets from10x Genomics (SC00) and others from Northwestern Division of Pulmonary and Critical Care lung transplant data sequenced on NextSeq500 and Hiseq4000 platforms.
- Programs: R was used for all data analysis and For single cell data analysis and visualization the package Seurat was used alongside the ggplot package to produce plots.
- Libraries were prepared following standard protocol found on 10x website before being sequenced 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 for each dataset.
- Differential gene tests were done using negative binomial distribution to compare cells sequenced on different platforms.
- For each pair of cells a gene was detected in, we took the ratio of scaled counts to get a counts 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 ratios was also examined.

RESULTS

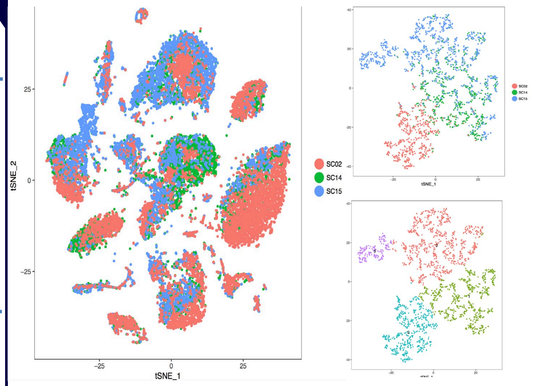


Figure 1: Initial discrepancies detected based on choice of sequencing platform shown at bulk (left) and individual cell type level (alveolar macrophages, right) This was demonstrated in all libraries analyzed (SC02 NextSeq SC14/15 HiSeq).

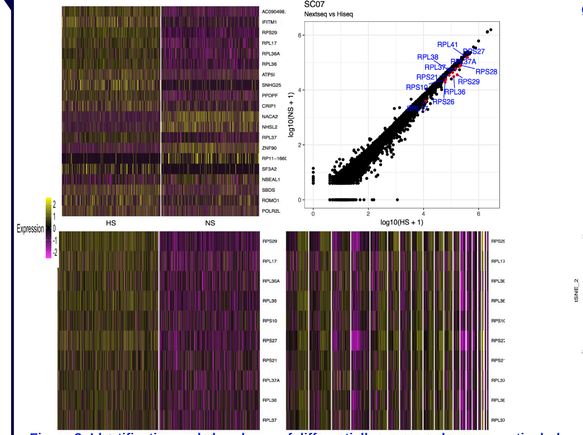


Figure 2: Identification and abundance of differentially expressed genes, particularly ribosomal genes, detected between sequencing platform. Heat maps show bias of ribosomal genes skewed towards Hiseq and splitting each cluster of cells in half. This demonstrates the Bias towards certain genes are platform specific and not influenced by abundance of expression.

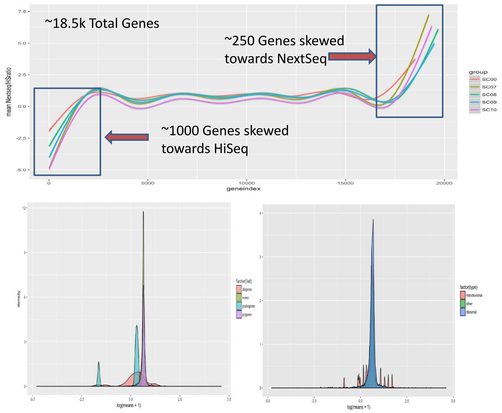


Figure 3: Density plots of genes used for building principle components used in clustering, genes differentially expressed between libraries, mitochondrial genes, and ribosomal genes along the curve of NextSeq to HiSeq ratios by gene above . DE genes appear at skewed towards HiSeq as they are skewed to the left tail end of the top plot. Genes used in principle component analysis to identify cell types are located in the middle region indicating the least skewing.

[1] "RPL36"	"FUS"	"PPIA"	"NBEAL1"
[5] "RPS29"	"SRSF5"	"RPS28"	"RPS3A"
[9] "RPL17"	"RP11-366C4.3"	"RP11-356C4.1"	"TCEB2"
[13] "NACA2"	"SAP18"	"PABPC5"	"MBNL1"
[17] "ATP5L2"	"MTRNR2L11"	"SNHG25"	"MIND51"
[21] "TIPIN"	"MF5D28"	"AC113404.1"	"NHSL2"
[25] "SH3BGRL"	"RP11-849F2.2"	"TXNDC5"	"AES"
[29] "RPL37A"	"HNRNPA1L2"	"CRIP1"	"EIF3A"
[33] "ATP5EP2"	"RPL36A"	"LINC01116"	"MTRNR2L12"
[37] "MAP1LC3B2"	"ATP6V0E1"	"ABCB5"	"EIF5AL1"
[41] "RP11-2H8.2"	"ACTR3"	"RPL37"	"MT-CO1"
[45] "COX6A1"			

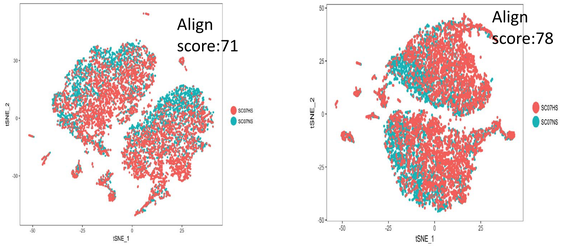


Figure 4:Genes used to correct cluster skewing and Before(left) and After(right) results showing improvement of alignment while minimizing the amount of information lost for downstream analysis as with methods that regress out all ribosomal genes or use Canonical Correlation Analysis to align datasets. Alignment 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 most had multiple isoforms.
- Almost all principle component genes are from the region with the smallest discrepancy between number of reads detected in NextSeq vs 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 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.