Singling Out Sources of Error in Novel-Tissue Single-Cell RNA Sequencing

Pranav K. Mishra[1], Sai Rama Krishna Meka[1], Michael Klüppel[1], Huseyin Ozkan[1], Esin Ozkan[1], Anna Spagnoli[1]
Rush University, Chicago, IL, United States
pranav_k_mishra@rush.edu

Disclosures: P.K. Mishra: None. S. Meka: None. M. Klüppel: None. H. Ozkan: None. E. Ozkan: None. A. Spagnoli: None.

INTRODUCTION: In the 15 years since it was first described, single-cell RNA sequencing (scRNA-seq) has become one of the most predominant transcriptome-wide analysis techniques in biomedical research. This technology "came of age" through the parallel expansion of computational capacity and selection of tissues examined in large atlas projects. Maturation of scRNA-seq has led to the development of boilerplate workflows, which were largely developed from terminally differentiated or mature tissues. Applying conventional workflows without first systematically filtering, using advanced algorithms, can erroneously minimize or eliminate critical data and/or lead to data misrepresentation. Here, we analyze scRNA-seq data using a novel stepwise, "human-in-the-loop" approach to establish a methodology to analyze scRNA-seq data obtained from isolated growth plates.

METHODS: Primary growth plates samples from post-natal day 5 Prx1CreER-GFP mice were isolated from the distal radius-ulna (RU), tibia, and humerus. Cells were FACS-sorted based on GFP expression, as imaging reporter for Prrx1 2.4 Kb limb enhancer activitytometry. 10,000 GFP-positive cells from each of the three aforementioned samples, along with a 10,000 cell GFP-negative RU control, were analyzed using chromium scRNA-seq (10X Genomics). Initial demultiplexing, upstream quality control (QC), and primary outcomes were analyzed using "conventional" approaches by the Research Bioinformatics Core at the University of Illinois at Chicago (UIC). In parallel, the authors independently performed a series of human-in-the-loop QC experiments to examine and compare with core's Louvain algorithm based methodology via *seurat*. Our group utilized tools from scverse project, including the *scvi-tools* for upstream QC and differential expression (DE) analysis, along with scanpy for further DE and visualization. First, we apply *solo*, a semi-supervised deep learning algorithm, to each of the four samples. A doublet-singlet score (DSS) is calculated, upon which we filter cells flagged with a DSS > 2.5. The 4 samples are integrated together, where we calculate, filter, and normalize based on QC metrics: unique molecular identifiers (UMI); total gene counts (TC); mitochondrial and ribosomal gene percentages; and highly variable genes. The Leiden algorithm via *scvi* performs community detection into cell clusters. We calculate probabilistic DE between samples and/or clusters and perform pathway analysis. Primary outcomes include the final QC metrics, sample and cluster wise DE, and top pathways identified by Gene Ontology (GO) and KEGG.

RESULTS: A total of 44,349 cells were detected by Cell Ranger across 4 sample groups, with 2 samples having a sample count exceeding the number of 10k cells fed into the pipeline per sample: 10,356 and 21,634 for GFP-positive Tibia and Humerus, respectively. Solo initially identified 11,234 cells as doublets (25.3%). We identified a DSS > 2.5 approximated an upper 2σ cutoff and was applied. 42,777 cells entered our quality control pathway, with 20,047 passing through (9% reduction over the core lab). Mitochondrial (0.71 ± 0.28 SD) and ribosomal (21.13 ± 4.97 SD) gene percentage were uniformly distributed across UMI and total cell counts (Figure 1). Therefore, we utilized TC, mitochondrial, and ribosomal gene percentages as continuous covariates while normalizing. Qualitatively, the Leiden algorithm performed tighter, discrete clustering. DE analysis in GFP-positive RU vs Tibia had a 20% absolute reduction in extraneous genes (e.g. 60S ribosomal unit genes) over the "conventional" protocol.

DISCUSSION: A human-in-the-loop approach to neural network based upstream QC and clustering provides a biologically representative cell set, which is normally distrusted without losing cellular nuance through overly aggressive filtration. Demystifying the "black-box" of scRNA-seq is critical in novel scRNA-seq analyses. For example, applying a "literature standard" filter for ribosomal genes of < 2% would eliminate the dataset, without considering the high cellular activity of the growth plate. Improving the DE analysis improved our downstream heatmaps and pathway analysis, allowing the authors to identify several genes and pathways which were not immediately apparent with the contemporary approach.

SIGNIFICANCE: As scRNA-seq technologies are stabilized and increasingly utilized, investigators should remain active throughout the experiment, as opposed to considering the technology as a complex "black box" feeding boilerplate inputs which may eliminate the biological nuance originally sought. This is especially crucial when analyzing developing tissues through "conventional" pipelines designed for mature, adult tissues.

ACKNOWLEDGEMENTS: We acknowledge the supports of Dr. Mark Maienschein-Cline, Director of the UIC Research Informatics Core, for the "conventional" analyses of the dataset and Dr. Zarema Arbieva, Director of the UIC Genomics Research Core, for the scRNA-seq (10X Genomics) studies. University of Illinois Research Informatics Core is supported in part by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1TR002003.

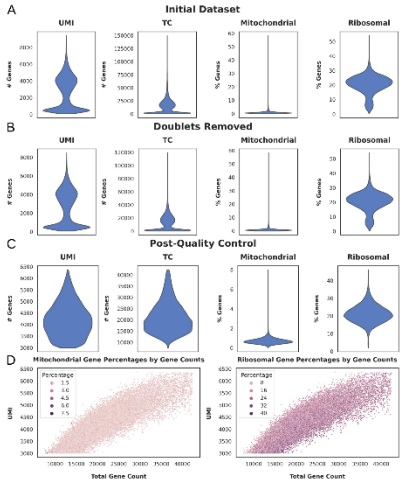

Figure 1 Human-in-the-loop exploration of scRNA-seq parameters: (A) Initial dataset: 44,349 cells identified by Cell Ranger. (B) Post-doublet removal using *solo*, with a DSS > 2.5: 42,777 cells. (C) Post-QC with 20,047 cells entering the Leiden-based *scvi* clustering. (D) Mitochondrial and ribosomal gene percentages are uniformly distributed across UMI vs. TC of genes in all samples.