Integrating bulk and single cell transcriptomics for accurate detection of tissue-specific gene expression

Alec Barrett[1], Erdem Varol[2], Mingfeng Li[3], Rebecca McWhirter[4], Seth R. Taylor[4], Alexis Weinreb[1,3], Nenad Sestan[3], Oliver Hobert[5,6], David M. Miller III[4,7], Marc Hammarlund[1,3]

[1]Yale University Department of Genetics, [2]Columbia University Department of Statistics, [3]Yale University Department of Neurobiology, [4]Vanderbilt University Department of Cell and Developmental Biology, [5]Columbia University Department of Biology, [6]Howard Hughes Medical Institute, [7]Vanderbilt University Program in Neuroscience

Abstract:

Advances in RNA-seq for bulk and single cell (sc) approaches have produced increasingly fine dissections of the *C. elegans* transcriptome. Although both techniques can yield transcriptomes for individual cell types, each comes with strengths and weaknesses. scRNA-Seq affords high resolution, but suffers from dropout, leading to false negatives. Bulk sequencing detects more genes, but suffers from contaminating cell types, resulting in false positives. In this work we integrated these orthogonal approaches to improve the accuracy of both methods. We used bulk samples collected for specific neuron types and sc datasets for all *C. elegans* neurons and additional non-neuronal cells (1). We used sc data to estimate contamination in each bulk sample, and developed novel methods for removing these gene counts and improving gene detection. In one approach we used linear histogram matching to scale sc counts, and subtracted putative contamination using data from non-neuronal clusters. In another approach, we performed an unweighted integration of single cell pseudobulk counts with the average subtracted bulk profile for each cell type, similar to the integration technique used in Seurat V3 (2). We assessed these approaches in two ways: 1) Measuring improvements in calling genes with known expression in all neurons; 2) Examining effects on eliminating genes expressed exclusively in contaminating tissues. We found that our analysis reduced false positives, while maintaining robust true positive detection, thus offering a novel strategy for utilizing complementary bulk and sc RNA-Seq data sets to enhance the accuracy of cell-specific expression profiling data.

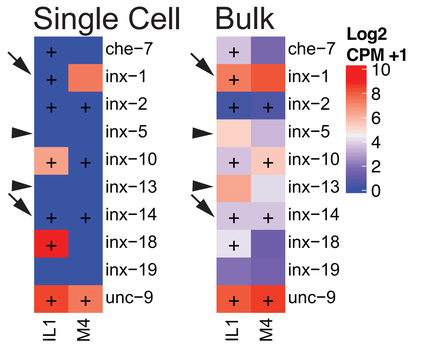

Figure 1: Bulk and Single Cell (sc) data have complimentary detection errors. scRNAseq provides low false positive rates (FPR), but high false negative rates (FNR) due to dropout (black arrows). Cell specific bulk RNAseq provides high FPR but low FNR due to noise and contamination (black arrowheads). Integrating these two datasets provides an opportunity to detect genes more accurately than either technique can achieve on its own.

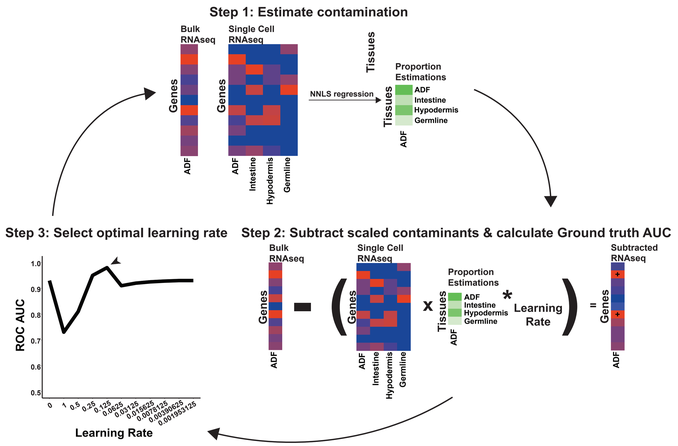

Figure 2: Subtraction algorithm schematic. For each bulk sample, a for loop is initiated that performs iterative small subtractions, with the goal of optimizing detection of known ground truth genes. Steps: 1) Estimate tissue level composition of bulk sample using single cell data. 2) scale single cell matrix using proportions and a learning rate, and subtract from the bulk data. 3) calculate the ROC AUC for all learning rates, select the subtraction with the highest AUC. If AUC does not improve. Break the loop

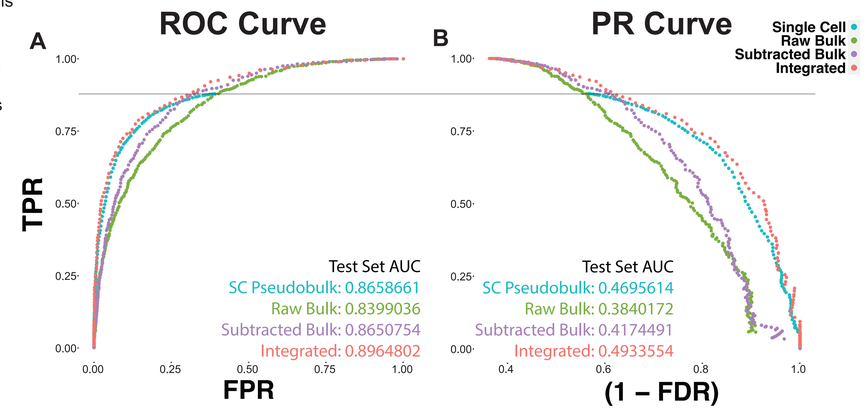

Figure 3: Integrated counts outperform bulk and single cell in test gene detection. For each dataset, we assessed their ability to capture the expression pattern of known ground truth marker genes at a wide range of thresholds. A) the receiver operating characteristic (ROC) curve for the single cell (sc) pseudobulk counts per million (CPM) (blue), raw bulk CPM (green), subtracted bulk CPM (purple), and Integrated (red) datasets. Showing the relationship between the true positive rate (TPR) and false positive rate (FPR), across all thresholds. B) The precision-recall (PR) curve for each dataset. Showing the relationship between the TPR, and the false discovery rate (FDR), across all thresholds. Black line indicates the test set TPR threshold used for comparisons in Figures 3 & 4.

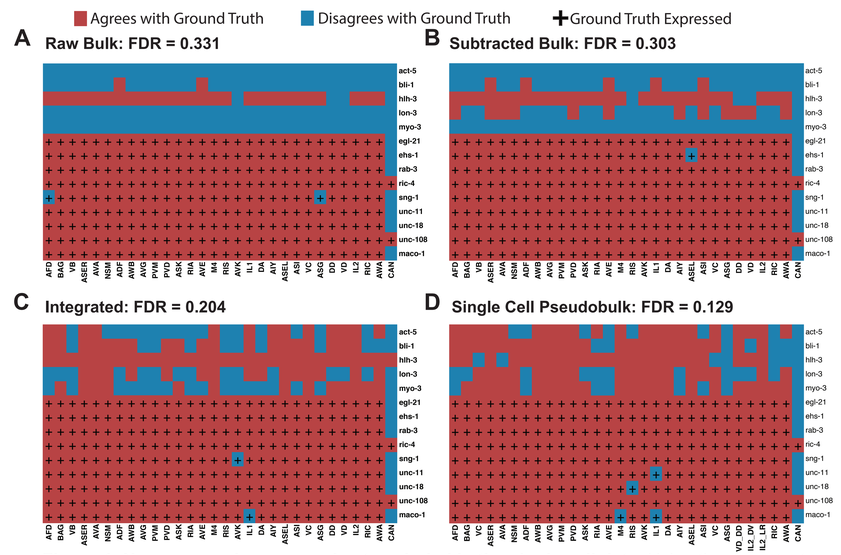

Figure 4: Non-neuronal genes are best excluded in the single cell data. Using the same test set true positive rate (TPR) for each sample (0.879), we compared the detection of pan-neuronal and non-neuronal genes in all four datasets. Red colors indicates agreement with ground truth expression, and blue represents disagreement with the ground truth. Plus sign indicates expression in ground truth.

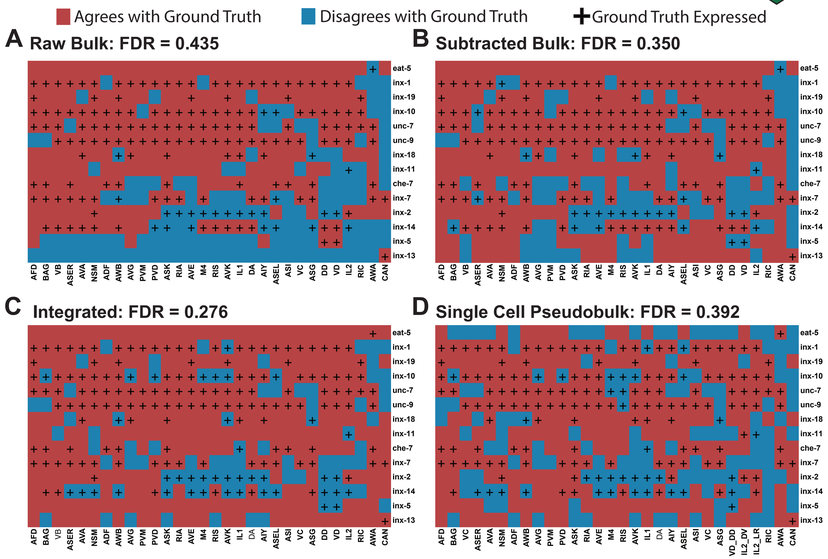

Figure 5: Integrated counts accurately detect innexin genes. Using the same test set true positive rate (TPR) for each sample (0.879), we compared the detection of innexin genes in all four datasets. Red colors indicates agreement with ground truth expression, and blue represents disagreement with the ground truth. Plus sign indicates expression in ground truth.

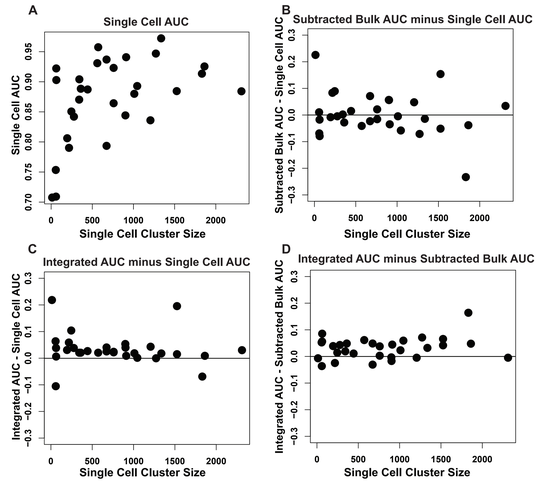

Figure 6: Integrated AUC does not degrade for lowly populated clusters. Relationship of single cell cluster size to ROC AUC for individual cell types.

Acknowledgements: This work was funded by NIH grant R01NS100547 and by Vanderbilt Trans-Institutional Program. Flow Cytometry experiments were performed in the Vanderbilt Flow Cytometry Shared Resource which is supported by the Vanderbilt Ingram Cancer Center (P30 CA68485) and the Vanderbilt Digestive Disease Research Center (DK058404). The Vanderbilt VANTAGE Core provided technical assistance for this work and is supported by CTSA Grant (5UL1 RR024975-03), the Vanderbilt Ingram Cancer Center (P30 CA68485), the Vanderbilt Vision Center (P30 EY08126), and NIH/NCRR (G20 RR030956). Some strains were provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). Bulk samples were sequenced at the Yale Center for Genomic Analysis

1. Taylor SR, Santpere G, Weinreb A, Barrett A, Reilly MB, Xu C, et al. Molecular topography of an entire nervous system. bioRxiv. 2020:2020.12.15.422897.
2. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888-902.e21.