Single-cell RNAseq and Deconvolution of bulk-cell RNAseq

# Significance

RNA-seq is a new but established technology for genome-wide transcriptomic analysis. It has been widely applied for understanding various diseases and enables discovery of gene clusters associated with common functions, as well as identification of novel transcripts with the same functions. At the same time, single-cell technologies can greatly expand upon the sensitivity and cell-type specificity. On the transcriptomic level, single-cell RNA-seq offers an unbiased measurement of the entire collection of mRNA transcripts produced in each cell. Despite its inherent sparsity, it complements bulk RNA-seq in characterizing the heterogeneity among different cell types. A promising approach for discovering new cell types is to perform unsupervised clustering on single-cell RNA-seq data and use it to inform the bul-cell RNAseq data.

We utilize single-cell measurements mRNA abundances using RNA-seq to deeply characterize populations of cells. We develop pipelines for single-cell RNA-seq that accounts for the technical variability for each gene and converts raw reads to gene counts in a biologically meaningful manner. After processing, we employ our previously developed methods to reduce the data into a few robust dimensions, and cluster the results into metagenes corresponding to pathways. These metagenes are subsequently analyzed by DREMI and DREVI in order to characterize novel pathways and their interactions in the cell populations. With the understanding of signaling and gene-interaction networks we characterize the pathways involved in different cell types, and understand their involvement in emerging phenotypes.

# Methods

## Single-Cell RNA-Seq Data

Single-cell RNA-seq has the possibility of offering an unbiased view of the pathways that are transcriptionally activated even when cells seem phenotypically similar. However, single-cell sequencing suffers from more technical noise as compared to bulk RNA-seq, arising largely from three sources: (1) sampling inefficiencies which result in only a small fraction of the total number of transcripts being captured; (2) cell-to-cell variations in sequencing efficiency, potentially due to differences in lysis between cells; and (3) amplification bias owing to the small amount of starting material for the RNA-seq. Attempts have been made to address these concerns [36, 37], however, there is no standard to reach robust transcript counts. The main steps of our method include (1) debarcoding and error correction, (2) Aligning reads from each unique molecular identifier (UMI), and (3) quantifying the biological noise in genes.

One of the keys to successfully extracting information from single-cell RNA-seq data is to be able to use the high-dimensionality of the data to bolster individual (especially low-abundance) gene dimensions that can suffer from dropout. We perform the following steps in order to be able to analyze and cluster single-cell RNA-seq data: (1) use non-linear dimensionality reduction and clustering on genes to form meta-genes, (2) value-impute based on cell clusters and meta-genes, and (3) use the value-imputed data to study gene-gene interactions.

## Deconvolution of cell-type signatures from bulk RNA-seq data

We identify the cell-type signatures from within RNAseq data from a mixture of cells using linear models. Given gene expression levels and cell type fractions for each sample, we can use a linear matrix model to identify cell types gene expression signatures. For instance, the sample’s *i*th gene expression level can be modeled as a linear superposition of the same gene’s expression levels of multiple cell type signatures; i.e., the *i*th gene expression level of *k*th individual person, *x(i,k)* is the linear combination of this gene’s expression levels of different cell type signatures; i.e., , where *s(i,j)* is the *i*th gene’s expression level in the *j*th cell type, and *w(i,k)* is the contributing weight of *j*th cell type to *k*th sample, which can be the *j*th cell type fraction of *k*th sample. If we rewrite this linear model in a matrix form, we have that *X*=*SW*, where *X* is the gene expression matrix where the rows and columns represent genes and samples, *W* is the cell type fraction matrix whose rows and columns represent cell types and samples, and *S* is the cell type signature matrix whose the rows and columns represent genes and cell types. Single-cell RNA-seq data described above yields counts of different cell types, providing the data required for matrix *W*.

Figure 1. Example deconvolution analysis using RNAseq from a mixture of cells in a sputum sample. Left, cell type signature correlation between samples from asthma patient and control for top genes that have highly differential expression levels. Right, all gene expression levels in cell type signatures (x-axis: control, y-axis: asthma). In this case cell population percentages were obtained by cell staining and light microscopy, but could also come from single-cell RNAseq.

 We also apply

advanced models to capture nonlinear effects from different cells to gene expression. For example, we can use machine-learning methods to investigate the gene markers from cell type gene expression signatures for both bulk data and single-cell type. In particular, we use the Denoising Autoencoder (DA), an unsupervised machine-learning framework to extract and characterize cell type signatures. DA is able to discover non-linear expression features from gene expression data using sigmoid transformation. We will apply DA to different sample clusters and compare their non-linear features, and find the genes that have features to most discriminate clusters.