Submitter Name: Guy Twa
PI Name: Jeremy Day

Submiited Email: gtwa@uab.edu
PI Email: jjday@uab.edu

Accurate sample deconvolution of pooled snRNA-seq using sex-dependent gene expression patterns

Guy Twa[1], Robert Phillips III[1,2], Nathaniel Robinson[1], Jeremy Day[1]

[1]University of Alabama at Birmingham, Birmingham AL
[2]Present institution: Lieber Institute for Brain Development, Baltimore MD

Single nucleus RNA sequencing (snRNA-seq) technology offers high resolution methods for studying cell type specific mechanisms. These technologies have high costs and technical limitations, often requiring pooling independent biological samples for sequencing. Pooling results in loss of individual data, and current methods to recover this information involve additional technologies and data modalities. Deconvolution of sample identity using inherent features would enable incorporation of pooled barcoding and sequencing protocols, thereby increasing data throughput and analytical sample size without requiring increases in experimental sample size and sequencing costs. In this study, we provide a proof of concept that sex-dependent gene expression patterns can be leveraged for deconvolution of snRNA-seq data from pooled sequencing of different sex samples. Using previously published ventral tegmental area snRNA-seq data, we trained a range of machine learning models to perform cell sex classification using genes that are differentially expressed in cells from male and female rats. Models utilizing sex-dependent gene expression were able to predict the cell sex with high accuracy (90-92%), and outperformed simple classification models using only sex chromosome gene expression (69-89%). Generalizability of these models to other brain regions was assessed using an additional published data set from the rat nucleus accumbens. Within this data set, model performance remained highly accurate in cell sex classification (89-90% accuracy). This work provides a model for future snRNA-seq studies to perform sample deconvolution using a two-sex pooled sample sequencing design, and benchmarks the performance of various machine learning approaches to deconvolve sample identification from inherent sample features.