Towards Whole Transcriptome Deconvolution Using Single-cell Data

James Lindsay
Computer Science and Engineering
University of Connecticut
james.lindsay@engr.uconn.edu

Craig E. Nelson
Molecular and Cell Biology
University of Connecticut
craig.nelson@uconn.edu

Ion I. Măndoiu
Computer Science and Engineering
University of Connecticut
ion@engr.uconn.edu

***Abstract*—Obtaining whole-transcriptome expression profiles of closely related cell types is a daunting task faced by stem-cell biologists. Here we present an approach that utilizes single-cell qPCR probing of a small number of genes to aid in the deconvolution of whole-transcriptome profiles of mixed samples.**

I. Introduction

The expression profiles of m genes measured in n mixtures of k cell types are modelled as $X = SC$, where X is a $m \times n$ matrix whose columns are the expression profiles of individual mixtures, S is $m \times k$ "signature" matrix whose columns are expression profiles of individual cell types, and C is a $k \times n$ "concentration" matrix whose columns represent the proportions of each cell type in individual mixtures. In this abstract we will assume that individual cell types as well as a reduced signature matrix $\hat{S}$ can be reliably inferred from single-cell qPCR data generated for a small subset of genes.

II. Methods

A. Constructing Reduced Cell-Type Signatures

1) Noise Reduction: Due to large biological and technical noise in single-cell qPCR data we applied a common technique where each sample was required to have .95 Pearson correlation with at-least one other sample, otherwise it is removed.

2) K-means Clustering: We chose to use k-means clustering to group the gene expression data from single-cell data because it explicitly allows us to control the number of theoretical cell-types. The average expression profile of each single-cell in a cluster is used to create the reduced cell-type signature matrix $\hat{S}$.

B. Estimate Mixing Proportions

The next task is to solve for the concentration matrix C. We utilize the same methodology described in [**?**] to compute the concentration matrix describing the mixtures. Each column of X, and hence also columns of the reduced expression matrix $\hat{X}$ obtained by retaining only rows of X corresponding to genes measured by qPCR, is a linear combination of single cell expression profiles with unknown concentrations. Let us denote a particular column in $\hat{X}$ as x and its corresponding column in C by c. Inferring c can be formulated as the following quadratic program that can be solved using standard constrained quadratic programming solvers:

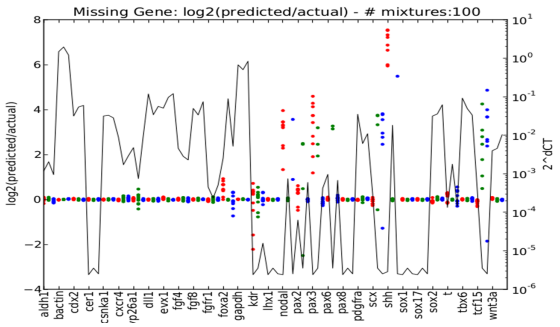

Fig. 1. On the left y-axis is the log2 ratio between predicted and actual gene expression signatures per cell-type. On the right y-axis is the average expression signature for each gene and cell-type. Each cell-type is a particular color.

$$\begin{aligned} \text{minimize} \quad & ||\hat{S}c - x||_2 \\ \text{subject to} \quad & \sum c = 1 \\ & c_i \geq 0 \ \forall i = 0...m \end{aligned}$$

C. Estimate Full Expression Signatures

It is still necessary to estimate the signatures of the full gene profile. Using the concentration matrix C inferred in previous step, the gene signature s of a gene not measured by qPCR can be inferred from the mixed gene expression data x using a similar least squares quadratic program:

$$\begin{aligned} \text{minimize} \quad & ||sC - x||_2 \\ & s_i \geq 0 \ \forall i = 0...k \end{aligned}$$

III. Preliminary Experimental Results

The above method was applied on qPCR expression data generated from mouse embryos at the 7-8 somite stage. Expression levels of 31 genes were characterized by RT-qPCR for 97 single cells and 12 mixed samples. In order to test the methods ability to estimate the concentration matrix and complete gene signature we ran a leave-one-out experiment on each gene. Figure 1 demonstrates that the method is able to accurately deconvolve expression levels of most genes, however particular genes seem to pose a challenge.

978-1-4673-1321-6/12/$31.00