ATCGes: A Tool for Curating Gene Expression Signatures from Microarray Data

Author: Louie Dinh
Email: LouieDinh@gmail.com

Domain and Problem

Studying genomics data is difficult because techniques for observing cellular processes in-vitro are severely limited. However, the capabilities of tools for gather data ex-vivo are exponentially increasing. Since the advent of sequencing, we have progressed to processing an entire human genome for approximately $1000. Furthermore, microarrays allow us to measure the gene activity of our 20,000 genes simultaneously.

A major deficiency in microarray gene expression data is the inability to measure relative cell proportions. We are capable of saying that there is more of gene A, but unable to separate the causes of increased expression versus increased cell proliferation. These are two very different things. Understanding cell population sizes is crucial to disease etiology. For example an increase in different subsets of immune cells can either mean your body is fighting or helping cancer progression[Ostrand-Rosenberg 2008]. Another reason to measure cell proportion is because it is a major confounder for statistical procedures. Models that don't take cell proportions into account lose much of their statistical power. Since measuring cell proportions is costly and time intensive, we would like to computationally determine the relative cell proportions.

Given the confluence of factors (cheaper sequencing, gene expression microarrays, methylation microarrays), we'd like deconvolute the mixture of cells back to their original proportions. This is known as the source separation problem. For a given matrix, this is easy if it has an inverse and can be solved analytically. However, in cells,
the measurements are noisy, there are many linearly dependent rows, thousands of gene measurements make inverting the matrix difficult and also computationally expensive O(N^3) even if all measurements were perfect and linearly independent.

The crux of all deconvolution problems is identifying a submatrix that has nice properties. Essentially we want differentially expressed regions that allow us to tease
apart the different cell proportions. Manually looking through thousands of genes is not feasible. We want a way to select a basis matrix.

Data Set

The data set comes from microarray luminance measurements, also known as spots, from the gene expression data of subsets of immune cells [Abbas et. al. 2005]. Each cell type is measured across 20,000 genes simultaneous.

Previous Experience

I have previous experience working with genomic data sets. Previously I have worked in Loren Riesberg's lab building phylogenies, Michael Brudno's lab doing genome assembly from high throughput sequencing data and Irmtraud Meyer's lab predicting RNA secondary structure. We are using the IRIS (Immune response in silico) dataset from Abbas et al. In this dataset the characterize the gene expression profiles of the immune constituents of whole blood.

Task Abstraction

Why

At the high level, the user is attempting to **produce** a basis matrix by **annotating** the genes whose expression can distinguish between the different cell types.

During the search, the **target is known**. The user is **looking for a specific feature**, a gene that is differentially expressed in one or two cells against the background cells. The **location is unknown** and we are searching through approximately 20,000 genes.

The user is attempting to **identify** a gene that acts as a good marker for the cell and does so by **comparing** the expression of a single gene across the different cell types.

Once a gene is found, it is recorded into the basis matrix for use in downstream analyses.

What

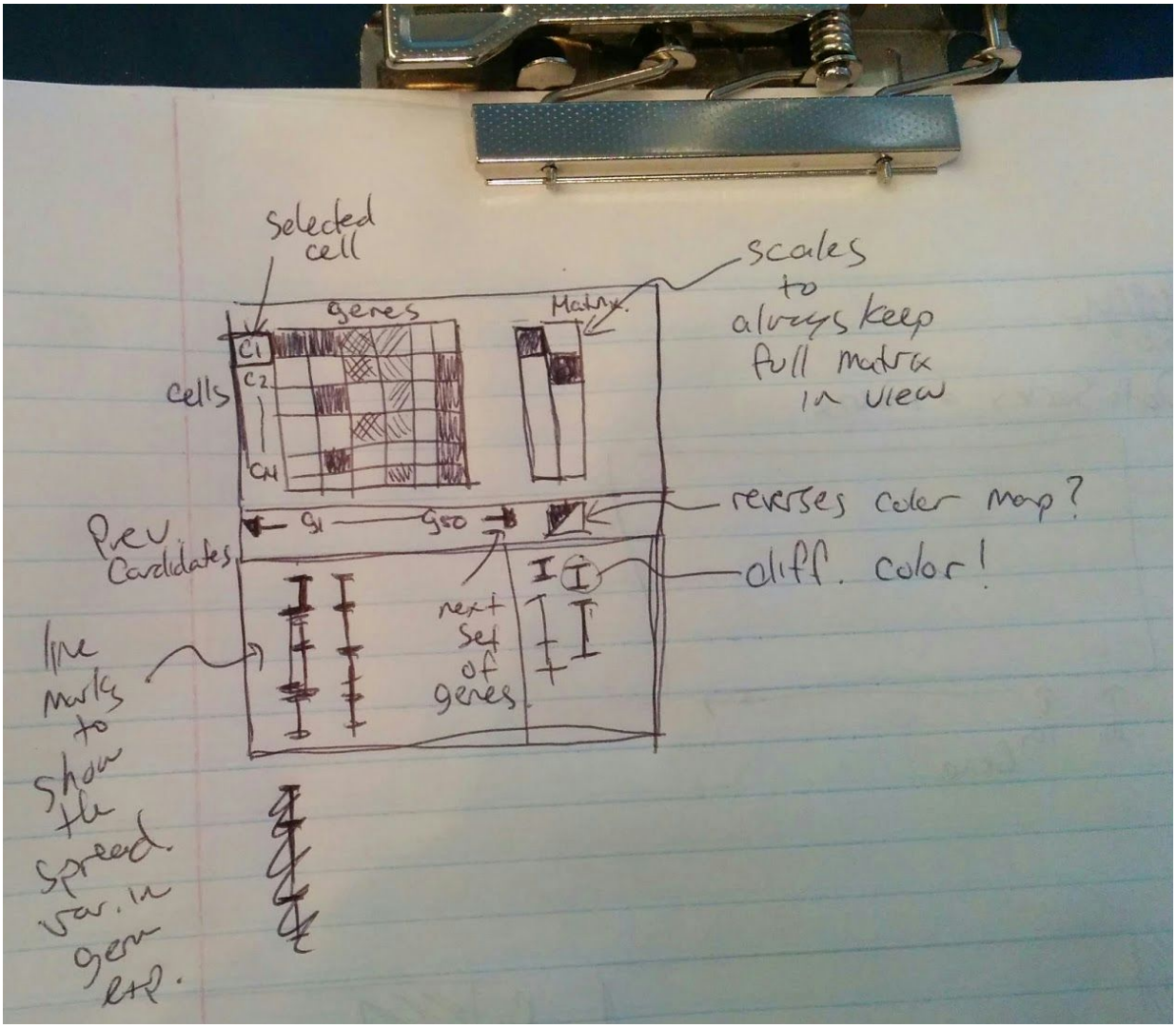
We are dealing with tabular data. The expression matrix has dimensions the number of cells (dozens) by the number of measured genes (tens of thousands). The keys are genes/cell-type pairs and the values are expression data that is measured as luminance on the microarray chip and then processed into approximate RNA abundance.

Proposed Solution

Preprocessing

The data needs to be preprocessed before visualizing because the gene expression levels are heterogenous (minimum=0.16, maximum=734038). Since we are only interested in comparing the same gene between cells to spot differential expression, we can normalize the data without any loss of information. We will follow the often used normalization of subtracting the mean expression and dividing by standard deviation (mahalanobis distance). This will allow us to compare genes by their deviations from the mean expression level rather than on an absolute scale.

Sketch



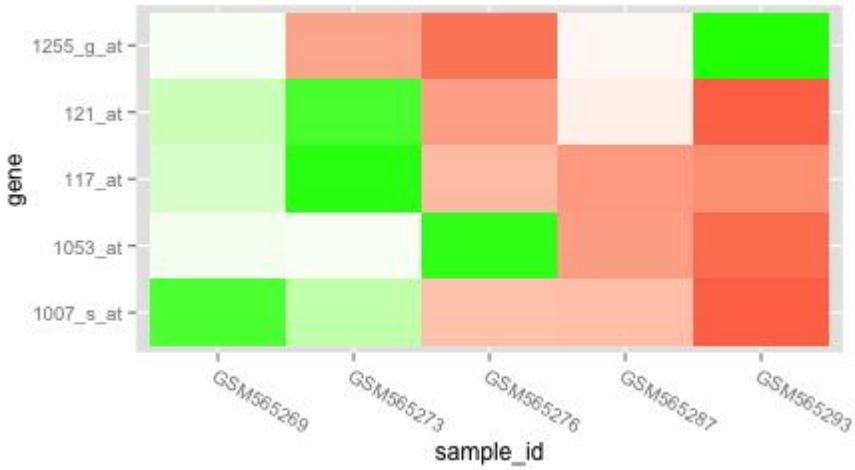
Description

This is the overview of the system. The left heatmap is the working area where candidate genes are being shown. When a cell is selected in the right column, the system filters out a set of candidate genes for the user to examine. The right view is the working basis matrix which scales to always keep the entire basis in view. The bottom pane is a modified box plot that shows the variance in expression levels on a per cell basis. The cell of interest will be color coded differently to provide visual pop-out so that the user can examine the amount of differential expression. I'm considering a button to reverse colour mappings.

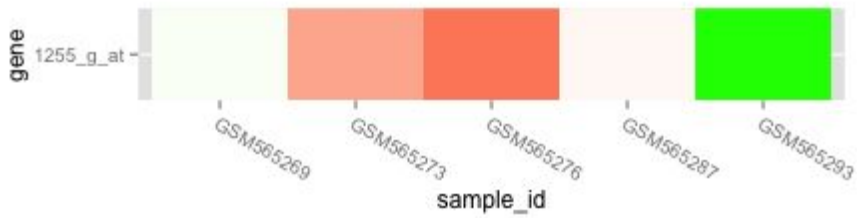
Scenario

User starts with a gene expression matrix and loads it into the system. We assume that they are attempting to create a basis matrix for every cell in the data file.

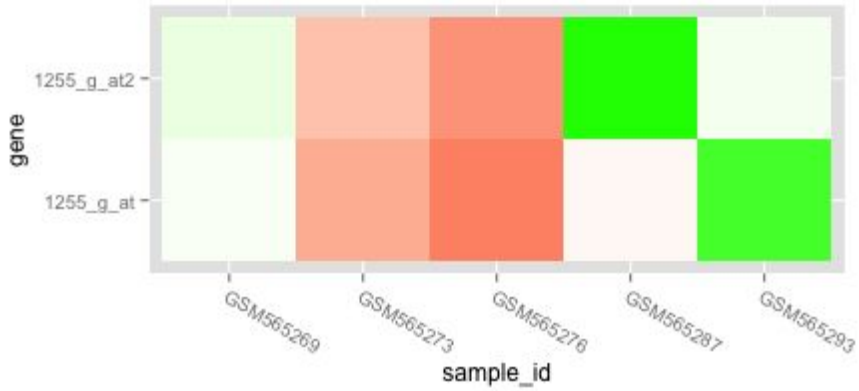
1. User decides on the cell that he would like to distinguish first and selects it to focus.
    a. System highlights cell by moving it to the top [See Sketch]. Filters out all bad candidate genes by finding a big gap between mean expression in cell of interest and other cells. System now shows approximately 2 dozen candidates.
2. If no likely candidates, then hitting next will show the next set of gene candidates.
3. Left with a heatmap like the following. The gene 1255_g_at seems like it provides good separation for GSM565293, so we will select it. This process is also facilitated by the variance box plots [See Sketch] that is directly below.



4. The system adds the selected gene into an adjacent view to the right that records the current basis matrix selected.



5. The user can continue selecting genes to distinguish the current cell, in which case we continue or the user selects a new cell to examine.
6. The user continues this process of selecting genes and building up the basis matrix. Here is the basis matrix after 2 genes have been selected.



7. Once the user has an appropriately sized basis matrix with a minimum of one distinguishing gene per cell type, the process is complete and the user is left with a basis matrix to be used in deconvolution.

Implementation

I will be using D3, which should be able to scale to ~20k genes and ~50 cell types. It will allow the scalability required while being accessible through the browser.

Milestones

November 23rd: Data loads into D3 and able to draw a heatmap of a subset of the data.
December 2nd: Candidate filtering algorithm and boxplots for variance in the data.
December 10th: Interactivity. Browse through candidates if current set doesn't offer any good genes. Selected genes moves into working basis matrix.
December 15th: Presentation Ready
December 18th: Submit Final Report

Related Works

There has been extensive work done on visualizing gene expression data. In a review Gehlenborg (2010) surveys the major visualization approaches for expression data. Common approaches include hierarchical clustering (Eisen 1998), dimensionality reduction (Bushati 2011), and self organizing maps (Tamayo 1999). Most often the goal is to understand how expression of different genes are related. The product of analyses are groups of genes that work together or inhibit each other in some way. The work suggested in this project has the opposite goal. It is attempting to identify genes as biomarkers that are unique to a particular cell rather than similar genes and their cooperative structure.

Citations

- Abbas AR, Baldwin D, Ma Y, Ouyang W et al. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. Genes Immun 2005 Jun;6(4):319-31.
- Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. Proc. Natl Acad. Sci. USA, 95, 14863–14868
- Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl Acad. Sci. USA, 96, 2907–2912
- Bushati N, Smith J, Briscoe J, Watkins C. An intuitive graphical visualization technique for the interrogation of transcriptome data.Nucleic Acids Res 2011;39:7380–7389.
- Gehlenborg,N., O'Donoghue,S.I., Baliga,N.S., Goesmann,A., Hibbs,M.A., Kitano,H., Kohlbacher,O., Neuweger,H., Schneider,R., Tenenbaum,D. et al. (2010) Visualization of omics data for systems biology. Nat. Methods, 7, S56–68.
- Ostrand-Rosenberg S (2008) Immune surveillance: a balance between protumor and antitumor immunity. Curr Opin Genet Dev 18:11–18