Correlated Feature Selection for Single-Cell Phenotyping

Geoff Stanley
Stanford University, Program in Biophysics

Abstract

Single cell transcriptome sequencing promises to provide a complete, unbiased picture of the cellular phenotypes comprising mammalian tissues: unprecedented amounts of data – expression levels of O(10000) genes, many of which are functionally annotated – are recovered per cell over hundreds of cells. However, functionally distinct cell types may be generated by very small differences in gene expression. Analysis methods must be developed that are sensitive to expression level differences in only O(0.1%) of expressed genes. I have sequenced the transcriptomes of 151 D1 and 86 D2 medium spiny neurons. I have found that they have highly similar gene expression patterns, differing by only ~13 out of ~7000 mutually expressed genes. Standard methods of dimensionality reduction (PCA, tSNE) on the entire feature set failed to differentiate the two. I found that by correlating genes to find consistent expression signatures, and reducing the feature set to only those signature-correlated genes, I was able to fully recapitulate the functionally relevant phenotypes of striatum. This work points toward an important

Introduction

In collaboration with Ozgun Gokce in the Sudhof lab, we have used single-cell transcriptome sequencing to fully describe the cell types and expression signatures of all of the cell types of the mouse striatum, a region of the brain that is involved in responses to reward and initiating motor action. We focused on two neuron types: the D1 and D2 medium spiny neurons (MSNs). These neurons are very similar; however, it has not been clear whether the D1 and D2 classification represents the full extent of phenotypic diversity in medium spiny neurons, as currently methods rely on a small number of genes selected in an ad hoc manner to differentiate them.

Goals

- Develop sensitive unbiased learning methods to distinguish between similar cell types.
- Discover the complete phenotypic structure of medium spiny neurons.
- Generate a list of genes that classify all known and novel MSN subtypes.

Data

We isolated 237 medium spiny neurons which were fluorescently labeled by their expression of Gfp or Tdtomato under the D2 or D1 gene promotor, respectively. The fluorescently-labeled neurons were captured and amplified on Fluidigm C1 chip and poly-A+ mRNA sequenced to a depth of >1 million 75bp paired-end reads/cell. The features of the data set are measured expression levels of 22,300 genes. Expression is measured as log2 TPM, where TPM (transcripts per million reads per cell) is calculated using Cufflinks transcriptome assembler on Bowtie/Tophat aligned data.

Methods

Unbiased analysis of single-cell gene expression data requires two basic components: dimensionality reduction/visualization, and classification. I used two pipelines that have been employed in previous single-cell papers:

PCA

PCA for dimension reduction and visualization, and k-means to assign phenotypes.
PCA was implemented with the R package FactoMineR [cit]. The samples were plotted on the first two PC dimensions. The number of subtypes *k* was inferred by looking at the plot, and *k*-means used to assign cells to groups.

tSNE

2) Kharchenko correlation as a cell-to-cell distance metric, t-Stochastic Neighbor Embedding for dimensionality reduction and visualization. The Kharchenko distance metric is a weighted correlation between cells over their gene expression, where the expression level (in TPM) of a gene in cells *k* and *l* is weighted by the posterior probability

$$\begin{align}
\mathrm{dist}_{\mathrm{K,ij}} &= \frac{\sum_i w_i(x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l)}{\mathrm{Var}(cell_k)\mathrm{Var}(cell_l)} \tag{1}\\
x_{ik} &= TPM_{ik} * w_i \tag{2}\\
w_i &= P(TPM_i^k, TPM_i^l; \theta) \tag{3}\\
P(TPM_i, TPM_j; \theta) &\sim (1-g)\mathrm{Pois}(\lambda_0 = 0.1) + g\mathrm{NB}(e) \tag{4}
\end{align}$$

that it's value is actually correlated to its concentration as mRNA in the cell. The weight for given expression values is the ratio of the value of the negative binomial to the Poisson at that expression level. The resulting cell-cell distances are mapped to a 2D space by minimizing the Kullback-Leibler divergence between the distances in high-D space and the distances in 2D space.

Correlated Feature Selection

A symmetric matrix was created of the correlations of gene *i* and gene *j* for all *i, j* that were expressed in at least one cell and had a nonzero variance across all cells.

$$\mathrm{Cor}_{\mathrm{ij}} = \mathrm{Cor}(\mathrm{gene}_{\mathrm{i}}, \mathrm{gene}_{\mathrm{j}}$$

Then I selected genes *k* such that

$$\begin{gathered}
\mathrm{max}_{\mathrm{i}}(\mathrm{Cor}_{ik}) > a\\
\mathrm{min}_{\mathrm{i}}(\mathrm{Cor}_{ik}) < -b\\
a, b \in (0,1)
\end{gathered}$$

i.e., genes that had a maximum positive correlation of at least *a* to at least one other gene *i* and a minimum negative correlation of at least -*b* to at least one other gene, where *a* and *b* were determined empirically. I then performed Ward hierarchical clustering on the resulting matrix.

I selected genes manually based on multiple genes forming a correlation "cluster" (red square outline in Figure 2a) – i.e., multiple genes all correlated to each other, that were all anti-correlated to another correlation "cluster" (green square outline in Figure 2a). I then ran both the PCA and tSNE pipelines on this reduced set of genes and assessed the accuracy of assignment as well as the appearance of any novel subtypes

Results

To visualize the phenotypic structure of the neurons, I compared the full feature set PCA and tSNE with PCA/tSNE on the reduced feature set from correlation analysis. I did not find any novel subtypes with any of the methods. Full-feature set pipelines were unable to distinguish the labelled D1/D2 MSN subtypes from each other (Figure 1).

PCA/tSNE on full feature set

PCA (Fig 1a,b) seems to find high-variance outliers; however they do not form any kind of obvious cluster or structure. tSNE (Fig 1c) actually appeared separated D1 and D2 cells; however, the separation was not detectable with k-Gaussians in the 2D embedded space. It did tend to separate a small (<10%) population of cells which it clustered together. I could not find any other way of distinguishing those cells from the rest, so I considered the population to be an artifact of tSNE.

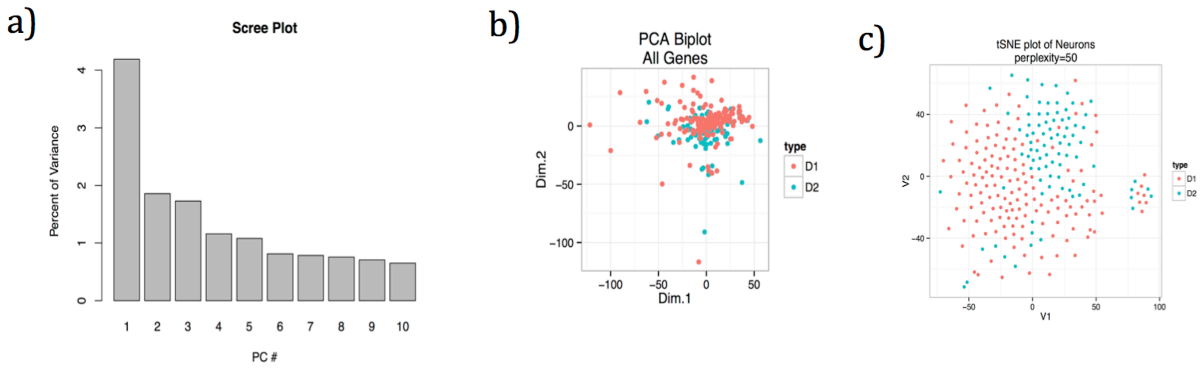

Figure 1. Unbiased analysis of full-feature set data from MSNs. a). Scree plot showing percent of variance (normalized to 100) covered by each principal component. b). Plot of samples on first 2 principal compenents; labeled D1 (red) and D2 (blue). c). tSNE 2D embedding of samples based on Kharchenko distance, same labels. Small cluster on right appears to be spurious.

Correlated Feature Selection

The symmetric heatmap of the correlation matrix shows only two clusters that are clearly anticorrelated to each other (Figure 2a). Reducing *a* or *b* did not introduce new anti-correlated clusters. I selected 13 genes from the two anticorrelated clusters, most of which are known D1/D2 markers (Drd1a, Tac1, Isl1, Pdyn, Sfxn1; Drd2, Adora2a, Penk, Gpr6).

I then ran the PCA/tSNE pipelines on all cells with expression values for only these 13 genes; PCA clearly separetes them into 2 populations (Figure 1c). It is difficult to determine *k* for the tSNE plot, so *k*-means would be ineffective in an unlabeled setting.

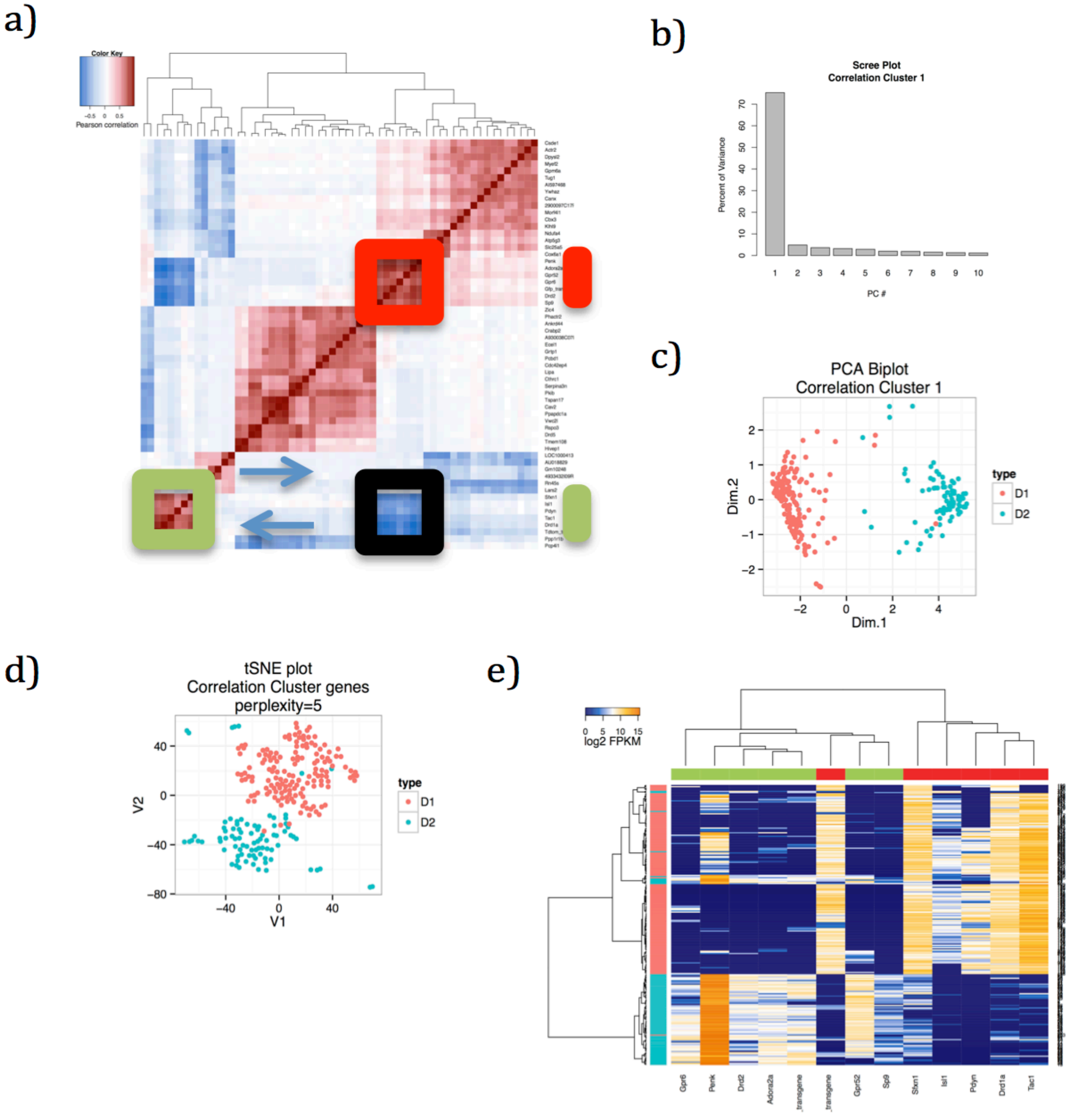

Fig 2. Correlated Feature Selection. a). Correlation heatmap of genes. Each cell represents the Spearman correlation between two genes over all cells. Red is highly correlated, blue is anticorrelated. b). Scree plot of PCA. c). PCA biplot on first two dimensions. The two types are well-separated by the method. d). tSNE performed on the 13 dimensions, embedded in 2 dimensions. Though samples are spatially separated, they do not form very distinct clusters that would be detectable with unlabeled data. e). Cell-gene heatmap of expression of the 13 genes, with genes and cells ordered by hierarchical clustering on correlation distance. Cells are rows, genes are columns. Clearly there are two distinct populations that have differential expression.

Accuracy of methods

	k	**%D1 Classified Correctly**
tSNE	2	46%
PCA	1	N/A
CFS+tSNE	unknown	N/A
CFS+PCA	2	98%

Discussion

I have demonstrated a method for effective feature reduction of single-cell RNAseq datasets. It works because genes that are exclusively expressed in one cell type will be correlated to each other and anticorrelated to genes that are exclusively expressed in another; the other 10,000 "noisy" genes are filtered out. D1 and D2 neurons have critical function differences in the brain, yet standard single-cell RNAseq analysis would label them as undifferentiated. Single-cell RNAseq promises to fully categorize the phenotypic diversity of multicellular organisms; to this end I will continue to develop more sensitive analysis methods.