Statistical methods for single-cell RNA sequencing data

Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison

http://www.biostat.wisc.edu/~kendzior/
Single-cell vs. bulk RNA-seq

Heterogeneous  Homogeneous  Sub-population
Features of single-cell RNA-seq data

- Abundance of zeros, increased variability, complex distributions

Challenges in scRNA-seq

- Normalization
- Technical vs. biological zeros
- Clustering; Identifying sub-populations
- De-noising
  - Adjusting for technical variability
  - Adjusting for biological variability (oscillatory genes)
- Identifying and characterizing differences in gene-specific expression distributions (aka. identifying differential distributions)
- Pseudotime reordering
- Network reconstruction
Challenges in scRNA-seq

- Normalization
- Technical vs. biological zeros
- Clustering; Identifying sub-populations
- De-noising
  - Adjusting for technical variability
  - Adjusting for biological variability (oscillatory genes)
- Identifying and characterizing differences in gene-specific expression distributions (aka. identifying differential distributions)
- Pseudotime reordering
- Network reconstruction
Challenges in scRNA-seq

- **Normalization**
  - Bacher, Chu et al., *Nature Methods*, 2017

- **Technical vs. biological zeros**

- **Clustering; Identifying sub-populations**

- **De-noising**
  - Adjusting for technical variability
  - Leng et al. *Bioinformatics*, 2016
  - Adjusting for biological variability (oscillatory genes)
  - Leng, Chu et al., *Nature Methods*, 2015

- **Identifying and characterizing differences in gene-specific expression distributions (aka. identifying differential distributions)**

- **Pseudotime reordering**
  - Korthauer et al., *Genome Biology*, 2016

- **Network reconstruction**
SCnorm: A quantile-regression based approach for robust normalization of single-cell RNA-seq data

Bacher, Chu et al., Nature Methods, 2017
Background

- Goal: correct for technical artifacts and/or gene-specific features
  - Sequencing depth
  - Length, GC content
  - Amplification and other technical biases
- Without UMIs/spike-ins, most single-cell methods calculate global scale factors as in bulk RNA-seq
  - One scale factor is calculated per sample and applied to all genes in that sample.
Bulk: Global scale-factor normalization for sequencing depth
Expression vs. depth varies with expression in scRNA-seq
We see the count-depth relationship varying with expression in many datasets
Overview of SCnorm

- Identify gene groups based on the count-depth relationship.

Within each group,

- Quantile polynomial regression is used to quantify the group-specific relationship between expression and sequencing depth. The quantile is chosen iteratively.

- Predicted values are used to calculate group-specific scale factors for each cell.
SConorm

- Filter: genes having greater than 10% expression values nonzero and median nonzero expression greater than 2.
- Let $Y_g = (y_{g1}, \ldots, y_{gJ})$ denote log non-zero expression for gene $g$ in cell $j$; $X_j$ denote log sequencing depth.
- The gene-specific count-depth relationship is estimated by:
  $$Q^{0.5} (Y_{g,j} | X_j) = \beta_{g,0} + \beta_{g,1}X_j$$
- Genes are split into $K$ groups. The group specific count-depth relationship is estimated by:
  $$Q^{\tau_k,d_k} (Y_j | X_j) = \beta_0^{\tau_k} + \beta_1^{\tau_k}X_j + \cdots + \beta_d^{\tau_k}X_j^{d_k}$$
- Estimates of $\tau_k$ and $d_k$ minimize $|\hat{\eta}_1^{\tau_k} - \text{mode} \hat{\beta}_{g,1}|$; where $\hat{\eta}_1^{\tau_k}$ represents the count-depth relationship among predicted values.
- $K$ is chosen so that the absolute value of the maximum normalized slope mode is $< 0.1$ within each of ten groups.
Bulk RNA-seq

SCnorm

Figure 1
Single-cell RNA-seq

SCnorm

CK  SAGES 2017
H1 - 1 (~ 1 million reads per cell)
H1 - 4 (~4 million reads per cell)
Implications for DE analysis

Un-normalized

Normalized Expression

SCnorm

Normalized Expression

Global Scale Factor

Normalized Expression
**FC= H1-1/H1-4**

- H1-1: ~100 H1 cells profiles at ~1 million reads per cell
- H1-4: Same H1 cells profiled at ~4 million reads per cell
- Prior to normalization, H1-1/H1-4 should be about $\frac{1}{4}$
- Post normalization, H1-1/H1-4 should be about 1
- If over-normalization is going on, H1-1/H1-4 will be greater than 1.
**FC= H1-1/H1-4**

- H1-1: ~100 H1 cells profiles at ~1 million reads per cell
- H1-4: Same H1 cells profiled at ~4 million reads per cell
Normalization via SConv

Figure 5: A density curve of the slopes from a quantile regression of log nonzero expression versus sequencing depth are plotted for LCR normalization on five public datasets. Genes are divided into 10 groups based on nonzero median raw expression. Genes were filtered for each dataset as those having greater than 10% nonzero expression. LCR was run with k = 7, 7, 5, 7, and 3 for IslamES, IslamEF, BuettnerG1, BuettnerS, and BuettnerG2M, respectively.
Challenges in scRNA-seq

- Normalization
- Technical vs. biological zeros
- De-noising
- Clustering; Identifying sub-populations
- Identifying oscillatory genes
- Identifying and characterizing differences in gene-specific expression distributions (aka. identifying differential distributions)
- Pseudotime reordering
- Network reconstruction
scDD: A Dirichlet mixture model based approach for identifying differential distributions in scRNA-seq experiments

Korthauer et al., Genome Biology, to appear, 2016
Gene-specific multi-modality

(A) Expression States of Gene X for Individual Cells Over Time

Low Expression State: $\mu_1$  
High Expression State: $\mu_2$

(B) Snapshot of Population of Single Cells

(C) Histogram of Observed Expression Level of Gene X
Many genes show multi-modal expression distributions

![Graph showing multi-modal expression distributions with histograms and bar charts. The x-axis represents the number of clusters, and the y-axis represents the count or proportion. The graphs display data for different datasets, with each dataset color-coded for easy comparison.](image_url)
Opportunity to identify differences beyond traditional DE

Differential expression (DE)

Differential proportions (DP)

Differential modes (DM)

Both DM and DE
scRNA-seq DE Analysis

- Recent methods use mixture modeling to account for ‘on’ and ‘off’ components
  - Shalek et al. (2014)
  - SCDE (Kharchenko et al., 2014)
  - MAST (Finak et al., 2015)

- When detected, each gene has a latent level of expression within a biological condition, and measurements fluctuate around that level due to biological and technical sources of variability
scDD: Goal

- Model expression profiles while accommodating the often multimodal distributions in the detected cells

- Find genes with Differential Distributions (DD) of expression across two conditions:
  - differential means
  - differential proportion within modes
  - differential modality (number of modes)
  - combination thereof
  - differential zeroes (detection rate)
scDD: Overview

- Assume that log non-zero normalized, de-noised, expression measurements $Y_g = (y_{g1}, ..., y_{gJ})$ for gene $g$ in $J$ cells arise from a conjugate Dirichlet Process Mixture (DPM) of normals model:

  \[
  y_j \sim N(\mu_j, \tau_j) \\
  \mu_j, \tau_j \sim G \\
  G \sim DP(\alpha, G_0) \\
  G_0 = NG(m_0, s_0, a_0/2, 2/b_0)
  \]

- Let $K$ denote the number of components (unique values in $\{\mu_j, \tau_j\}, j=1, ..., J$). Of primary interest is the posterior of $(\mu, \tau)$, which is intractable for moderate sample sizes.

- Let $Z = (z_1, ..., z_J)$ denote component memberships. Then $f(Y \mid Z)$ is a PPM.

  \[
  f(Y \mid Z) = \prod_{k=1}^{K} f(y^{(k)}) \\
  \propto \prod_{k=1}^{K} \frac{\Gamma(a_k/2)}{(b_k/2)^{a_k/2} s_k^{-1/2}}
  \]
scDD: Overview (continued)

- To quantify the evidence of DD for gene $g$, obtain MAP partition estimate, $\hat{Z}$, and evaluate $f(Y, \hat{Z})$ under competing hypotheses:
  - ignoring condition ($\mathcal{M}_{ED}$: equivalent distributions)
  - separately within condition ($\mathcal{M}_{DD}$: differential distributions)

- Evaluate $\mathcal{M}_{DD}$ using a pseudo-Bayes Factor score:

$$Score_g = \log \left( \frac{f(Y_g, \hat{Z}_g | M_{DD})}{f(Y_g, \hat{Z}_g | M_{ED})} \right)$$

- Assess significance via permutation.
scDD: Evaluation via simulation studies

- **8000 ED genes:**
  - 4000 from single Negative Binomial component
  - 4000 from two component mixture of Negative Binomial
- **2000 DD genes:**
  - 500 DE genes
  - 500 DP genes (0.33/0.66 proportion difference)
  - 500 DM genes (0.50 belong to second mode)
  - 500 DB genes (mean in second condition is average of means in the first)
- Sample sizes varied $\in \{50, 75, 100\}$
- Component distances $\Delta_\mu$, for multimodal conditions varied $\in \{2, 3, 4, 5, 6\}$ SDs
- Means, variances, and detection rates sampled empirically

Evaluate: Power to identify DD genes
- Rate at which DD genes are correctly classified
- Rate at which correct # components are identified
**scDD: Power to detect DD genes within each category**

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Method</th>
<th>True Gene Category</th>
<th>Overall (FDR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>scDD</td>
<td>0.893 <strong>0.418</strong> 0.898 0.572</td>
<td><strong>0.695</strong> (0.030)</td>
</tr>
<tr>
<td></td>
<td>SCDE</td>
<td>0.872 0.026 0.816 0.260</td>
<td>0.494 (0.004)</td>
</tr>
<tr>
<td></td>
<td>MAST</td>
<td><strong>0.908</strong> 0.400 0.871 0.019</td>
<td>0.550 (0.026)</td>
</tr>
<tr>
<td>75</td>
<td>scDD</td>
<td>0.951 0.590 <strong>0.960</strong> 0.668</td>
<td><strong>0.792</strong> (0.031)</td>
</tr>
<tr>
<td></td>
<td>SCDE</td>
<td>0.948 0.070 0.903 0.387</td>
<td>0.577 (0.003)</td>
</tr>
<tr>
<td></td>
<td>MAST</td>
<td><strong>0.956</strong> <strong>0.632</strong> 0.942 0.036</td>
<td>0.642 (0.022)</td>
</tr>
<tr>
<td>100</td>
<td>scDD</td>
<td>0.972 0.717 <strong>0.982</strong> 0.727</td>
<td><strong>0.850</strong> (0.033)</td>
</tr>
<tr>
<td></td>
<td>SCDE</td>
<td>0.975 0.125 0.946 0.478</td>
<td>0.631 (0.003)</td>
</tr>
<tr>
<td></td>
<td>MAST</td>
<td><strong>0.977</strong> <strong>0.752</strong> 0.970 0.045</td>
<td>0.686 (0.022)</td>
</tr>
<tr>
<td>500</td>
<td>scDD</td>
<td><strong>1.000</strong> 0.985 <strong>1.00</strong> 0.903</td>
<td><strong>0.972</strong> (0.034)</td>
</tr>
<tr>
<td></td>
<td>SCDE</td>
<td><strong>1.000</strong> 0.858 0.998 0.785</td>
<td>0.910 (0.004)</td>
</tr>
<tr>
<td></td>
<td>MAST</td>
<td><strong>1.000</strong> <strong>0.992</strong> <strong>1.00</strong> 0.174</td>
<td>0.792 (0.021)</td>
</tr>
</tbody>
</table>
Comparison of hESCs

Undifferentiated

Differentiated

Number of DD genes identified in each cell type comparison

<table>
<thead>
<tr>
<th>Comparison</th>
<th>DE</th>
<th>DP</th>
<th>DM</th>
<th>DB</th>
<th>DZ</th>
<th>Total</th>
<th>SCDE</th>
<th>MAST</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1 vs NPC</td>
<td>1342</td>
<td>429</td>
<td>739</td>
<td>406</td>
<td>1590</td>
<td>4506</td>
<td>2938</td>
<td>5729</td>
</tr>
<tr>
<td>H1 vs DEC</td>
<td>1408</td>
<td>404</td>
<td>939</td>
<td>345</td>
<td>880</td>
<td>3976</td>
<td>1581</td>
<td>3523</td>
</tr>
<tr>
<td>NPC vs DEC</td>
<td>1245</td>
<td>449</td>
<td>700</td>
<td>298</td>
<td>2052</td>
<td>4744</td>
<td>1881</td>
<td>5383</td>
</tr>
<tr>
<td>H1 vs H9</td>
<td>194</td>
<td>84</td>
<td>55</td>
<td>32</td>
<td>145</td>
<td>510</td>
<td>102</td>
<td>1091</td>
</tr>
</tbody>
</table>

scDD only: 2% 21% 38% 24% 15%
Genes identified in H1 vs. NPC comparison

Not DE by SCDE or MAST
Acknowledgements

Rhonda Bacher
Jeea Choi
Ziyue Wang
Jacob Maronge

Ning Leng, PhD
Keegan Korthauer, PhD
Shuyun Ye, PhD

Jamie Thomson, VMD, PhD
Ron Stewart, PhD
Li-Fang Chu, PhD
Scott Swanson, MS

Biomedical Computing Group
NIH R01GM076274
NIH U54 AI117924