Predicting Cell Phenotype using Single Cell RNASEq

Andrew McDavid

Department of Statistics, University of Washington
and
Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center
anmcd@uw.edu

June 10, 2016
One gene expression experiment to predict phenotypes in another?

- Want reproducibility. But technical variability and confounding is ubiquitous [Hicks et al., 2015].
- Phenotypic labels $Y$ in experiment $E$ with expression $X$.
- Want to generalize to new experiment $(E', X)$.
- $E$ and $E'$ may differ in technology, batch, tissue type, etc. Different distribution of $X$.
- Maybe different conditional distribution $Y|X$. 
Why bother?

- Not cost or experimentally effective to sort cells to desired granularities
- Sorting may paradoxically increase batch variability
- Leverage previous experiments
- Want to learn something new (unsupervised clustering)
Current approaches

- Vast literature for bulk [van’t Veer et al., 2002, Dudoit et al., 2002]
- Spatial localization [Satija et al., 2015] using linear discriminant analysis
- Cell cycle prediction [Scialdone et al., 2015] using PCA
- Lots of other machine learning tricks we could try: multinomial regression, neural networks, CART, SVM, random forests...
Regression-based approaches

- Regression-based classifiers have some advantages

\[ P(y_i = k|x_i, \beta) = f(\beta^T x_i) \]

- Interpretable classifiers (\( \beta \) might give us an odds ratio).
- Statistical calibration
Case study I (Bill Robinson and Daniel Liu, unpublished)

- B-cells sorted for **specificity** to Rheumatoid antigens; or antigen-negative. Several donors.
- Want to compare specificities.
- Heterogeneous population. (Potentially confounding) **subtype** differences: *naive, memory, plasmablast*...
- Some cells labeled with subtype; not feasible to sort all cells.
- Want to use the labeled subset to predict **subtype** elsewhere.
- $\sim 7000$ genes expressed in $> 5\%$ of cells.
Labeled subset

- Bulk: 3 patients with some technical replicates
- Single cell: 144 replicates, 1 patient
Bulk: 3 patients with some technical replicates

Single cell: 144 replicates, 1 patient

One specificity: antigen-negative.
Labeled subset

- Bulk: 3 patients with some technical replicates
- Single cell: 144 replicates, 1 patient
- One specificity: antigen-negative.
- Unique molecular identifiers (UMI) to label mRNAs
Method: sparse multinomial regression (\texttt{glmnet})

- Predict

\[ P(y_i = k | B(x_i), \beta) = \frac{e^{\beta_k^T B(x_i)}}{\sum_{l=1}^L e^{\beta_l^T B(x_i)}} , \]

- \( B(x_i) \) some basis expansion of log counts per million vector \( x \).

- Choices for \( B \):
  1. Raw \textbf{continuous} values \( x_{ij} \)
  2. \textbf{Discretized} threshold \( v_{ij} = 1_{x_{ij}>a} \).
  3. \textbf{Robust Z-scores} \( z_{ij} = (x_{ij} - m_i) / s_i \)
  4. \textbf{Ranks} \( r_{ij} = \sum_l 1_{x_{il}>x_{ij}} \)
  5. Combinations of these

- Vary tuning parameters \( \lambda \) and \( \alpha \), basis expansion \( B \), then choose best model. Should we believe the accuracy of this model?
Method: sparse multinomial regression (glmnet)

- Predict

\[
P(y_i = k | B(x_i), \beta) = \frac{e^{\beta_k^T B(x_i)}}{\sum_{l=1}^{L} e^{\beta_l^T B(x_i)}},
\]

\(B(x_i)\) some basis expansion of log counts per million vector \(x\).

- Choices for \(B\):
  1. Raw **continuous** values \(x_{ij}\)
  2. **Discretized** threshold \(v_{ij} = 1_{x_{ij}>a}\).
  3. **Robust Z-scores** \(z_{ij} = (x_{ij} - m_i)/s_i\)
      
      Cell location/scale transformation: \(x_{ij} \mapsto b_i x_{ij} + a_i\). Somewhat redundant with normalization.
  4. **Ranks** \(r_{ij} = \sum_{l} 1_{x_{il}>x_{ij}}\)
  5. **Combinations** of these

- Vary tuning parameters \(\lambda\) and \(\alpha\), basis expansion \(B\), then choose best model. Should we believe the accuracy of this model?
Method: sparse multinomial regression (glmnet)

- Predict

\[
P(y_i = k | B(x_i), \beta) = \frac{e^{\beta_k^T B(x_i)}}{\sum_{l=1}^{L} e^{\beta_l^T B(x_i)}},
\]

\(B(x_i)\) some basis expansion of log counts per million vector \(x\).

- Choices for \(B\):
  1. Raw **continuous** values \(x_{ij}\)
  2. **Discretized** threshold \(v_{ij} = 1_{x_{ij} > a}\).
  3. **Robust Z-scores** \(z_{ij} = (x_{ij} - m_i) / s_i\)
     Cell location/scale transformation: \(x_{ij} \mapsto b_i x_{ij} + a_i\). Somewhat redundant with normalization.
  4. **Ranks** \(r_{ij} = \sum_{l} 1_{x_{il} > x_{ij}}\)
     Cell monotone transformation: \(x_{ij} \mapsto g_i(x_{ij})\)
  5. Combinations of these

- Vary tuning parameters \(\lambda\) and \(\alpha\), basis expansion \(B\), then choose best model. Should we believe the accuracy of this model?
The First Paradox of Statistics

Optimization is like oxygen: the model needs it to function, but excess is damaging—or explosive.
Cross validation to assess the accuracy

Partition data into $K$ non-overlapping subsets. Train on $K - 1$, test on 1. Repeat $K$ times.

Independence of each test and training set ensures that the accuracy estimation is not optimistic.

from Elements of Statistical Learning
## In-sample results

<table>
<thead>
<tr>
<th>method</th>
<th>alpha</th>
<th>train</th>
<th>accuracy</th>
<th>seAccuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>dic:cont</td>
<td>0.50</td>
<td>single</td>
<td>0.97</td>
<td>0.01</td>
</tr>
<tr>
<td>rank</td>
<td>0.50</td>
<td>single</td>
<td>0.96</td>
<td>0.01</td>
</tr>
<tr>
<td>dic</td>
<td>0.50</td>
<td>single</td>
<td>0.96</td>
<td>0.00</td>
</tr>
<tr>
<td>rank</td>
<td>0.50</td>
<td>single</td>
<td>0.95</td>
<td>0.01</td>
</tr>
<tr>
<td>dic:cont:robustZ</td>
<td>0.70</td>
<td>single</td>
<td>0.95</td>
<td>0.01</td>
</tr>
<tr>
<td>dic:cont:robustZ</td>
<td>0.50</td>
<td>single</td>
<td>0.95</td>
<td>0.01</td>
</tr>
<tr>
<td>rank</td>
<td>0.70</td>
<td>single</td>
<td>0.95</td>
<td>0.02</td>
</tr>
<tr>
<td>dic</td>
<td>0.50</td>
<td>single</td>
<td>0.95</td>
<td>0.02</td>
</tr>
<tr>
<td>dic:cont:robustZ</td>
<td>0.50</td>
<td>single</td>
<td>0.95</td>
<td>0.01</td>
</tr>
<tr>
<td>dic:cont</td>
<td>0.70</td>
<td>single</td>
<td>0.94</td>
<td>0.02</td>
</tr>
</tbody>
</table>

### Key

dic = discrete, cont=continuous, robustZ = z-transformed

dic:cont is combination
## Validation results

<table>
<thead>
<tr>
<th>method</th>
<th>alpha</th>
<th>train</th>
<th>accuracy</th>
<th>seAccuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>dic:cont:robustZ</td>
<td>0.50</td>
<td>single</td>
<td>0.69</td>
<td>0.03</td>
</tr>
<tr>
<td>rank</td>
<td>0.50</td>
<td>single</td>
<td>0.68</td>
<td>0.03</td>
</tr>
<tr>
<td>dic</td>
<td>0.50</td>
<td>single</td>
<td>0.68</td>
<td>0.02</td>
</tr>
<tr>
<td>dic:cont:robustZ</td>
<td>0.70</td>
<td>single</td>
<td>0.68</td>
<td>0.03</td>
</tr>
<tr>
<td>dic:cont:robustZ</td>
<td>0.70</td>
<td>single</td>
<td>0.68</td>
<td>0.03</td>
</tr>
<tr>
<td>dic:cont:robustZ</td>
<td>0.50</td>
<td>single</td>
<td>0.68</td>
<td>0.03</td>
</tr>
<tr>
<td>dic</td>
<td>0.50</td>
<td>single</td>
<td>0.67</td>
<td>0.02</td>
</tr>
<tr>
<td>rank</td>
<td>0.50</td>
<td>single</td>
<td>0.67</td>
<td>0.04</td>
</tr>
<tr>
<td>dic:cont</td>
<td>0.50</td>
<td>single</td>
<td>0.67</td>
<td>0.03</td>
</tr>
<tr>
<td>dic:cont:rank</td>
<td>0.70</td>
<td>single</td>
<td>0.66</td>
<td>0.03</td>
</tr>
<tr>
<td>method</td>
<td>alpha</td>
<td>train</td>
<td>accuracy</td>
<td>seAccuracy</td>
</tr>
<tr>
<td>----------</td>
<td>-------</td>
<td>-------</td>
<td>----------</td>
<td>------------</td>
</tr>
<tr>
<td>dic:cont</td>
<td>1.00</td>
<td>single</td>
<td>0.61</td>
<td>0.02</td>
</tr>
<tr>
<td>dic:cont</td>
<td>1.00</td>
<td>single</td>
<td>0.58</td>
<td>0.02</td>
</tr>
<tr>
<td>rank</td>
<td>0.70</td>
<td>single</td>
<td>0.57</td>
<td>0.03</td>
</tr>
<tr>
<td>rank</td>
<td>1.00</td>
<td>single</td>
<td>0.57</td>
<td>0.04</td>
</tr>
<tr>
<td>rank</td>
<td>1.00</td>
<td>single</td>
<td>0.56</td>
<td>0.02</td>
</tr>
<tr>
<td>rank</td>
<td>0.70</td>
<td>single</td>
<td>0.54</td>
<td>0.02</td>
</tr>
<tr>
<td>dic:cont</td>
<td>0.70</td>
<td>single</td>
<td>0.53</td>
<td>0.05</td>
</tr>
<tr>
<td>rank</td>
<td>0.50</td>
<td>single</td>
<td>0.53</td>
<td>0.02</td>
</tr>
<tr>
<td>rank</td>
<td>0.50</td>
<td>single</td>
<td>0.53</td>
<td>0.02</td>
</tr>
<tr>
<td>dic:cont</td>
<td>0.70</td>
<td>single</td>
<td>0.49</td>
<td>0.03</td>
</tr>
<tr>
<td>Symbol</td>
<td>M</td>
<td>N</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>ELK2AP</td>
<td>-1.67</td>
<td>0.00</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td>RNA28S5</td>
<td>0.00</td>
<td>-2.16</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>IGJ</td>
<td>0.00</td>
<td>0.00</td>
<td>2.03</td>
<td></td>
</tr>
<tr>
<td>CIITA</td>
<td>0.00</td>
<td>0.39</td>
<td>-0.99</td>
<td></td>
</tr>
<tr>
<td>MIPOL1</td>
<td>-0.48</td>
<td>0.78</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>MS4A1</td>
<td>0.00</td>
<td>0.00</td>
<td>-1.22</td>
<td></td>
</tr>
<tr>
<td>RNA45S5</td>
<td>0.00</td>
<td>-0.19</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td>p12953</td>
<td>-1.15</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>BTG1</td>
<td>0.00</td>
<td>1.01</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>ACTG1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.86</td>
<td></td>
</tr>
</tbody>
</table>
Comparison to population estimates
Can we use the population-level estimates as a prior?

The $k = 1, \ldots, 6$ patients have differing B-cell flow fractions $\pi_k$.

**Multinomial probability**

\[
P(y_i = k|B(x_i), \beta, \alpha) = \frac{e^{\alpha_k + \beta_k^T B(x_i)}}{\sum_{l=1}^L e^{\alpha_l \beta_l^T B(x_i)}},
\]

Find $\alpha_k$ such that $P(y_i = k) = \sum_x P(y_i = k|B(x_i), \beta) = \pi_k$
Comparison to population estimates with margin adjustment
Could we discriminate with the unsupervised analysis? PCA, All genes

[Diagram showing scatter plots and density plots for PC1, PC2, PC3, and PC4, with data points and contour lines indicating distribution across the principal components.]
PCA, 53 classifier genes

53 genes (rank transformed)
TSNE, All genes

All genes (rank transformed)
TSNE, 53 classifier genes

53 genes (rank transformed)
333 genes, 930 cells in three cell lines (H9/MB-231/PC3), sorted by cell cycle (G0/G1, S, G2/M)

119 known, ranked genes associated with cell cycle from a bulk expression data base (cyclebase.org)

Can we predict cell cycle using this panel of genes? Does such a prediction generalize across cell lines?
Cross-validated and out-of-cell line accuracy
Conclusions

- Rank transformation is not uniformly best, but may be a good compromise
  Scialdone et al. [2015] finds a similar result.
- Don’t believe your training error!
- Unsupervised clustering is challenging; success depends on the nature of the signal.
Conclusions

- Rank transformation is not uniformly best, but may be a good compromise
  Scialdone et al. [2015] finds a similar result.
- Don’t believe your training error!
- Unsupervised clustering is challenging; success depends on the nature of the signal.
- Reproducibility of unsupervised clustering?
A call to arms

Broad interest in the field in supervised and unsupervised classification.

Reference data sets as comparators?

It’s hard to tune other people’s methods. UCI machine learning DB.

Desiderata:

- Continuous and ordinal phenotypes
  (time series, cell cycle, spatial location)
- Categorical
  Cell lineages, antigen stimulations
- Unsupervised clustering
Acknowledgments

**Stanford Medicine**
Dan Lui
Bill Robinson

**Nanostring Technologies**
Lucas Dennis
Patrick Danaher

**Fred Hutchinson**
Raphael Gottardo
Greg Finak

