Feature selection, dimensionality reduction, clustering, marker gene identification, and visualization in scRNA-Seq

Cihan Oguz
Bioinformatics Analyst
NIAID Collaborative Bioinformatics Resource (NCBR)
Leidos Biomedical Research, Inc.
October 3, 2019
Feature (gene) selection → Dimensionality reduction → Clustering → Marker gene identification
Feature (gene) selection

- Prior to dimensionality reduction, genes with highest expression variability (with enough read counts/above background) are identified.

- Typical input: Data normalized with the total expression in each cell, multiplied by a factor (e.g., 10,000) and log-transformed (not scaled data).

- 1000-5000 genes with the highest expression variability are selected.

- In robust workflows (e.g., Seurat and Scanpy), downstream analysis is not very sensitive to the exact number of selected genes.

- Expanded selection can help identify novel clusters with the risk of introducing additional noise into downstream analysis.

- Ideally, gene selection is done after batch correction.

- The goal is making sure genes variable only among batches (rather than cell groups within batches) do not dominate downstream results.
Dimensionality reduction of scRNA-Seq data

- scRNA-Seq data is inherently low-dimensional.

- Information in the data (expression variability among genes/cells) can be reduced from the number of total genes (1000s) to a much lower number of dimensions (10s).

- Dimensionality reduction generates linear/non-linear combinations of gene expression vectors for clustering & visualization.

- Major dimensionality reduction techniques for scRNA-Seq:
  - Principal component analysis (PCA)
  - Most commonly used ones: UMAP and t-SNE (inputs: PCA results)
  - UMAPs typically preserve more of global structure with shorter run times
  - Other alternatives: Diffusion Maps & force-directed layout with k-nearest neighbors
Scaling normalized data & performing PCA

- PCA is performed on the scaled data.
- Scaled data represented as z-scores.
- Mean=0 & variance=1 for each gene.
- z-scoring makes sure that highly-expressed genes do not dominate.

Elbow plots show the number of PCs to include moving forward (URD tool in R can automatically detect the elbow).

PC score plots show genes that dominate each PC

PC heatmaps visualize anti-correlated gene sets (yellow: higher expression)

Jackstraw analysis generates a p-value (significance) of each PC 1% of the data is randomly permuted, PCA is rerun, ‘null distribution’ of gene scores constructed (these steps repeated many times).

‘Significant’ PCs have a strong enrichment of low p-value genes.
The effect of regressing out cell cycle phase scores on PCA results

• What if majority of variance in top PCs are dominated by a specific set of genes that are not of biological interest?

• Example: Cell cycle heterogeneity in a murine hematopoietic progenitor data set.

• Scores computed for each phase based on canonical markers.

• Each cell mapped to a cell cycle phase (highest scoring phase)

• To differentiate between cycling and non-cycling cells, |G2/M-S| score difference is regressed out.
t-SNE: t-Distributed Stochastic Neighbor Embedding

- One of the two most commonly used nonlinear dimensionality reduction techniques
- Used for embedding high-dimensional data for onto a low-dimensional space of 2 or 3 dimensions
- Can reveal local data structures efficiently without overcrowding
- t-SNE creates a probability distribution using the Gaussian distribution (defines the relationships between points in high-dimensional space).
- Uses the Student t-distribution to recreate the probability distribution in low-dimensional space.
- This prevents the crowding problem (points tend to get crowded in low-dimensional space/curse of dimensionality).
- t-SNE optimizes the embeddings directly using gradient descent.
- Cost function is non-convex with risk of getting stuck in local minima (t-SNE avoids poor local minima).
- Perplexity parameter serves as a knob that sets the number of nearest neighbors

**t-SNE** tries to recreate a low dimensional space that follows the probability distribution dictating the relationships between various neighboring points in higher dimensions.
UMAP: Uniform Manifold Approximation and Projection

- UMAP is another manifold learning technique for dimensionality reduction
- Four major UMAP parameters control its topology

**Number of neighbors per cell on the UMAP**
- Balances local vs global structure in the data by constraining the size of the local neighborhood
- Low values force UMAP to concentrate on very local structure (potentially to the detriment of the big picture),
- Large values will push UMAP to look at larger neighborhoods of each point when estimating the manifold structure of the data, losing fine detail structure

**Minimum distance between cells on the UMAP**
- Controls how tightly cells are packed together
- Effective minimum distance between embedded points
- Low values lead to clumped nearby cells (finer topological structure)
- High values prevent packing points together (clusters get closer)
- High values preserves broad topological structure at the expense of finer topological details
Number of UMAP dimensions

- Reduced data can be embedded into 2, 3, or higher dimensions

Cosine similarity & Pearson/Spearman correlation are scale invariant (driven by relative differences between cells, robust to library or cell size differences)

UMAP distance metric (cell to cell)

- The metric used to measure distance between cells in the input space
- Examples: Euclidean, Manhattan, and Minkowski
- Angular metric: Cosine similarity
- Pearson and Spearman correlation based metrics
Diffusion map: Each DC (diffusion map dimension) highlights the heterogeneity of a different cell population.

Connectivity \sim \text{probability of walking between the points in one step of a random walk (diffusion)}

Force-directed graph layout via ForceAtlas2

Nodes repulse each other like charged particles, while edges attract their nodes, like springs.

Various dimensionality reduction methods applied to mouse intestinal epithelium data

Clustering cells with similar expression profiles together

• Unsupervised machine learning problem
  – Input: distance matrix (cell-cell distances)
  – Output: Cluster membership of cells

• Cells grouped based on the similarity of their gene expression profiles
  – Distance measured in dimensionality-reduced gene expression space (scaled data)

• k-means clustering divides cells into k clusters
  – Determines cluster centroids
  – Assigns cells to the nearest cluster centroid
  – Centroid positions iteratively optimized (MacQueen, 1967).
  – Input: number of expected clusters (heuristically calibrated)

• k-means can be utilized with different distance metrics
• Alternatives to standard Euclidean distance:
  – Cosine similarity (Hagghverdi et al, 2018)
  – Correlation-based distance metrics (Kim et al, 2018)
  – SIMLR method learns a distance metric using Gaussian kernels (Wang et al, 2017)
Using community detection & modularity optimization for finding clusters

• Community detection methods utilize graph representation derived from k-nearest neighbors (kNN)

• Then, the modularity function is optimized to determine clusters.

• Typical range of k is 5-100

• Densely sampled regions of expression space are represented as densely connected regions of the graph.

• Community detection is often faster than clustering as only neighboring cell pairs have to be considered as belonging to the same cluster.

• Optimized modularity function includes a resolution parameter, which allows the user to determine the scale of the cluster partition.
Number of clusters and biological context

- Number of clusters is a function of the resolution parameter.

- Multiple resolution values can be explored to see the interplay between resolution and UMAP or t-SNE plots for a given data set.

- Biological context can be used for guidance.

- Examples: Expected number of major cell types or subtypes.

- Isolating a cluster to identify sub-clusters can generate useful biological insights (e.g., differential expression between cellular subtypes in a cluster).

- If cluster-specific markers for multiple clusters overlap (e.g., ribosomal genes), these clusters can be merged without losing much information regarding cell subtypes.
### Clustering methods for scRNA-Seq data

<table>
<thead>
<tr>
<th>Name</th>
<th>Year</th>
<th>Method type</th>
<th>Strengths</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>scanpy</td>
<td>2018</td>
<td>PCA + graph-based</td>
<td>Very scalable</td>
<td>May not be accurate for small data sets</td>
</tr>
<tr>
<td>Seurat</td>
<td>2016</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phenograph</td>
<td>2015</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SC3</td>
<td>2017</td>
<td>PCA + k-means</td>
<td>High accuracy through consensus, provides estimation of k</td>
<td>High complexity, not scalable</td>
</tr>
<tr>
<td>SIMLR</td>
<td>2017</td>
<td>Data-driven dimensionality reduction + k-means</td>
<td>Concurrent training of the distance metric improves sensitivity in noisy data sets</td>
<td>Adjusting the distance metric to make cells fit the clusters may artificially inflate quality measures</td>
</tr>
<tr>
<td>CIDR</td>
<td>2017</td>
<td>PCA + hierarchical</td>
<td>Implicitly imputates dropouts when calculating distances</td>
<td></td>
</tr>
<tr>
<td>GiniClust</td>
<td>2016</td>
<td>DBSCAN</td>
<td>Sensitive to rare cell types</td>
<td>Not effective for the detection of large clusters</td>
</tr>
<tr>
<td>pcaReduce</td>
<td>2016</td>
<td>PCA + k-means + hierarchical</td>
<td>Provides hierarchy of solutions</td>
<td>Very stochastic, does not provide a stable result</td>
</tr>
<tr>
<td>Tasic et al.</td>
<td>2016</td>
<td>PCA + hierarchical</td>
<td>Cross validation used to perform fuzzy clustering</td>
<td>High complexity, no software package available</td>
</tr>
<tr>
<td>TSCAN</td>
<td>2016</td>
<td>PCA + Gaussian mixture model</td>
<td>Combines clustering and pseudotime analysis</td>
<td>Assumes clusters follow multivariate normal distribution</td>
</tr>
<tr>
<td>mpath</td>
<td>2016</td>
<td>Hierarchical</td>
<td>Combines clustering and pseudotime analysis</td>
<td>Uses empirically defined thresholds and a priori knowledge</td>
</tr>
<tr>
<td>BackSPIN</td>
<td>2015</td>
<td>Biclustering (hierarchical)</td>
<td>Multiple rounds of feature selection improve clustering resolution</td>
<td>Tends to over-partition the data</td>
</tr>
<tr>
<td>RaceID1, RaceID2, RaceID3</td>
<td>2015</td>
<td>k-Means</td>
<td>Detects rare cell types, provides estimation of k</td>
<td>Performs poorly when there are no rare cell types</td>
</tr>
<tr>
<td>SINCERA</td>
<td>2015</td>
<td>Hierarchical</td>
<td>Method is intuitively easy to understand</td>
<td>Simple hierarchical clustering is used, may not be appropriate for very noisy data</td>
</tr>
<tr>
<td>SNN-Cliq</td>
<td>2015</td>
<td>Graph-based</td>
<td>Provides estimation of k</td>
<td>High complexity, not scalable</td>
</tr>
</tbody>
</table>

Each method with own strengths & limitations.

Seurat, Phenograph, and scanpy are the most popular methods (only limitation: accuracy for small data sets)

Other methods are mainly limited in their scalability, stability (stochastic), and ability to handle very noisy data.

---

Marker gene identification

Differential expression approaches for marker identification:
- Wilcoxon rank sum test and student’s t-test
- Logistic regression
- DESeq2: Negative binomial generalized linear models (read counts) & Wald test for significance.
- MAST: GLMs in which cellular detection rate is treated as a covariate
- GLMs are flexible and do not make assumptions (homogenous distributions of residuals/fitting errors or normally distributed variances).

Classifier based approach for marker identification:
- Classifiers built with normalized expression levels (one classifier per gene).
- Genes ranked with respect to their ability of each gene to distinguish between two groups of cells (e.g. KO vs WT, cluster 1 vs 2, or cluster 1 vs all clusters).
- Area under each ROC curve represents the predictive power of the gene.

MAST (Model-based Analysis of Single-cell Transcriptomics)

- Accounts for the fact that the number of cells expressing a gene varies from gene to gene.

- The fraction of genes expressed, or cellular detection rate (CDR) correlates with top PCs of variation.

- Modeling CDR as a covariate controls for differences in abundance due to cell size and other extrinsic biological and technical effects.

- MAST has been tested against differential expression methods developed for bulk RNA-Seq (limma, edgeR, and DESeq) in Finak et al (2015).

- MAST was found to generate GO enrichment profiles biologically more relevant to mucosal-associated invariant T cell activation and LPS-stimulated myeloid dendritic cells in Finak et al (2015).

• Identifying “markers” of clusters using ROC analysis.
• For each gene, a classifier built on that gene alone to classify between two groups of cells.
• Classifier performance evaluated using AUC.
• TP, FP, TN, and FN are computed at different expression thresholds.
• An AUC value of 1 means that expression values for this gene alone can perfectly classify the two groupings.
• All cells in C1 exhibit higher expression than all cells in C2 (AUC=1).
• A value of 0.5 implies that the gene has no predictive power to classify the two clusters.

Cluster 1 vs Cluster 2
Positive class: C1, Negative class: C2

True positive (TP)
True membership: C1, Prediction: C1

False positive (FP)
True membership: C2, Prediction: C1

Can the expression level of a gene correctly predict the cluster membership at different expression thresholds?
A closer look at the ROC calculations

Cluster 1 vs Cluster 2
Positive class: C1, Negative class: C2

True positive (TP)
True membership: C1, Prediction: C1

False positive (FP)
True membership: C2, Prediction: C1

False negatives (i.e., those members of the “positive” population who tested negative because their test results happened to be lower than the positive cut-off threshold)

False positives (i.e., those members of the “negative” population who tested positive because their test results happened to be higher than the positive cut-off threshold)
### A typical differential expression analysis output

<table>
<thead>
<tr>
<th></th>
<th>p_val</th>
<th>avg_logFC</th>
<th>pct.1</th>
<th>pct.2</th>
<th>p_val_adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>S100A9</td>
<td>0</td>
<td>3.860873</td>
<td>0.996</td>
<td>0.215</td>
<td>0</td>
</tr>
<tr>
<td>S100A8</td>
<td>0</td>
<td>3.796840</td>
<td>0.975</td>
<td>0.121</td>
<td>0</td>
</tr>
<tr>
<td>LGALS2</td>
<td>0</td>
<td>2.634295</td>
<td>0.908</td>
<td>0.059</td>
<td>0</td>
</tr>
<tr>
<td>FCN1</td>
<td>0</td>
<td>2.352693</td>
<td>0.952</td>
<td>0.151</td>
<td>0</td>
</tr>
<tr>
<td>CD14</td>
<td>0</td>
<td>1.951644</td>
<td>0.667</td>
<td>0.028</td>
<td>0</td>
</tr>
<tr>
<td>TYROBP</td>
<td>0</td>
<td>2.111879</td>
<td>0.994</td>
<td>0.285</td>
<td>0</td>
</tr>
</tbody>
</table>

#### Comparing gene expression in different cell groups:
- Cluster 1 vs Cluster 2
- Cluster 1 vs all other clusters
- Cluster1_KO vs Cluster1_WT

#### Tips for marker identification
- Marker identification can take time with thousands of cells and genes
- Prefiltering cells and genes can reduce the computational time significantly
- Genes rarely detected in either group of cells, are not likely to be differentially expressed
- Genes with small fold-change can also be excluded
- Typically, only upregulated genes (>1 FC) are relevant for cluster-specific marker discovery

### Average log FC
- Ratio of expression in log-space
- pct.1 = percent of cells in Cluster 1 in which the gene is detected
- pct.2 = percent of cells in Cluster 2
- p_val_adj = FDR

### Classifier based marker identification: AUC values replace the p-values
Visualization with violin plots, heatmaps, and ridgeline plots

Violin plots show smoothed probability density distributions of expression

Ridgeline plots visualize finer details (bimodality) of expression distributions

Heatmaps visualize significant expression differences between clusters through contrasting colors

Subsampling of cells: choosing a subset of the whole cell population to avoid having to draw extremely large heatmaps
Visualization with dot plots and feature plots

Dot plots show the average expression levels

- Visualizing scaled vs normalized expression levels
- Scaled data typically magnifies the differences between clusters
- Size of dots proportional to the percentage of cells that express the gene

Feature plots

- Visualizing how the expression of the gene is distributed among cells in the reduced space
- Can choose cells to plot, reduction method to use (PCA, t-SNE, or UMAP), or a quantile expression cut-off.
Visualization with dot plots and feature plots

Feature plots split by different experimental conditions

Feature plots split by different experimental conditions & clusters
More ways to visualize cluster-specific expression patterns

Changes in average expression in different clusters or varying experimental conditions

Visualizing co-expression among gene pairs

Cluster-specific co-expression of genes can provide insights regarding activation or inhibition of pathways in cell subpopulations
Visualization of differential expression & enrichment results

Volcano plot

log2FC vs. significance (-log10(p-value))

https://galaxyproject.github.io/training-material/topics/transcriptomics/tutorials/rna-seq-viz-with-volcanoplot/tutorial.html

Cluster-specific enriched pathways or GO terms (circle size ~ fraction of cluster-specific markers in the enriched pathway)

Dot plot (Clusterprofiler in R)
Wrapping-up

- Feature (gene) selection
  - Identifying genes with high expression variance
- Dimensionality reduction (PCA, UMAP, and t-SNE)
- Clustering
  - K-nearest neighbors, community detection, and modularity optimization
- Marker gene identification
  - Differential expression and classifier based approaches
- Visualization
  - Violin plots, dot plots, heatmaps, volcano plots, and scatter plots with the t-SNE/UMAP coordinates of cells

Acknowledgements:

NIAID Collaborative Bioinformatics Resource (NCBR)

Justin Lack (Lead), Arun Boddapati, Susan Huse, Vasu Kuram, Tovah Markowitz, Paul Schaughency