Slingshot:

Kelly Street

1 Abstract

Recently, single-cell RNA-Seq has afforded researchers an unprecedentedly detailed view of cellular transcription. Communities of heterogeneous cells that could previously only be interrogated collectively can now reveal multiple functionally distinct groups with complex relationships. One common target for these studies has been stem cells and their descendants. Mapping transcriptional progression from stem cell populations to specialized cell types has become crucial for properly understanding these systems and many statistical and computational methods have been proposed. Slingshot is a uniquely robust and flexible tool for inferring developmental lineages and ordering cells to reflect continuous differentiation processes. It constructs a differentiation tree using clusters of cells as nodes, which provides stability and reduces the complexity of the inferred lineages. This map is used to assign individual cells to one or more developmental lineages, which are represented by smooth curves in a reduced dimension space. These curves provide discerning power not found in methods based on piecewise linear trajectories while also adding stability over a range of possible dimensionality reduction and clustering techniques. [real and simulated data, compare to X, Y, and Z methods]

2 Introduction

Bio context, translation into stat question, lit review

Traditional/bulk RNA-Seq affords us a bird's-eye view of transcription, often mixing different cell types into one sample. Single-cell RNA-Seq can give us a much more detailed picture. Higher resolution means the ability to distinguish closely related populations of cells and characterize their relationships.

For some cell types, there may not be a clear distinction, but rather a smooth transition with individual cells existing on a continuum between known states. In these cases, we can model the relationship between states as a continuous process dependent on a single underlying variable, often called "pseudotime."

3 Methods

3.1 Normalization

After read mapping and expression quantification, a typical RNA-Seq analysis will perform some sort of normalization. Raw gene expression profiles need to be normalized prior to lineage reconstruction to remove unwanted technical effects that bias expression measures. We used SCONE to compare various normalization methods and select the one that performs best on our data.

3.2 Clustering

Using unsupervised clustering to find biologically distinct classes of cells is another feature of many single-cell RNA-Seq data analysis pipelines. Clustering cells provides stability in downstream analysis by dramatically reducing the number of possible relationships between different cells and lessening the potential impact of outliers on the final ordering. We denote these clusters as non-overlapping subsets of cells, $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$.

Our data were clustered with the `clusterExperiments` package, which uses an iterative procedure to find tight, stable clusters. To increase the stability of the clustering process and lineage reconstruction, a small percentage of cells may be left unclustered. These cells can be ordered by the final set of lineages, but will not be used in their construction.

3.3 Dimensionality Reduction

Before constructing lineageas, we recommend finding a suitable low-dimensional representation of the data. There are a wide variety of methods available for this purpose and some may be more appropriate than others for particular datasets. We note only that since Slingshot's default lineage reconstruction method is nonlinear, there is no reason to believe that linear dimensionality reduction methods cannot perform well. By way of example, we use PCA for our data [or for a particular dataset]. We will format this reduced-dimensionality dataset in an $n \times p$ matrix denoted by $\mathbf{X}$.

3.4 Mapping Cell Types

Slingshot identifies relationships between clusters by treating them as nodes in a graph and drawing a minimum spanning tree (MST). Distances between clusters depend on both the Euclidean distance between their centers and their respective shapes. By default, Slingshot will use the full covariance matrix of both clusters in the reduced dimensionality space to calculate the distance:

$$d(\mathcal{C}_i, \mathcal{C}_j) = (\bar{x}_i - \bar{x}_j)^T (S_i + S_j)^{-1} (\bar{x}_i - \bar{x}_j)$$

Where $\bar{x}_i$ represents the center (mean) of cluster i and S_i is the empirical covariance matrix. In the presence of small clusters, the matrix $S_i + S_j$ may not be invertible and we instead use $\tilde{S}_i + \tilde{S}_j$, where $\tilde{S}_i$ is 0 for all off-diagonal elements involving dimensions higher than the minimum cluster size.

This allows Slingshot to draw trees that are better covered by and representative of the cells in a dataset. For datasets in which there are outlying clusters or more than one tree is appropriate, Slingshot includes a granularity parameter that effectively limits the maximum edge length in the tree. [By default, we look for a large jump (more than 3 times the average jump) in the ordered edge lengths of the MST and if there is one, we set the limit at the halfway point of that jump].

3.5 Lineage Identification

Lineages are defined as ordered sets of clusters created by tracing paths through the MST. Given a pre-specified origin cluster, every direct path from this cluster to a leaf node will be returned as a lineage. In the simple case where the MST has only two leaf nodes, this results in one lineage if the origin cluster is a leaf and two lineages if it is an interior node. Clusters with more than

two connections will create bifurcations and produce additional lineages. In the absence of a pre-specified origin, Slingshot will select an origin based on parsimony, producing a set of lineages with the maximal number of clusters shared between them.

3.6 Ordering and Pseudotime Calculation

The next step is to model each of these lineages with a smooth curve. This is achieved by an iterative procedure similar to the principal curve algorithm. The piecewise linear path through the cluster centers is taken as the initial curve and pseudotime values are obtained by each points orthogonal projection onto the curve. The curve is updated using a smoothing spline to predict each dimension as a function of pseudotime and this process is repeated until convergence. When there are branching lineages with the same origin, a shrinkage step will be included at each iteration that forces a degree of similarity between the curves in the neighborhood of shared clusters. This typically ensures that the final curves will form a tree structure. Final pseudotime values are derived from the each points orthogonal projection onto the curves. Thus, cells belonging to clusters that are included in multiple lineages will have multiple pseudotime values, but these values should agree fairly well.

3.7 DE Genes

4 Datasets

5 Results

6 Discussion