Analysis of single-cell RNA-seq data with R and *Bioconductor*.

Fanny Perraudeau, Kelly Street, Davide Risso

Tasks and packages

- Dimensionality reduction with zinbwave.
  - bioconductor.org/packages/zinbwave
- Cluster analysis with clusterExperiment.
  - bioconductor.org/packages/clusterExperiment
- Lineage inference and trajectory analysis with slingshot.
  - github.com/kstreet13/slingshot

Workshop material:
github.com/fperraudeau/bioc2017singlecell

Acknowledgements

UC Berkeley

Experiments and analyses

Russell Fletcher
Diya Das
John Ngai Lab

Methods development

Sandrine Dudoit
Elizabeth Purdom
Nir Yosef
Svetlana Gribkova
JP Vert

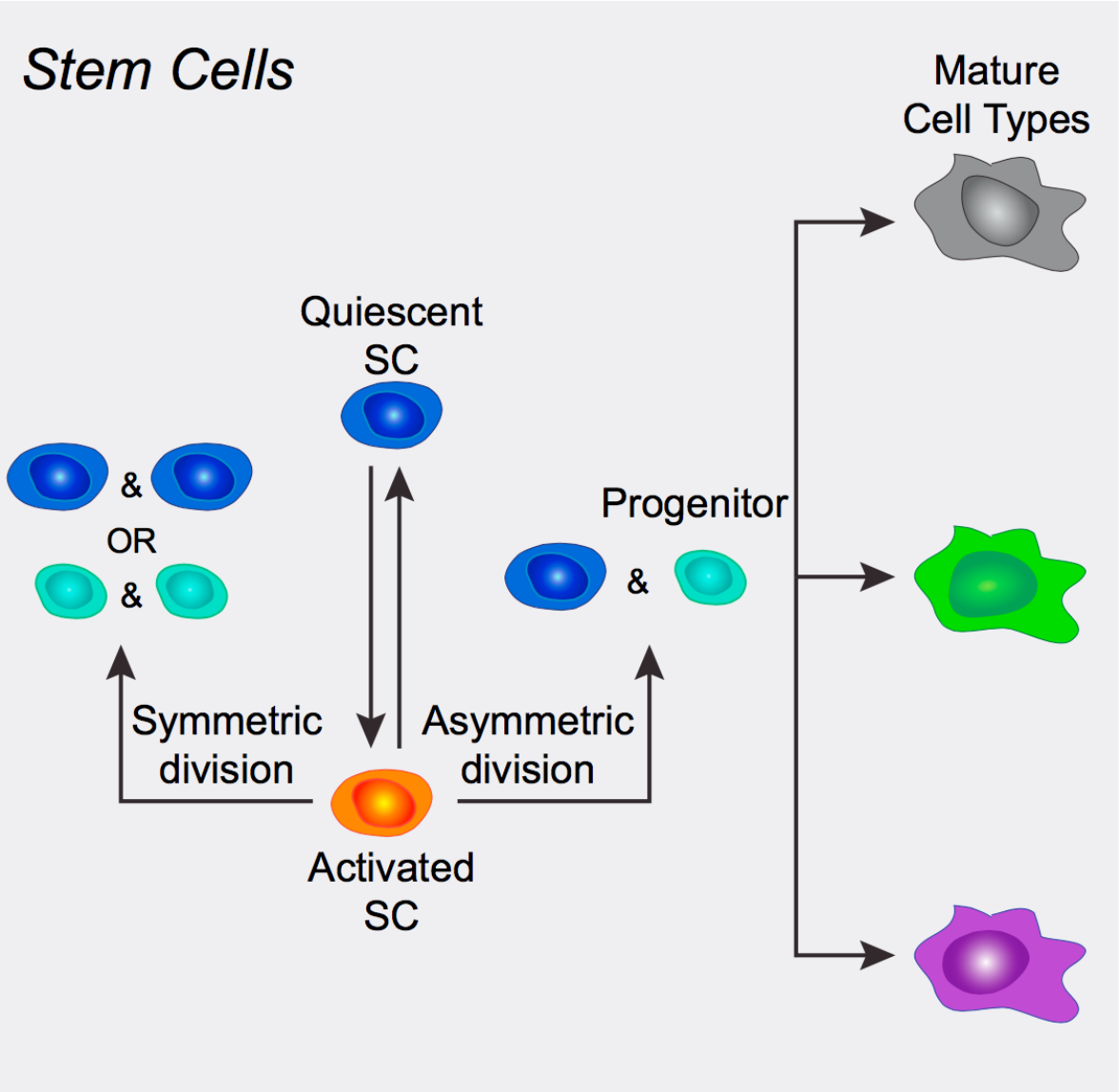
Core Facilities

QB3 Functional Genomics Laboratory
QB3 Genomics Sequencing Laboratory
Cancer Research Laboratory

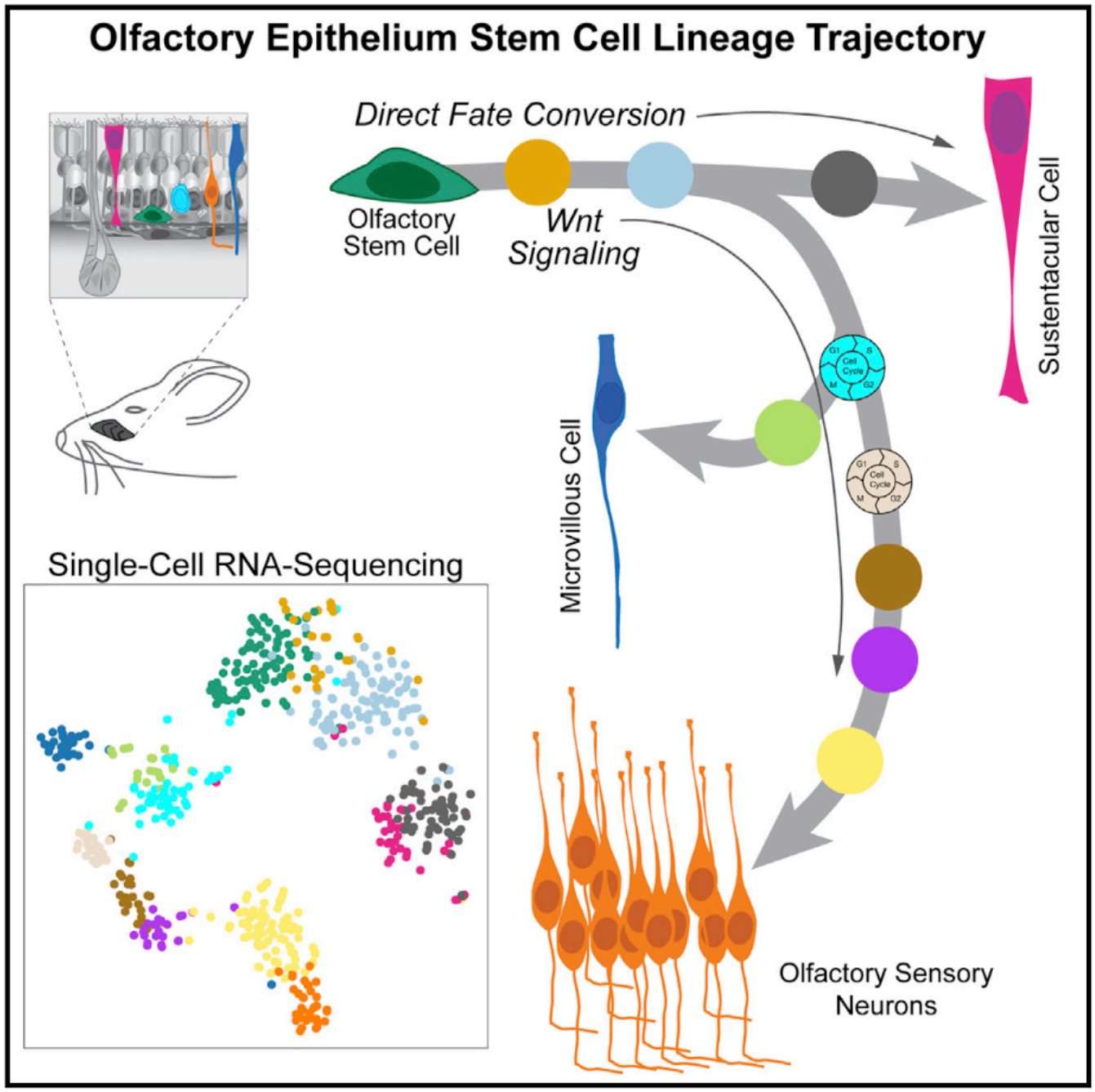
Funding

NIH BRAIN Initiative Cell Census Consortium

Adult stem cells maintain and regenerate tissues



Bond, Ming, Song (2015) Cell Stem Cell



Fletcher et al. (2017)
Cell Stem Cell

Useful links

- ZINB-WaVE preprint:
  - https://doi.org/10.1101/125112
- Slingshot preprint:
  - https://doi.org/10.1101/128843

- Please submit bug reports and other issues at:
  - github.com/drisso/zinbwave/issues
  - github.com/epurdom/clusterExperiment/issues
  - github.com/kstreet13/slingshot/issues

Contact us!

- E-mail
  - fperraudeau@berkeley.edu
  - kstreet@berkeley.edu
  - risso.davide@gmail.com
- Twitter:
  - @fannyperraudeau
  - @KStreetBerkeley
  - @drisso1893

RSEC: Resampling-based Sequential Ensemble Clustering

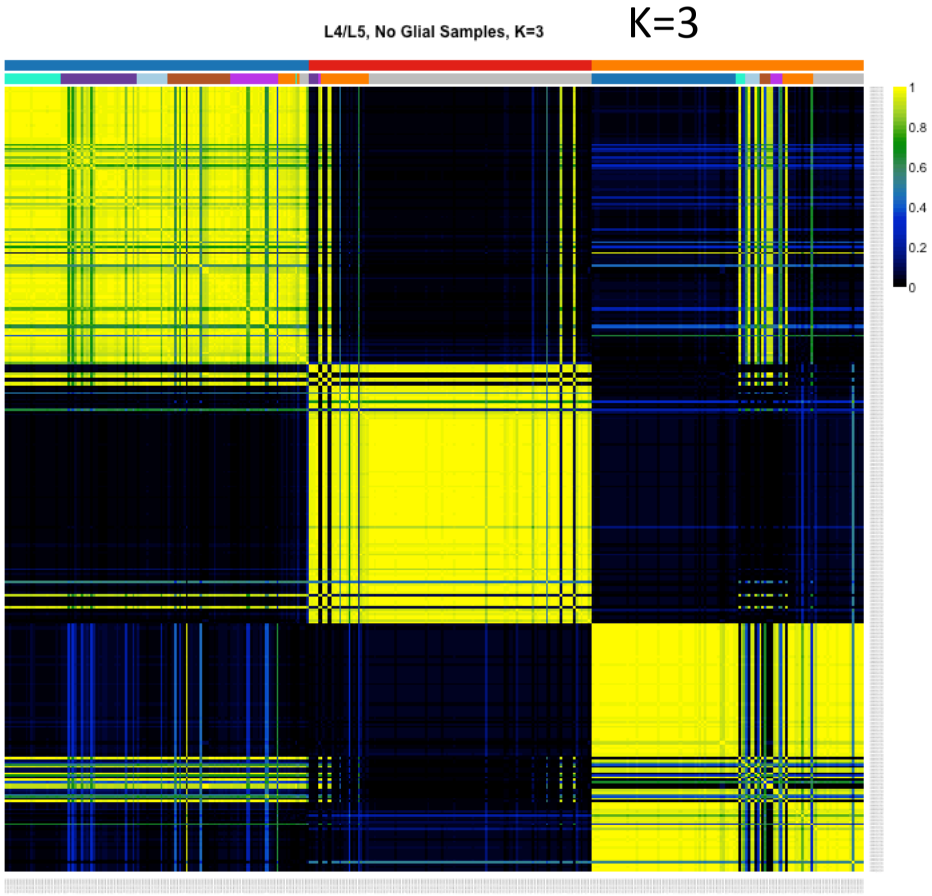
- Use a clustering routine that finds a large number of small, coherent clusters:
  - Subsampling of data to find robust clusters
  - Sequential clustering → find a group of coherent samples, remove them, start over
- Perform this routine over *many* different parameters
- Find a single consensus over the clusterings
- Merge together non-differential clusters
- Find biomarkers via differential expression with targeted comparisons
- Implemented with visualization tools in `clusterExperiment` package

What mean by subsample clustering?

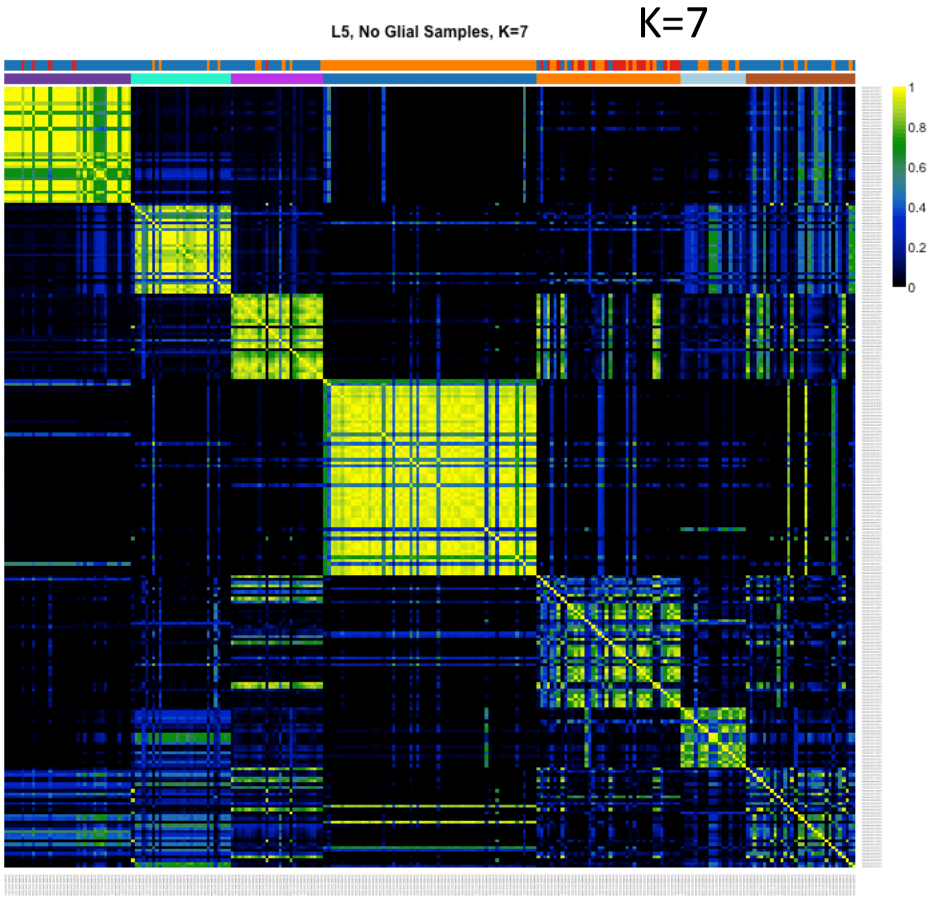
- Pick an underlying clustering strategy (e.g. kmeans or PAM with particular choice of K)
- Repeat the following:
  - Subsample the data, e.g. 70% of samples
  - Find clusters on the subsample
- Create coClustering matrix D: % of subsamples where samples were in same cluster

Examples of matrix D

K=3



K=7



Note, here forced the samples in order given by PAM
Also used kmeans in resampling, rather than PAM

What mean by subsample clustering?

- Pick an underlying clustering strategy (e.g. kmeans or PAM with particular choice of K)
- Repeat the following:
  - Subsample the data, e.g. 70% of samples
  - Find clusters on the subsample
- Create coClustering matrix D: % of subsamples where samples were in same cluster
- Cluster matrix D for final clustering
  - Could be with original clustering strategy, or different one
  - "Right" K may not be the K used in original clustering (e.g. kmeans)
  - We use a more flexible approach of hierarchical clustering and picking clusters so have at least 1-α similarity
  - Change from picking K to picking α, more intuitive choice
  - Not all samples get clustered

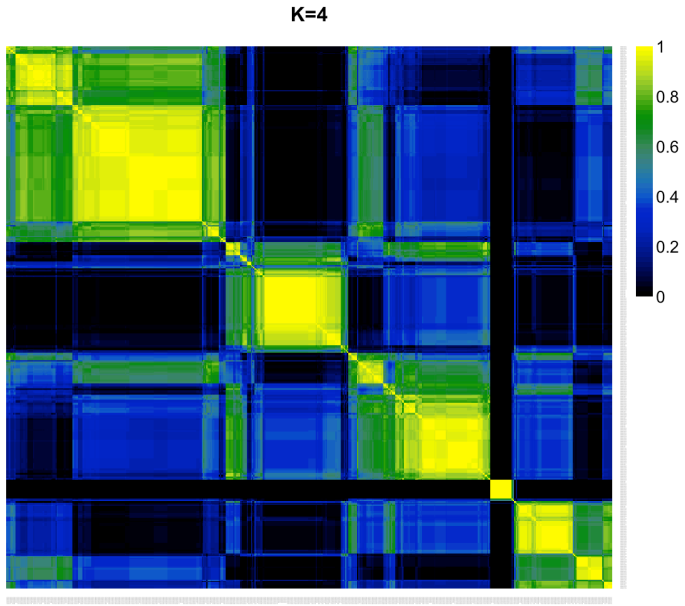
What mean by sequential clustering?

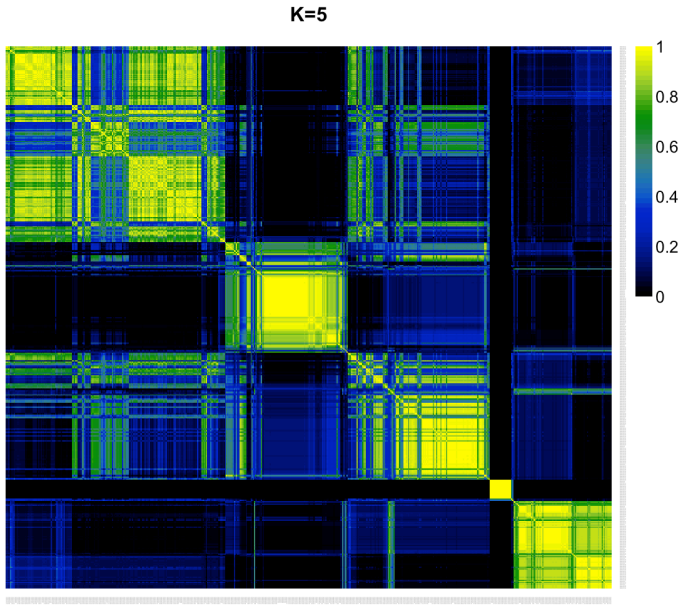
- Over range of starting parameters, do clustering
- The cluster that stays at least β similar, identify as cluster and remove
- Repeat until no more such clusters found, or not enough samples left
- Draws on ideas of "tight clustering" of genes of Tseng and Wong (2005)

Specifically,

- We range over K in underlying PAM in subsampling
- We find clusters based on results of subsample clustering (so may not be same K as input parameter)

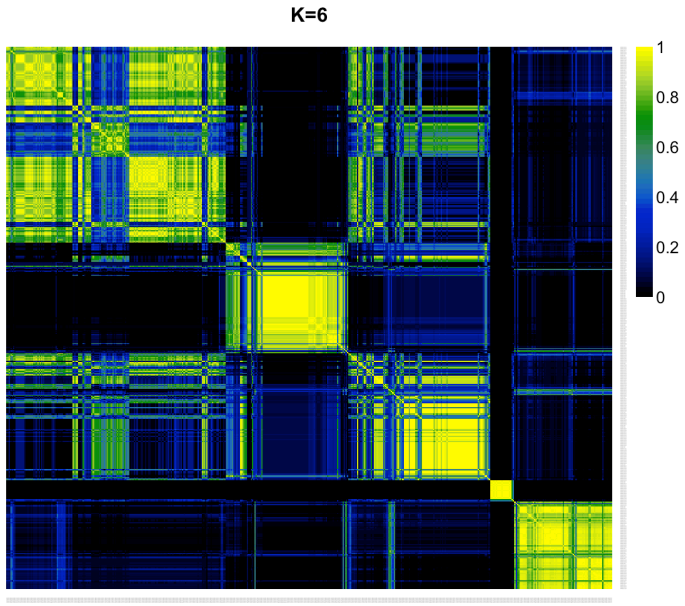
K=4



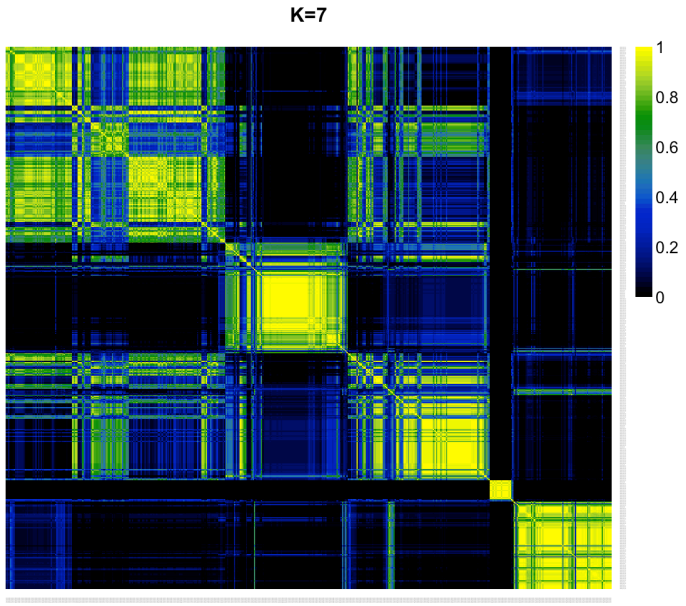


K=5

Because of subsampling, changing K is more a perturbation



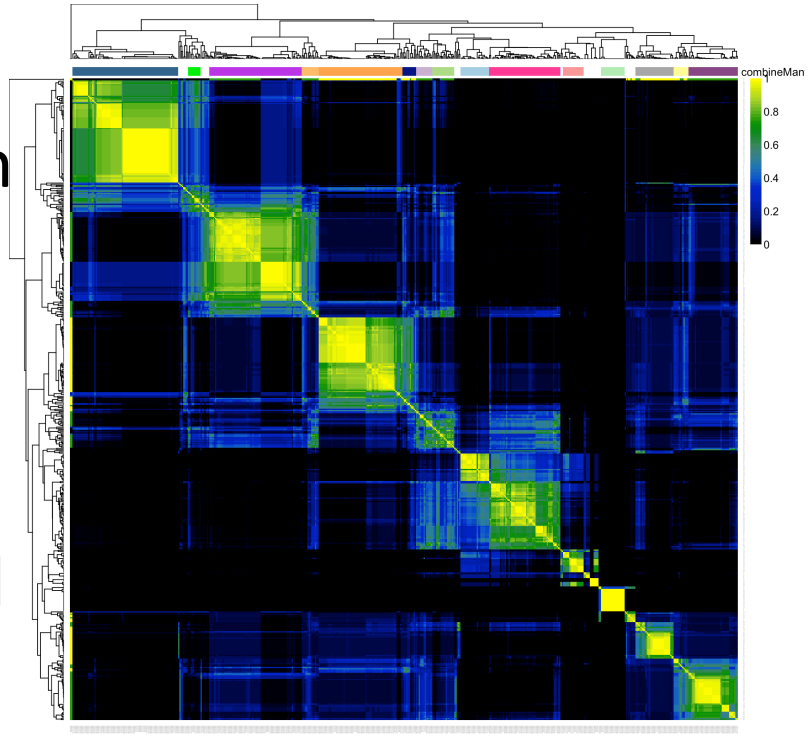
K=6



K=7

Find consensus cluster

- Create a co-Clustering matrix D of how many times co-cluster together and cluster D
  - Like with subsampling, on now across different parameters
  - Again, important that clusters are largely perturbations, not radical different clusters



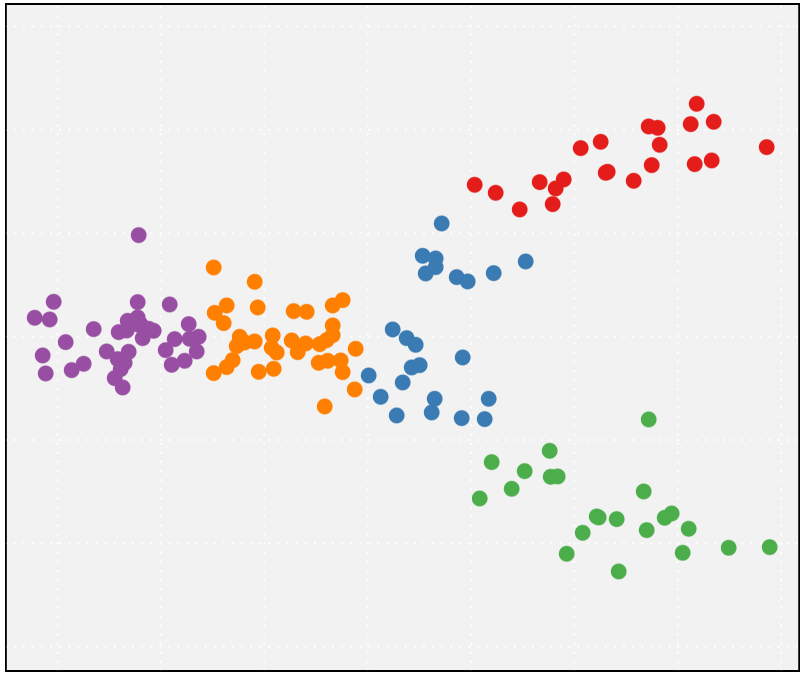
Bioconductor workflow for single-cell RNA-seq data analysis: dimensionality reduction, clustering, and pseudotime ordering.

Fanny Perraudeau, Davide Risso, Kelly Street

BioC2017

July 28th, 2017

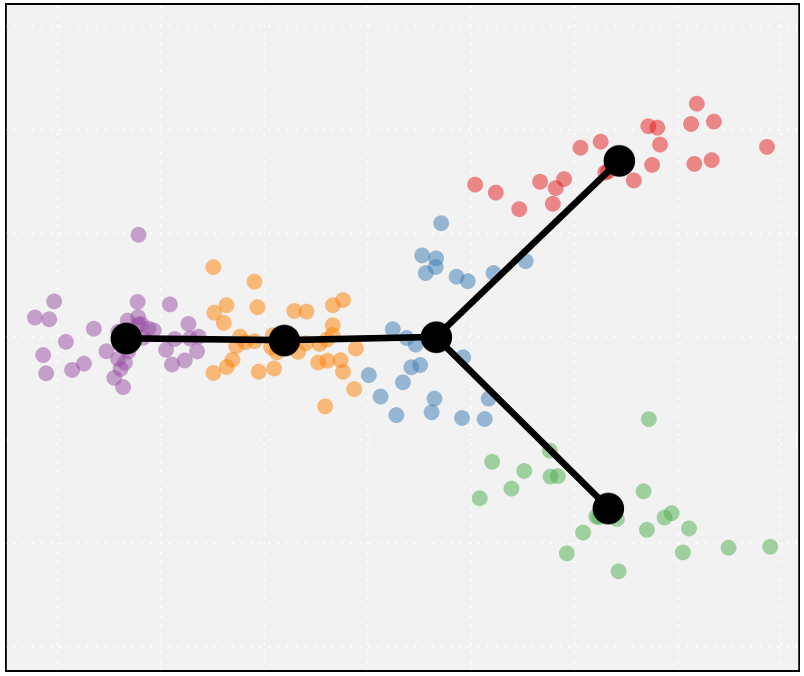
Step 0: Input



Clustered data in low-dimensional space.

We consider dimensionality reduction and clustering to be separate problems, but generally prefer PCA and RSEC.

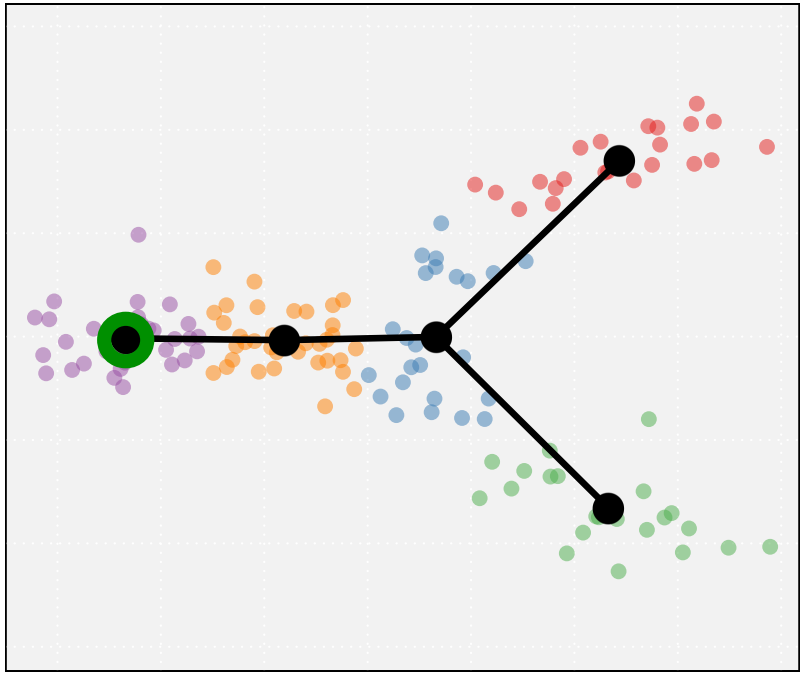
Step 1: Cluster MST



Construct minimum spanning tree (MST) on clusters.

The nodes are clusters, not cluster centers. Requires a distance metric.

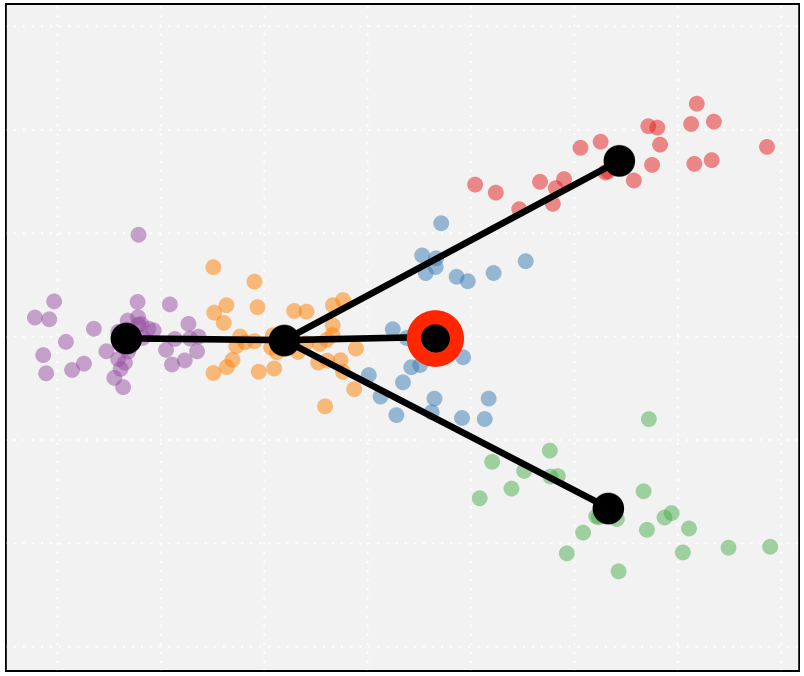
Step 1: Cluster MST



Construct minimum spanning tree (MST) on clusters.

Select starting cluster based on marker genes or parsimony.

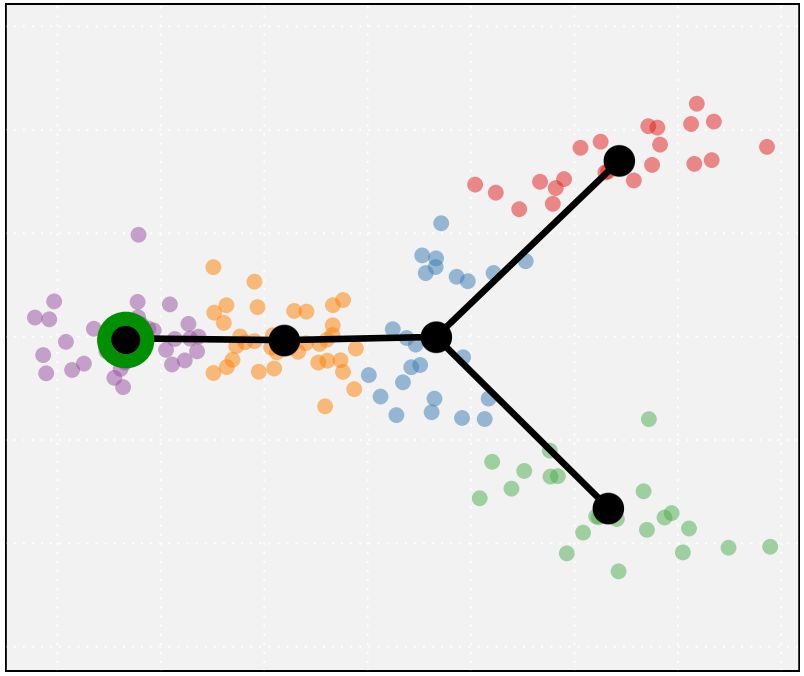
Step 1: Cluster MST



Specify known terminal clusters for additional supervision.

This results in the construction of a constrained MST.

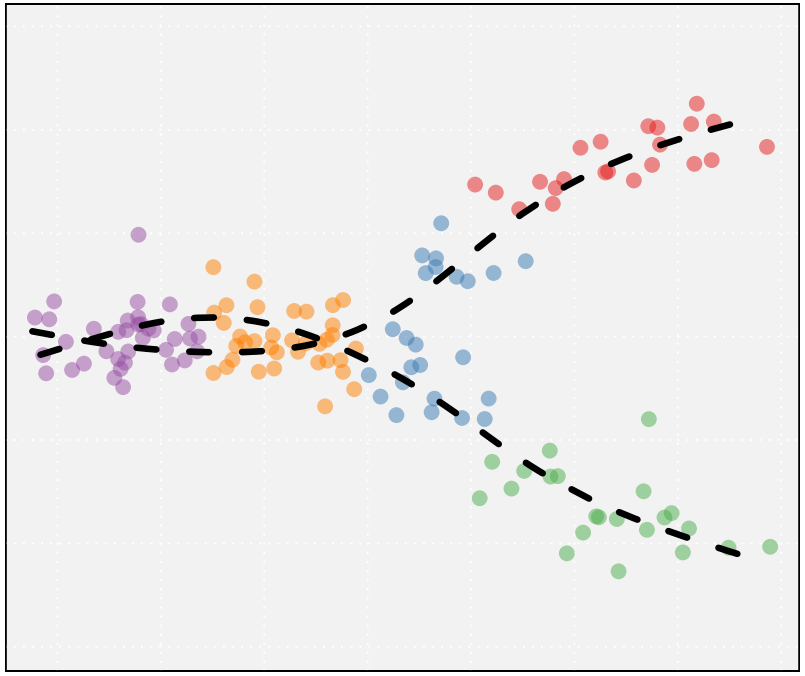
Step 1: Cluster MST



Construct minimum spanning tree (MST) on clusters.

Select starting cluster based on marker genes or parsimony.

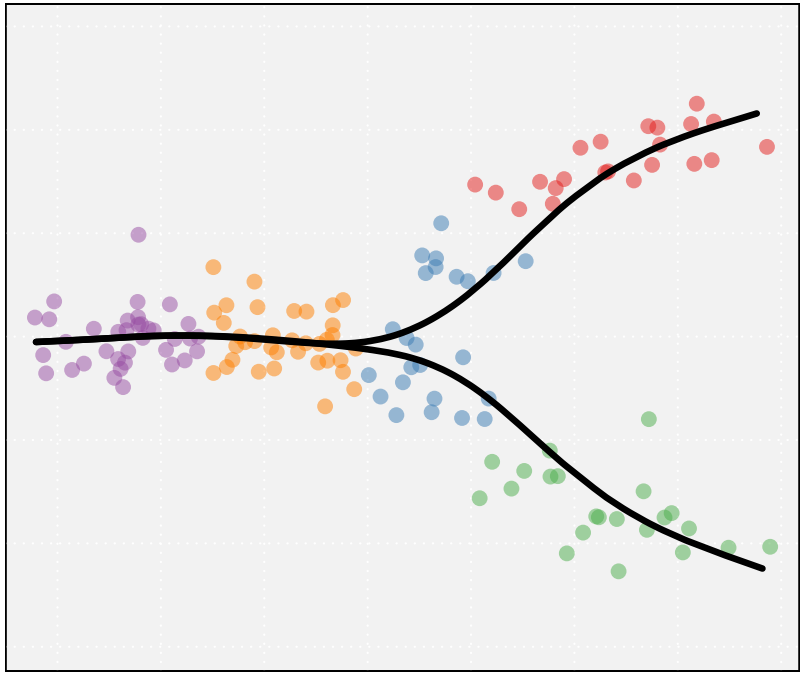
Step 1.5: Principal Curves



Highly stable, nonlinear generalization of principal components. Fits a curve to the “middle” of the data.

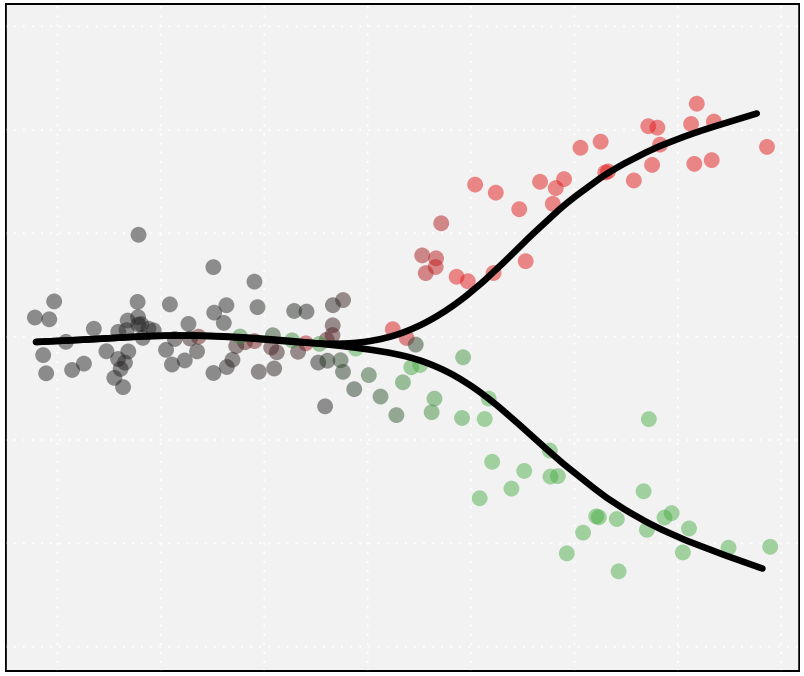
Inconsistent, fails to reflect underlying biology (ie. branching).

Step 2: Simultaneous Principal Curves



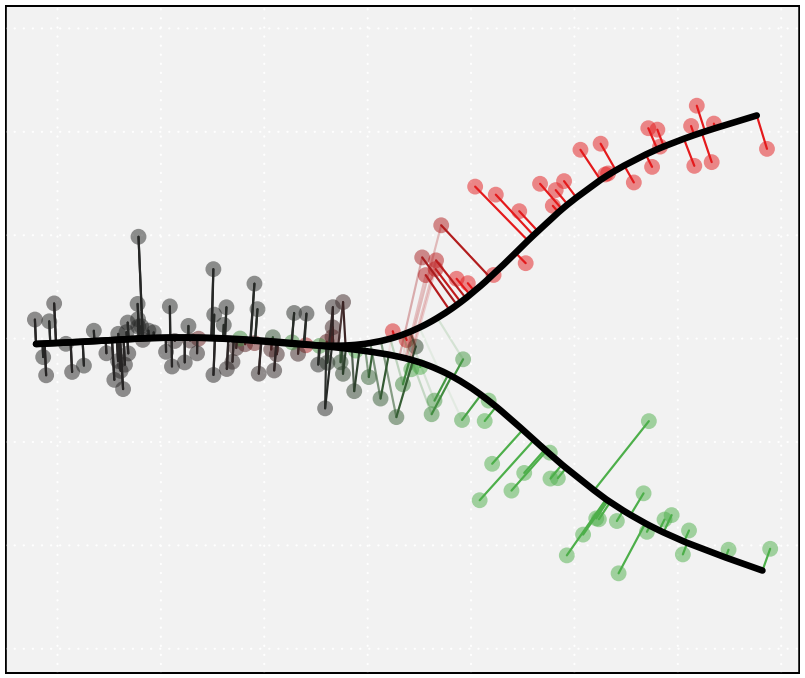
Still highly stable and nonlinear. Extends the concept of principal curves to multiple, branching curves with common origin (smooth tree structures).

Step 2: Simultaneous Principal Curves



Still highly stable and nonlinear. Extends the concept of principal curves to multiple, branching curves with common origin (smooth tree structures).

Step 2: Simultaneous Principal Curves



Project cells onto curves to obtain pseudotime values.

Like linear principal components, the curves seek to minimize squared projection distance (subject to some constraints).