scRNAseq clustering tools

Åsa Björklund

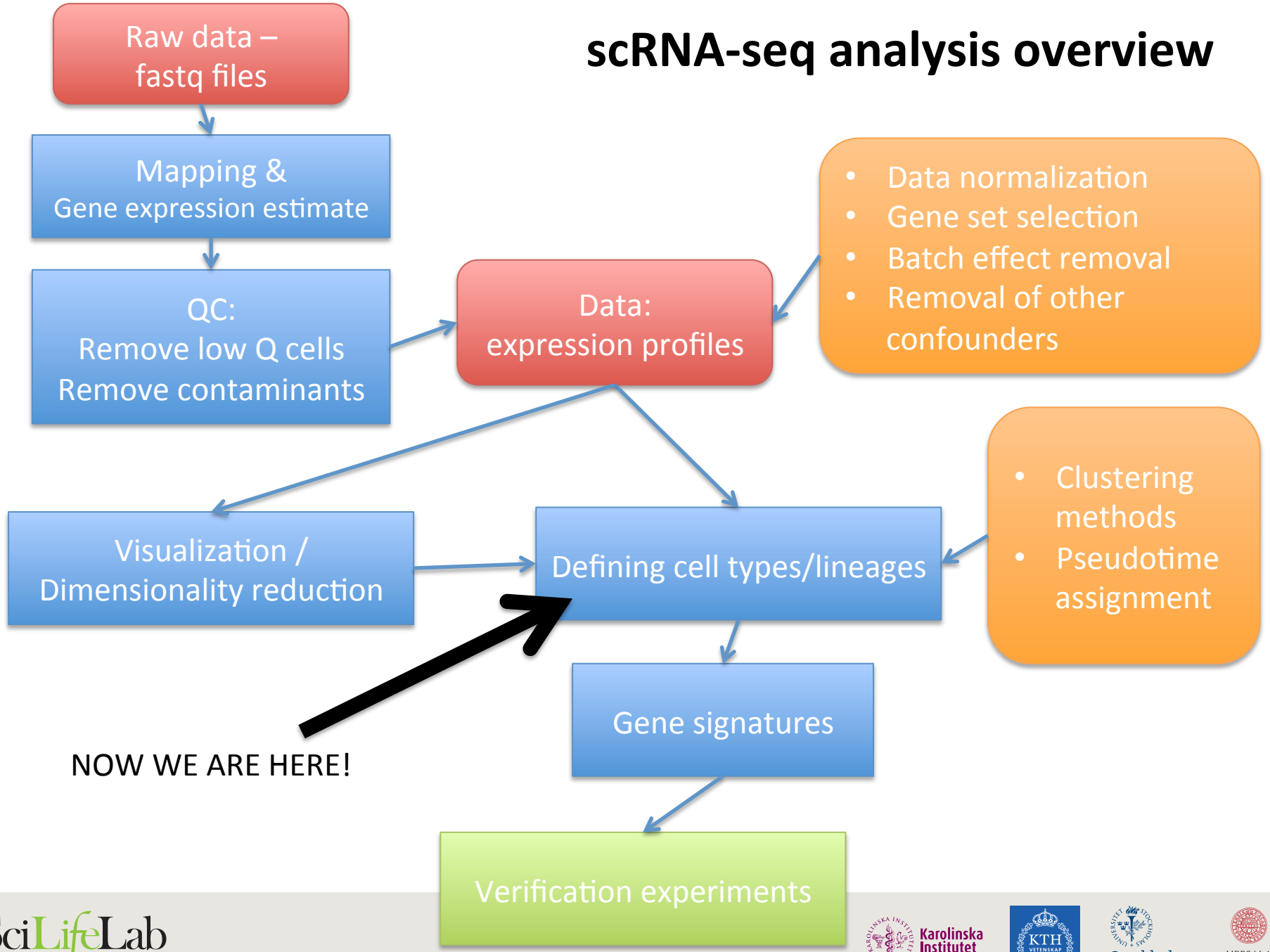
asa.bjorklund@scilifelab.se

What is a celltype?

What is a cell type?

- A cell that performs a specific function?
- A cell that performs a specific function at a specific location/tissue?
- Not clear where to draw the line between cell types and **subpopulations** within a cell type.
- Also important to distinguish between **cell type** and **cell state**.
  - A cell state may be infected/non infected
  - Metabolically active/inactive
  - Cell cycle stages
  - Apoptotic

scRNA-seq analysis overview



Outline

- Basic clustering theory
- Examples of different toolkits for clustering
- Pseudotime analysis

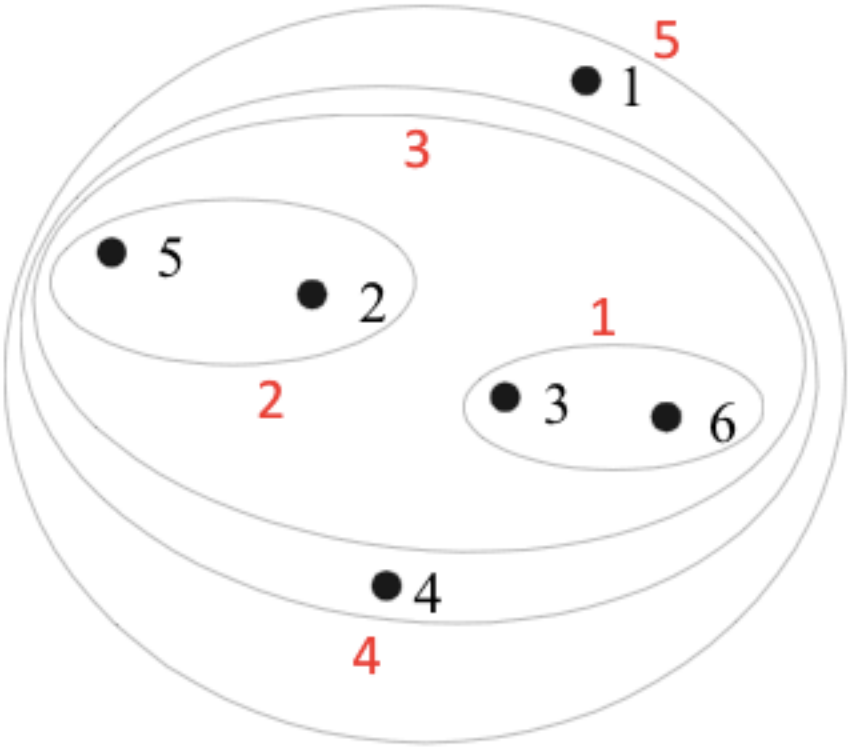
What is clustering?

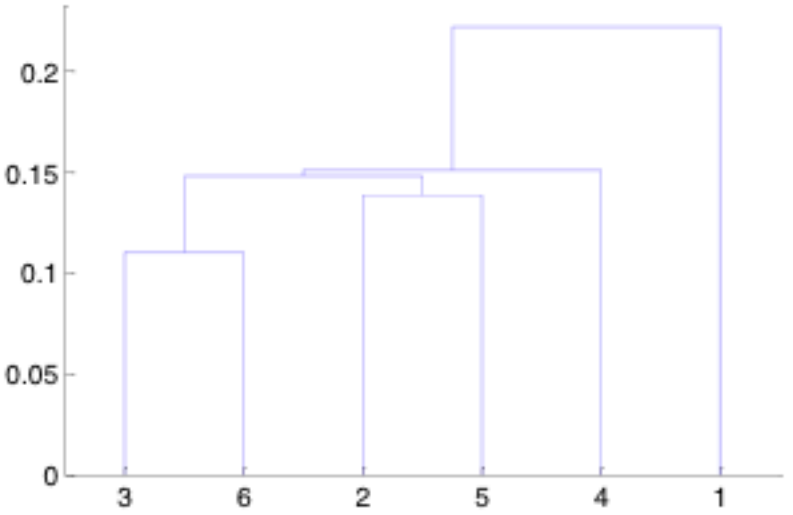
- “The process of organizing objects into groups whose members are similar in some way”
- Typical methods are:
  - Hierarchical clustering
  - K-means clustering
  - Graph based clustering

Hierarchical clustering

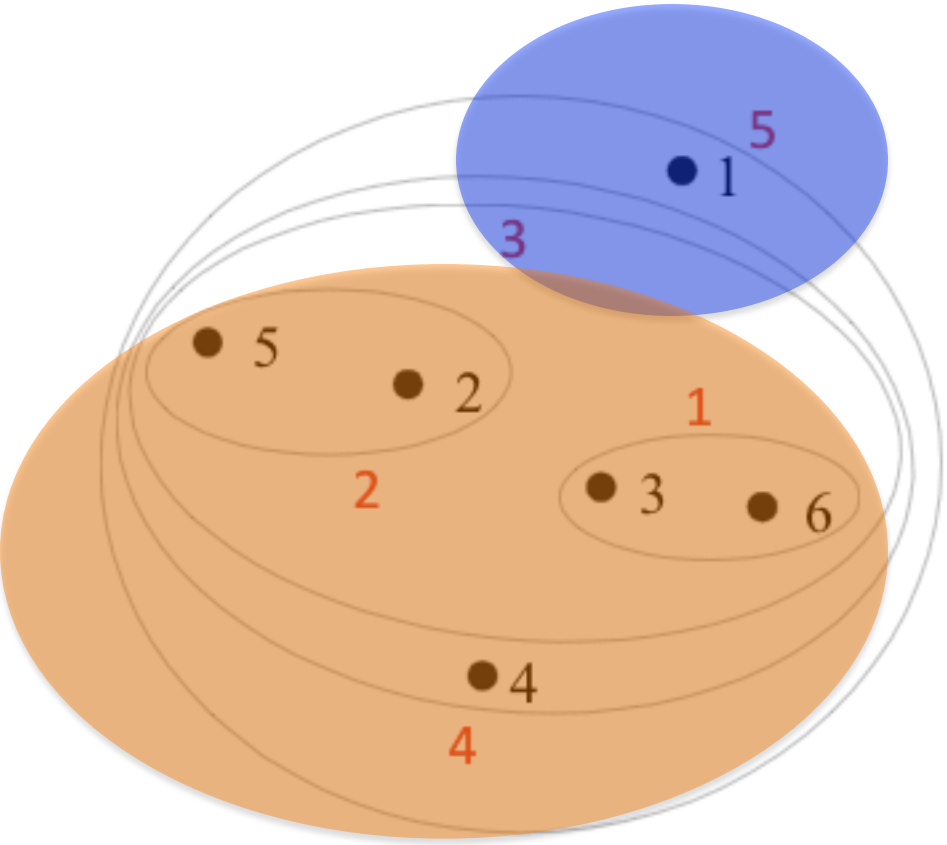
- Builds on **distances** between data points
- **Agglomerative** – starts with all data points as individual clusters and joins the most similar ones in a bottom-up approach
- **Divisive** – starts with all data points in one large cluster and splits it into 2 at each step. A top-down approach
- Final product is a **dendrogram** representing the decisions at each merge/division of clusters

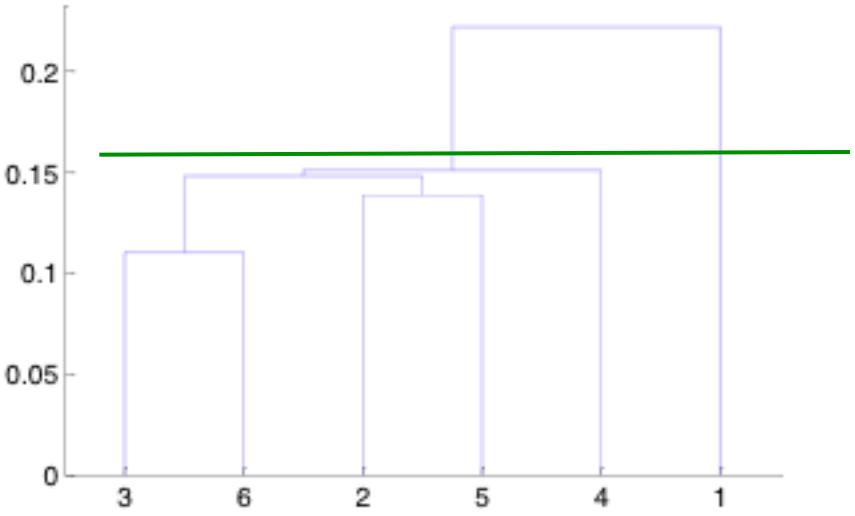
Hierarchical clustering





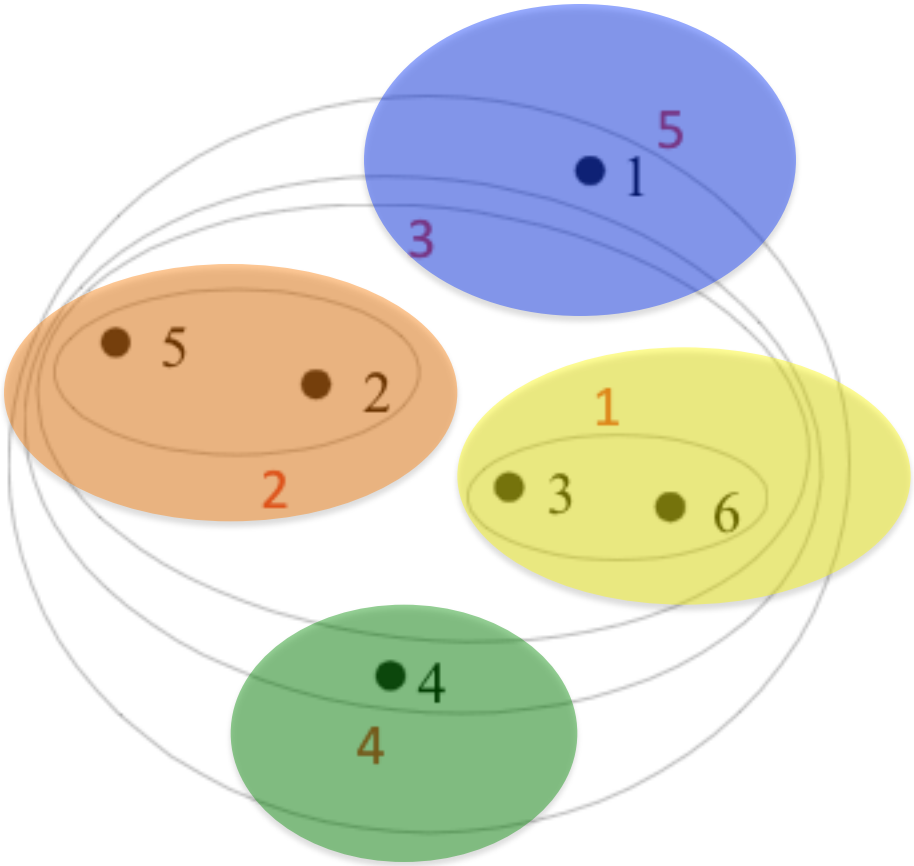
Hierarchical clustering

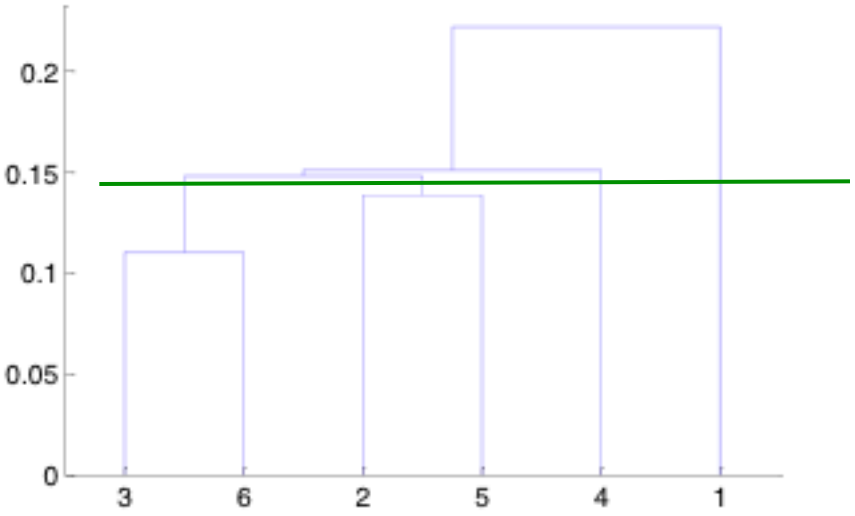




Clusters are obtained by cutting the tree at a desired level

Hierarchical clustering





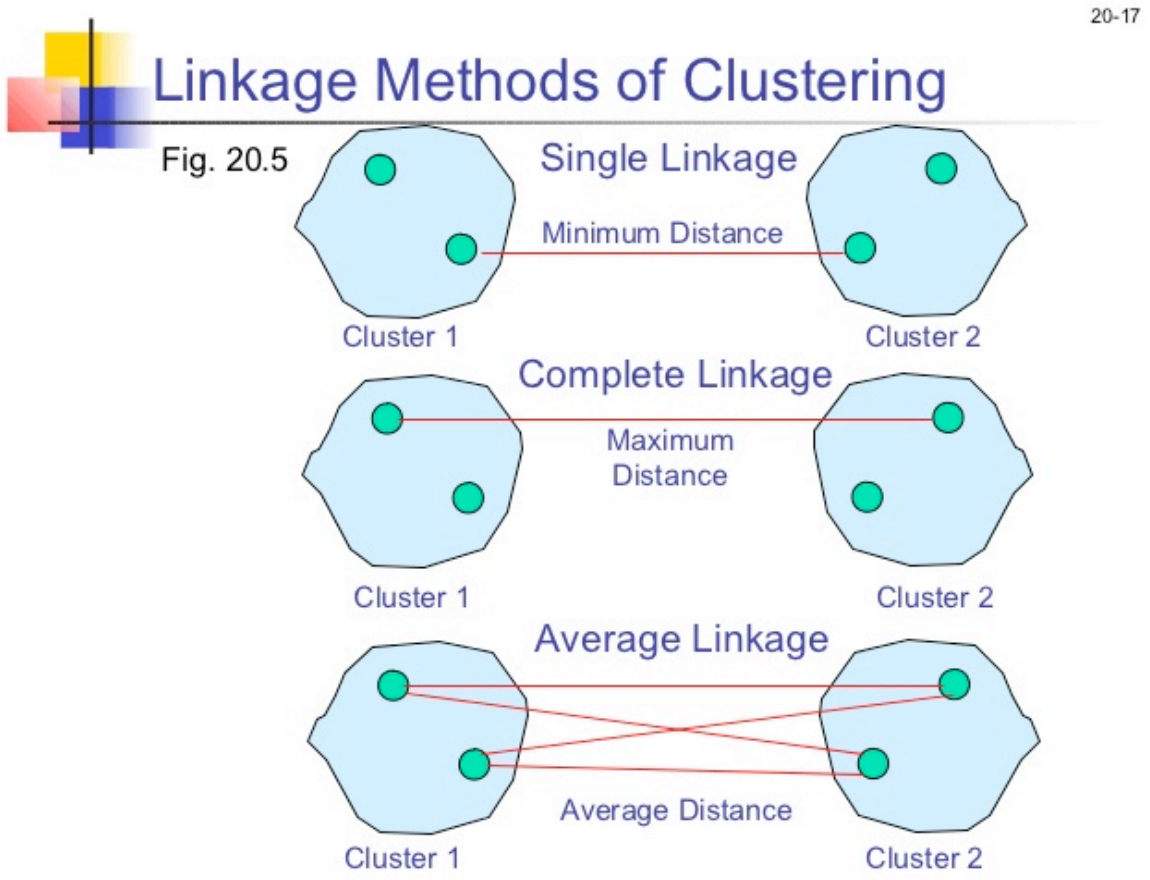
Clusters are obtained by cutting the tree at a desired level

Different distance measures

- Most commonly used in scRNA-seq:
  - Euclidean distance
    - In multidimensional space
    - In PCA/tSNE or other reduced space
  - Inverted pairwise correlations (1-correlation)
- Others include:
  - Manhattan distance
  - Mahalanobis distance
  - Maximum distance

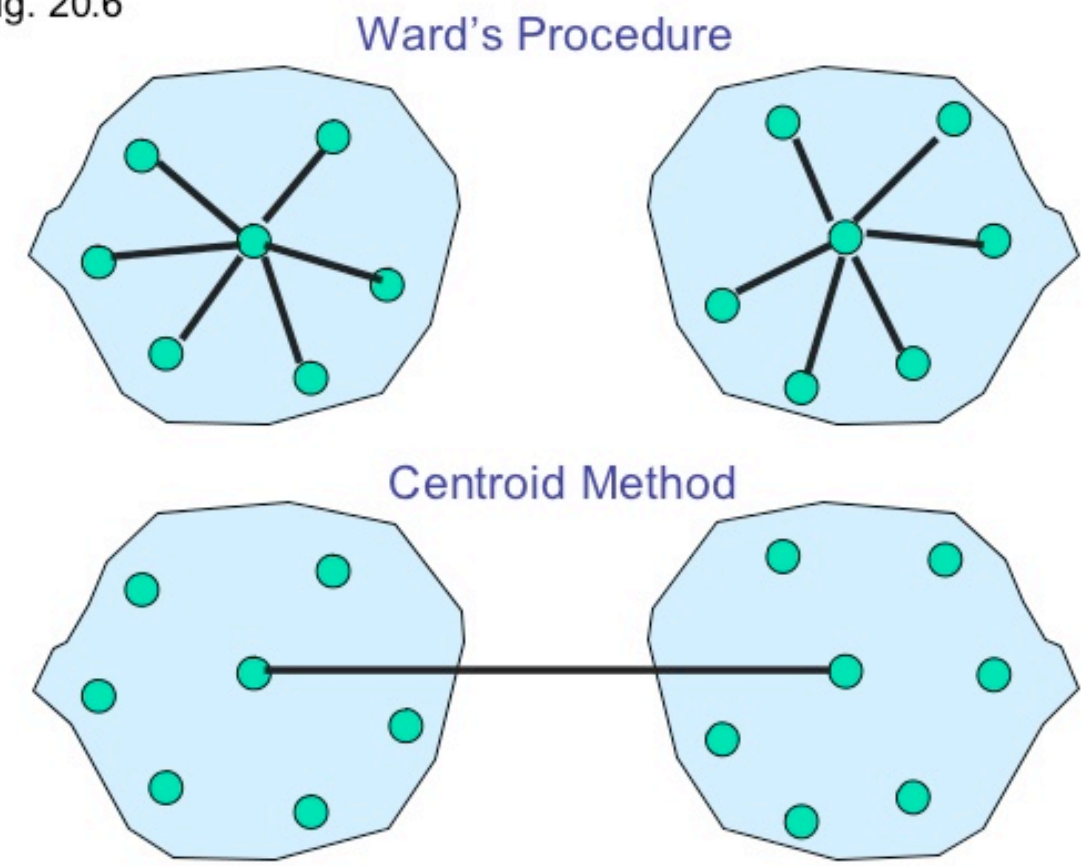
Linkage criteria

- Calculation of similarities between 2 clusters (or a cluster and a data point)



Other Agglomerative Clustering Methods

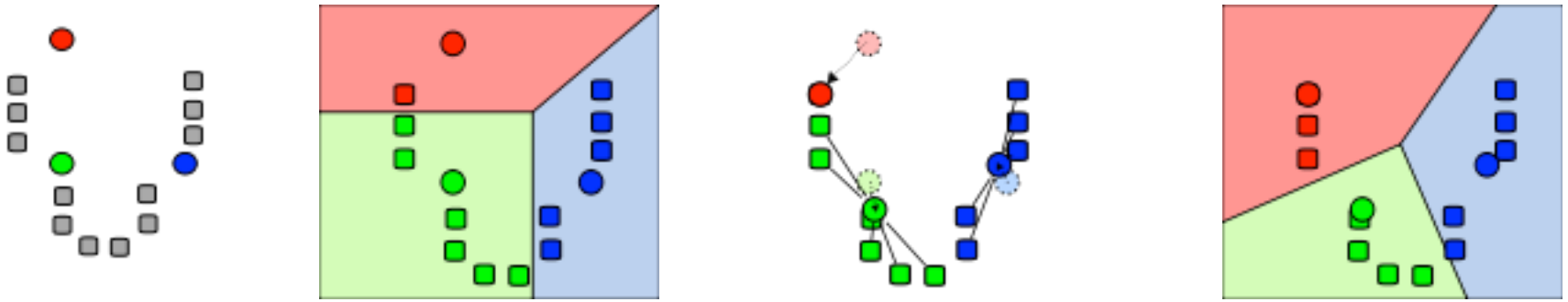
Fig. 20.6



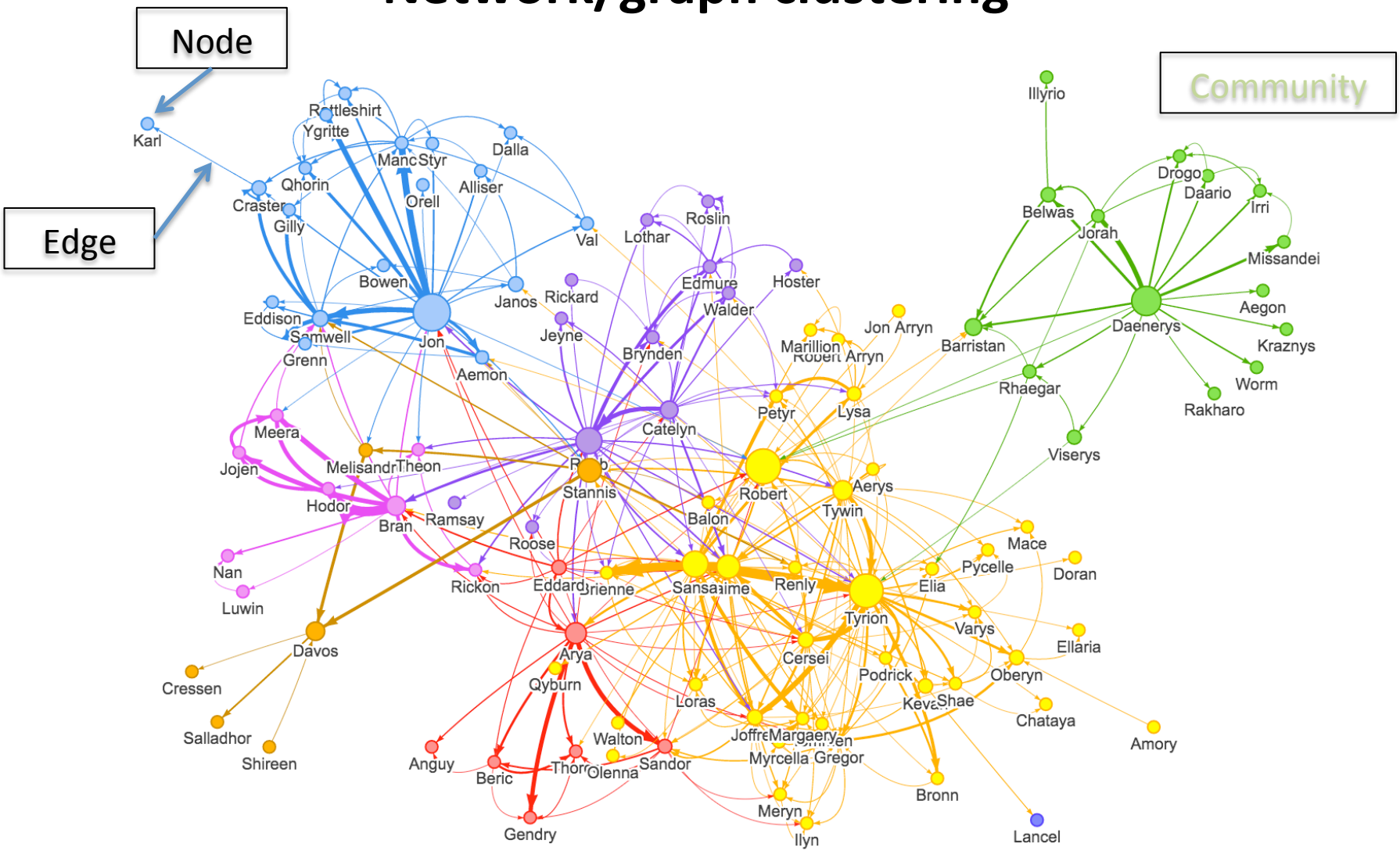
- Ward (minimum variance method). Similarity of two clusters is based on the increase in squared error when two clusters are merged.

K-means clustering

1. Starts with random selection of cluster centers (centroids)
2. Then assigns each data points to the nearest cluster
3. Recalculates the centroids for the new cluster definitions
4. Repeats steps 2-3 until no more changes occur.



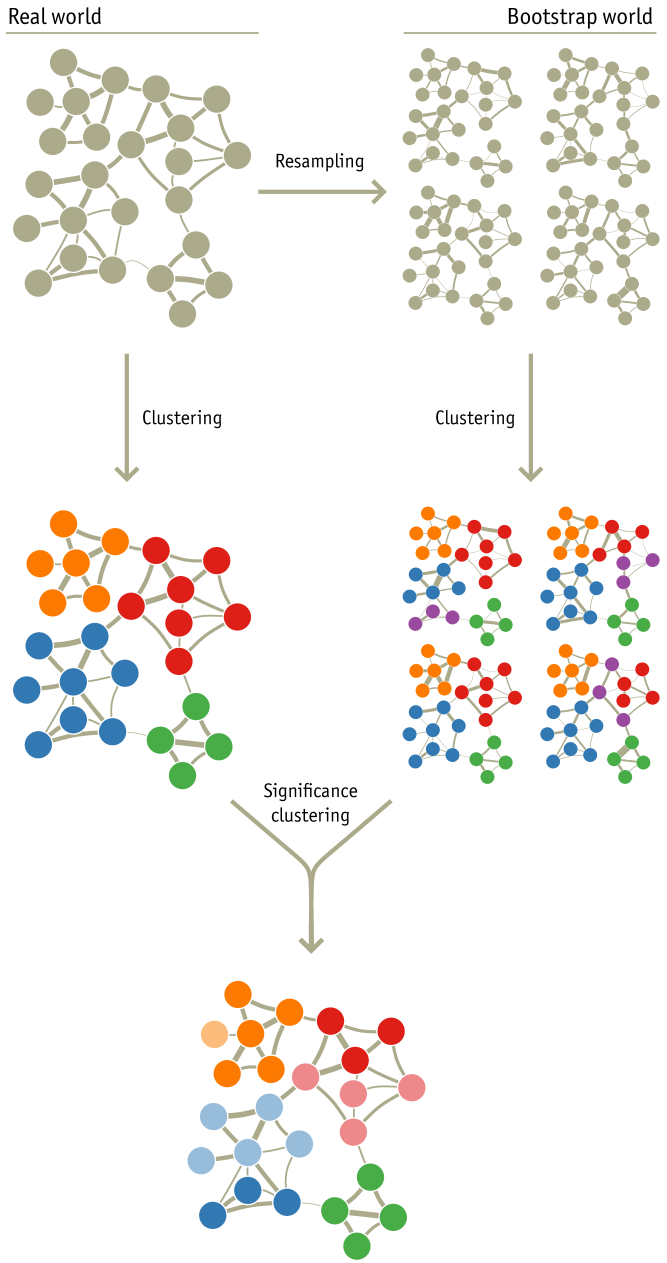
Network/graph clustering



(http://www.lyonwj.com/2016/06/26/graph-of-thrones-neo4j-social-network-analysis/)

Bootstrapping

- How confident can you be that the clusters you see are real?
- You can always take a random set of cells from the same cell type and manage to split them into clusters.
- Most scRNAseq packages do not include any bootstrapping

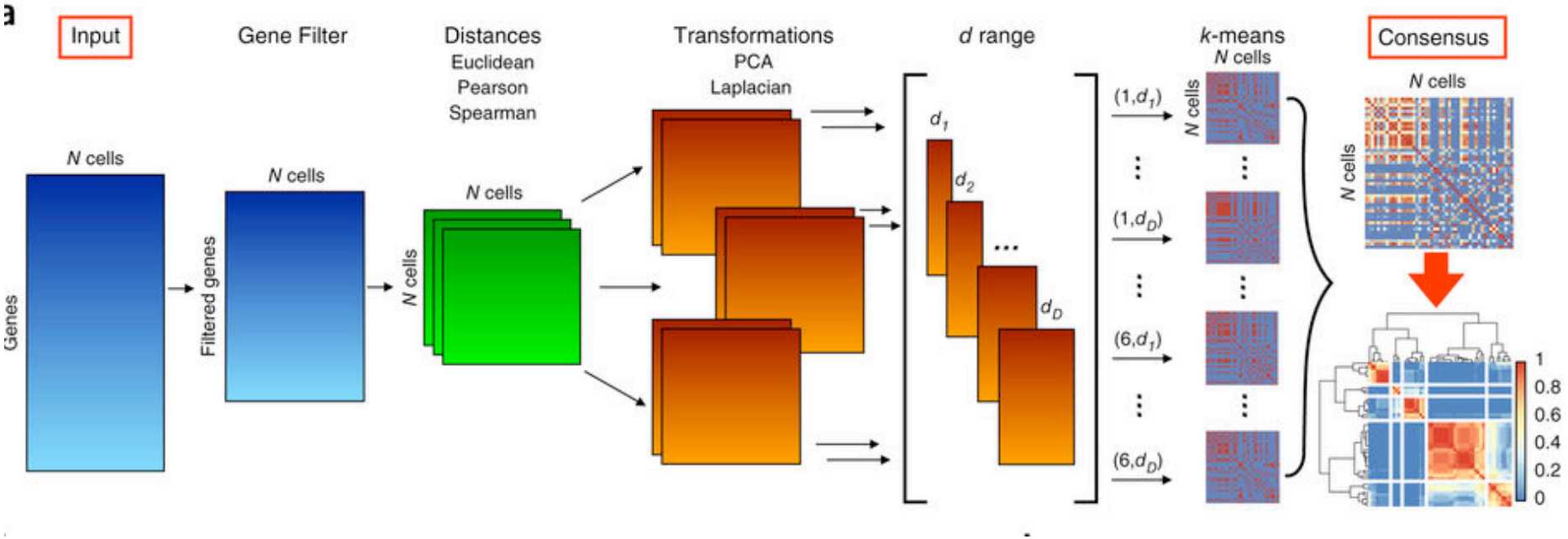


(Rosvall et al. *Plos One* 2010)

scRNAseq clustering

- Easy case with distinct celltypes:
  - rpkms/counts – Euklidean or correlation distances
  - PCA, tSNE or other dimensionality reduction method
- Examples of programs for clustering (many more out there):
  - WGCNA
  - BackSPIN
  - Pagoda
  - SC3
  - pcaReduce
  - SNNcliq
  - Seurat

Single Cell Consensus Clustering – SC3



(Kiselev et al *Nat. Methods* 2017)

Single Cell Consensus Clustering – SC3

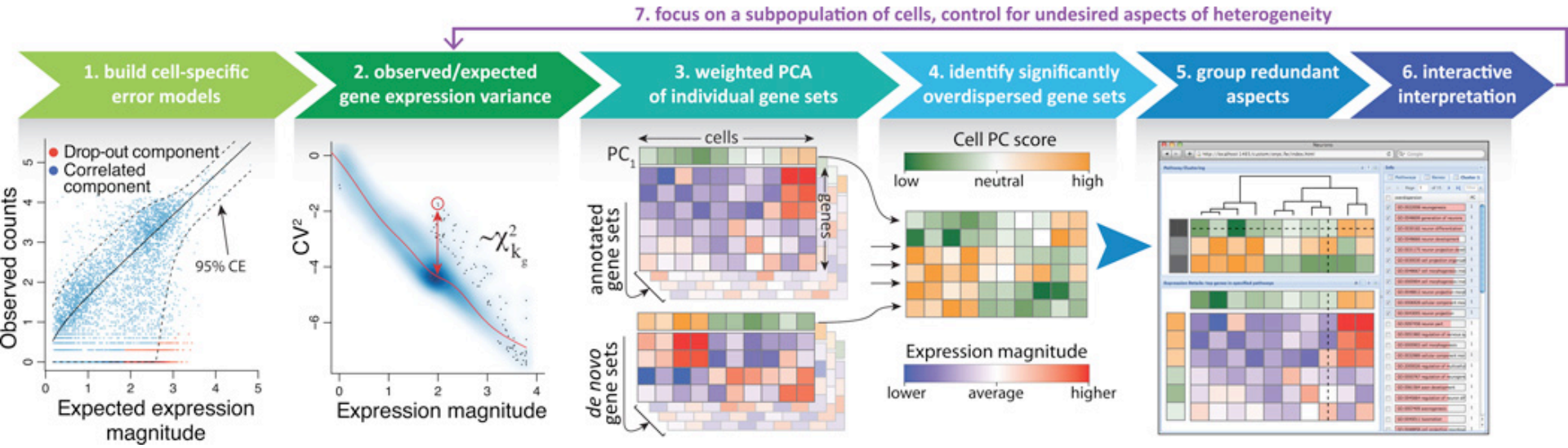
1. Gene filtering – rare and ubiquitous genes
2. Distance matrices (DM) – Euklidean, Spearman, Pearson
3. Transformation of DM with PCA or Laplacian
4. K-means clustering with first *d* eigenvectors
5. Consensus clustering – distance 1/0 for cells in same/ different clusters -> hierarchical clustering on average distances.

Differential expression with nonparametric Kruskal–Wallis test.

Marker genes with areas under the ROC curve (AUROC) from 100 permutations of cell cluster labels and P-values from Wilcoxon signed-rank test.

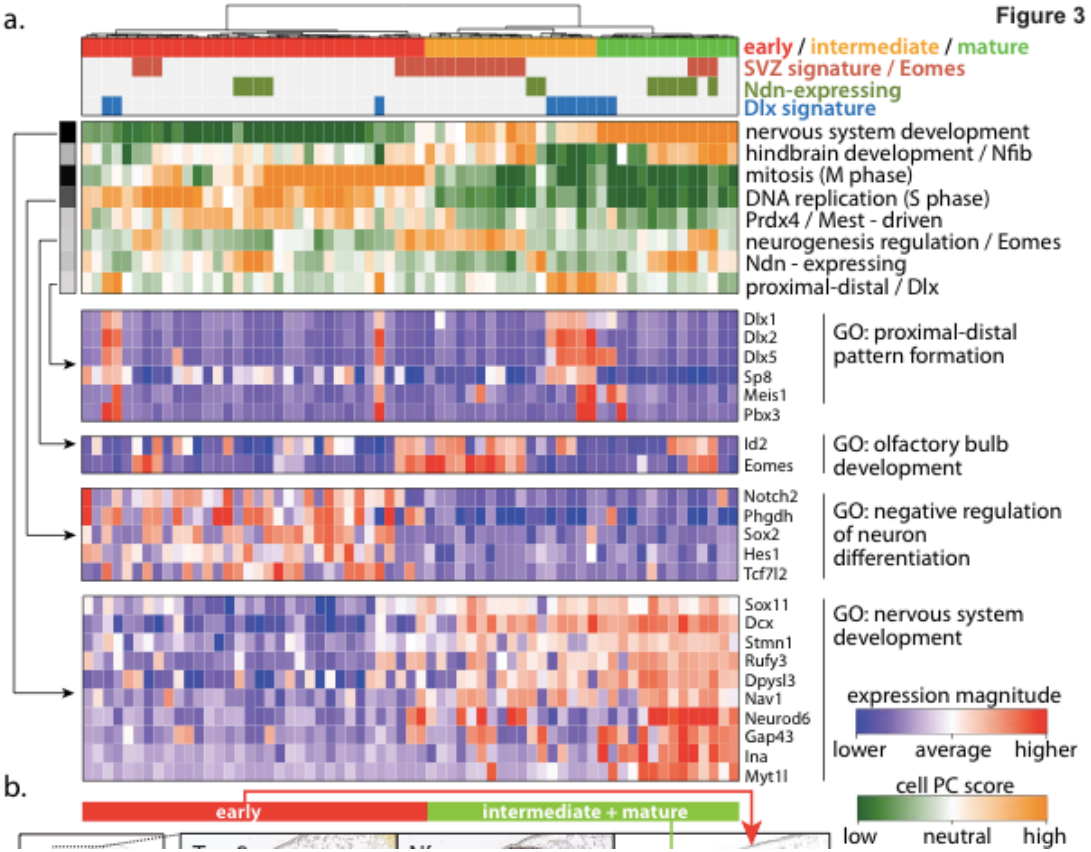
Pagoda – Pathway And Geneset OverDispersion Analysis

Implemented in the SCDE package



(Fan et al. *Nature Methods* 2016)

Pagoda – Pathway And Geneset OverDispersion Analysis



- Helps with biological interpretation of data
- Important to have good and relevant gene sets
- High memory consumption when running Pagoda
- Also has methods for removing batch effect, detected genes, cell cycle etc

(Fan et al. *Nature Methods* 2016)

BackSPIN - Biclustering

- Simultaneous clustering genes and cells.
- An iterative, biclustering method based on sorting points into neighborhoods (SPIN) to find shapes in a reduced space
  1. ordering of samples using genes as features,
  2. ordering of genes using samples as features and
  3. zooming in on subsets of the original expression matrix to order objects in a reduced subspace.
- Clusters both genes and cells to identify subpopulations as well as potential markers for each subpopulations.
- Implemented in Python.

Shared nearest neighbor (SNN)-Cliq

- Similarity matrix using Euclidean distance (can use other distances)
- List the *k*-nearest-neighbors (KNN)
- Edge between cells if at least one shared neighbor
- Weights based on ranking of the neighbors
- Graph partition by finding cliques
- Identify clusters in the SNN graph by iteratively combining significantly overlapping subgraphs
- Implemented in Matlab and Python

Seurat

- Developed for drop-seq analysis − compatible with 10X output files.
- First construct a KNN (k-nearest neighbor) graph based on the euclidean distance in PCA space.
- Refine the edge weights between any two cells based on the shared overlap in their local neighborhoods (Jaccard distance).
- To cluster the cells, modularity optimization techniques to iteratively group cells together.

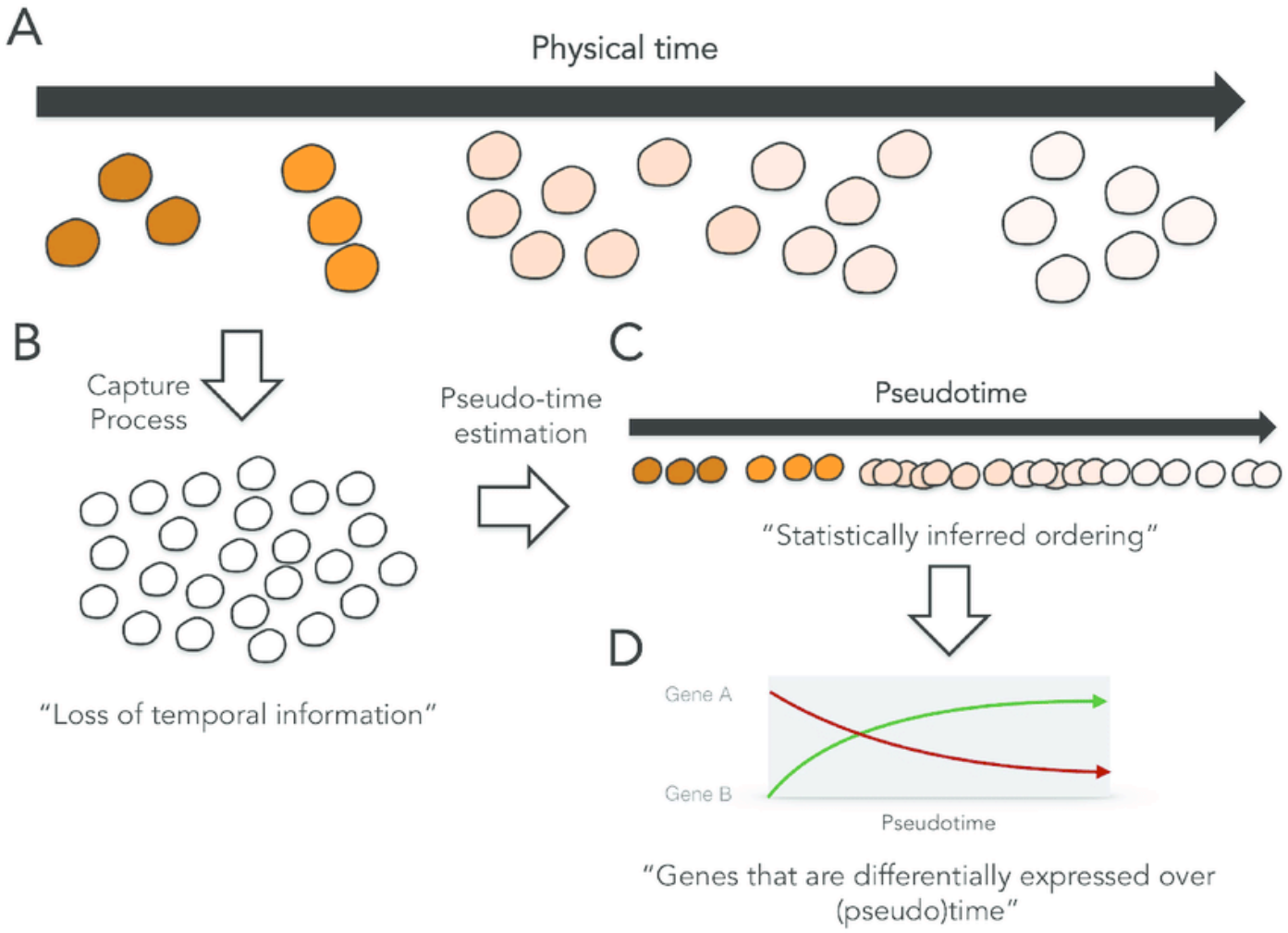
Seurat

- Also contains functions for:
  - Spatial reconstruction of single cell data using *in situ* references (Zebrafish embryos)
  - Integrated analysis across platforms
- Differential expression tests:
  - ROC test
  - t-test
  - Likelihood-ratio test (LRT) test based on zero-inflated data
  - LRT test based on tobit-censoring models
- OBS! Earlier versions of Seurat uses “spectral tSNE” and DBScan density clustering.

Which clustering method is best?

- Depends on the input data
- Consistency between several methods gives confidence that the clustering is robust
- The clustering method that is most consistent – best bootstrap values is not always best
- In a simple case where you have clearly distinct celltypes, simple hierarchical clustering based on euklidean or correlation distances will work fine.

Pseudotime/trajectory analysis



(Kieran et al. *Plos Comp Biol*. 2017)

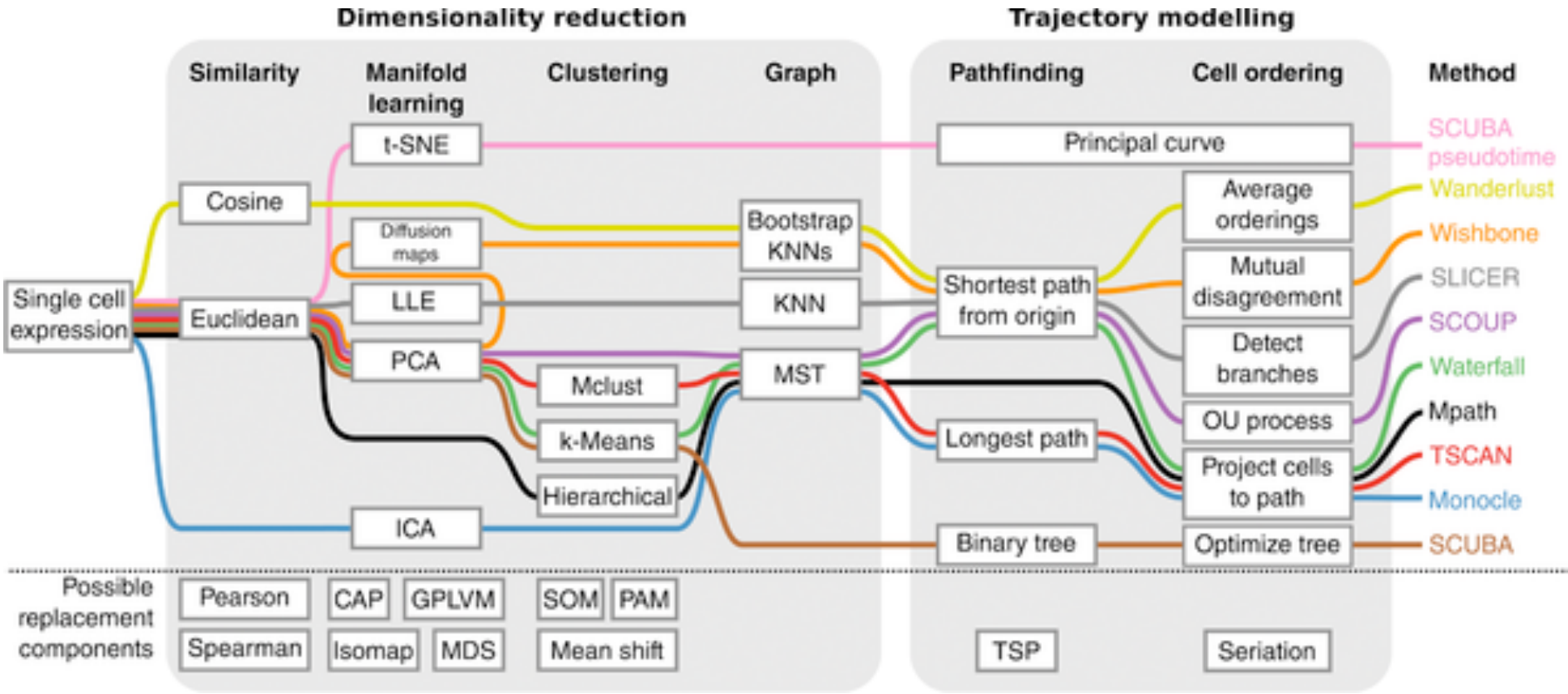
Should you run trajectory analysis?

- Are you sure that you have a developmental trajectory?
- Do you believe that you have branching in your trajectory?

- Be aware, any dataset can be forced into a trajectory without any biological meaning!
- First make sure that gene set and dimensionality reduction captures what you expect.

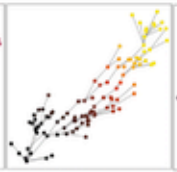
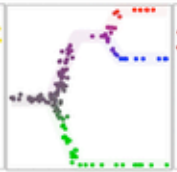
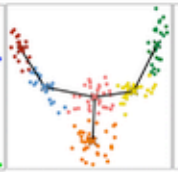
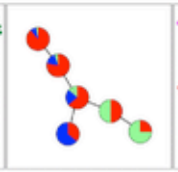
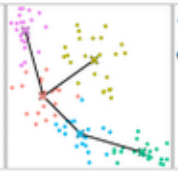
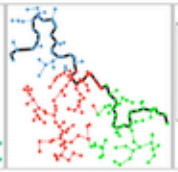
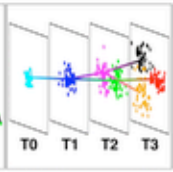
Trajectory analysis – main steps

1. Gene set selection
2. Dimensionality reduction
3. Infer trajectories (branched or straight)
4. Order cells
5. Discover interesting gene patterns



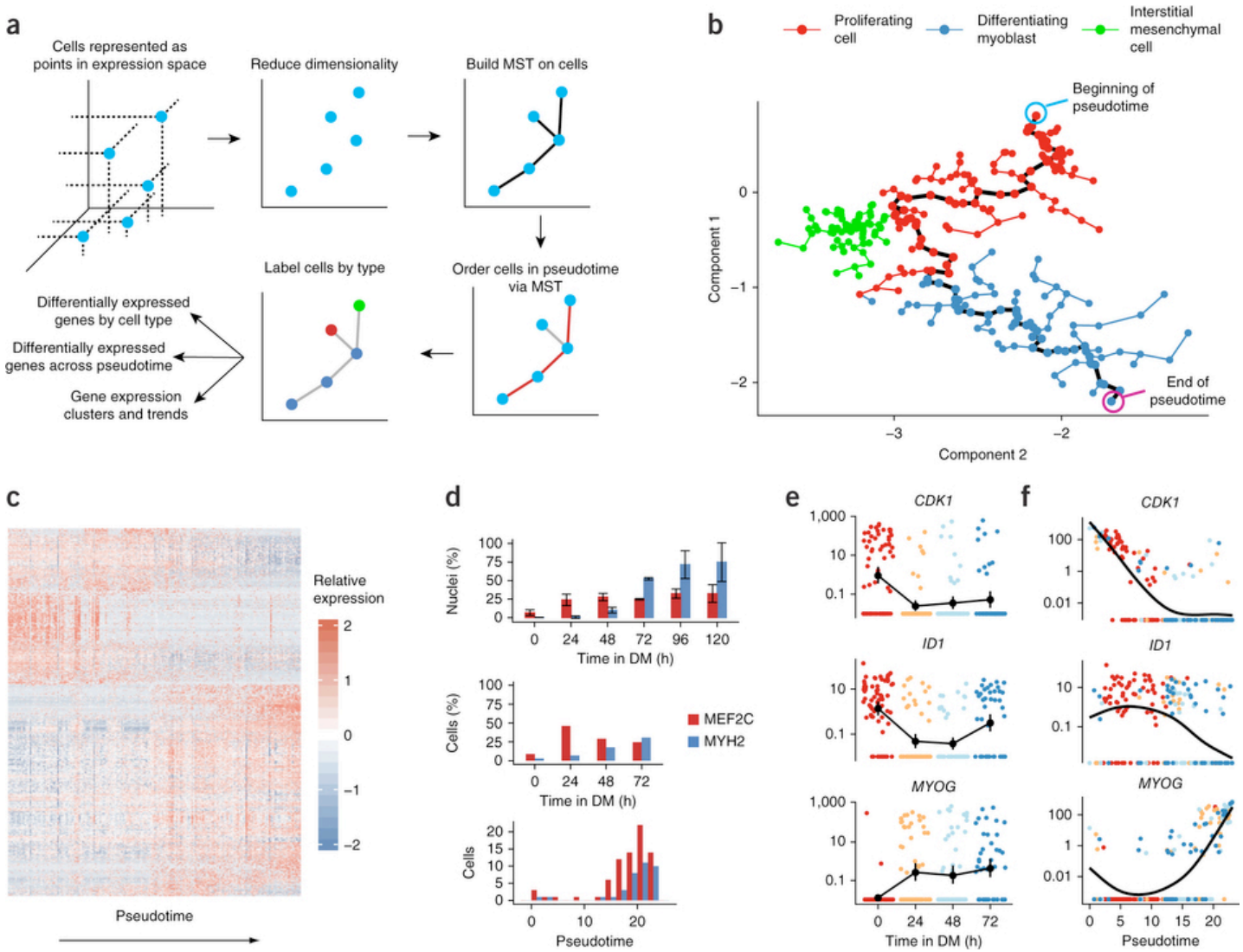
(Cannoodt et al. *EJI* 2016)

Summary of pseudotime tools

Method	SCUBA pseudotime	Wanderlust	Wishbone	SLICER	SCOUP	Waterfall	Mpath	TSCAN	Monocle	SCUBA
Visual abstract										
Structure	Linear	Linear	Single bifurcation	Branching	Branching	Linear	Branching	Linear	Branching	Branching
Robustness strategy	Principal curves	Ensemble, starting cell	Ensemble, starting cell	Starting cell	Starting population	Clustering of cells	Clustering of cells using external labelling	Clustering of cells	Differential expression	Simple model
Extra input requirements	None	Starting cell	Starting cell	Starting cell	Starting population	None	Time points	None	Time points	Time points
Unbiased	+	±	±	±	±	+	−	+	−	−
Scalability w.r.t. cells	−	−	±	±	−	±	+	+	−	±
Scalability w.r.t. genes	+	+	+	+	−	+	±	±	±	+
Code and documentation	−	±	+	±	+	±	+	+	+	±
Parameter ease-of-use	+	+	+	+	−	±	−	+	+	+

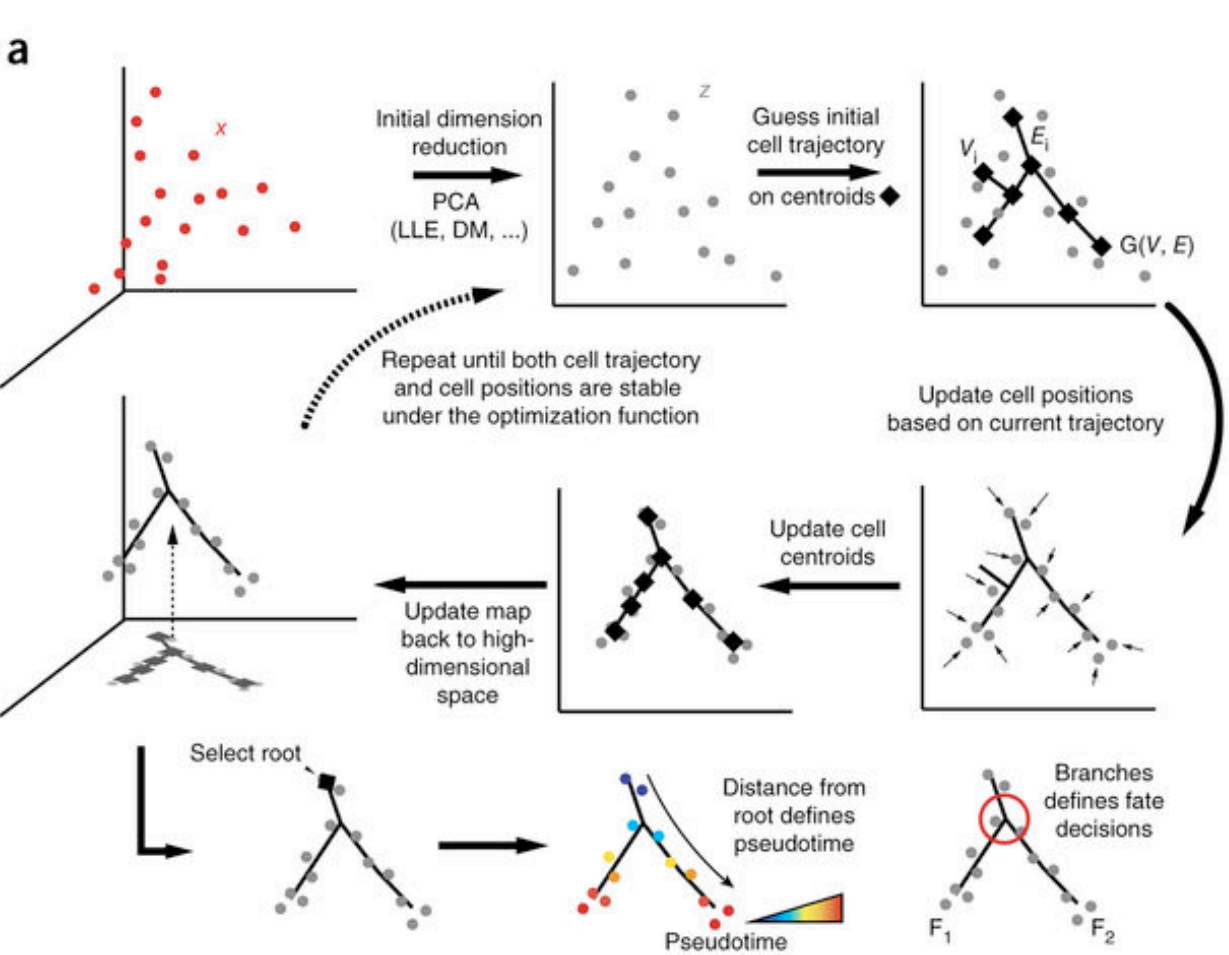
(Cannoodt et al. *EJI* 2016)

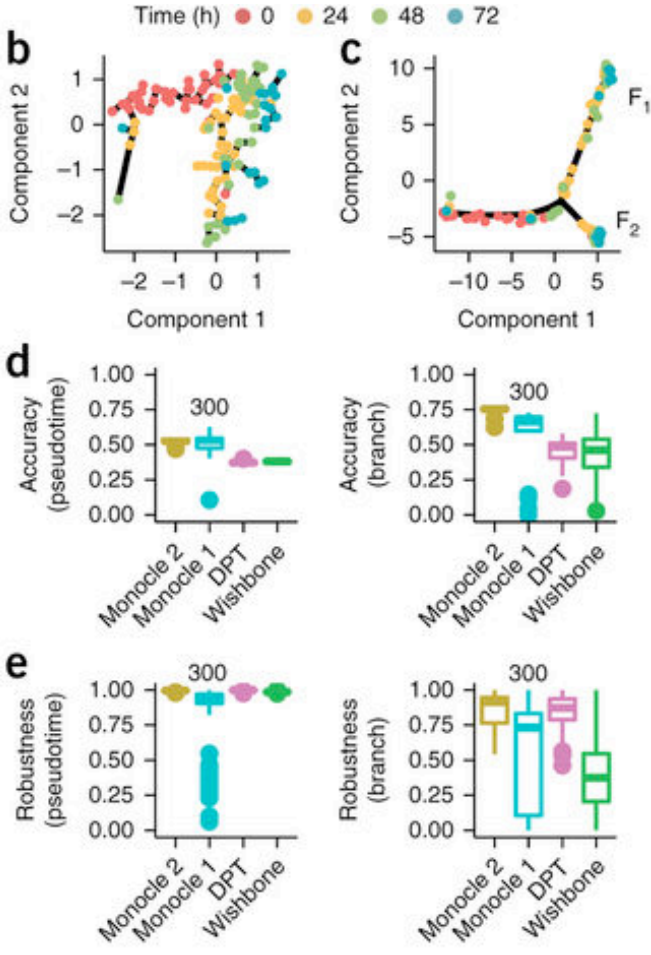
Pseudotime ordering – Monocle1



(Trapnell et al. *Nature Biotech* 2014)

Monocle2 − reversed graph enbedding

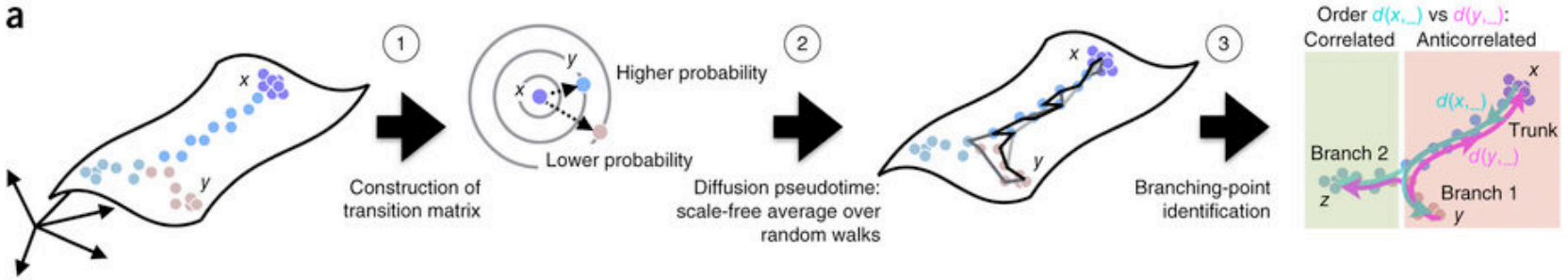


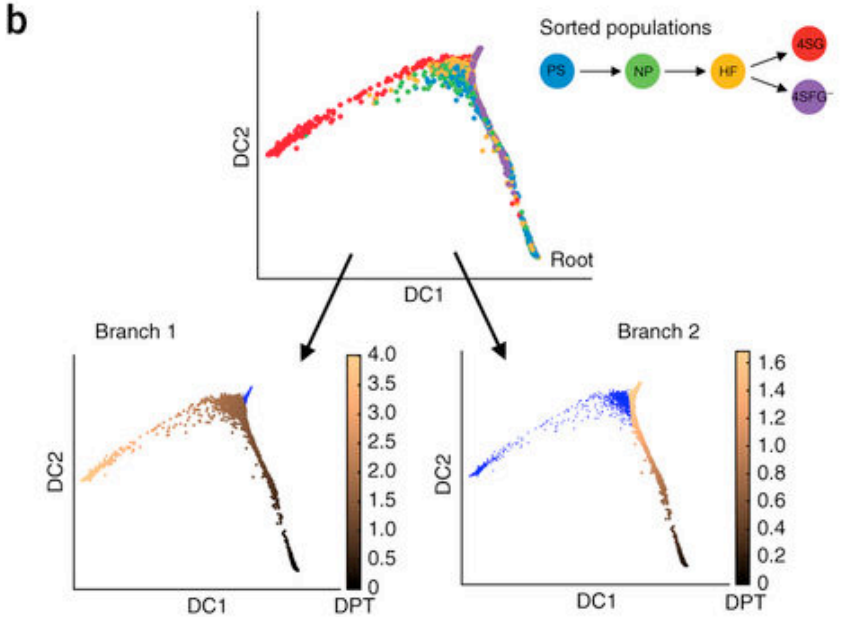


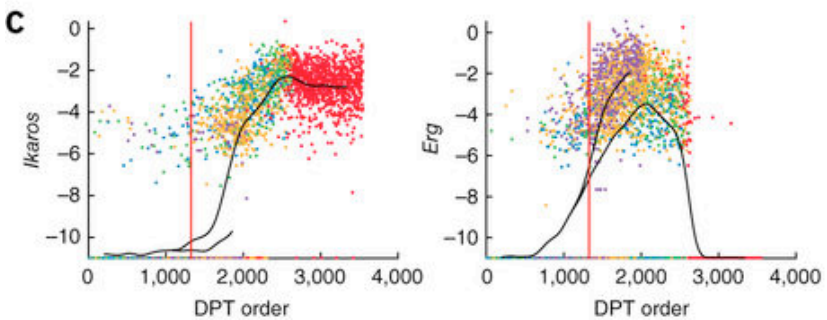
(Qiu et al. *Nat Methods* 2017)

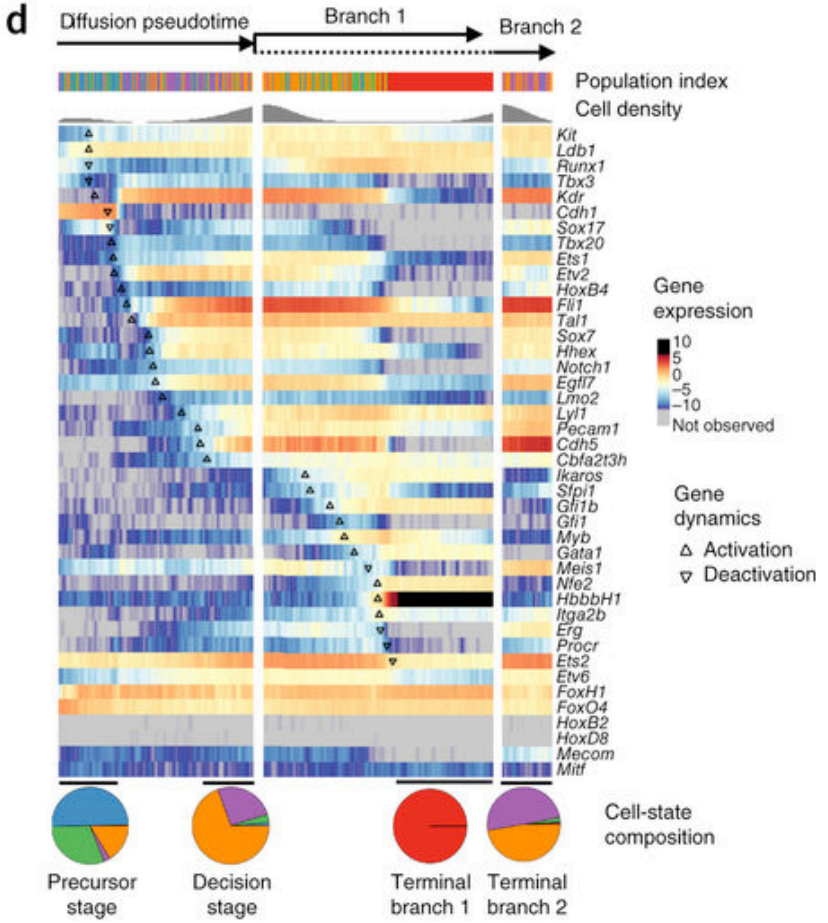
Diffusion pseudotime (DPT)

(Haghvedi et al *Nature Methods* 2016)









How many clusters do you really have?

- It is hard to know when to stop clustering − you can always split the cells more times.
- Can use:
  - Do you get any/many more significant DE genes from the next split?
  - Some tools have automated predictions for number of clusters − may not always be biologically relevant
- Always check back to QC-data − is what your splitting mainly related to batches, qc-measures (especially detected genes)

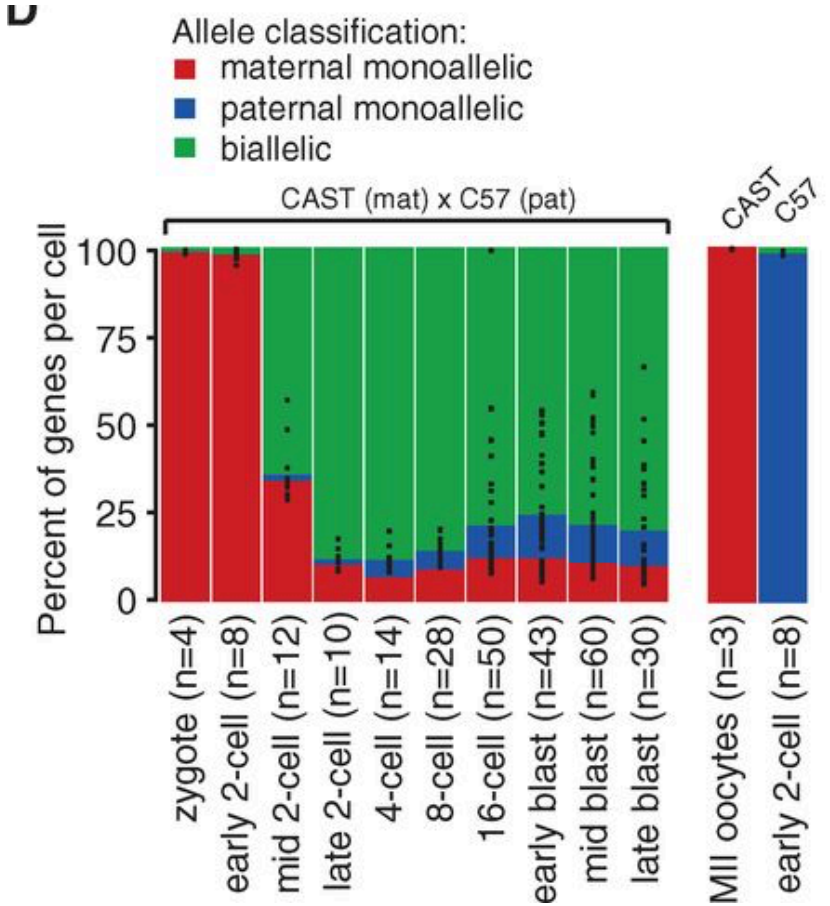
Additional analyses

- Copy-number variation
- Allelic expression
- Variant calling
- Alternative splicing

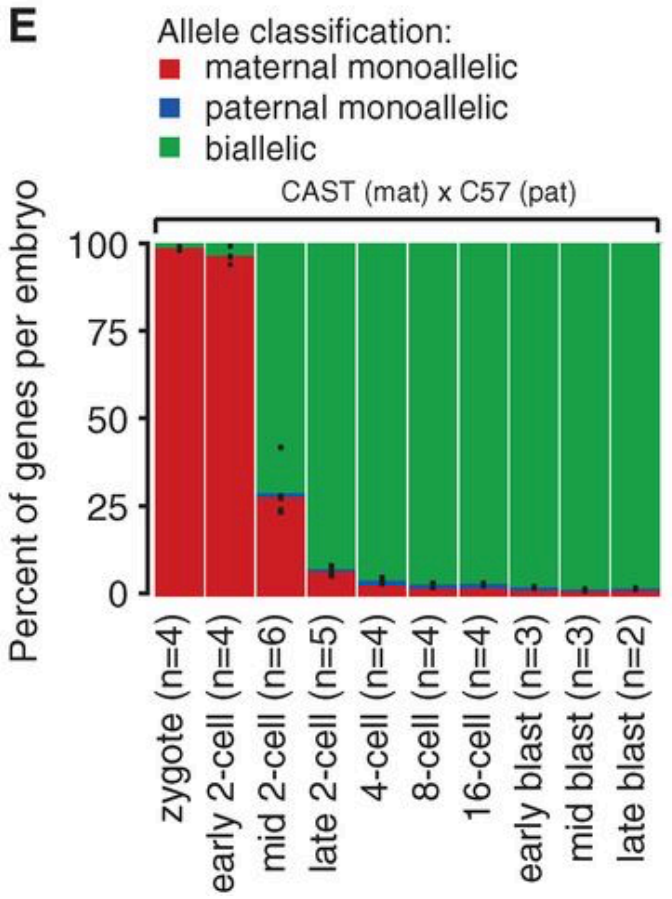
- The last 3 require full length methods
  - But only works for highly expressed genes with good read coverage
  - Must be careful to take into consideration the drop-out rate, a unique splice form/allele in a single cell may actually be a detection issue.

Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells

Single cells

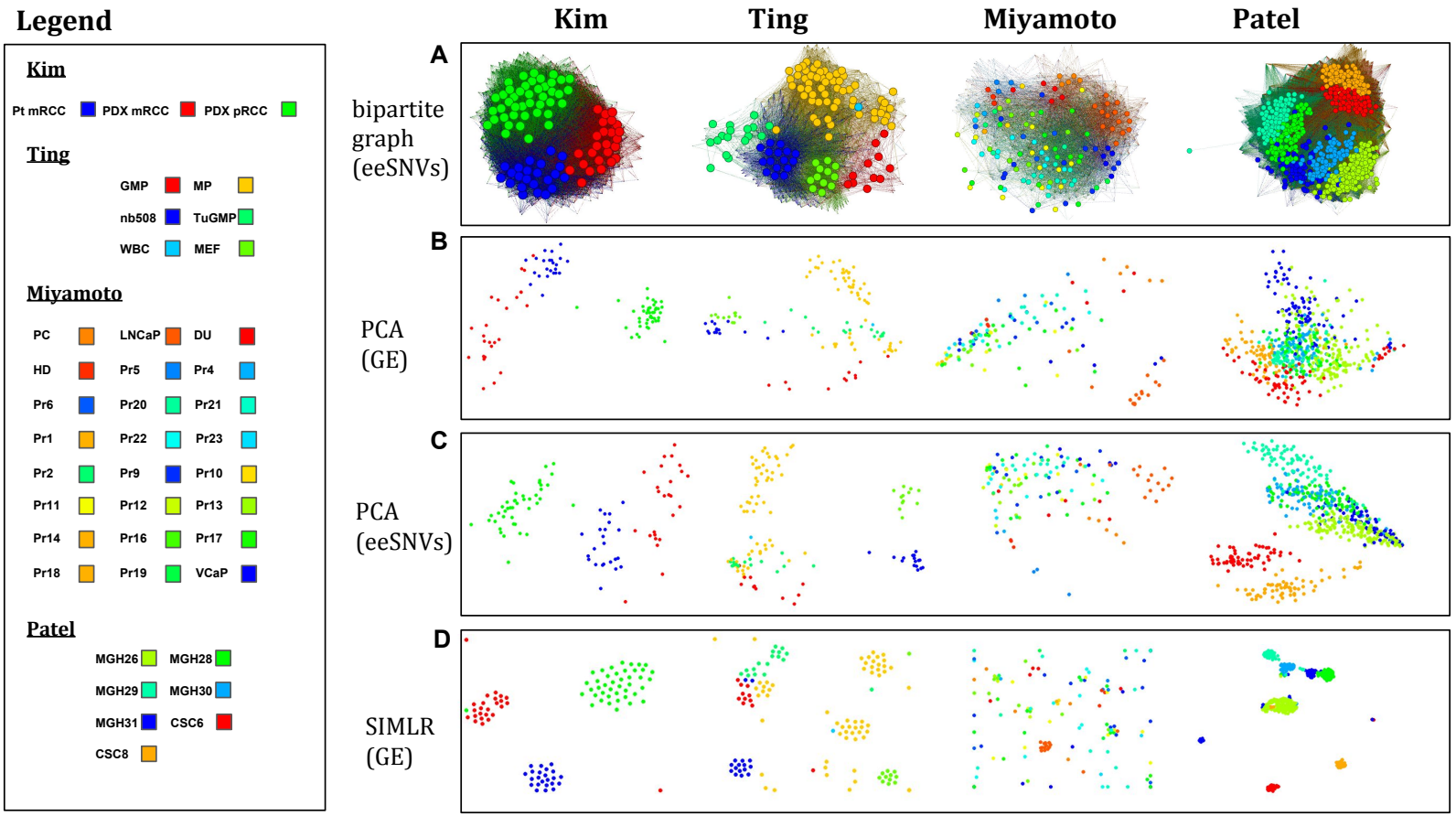


Pooled embryos



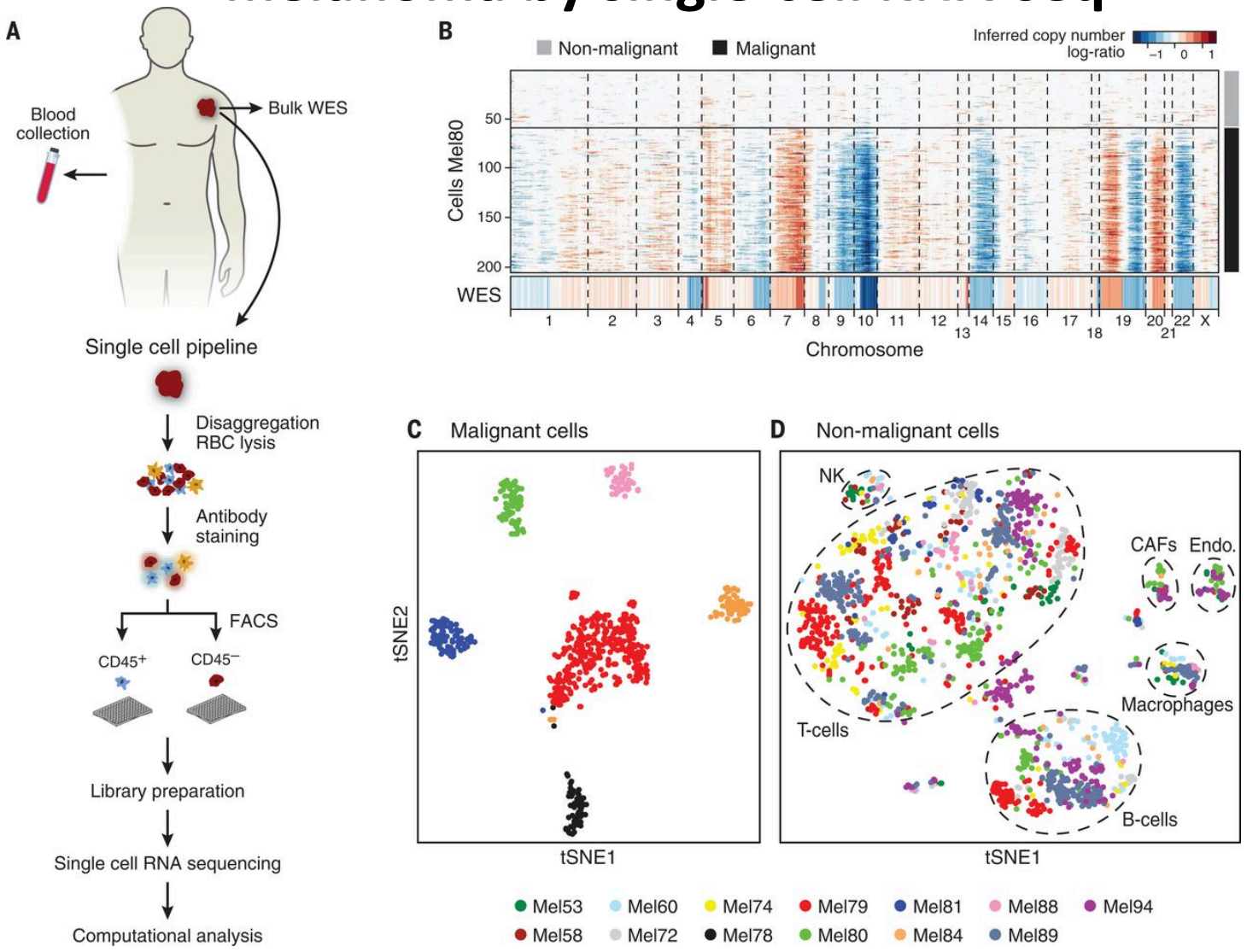
(Deng et al. *Science* 2014)

Using Single Nucleotide Variations in Cancer Single-Cell RNA-Seq Data for Subpopulation Identification and Genotype-phenotype Linkage Analysis



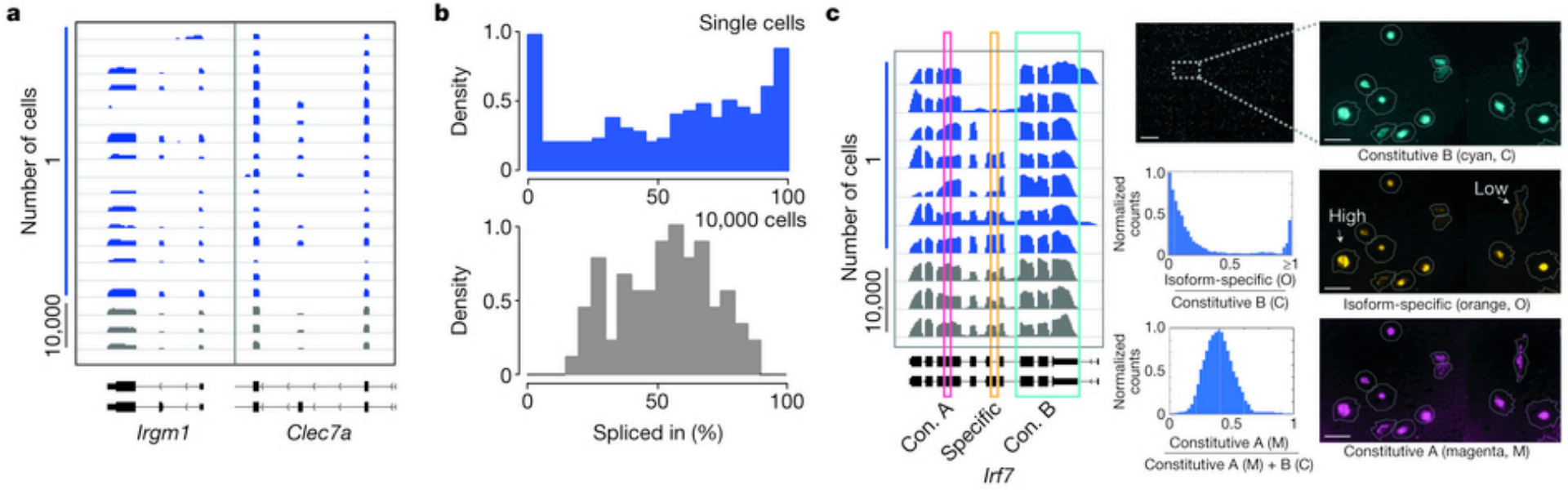
(Poiron et al. *BioRxiv* 2016)

Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq

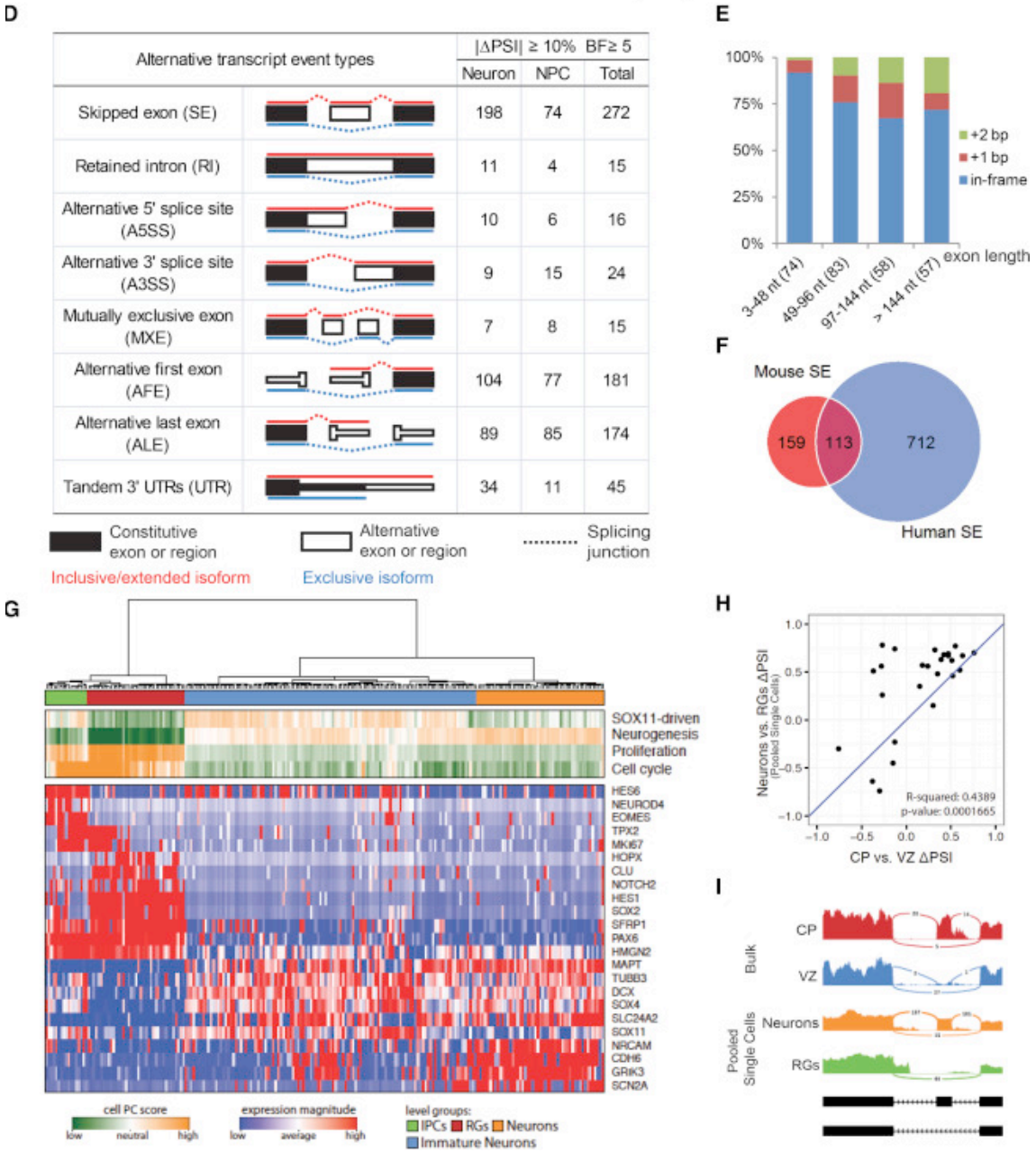


(Tirosh et al. *Science* 2016)

Cell specific alternative splicing



(Shalek et al. *Nature* 2013)



(Zhang et al *Cell* 2016)