UAB MEDICINE
Knowledge that will change your world

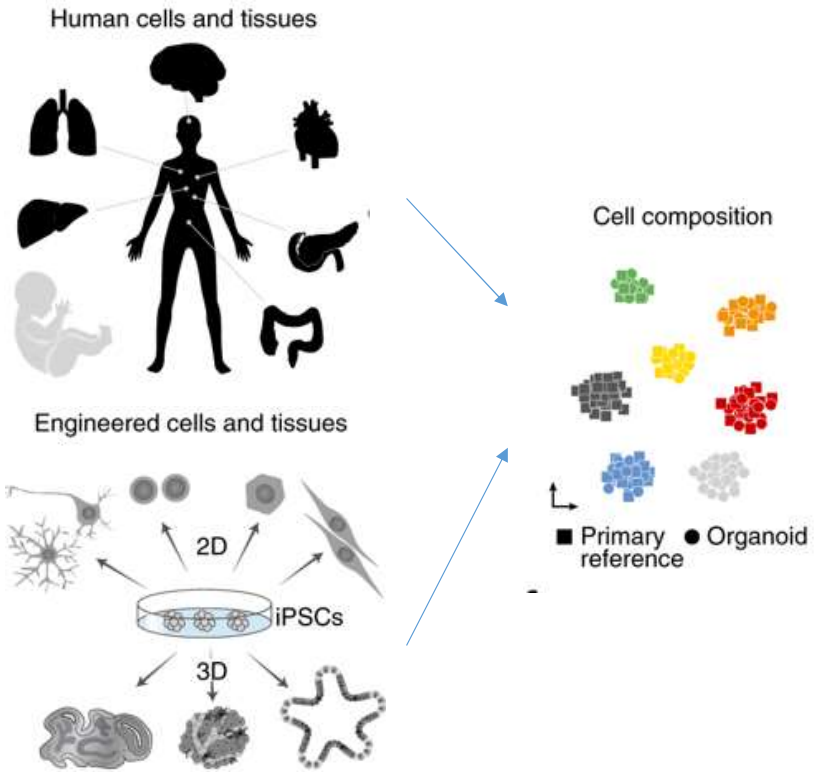
INFORMATICS INSTITUTE

Computational techniques in single-cell genomics data analysis

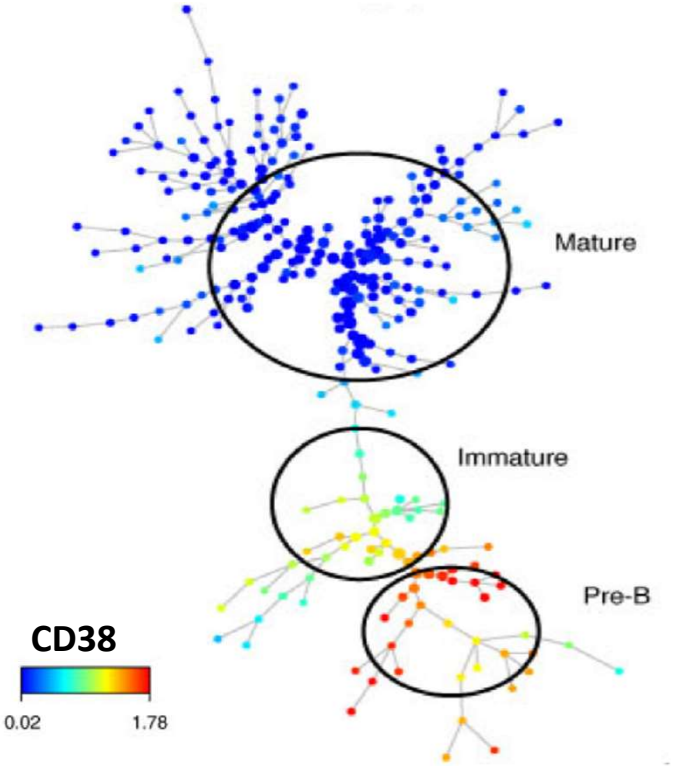
Jake Y. Chen, PhD
jakechen@uab.edu

March 29th 2019

Single-cell genomic analysis allows us to understand cellular compositions and their relationships



Nature Methods **vol. 15**, pp. 661–667 (2018)

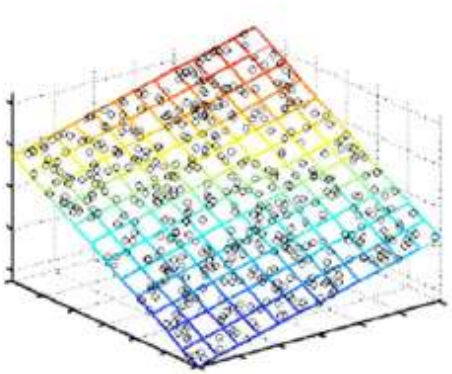


Nature Protocols. 11.7: p1264+ (2016)

Outline

- **Dimensionality reduction techniques**
- 2-D visualization of single-cell expression data
- Pseudo-time developmental trajectory analysis
- Single-cell genomic analysis pipelines

Dimensionality reduction for single-cell analysis



$3d \Rightarrow 2d$

- Each single cell involves huge numbers of features (dimensions)
  - $10^{2\text{-}3}$ curated gene categories
  - 2000-5000 gene transcripts
  - 10^6 genetic variations
- High dimensionality has a high cost
  - Redundant and irrelevant features degrade algorithm performance
  - Difficulty in interpretation and visualization
  - Computation may become infeasible
  - Curse of dimensionality: “overfitting” problem

Project *n*-dimensional data onto a *k*-dimensional space (*k* << *n) to reduce noise and help with data explorations*

Approaches to dimensionality reduction

- Feature selection
  - Select subset of existing features (without modification)
  - Could introduce bias
- Model regularization
  - reduces *effective & actual* dimensionality
  - Not always feasible due to characteristics of data
- Map existing features into smaller number of new features
  - Linear combination methods(projection)
  - Nonlinear combination methods

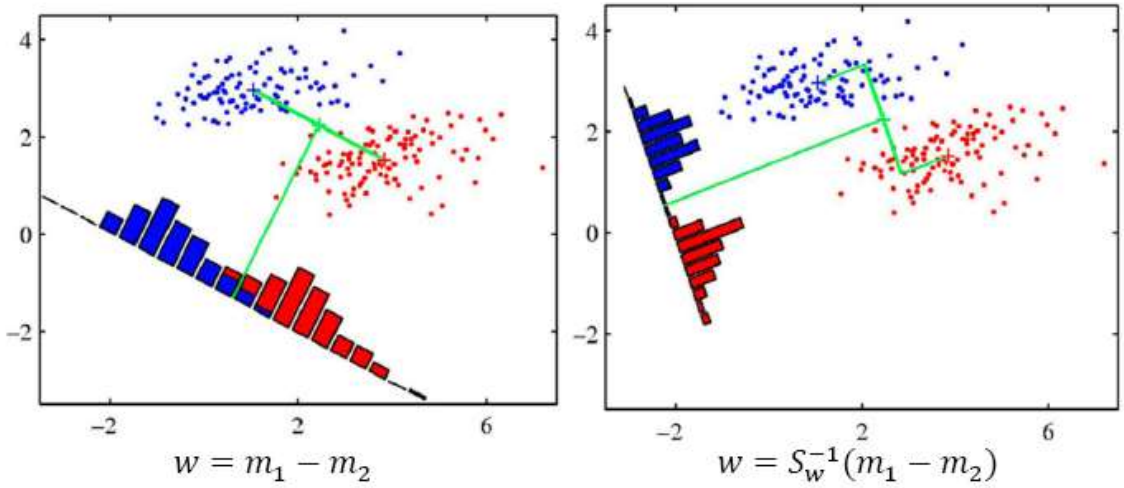
May be
* **Supervised**, e.g., classification using linear discriminant analysis (LDA)
* **Unsupervised**, e.g, principal component analysis (PCA), Multidimensional Scaling (MDS), random space projections (RSP)

Linear Discriminant Analysis (*LDA*)

a supervised classification method for dimensionality reduction

$$\mathbf{w} = S_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

- Projecting data onto one dimension that maximizes the ratio of between-class scatter and total within-class scatter

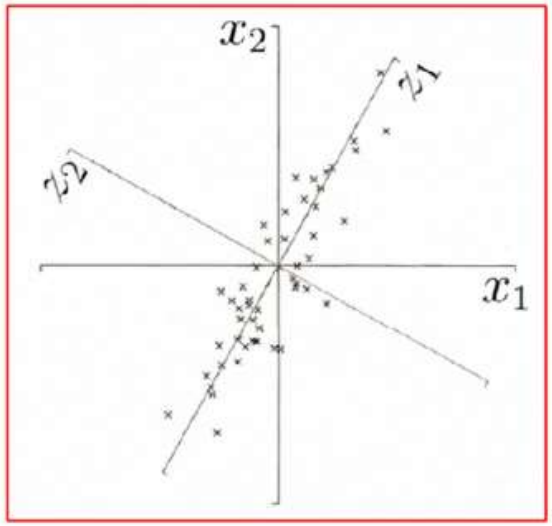


Principle component analysis (*PCA*)

an unsupervised method for dimensionality reduction

GOAL: account for variance of data in as few dimensions as possible (using linear projection)

- PC1 is the projection direction that maximizes the variance of the projected data
- PC2 is the projection direction that is orthogonal to PC1 and maximizes variance of the projected data

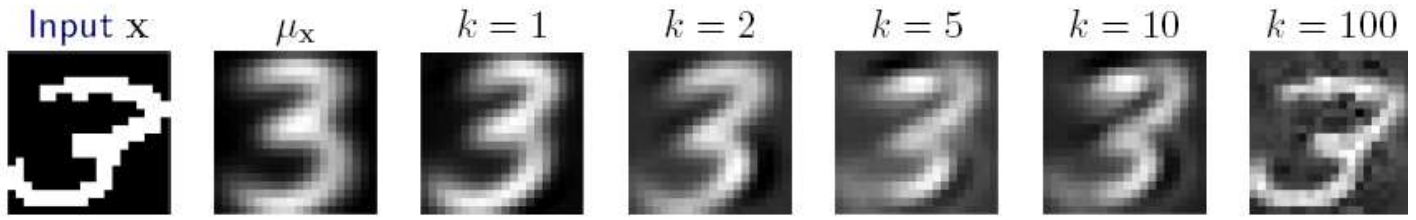


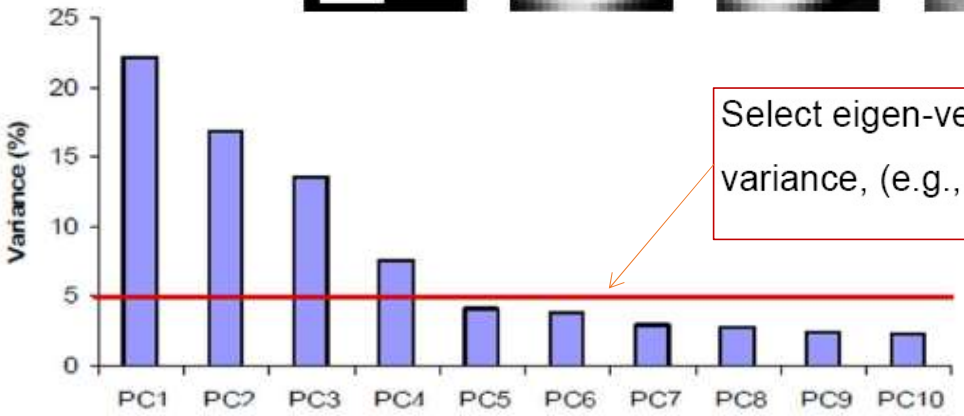
How PCA Works

- Mean center the data
- Compute covariance matrix Σ
- Calculate eigenvalues and eigenvectors of Σ
  - Eigenvector with largest eigenvalue λ_1 is 1st PC
  - Eigenvector with k^{th} largest eigenvalue λ_k is k^{th} PC
  - λ_k / Σ_i λ_i = proportion of variance captured by k^{th} PC

PCA: k-dimension reduction and limitations

Choosing K dimensions





Select eigen-vectors that retain a fixed percentage of the variance, (e.g., 80%, the smallest d such that $\frac{\sum_{i=1}^{d} \lambda_i}{\sum_i \lambda_i} \geq 80\%$)

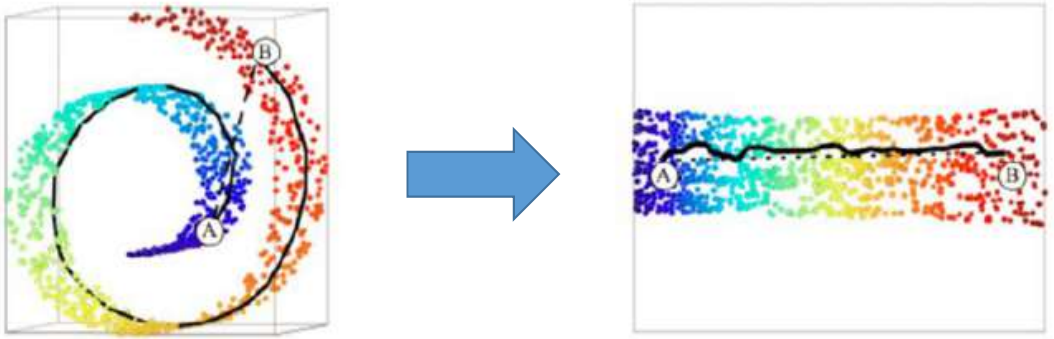
PCA Limitations

- Slow to calculate covariance matrix n x n
- Fails when data consists of multiple separate clusters.
- Directions of greatest variance may not be most informative.

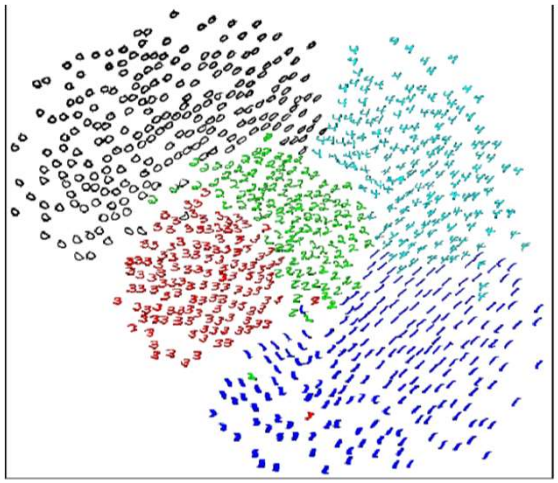
Isometric Feature Mapping (*ISOMAP*)

A non-linear dimensionality reduction method

- Data often lies on or near a nonlinear low-dimensional surface called *manifolds*.
- Aims to preserve the global nonlinear geometry of the data by preserving the geodesic distances
- Geodesic distance: the shortest route between two points on the surface of the manifold, e.g., A to B not following Eucleadian distance



t-Stochastic Neighbor Embedding (*t-SNE*)



- Reduce dimensionality while preserving local similarity
- A heuristic method to reveal a map structure at many different scales.
- Based on earlier work of “Stochastic neighbor embedding” (SNE)
- Good if high-dimensional data lie on low-dimensional manifolds

Stochastic Neighbor Embedding (SNE)

SNE starts by converting the Euclidean distances between high-dimensional datapoints into conditional probabilities that represent similarity. It can be described as:

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)},$$

For the low-dimensional counterparts y_i and y_j of the high-dimensional datapoints x_i and x_j,

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}.$$

11

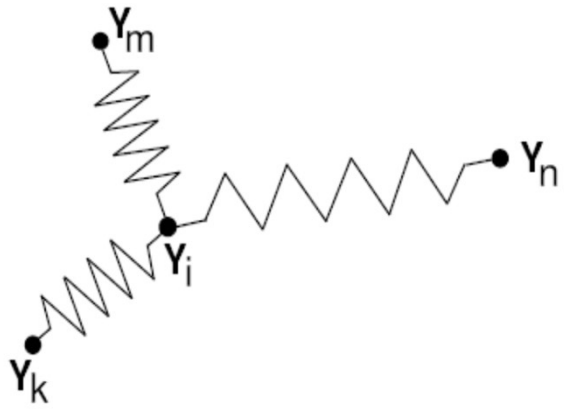
Kullback-Leiber (KL) Divergence as the faithfulness measure

- SNE aims to find a low-dimensional data representation that minimizes the mismatch between $p_{j|i}$ and $q_{j|i}$.
- Kullback-Leibler (KL) divergence is used to measure the faithfulness in which $q_{j|i}$ models $p_{j|i}$

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}},$$

$P_i = \{p_{1|i}, p_{2|i}, ..., p_{n|i}\}$ and $Q_i = \{q_{1|i}, q_{2|i}, ..., q_{n|i}\}$ are the distributions on the neighbors of datapoint i

SNE Gradient Descent Optimization



Physical interpretation is spring force models of y_i to each of the other points (y_m, y_n, y_k as shown). The spring between i and j exerts a force proportional to its length (y_i-y_j) and

$$(p_{j|i} - q_{j|i}) + (p_{i|j} - q_{i|j})$$

The similarity measure using KL divergence

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}},$$

The similarity measure using KL divergence

$$\frac{\delta C}{\delta y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j).$$

In order to speed up the optimization and to avoid poor local minima, add a momentum term

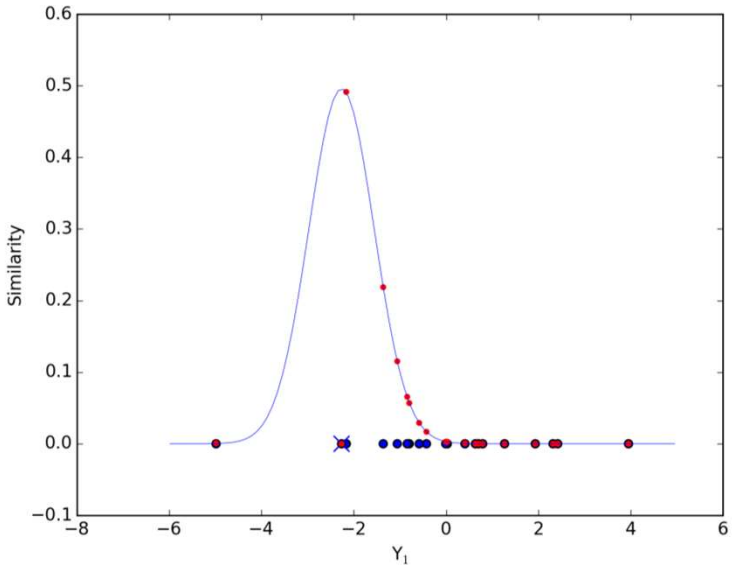
$$\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t) \left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)} \right)$$

momentum term

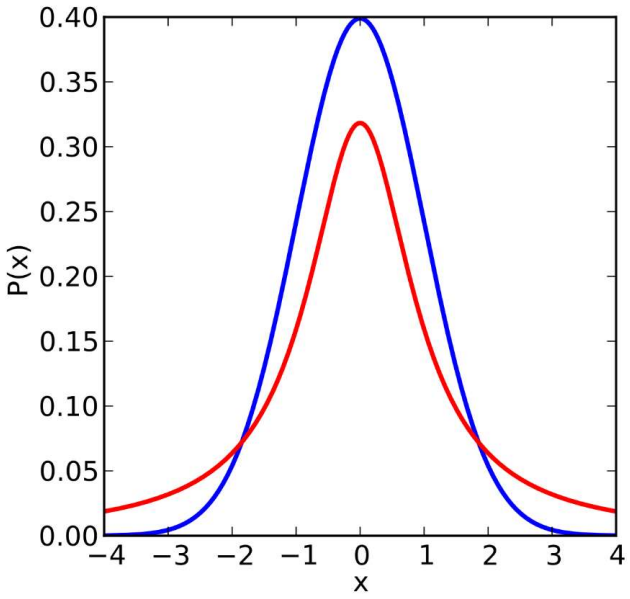
The “crowding” problem and solution

using asymmetric student t-distribution

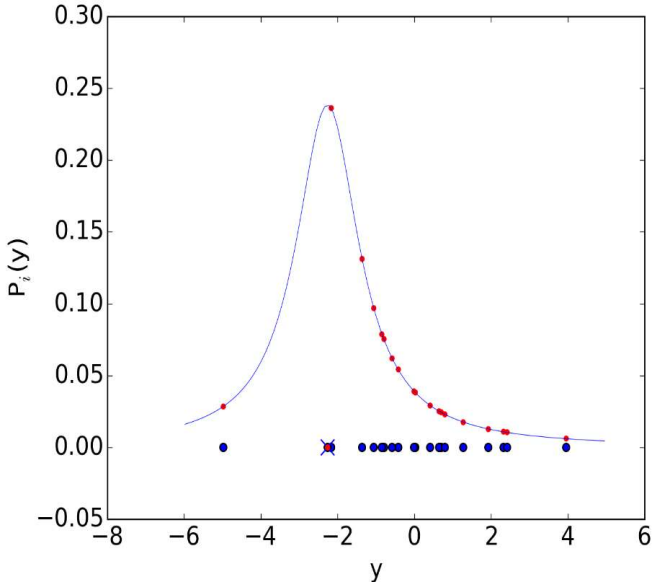
Standard normal distribution (blue)
t-distribution with df =1 (red)



There is much more space in high dimensions than in low dimensions.



Student-t distribution has heavier tails.



The t-distribution’s heavier tails can reduce “crowding” in low dimensions.

t-SNE: less crowding, faster than SNE

- It uses a symmetrized version of the SNE cost function to improve performance
- It uses a t-distribution rather than a Gaussian to compute the similarity between two points in the low-dimensional space.

SNE ⇒ Symmetric SNE ⇒ t-SNE

Modelisation:

SNE:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)}$$

$$q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k\neq i}\exp(-\|y_i - y_k\|^2)}$$

Symmetric SNE:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

$$q_{ij} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k\neq l}\exp(-\|y_k - y_l\|^2)}$$

t-SNE:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

$$q_{ij} = \frac{(1+\|y_i - y_j\|^2)^{-1}}{\sum_{k\neq l}(1+\|y_k - y_l\|^2)^{-1}}$$

Cost Function:

SNE:

$$C = \sum_i KL(P_i||Q_i)$$

Symmetric SNE:

$$C = KL(P||Q)$$

t-SNE:

$$C = KL(P||Q)$$

Derivatives:

SNE:

$$\frac{dC}{dy_i} = 2\sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

Symmetric SNE:

$$\frac{dC}{dy_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j)$$

t-SNE:

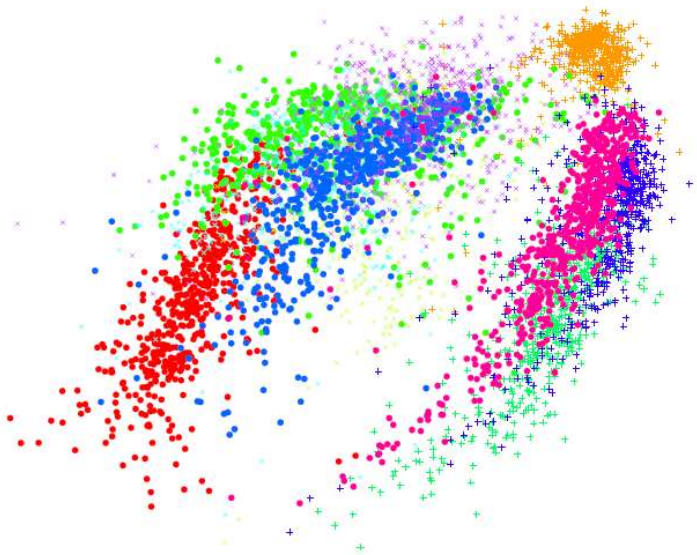
$$\frac{dC}{dy_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j)(1+\| y_i - y_j \|^2)^{-1}$$

The asymmetric cost function is difficult to optimize

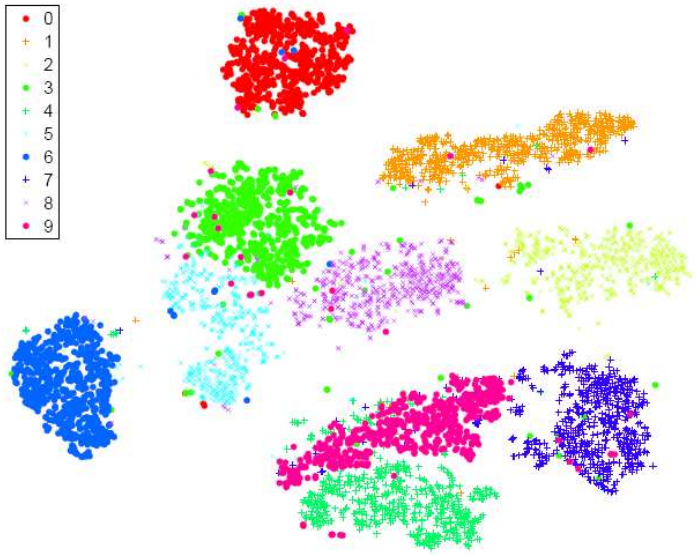
The use of t-statistic

15

t-SNE vs. ISOMAP



ISOMAP



t-SNE

Visualization of classes in MNIST data

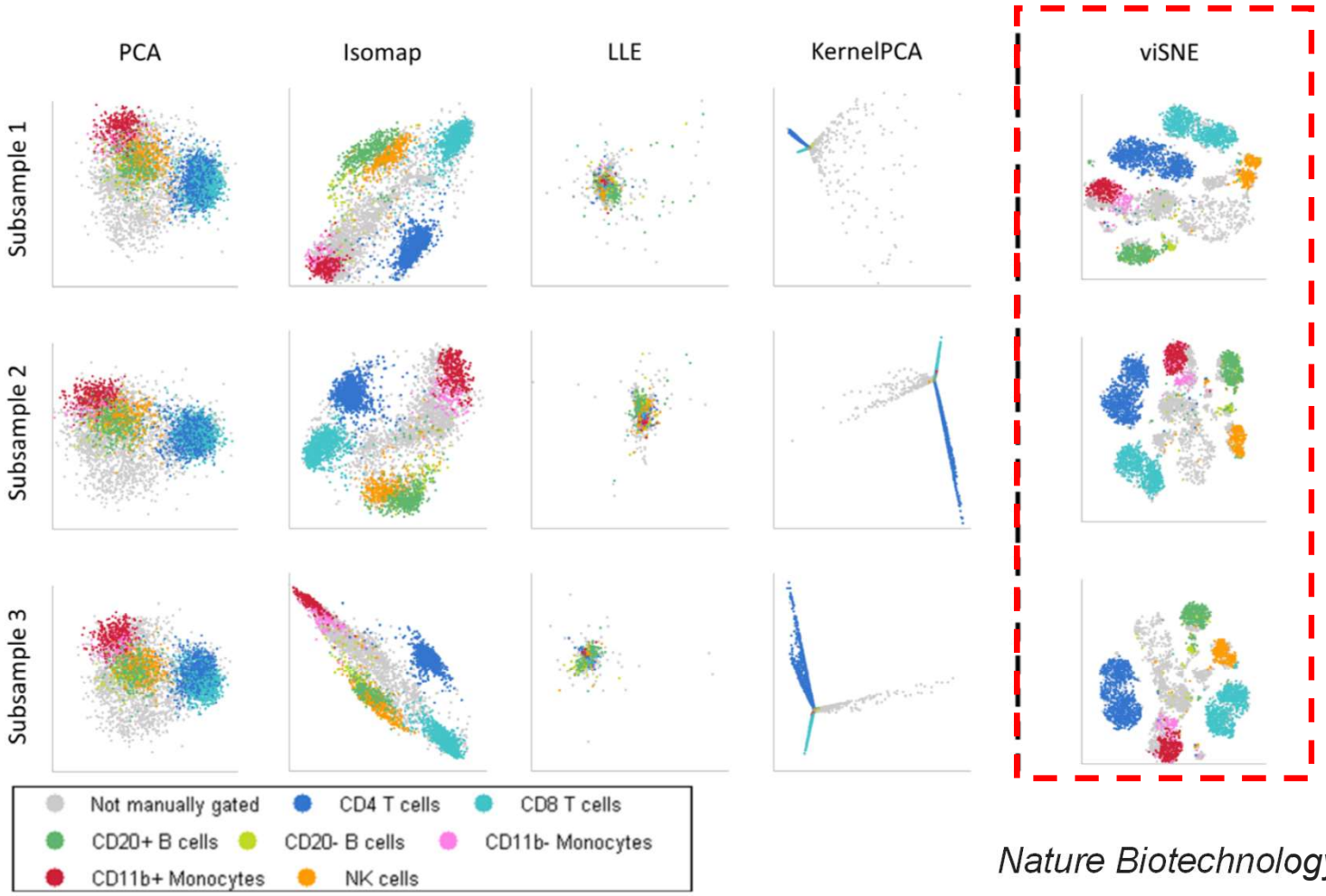
Literature references and further reading

- "Dimensionality reduction: a comparative review"
- MATLAB toolbox for dimensionality reduction
- "An Introduction to Statistical Learning, with applications in R" (Springer, 2013)
- G.Hinton and S.Roweis. *Stochastic neighbor embedding*. NIPS03(15) : 833-840.
- J. Cook, I.Sutskever,A.Mnih and G.Hinton. *Visualizing similarity data with a mixture of maps*, In proceedings of the 11th international Conference on Artificial Intelligence and Statistics,2007(2):67-74.
- L.van der Matten and G.Hinton. *Visualizing data using t-SNE*. Journal of Machine Learning Research,2008(9):2579-2605.
- Xiaohong Chen "Stochastic Neighbor Embedding and Its Variants"

Outline

- Dimensionality reduction techniques
- **2-D visual clustering of single-cell expression data**
- Pseudo-time developmental trajectory analysis
- Single-cell genomic analysis pipelines

viSNE: an implementation of t-SNE for scRNA-seq



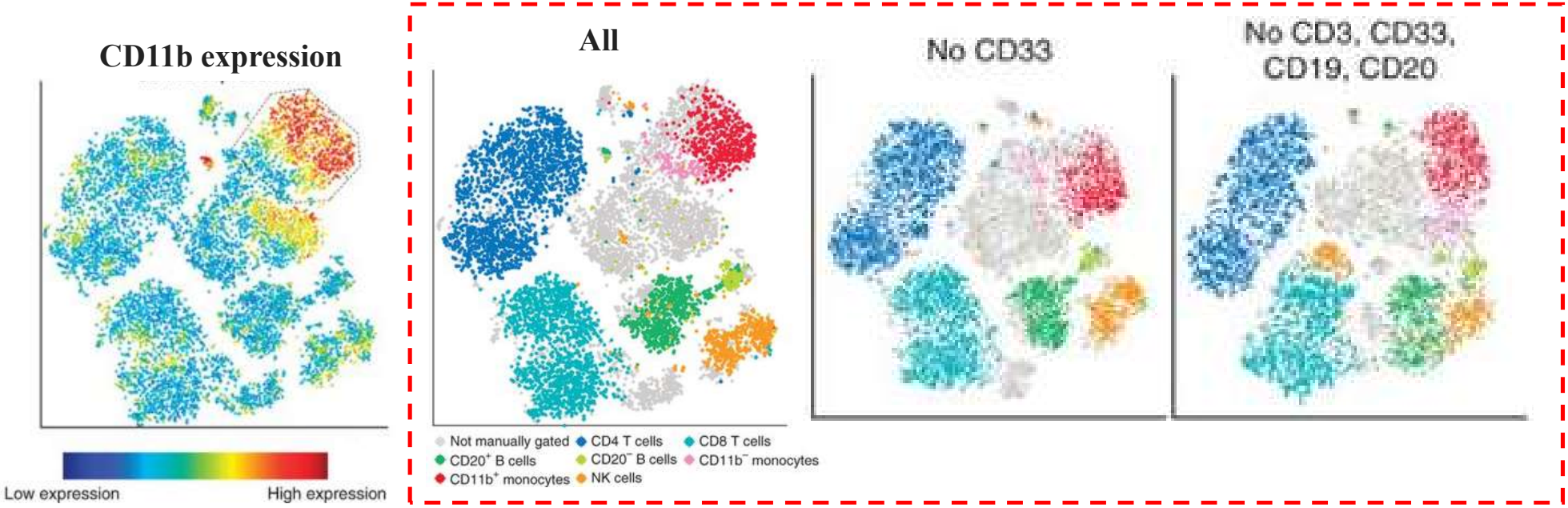
viSNE

- Stochastic layout
- Well separated subpopulations in clusters
- Need to combine samples to layout together if stable layout for comparison is desired

Nature Biotechnology **Vol 31**, pp545–552 (2013)

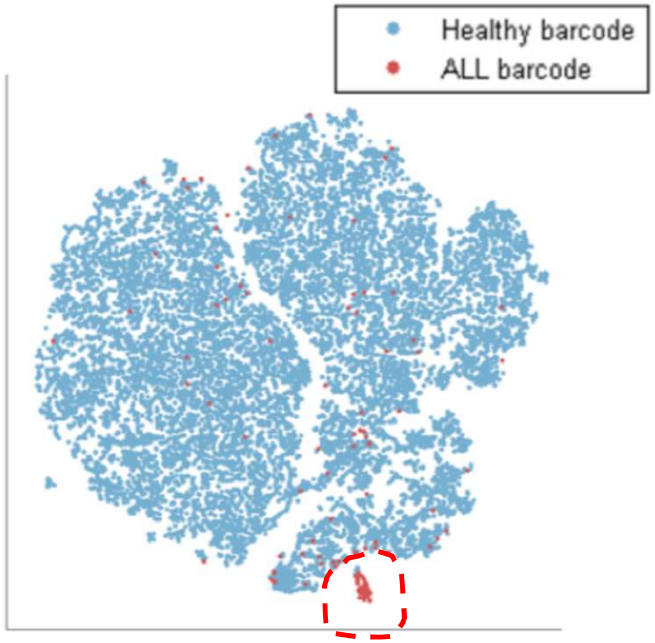
viSNE clusters are robust, against +/- of cytometry manually gated biomarkers

Layout is stable, not dominated by marker genes that defines the cell types

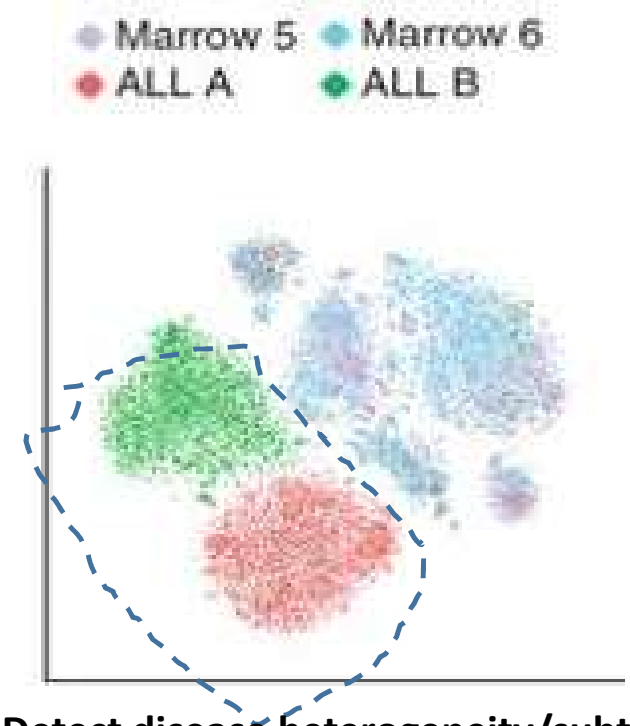


Sample: healthy human bone marrow, stained with 13 markers and measured with mass cytometry

viSNE in disease diagnosis and subtyping



Detect minimal residual disease (outlier detection)

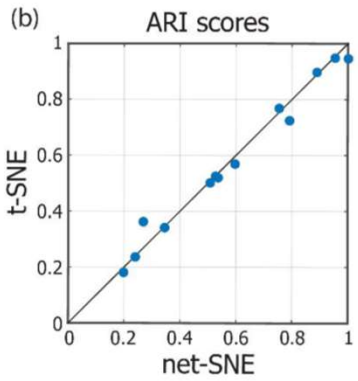


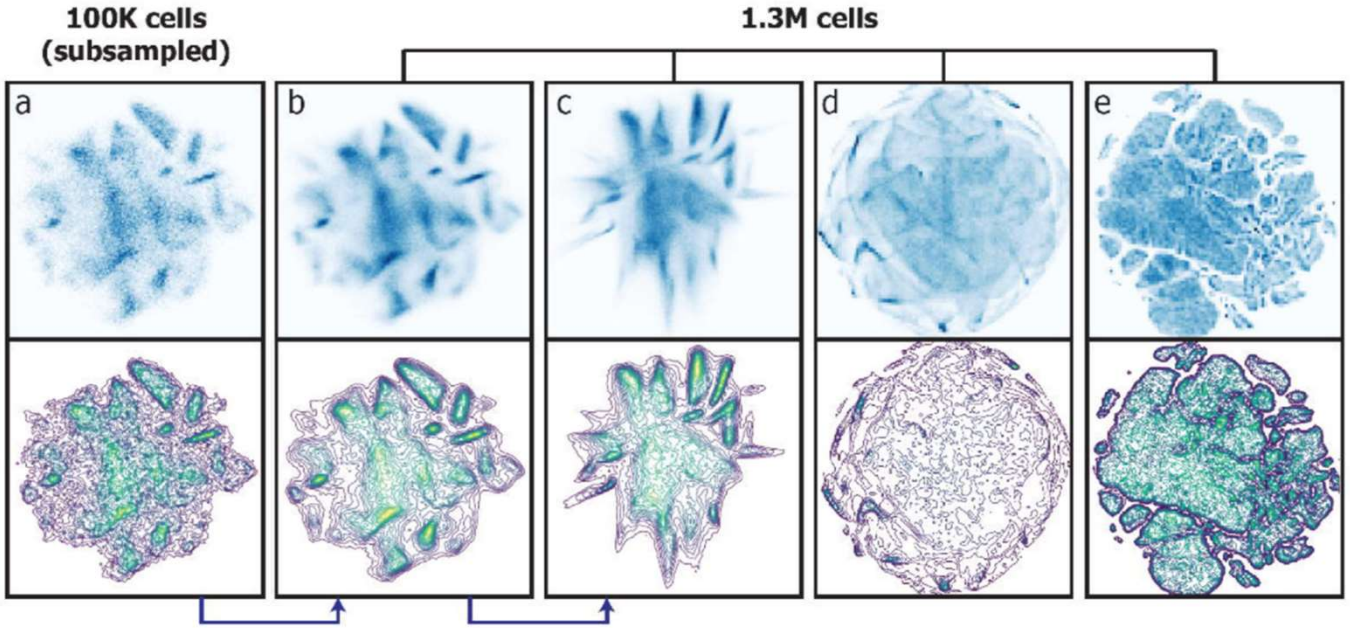
Detect disease heterogeneity/subtype

Neural t-SNE (*net-SNE*): fast, supervised embedding

Use two-hidden layer neural network to learn the mapping function parameters from high-dimensional space to low-dimensional space.

Adj. Rand Index (ARI)



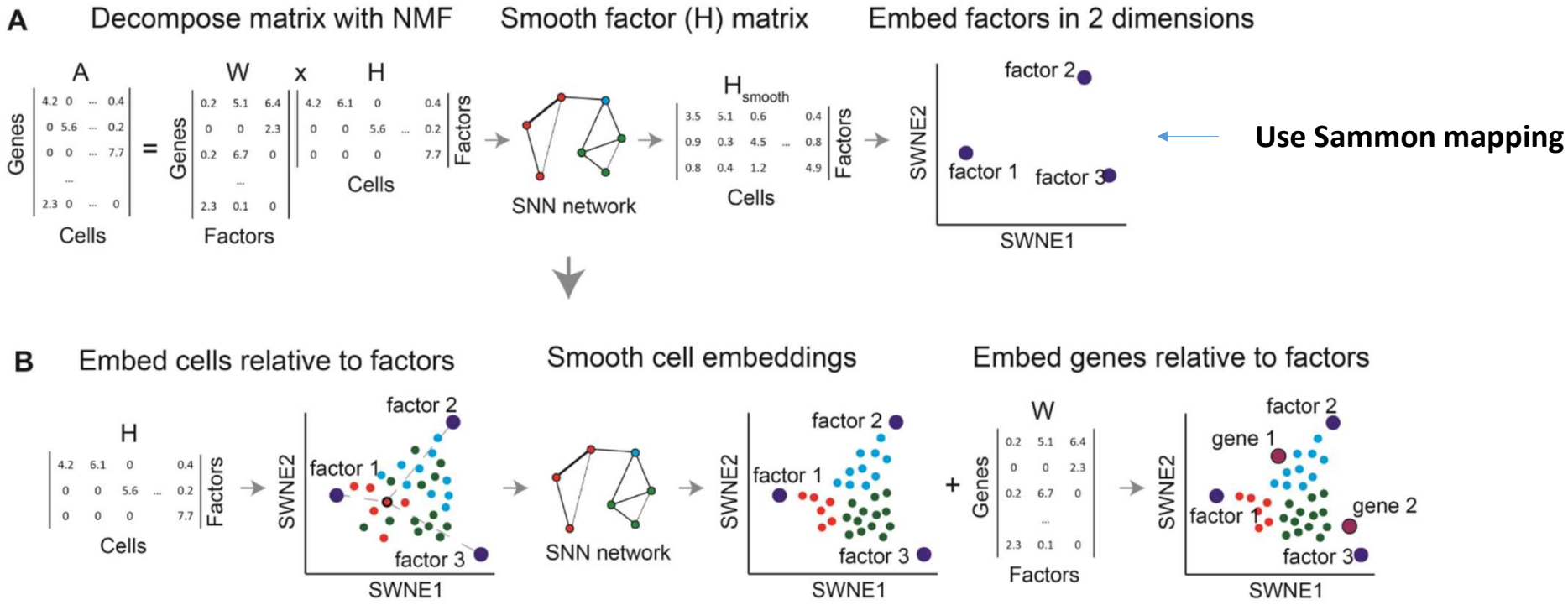


Method	net-SNE	generalize	net-SNE	further optimize	net-SNE	t-SNE (default)	t-SNE (10x Genomics)
Runtime	20 mins		-		45 mins	13 hrs	1.5 days*
Objective	-		7.01		6.51	8.00	6.54

doi: https://doi.org/10.1101/289223

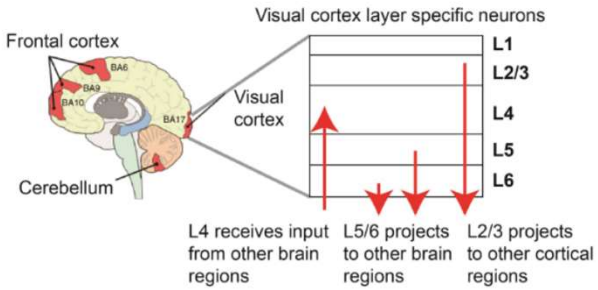
Similarity Weighted Nonnegative Embedding (*SWNE*)

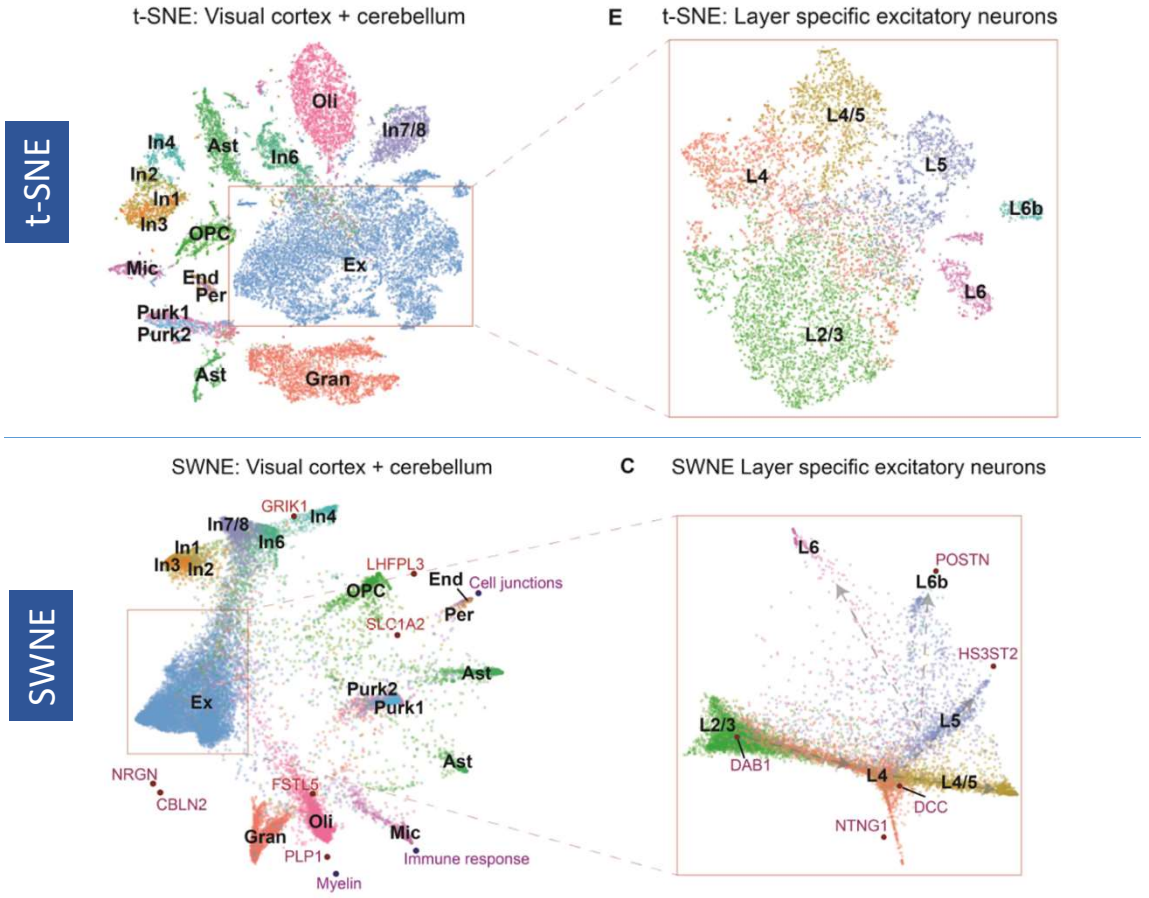
visualize cells and marker genes together for best interpretation



Cell Systems, Vol. 7, No. 6 (2018) pp. 565-566

SWNE helps interpret multi-scale information among cell populations next to marker genes

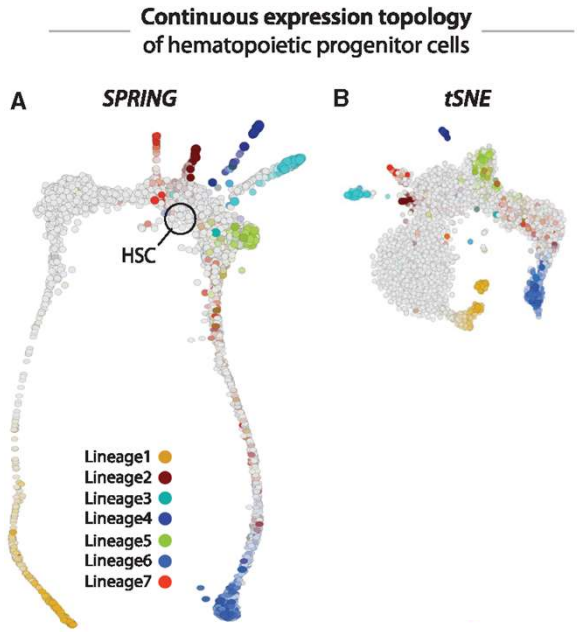




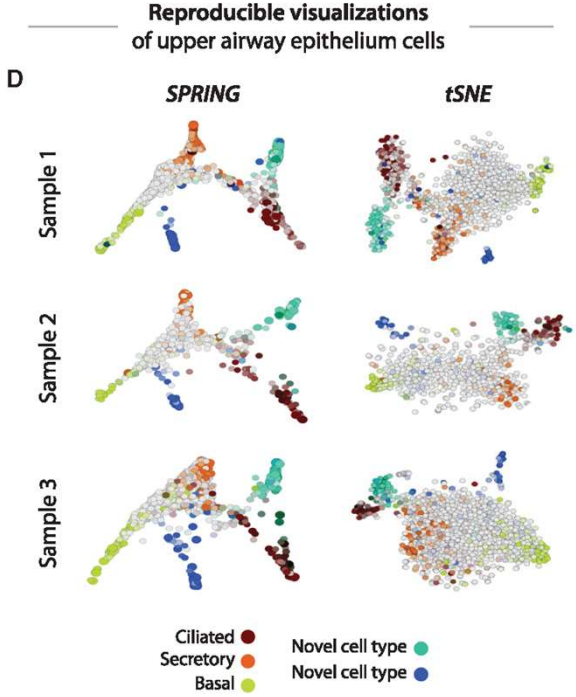
SPRING: stable kNN graph towards developmental trajectory analysis

SPRING uses kNN graph
each cell is a node that extends edges to the k other nodes with most similar gene expression.

Showing global features

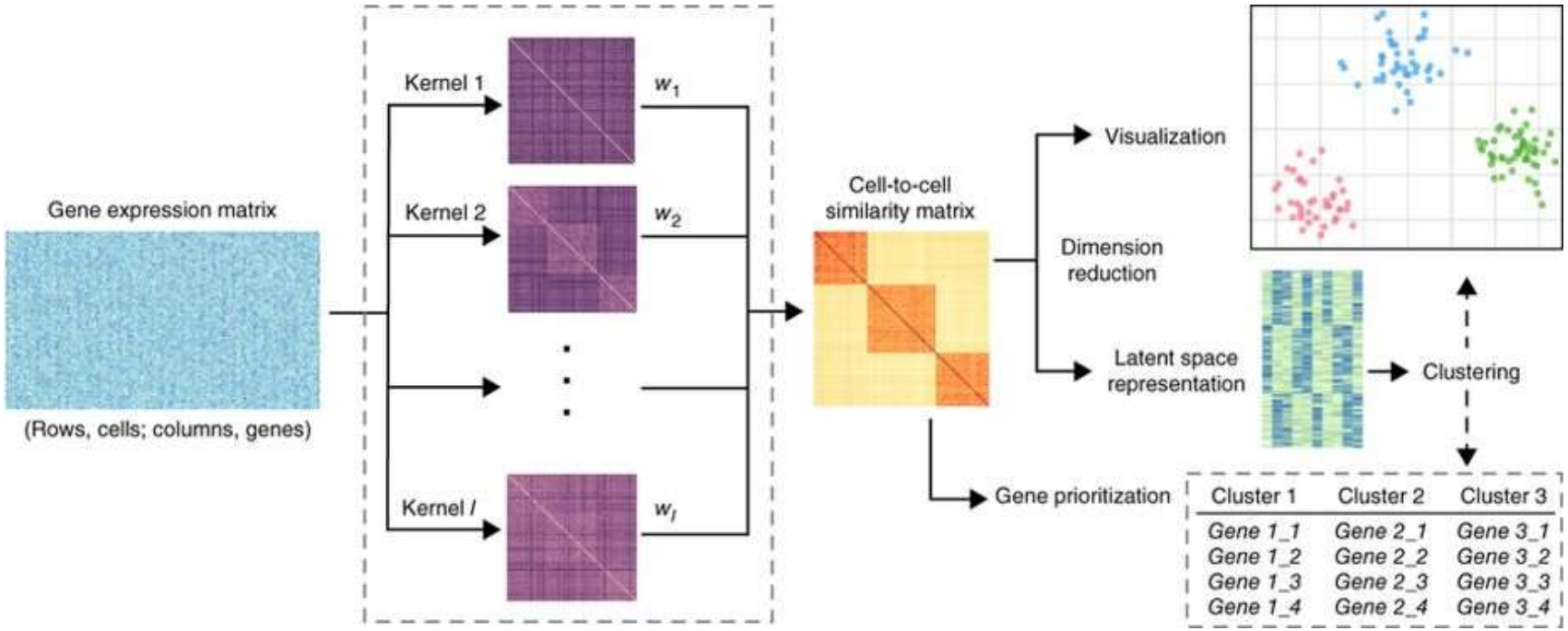


More stable embedding



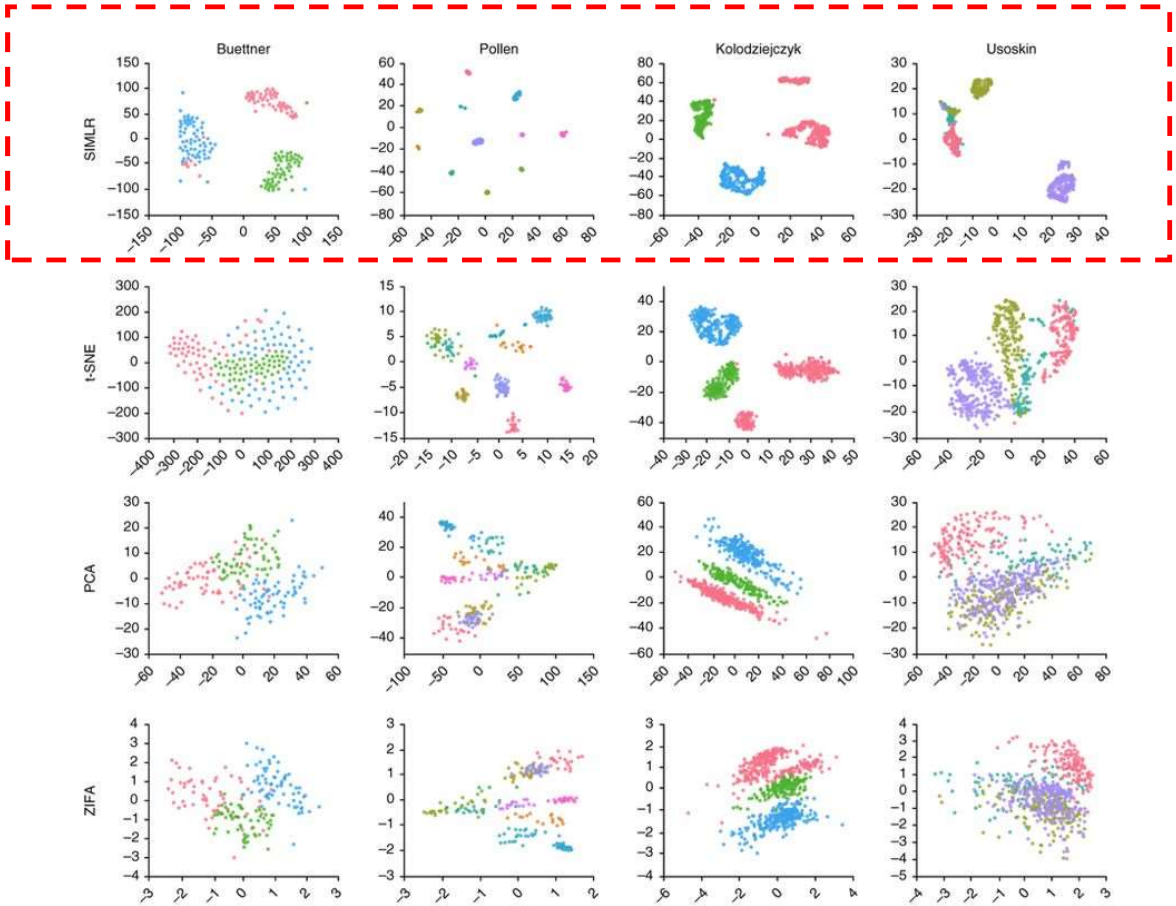
Bioinformatics, Vol. 34, No. 7 (2018) pp. 1246–1248

SIMLR: Added multi-kernel learning to enhance clusters



Nature Methods **vol. 14**, pp414–416 (2017)

SIMILR can succeed where t-SNE fails

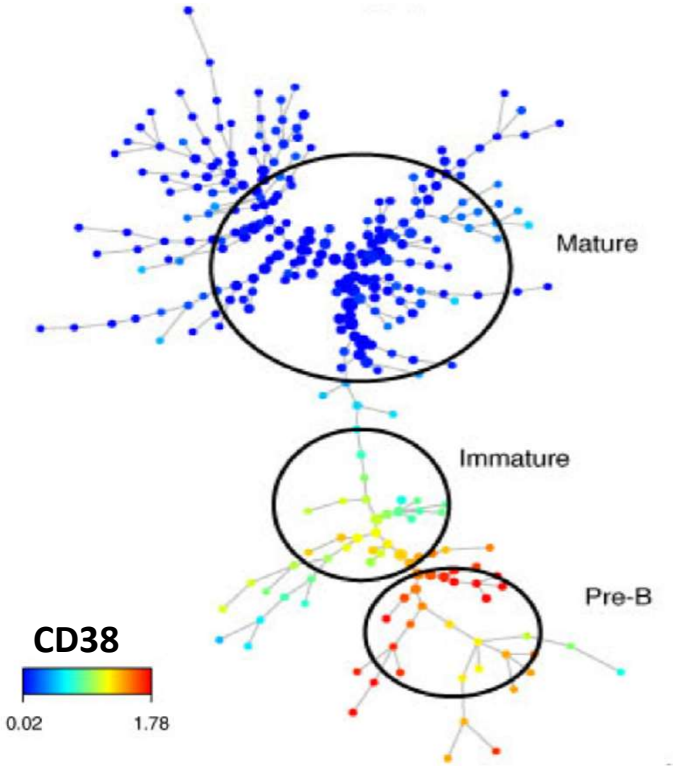


Outline

- Dimensionality reduction techniques
- 2-D visualization of single-cell expression data
- **Pseudo-time developmental trajectory analysis**
- Single-cell genomic analysis pipelines

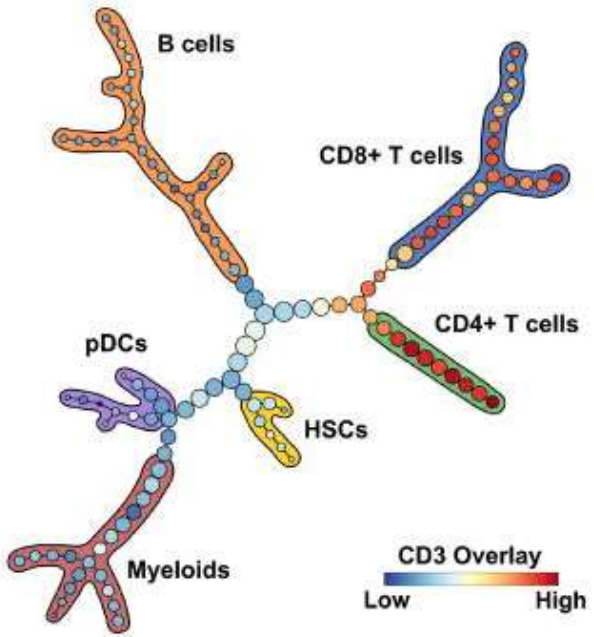
Pseudo-time Developmental Trajectory Analysis

SPADE Minimum Spanning Tree



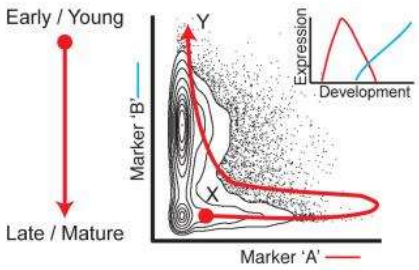
Nature Protocols. 11.7 (2016): p1264+

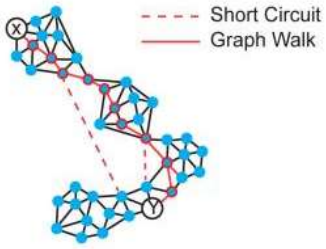
p-Creode multi-branched graph

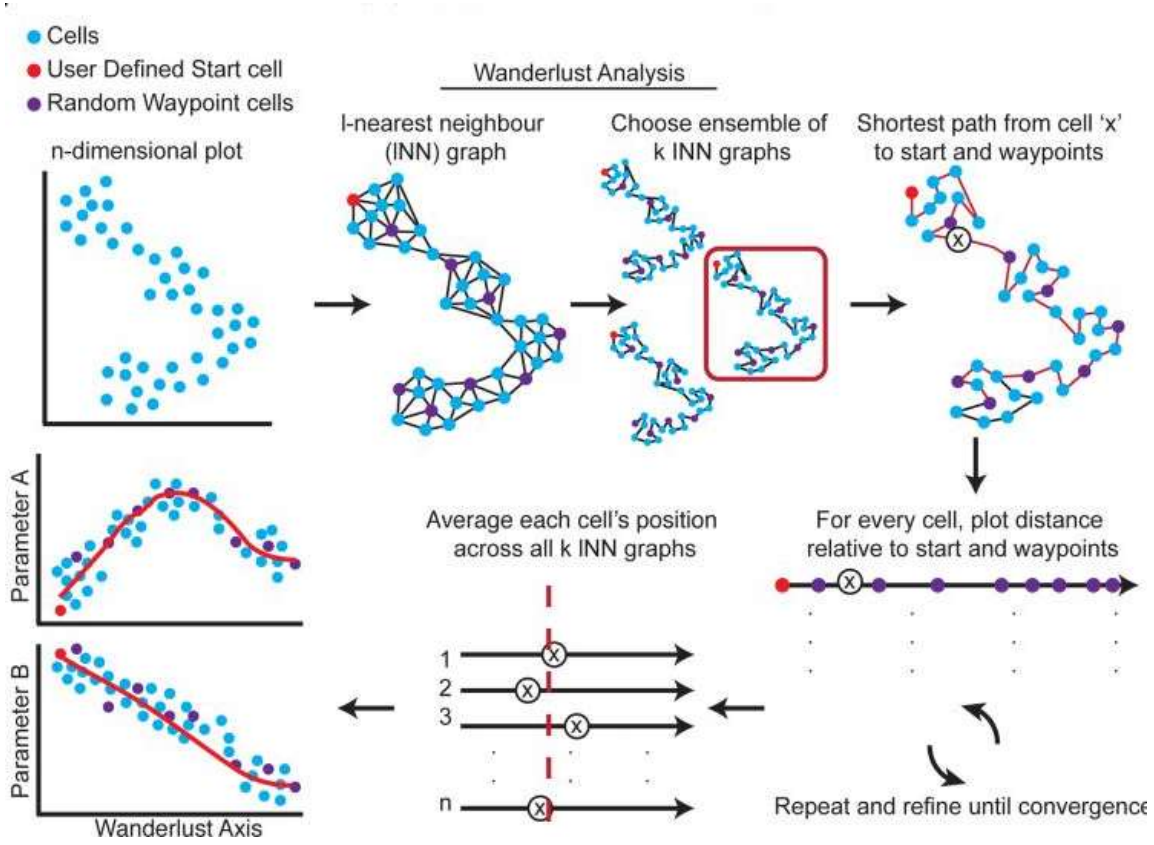


Cell Syst. 2018 Jan 24; 6(1): 37–51.e9.

The simple "*Wanderlust*" algorithm



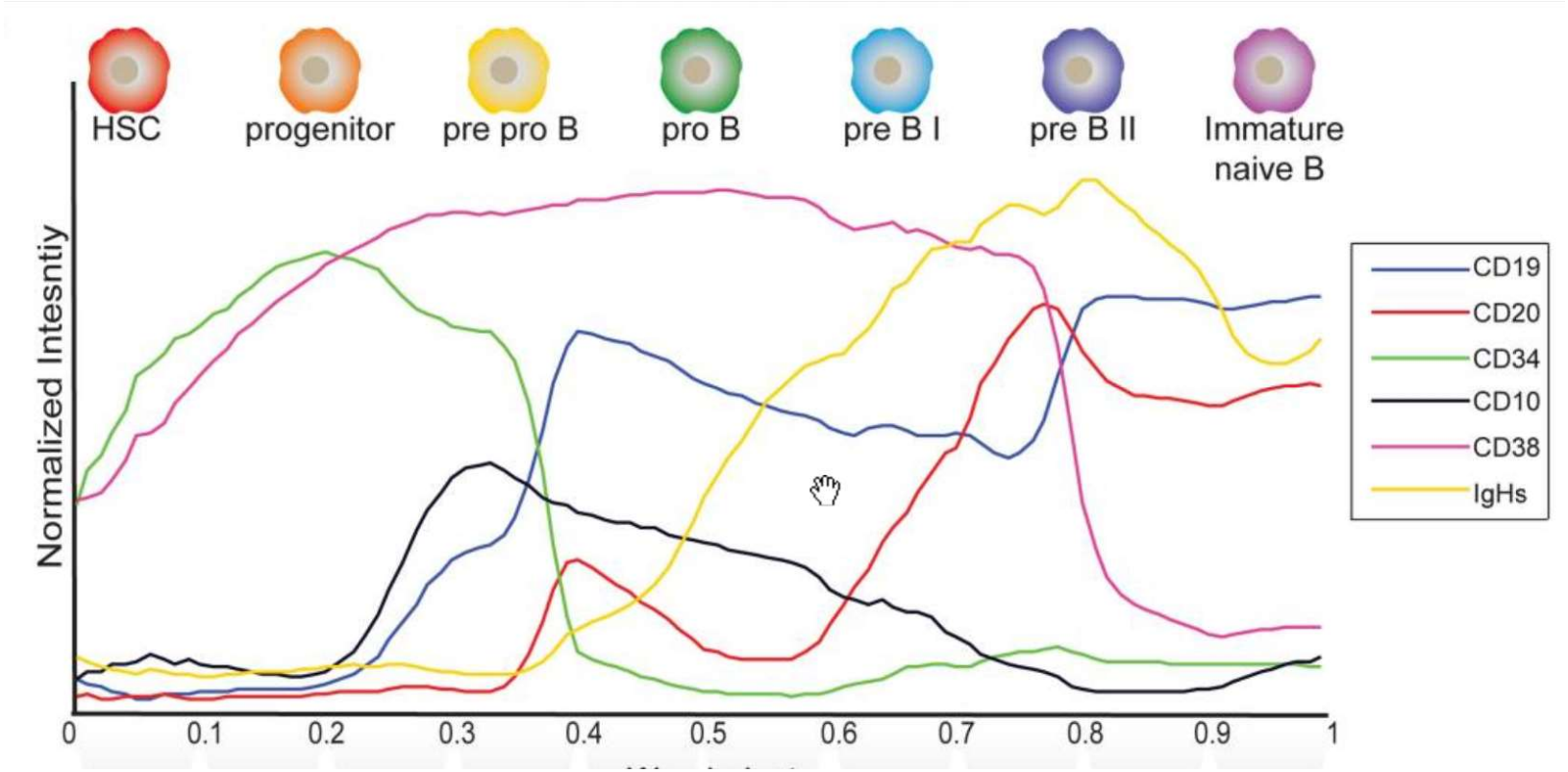




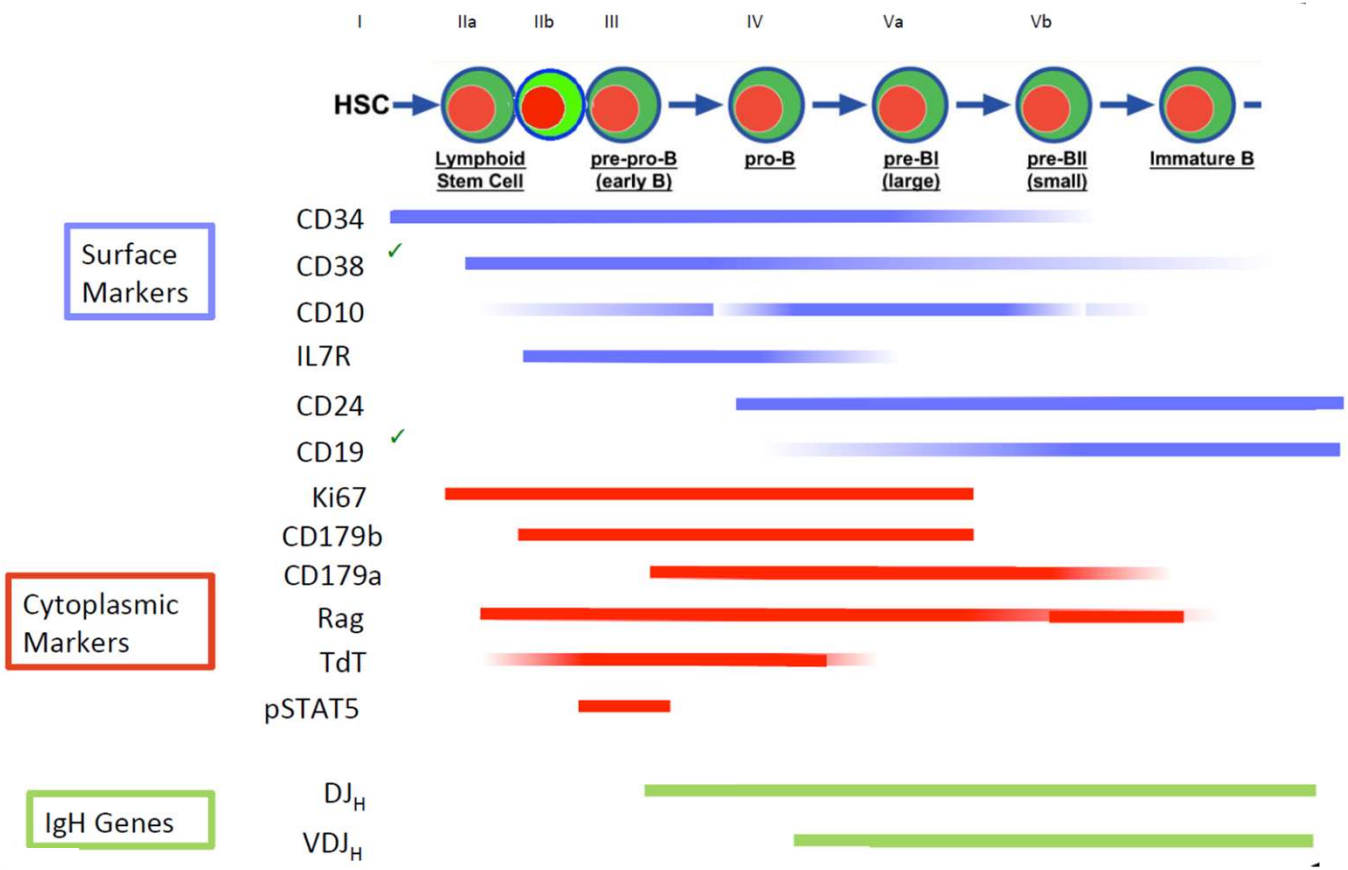
No bifurcations

Cell. 2014 Apr 24; 157(3): 714–725.

The predicted trajectory by the Wanderlust algorithm *rediscovers* human B cell development



New insights and resolution in human B cell development from scRNA-seq

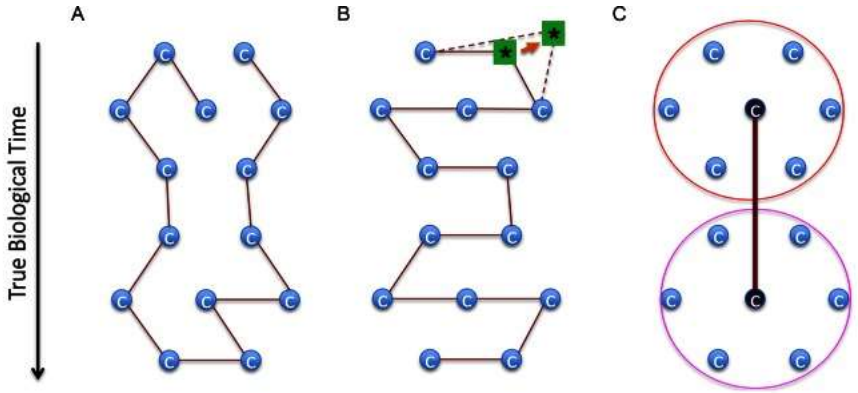


Advanced Trajectory Analysis

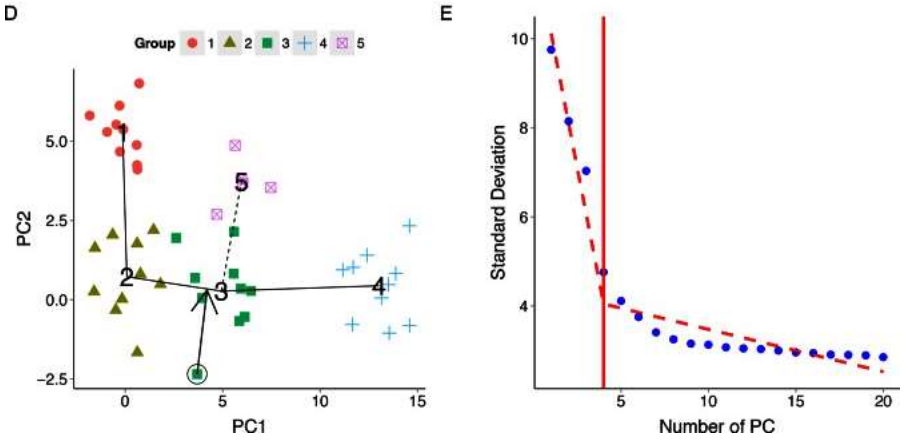
Use data-driven arrangement of cell states into pseudo-time progression trajectories to infer cellular transitions

- **Minimum Spanning Tree (MST) approaches**
  - TSCAN (Nucleic Acids Res. 2016;44:e117.), **SPADE** (Nat. Protoc. 2016;11:1264–1279)
  - Unstable, under-performing in less-defined systems
- **Non-linear embedding approaches**
  - Diffusion maps (Bioinformatics. 2015;31:2989–2998), Wishbone (Nat. Biotechnol. 2016;34:637–645), SLICER (Genome Biol. 2016;17:106), **DensityPath** (Bioinformatics, 2019, 1–9 doi: 10.1093/bioinformatics/bty1009)

TSCAN constructs Minimum Spanning Tree (MST) after clustering of cells



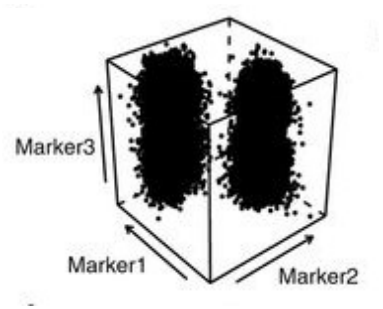
Clustering improves the chance of sorting order to follow biological time



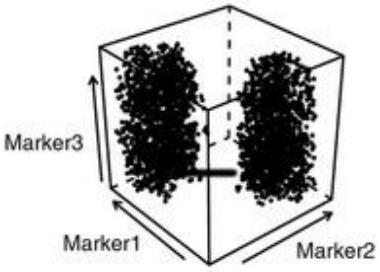
TSCAN works by connecting clustering centroid by MST

Nucleic Acids Res. 2016 Jul 27; 44(13): e117.

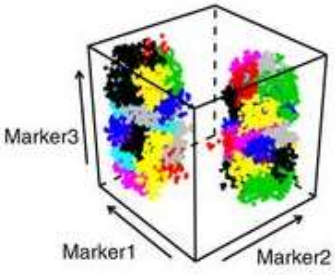
Spanning-tree Progression Analysis of Density-normalized Events (SPADE)



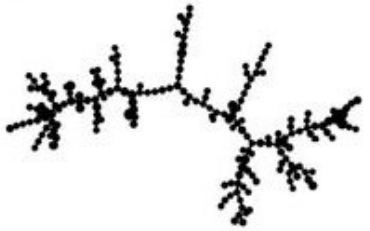
3-D mass cytometry data



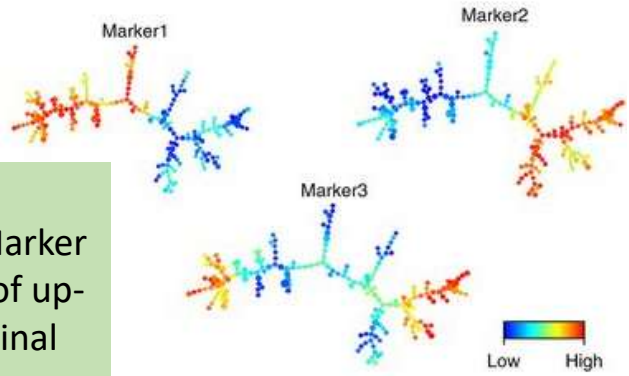
Density-dependent Down-sampled data



Agglomerative clusters



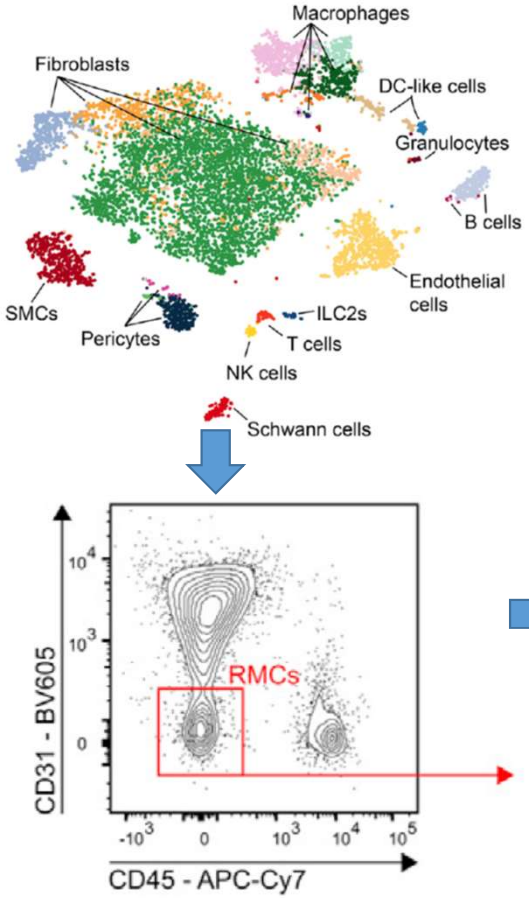
SPADE Minimum Spanning Tree



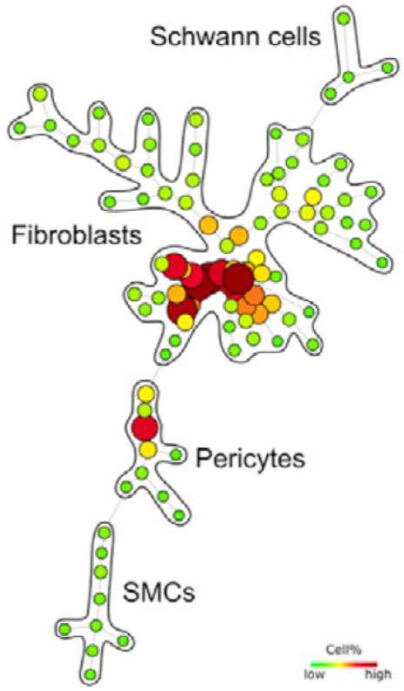
SPADE Tree colored by Marker Expressions of up-sampled original cell populations

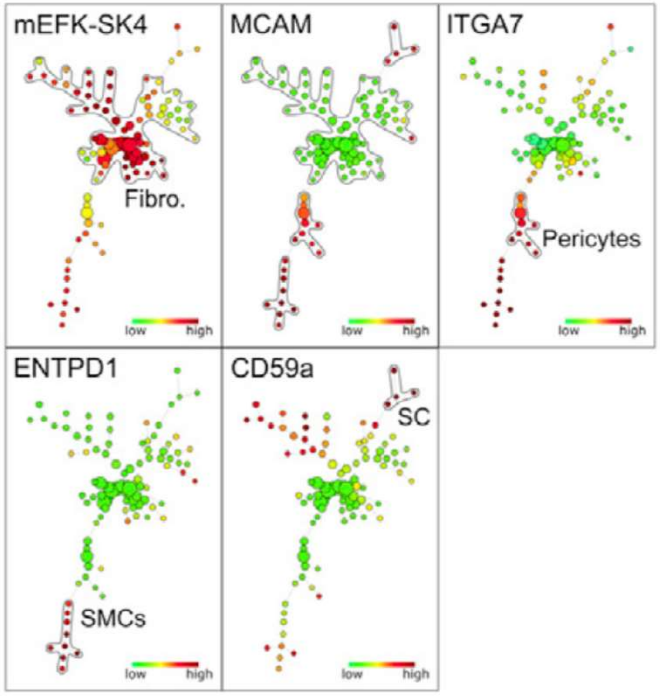
Nature Protocols. 11.7 (2016): p1264

Use SPADE to characterize the cardiac mural cell subpopulations from the cardiac cellulome



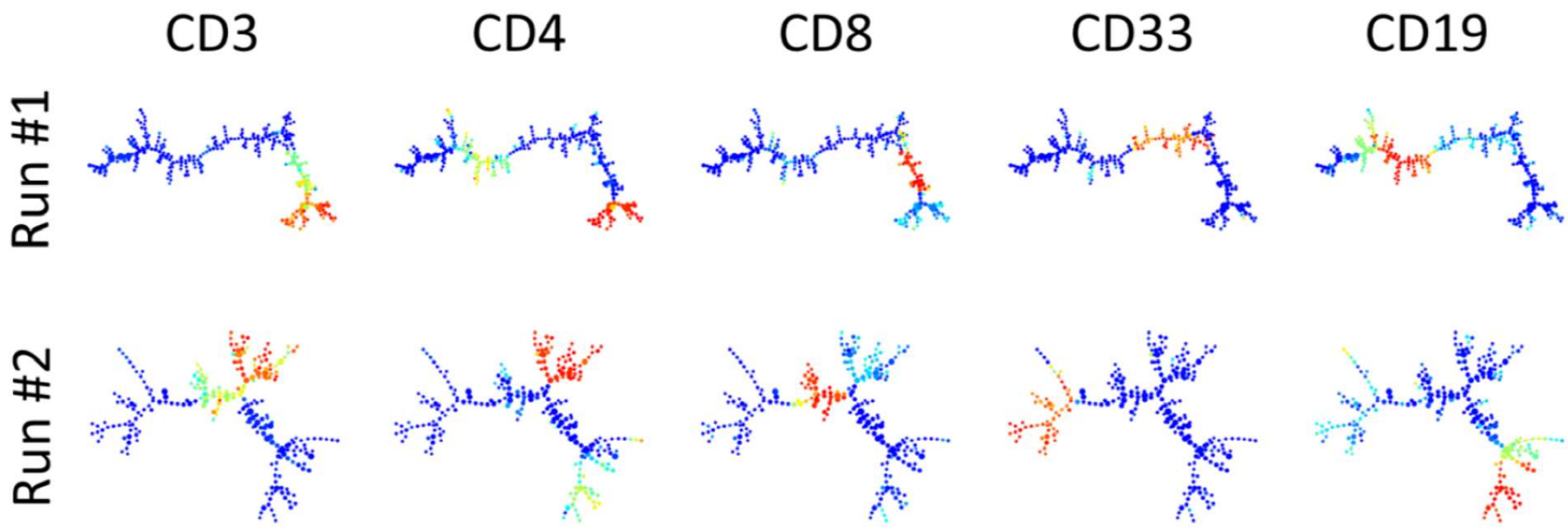
SPADE graph of SMCs, pericytes, fibroblasts, and Schwann cells





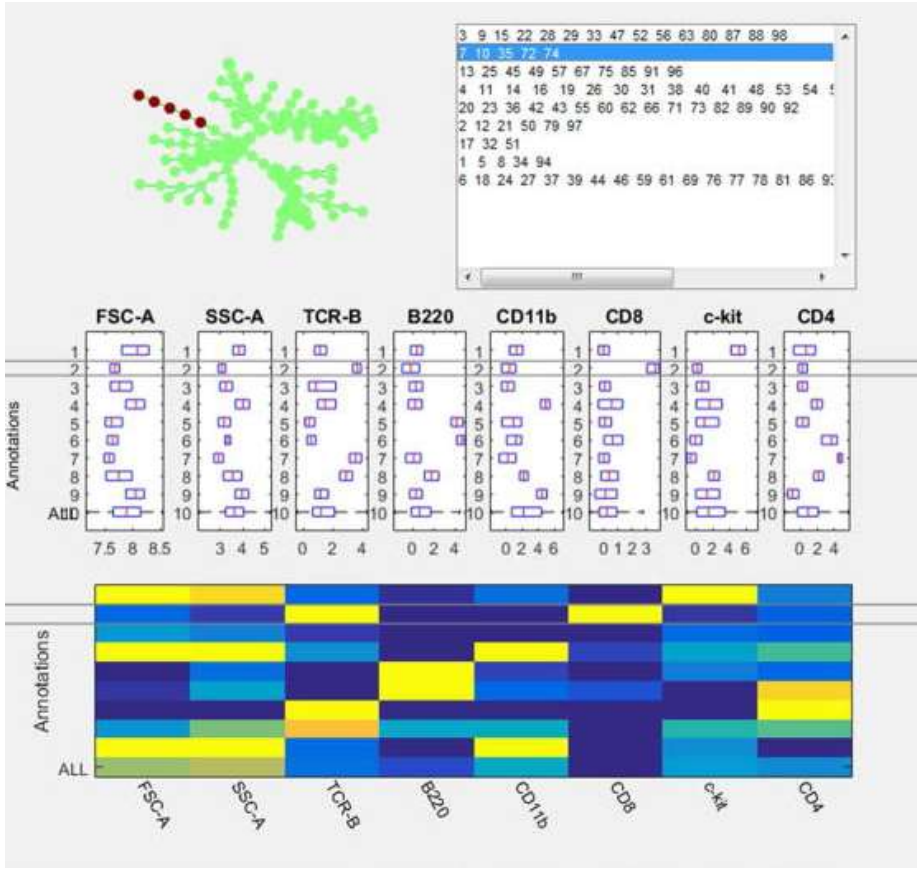
Skelly et al., 2018, Cell Reports 22, 600–610

Limitation: Different SPADE runs → unstable results



In SPADE, clusters group by immune subtypes (shorter range) are Okay, but longer range distances are less conserved between runs.

Deterministic SPADE

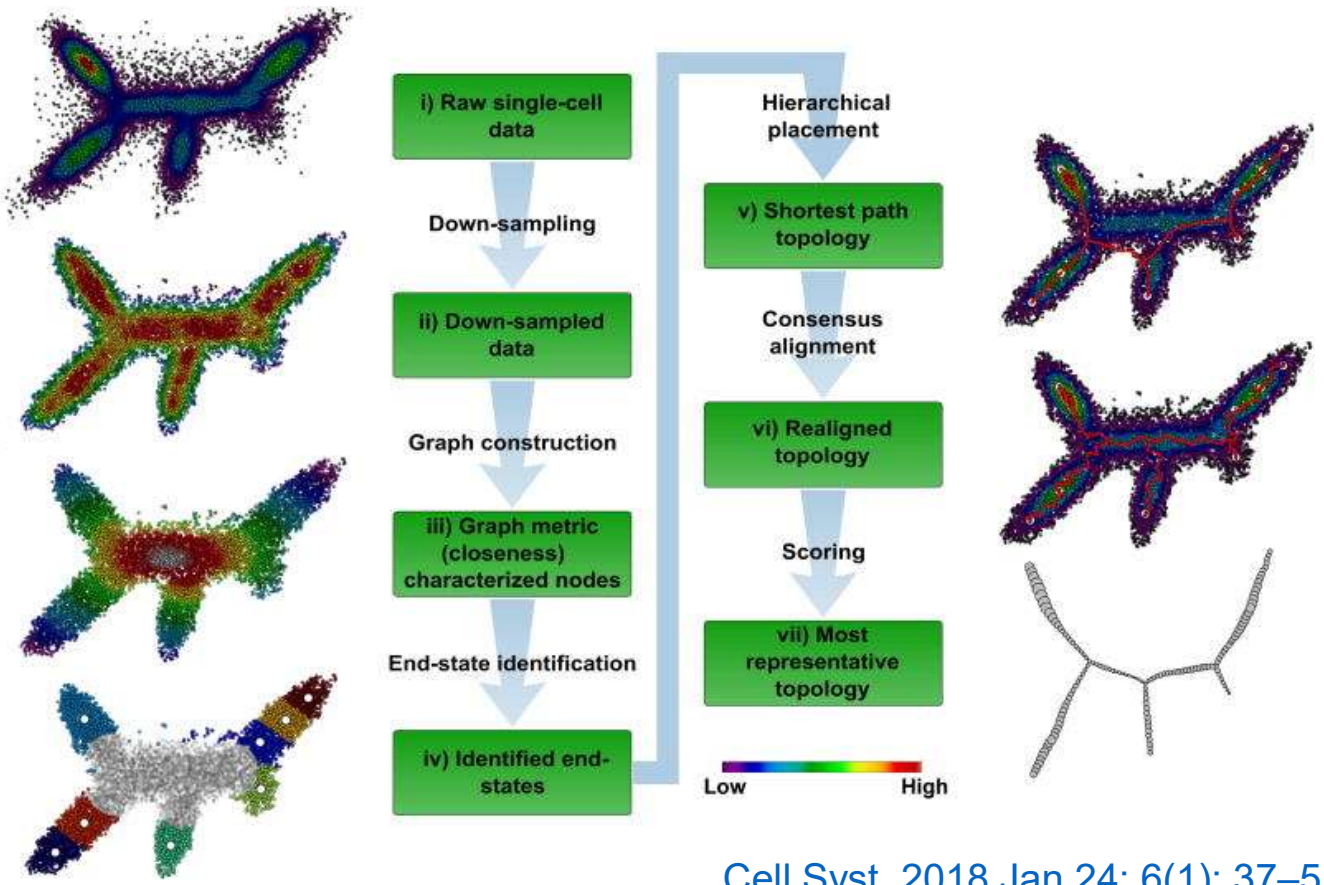


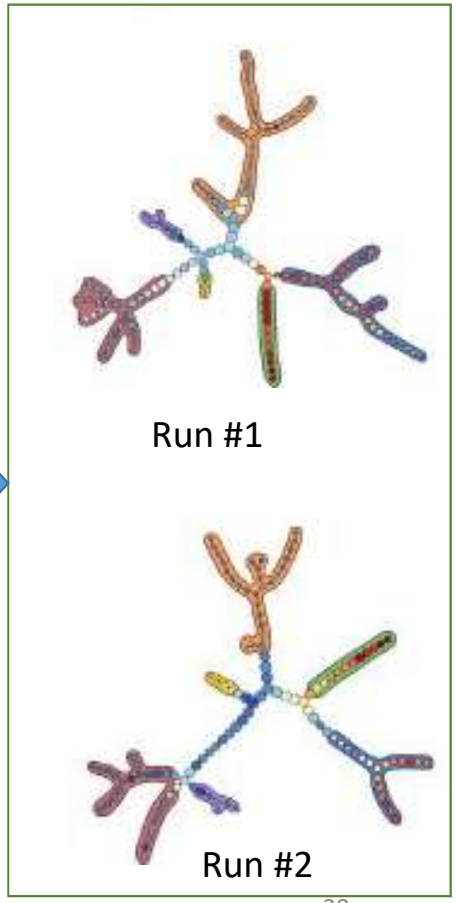
Features

- Combine stochastic down-sampling from SPADE with faithful down-sampling from SamSPECTRAL (BMC Bioinformatics. 2010;11:403)
- Use deterministic k-means clustering
- Semi-automated graph partitioning

Cytometry A. 2017 Mar; 91(3): 281–289

p-Creode can build more stable branches than SPADE

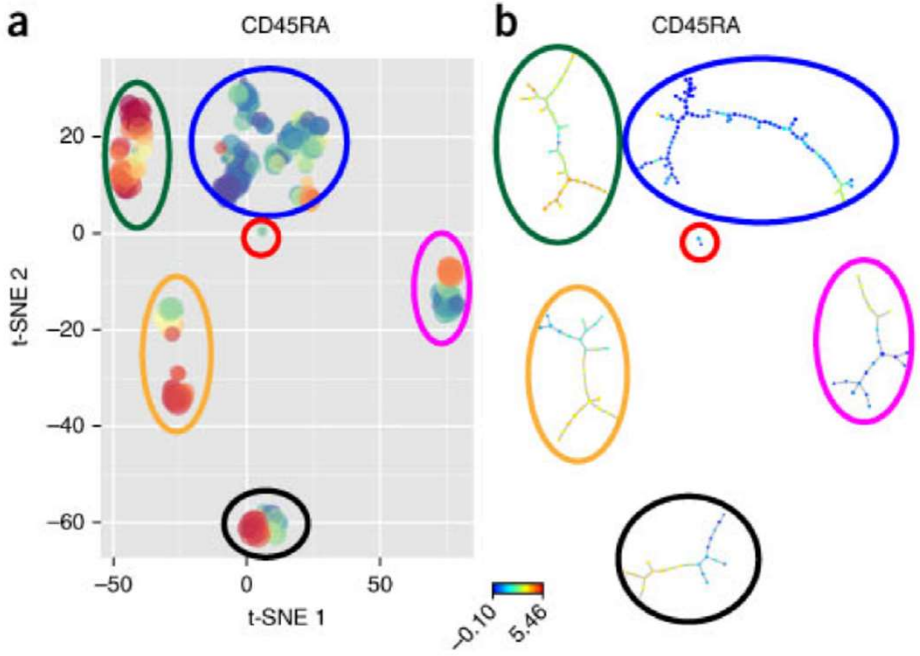




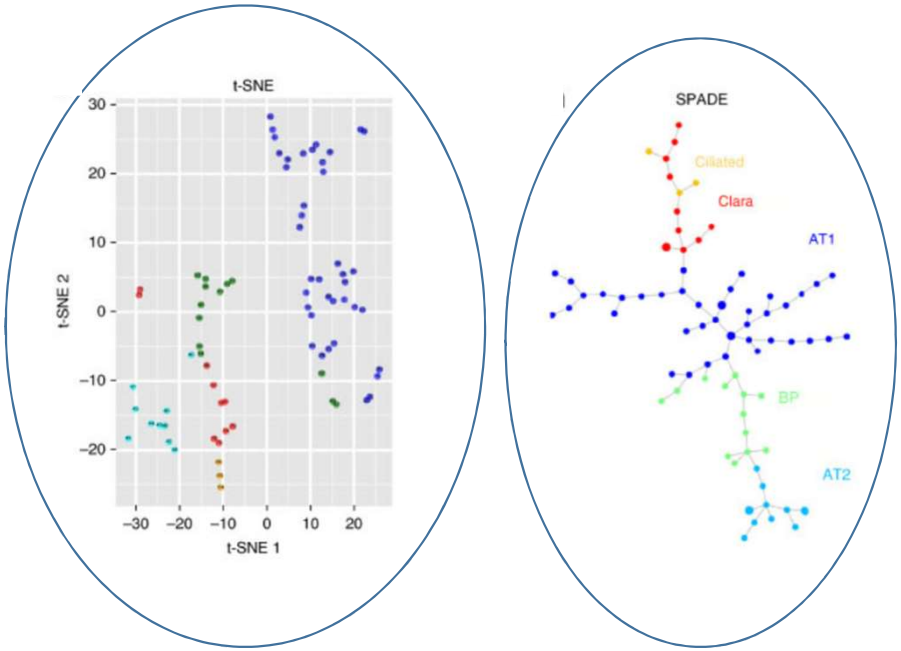
Cell Syst. 2018 Jan 24; 6(1): 37–51.e9.

Integrating dimensionality reduction, clustering, and trajectory analysis

Use t-SNE to obtain large clustered cell subpopulations



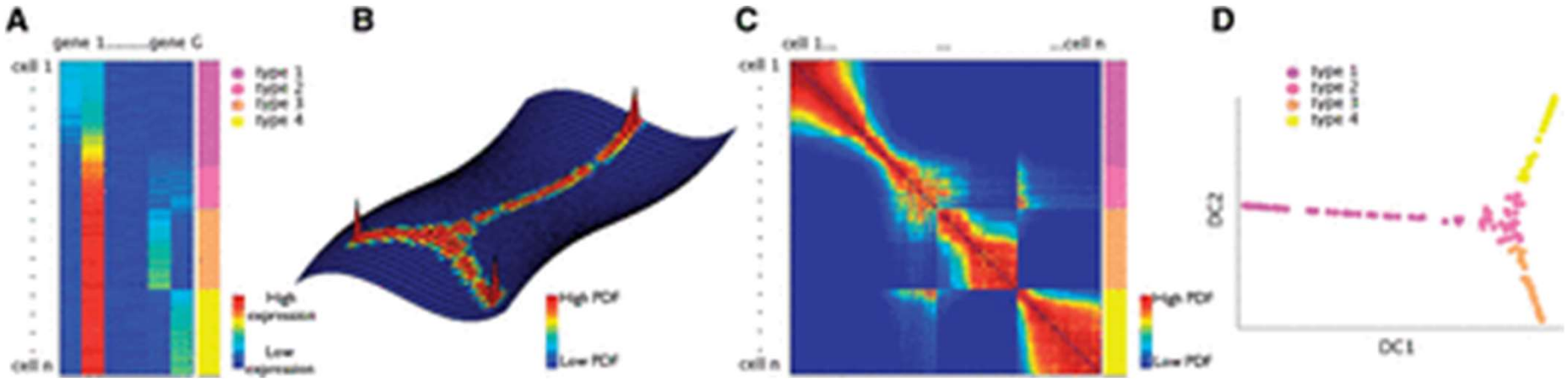
Use SPADE etc to differentiate cell lineage within a related subpopulation



Nat Protoc. 2016 Jul;11(7):1264-79.

Diffusion Maps:

Spectral clustering + global distance-based embedding



Gene vs cell matrix with cell type labels

Data paths as interfering Gaussians in the *d*-dimensional gene space

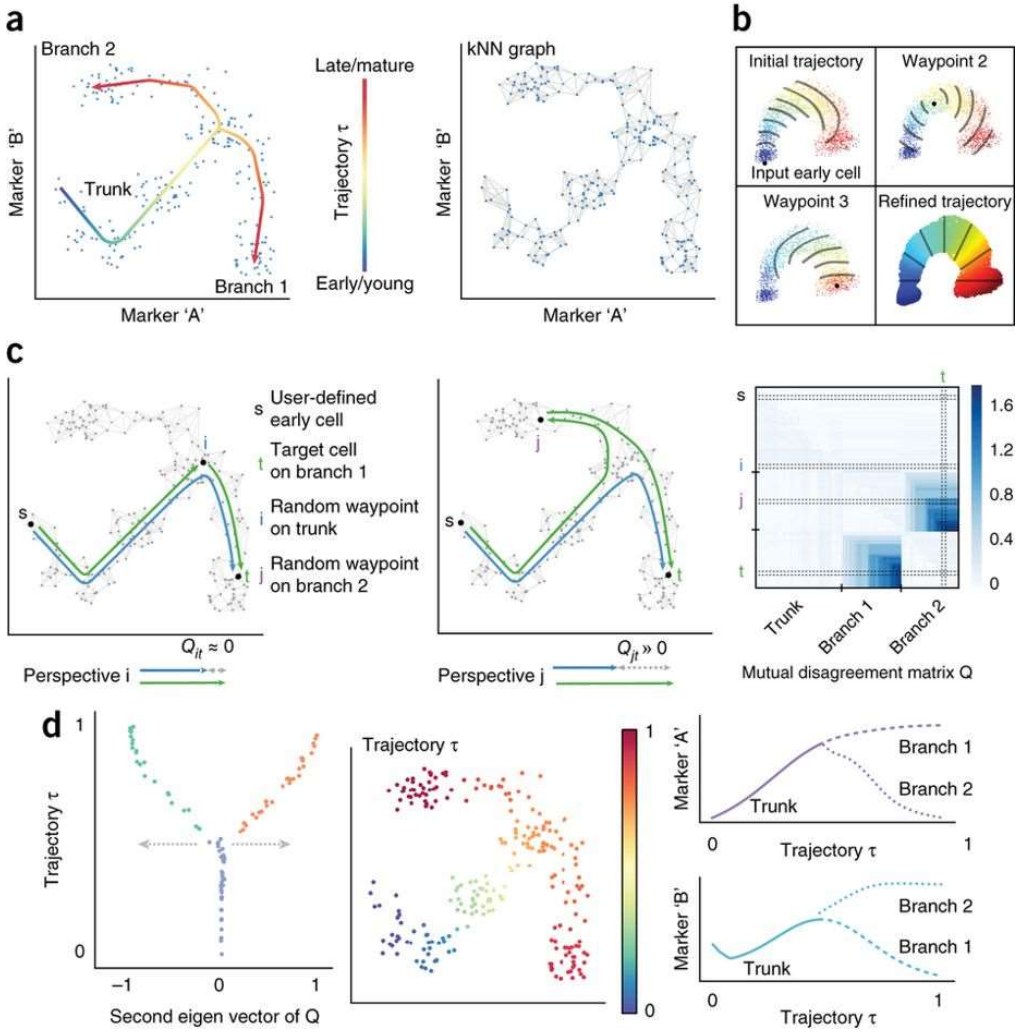
$n \times n$ Markovian transition probability matrix

The embedding on the two largest eigenvectors of the Markovian transition matrix (DC1 and DC2) which correspond to the largest DC of the data manifold.

Bioinformatics, Vol 31, No. 18 (2015) pp2989-2998

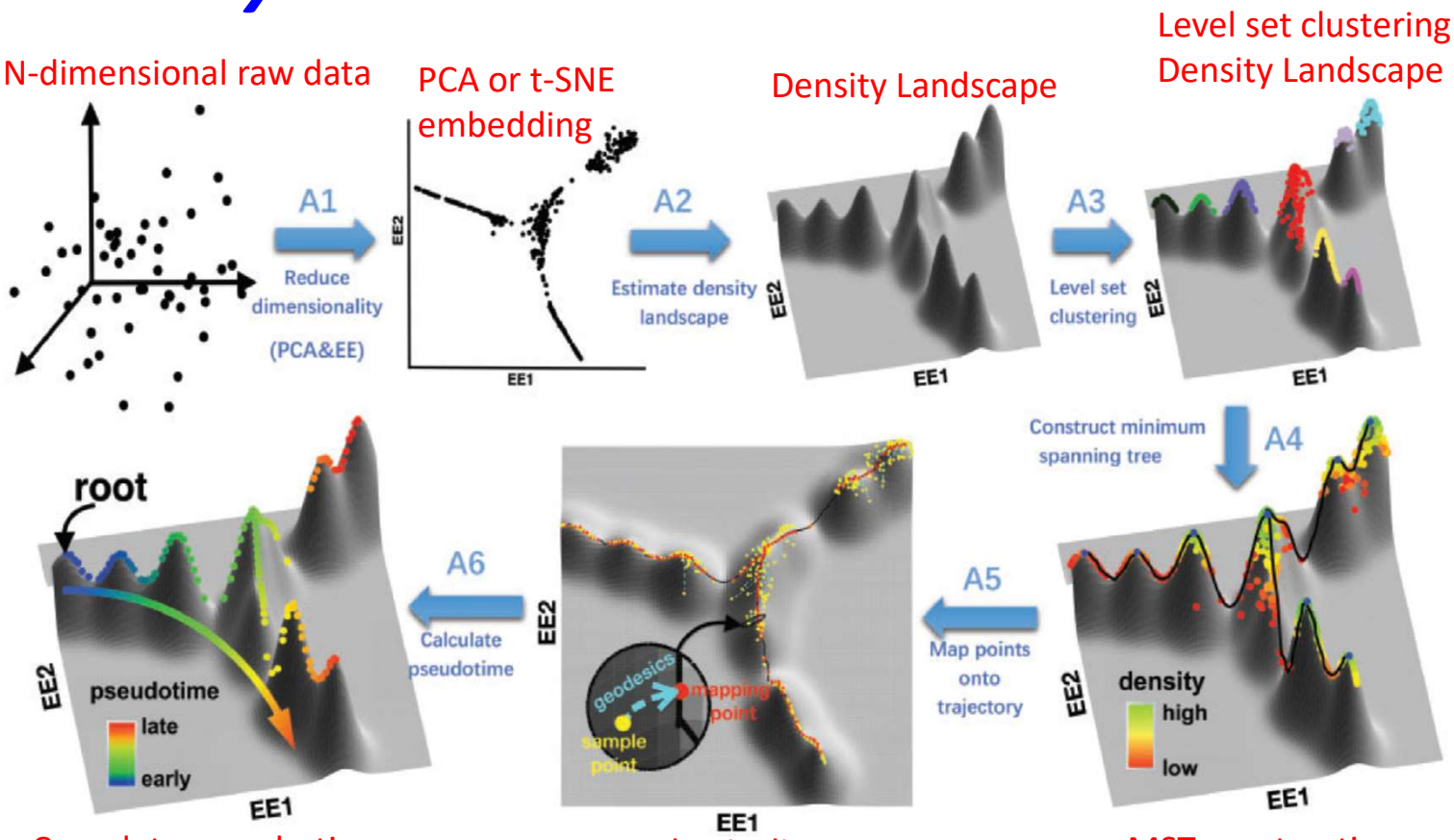
Wishbone can detect bifurcating trajectories

- kNN graph, guiding waypoints, and 2nd eigenvector of mutual disagreement matrix Q
- Waypoints arbitrary
- Only bifurcating events



Nature Biotechnology **vol.34** (2016) pp. 637–645

DensityPath



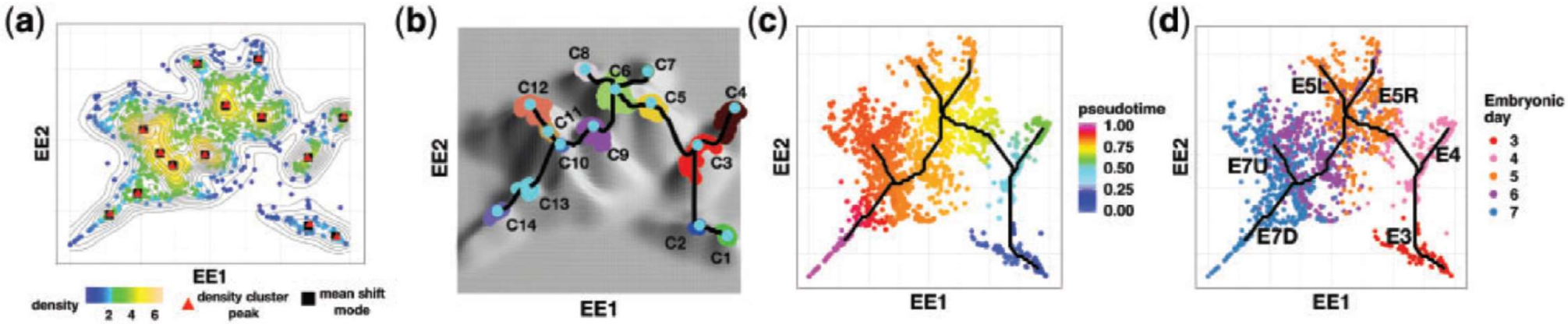
Complete pseudo-time estimation from root

Use geodestic distance to build trajectory

MST construction

Bioinformatics, 2019, 1–9 doi: 10.1093/bioinformatics/bty1009

DensityPath shows good multi-level clustering and pseudo-time analysis results



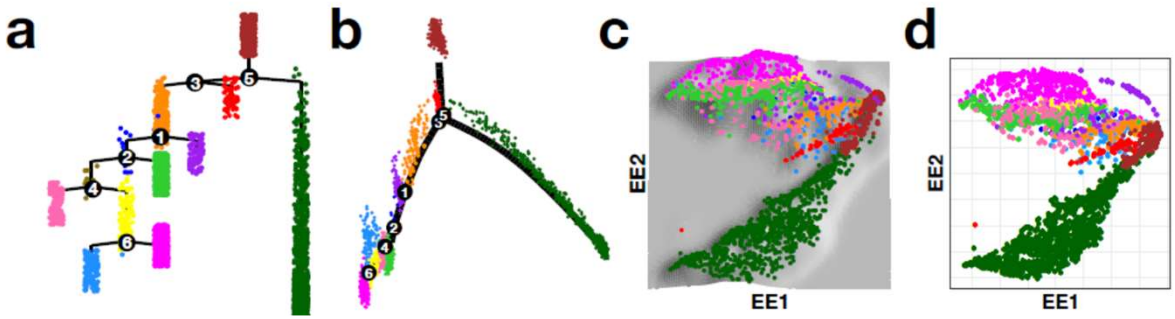
Density Landscape & LSC Results

Constructed trajectory showing bi-/tri-furcating events among 14 clusters

Pseudo-time labels

Real embryonic cells day 3-7 labelled

DensityPath reveals more refined multi-scale information

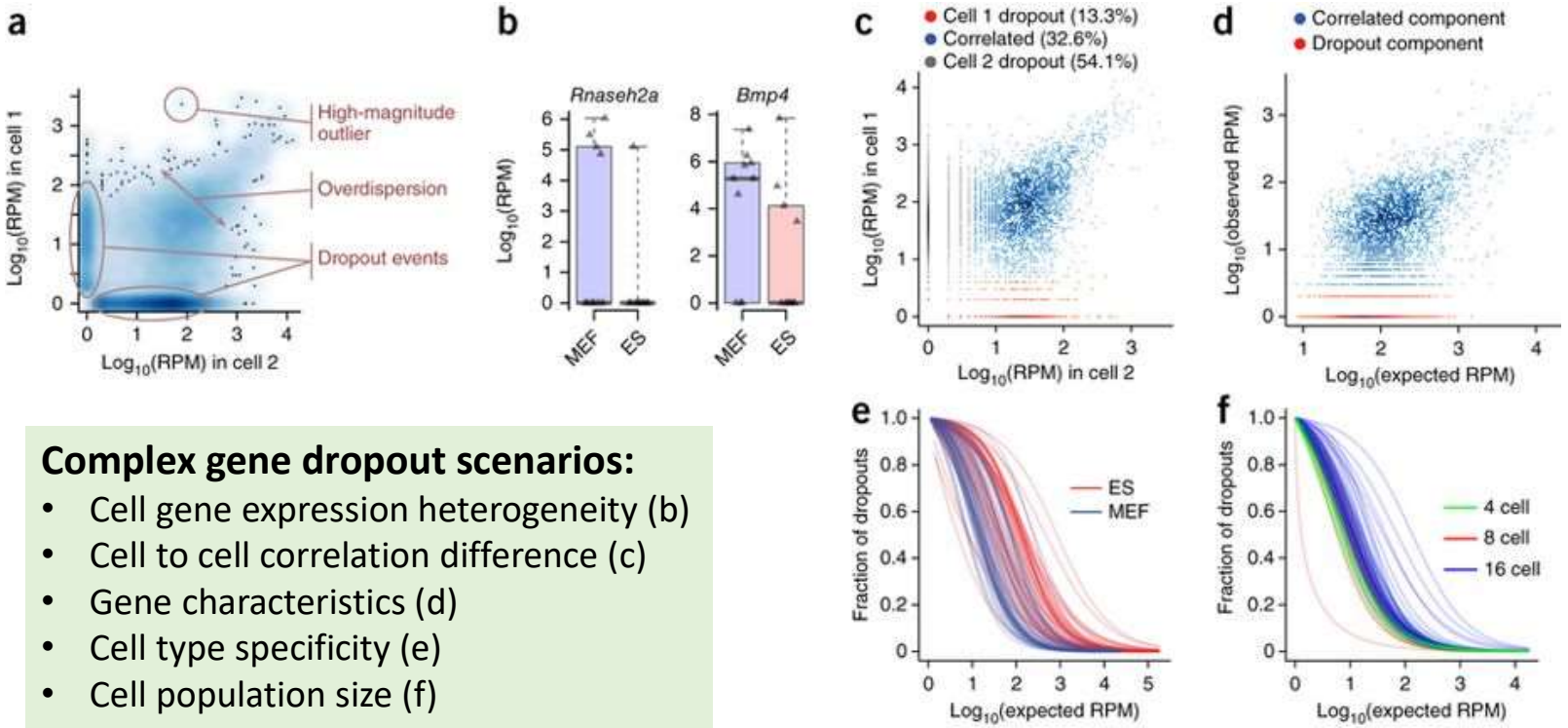


Monocle

Outline

- Dimensionality reduction techniques
- 2-D visualization of single-cell expression data
- Pseudo-time developmental trajectory analysis
- **Single-cell genomic analysis pipelines**

SCEA: strong in QC analysis

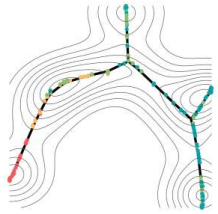


Complex gene dropout scenarios:

- Cell gene expression heterogeneity (b)
- Cell to cell correlation difference (c)
- Gene characteristics (d)
- Cell type specificity (e)
- Cell population size (f)

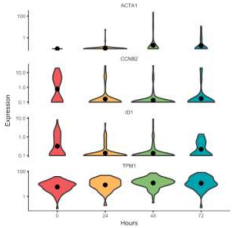
Nature Methods **Vol. 11**, pp. 740–742 (2014)

Monocle: among 1st for pseudo-time analysis

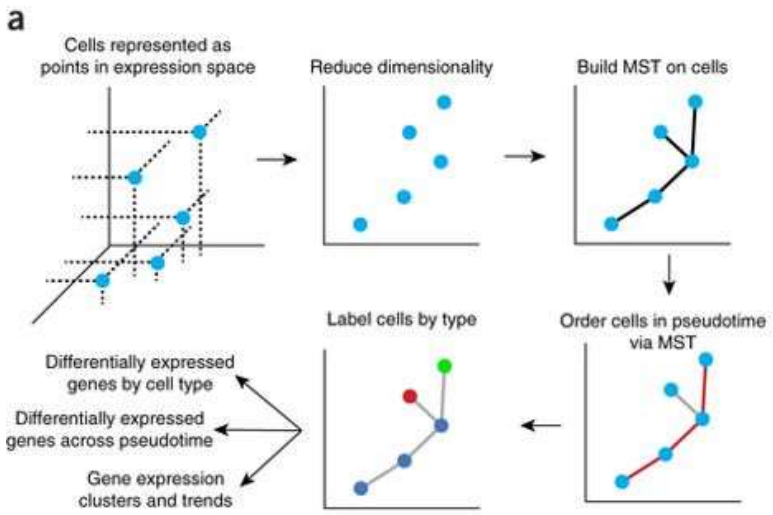


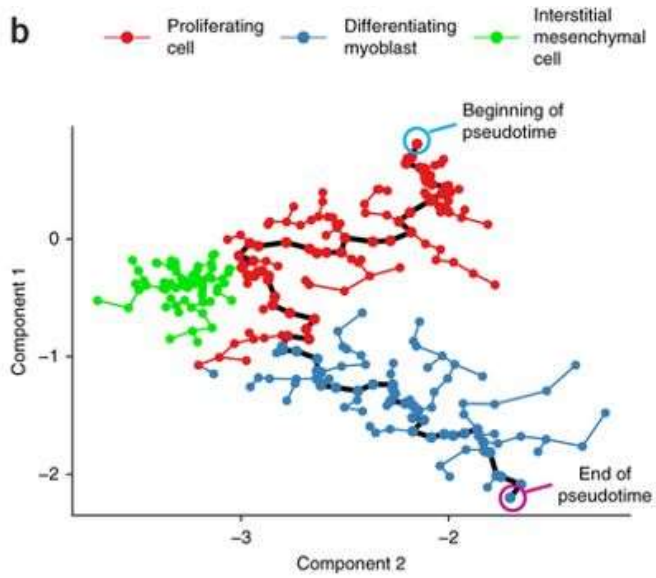
Pseudotime

Clustering



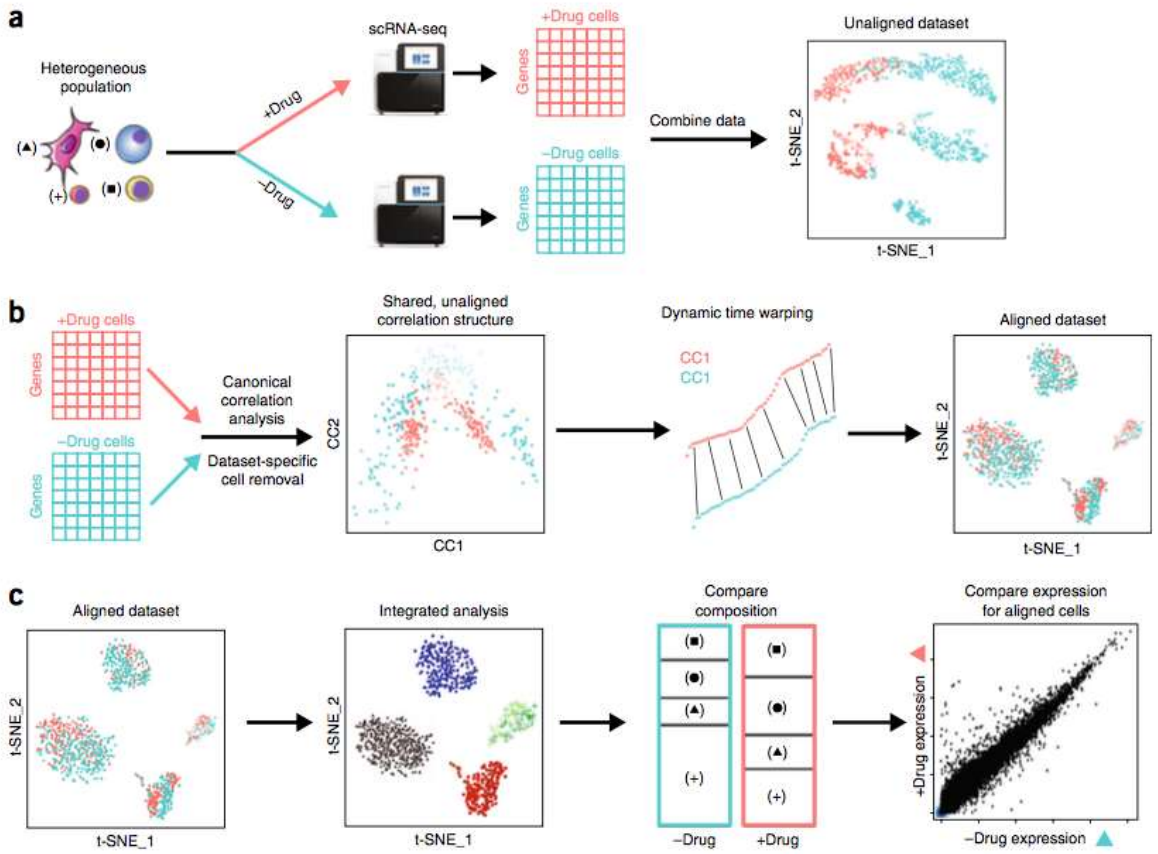
Differential expression





Nature Biotechnology **vol. 32**, pp381–386 (2014)

Seurat: R packages strong in multi-platform integrations



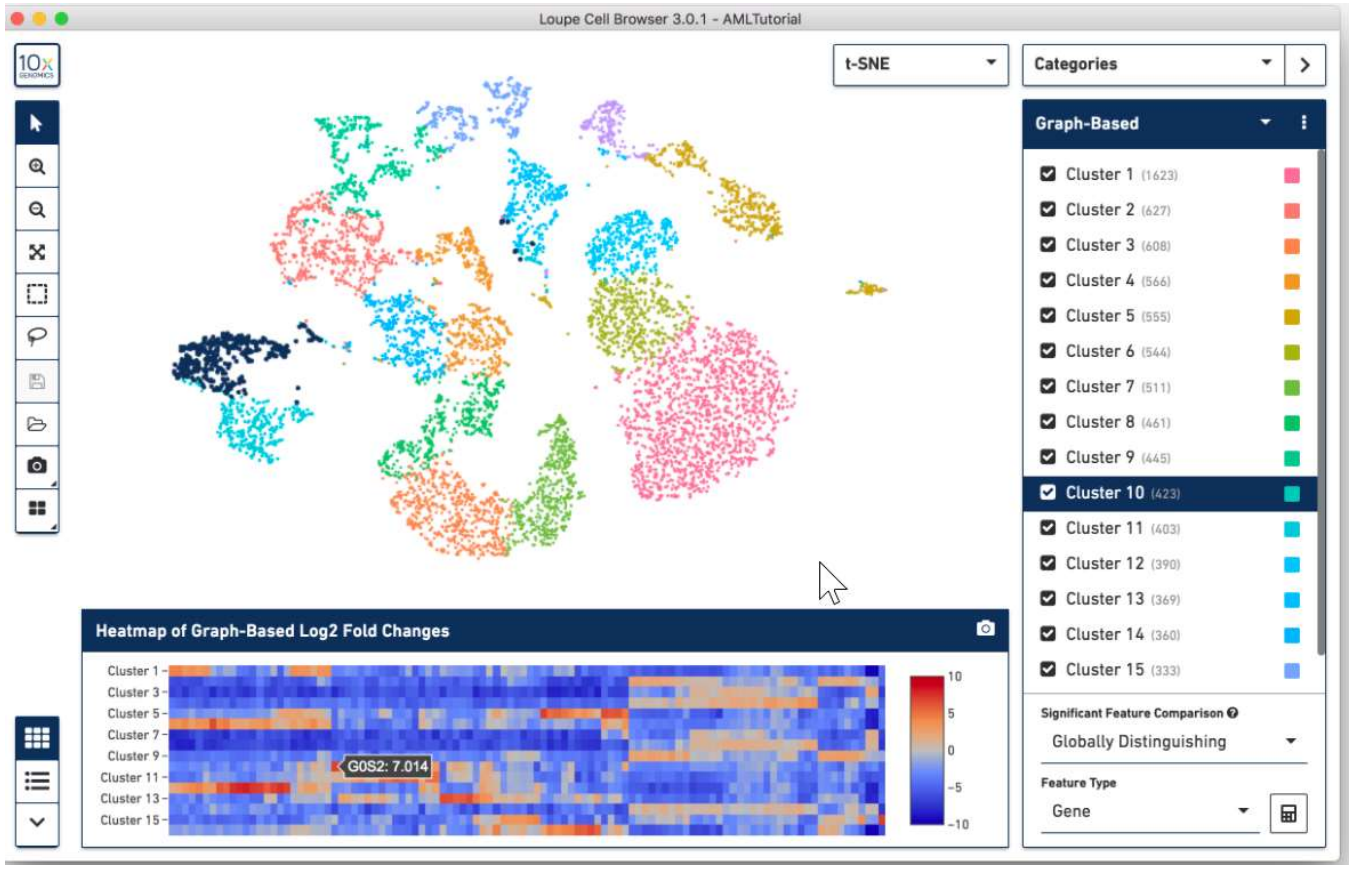
find linear combinations of the X's and linear combinations of the Y's that maximize the canonical correlation.

*i*th canonical pair (U_i,V_i)

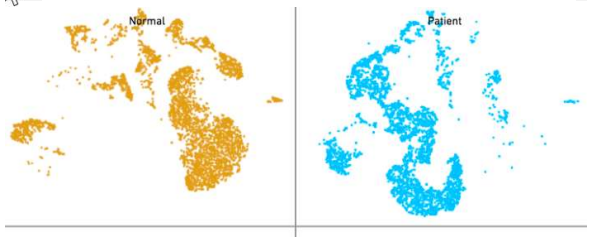
$$\rho_i^* = \frac{\mathrm{cov}(U_i, V_i)}{\sqrt{\mathrm{var}(U_i)\mathrm{var}(V_i)}}$$

Nature Biotechnology **vol. 36**, pp. 411–420 (2018)

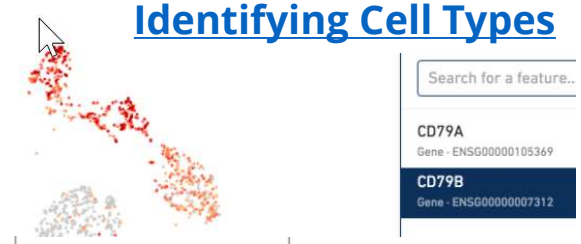
Loupe Cell Browser (commercial)



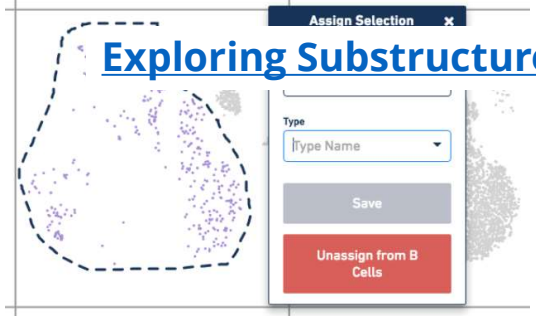
Finding Significant Genes



Identifying Cell Types

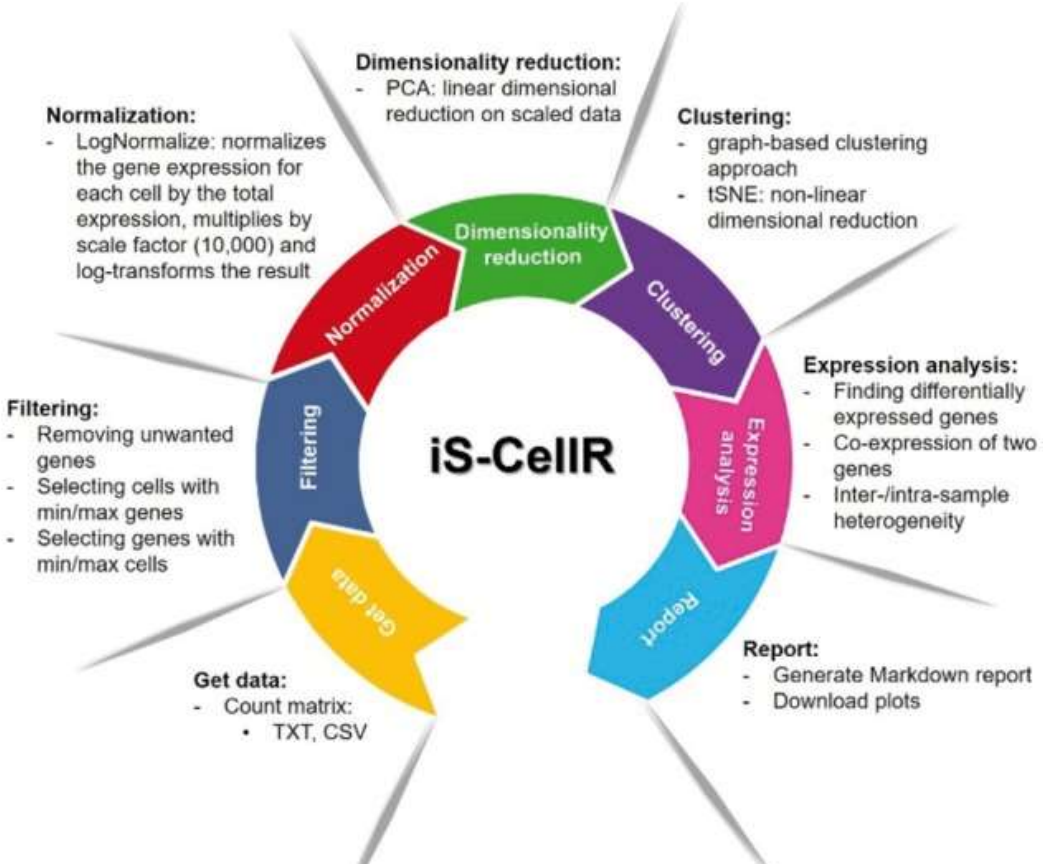


Exploring Substructure



Source: 10xgenomics.com

is-CellR: Web-based Open Source Dockerized Software



Bioinformatics. 2018 Dec 15; 34(24): 4305–4306.