6.874, 6.802, 20.390, 20.490, HST.506
Computational Systems Biology
Deep Learning in the Life Sciences

Lecture 11: Dimensionality Reduction

Prof. Manolis Kellis
Guest lecture: Joshua Welch

Massachusetts Institute of Technology

http://mit6874.github.io

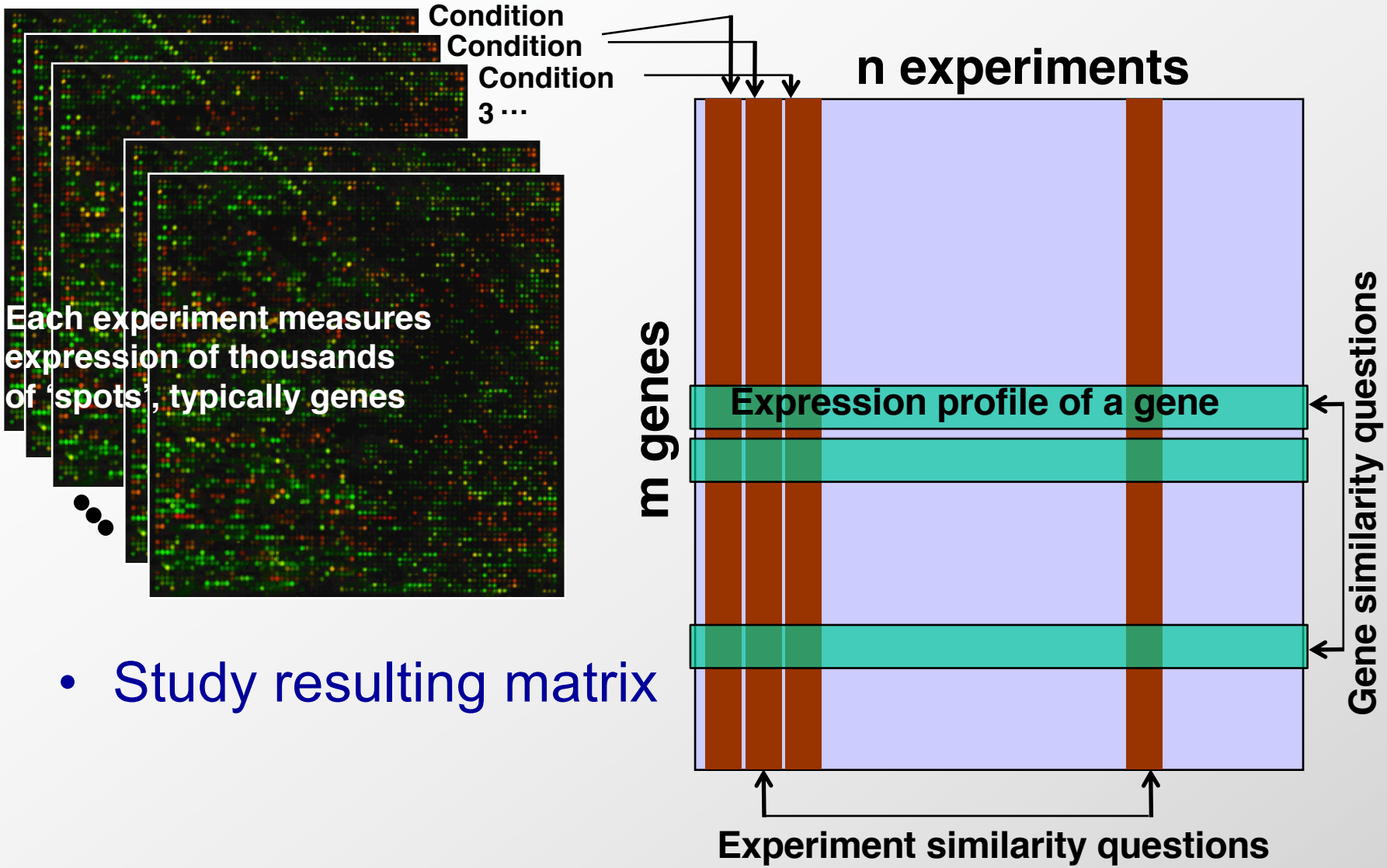
Slides credit: Soheil Feizi, Geoff Hinton

Today: Dimensionality Reduction for scRNA-seq

- Supervised (Classification) vs. unsupervised (Clustering)
- Supervised: Differential expression analysis
- Unsupervised: Embedding into lower dimensional space
- Linear reduction of dimensionality
  - Principle Component Analysis
  - Singular Value Decomposition
- Non-linear dimensionality reduction: embeddings
  - t-distributed Stochastic Network Embedding (t-SNE)
  - Building intuition: Playing with t-SNE parameters
- Deep Learning embeddings
  - Autoencoders

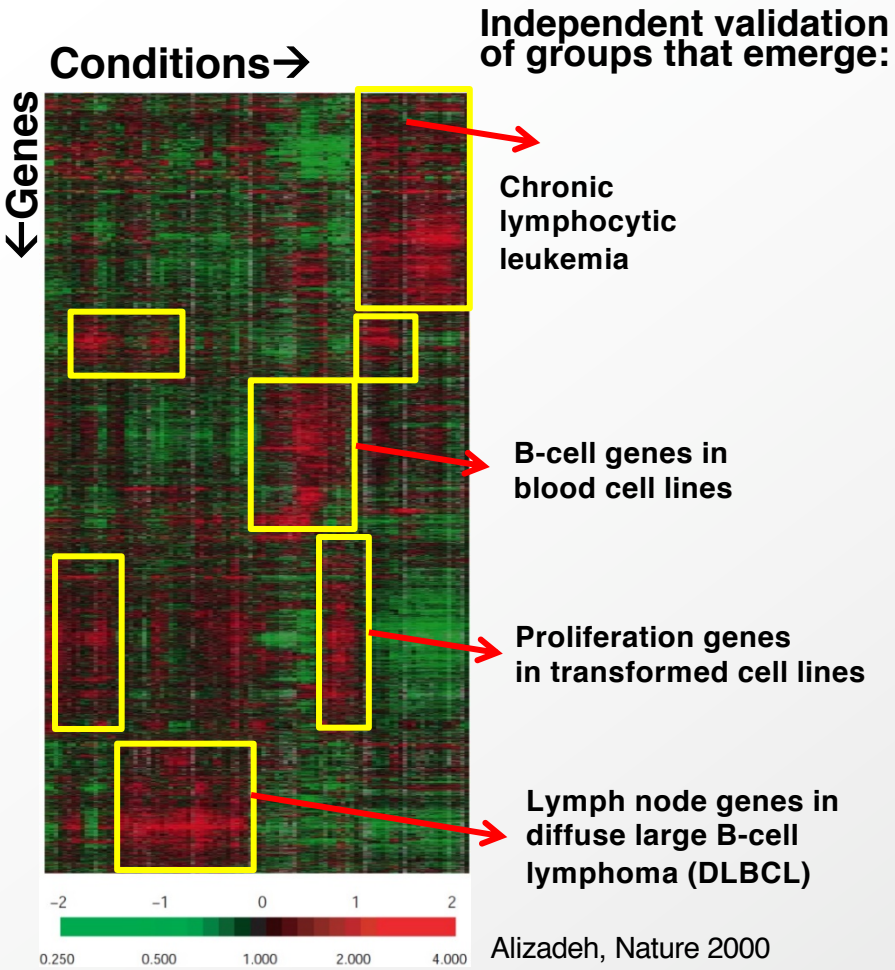
Expression Analysis Data Matrix

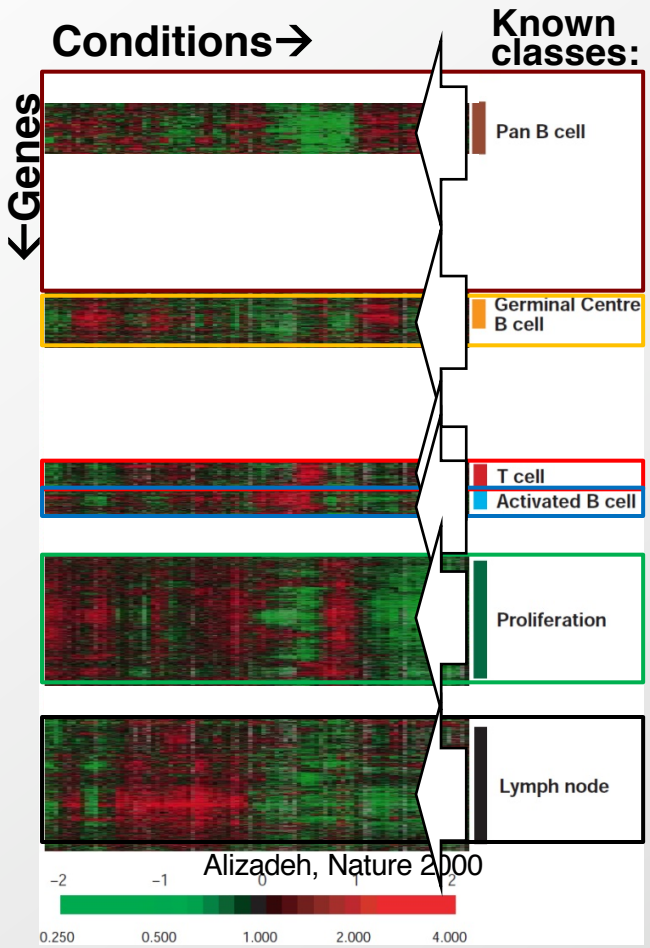
- Measure 20,000 genes in 100s of conditions



- Study resulting matrix

Clustering vs. Classification





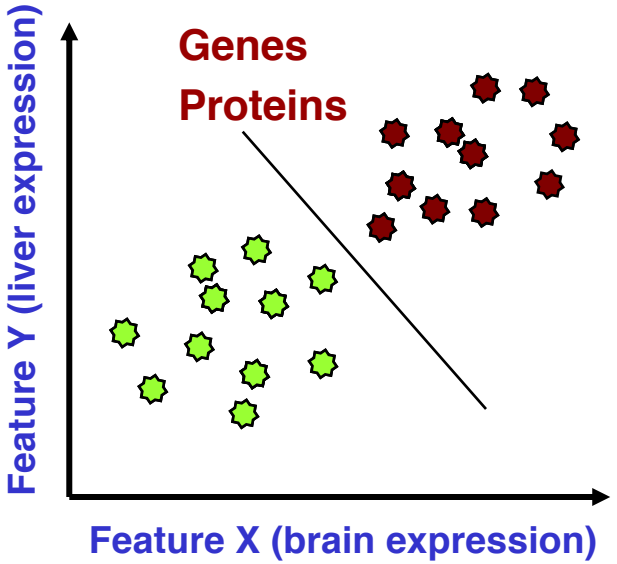
Goal of Clustering: **Group similar items** that likely come from the same category, and in doing so **reveal hidden structure**

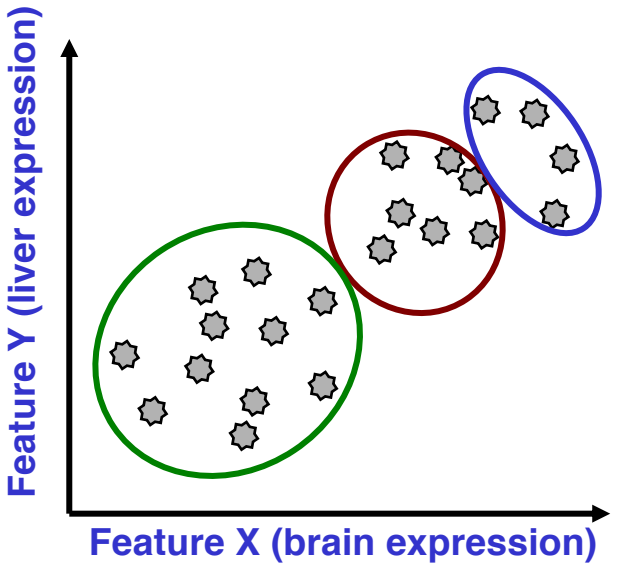
Goal of Classification: Extract features from the data that best **assign new elements** to ≥1 of **well-defined classes**

- **Unsupervised learning**
- **Supervised learning**

Clustering vs Classification

- Objects characterized by one or more features
- **Classification (supervised learning)**
  - Have labels for some points
  - Want a "rule" that will accurately assign labels to new points
  - Sub-problem: Feature selection
  - Metric: Classification accuracy
- **Clustering (unsupervised learning)**
  - No labels
  - Group points into clusters based on how "near" they are to one another
  - Identify structure in data
  - Metric: independent validation features

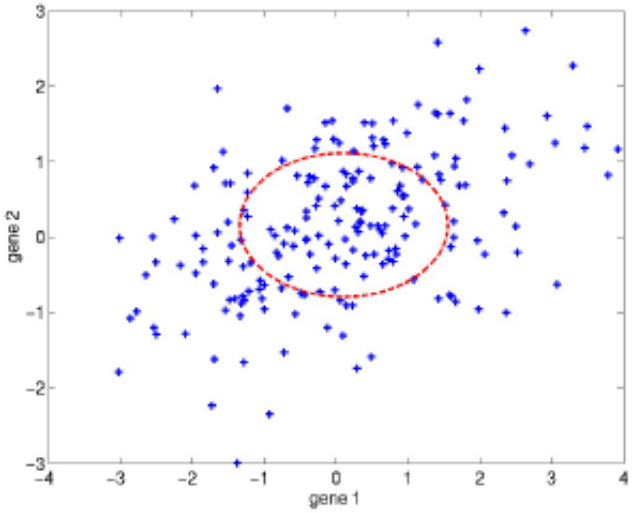




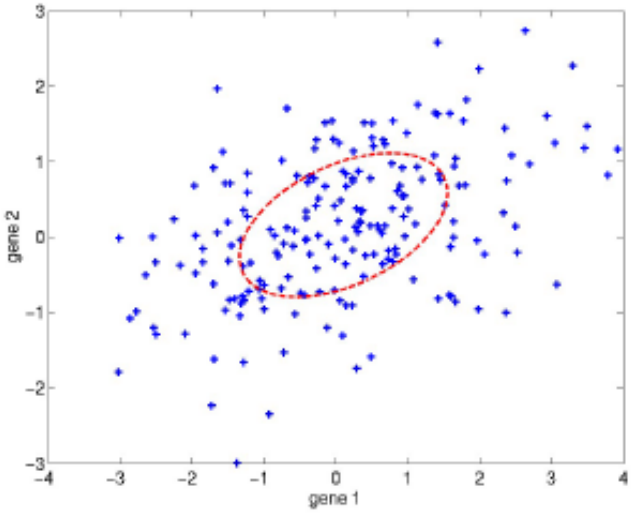
Supervised learning: differential gene expression

Statistical tests: example

- The alternative hypothesis H_1 is more expressive in terms of explaining the observed data



null hypothesis



alternative hypothesis

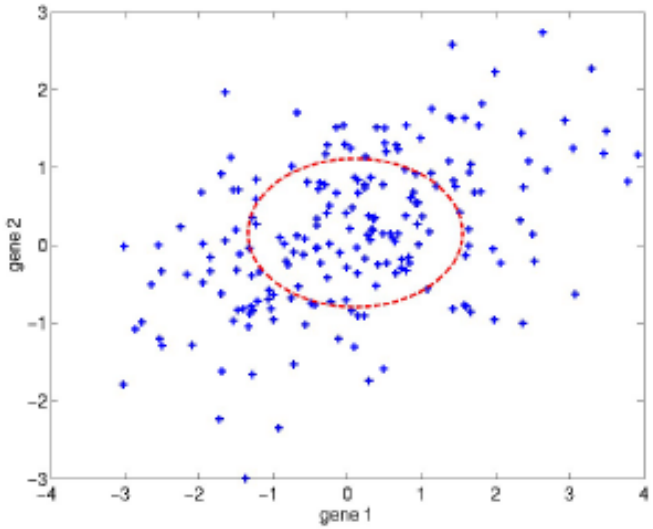
- We need to find a way of testing whether this difference is significant

Degrees of freedom

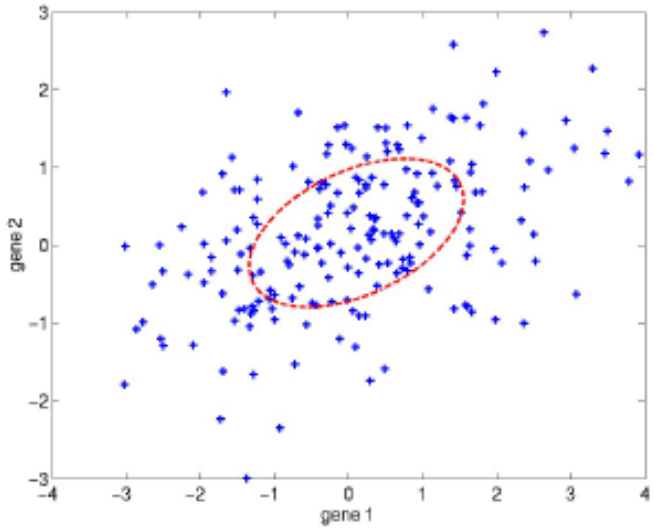
- How many degrees of freedom do we have in the two models?

$$H_0: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right)$$

$$H_1: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$



H_0



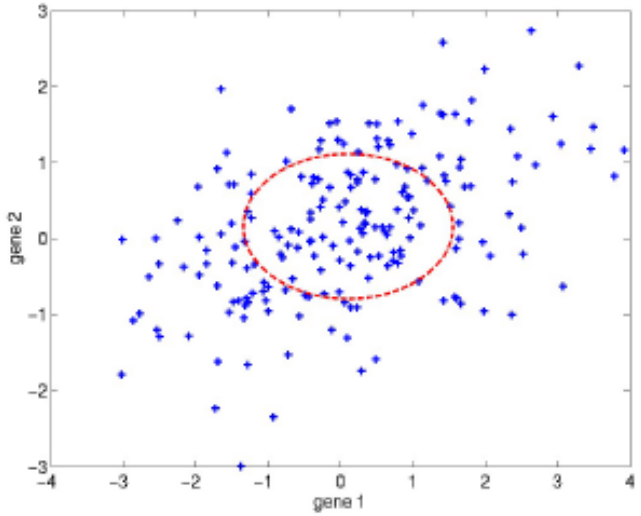
H_1

Degrees of freedom

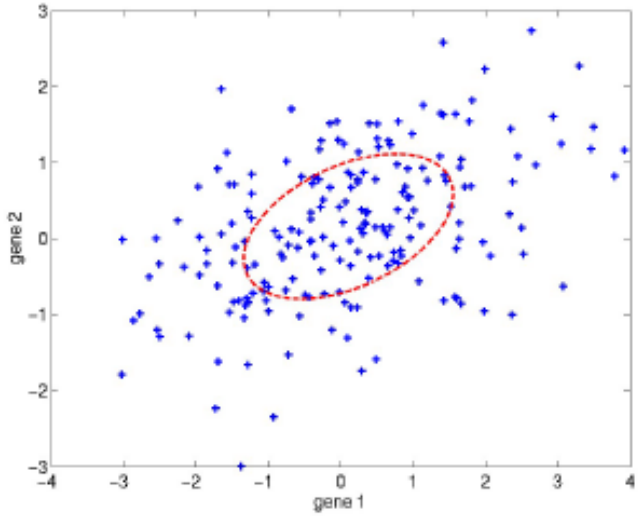
- How many degrees of freedom do we have in the two models?

$$H_0: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right)$$

$$H_1: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$



H_0



H_1

- The observed data overwhelmingly supports H_1

Test statistic

- Likelihood ratio statistic

$$T(X^{(1)}, \ldots, X^{(n)}) = 2 \log \frac{P(X^{(1)}, \ldots, X^{(n)} | \hat{H}_1)}{P(X^{(1)}, \ldots, X^{(n)} | \hat{H}_0)} \qquad (1)$$

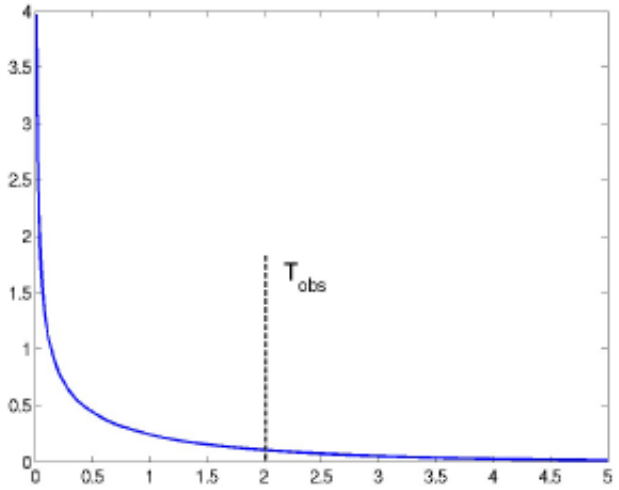
  Larger values of T imply that the model corresponding to the null hypothesis H_0 is much less able to account for the observed data

- To evaluate the P-value, we also need to know the sampling distribution for the test statistic

  In other words, we need to know how the test statistic $T(X^{(1)}, \ldots, X^{(n)})$ varies if the null hypothesis H_0 is correct

Test statistic cont'd

- For the likelihood ratio statistic, the sampling distribution is χ^2 with degrees of freedom equal to the difference in the number of free parameters in the two hypotheses



- Once we know the sampling distribution, we can compute the P-value

$$p = Prob(\, T(X^{(1)}, \ldots, X^{(n)}) \geq T_{obs} \,|\, H_0 \,) \qquad (2)$$

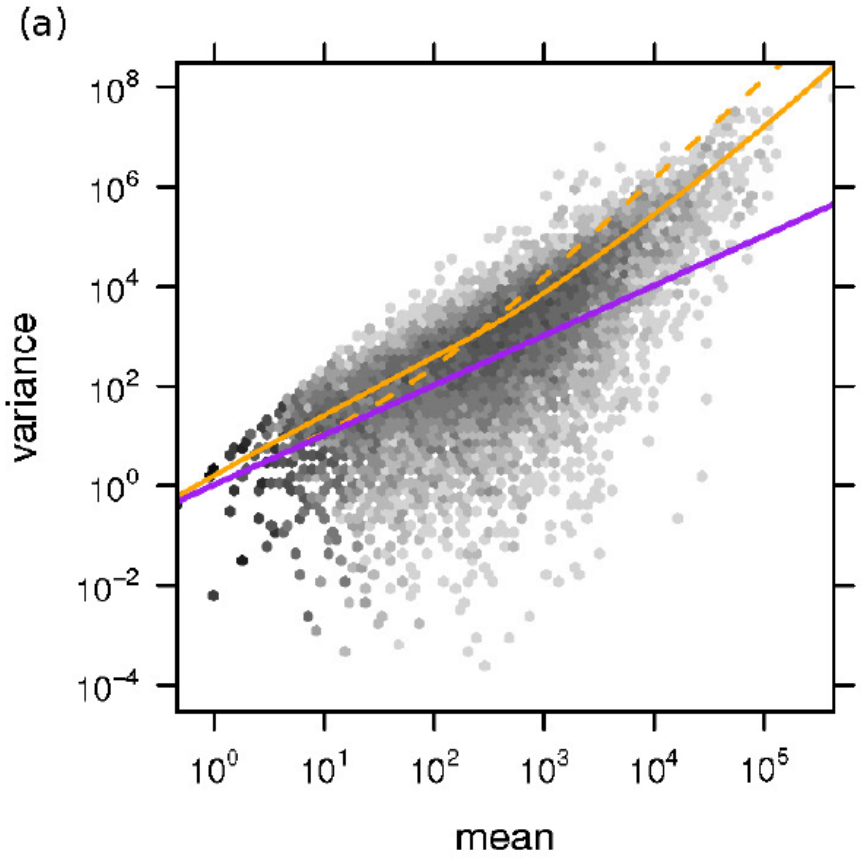
What is the right distribution for modeling read counts?

$$\lambda = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$f(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Poission?

Read count data is overdispersed for a Poission
Use a Negative Binomial instead



Orange Line – DESeq
Dashed Orange – edgeR
Purple - Poission

$$\sigma_{ij}^2 = \mu_{ij} + s_j^2 v_p\left(q_{ip(j)}\right)$$

Anders and Huber *Genome Biology* 2010, **11**:R106
http://genomebiology.com/2010/11/10/R106

A Negative Binomial distribution is better (DESeq)

- i gene or isoform p condition
- j sample (experiment) p(j) condition of sample j
- m number of samples
- K_{ij} number of counts for isoform i in experiment j
- q_{ip} Average scaled expression for gene i condition p

$$q_{ip} = \frac{1}{\#\text{ of replicates}} \sum_{\text{j in replicates}} \frac{K_{ij}}{s_j}$$

$$\mu_{ij} = q_{ip(j)} s_j \qquad \sigma_{ij}^2 = \mu_{ij} + s_j^2 v_p\left(q_{ip(j)}\right)$$

$$K_{ij} \sim NB\left(\mu_{ij}, \sigma_{ij}^2\right)$$

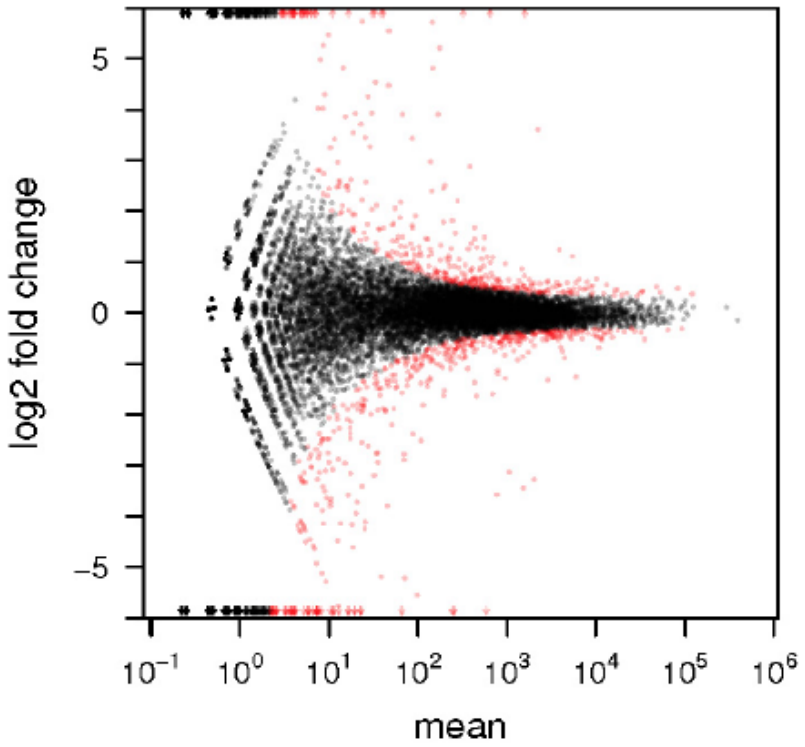


Figure 3 Testing for differential expression between conditions *A* and *B*: Scatter plot of $\log_2$ ratio (fold change) versus mean. The red colour marks genes detected as differentially expressed at 10% false discovery rate when Benjamini-Hochberg multiple testing adjustment is used. The symbols at the upper and lower plot border indicate genes with very large or infinite log fold change. The corresponding volcano plot is shown in Supplementary Figure S8 in Additional file 2.

Hypergeometric test for gene set overlap significance

N – total # of genes	1000
n1 - # of genes in set A	20
n2 - # of genes in set B	30
k - # of genes in both A and B	3

$$P(k) = \frac{\binom{n1}{k}\binom{N-n1}{n2-k}}{\binom{N}{n2}}$$

0.017

$$P(x \geq k) = \sum_{i=k}^{\min(n1,n2)} P(i)$$

0.020

Bonferroni correction

- Total number of rejections of null hypothesis over all N tests denoted by R.

$$Pr(R>0) \tilde{=} N\alpha$$

- Need to set α' = Pr(R>0) to required significance level **over all tests**. Referred to as the **experimentwise error rate**.
- With 100 tests, to achieve overall experimentwise significance level of α'=0.05:

$$0.05 = 100\alpha$$

$$\rightarrow \alpha = 0.0005$$

- **Pointwise** significance level of 0.05%.

Example - Genome wide association screens

- Risch & Merikangas (1996).
- 100,000 genes.
- Observe 10 SNPs in each gene.
- 1 million tests of null hypothesis of no association.
- To achieve experimentwise significance level of 5%, require pointwise p-value less than 5×10^{-8}

Bonferroni correction - problems

- Assumes each test of the null hypothesis to be **independent**.
- If not true, Bonferroni correction to significance level is **conservative**.
- Loss of power to reject null hypothesis.
- Example: genome-wide association screen across linked SNPs – correlation between tests due to LD between loci.

Benjamini Hochberg

- Select False Discovery Rate α
- Number of tests is *m*
- Sort p-values $P_{(k)}$ in ascending order (most significant first)
- Assumes tests are uncorrelated or positively correlated

1. For a given α, find the largest k such that $P_{(k)} \leq \frac{k}{m}\alpha$.
2. Reject the null hypothesis (i.e., declare discoveries) for all $H_{(i)}$ for $i = 1, \ldots, k$.

Unsupervised learning:
dimensionality reduction

Dimensionality reduction has multiple applications

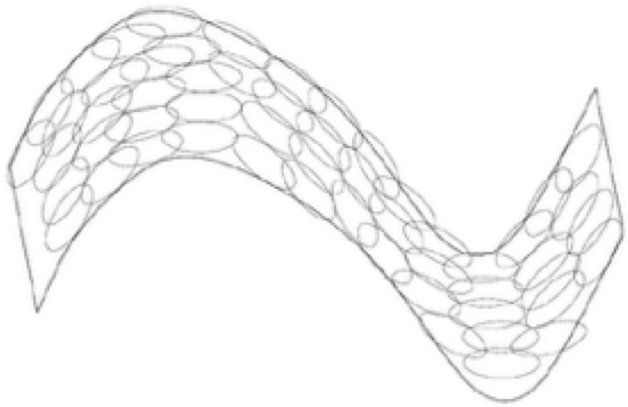
- Uses:
  - Data Visualization
  - Data Reduction
  - Data Classification
  - Trend Analysis
  - Factor Analysis
  - Noise Reduction

- Examples:
  - How many unique “sub-sets” are in the sample?
  - How are they similar / different?
  - What are the underlying factors that influence the samples?
  - Which time / temporal trends are (anti)correlated?
  - Which measurements are needed to differentiate?
  - How to best present what is “interesting”?
  - Which “sub-set” does this new sample rightfully belong?

A manifold is a topological space that locally resembles Euclidean space near each point

A manifold embedding is a structure preserving mapping of a high dimensional space into a manifold

Manifold learning learns a lower dimensional space that enables a manifold embedding

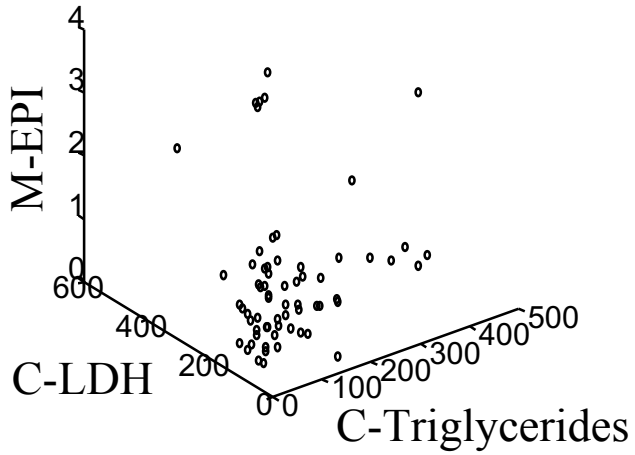


Principal Component Analysis

Example data

- Example: 53 Blood and urine measurements (wet chemistry) from 65 people (33 alcoholics, 32 non-alcoholics)

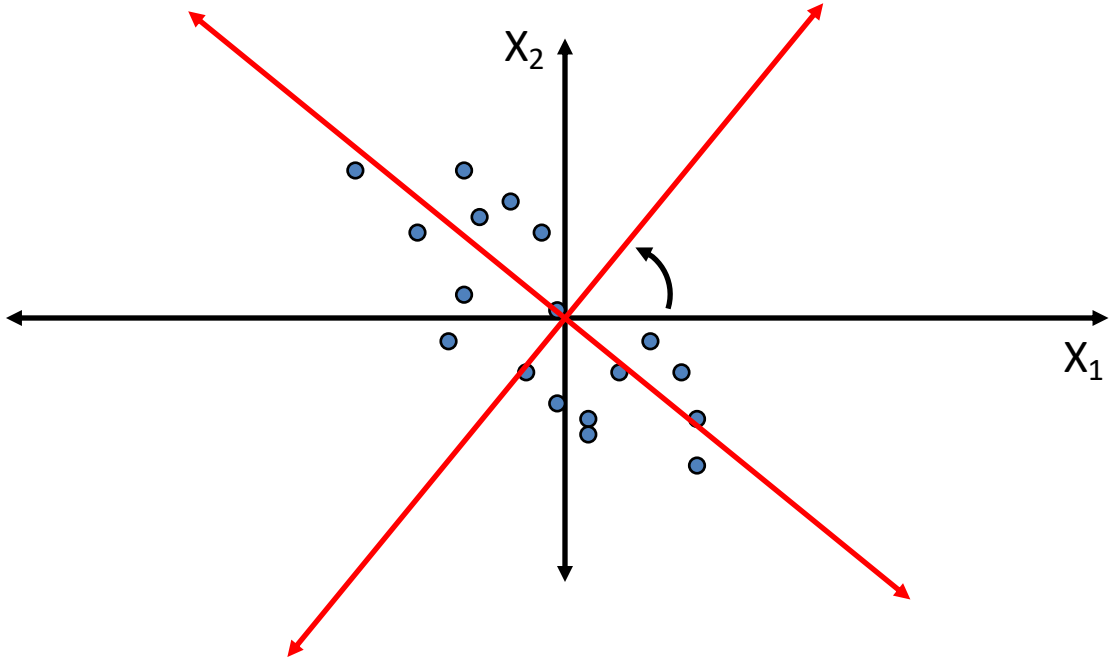
- Trivariate plot



	H-WBC	H-RBC	H-Hgb	H-Hct	H-MCV	H-MCH	H-MCHC
A1	8.0000	4.8200	14.1000	41.0000	85.0000	29.0000	34.0000
A2	7.3000	5.0200	14.7000	43.0000	86.0000	29.0000	34.0000
A3	4.3000	4.4800	14.1000	41.0000	91.0000	32.0000	35.0000
A4	7.5000	4.4700	14.9000	45.0000	101.0000	33.0000	33.0000
A5	7.3000	5.5200	15.4000	46.0000	84.0000	28.0000	33.0000
A6	6.9000	4.8600	16.0000	47.0000	97.0000	33.0000	34.0000
A7	7.8000	4.6800	14.7000	43.0000	92.0000	31.0000	34.0000
A8	8.6000	4.8200	15.8000	42.0000	88.0000	33.0000	37.0000
A9	5.1000	4.7100	14.0000	43.0000	92.0000	30.0000	32.0000

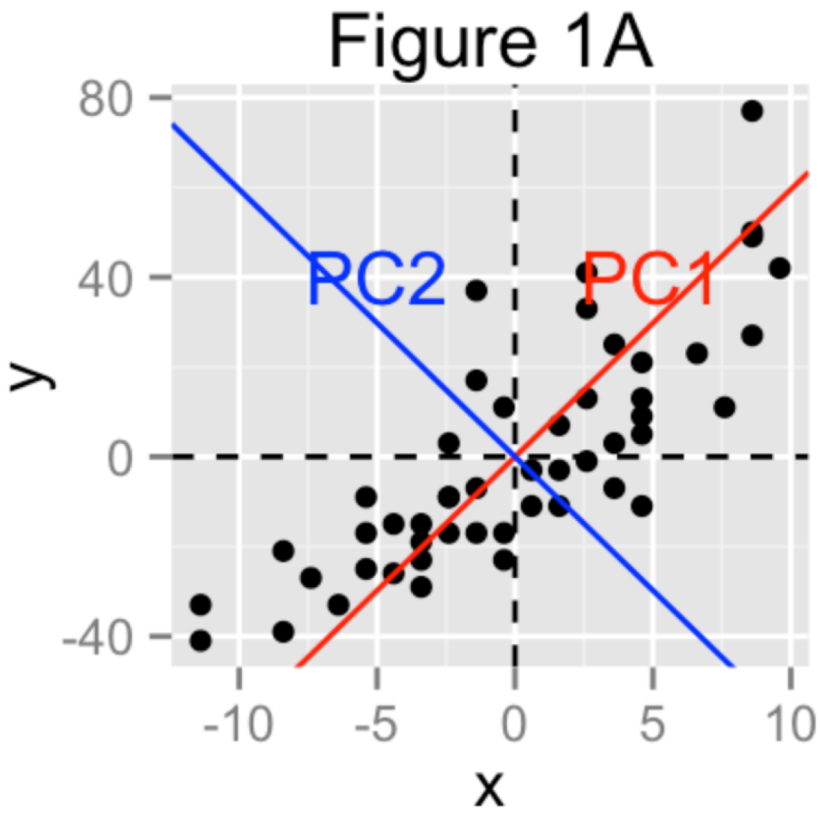
Principal Component = axis of greatest variability

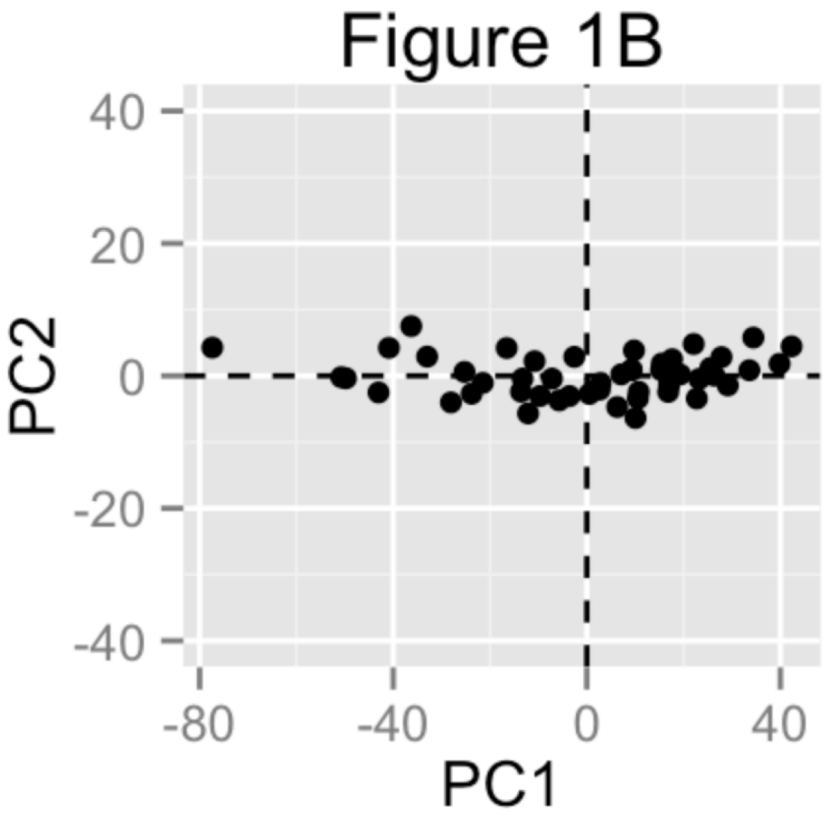
Suppose we have a population measured on p random variables $X_1,\ldots,X_p$. Note that these random variables represent the p-axes of the Cartesian coordinate system in which the population resides. Our goal is to develop a new set of p axes (linear combinations of the original p axes) in the directions of greatest variability:



This is accomplished by rotating the axes.

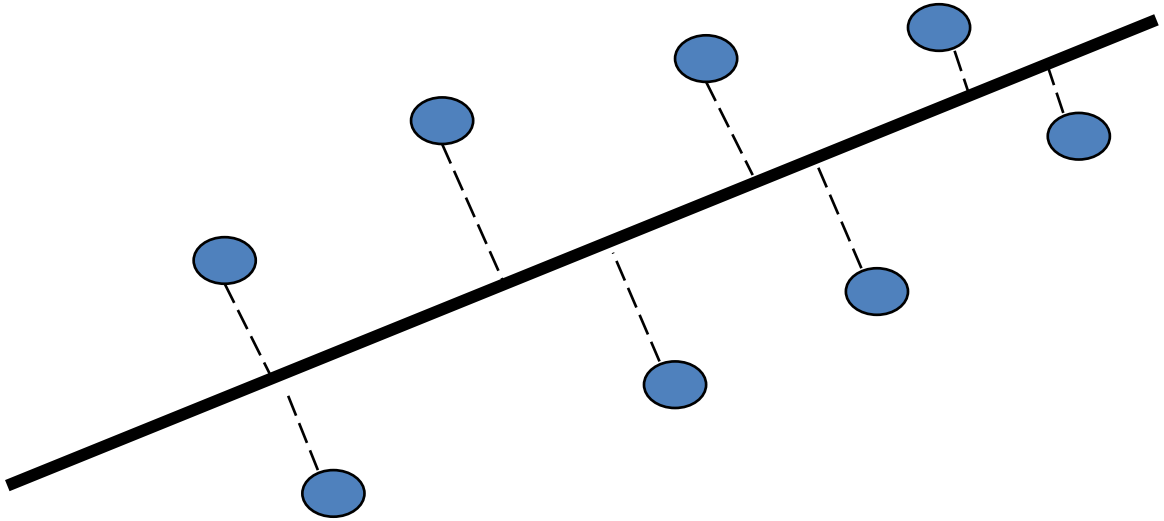
Data projected onto PC1





Selecting Principal Components

- Given m points in a n dimensional space, for large n, how does one project on to a 1 dimensional space?
- Formally, minimize sum of squares of distances to the line.



- Why sum of squares? Because it allows fast minimization, assuming the line passes through 0

Linear Algebra Review

- **Eigenvectors** (for a square $m \times m$ matrix $\mathbf{S}$)

$$\mathbf{S}\mathbf{v} = \lambda\mathbf{v}$$

(right) eigenvector $\mathbf{v} \in \mathbb{R}^m \neq \mathbf{0}$

eigenvalue $\lambda \in \mathbb{R}$

Example

$$\begin{pmatrix} 6 & -2 \\ 4 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

- How many eigenvalues are there at most?

$$\mathbf{S}\mathbf{v} = \lambda\mathbf{v} \iff (\mathbf{S} - \lambda\mathbf{I})\,\mathbf{v} = \mathbf{0}$$

only has a non-zero solution if $|\mathbf{S} - \lambda\mathbf{I}| = 0$

this is a m-th order equation in λ which can have **at most m distinct solutions** (roots of the characteristic polynomial) - can be complex even though **S** is real.

Eigenvalues & Eigenvectors

- For symmetric matrices, eigenvectors for distinct eigenvalues are **orthogonal**

$$Sv_{\{1,2\}} = \lambda_{\{1,2\}} v_{\{1,2\}}, \text{and } \lambda_1 \neq \lambda_2 \Rightarrow v_1 \bullet v_2 = 0$$

- All eigenvalues of a real symmetric matrix are **real**.

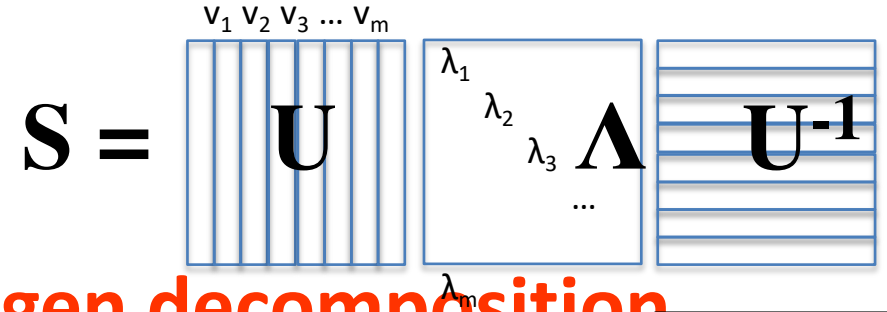
$$\text{for complex } \lambda, \text{if } |S - \lambda I| = 0 \text{ and } \mathrm{S} = \mathrm{S}^{\mathrm{T}} \Rightarrow \lambda \in \Re$$

- All eigenvalues of a positive semidefinite matrix are **non-negative**

$$\forall w \in \Re^n, w^T S w \geq 0, \text{then if } \mathrm{Sv} = \lambda \mathrm{v} \Rightarrow \lambda \geq 0$$

Eigen/diagonal Decomposition

- Let $\mathbf{S} \in \mathbb{R}^{m \times m}$ be a **square** matrix with ***m*** **linearly independent eigenvectors** (a "non-defective" matrix)



- **Theorem**: Exists an **eigen decomposition**

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$$

diagonal

Unique for distinct eigen-values

  - (cf. matrix diagonalization theorem)

- Columns of ***U*** are **eigenvectors** of ***S***
- Diagonal elements of $\mathbf{\Lambda}$ are **eigenvalues** of $\mathbf{S}$

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m), \quad \lambda_i \geq \lambda_{i+1}$$

Symmetric Eigen Decomposition

- If $S \in \mathbb{R}^{m \times m}$ is a **symmetric** matrix:
- **Theorem**: Exists a (unique) **eigen decomposition** $S = Q \Lambda Q^T$
- where ***Q*** is **orthogonal:**
  - $Q^{-1} = Q^T$
  - Columns of ***Q*** are normalized eigenvectors
  - Columns are orthogonal.
  - (everything is real)

Singular value decomposition (general m x n matrices)

Singular Value Decomposition

For an $m \times n$ matrix $\mathbf{A}$ of rank r there exists a factorization (Singular Value Decomposition = **SVD**) as follows:

$$A = U\Sigma V^T$$

$m \times m$ $\quad$ $m \times n$ $\quad$ V is $n \times n$

The columns of $\boldsymbol{U}$ are orthogonal eigenvectors of $\boldsymbol{AA^T}$.

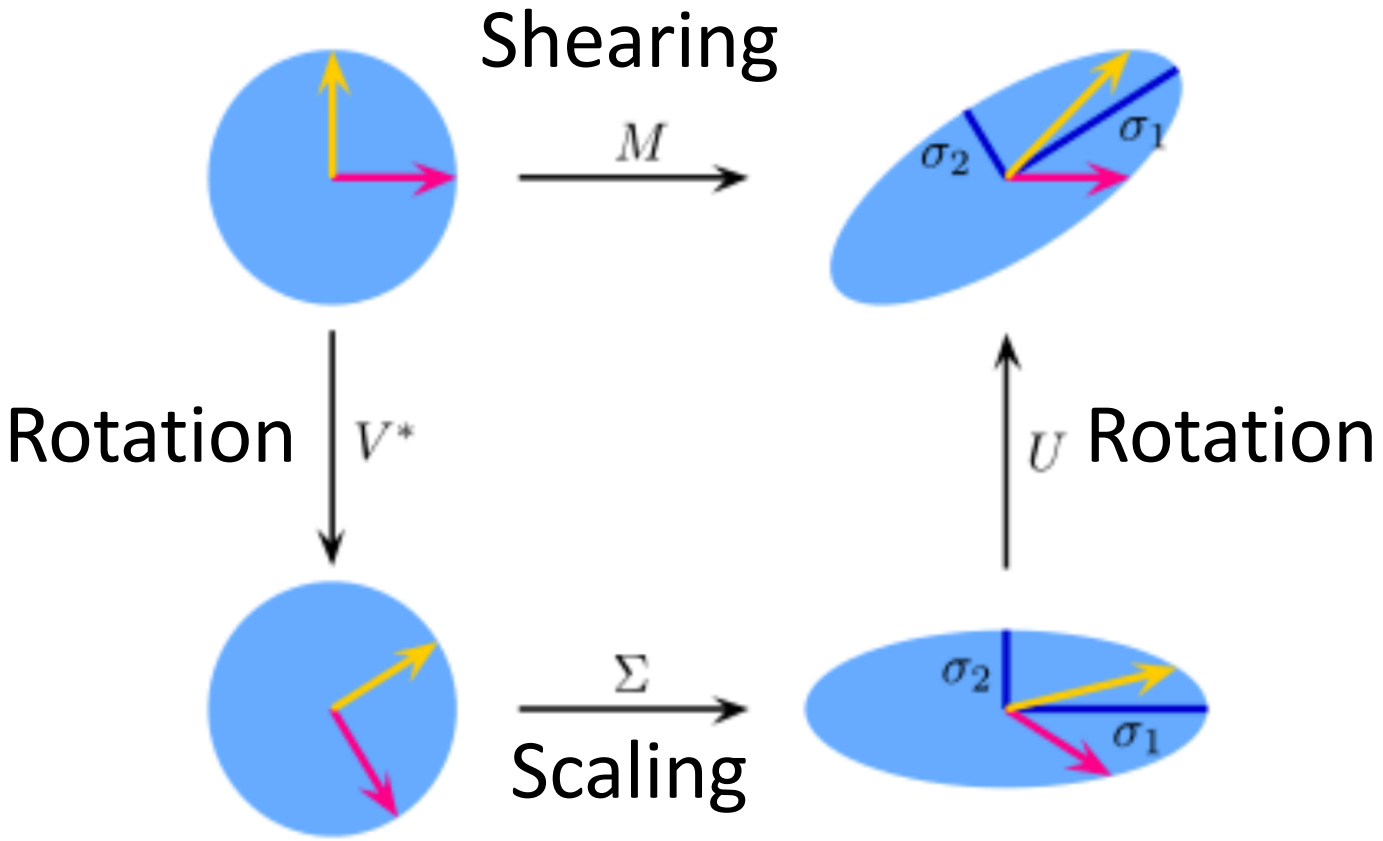
The columns of $\boldsymbol{V}$ are orthogonal eigenvectors of $\boldsymbol{A^TA}$.

Eigenvalues $\lambda_1 \ldots \lambda_r$ of $\boldsymbol{AA^T}$ are the eigenvalues of $\boldsymbol{A^TA}$.

$$\sigma_i = \sqrt{\lambda_i}$$

$$\Sigma = diag(\sigma_1 \ldots \sigma_r)$$ ← *Singular values.*

Geometric interpretation of SVD

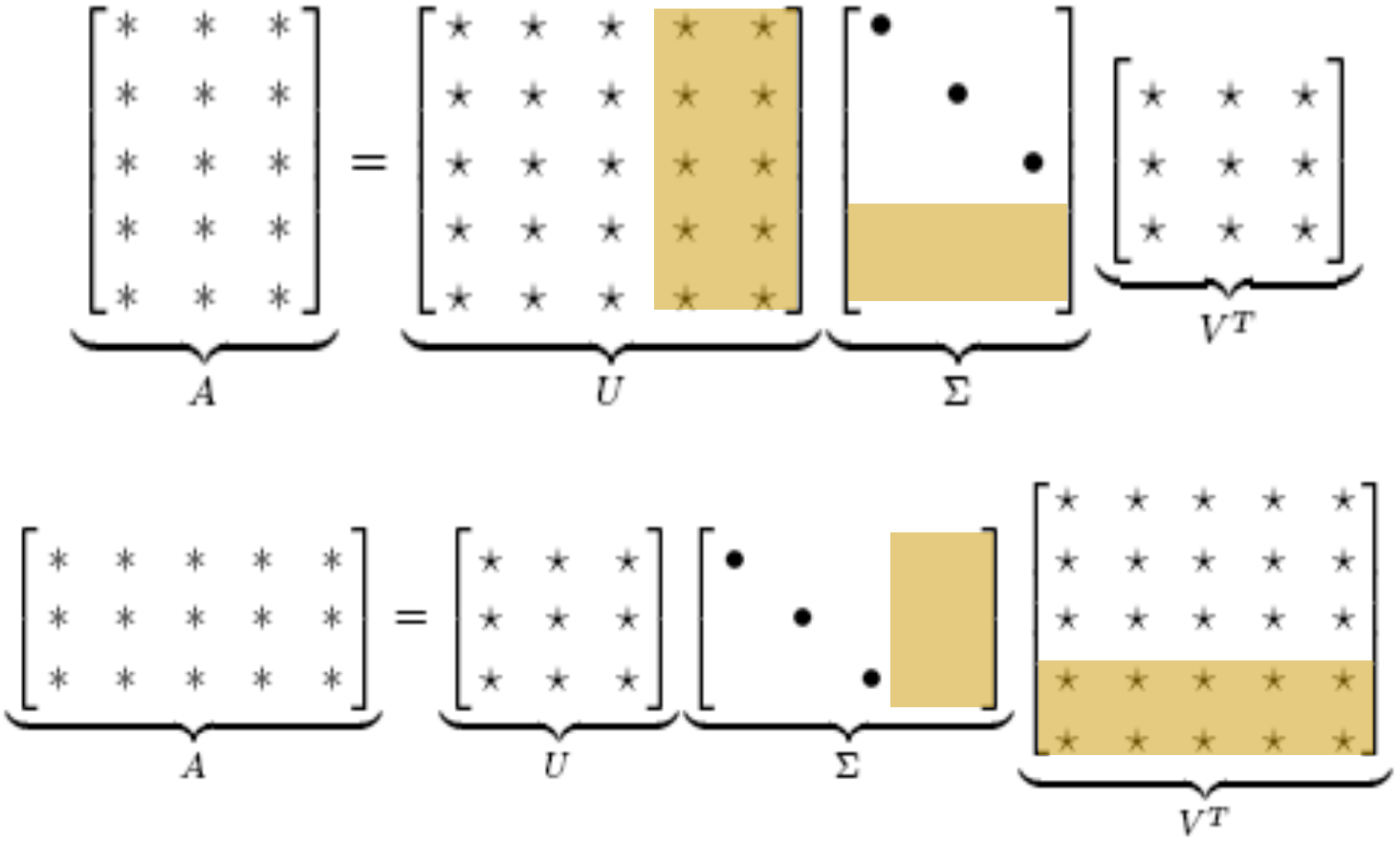


$$M = U \cdot \Sigma \cdot V^*$$

Mx = M(x) = U(S(V*(x)))

Singular Value Decomposition

- Illustration of SVD dimensions and sparseness



Singular Value Decomposition-example

- Let $A = \begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$

Thus m=3, n=2. Its SVD is

$$\begin{bmatrix} 0 & 2/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & 1/\sqrt{6} & -1/\sqrt{3} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{3} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

Typically, the singular values arranged in decreasing order.

Low-rank Approximation

- SVD can be used to compute optimal **low-rank approximations**.
- Approximation problem: Find $\boldsymbol{A_k}$ of rank $\boldsymbol{k}$ such that

$$A_k = \min_{X:rank(X)=k} \|A - X\|_F$$

Frobenius norm (aka Euclidian norm)

$$\|A\|_F \equiv \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}.$$

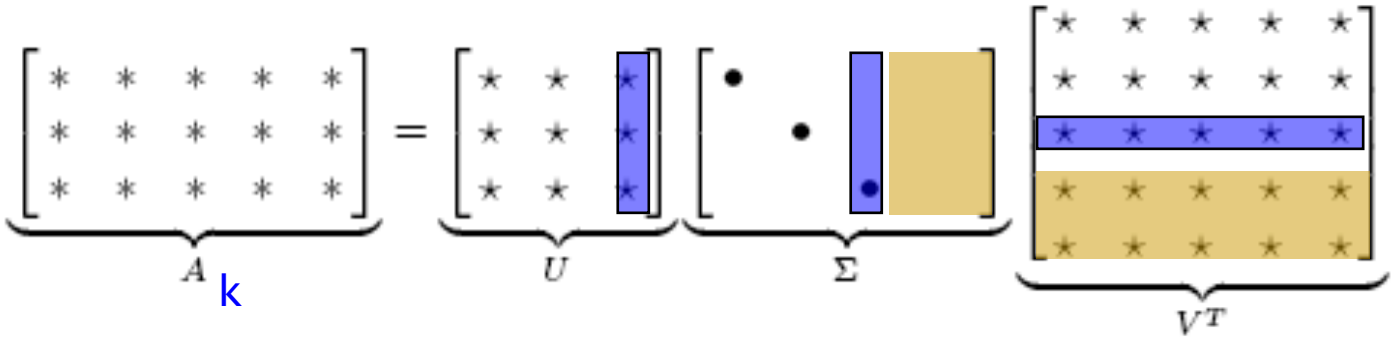
A_k and X are both $m \times n$ matrices.

Typically, want $k << r$.

Low-rank Approximation

- Solution via SVD

$$A_k = U\ \mathrm{diag}(\sigma_1,\ldots,\sigma_k,\underbrace{0,\ldots,0}_{\text{set smallest r-k singular values to zero}})V^T$$



$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$ ← *column notation: sum of rank 1 matrices*

- Error: $\min_{X:rank(X)=k} \|A-X\|_F = \|A-A_k\|_F = \sigma_{k+1}$

Principle Component Analysis (PCA)

- How do we find the eigenvectors v_i?
- We use singular value decomposition to decompose Σ into an orthogonal rotation matrix U and a diagonal scaling matrix S:

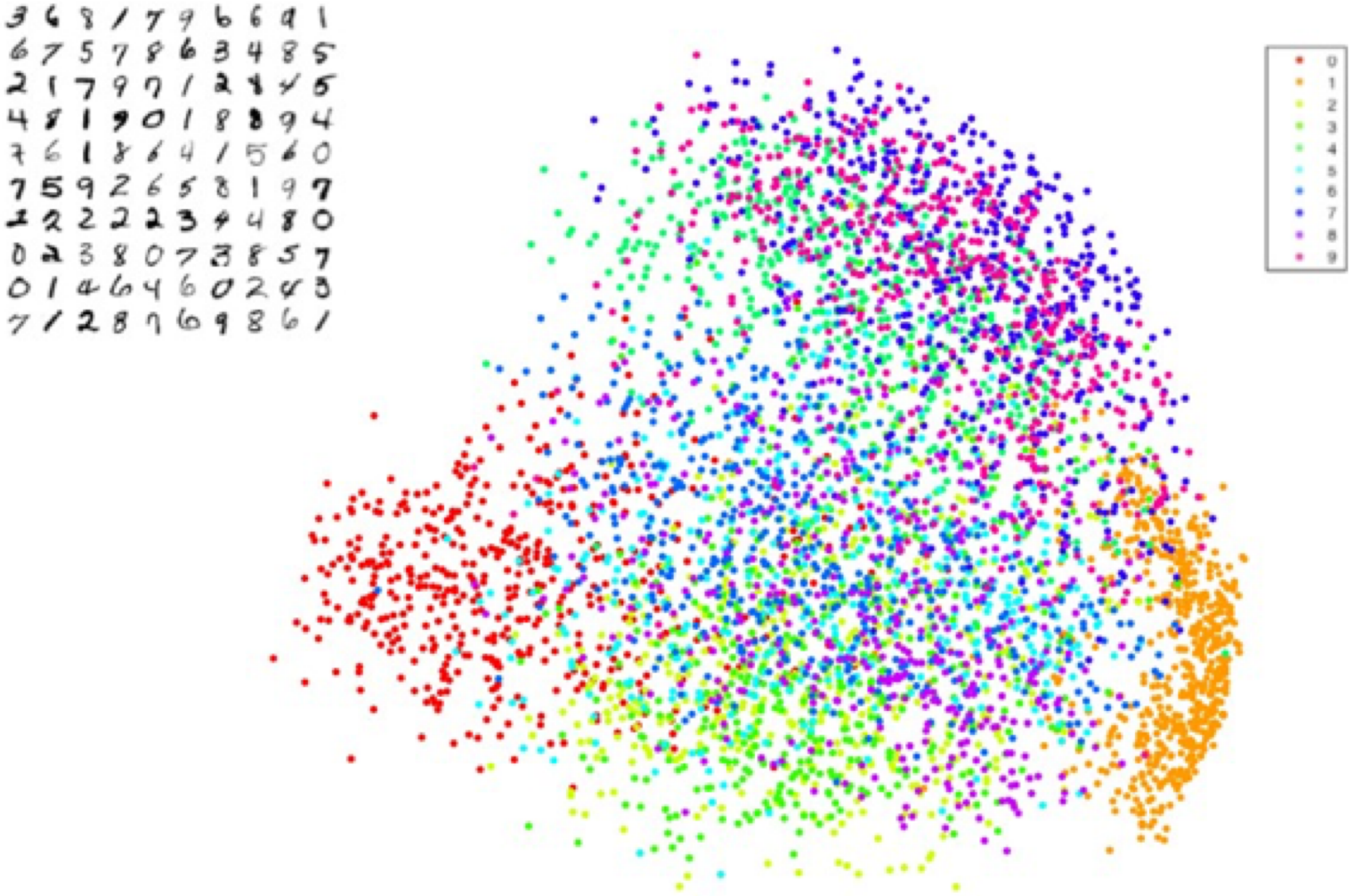
$$\Sigma = USU^T \tag{22}$$

$$\Sigma U = (USU^T)U \tag{23}$$

$$= US \tag{24}$$

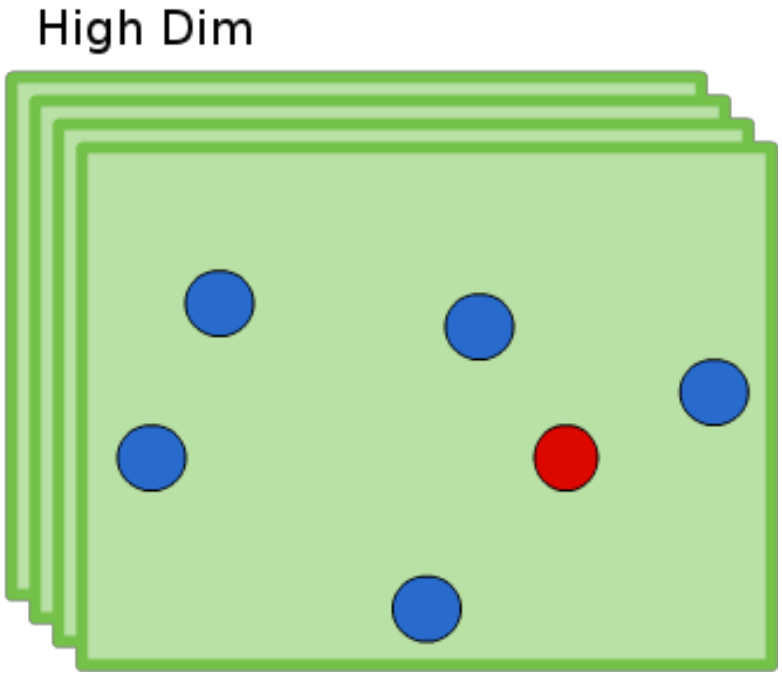
- The columns of U are the v_i, and S is the diagonal matrix of eigenvalues λ_i^2

PCA of MNIST digits

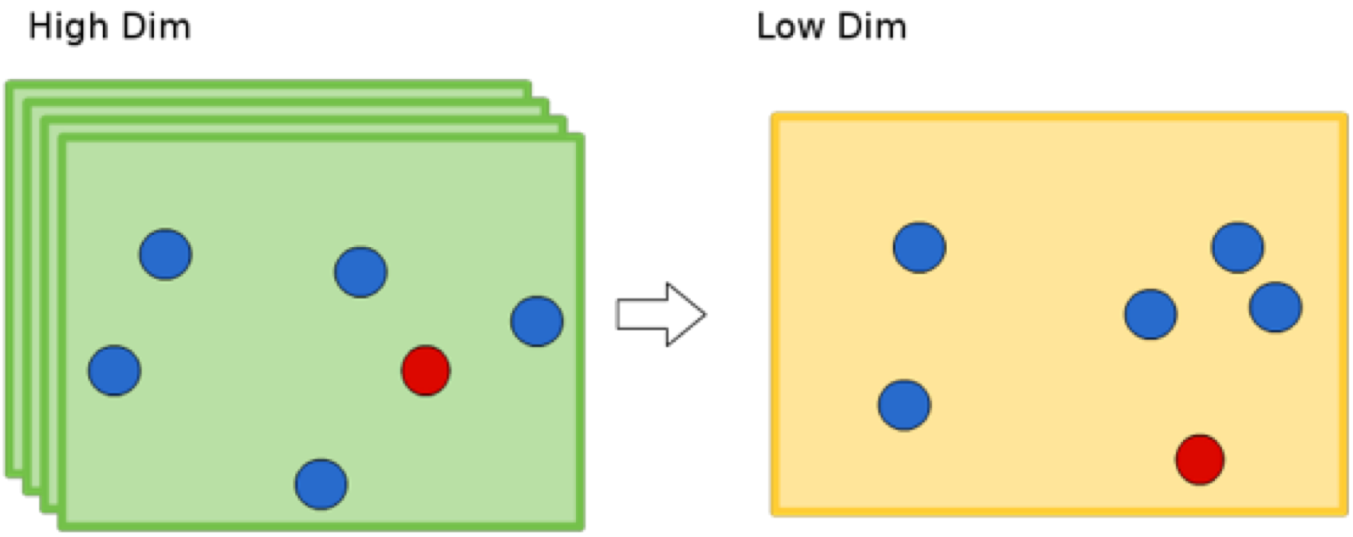


Non-linear embeddings: t-SNE

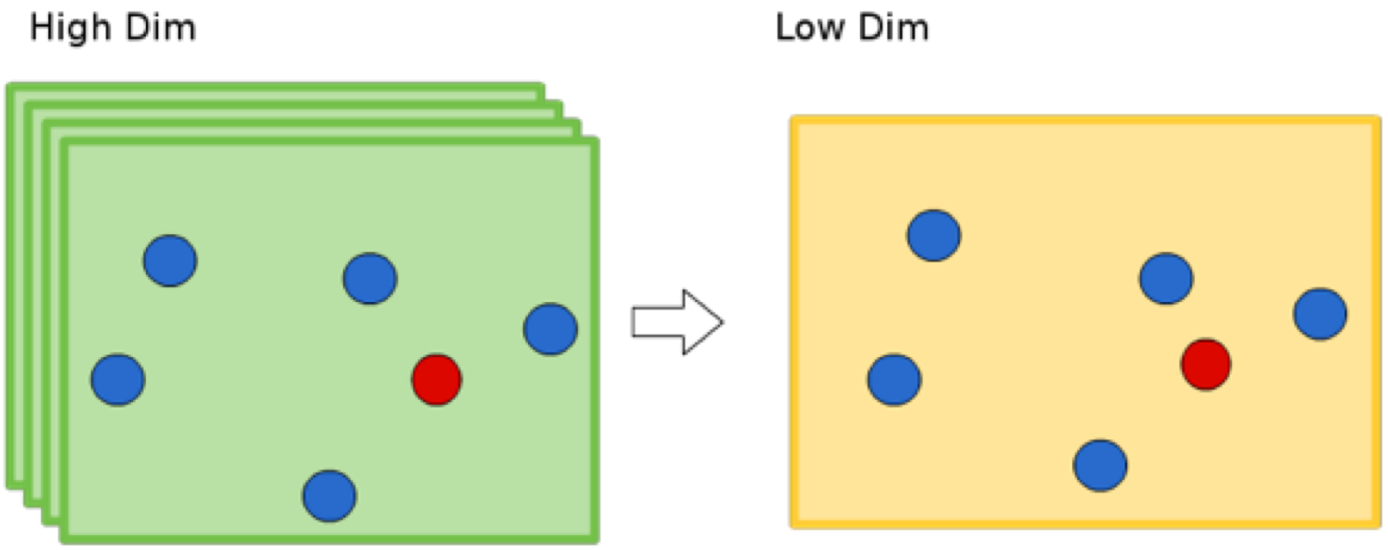
Distance Preservation Neighbor Preservation



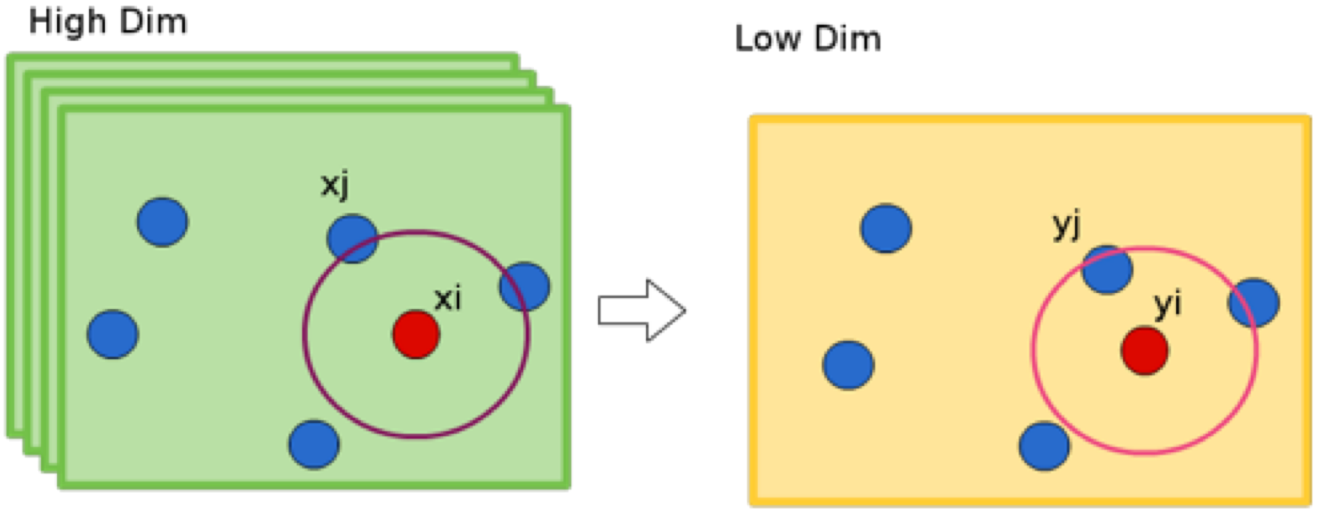
Neighborhood not preserved



Neighborhood preserved



Measure pairwise distances in high dimensional space



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

Set the bandwidth σ_i such that the conditional has a fixed perplexity (effective number of neighbors) $Perp(P_i) = 2^{H(P_i)}$, typical value is about 5 to 50

Shannon entropy of P_i

We want to choose an embedding that minimizes divergence between low and high dimension similarities

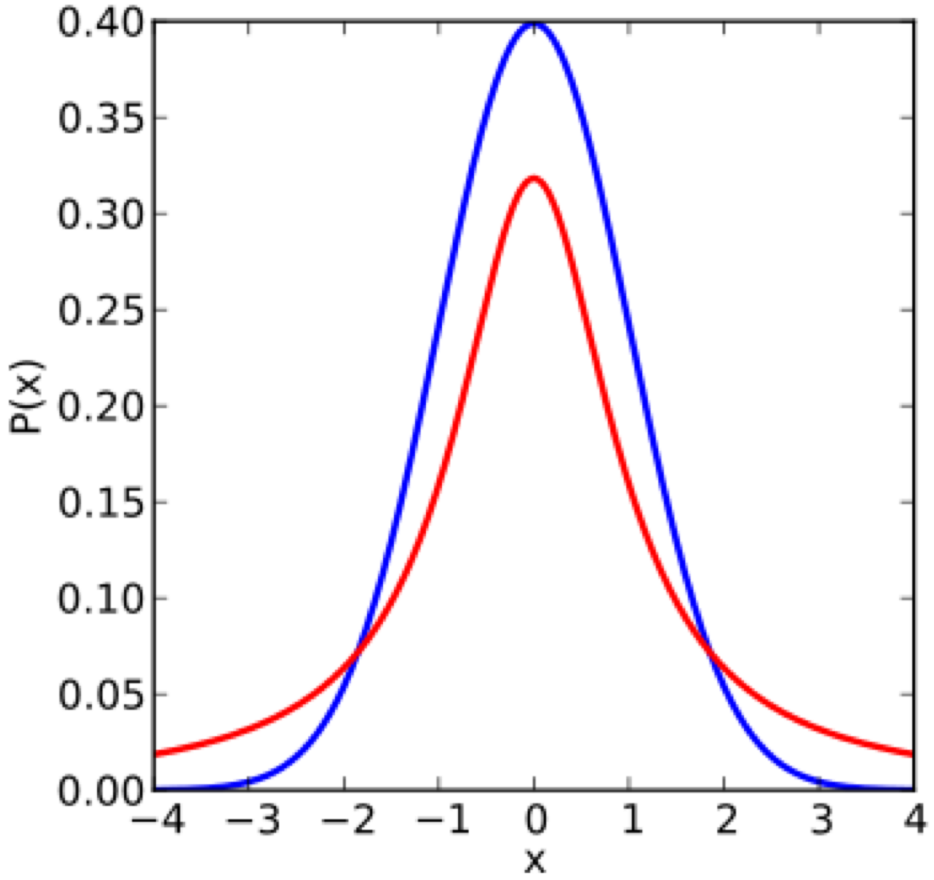
- Similarity of datapoints in High Dimension

$$p_{ij} = \frac{exp(-||x_i - x_j||^2/2\sigma^2)}{\sum_{k \neq l} exp(-||x_l - x_k||^2/2\sigma^2)}$$

- Similarity of datapoints in Low Dimension

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

Low dimensional embedding using a Student t-distribution to avoid overcrowding



Red – Student t-distribution (1 degree of freedom)
Blue - Gaussian

We can use gradient methods to find an embedding

- Cost function

p_{ij} = New (low) dimension distance

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

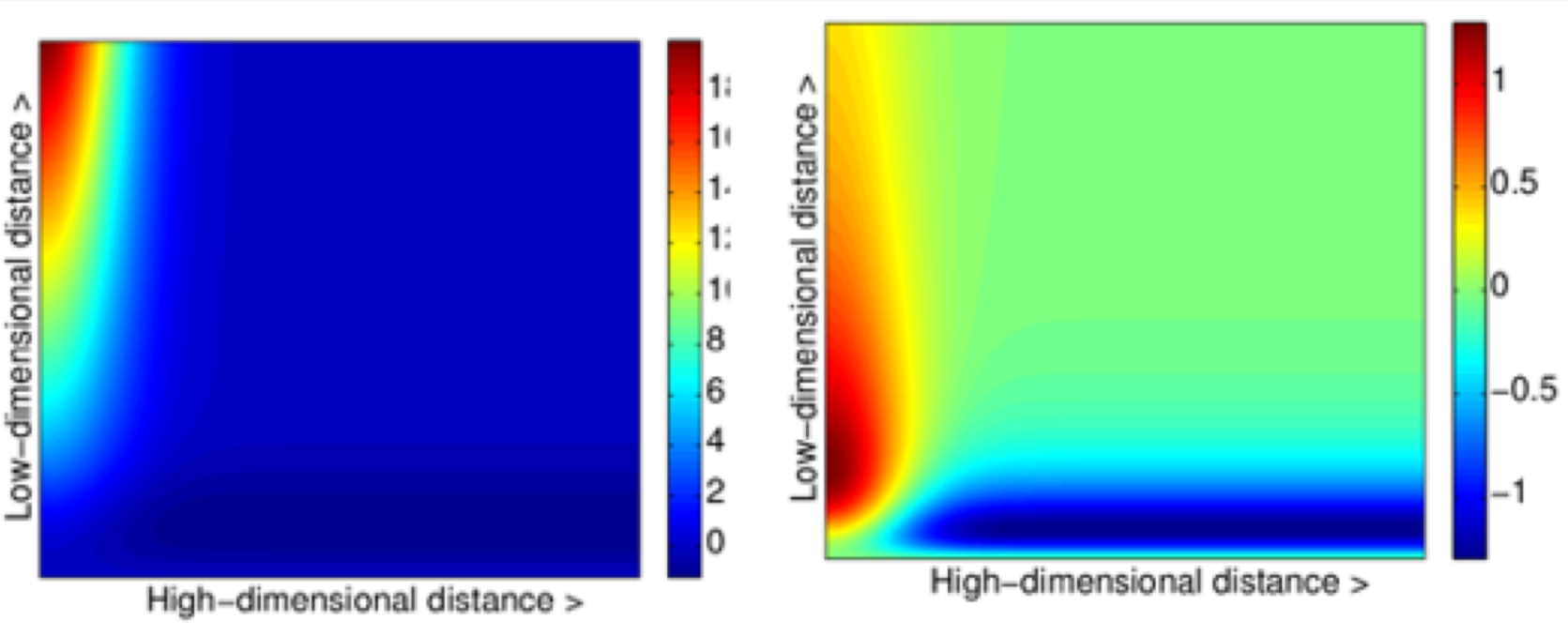
q_{ij} = Original (high) dimension D

  - Large p_{ij} modeled by small q_{ij}: Large penalty (not okay to bring distant points closer)
  - Small p_{ij} modeled by large q_{ij}: Small penalty (okay to separate nearby points)
  - t-SNE mainly preserves local similarity structure of the data

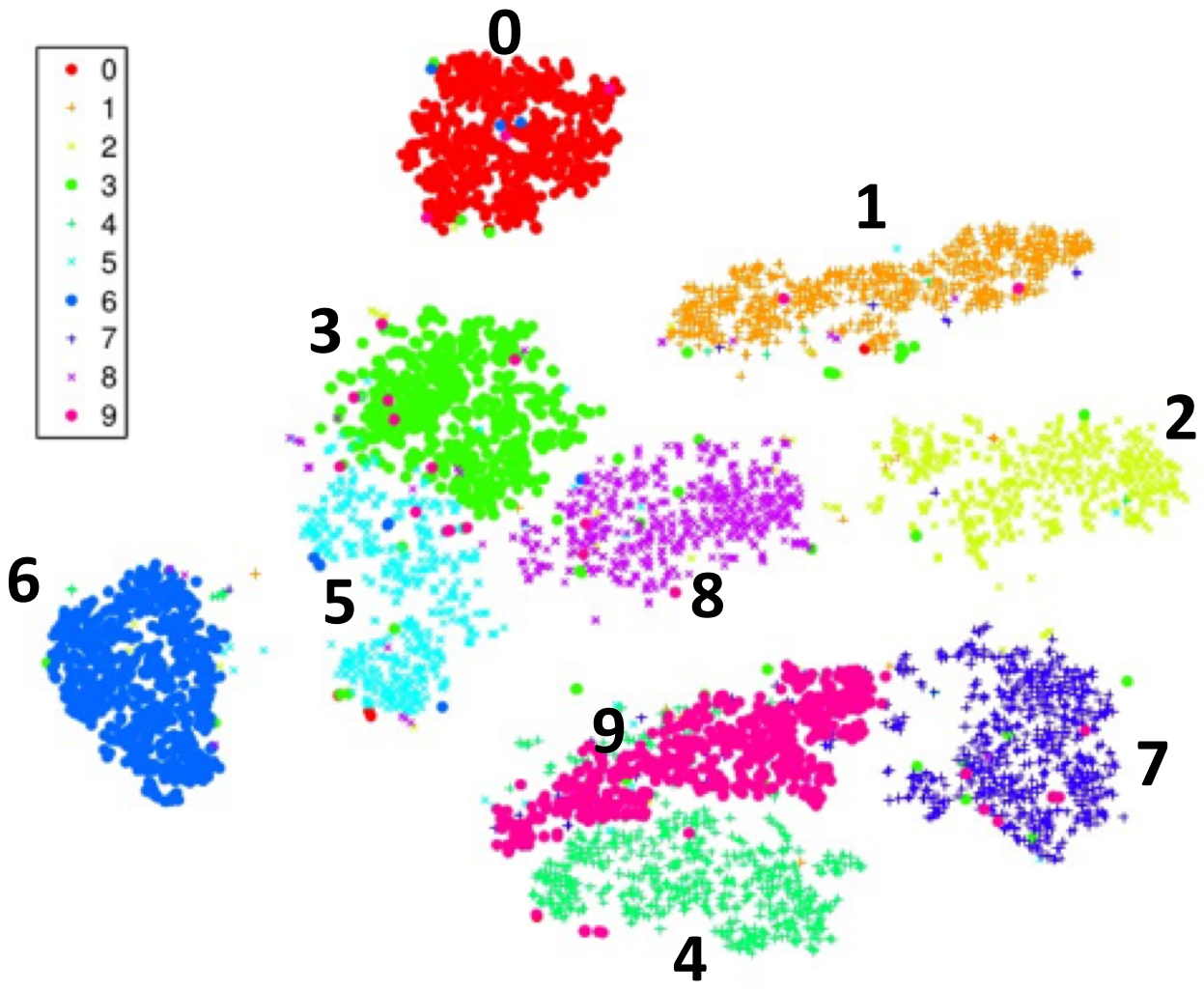
- Gradient

$$\frac{\partial C}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(1 + ||y_i - y_j||^2)^{-1}(y_i - y_j)$$

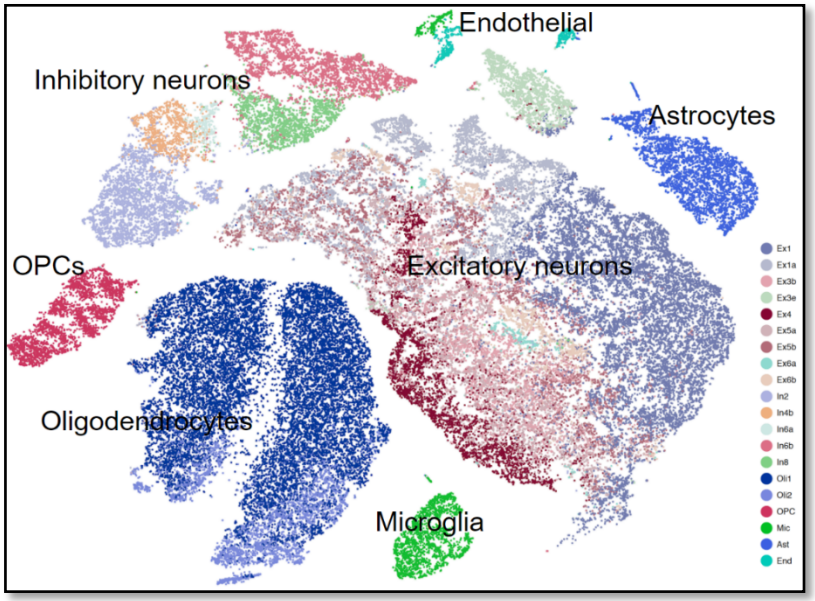
Interpretation of SNE (left) and t-SNE (right) gradients



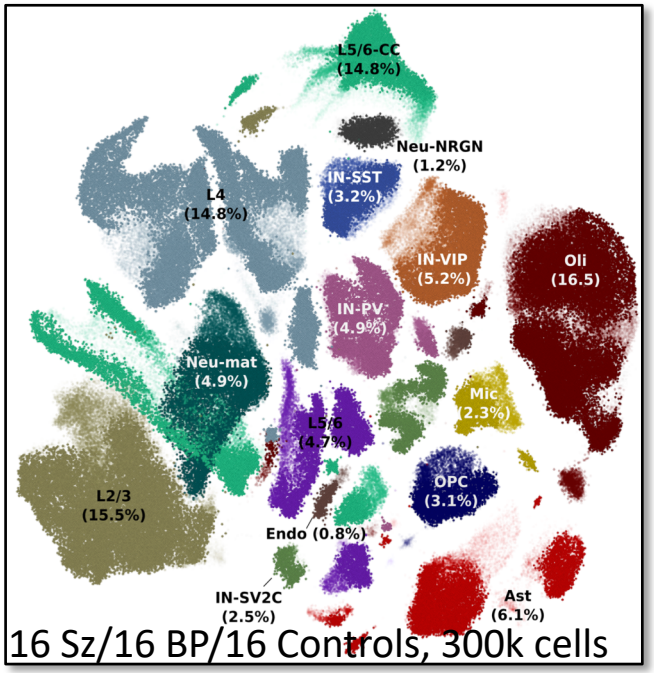
t-SNE of MNIST digits



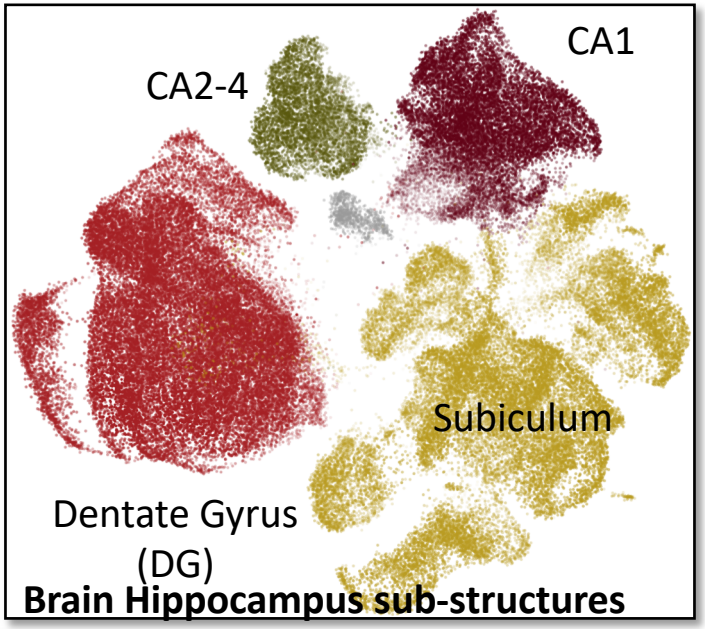
t-SNEs of single-cell Brain data



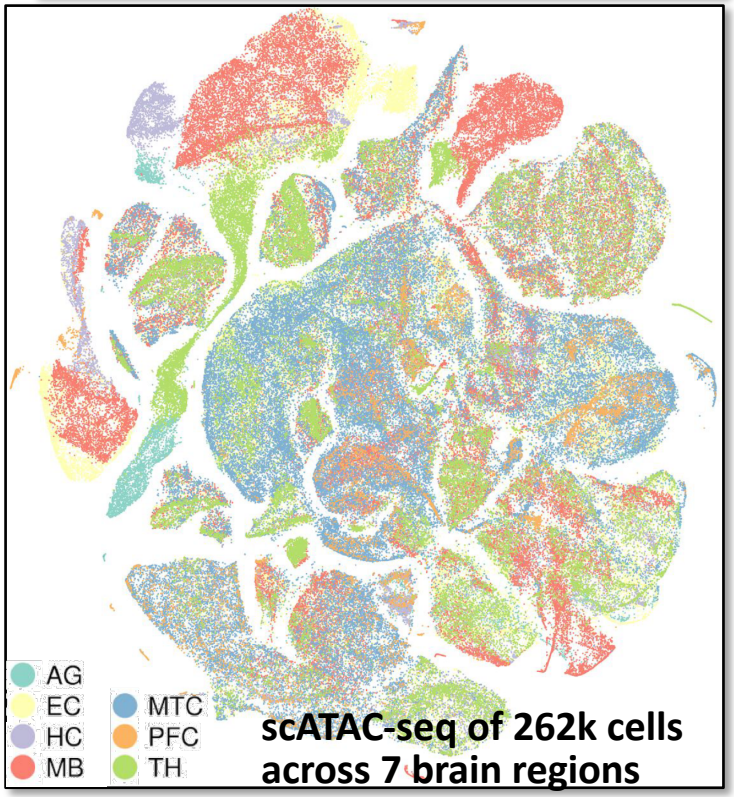
scRNA-seq in 48 individuals, 84k cells, Nature, 2019



16 Sz/16 BP/16 Controls, 300k cells



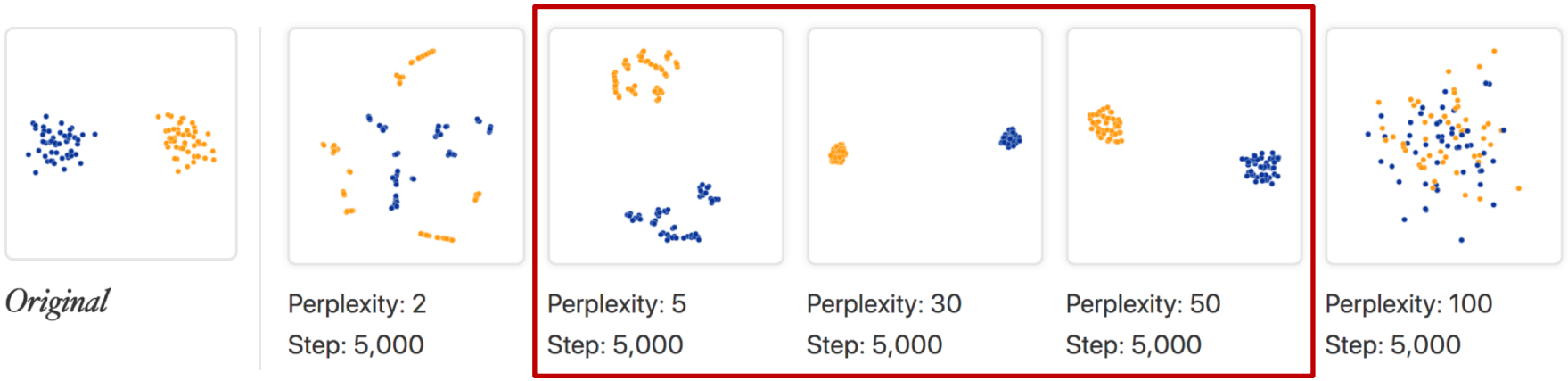
Brain Hippocampus sub-structures



scATAC-seq of 262k cells across 7 brain regions

Playing with t-SNE parameters

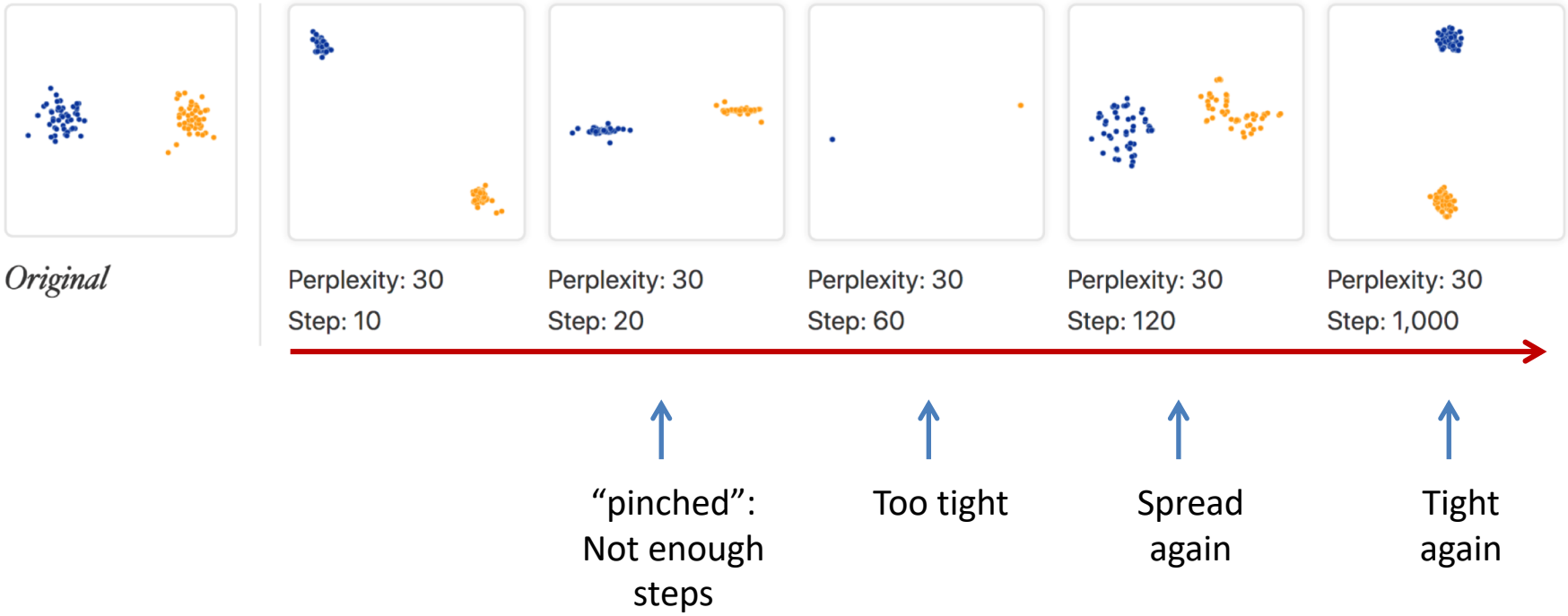
Perplexity matters



Recommended range by Van Der Maaten and Hinton

https://distill.pub/2016/misread-tsne/

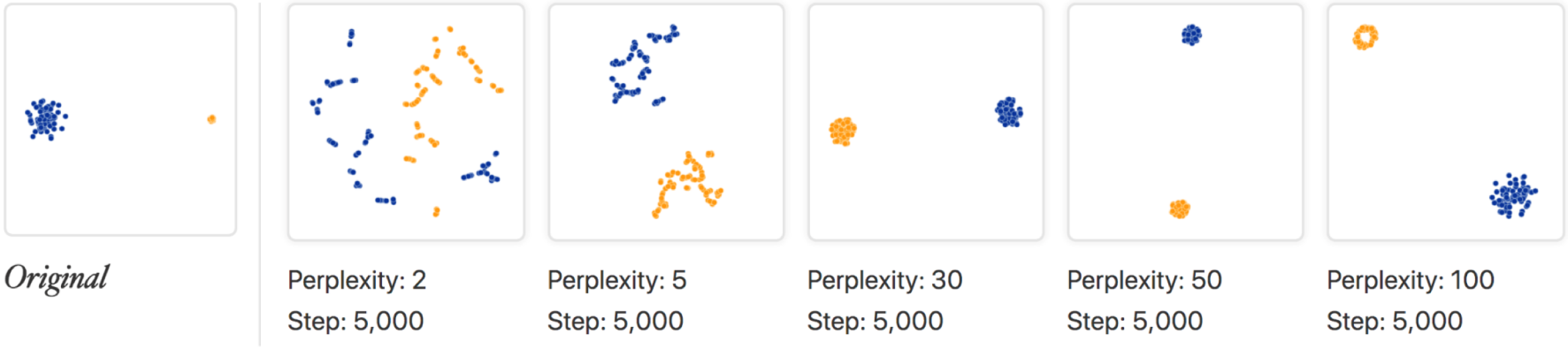
Number of steps matter



https://distill.pub/2016/misread-tsne/

Cluster sizes are not meaningful

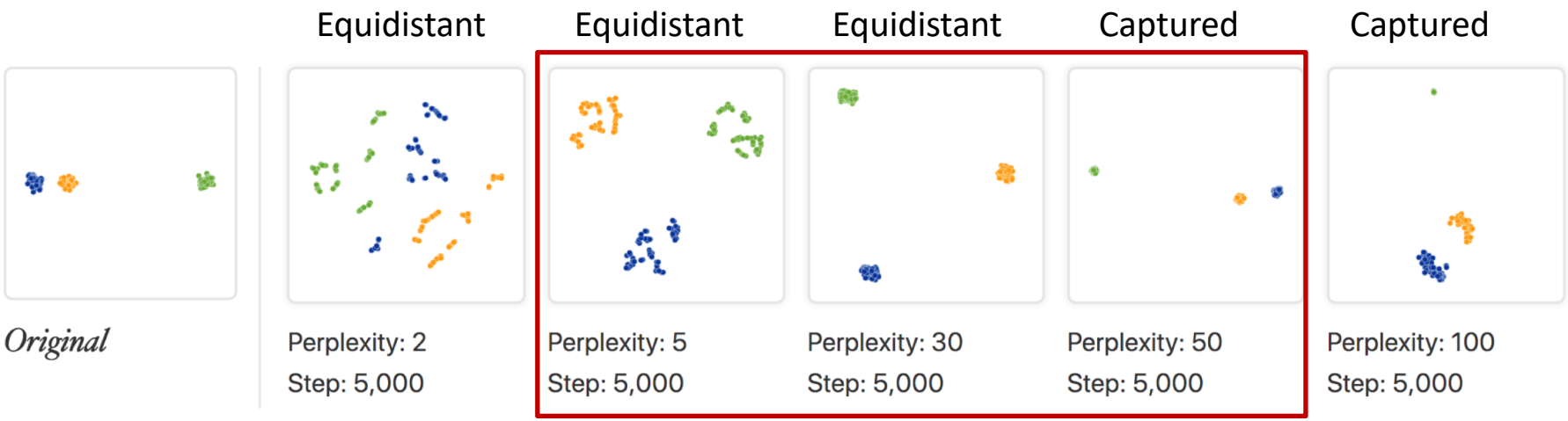
Original data: 2 Gaussians
Widely different (10-fold) dispersion



t-SNE loses that notion of distance.
By design, it adapts to regional variations in distance.

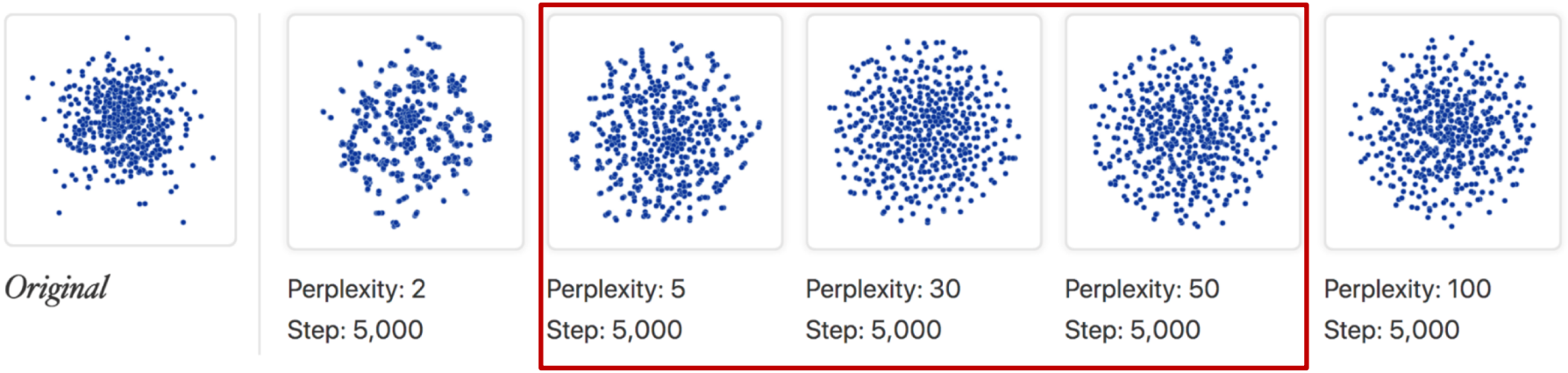
https://distill.pub/2016/misread-tsne/

Between-cluster distance is not always preserved



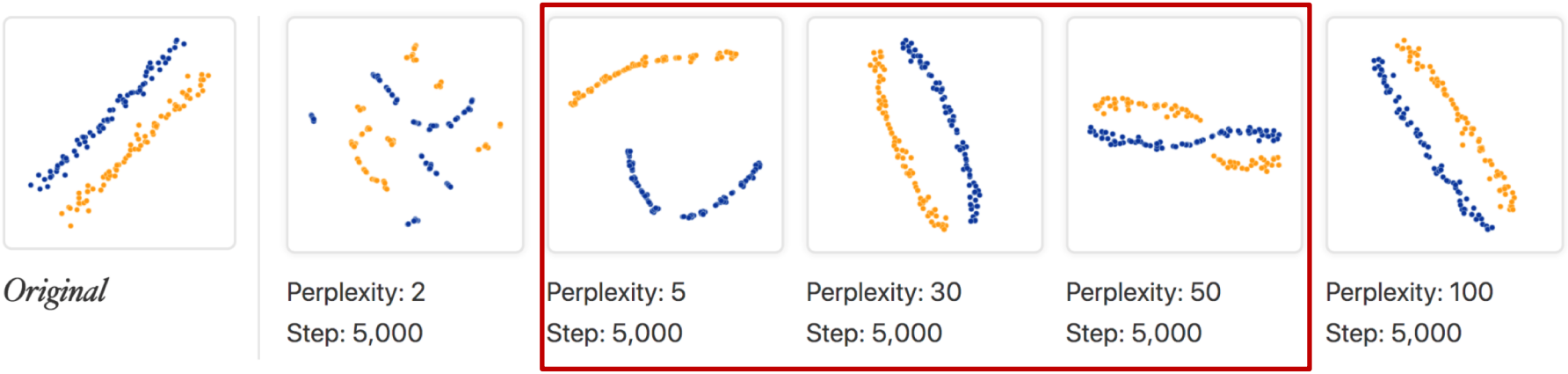
https://distill.pub/2016/misread-tsne/

False clusters may appear



https://distill.pub/2016/misread-tsne/

Relationships are not always preserved



https://distill.pub/2016/misread-tsne/

Different runs produce surprisingly similar results…

Original

Perplexity: 50
Step: 5,000

Perplexity: 50
Step: 5,000

Perplexity: 50
Step: 5,000

Perplexity: 50
Step: 5,000

Perplexity: 50
Step: 5,000

(… but not at very low perplexity)

Original

Perplexity: 2
Step: 5,000

Perplexity: 2
Step: 5,000

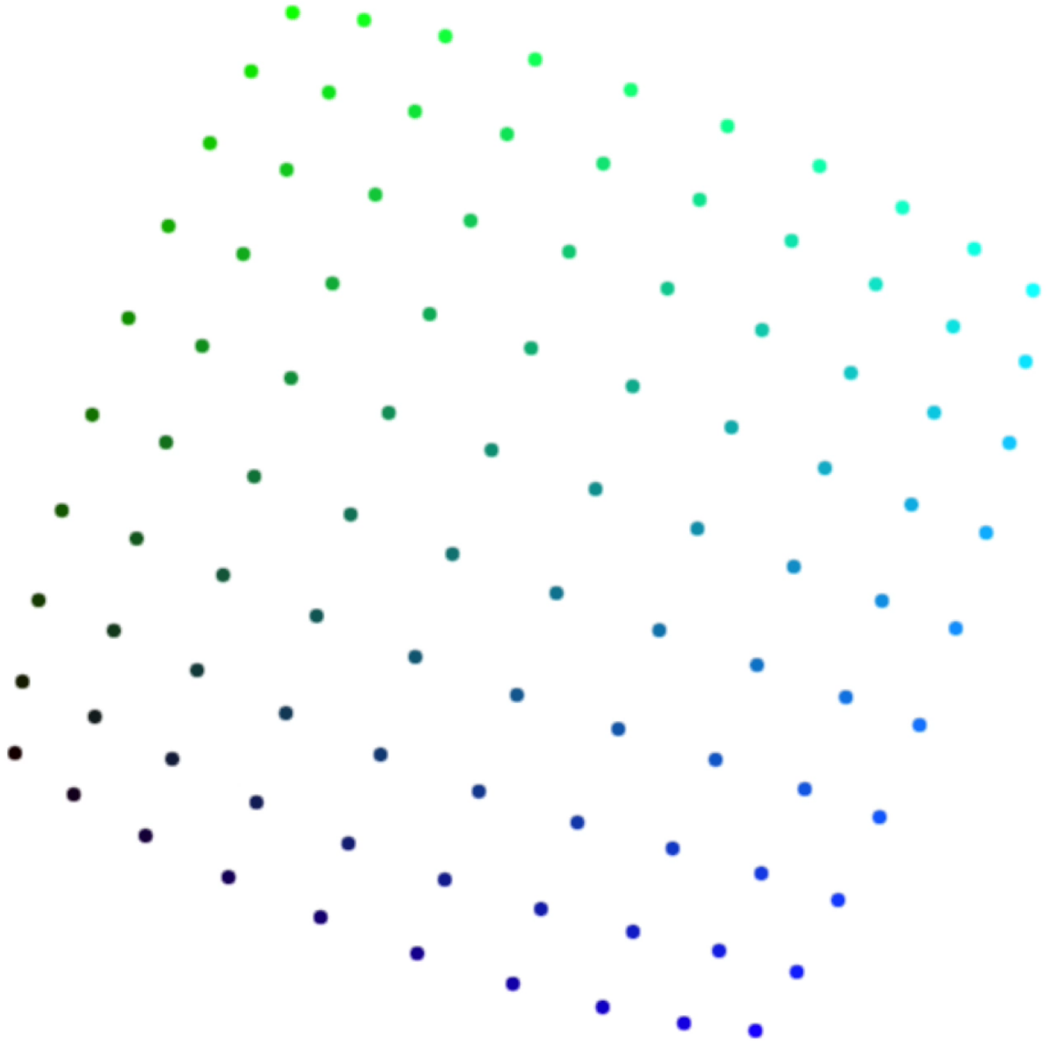
Perplexity: 2
Step: 5,000

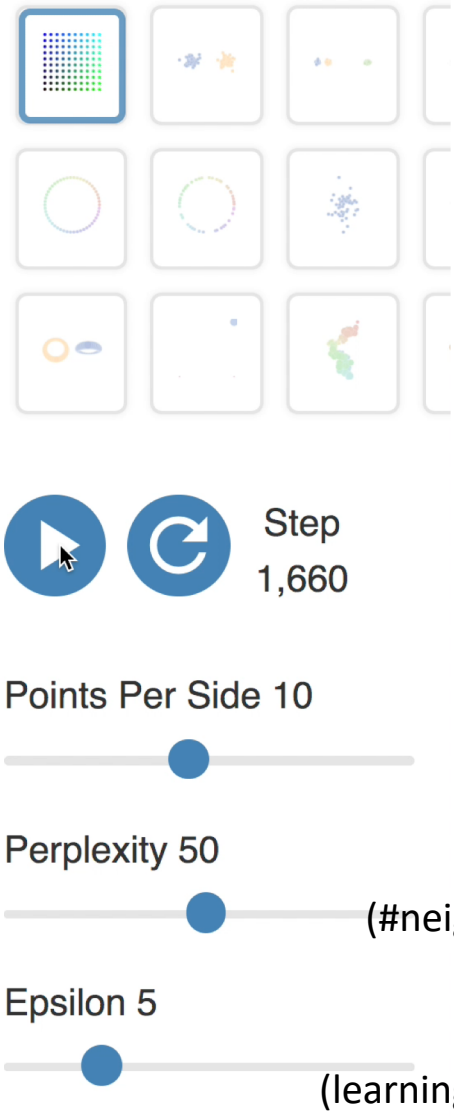
Perplexity: 2
Step: 5,000

Perplexity: 2
Step: 5,000

https://distill.pub/2016/misread-tsne/

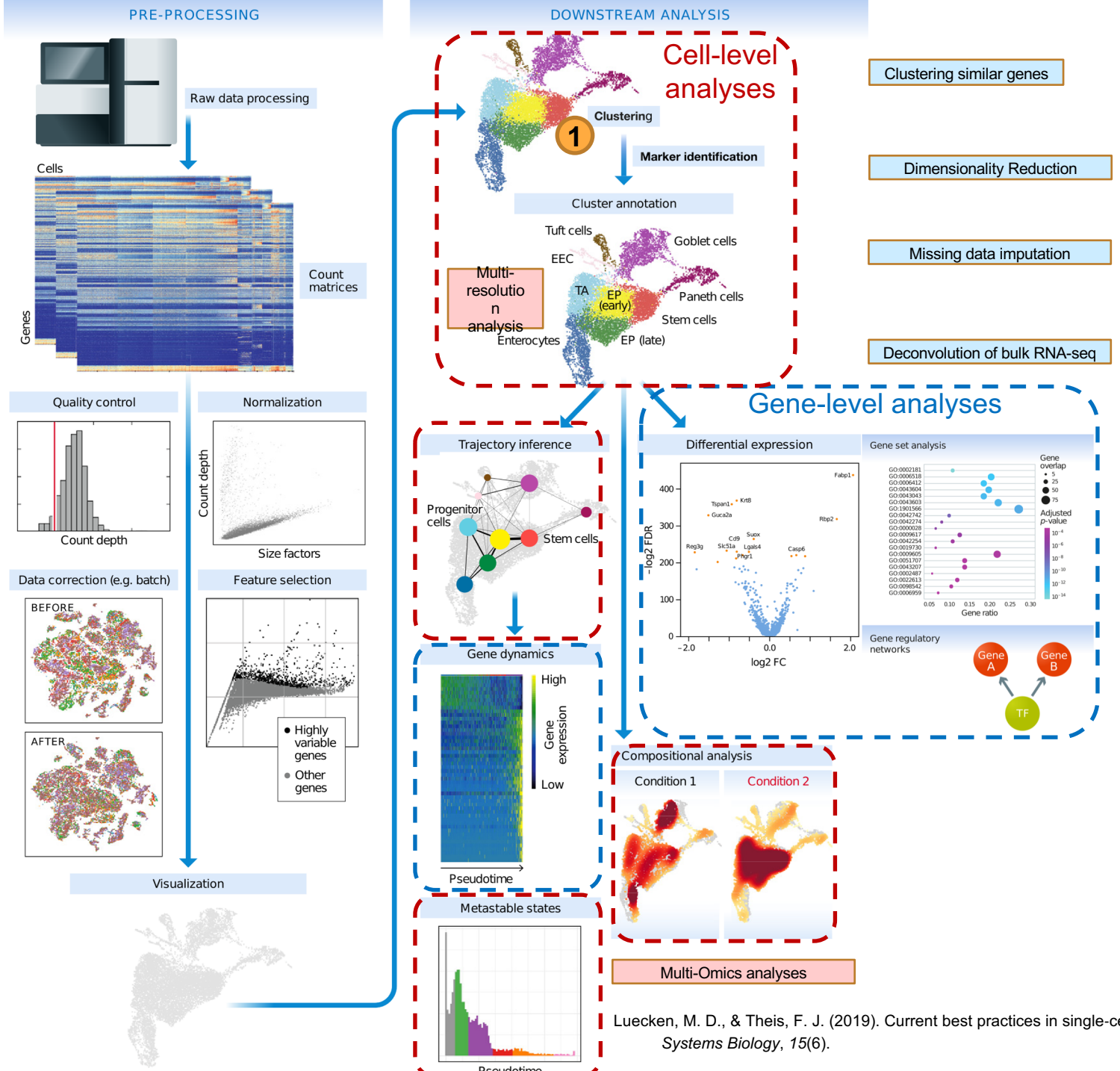
Demo of t-SNE in action





https://distill.pub/2016/misread-tsne/

Computational challenges in single-cell data analysis



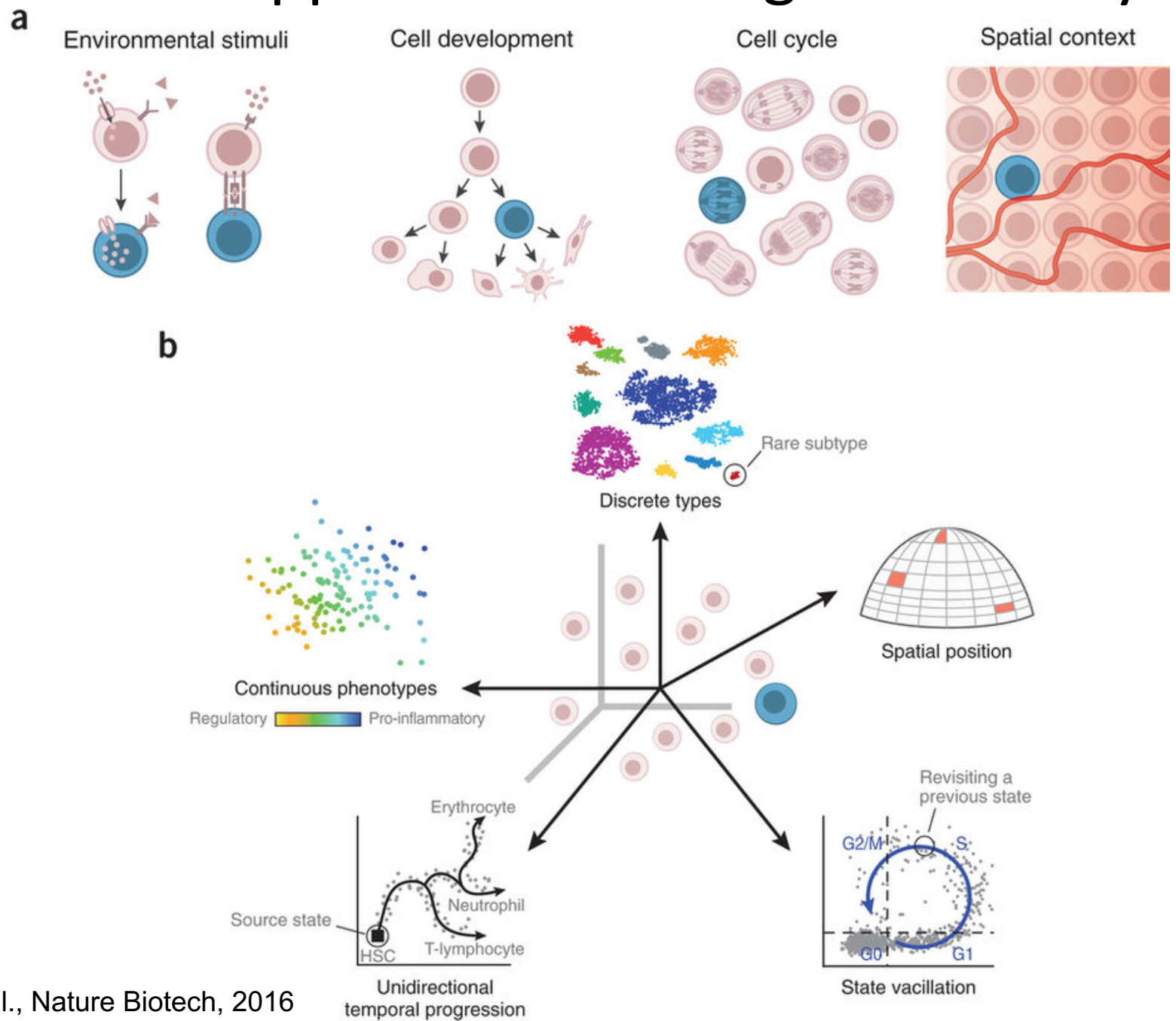
Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, *15*(6).

Extracting biological insights from scRNA-seq data

- Cell-to-cell correlation
- Gene-to-gene correlation
- Imputation of missing values
- Cellular trajectories and differentiation

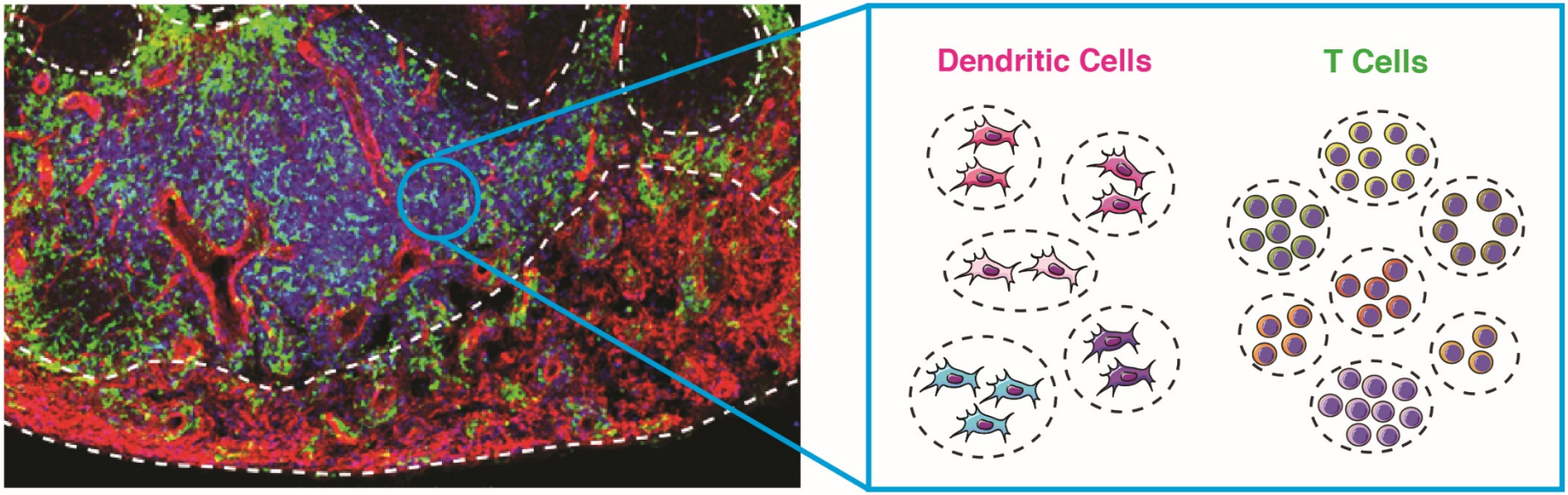
Clustering similar cells

Methods + applications of single-cell analysis



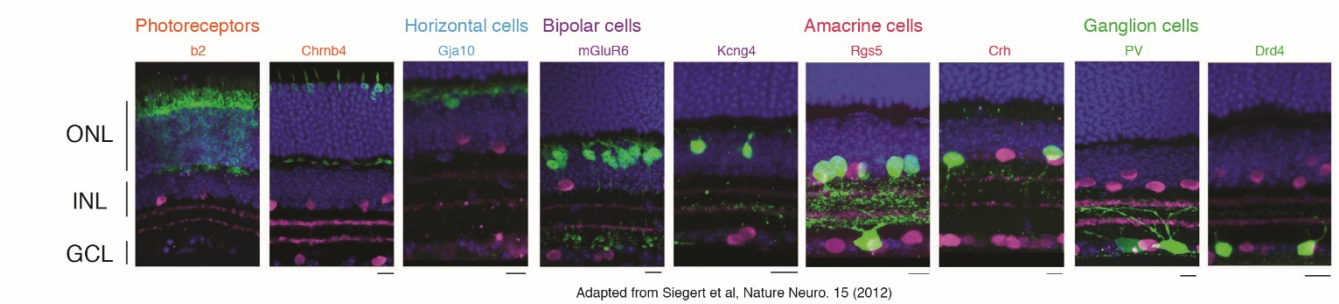
Wagner et al., Nature Biotech, 2016

From Complex Tissues to individual cell types

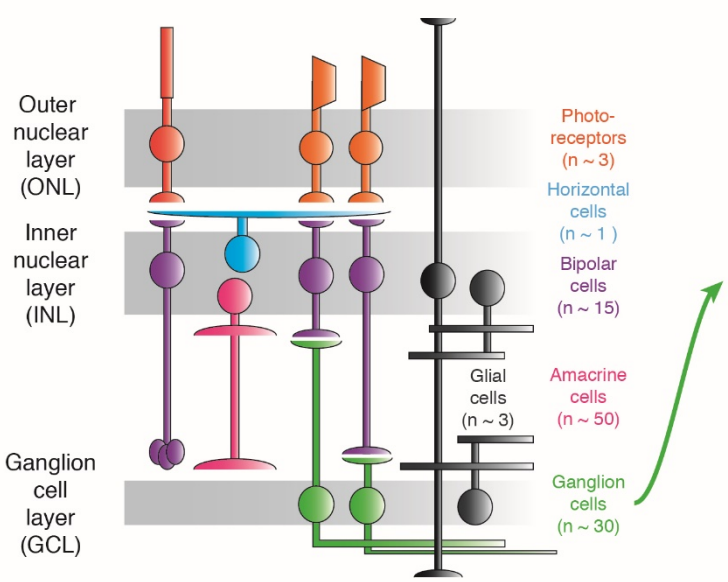


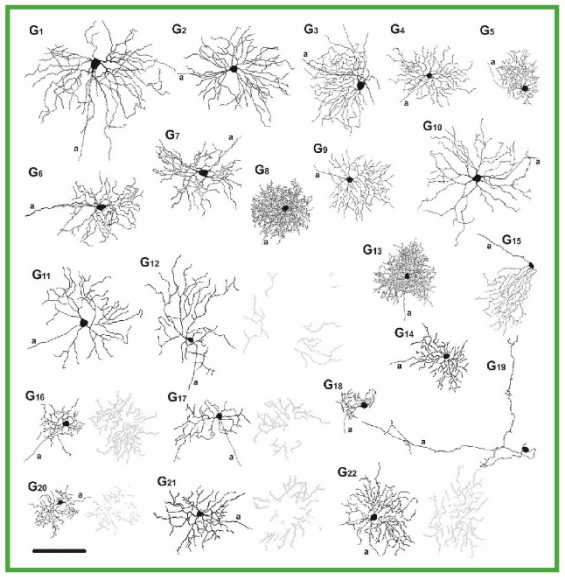
Can we identify the different cell types/states in a complex tissue?

Brain Case Study: The Mouse Retina

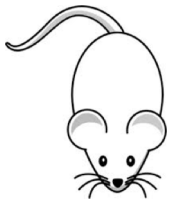


Adapted from Siegert et al, Nature Neuro. 15 (2012)





Adapted from Volgyi et al, *J. Comparative Neurology* **512** (2009)

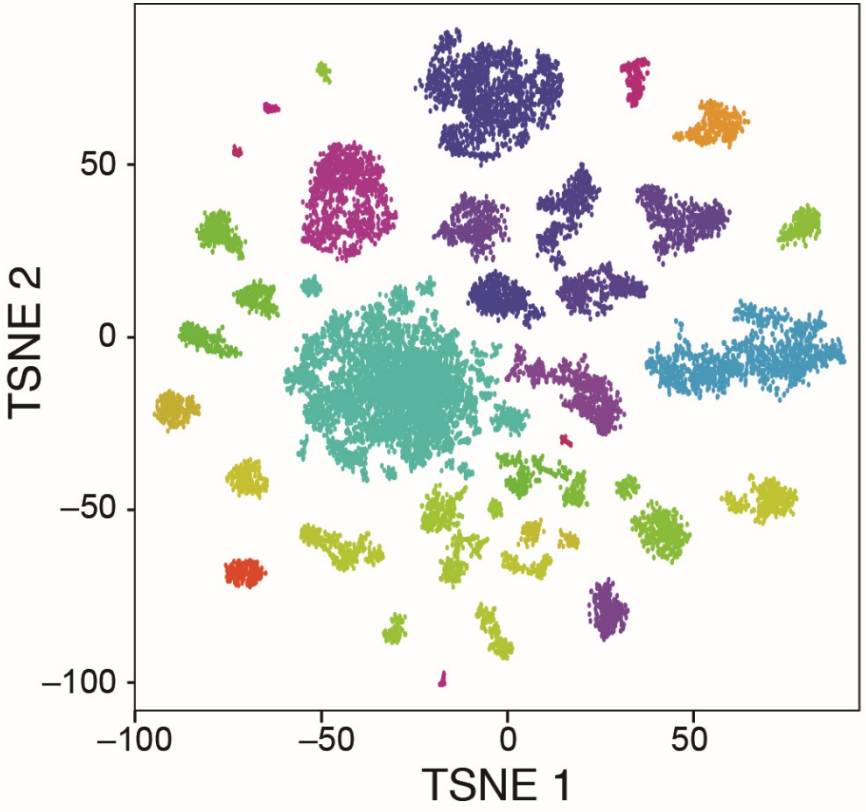


~100 cell subtypes, only some with molecular markers

Macosko et al, *Cell*, **161**, 2015

49,300 Retina Cells Grouped Into 39 Clusters

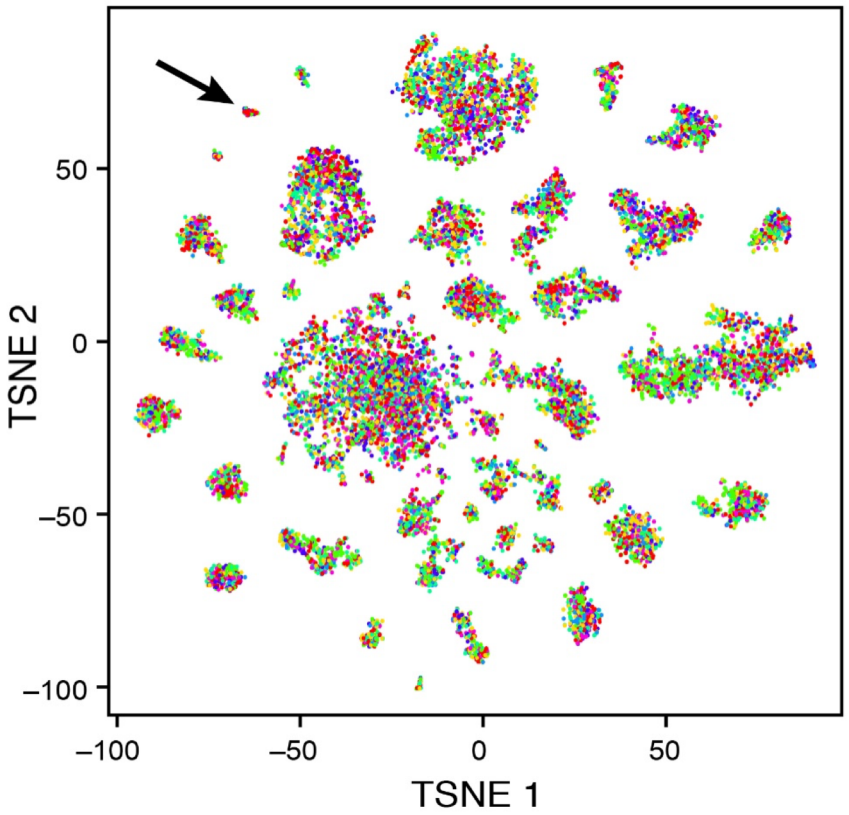
- **Drop-Seq:** 49,300 cells from dissociated mouse retina (P14) (~15k reads per cell)
- **Computational pipeline**: Select 25% best coverage cells, Dimensionality reduction (PCA+tSNE), Project remaining cells, Identify cell types (density clustering), Refine clusters (differential expression)



Macosko et al, *Cell*, **161**, 2015

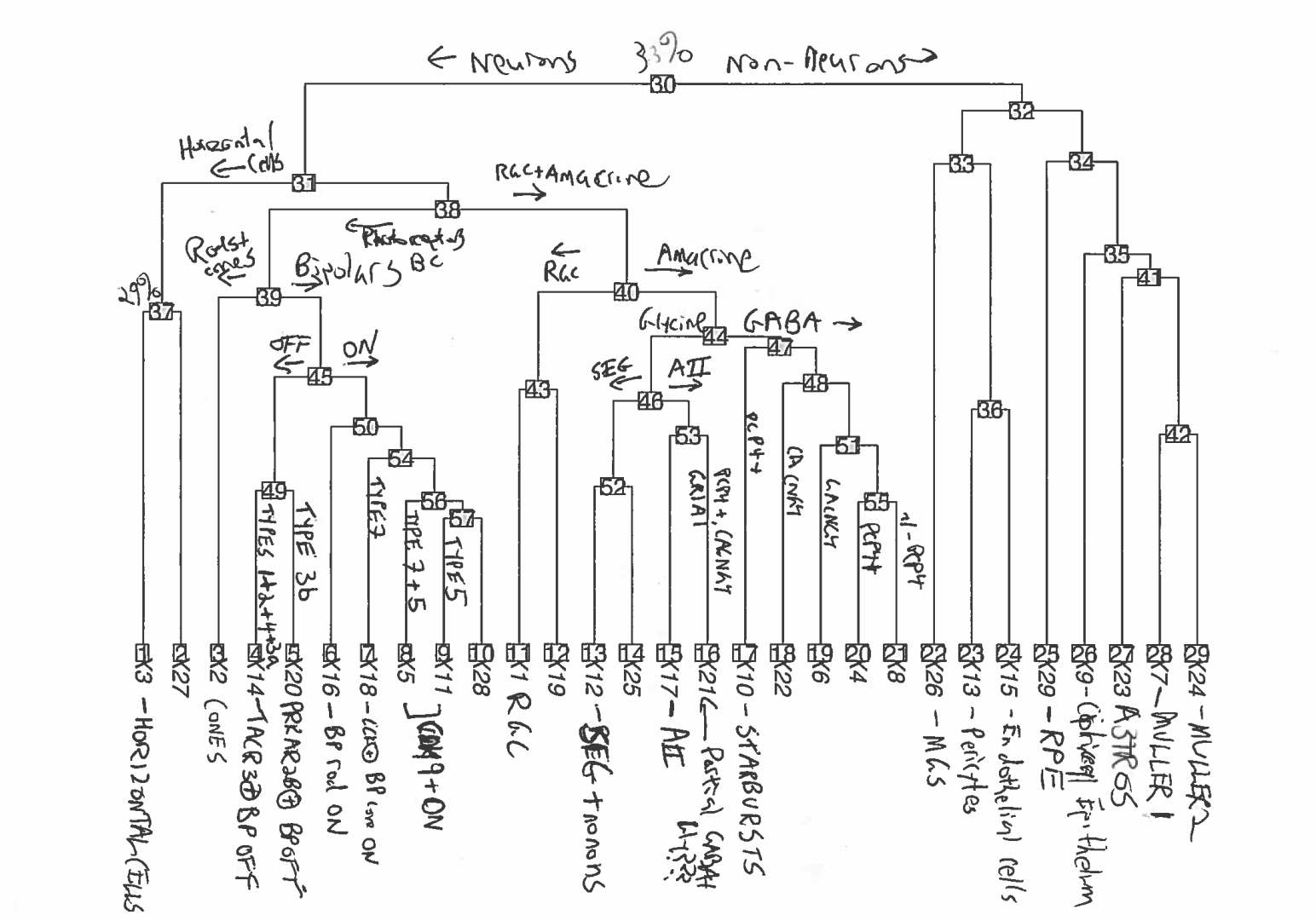
49,300 Retina Cells Grouped Into 39 Clusters

- **Drop-Seq:** 49,300 cells from dissociated mouse retina (P14) (~15k reads per cell)
- **Computational pipeline**: Select 25% best coverage cells, Dimensionality reduction (PCA+tSNE), Project remaining cells, Identify cell types (density clustering), Refine clusters (differential expression)



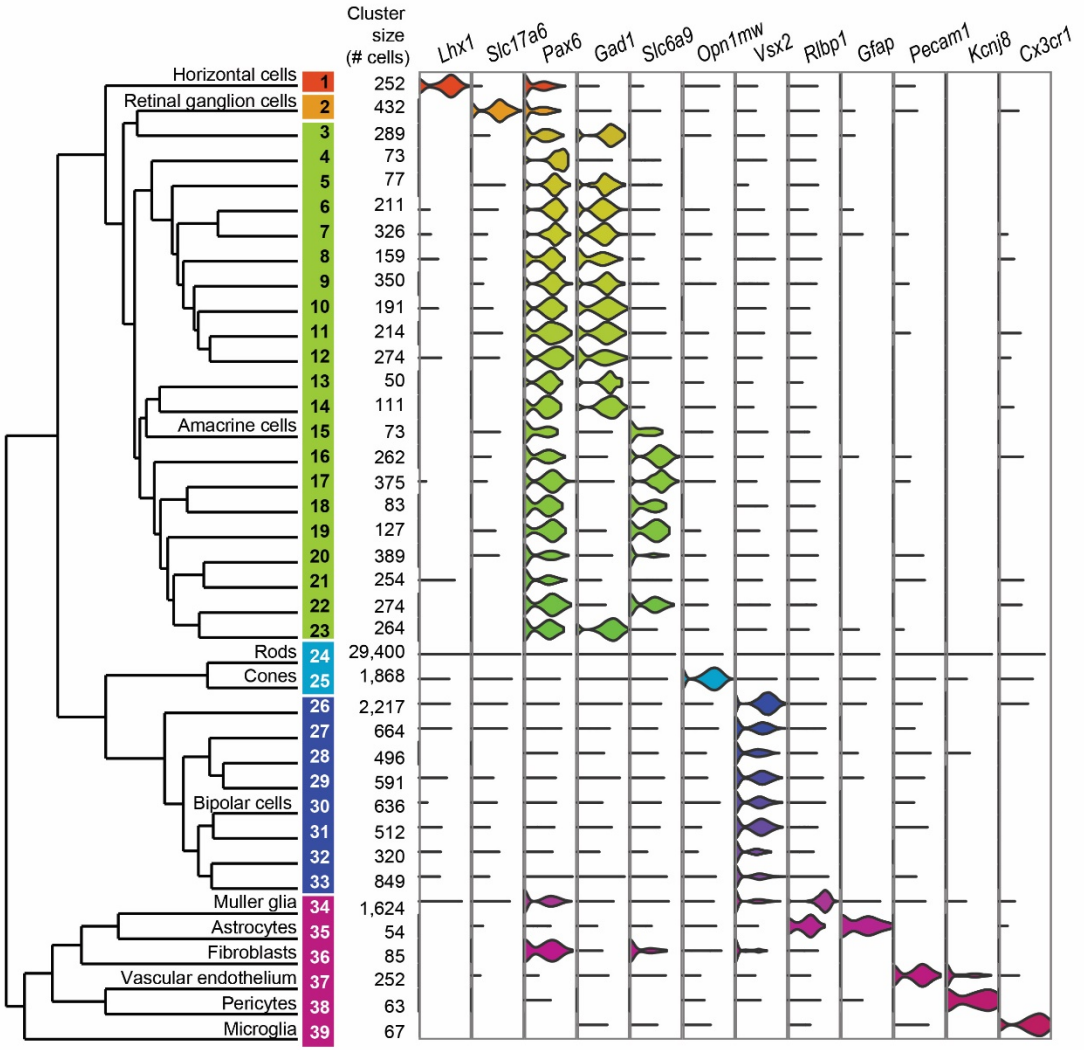
Macosko et al, *Cell*, **161**, 2015

Annotating A Cell Atlas



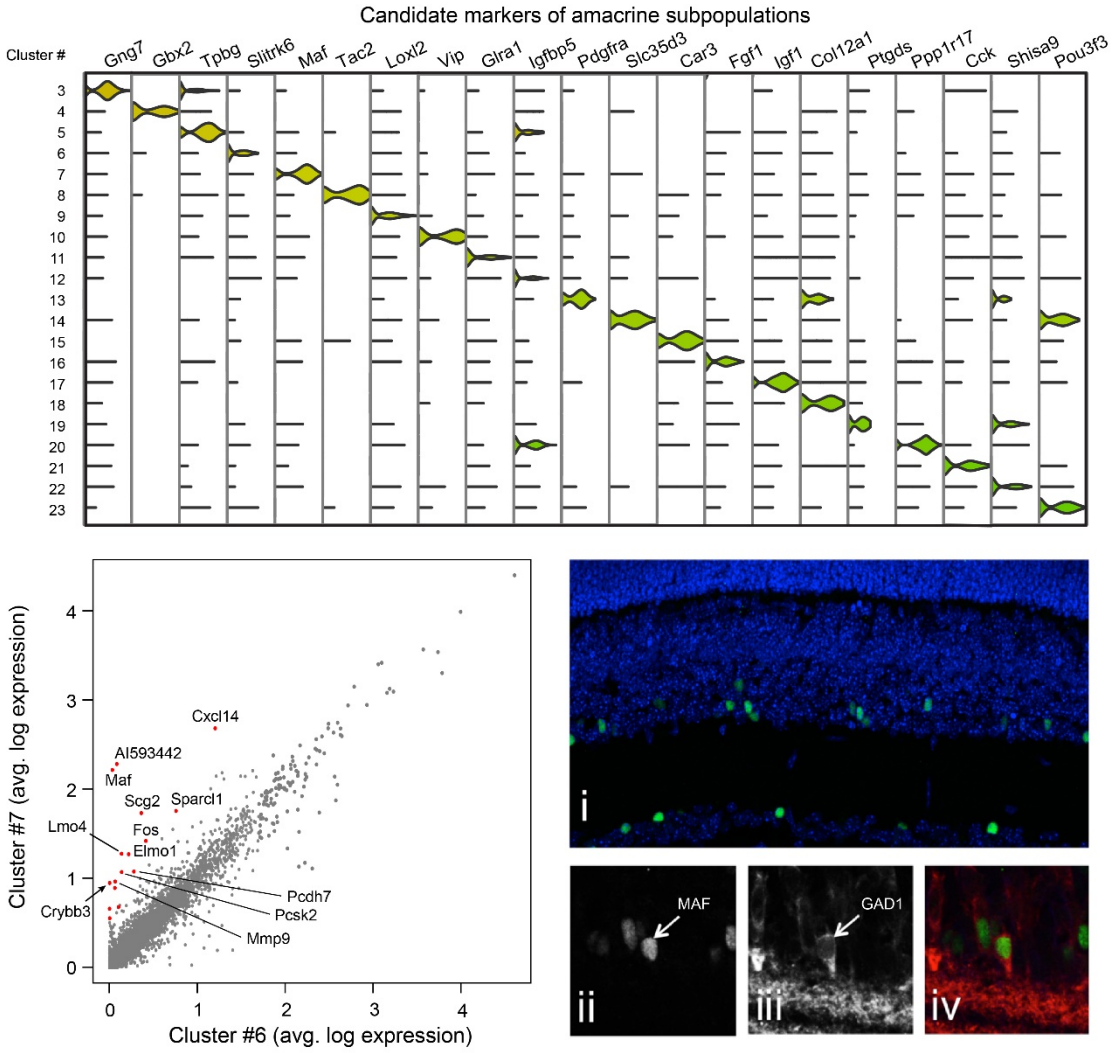
Macosko et al, *Cell*, **161**, 2015

39 Clusters: Known Cell Types & Relationships



Macosko et al, *Cell*, **161**, 2015

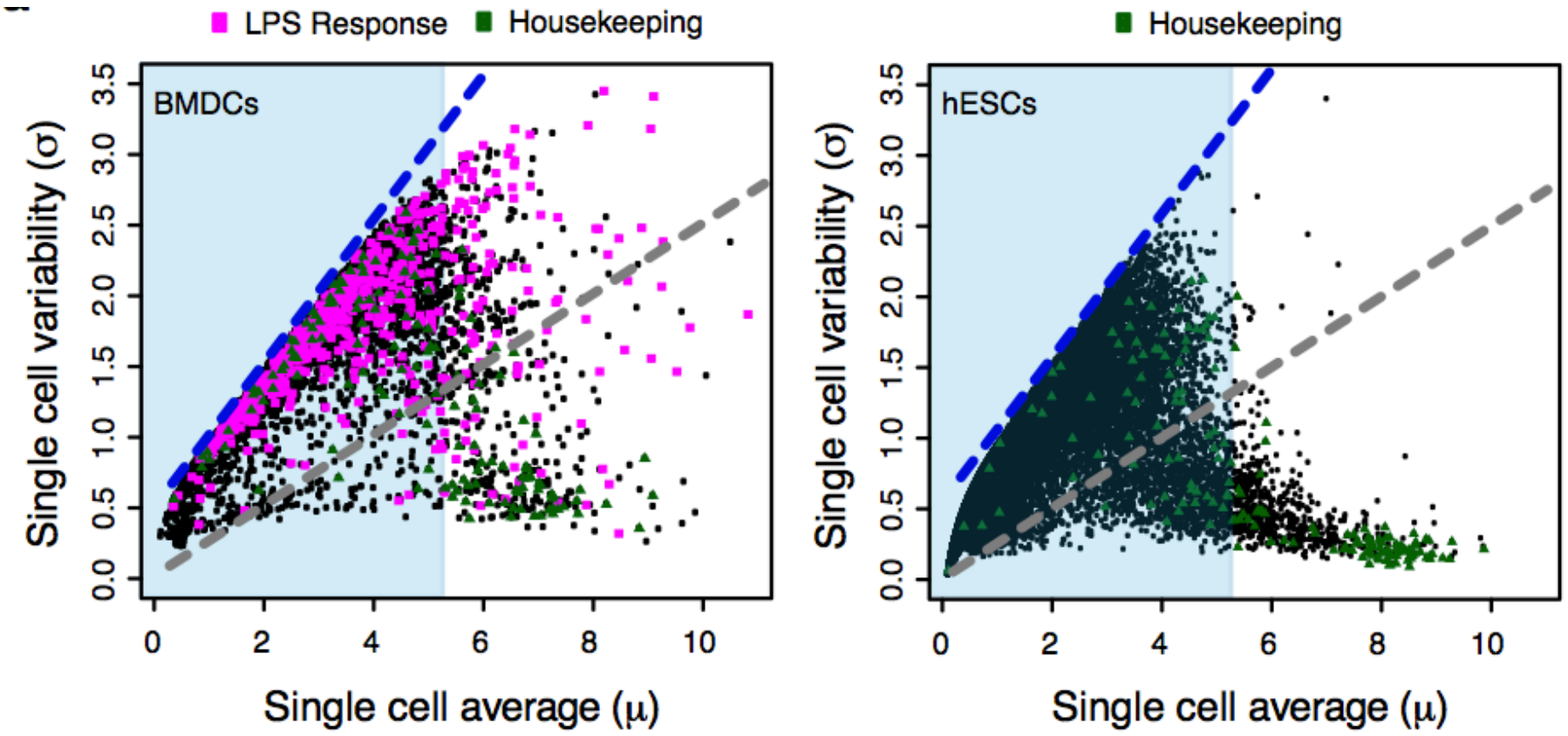
39 Clusters: New Markers That Can Be Validated!



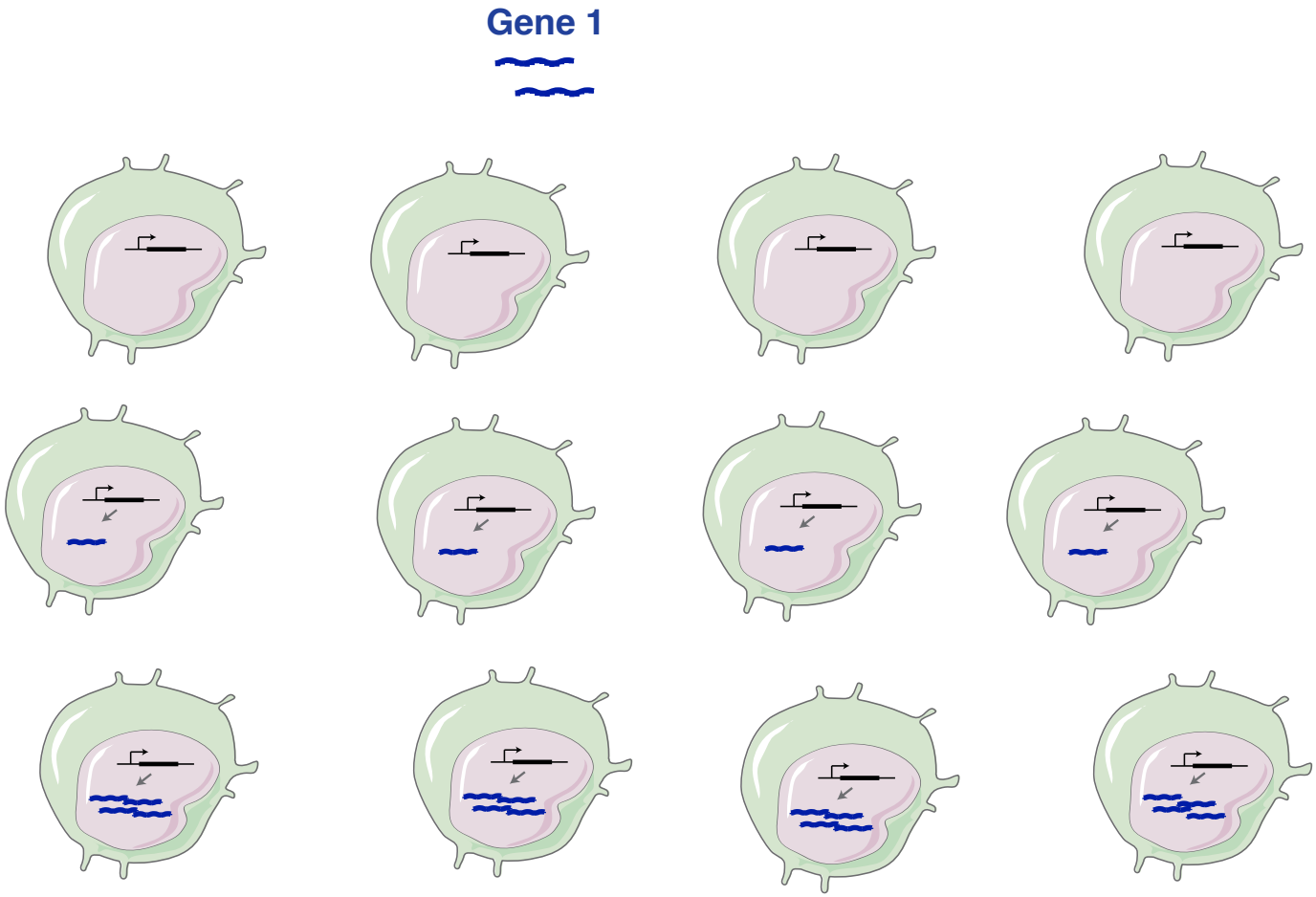
Macosko et al, *Cell*, **161**, 2015

Clustering similar genes

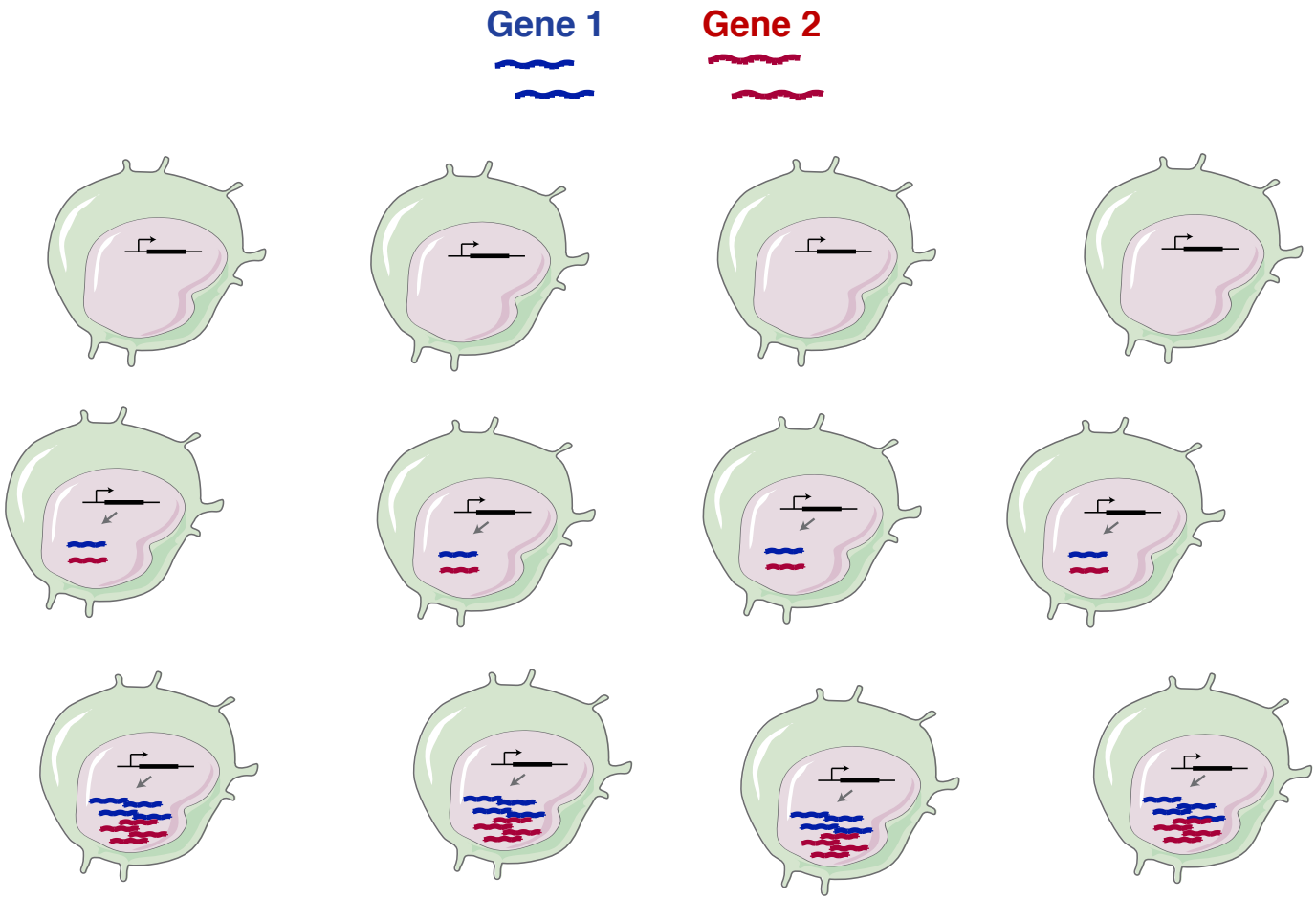
Identifying ‘variable’ genes



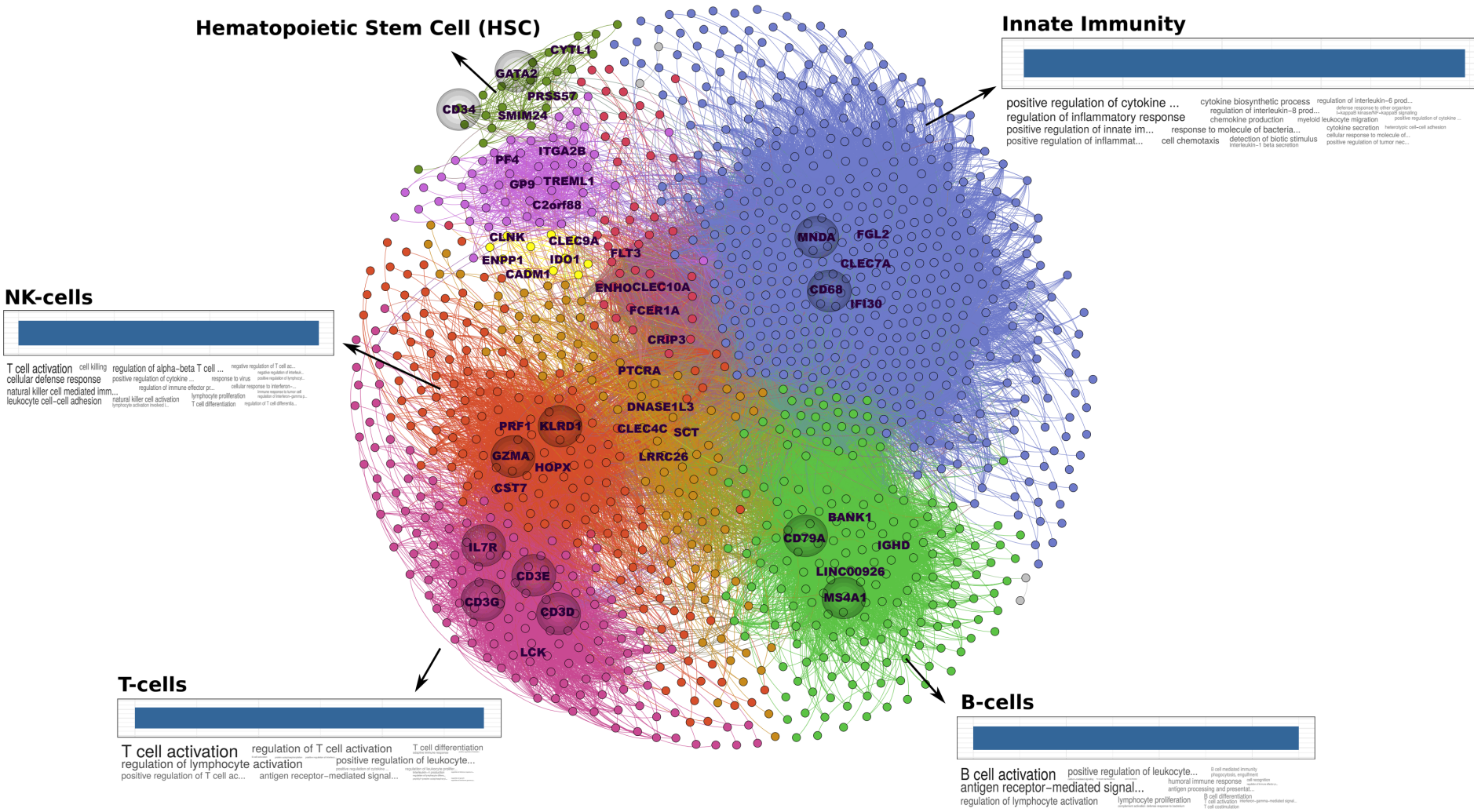
Variation is interesting



Co-variation implies co-regulation

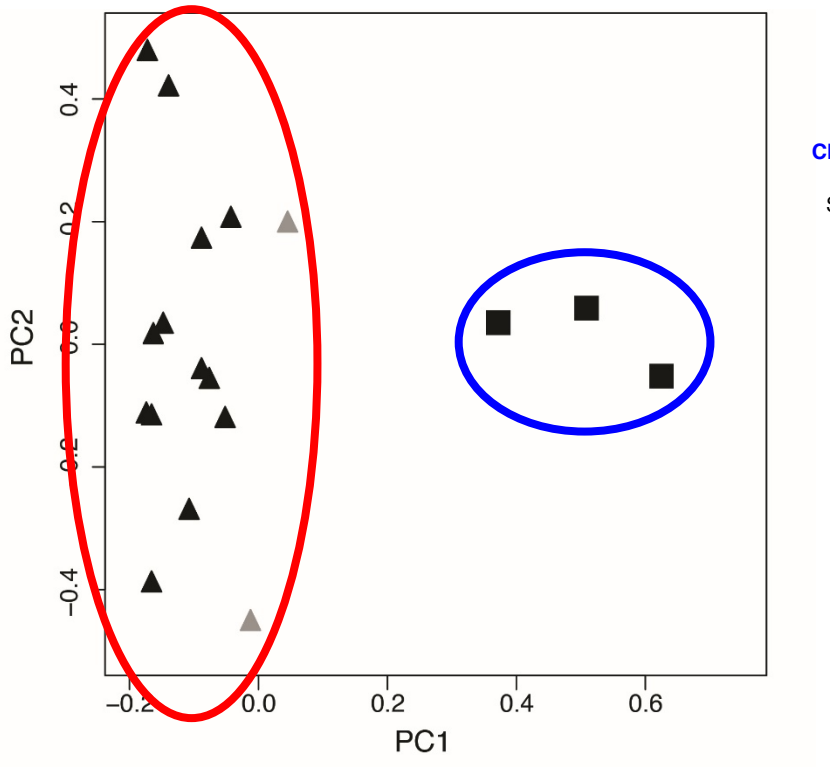


Sparsity-based gene network inference



CO-Dependency network of genes

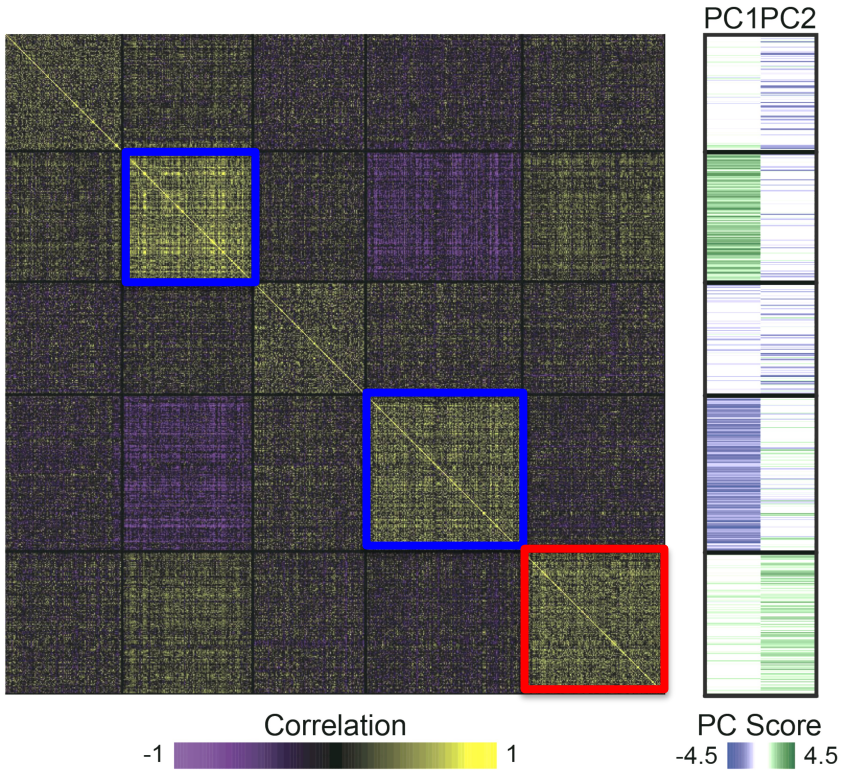
Genes Co-Vary Across Single Cells



Cluster-disrupted
Ccr7, Ccl22,
Serpinb9, Cd83

Maturing
Tnf, Il1a,
Cxcl10, Cd88

Antiviral
Ifit1, Ifih1,
Stat2, Irf7



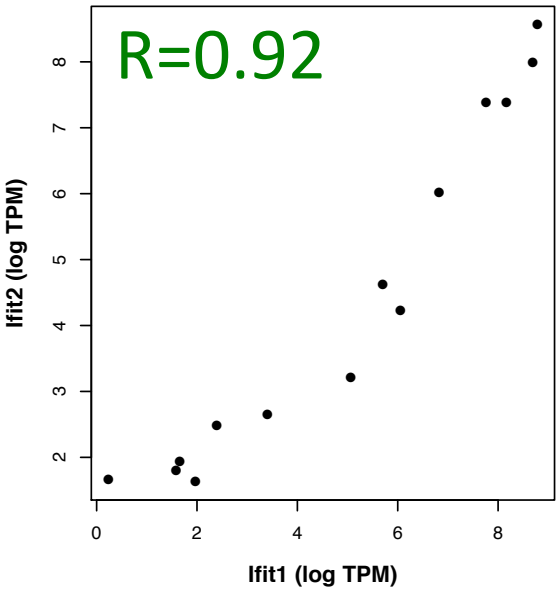
We can uncover cell states and circuits, as well as their markers and drivers, from structures in cell-to-cell variation

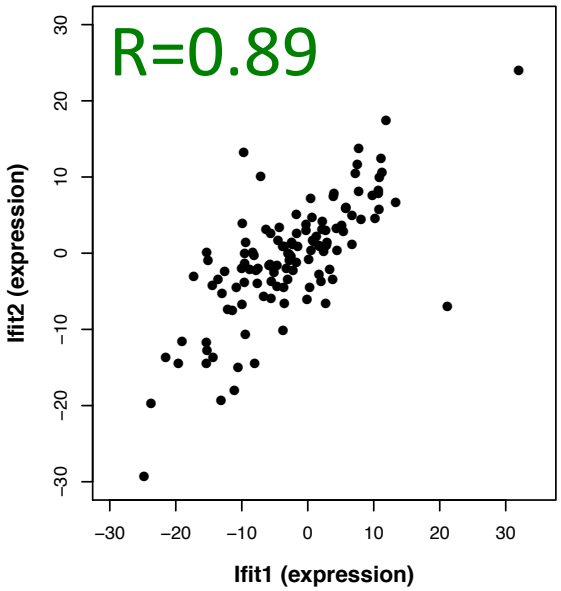
Shalek*, Satija* et al, *Nature*, **498**, (2013)

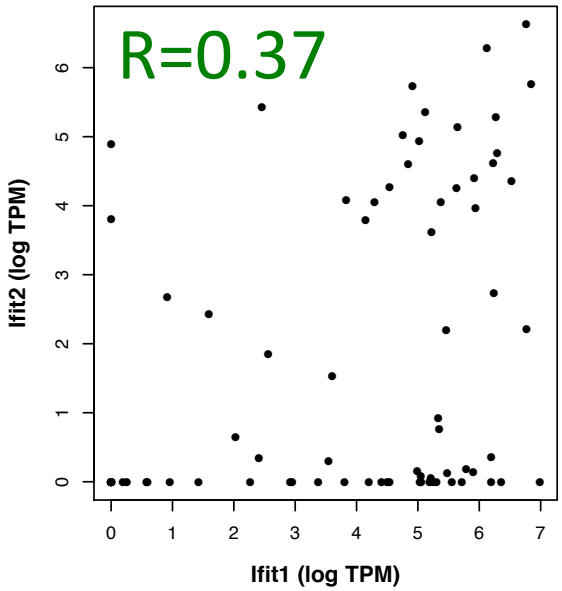
Correlation is not well-suited for single-cell analysis

POPULATIONS

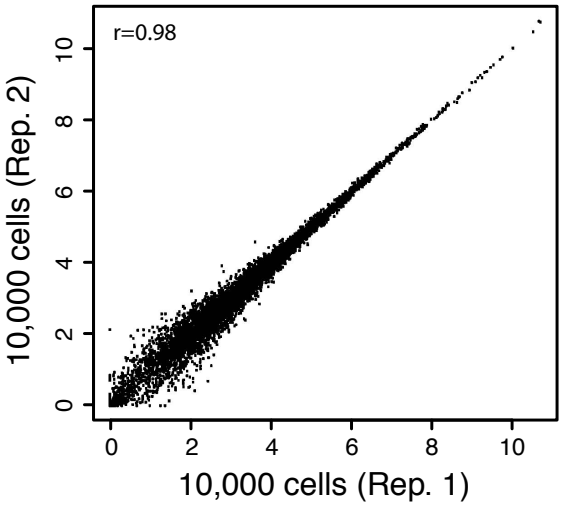
SINGLE CELLS

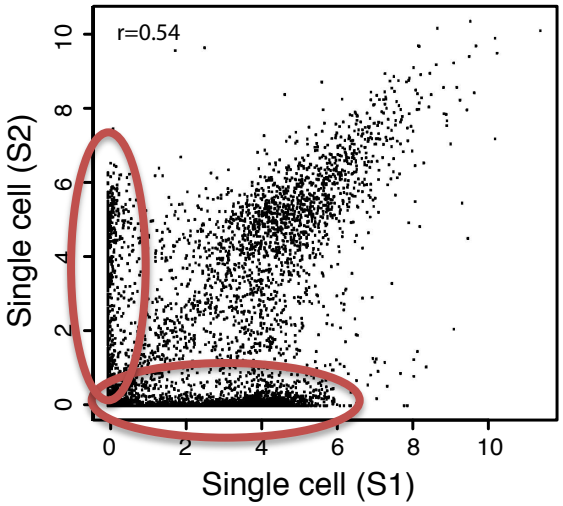


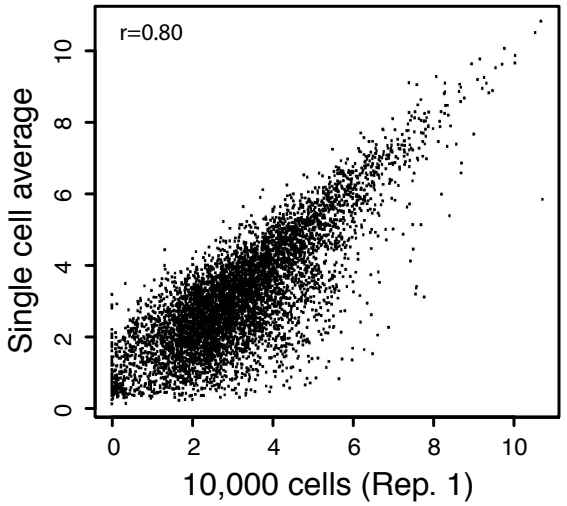




scRNA-Seq data has many many zeros





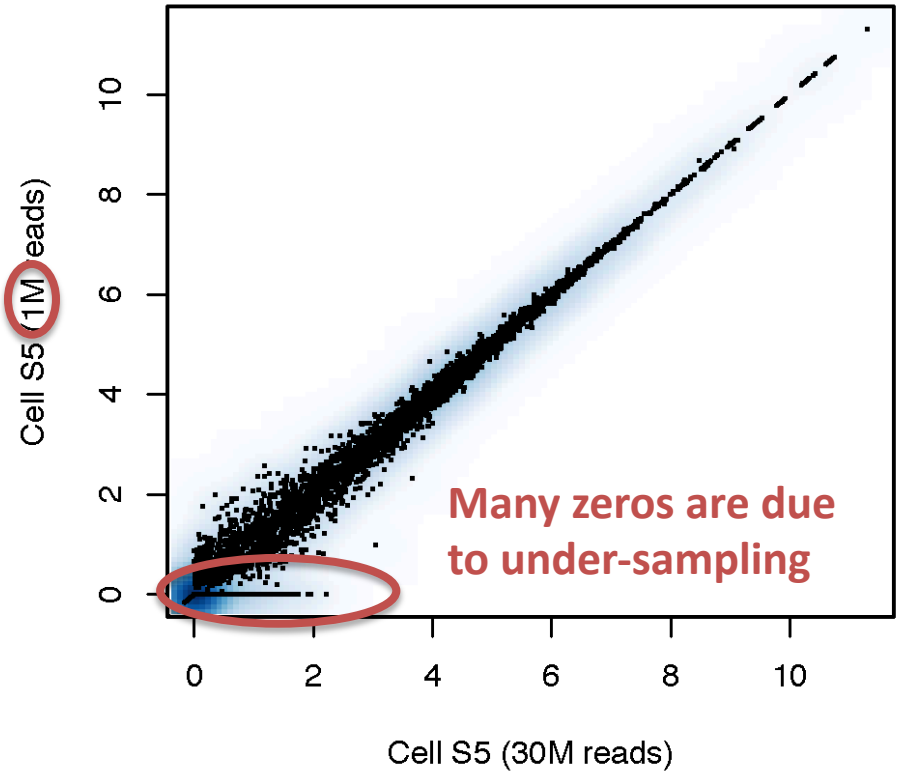


Transcriptome-wide, single cells are very different.

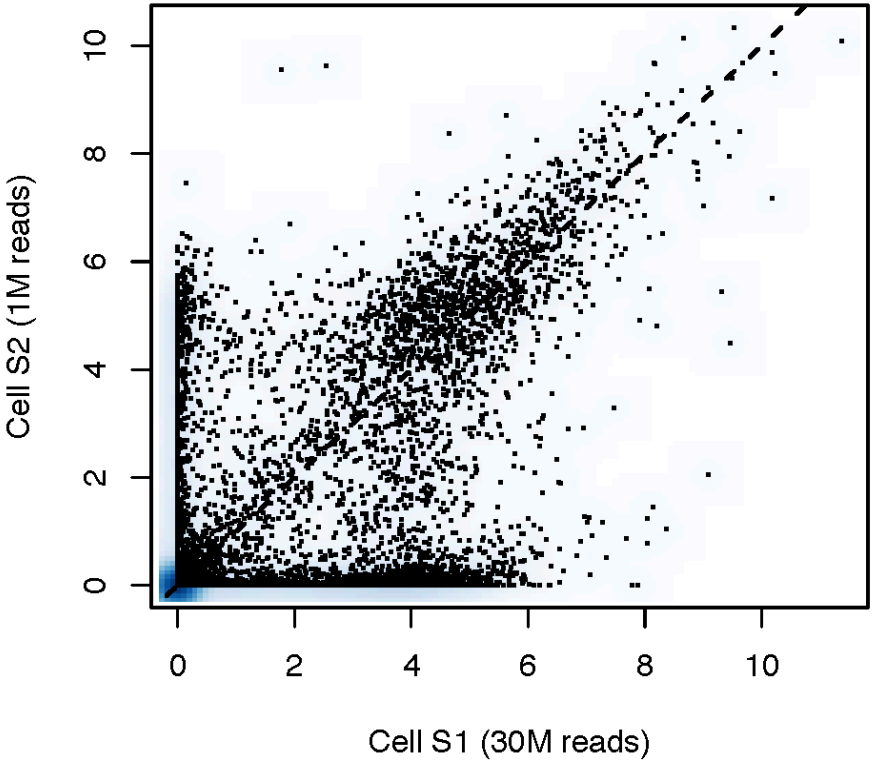
Shalek*, Satija* et al, *Nature*, **498**, (2013)

Variability due to sampling vs. biology

Sampling Bias



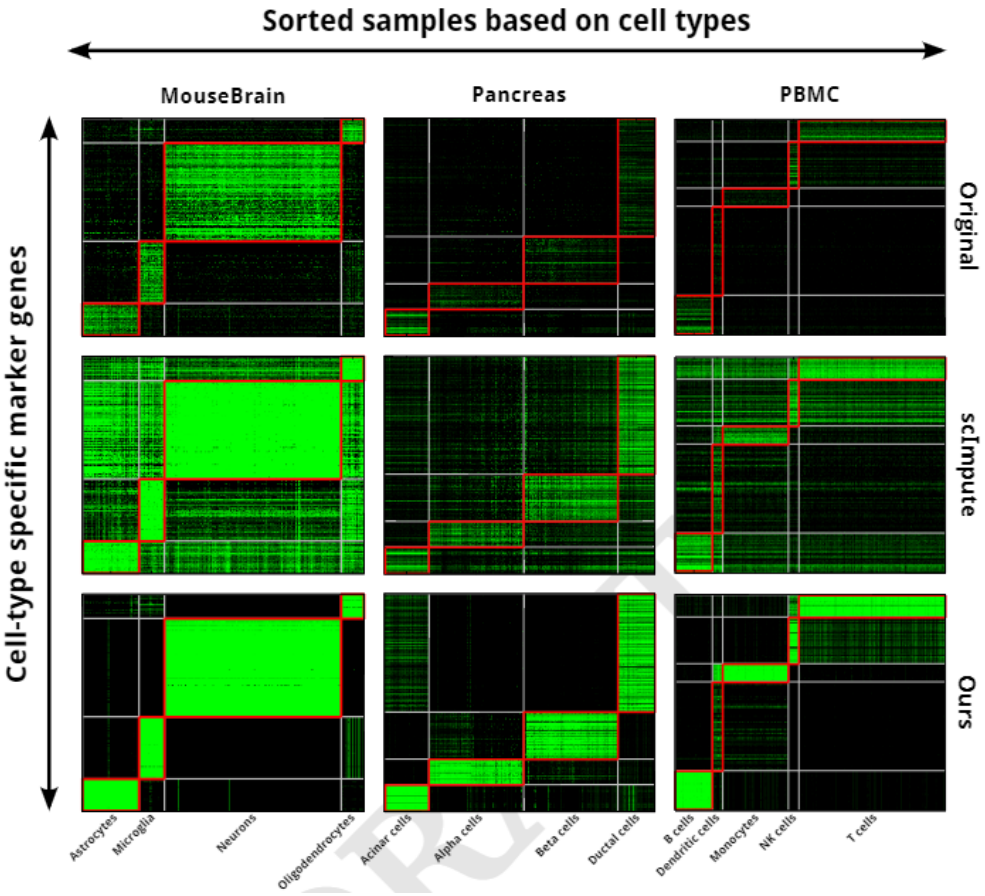
Cellular Variation



Shalek*, Satija* et al., Nature, 2013

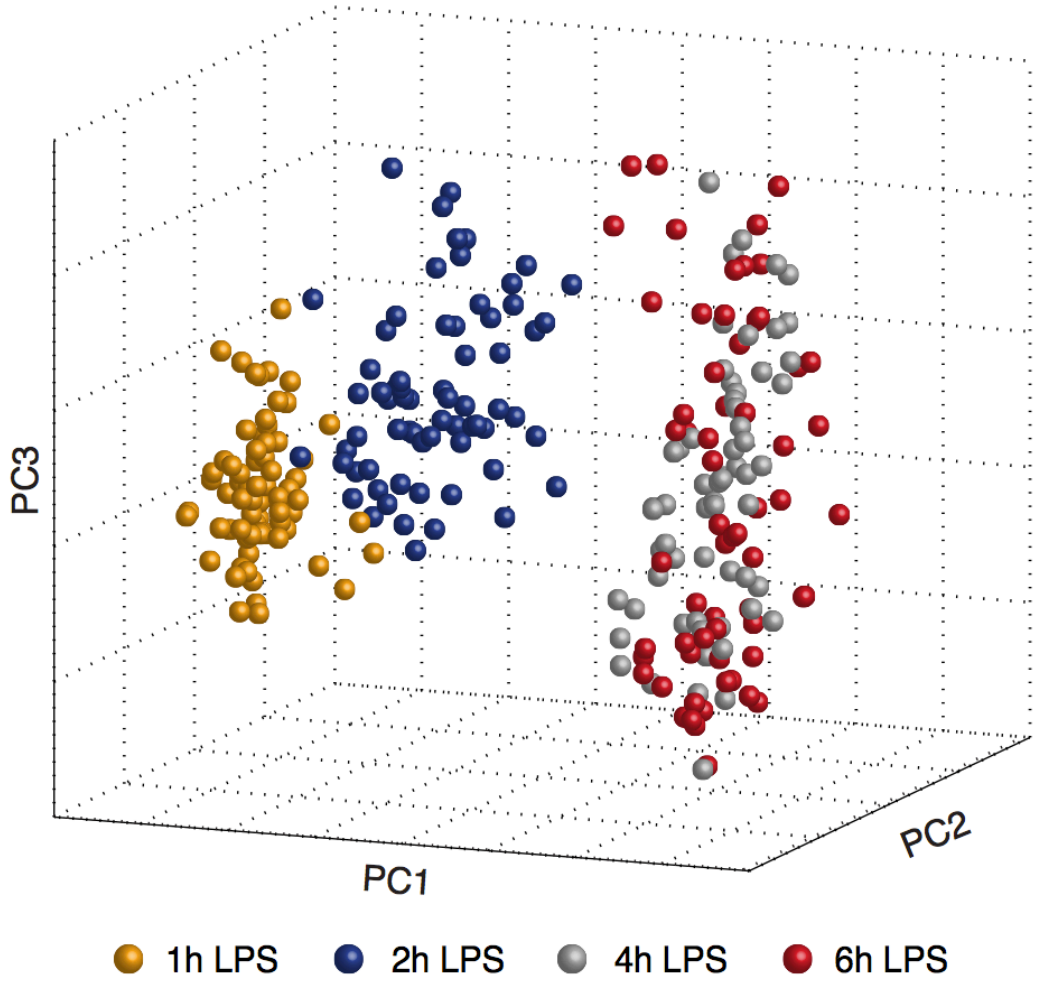
Dimensionality reduction

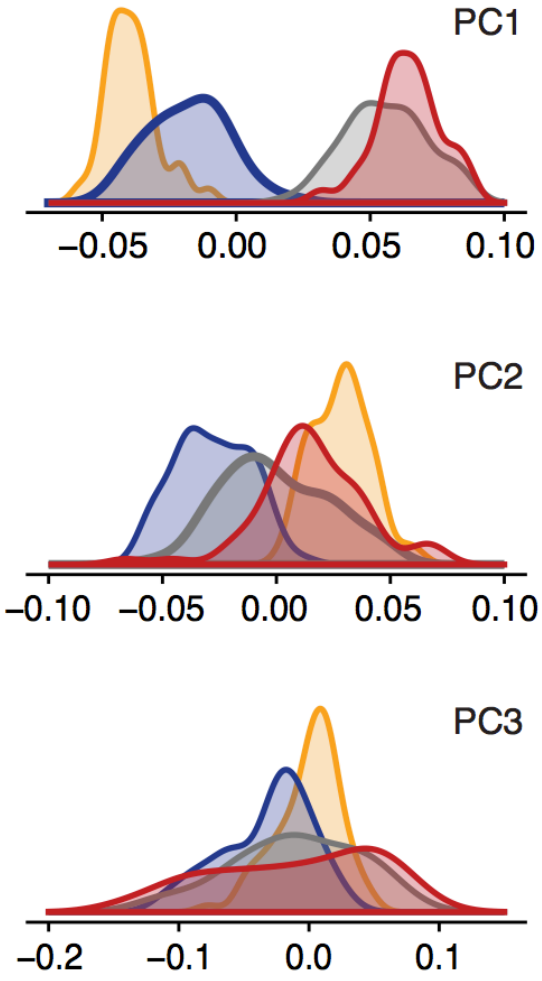
Dimensionality Reduction



- Curse of dimensionality
- Easier to visualize/process
- Reduce noise
- Linear methods: *PCA*
  - Identifying batch/cell cycle effects effects
- Nonlinear methods: t-distributed stochastic neighbor embedding (*t-SNE*)
  - Exploratory data analysis

PCA – 300 cell dataset



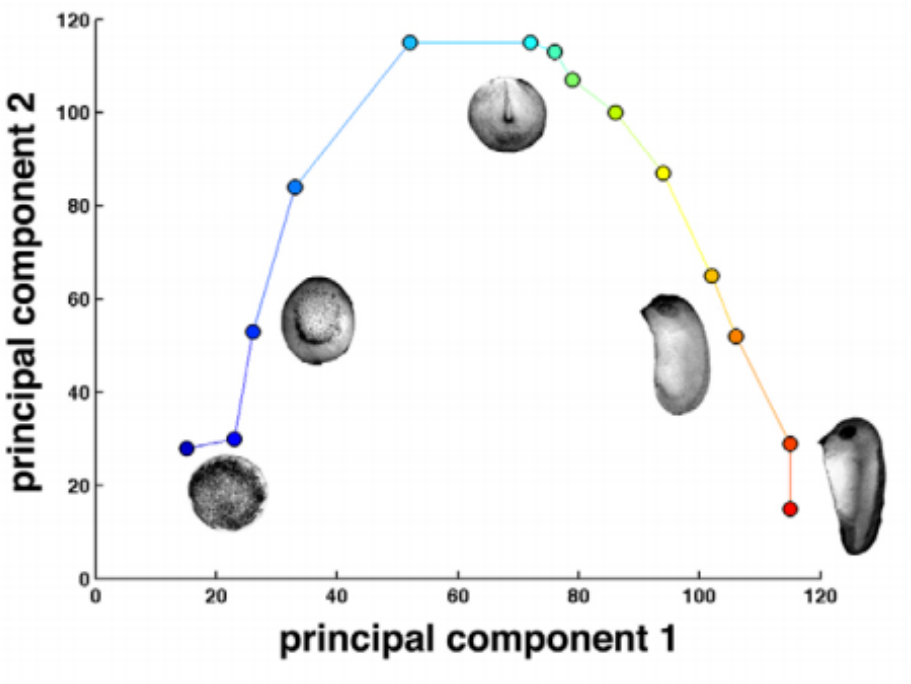


Important consideration for PCA

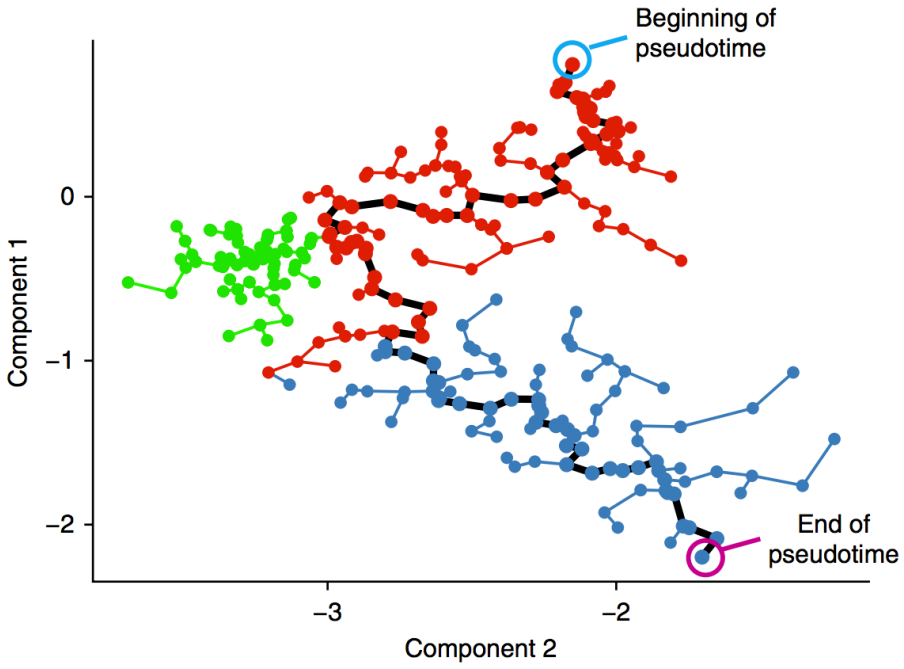
- **Input gene list**
  - Can dramatically alter output
- **Interpretation:**
  - 'Assigning 'biology' or function requires prior knowledge
  - PCs often correlate with technical quality
  - Not all PCs are significant (Chung, Storey, arXiv.org)
- **Limitations/extensions:**
  - PCs represent **linear** combination of individual features

Interpreting dimensionality reduction

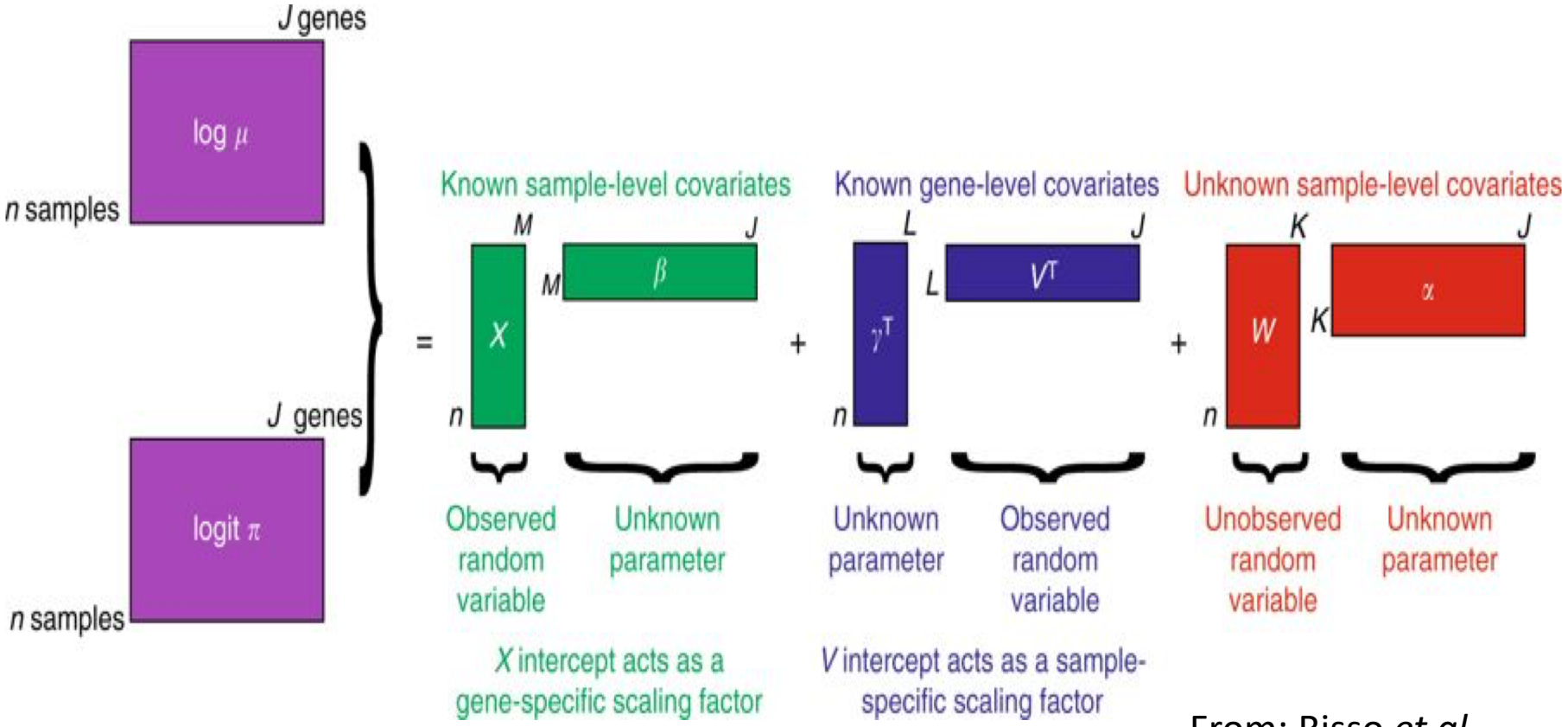
Anavy et al, Development, 2014



Trapnell et al, Nat. Biotech., 2014



Zero-inflated negative binomial model (ZINB-WaVE)

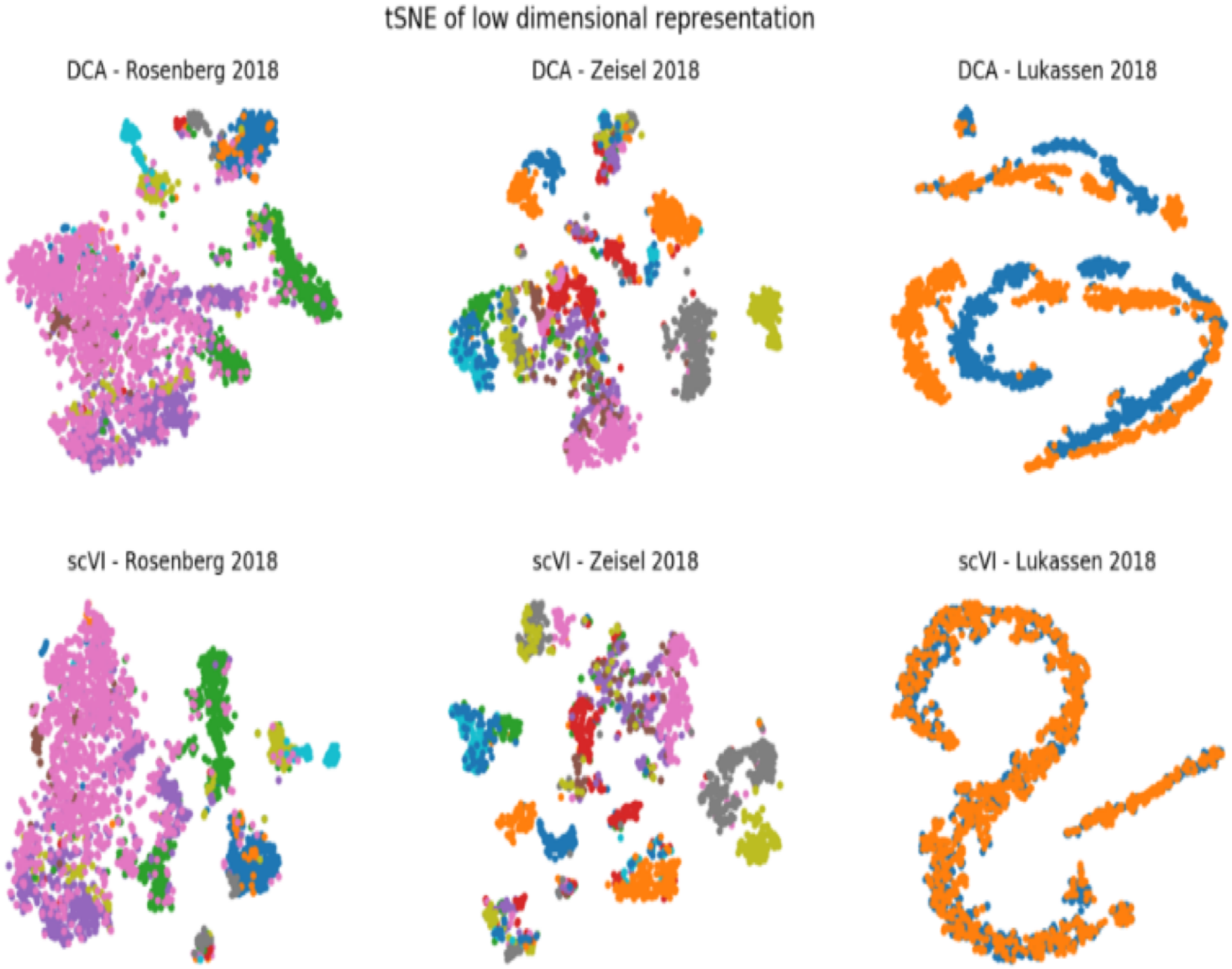


From: Risso *et al., 2018*

- A generalized linear factor analysis model

Dropping factor analysis in favor of deep autoencoders

- Single-cell Variational Inference (scVI)
- Deep count autoencoder (DCA)



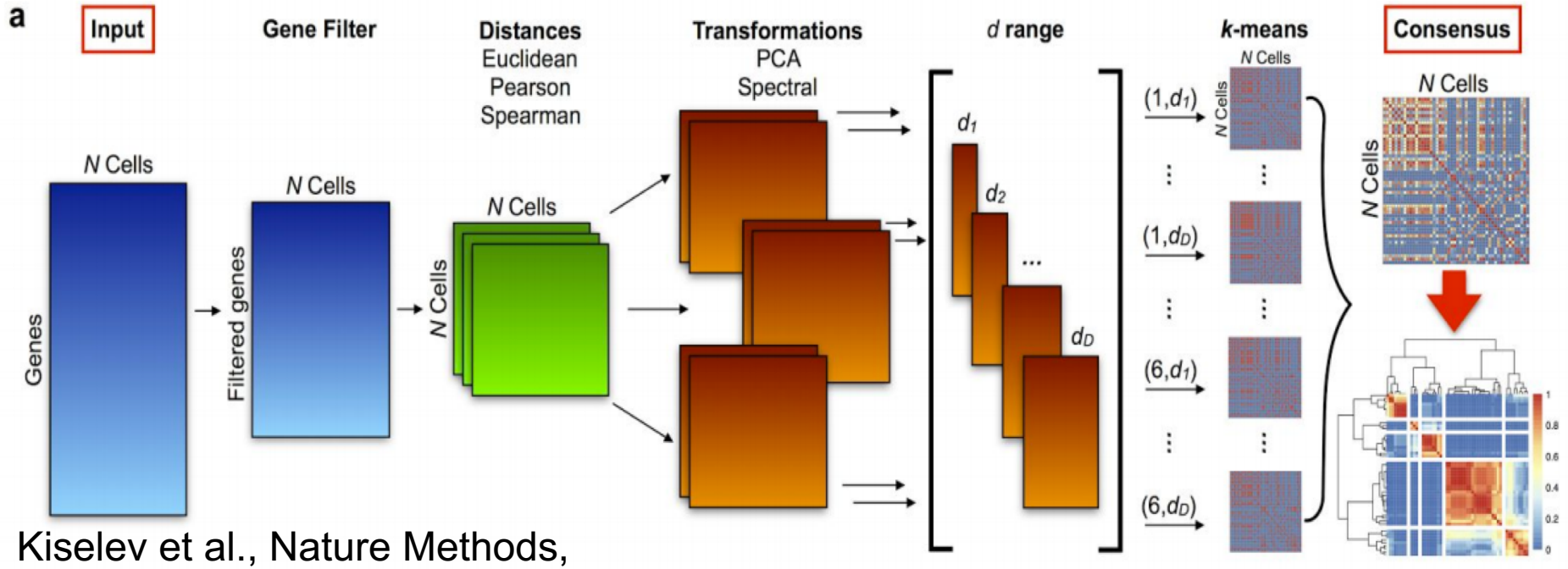
From: V. SVENSSON, 2018

Dinstinguishing different cell types

Discrete cell type identification

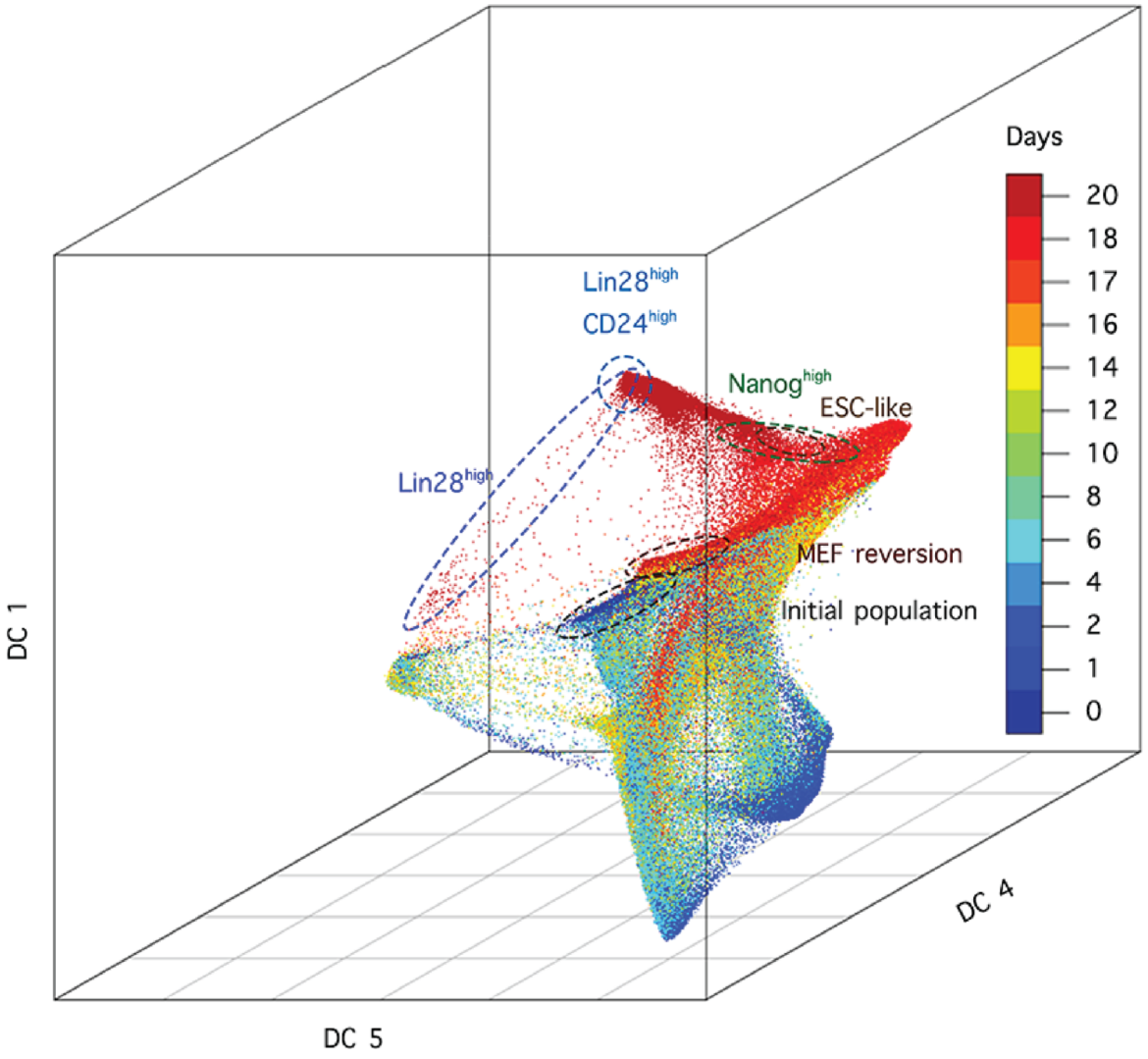
- **Based on traditional clustering approaches:**
  k-means, hierarchical, and graph-based clustering techniques
- **tSNE + k-means (traditional)**
- **SINCERA (Guo et al. 2015)**
  - Based on hierarchical clustering
  - Data is converted to z-scores before clustering
- **SNNCliq (C. Xu and Su 2015)**
  - Identifies the k-nearest-neighbours of each cell according to the distance measure.
  - Clusters are defined as groups of cells with many edges between them using a “clique” method.
- **PCAReduce (žurauskienė and Yau 2016)**
  - Combines PCA, k-means and “iterative” hierarchical clustering.
  - Starting from a large number of clusters pcaReduce iteratively merges similar clusters
  - After each merging event it removes the principle component explaning the least variance in the data.
- **SC3 (Kiselev et al. 2017)**
  - Based on PCA and spectral dimensionality reductions
  - Utilises k-means
  - Additionally performs the consensus clustering

Single-Cell Consensus Clustering (SC3)



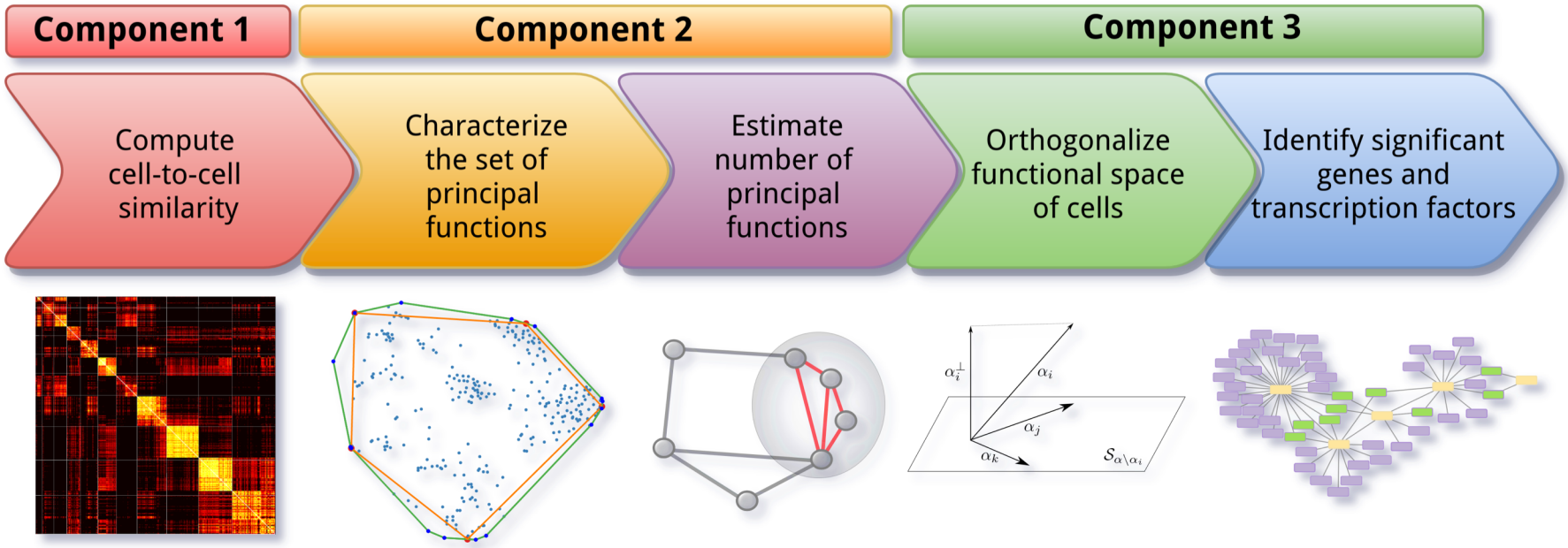
Kiselev et al., Nature Methods, 2017

Continuous cell states: diffusion map



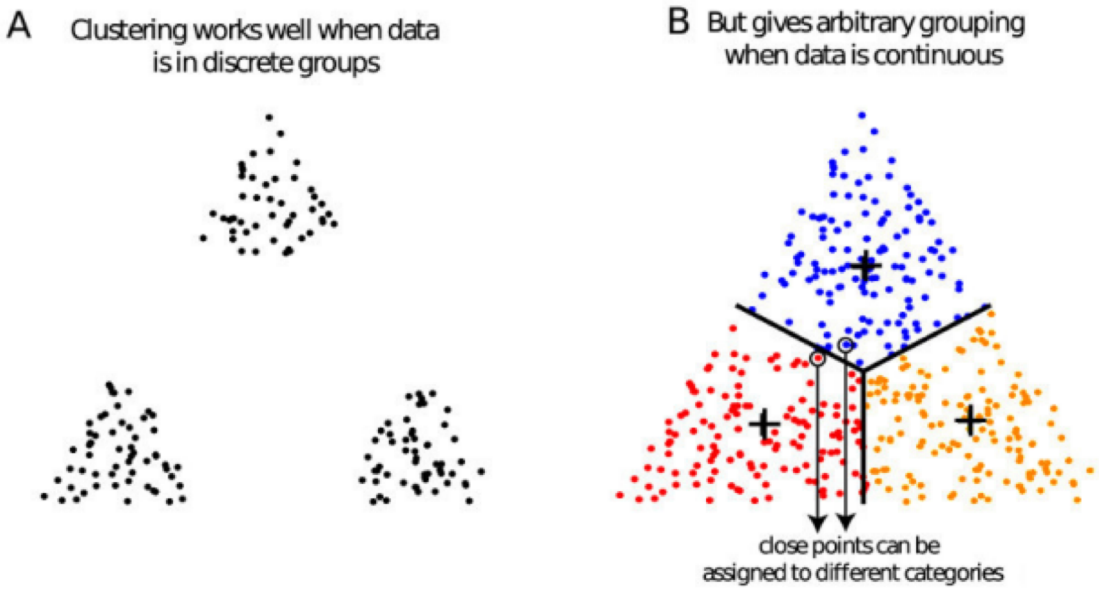
Angerer *et al*, Bioinformatics 2016

Archetypal-analysis for Cell type indentificaTION (ACTION)



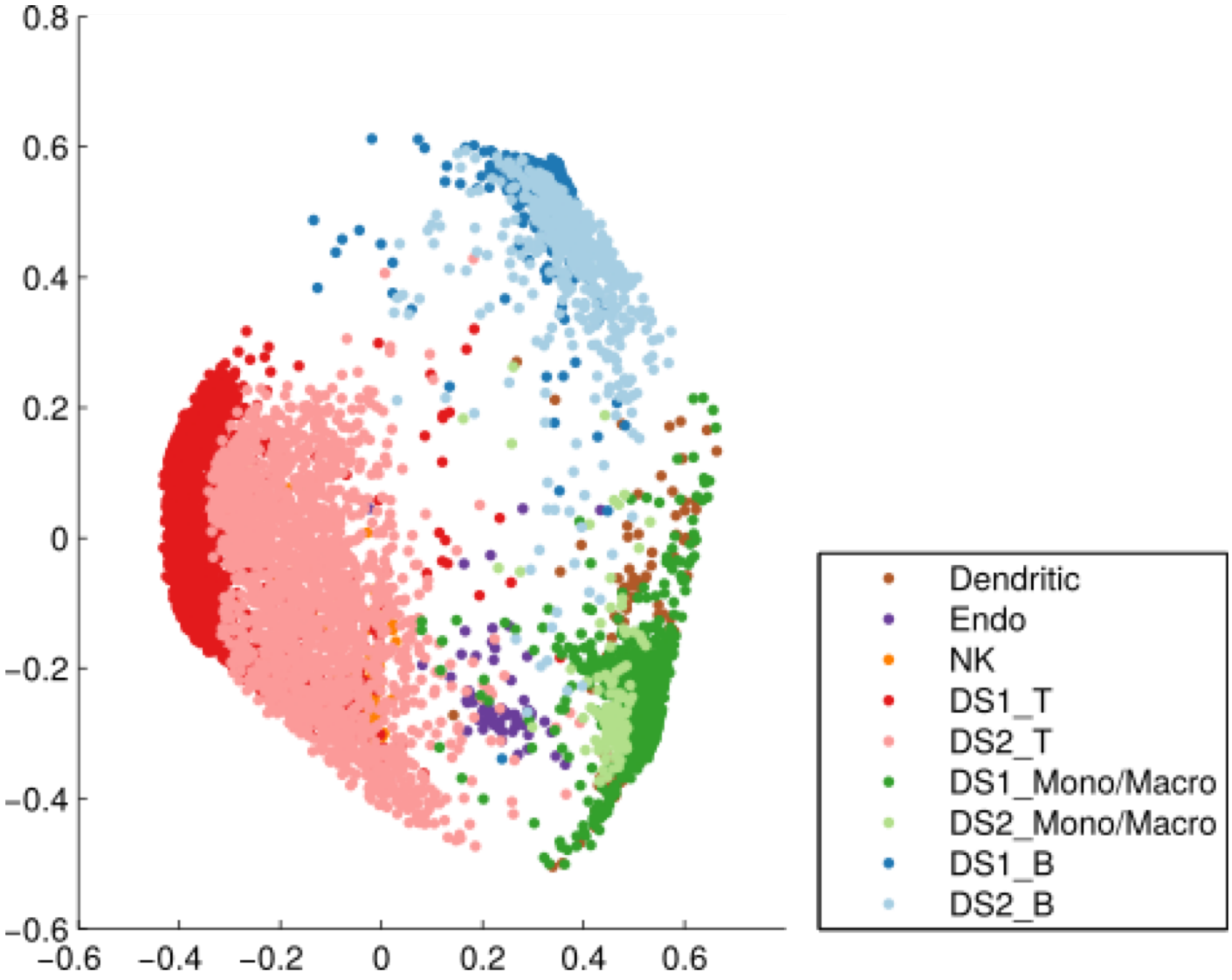
Mohammadi *et al.*, BioRxiv 2016, Nature Communications, under review

Combine discrete + continuous: archetype analysis



Hart *et al*, Nature Methods 2015

Matching cell types across datasets



Alignment of PBMC vs. Tumor scRNA

Multi-resolution analysis

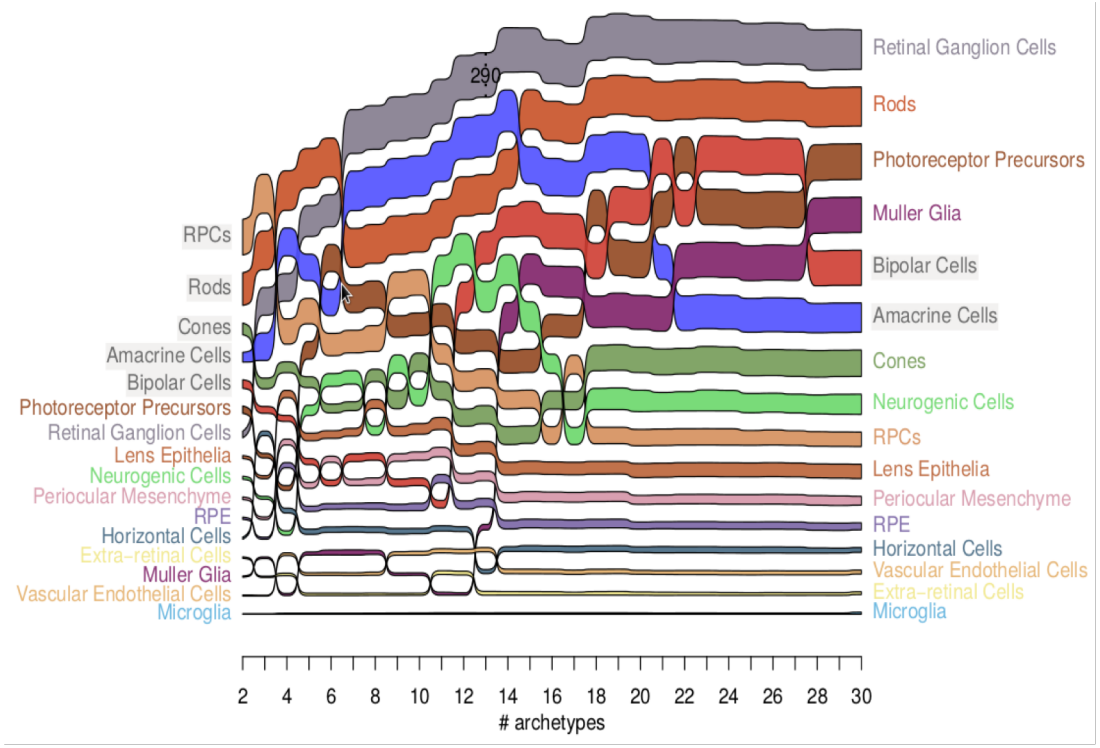
ACTIONet

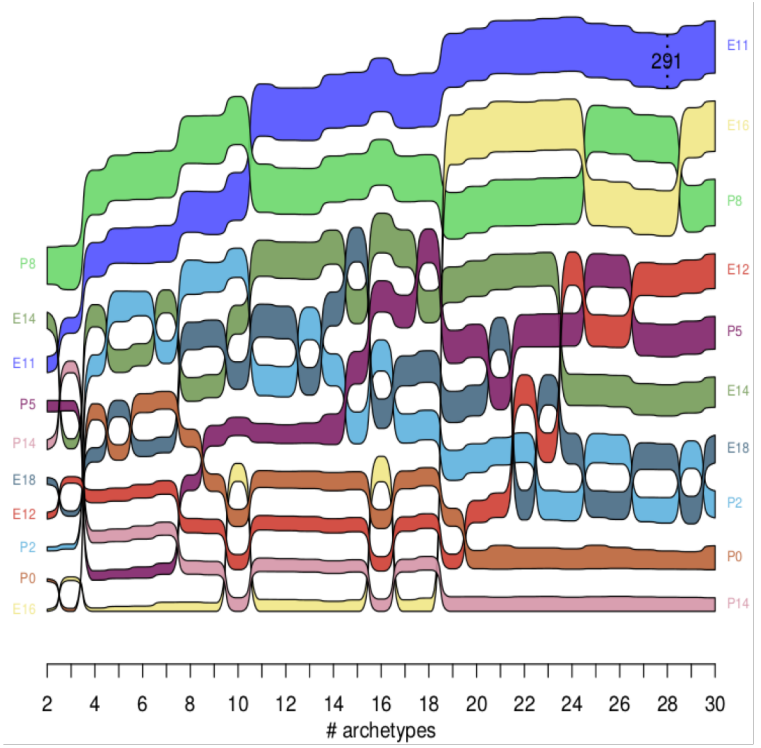
Main issues with parametric methods

How many archetypes?
How many factors?
How many clusters?

Optimal number of factors differs by celltype/age

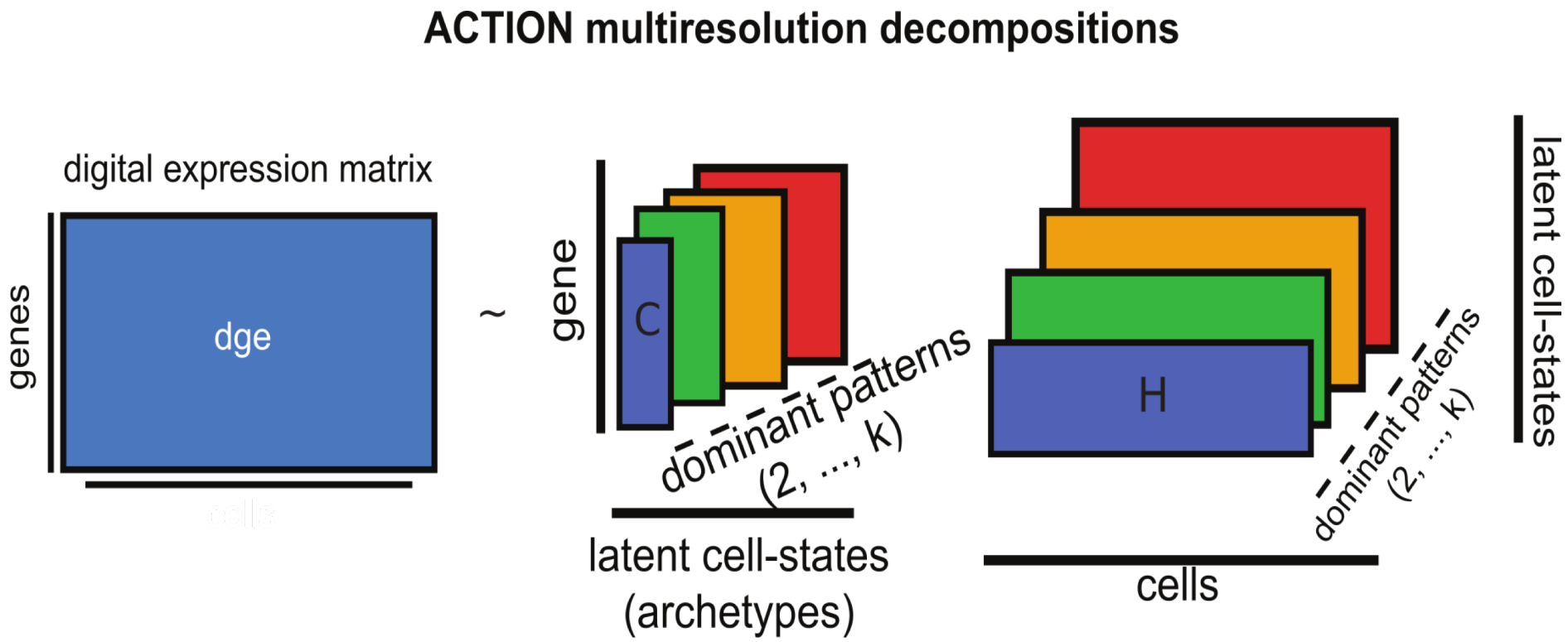
(Ex.: Mouse retina development -- similar results with other species/tissues)





Choosing one “optimal” k is dominated by the major cell type
(defeating the whole purpose of single-cell analysis)

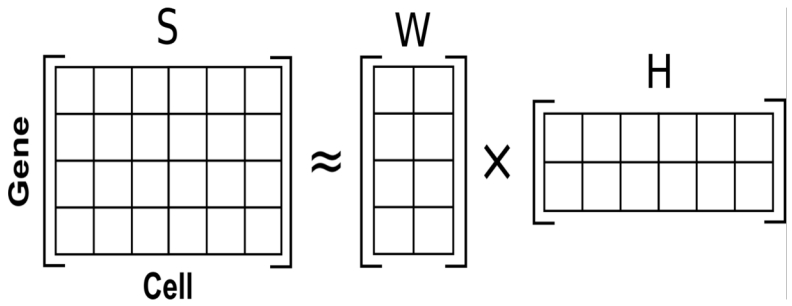
Recently developed method: ACTIONet



Complementary approaches

Factor analysis

ACTION



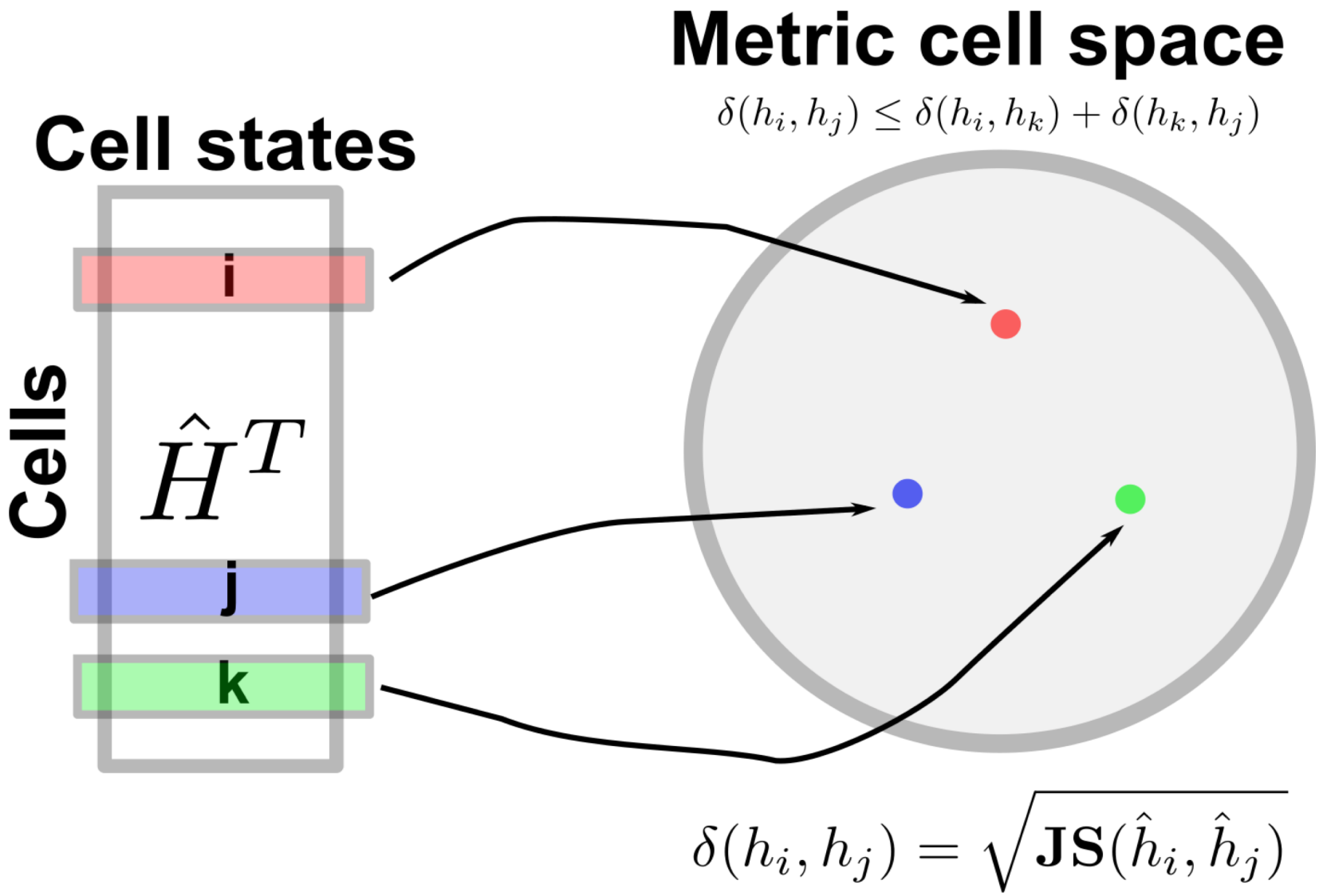
- Identifies hidden cell states/gene programs
  - Biological
  - Technical
- Deconvolves complex biological processes
- Uncovers discriminating/marker genes

Network analysis

- Reconstruct the topography of cell space
- Rich set of graph-based algorithms
  - Visualization (UMAP)
  - Clustering (Louvain/Leiden)
  - Imputation (PageRank)

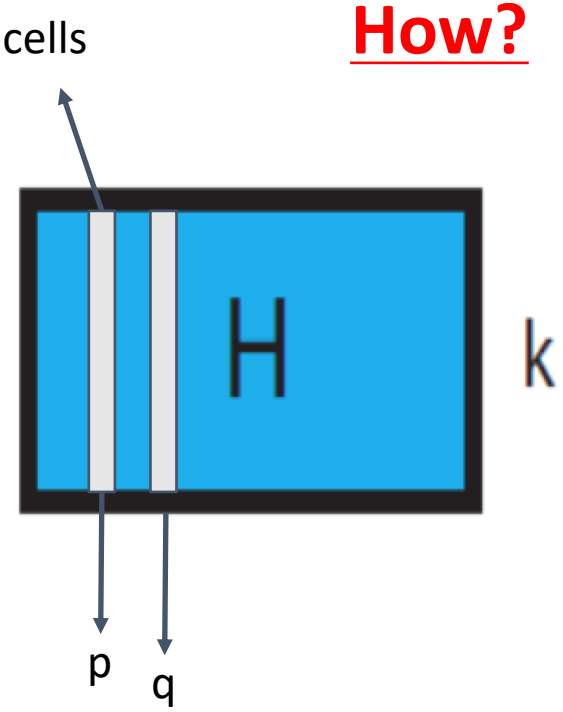
ACTIONet

Step 1: Define a metric cell space



Step 1: Define a metric cell space

How?



Kullback–Leibler divergence

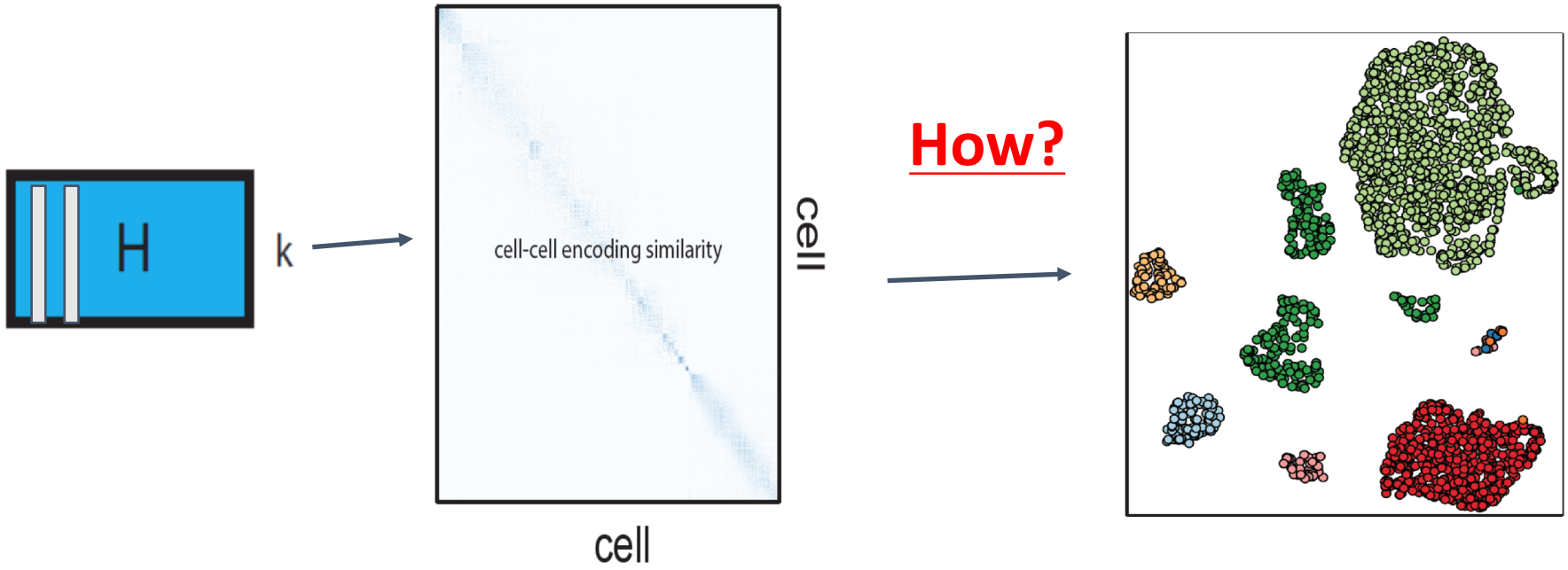
$$\mathbf{KLD}(p \parallel q) = \sum x_x p(x) \mathbf{log}\left(\frac{q(x)}{p(x)}\right)$$

Jensen-Shannon divergence

$$\mathbf{JSD}(p \parallel q) = \frac{1}{2}\{\mathbf{KLD}(p \parallel m) + \mathbf{KLD}(q \parallel m)\}, m = \frac{1}{2}(p+q)$$

Square-root of JSD is a metric (we love metric space ... Triangle inequality rocks! => Efficient proximity search)

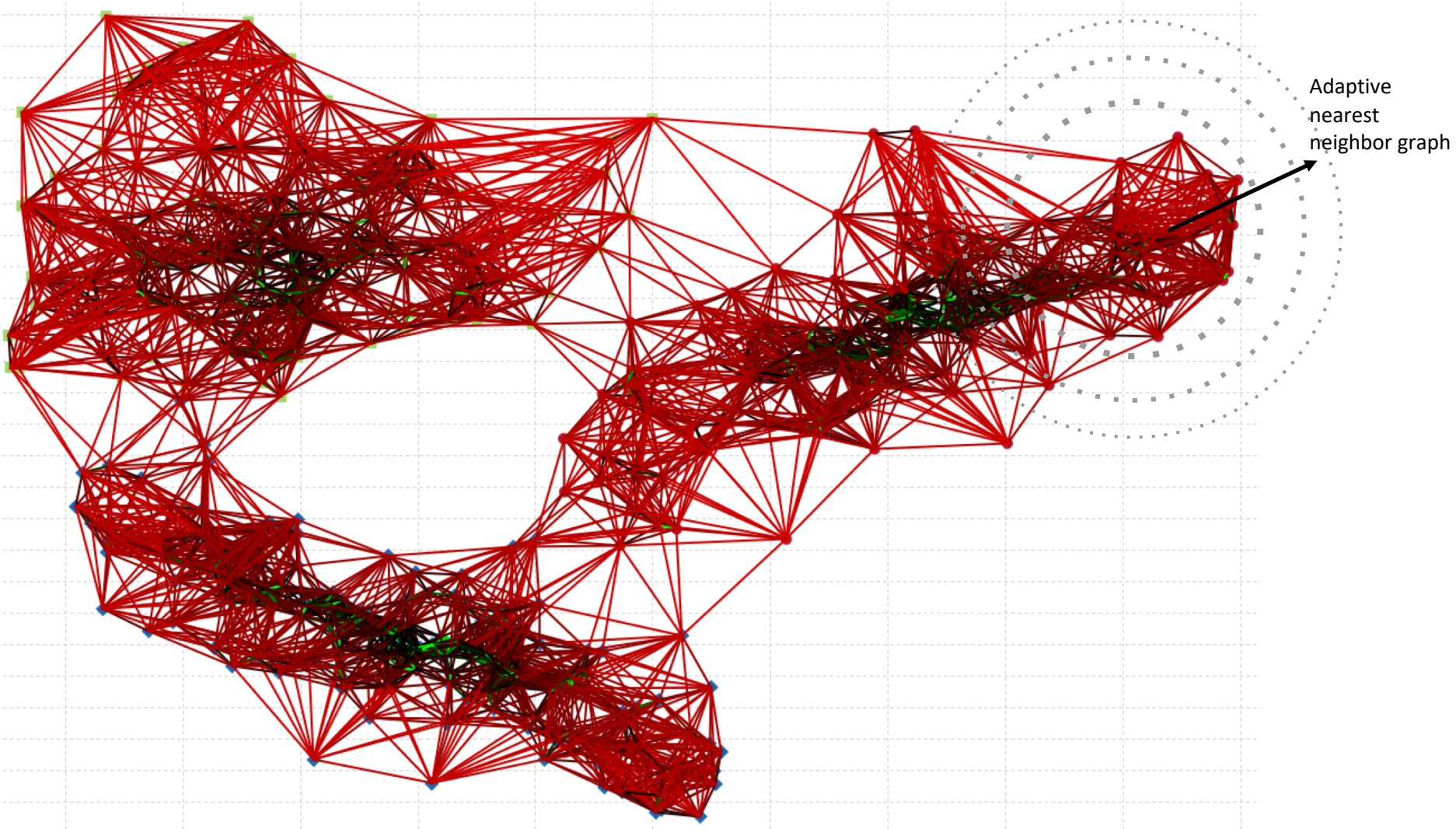
Step 2: Construct a network representation of the cell space



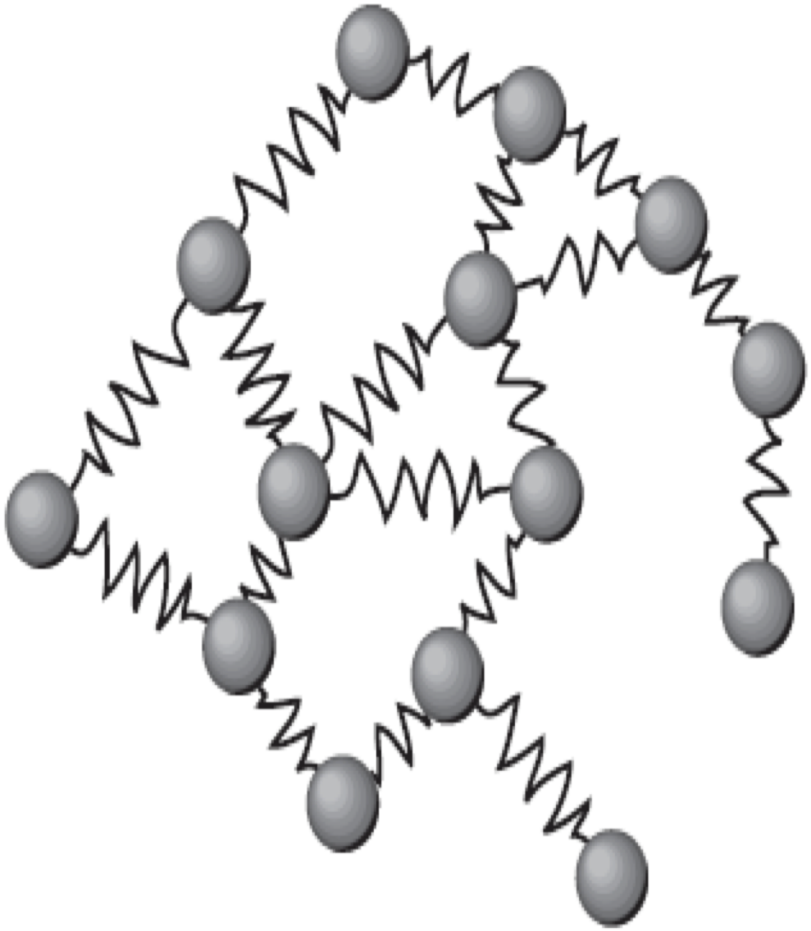
Density-dependent adaptive nearest neighbor graph

- Uses *k*-nearest neighbor* algorithm
- Automatically identifies an optimal number of nearest neighbors for each cell
  - Depends on the heterogeneity of the neighbors

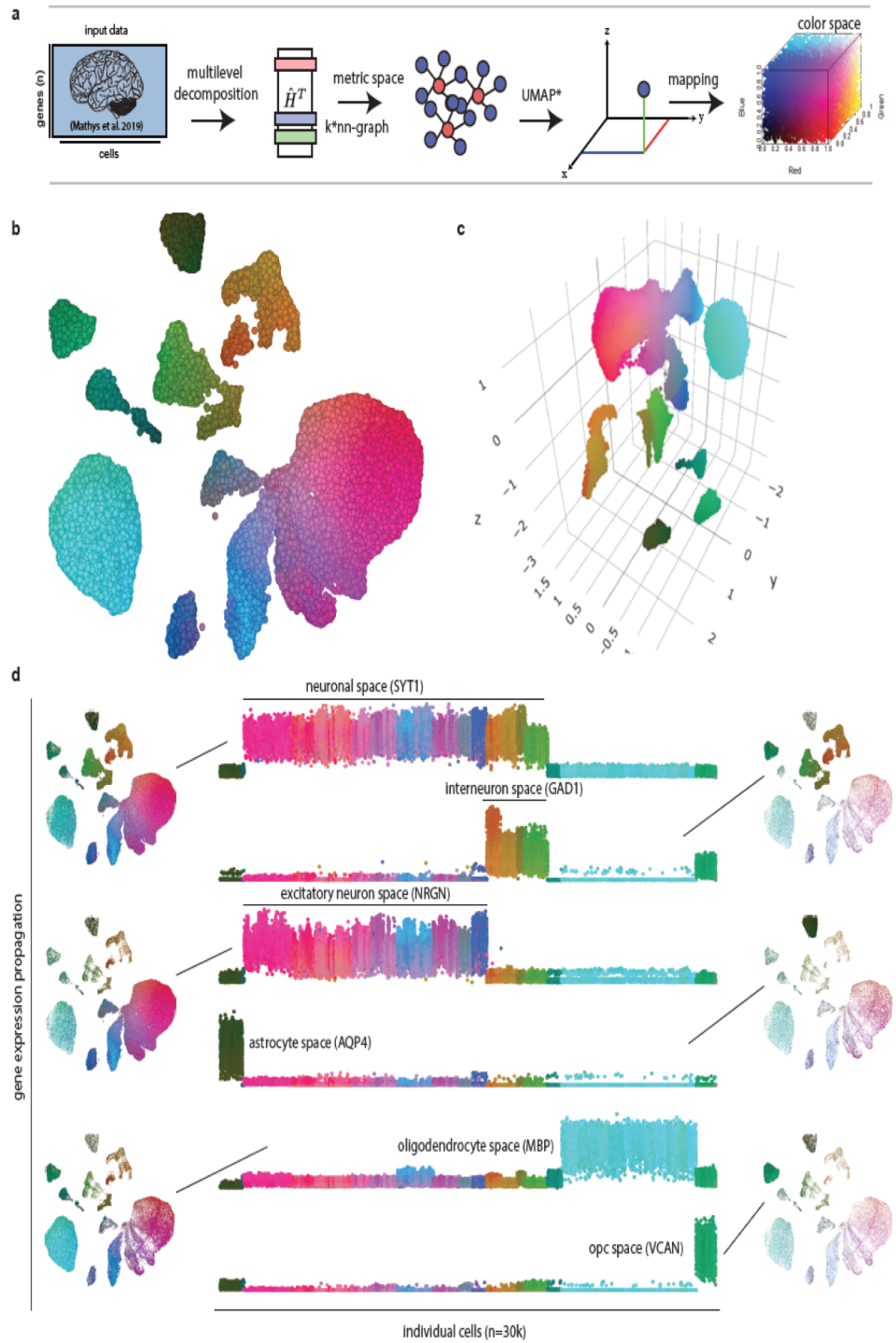
Step 2: Construct a network representation of the cell space



Step 3: Visualize cell-cell network (layout)



- Adopted from UMAP and reimplemented to work with the ACTIONet graph
- Force-directed layout
  - Stochastic-gradient descent (SGD)-based

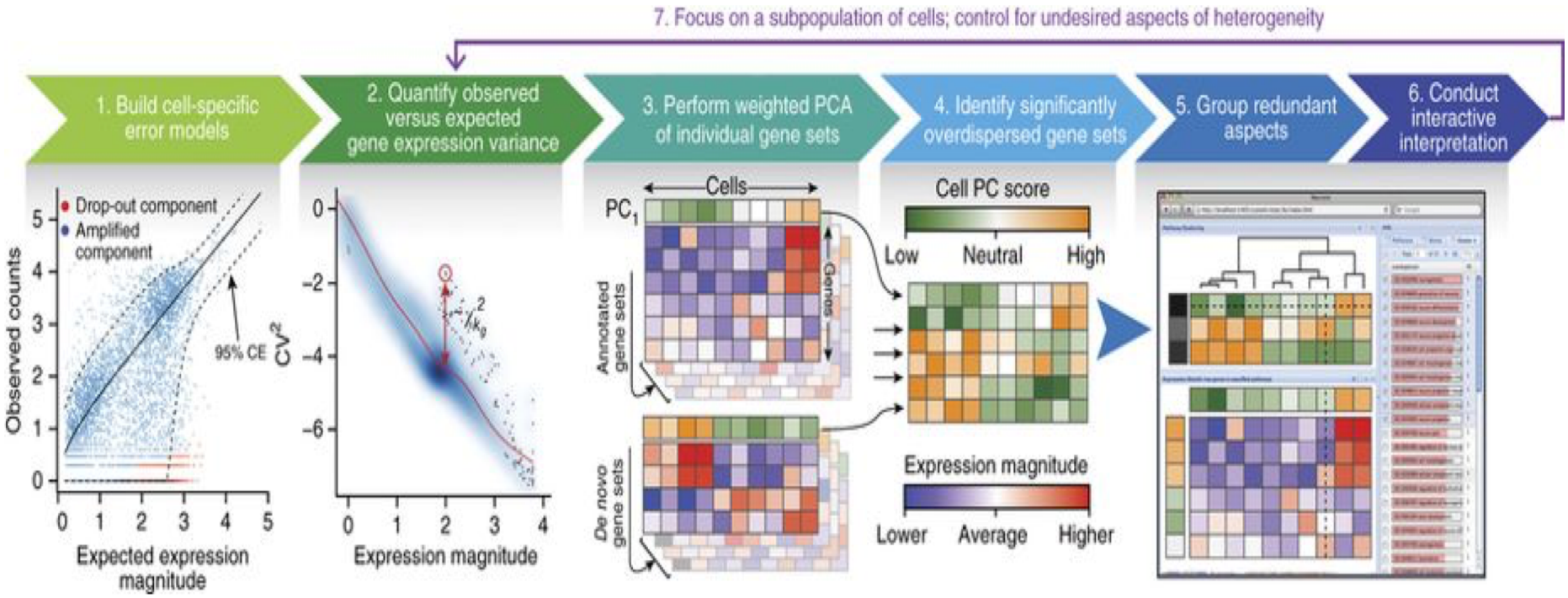


Step 4: Color-coding cells

- Idea: Use *de novo* coloring to fill the gap between 2D and 3D embeddings
- Projecting 3D coordinates onto a Perceptually uniform color space
  - CIE L*a*b*

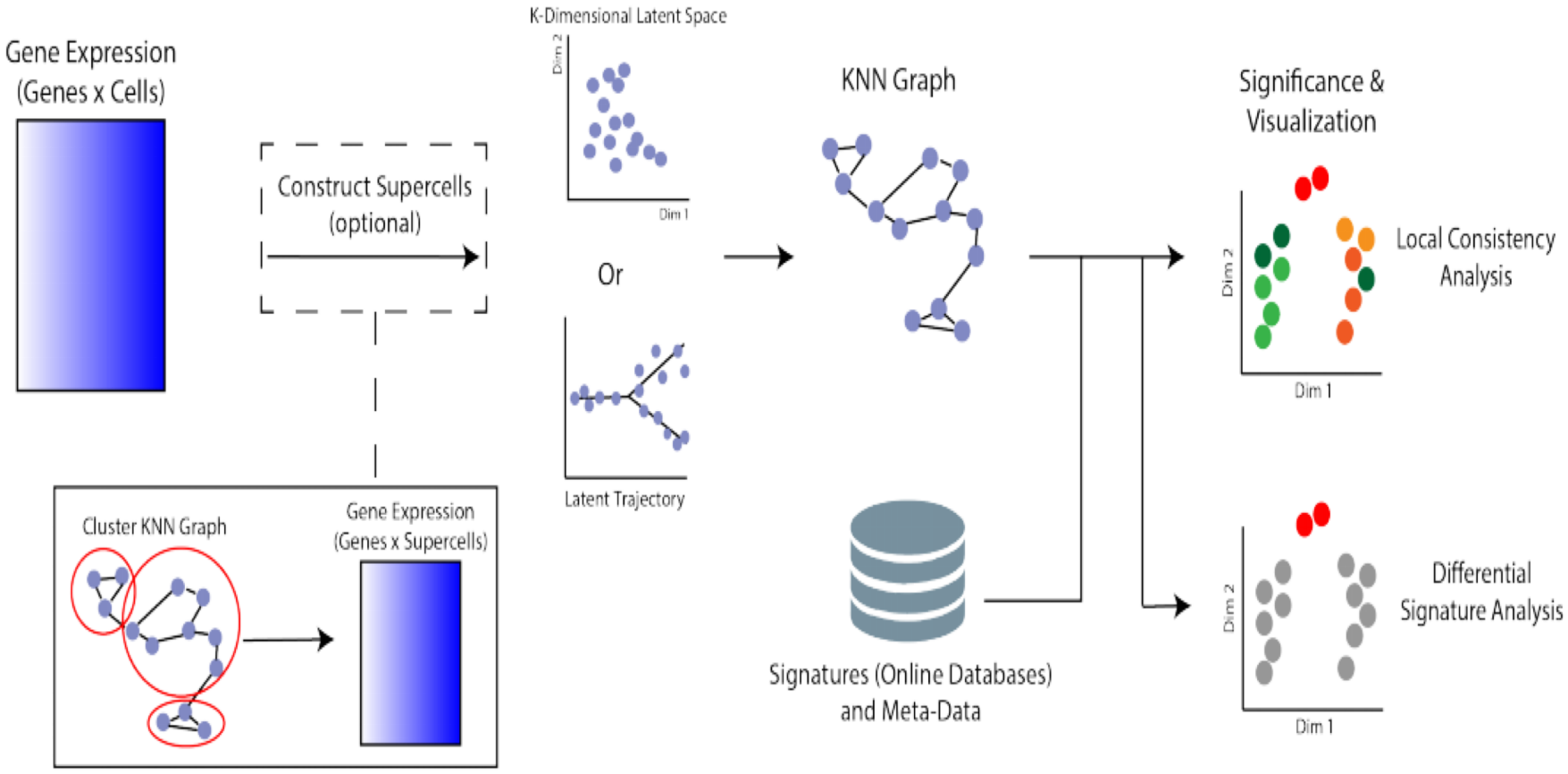
Interpreting cell-to-cell variabilities using known genesets/pathways

Pathway and gene set overdispersion analysis (PAGODA)



From: Fan *et al.,* 2016

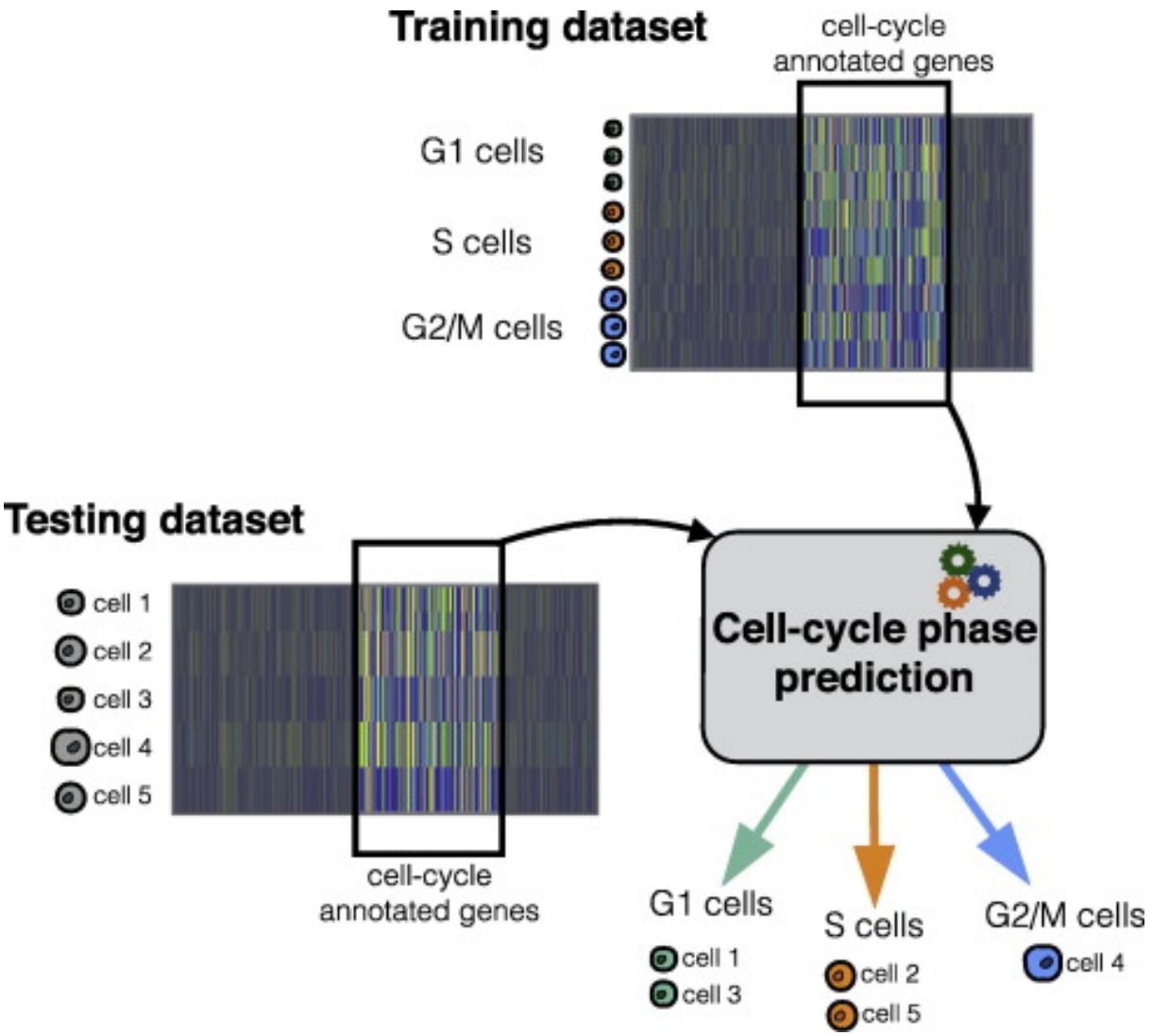
VISION method



From: DeTomaso *et al.*, 2018

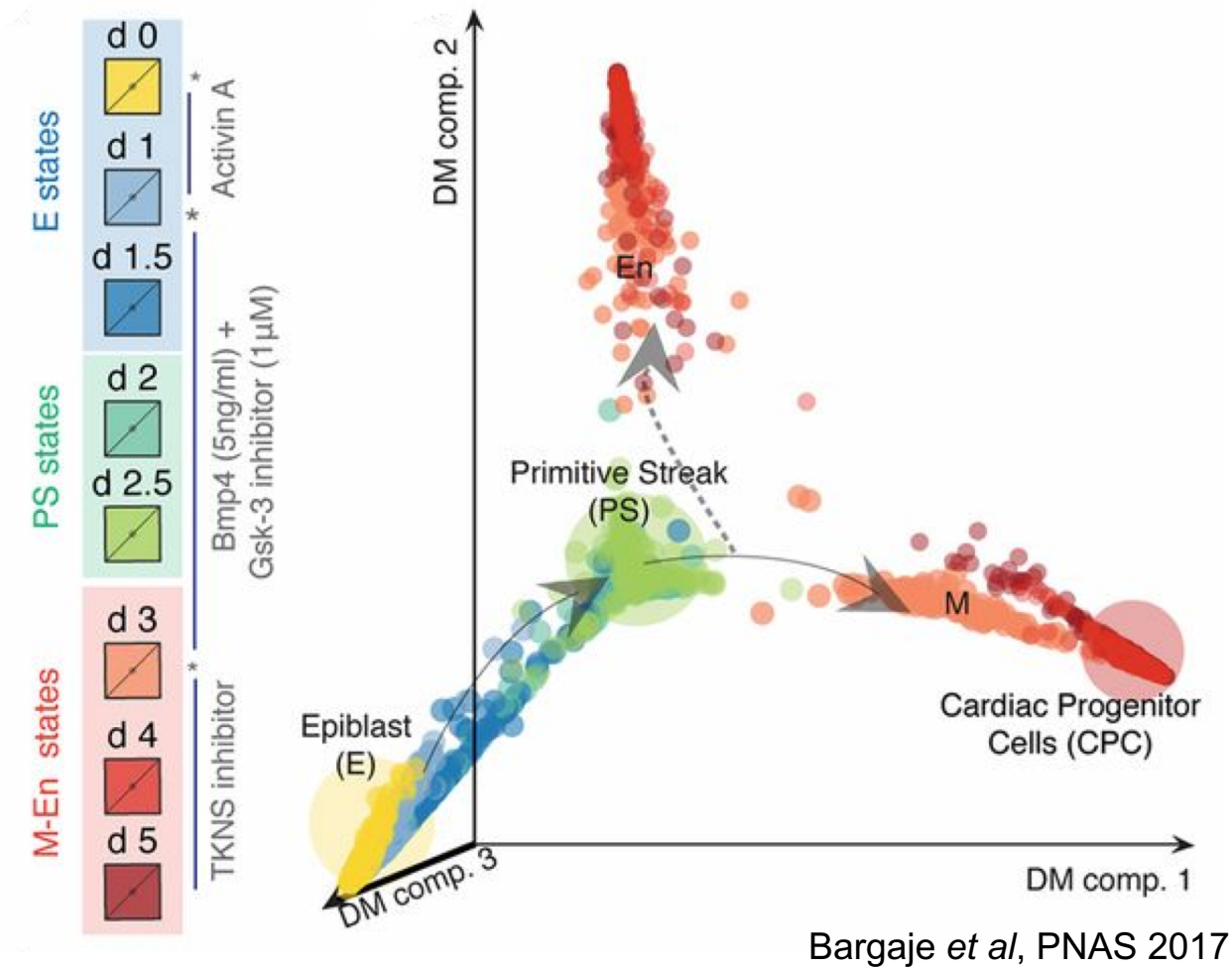
Trajectories through cell space

Cell-cycle phase prediction



Scialdone *et al.*, 2015

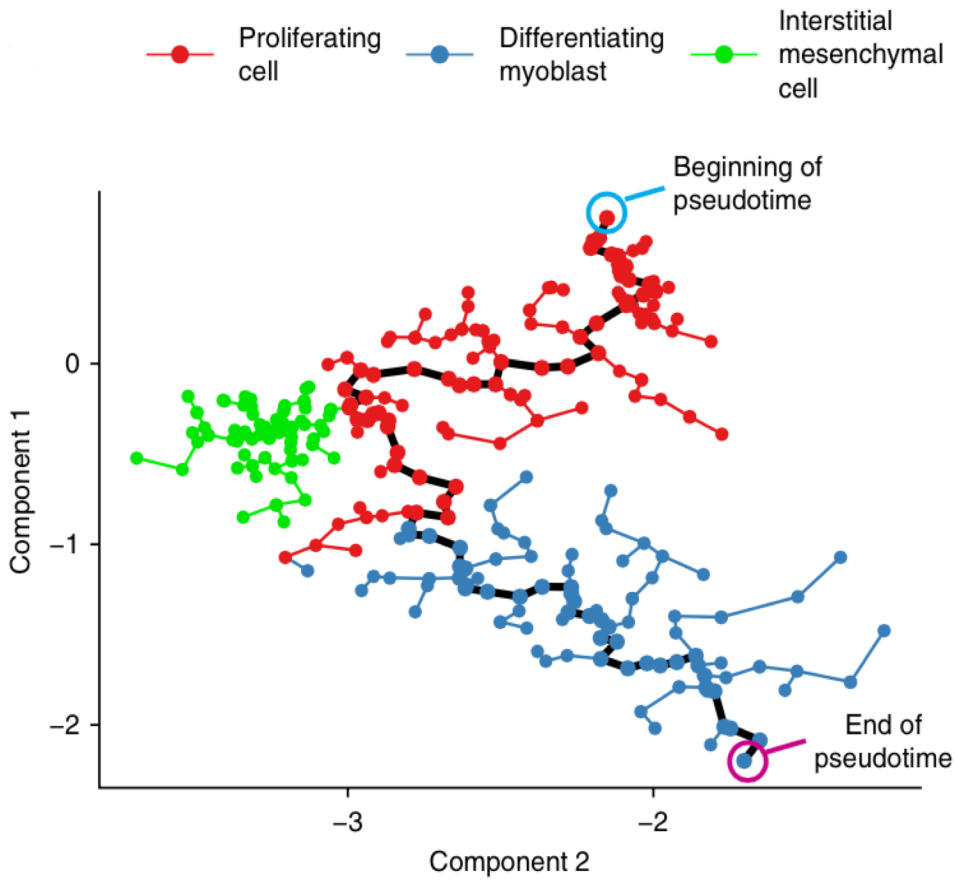
Trajectory inference



Bargaje *et al*, PNAS 2017

- Identify key branching points in development/disease
- Regulatory circuits that drive these transitions

Trajectory inference methods

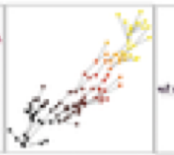
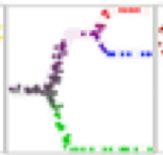
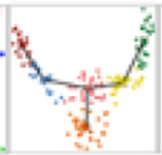
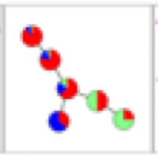
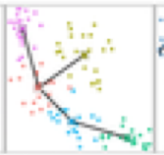
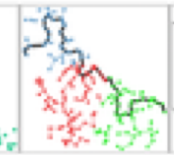
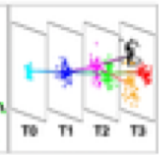


- Start with dimension reduction
- Build a graph among cells/inferred cell types
  - Typically underlying structure is based on minimum spanning tree (**MST**) or k-nearest neighborhood (**kNN**) graph.
- Either infer a linear (pseudo-time) ordering, or identify branching points

TSCAN pseudotime reconstruction with monocle

Ji *et al*, NAR 2016

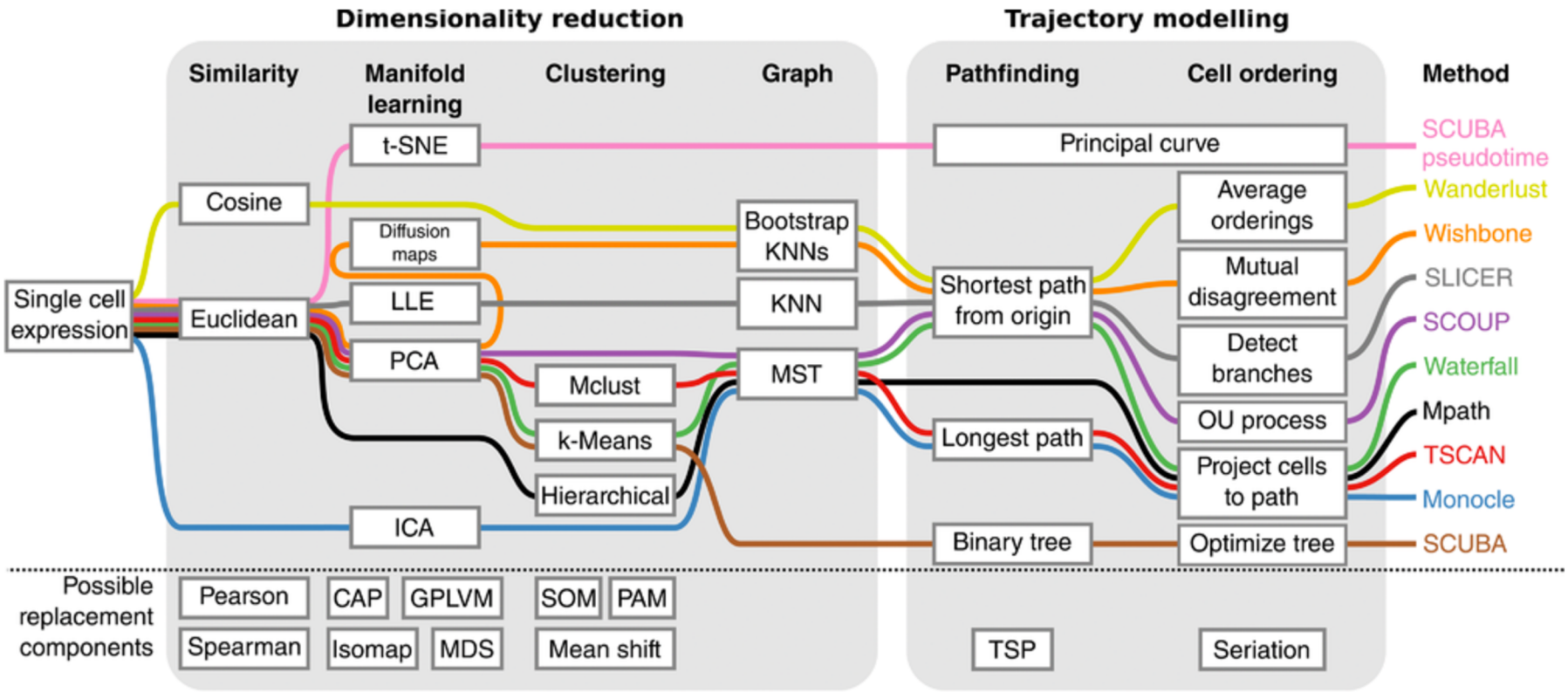
Overview of trajectory identification methods

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	+	+	+	+	−	±	−	+	+	+

First Author	Marco	Bendall	Setty	Welch	Matsumoto	Shin	Chen	Ji	Trapnell	Marco
Last Author	GC Yuan	Dana Pe'er	Dana Pe'er	Hartemink, Prins	Kiryu	Hongjun Song	Poidinger	Ji	Rinn	GC Yuan
Journal	PNAS	Cell	Nature Biotechnology	Genome Biology	BMC Bioinformatics	Cell Stem Cell	Nature Communications	NAR	Nature Biotechnology	PNAS
Year	2014	2014	2016	2016	2016	2015	2016	2016	2014	2014

hemberg-lab.github.io/scRNA.seq.course

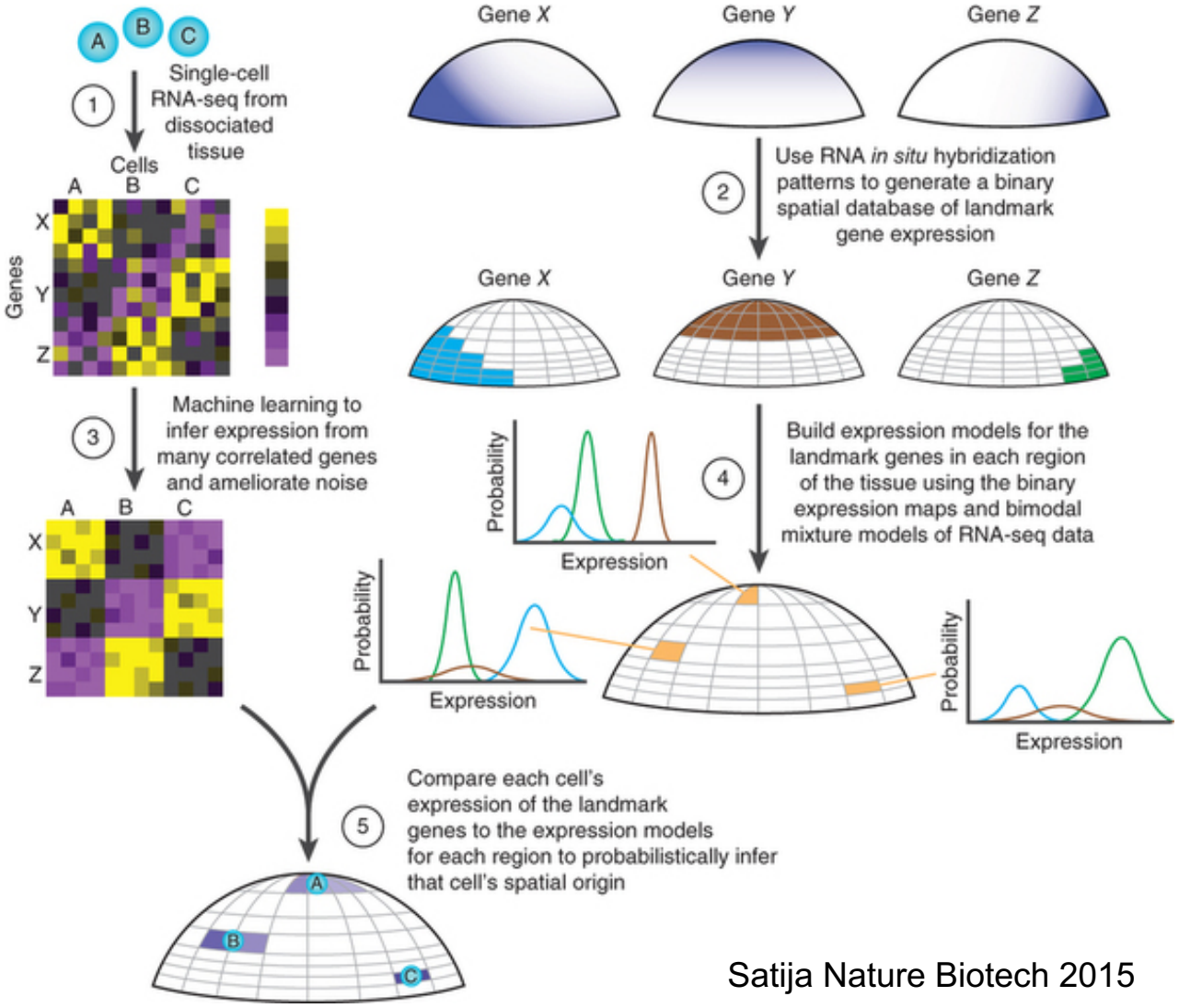
Trajectory identification: meta-method view



hemberg-lab.github.io/scRNA.seq.course

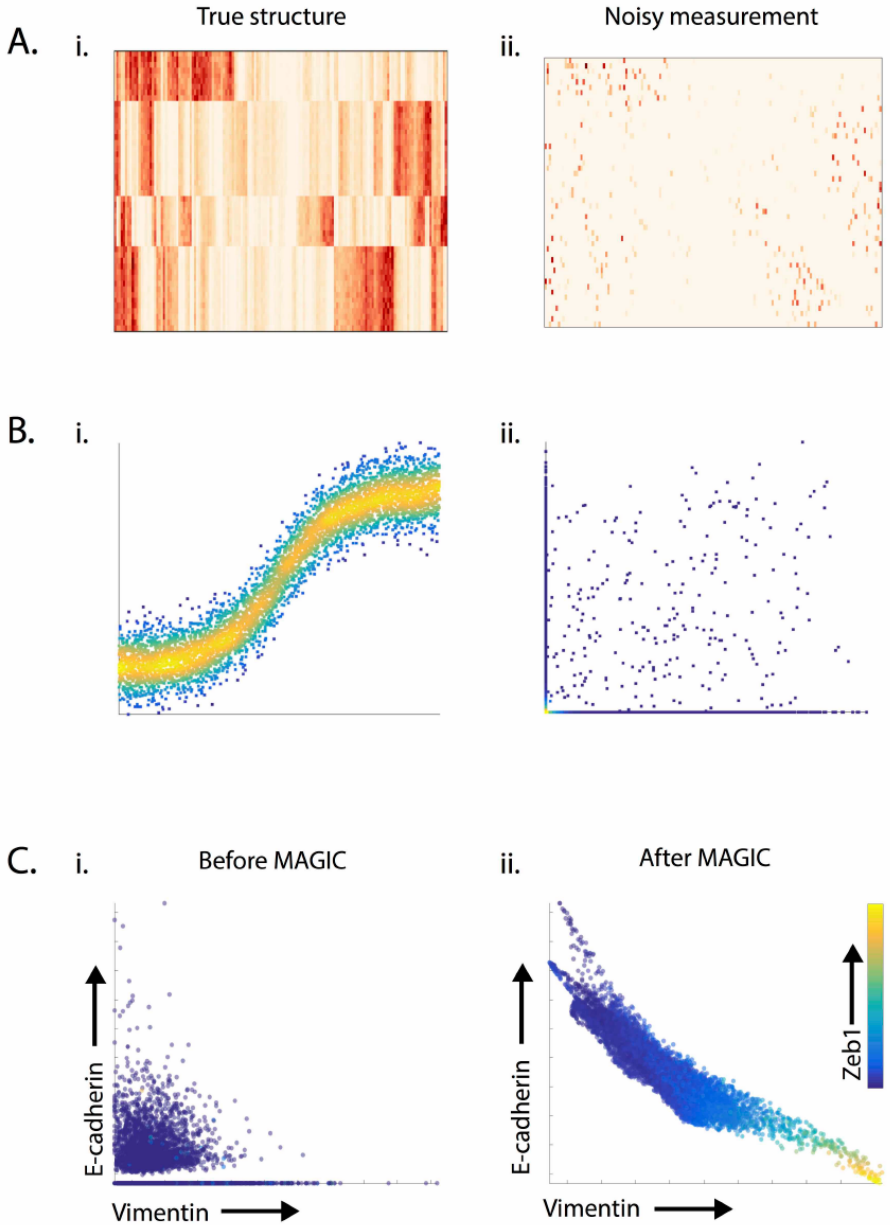
Dataset completion & missing data imputation

Spatial reconstruction of single-cell gene expression



Satija Nature Biotech 2015

Missing value imputation with MAGIC (Markov Affinity-based Graph Imputation of Cells)



Random walk on cell-cell similarity graph

- uses neighborhood-based Markov-affinity matrix
- shares weight information across cells
- generate an imputed count matrix