



Single-Cell Transcriptomics

Peter Kharchenko

Department of Biomedical Informatics,
Harvard Medical School

Regan Institute, September 14th 2017



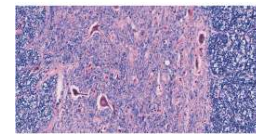
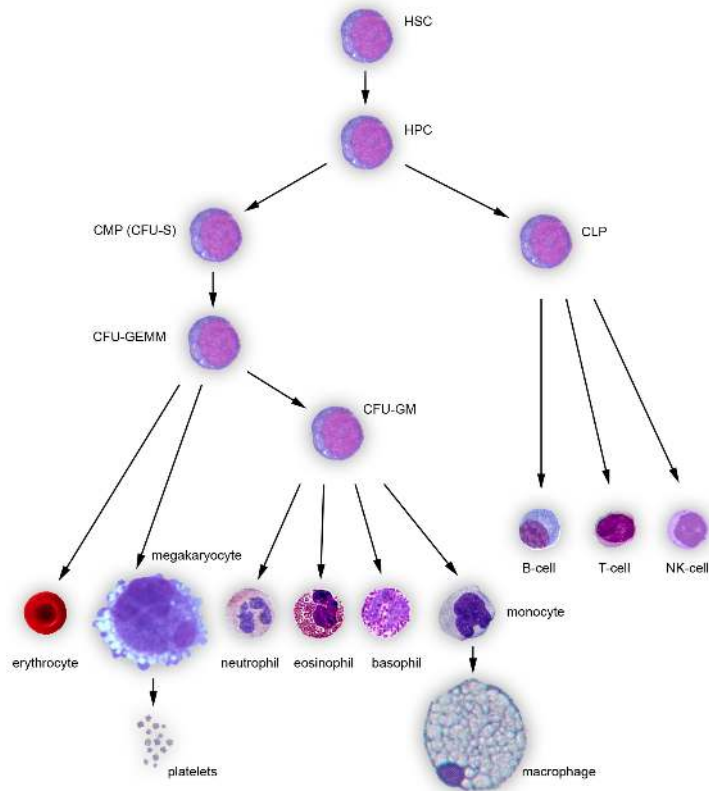
HARVARD
MEDICAL SCHOOL

DEPARTMENT OF
Biomedical Informatics

HSCI
HARVARD STEM CELL
INSTITUTE

Cells, Cell Types and Tissues

- Averages of 'bulk' samples
 - $\sim 10^7$ cells
- Complex tissue organization

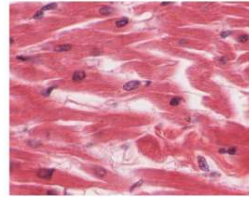
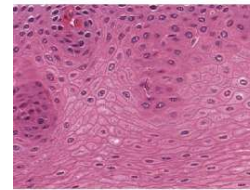


Nervous tissue

Brain
Spinal cord
Nerves

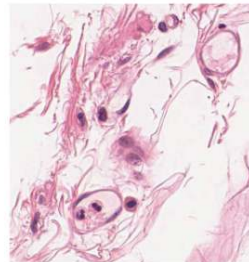
Epithelial tissue

Lining of GI tract organs
and other hollow organs
Skin surface (epidermis)



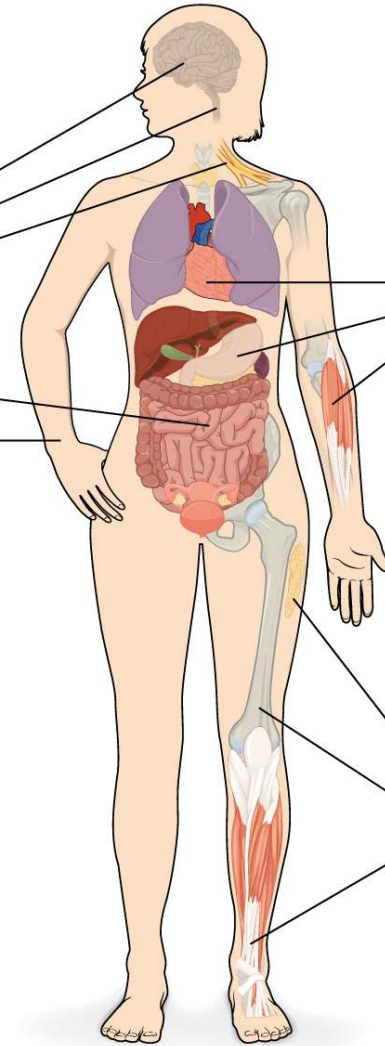
Muscle tissue

Cardiac muscle
Smooth muscle
Skeletal muscle



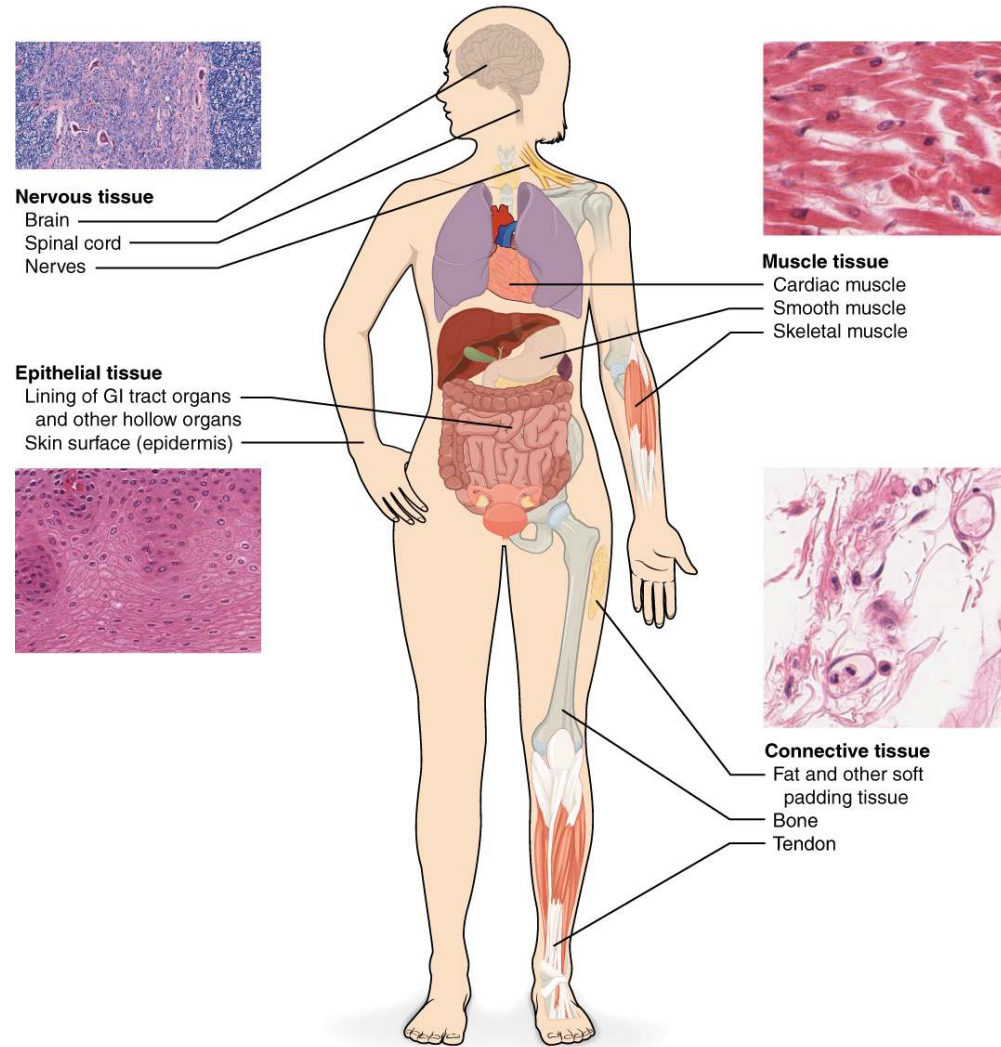
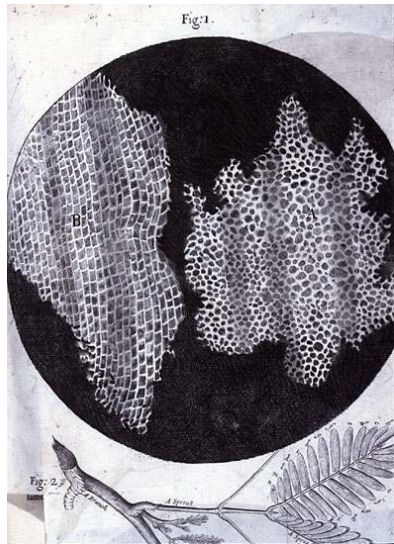
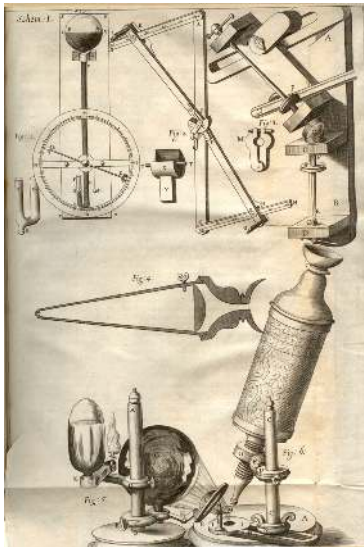
Connective tissue

Fat and other soft
padding tissue
Bone
Tendon



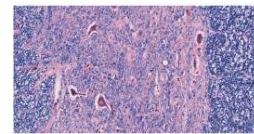
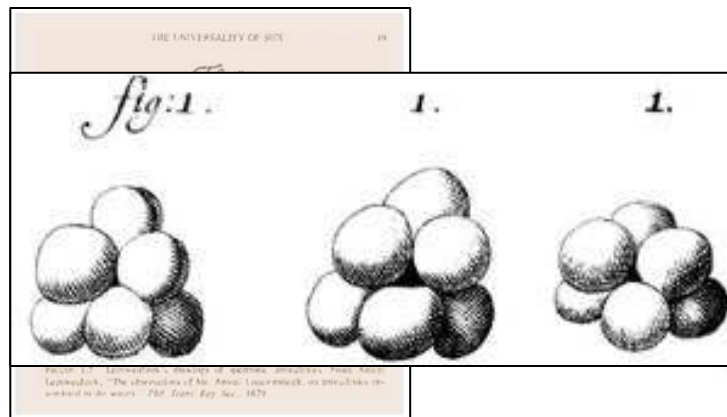
Cells, Cell Types and Tissues

- Averages of 'bulk' samples
 - $\sim 10^7$ cells
- Complex tissue organization
- Cell Types
 - Early Microscopy
 - Robert Hooke (1635-1703)



Cells, Cell Types and Tissues

- Averages of 'bulk' samples
 - $\sim 10^7$ cells
- Complex tissue organization
- Cell Types
 - Early Microscopy
 - Anton van Leeuwenhoek (1632-1723)

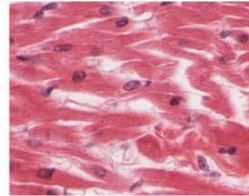
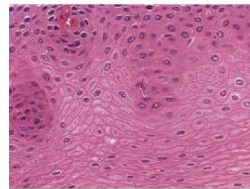


Nervous tissue

Brain
Spinal cord
Nerves

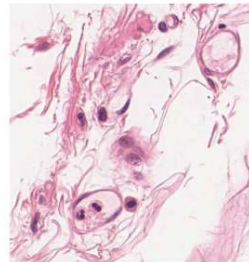
Epithelial tissue

Lining of GI tract organs
and other hollow organs
Skin surface (epidermis)



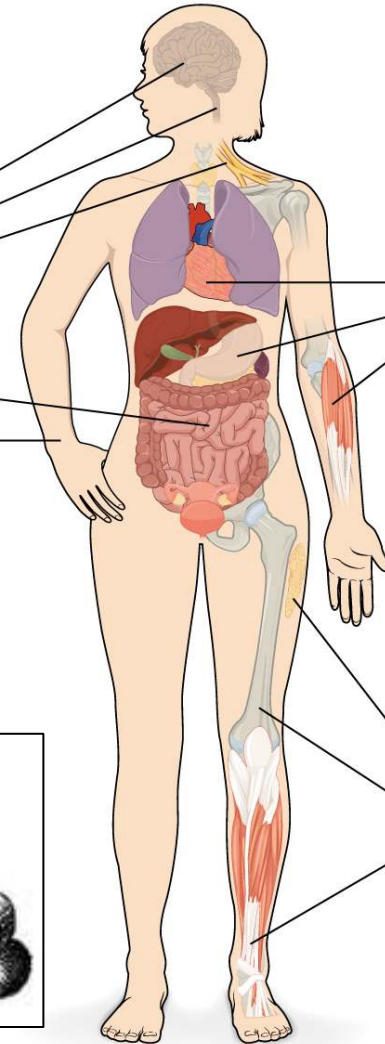
Muscle tissue

Cardiac muscle
Smooth muscle
Skeletal muscle



Connective tissue

Fat and other soft
padding tissue
Bone
Tendon

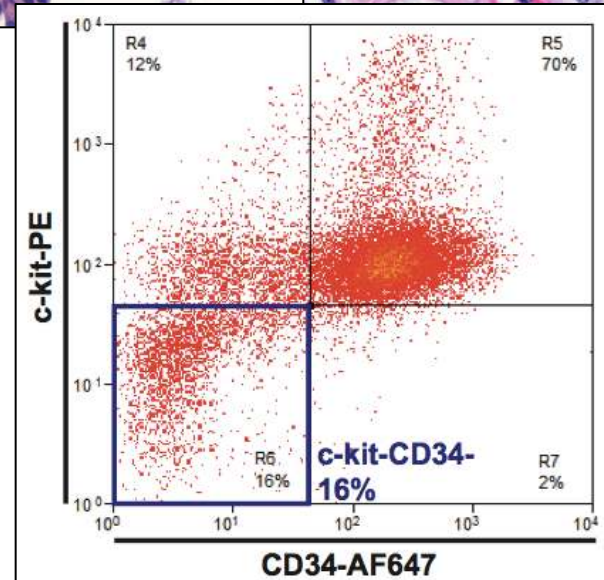
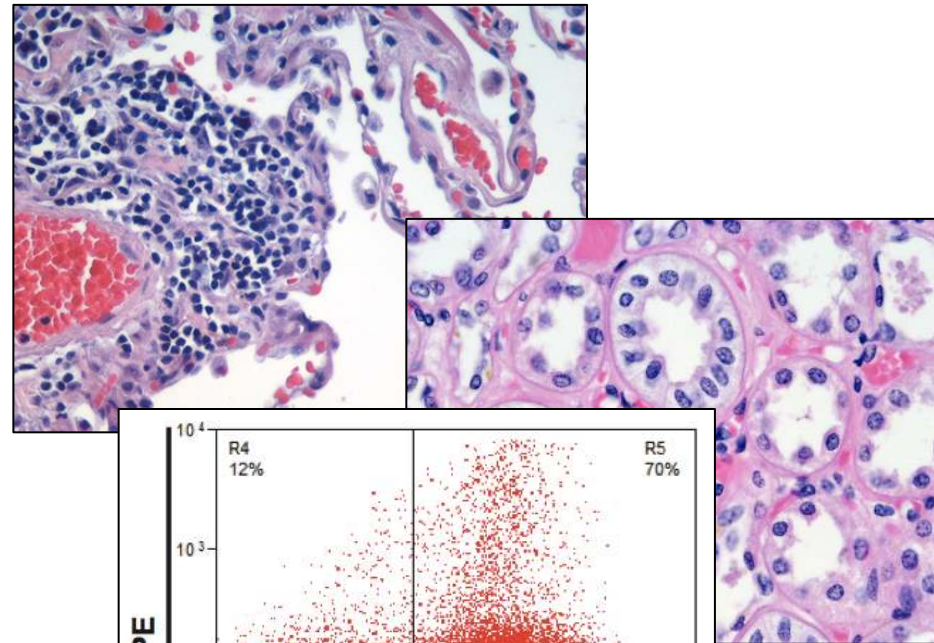


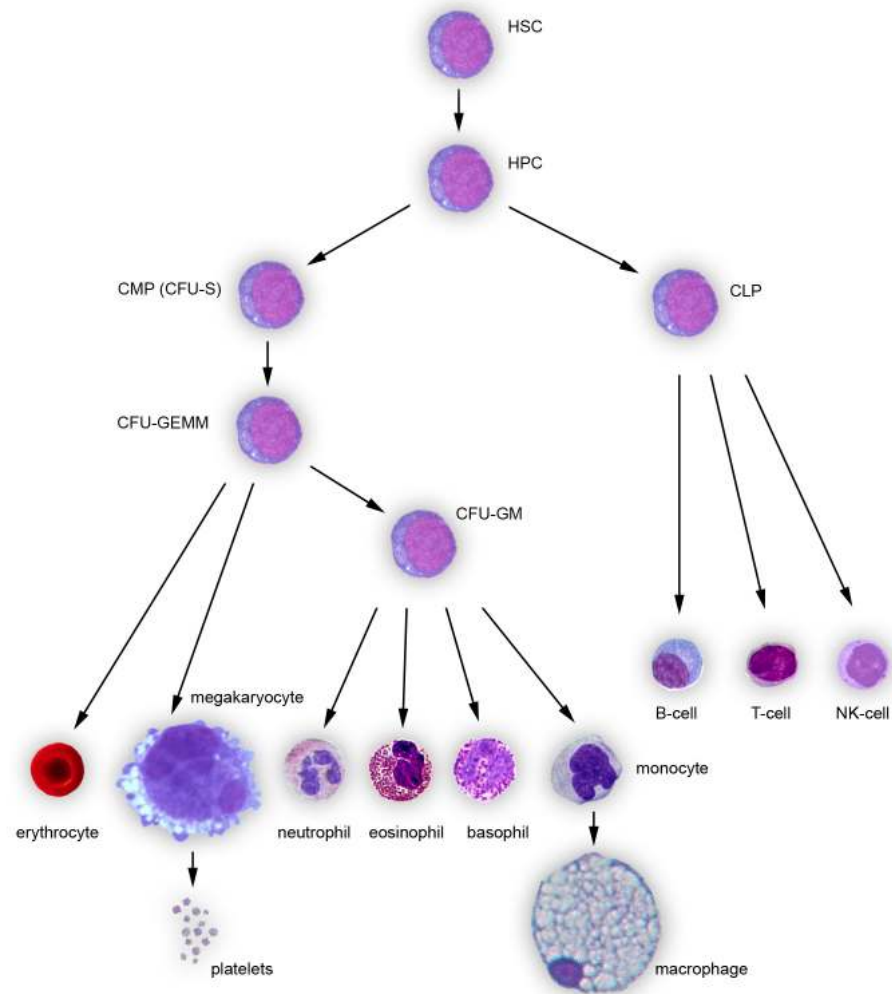
Cells, Cell Types and Tissues

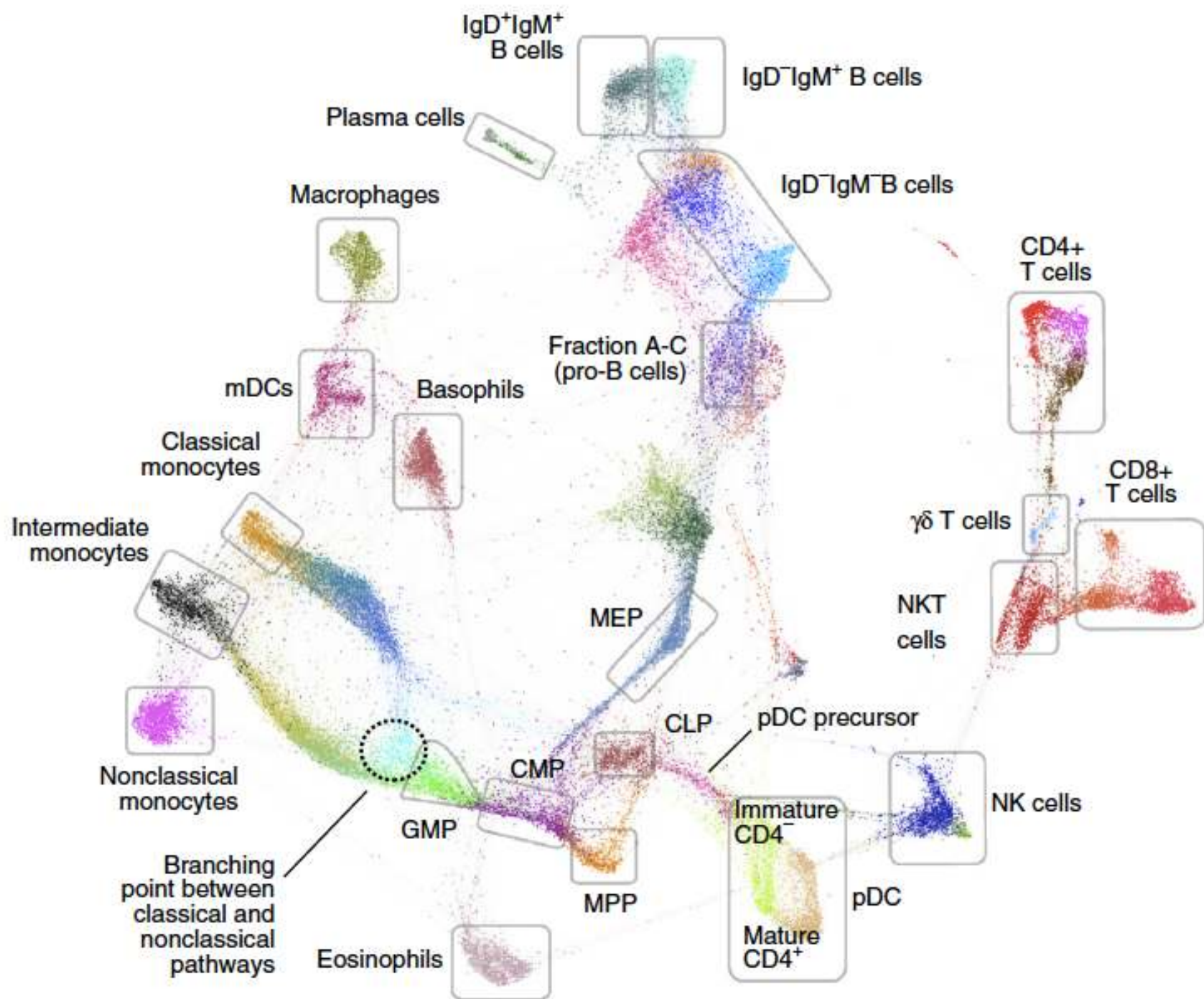
- Averages of 'bulk' samples
 - $\sim 10^7$ cells
- Complex tissue organization
- Cell Types
 - Early Microscopy
 - Anton van Leeuwenhoek (1632-1723)



Modern Histology







Single-Cell Genomics



■ Aims

- Classification: cell types, states, subpopulations
- Mechanisms: molecular differences, dynamics

■ Measurements

- Transcriptional state (single-cell RNA-seq, FISSEQ)
- Epigenetic state (DNA methylation, accessibility)
- Genomic sequence (somatic mutations)

■ Techniques

- Scaled-down sequencing assays
- Microfluidics: valve, droplet-based systems
- Microscopy, in-situ labeling, sequencing

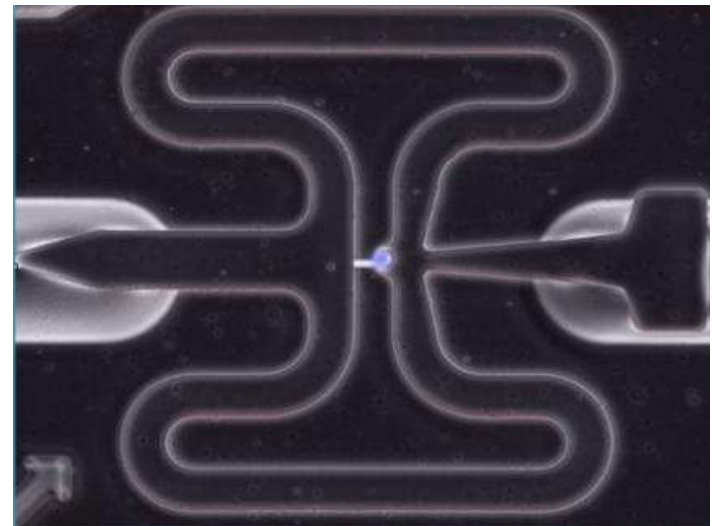
Single-cell analysis: Why?



- Variability between cells
 - Stereotypical subsets of cells
 - Cell types
 - Subclones
 - Composition of the sample
 - Cell state can reflect multiple concurrent processes
 - Cell cycle vs. cell type
 - Multiple axes of variation (cell types, cell states)
 - Gradual differences between cells
 - Differentiation trajectories
- Co-variation of individual characteristics
 - Transcriptional correlations
 - CNV co-occurrence

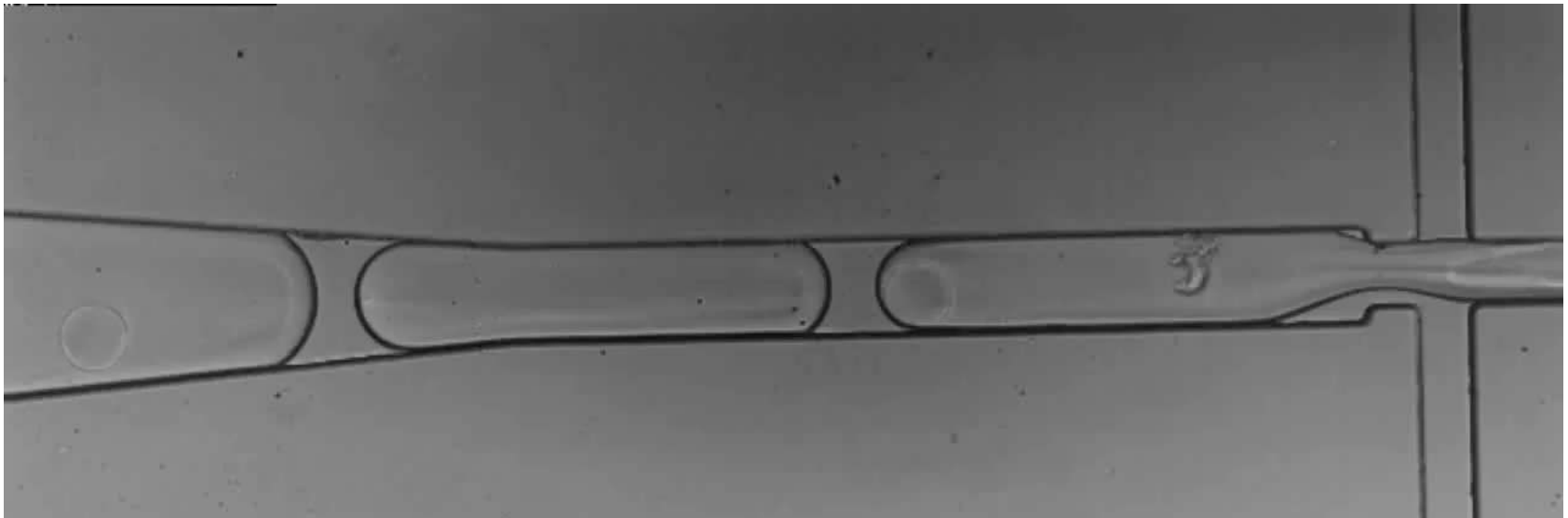
Progress in Single-Cell RNA Sequencing Techniques

- Cell separation and handling
 - Commercial C1 platform from Fluidigm



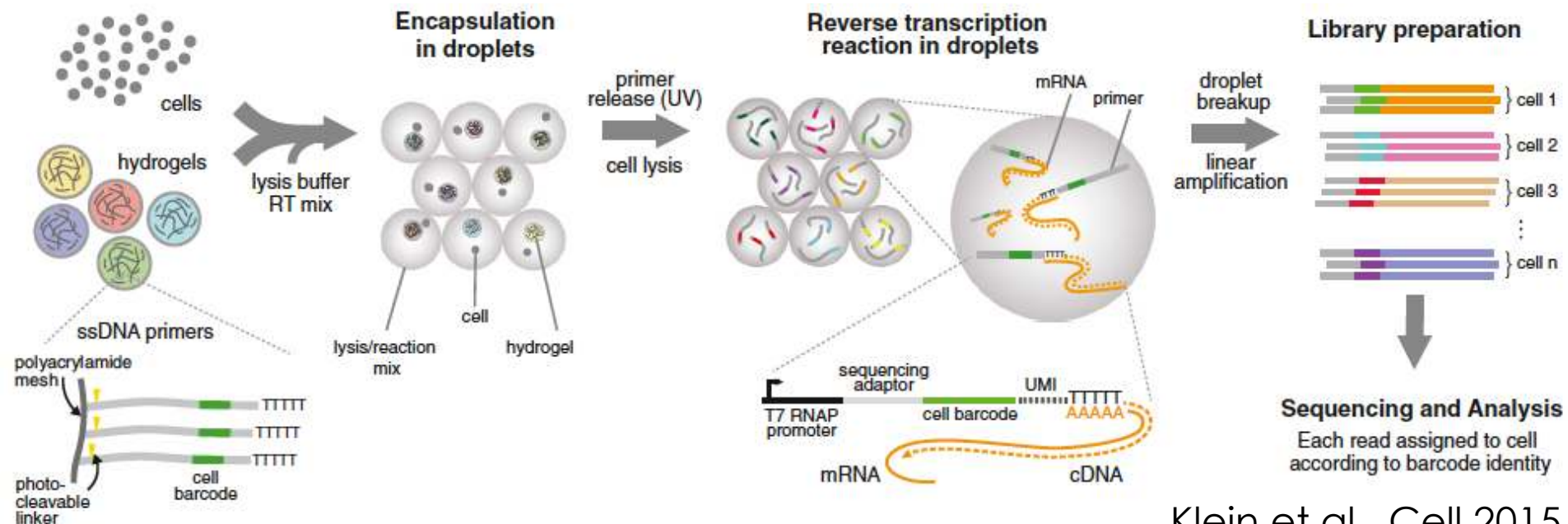
Progress in Single-Cell RNA Sequencing Techniques

- Cell separation and handling
 - Commercial C1 platform from Fluidigm
 - Using Cell Sorters to place cells
 - 384-well plates
 - Droplet microfluidics (~10K cells/run)



Progress in Single-Cell RNA Sequencing Techniques

- Cell separation and handling
 - Commercial C1 platform from Fluidigm
 - Using Cell Sorters to place cells
 - 384-well plates
 - Droplet microfluidics (~10K cells/run)

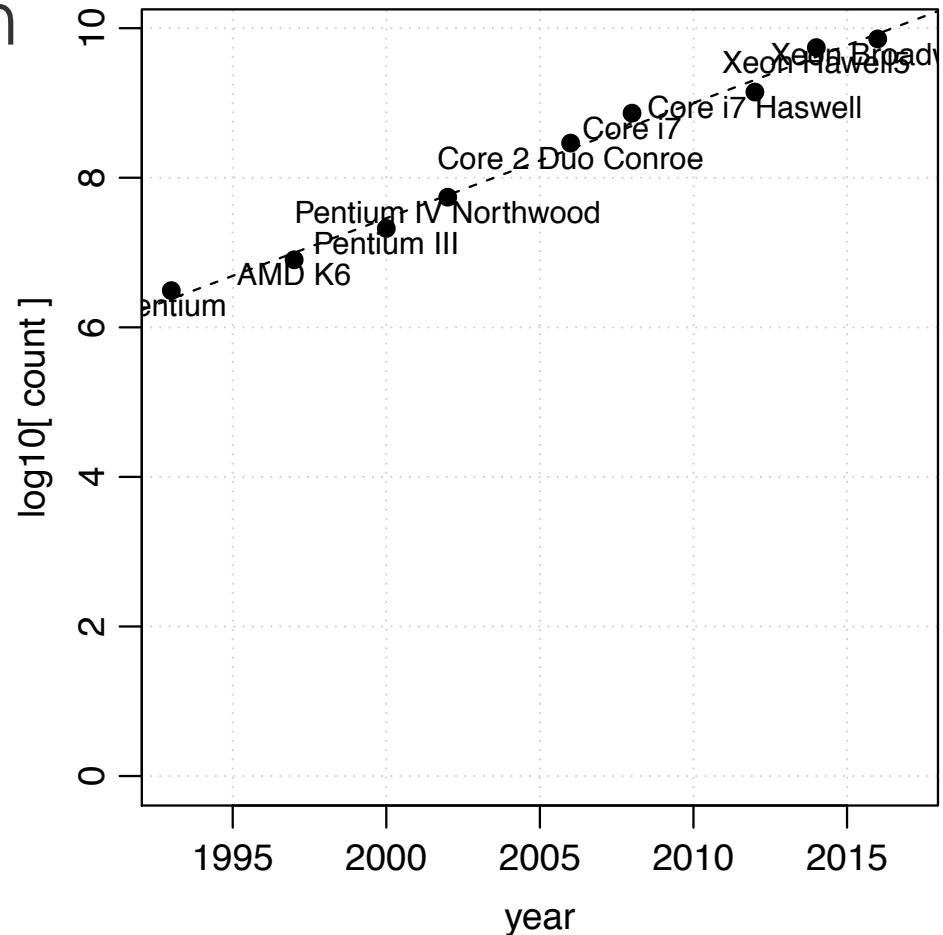


Progress in Single-Cell RNA Sequencing Techniques



■ Rapid data growth

■ Cell numbers

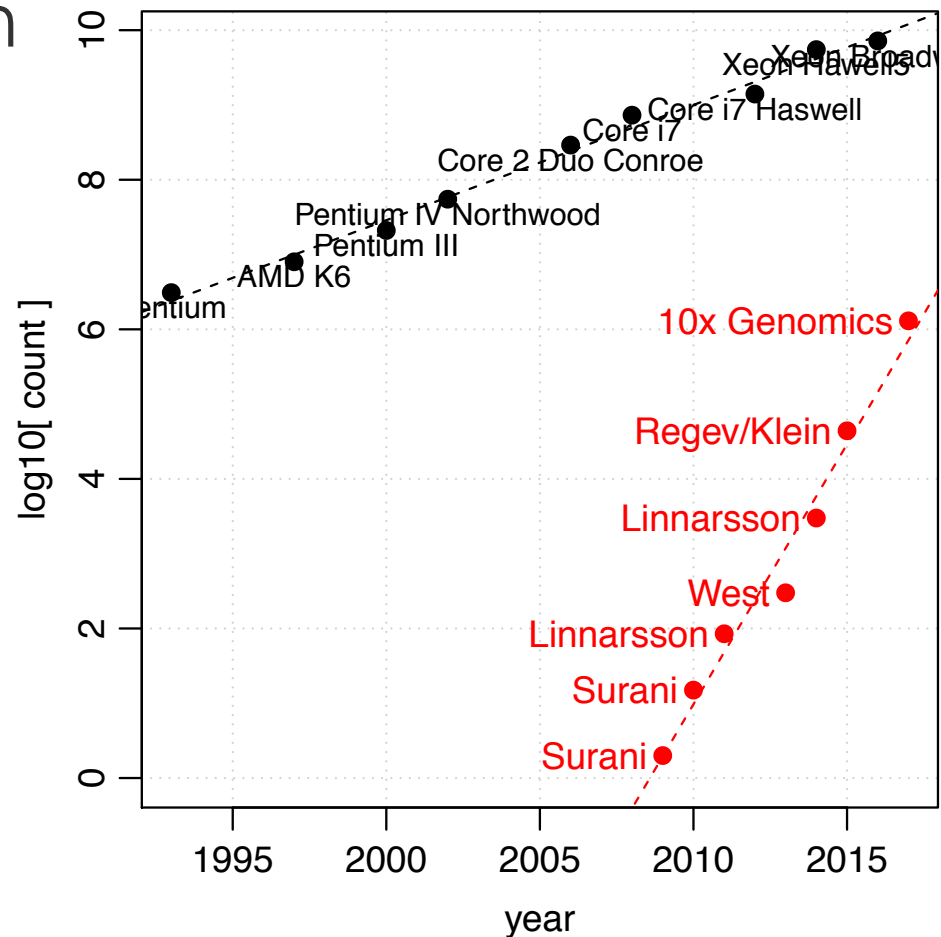


Progress in Single-Cell RNA Sequencing Techniques



■ Rapid data growth

■ Cell numbers



Progress in Single-Cell RNA Sequencing Techniques

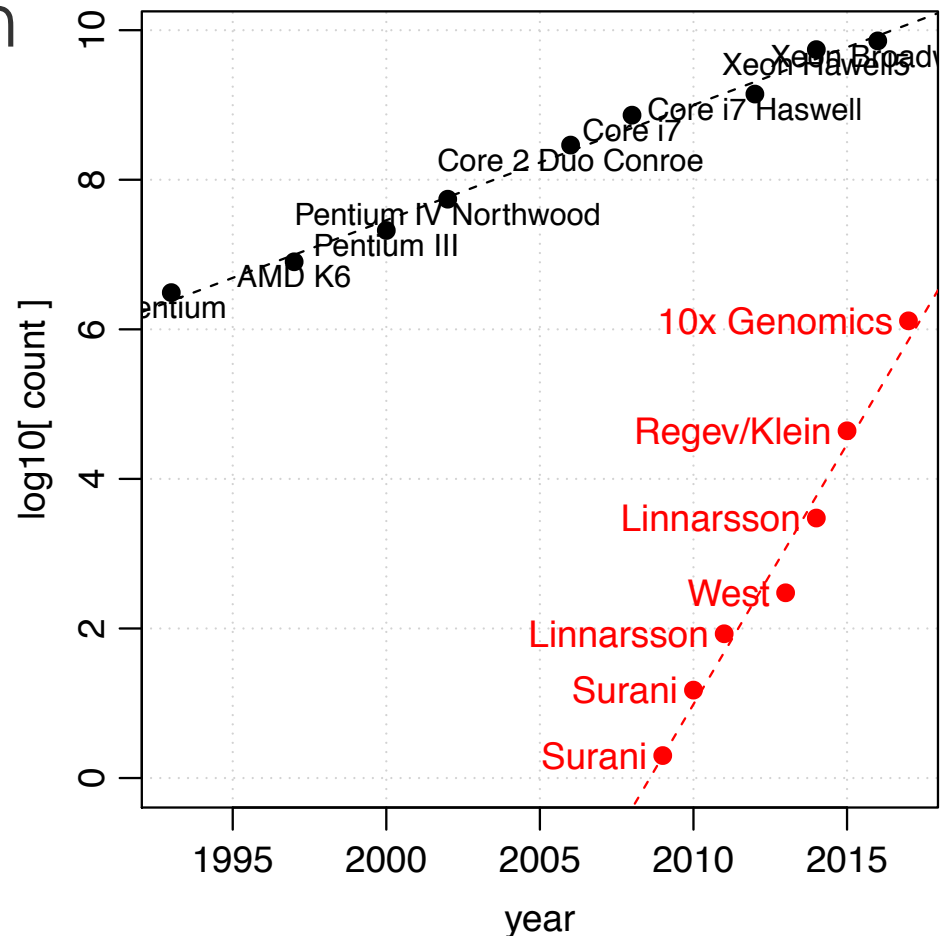


■ Rapid data growth

- Cell numbers
 - 10^5 cells by 10^4 genes
 - ~30GB sparse data
- High data volumes
- More samples
 - Multi-patient cohorts
 - Consortium-scale reference datasets

■ Computational task

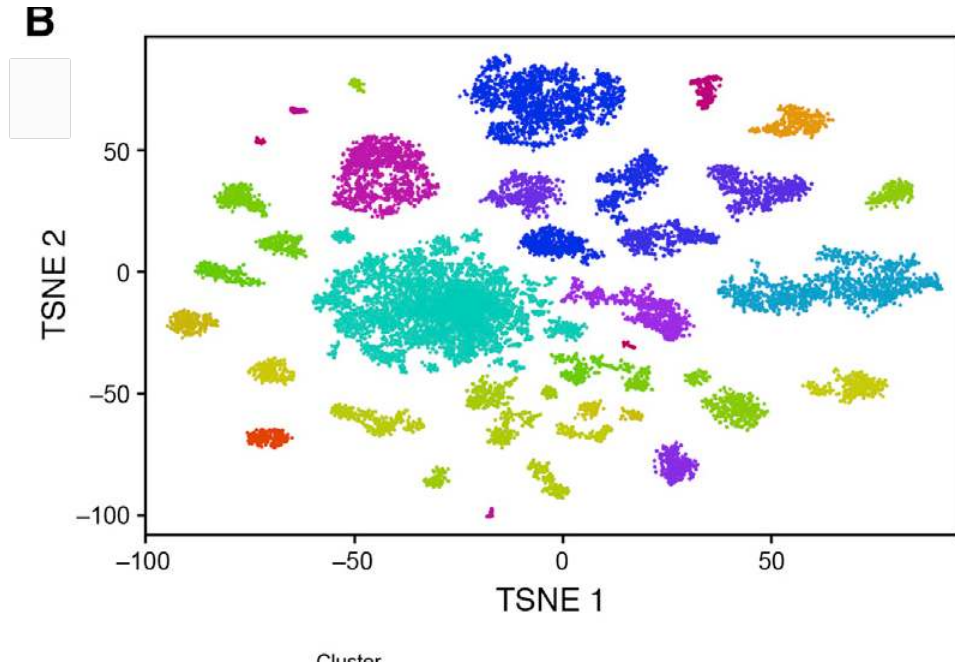
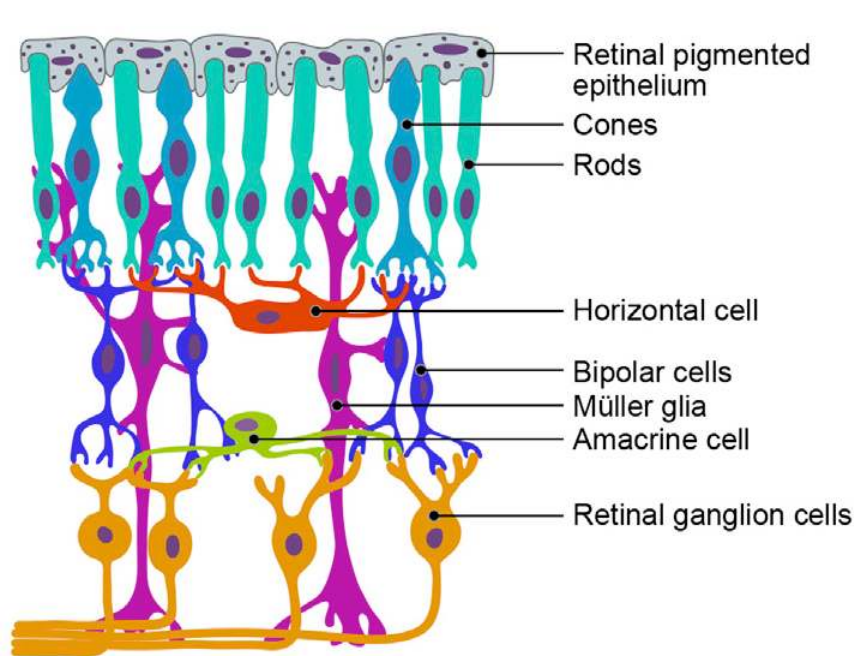
■ Statistical treatment



Single-Cell RNA-seq: tissue composition



- Mouse Retina: Macosco et al. Cell 2015



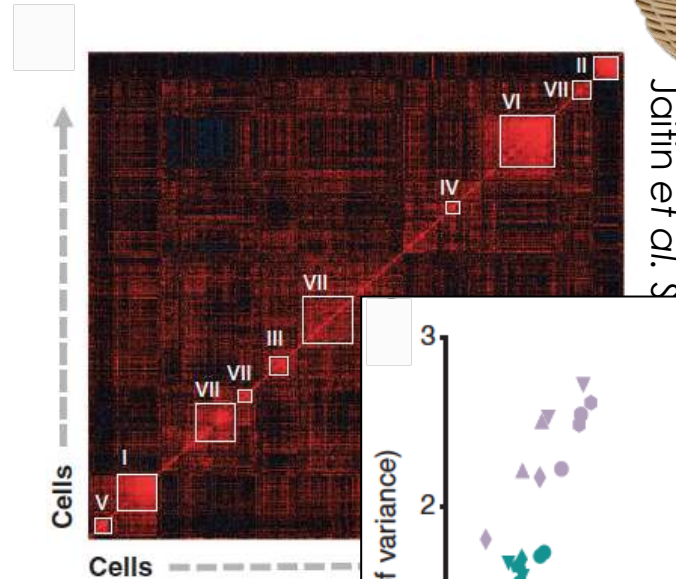
- 50 thousand cells collected using droplet method
- Recovers most known subtypes of cells

Identifying Transcriptionally Distinct Subpopulations



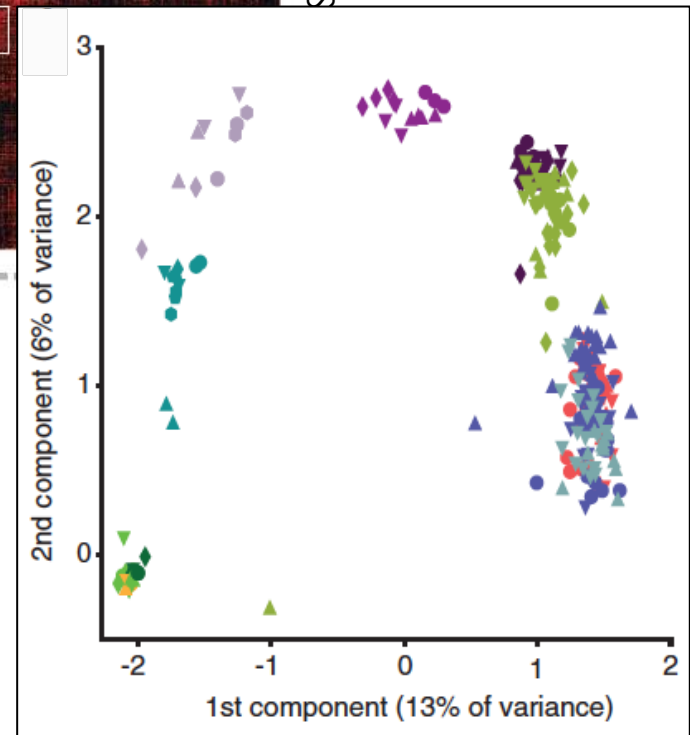
■ Clustering

- K-means, affinity
- Challenging topology
- Single, hard, partition

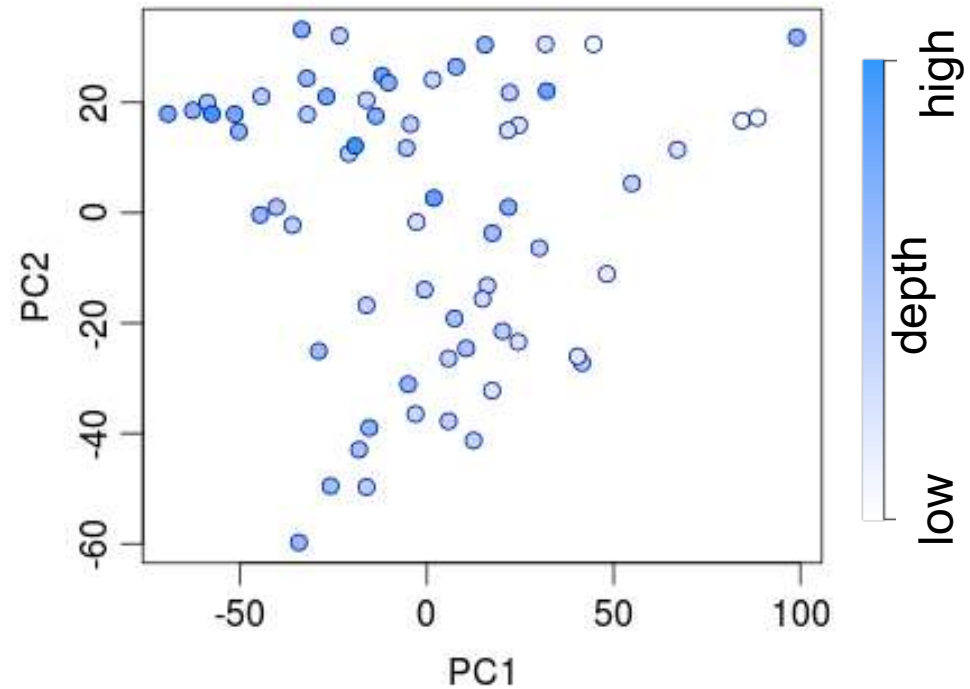
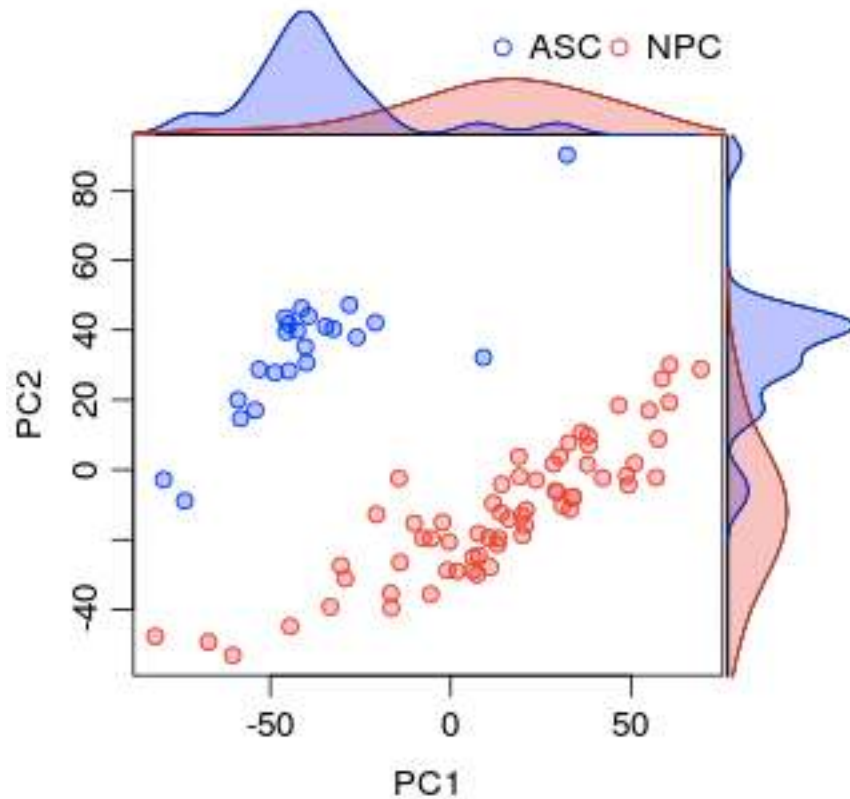


■ Dimensionality Reduction

- PCA, ICA
- Can detect gradual differences
- Sensitivity, significance



Easy and Challenging Mixtures



- Distant cell types easily separate
- Analysis of closely related cell types is challenging

Transcriptional Heterogeneity: Improving Statistical Sensitivity



- Get a better handle on technical noise
 - Account for possible drop-out events
 - Estimate true biological variability of a gene

Variability in single-cell RNA-seq data

- Differences between individual cells (of the same type)

- Overdispersion

- Measurement failures

- Cells vary in “quality”

- Problems for PCA, etc.

- Non-Gaussian

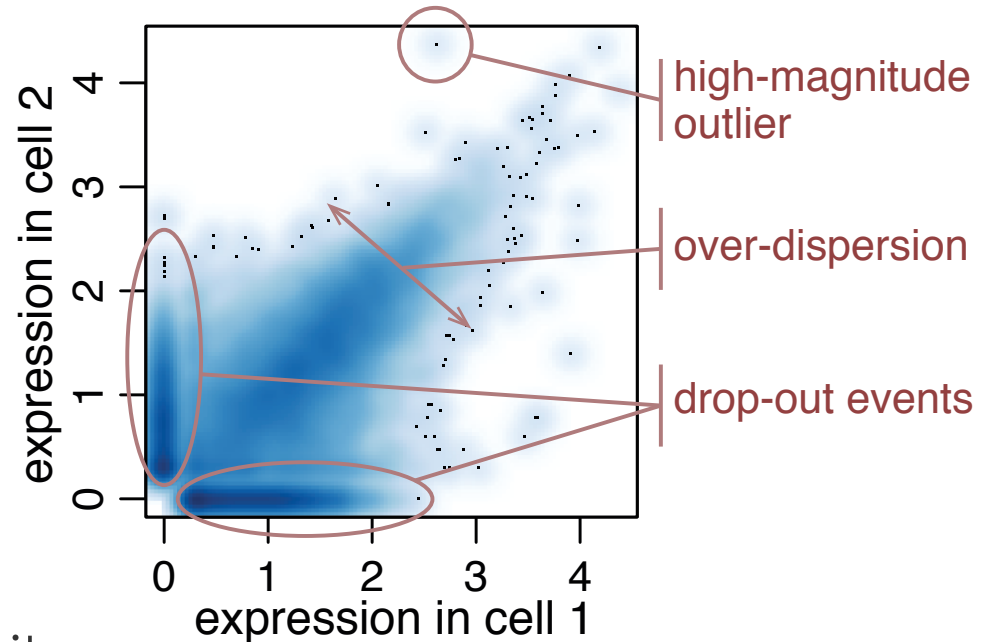
- Noisy cells form outliers

- Can mask true heterogeneity

- Biological and technical variability

- Cell-specific noise models

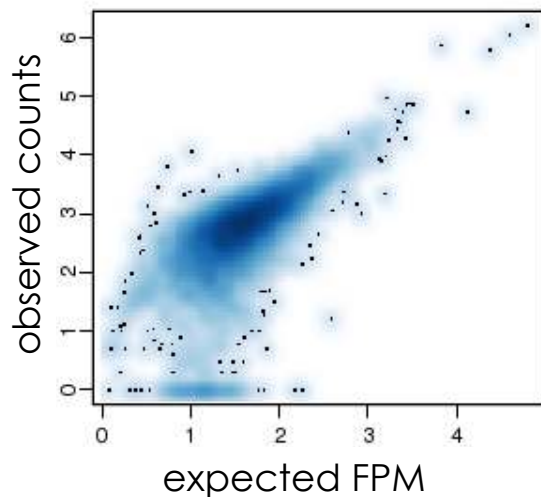
- Accurate variance normalization



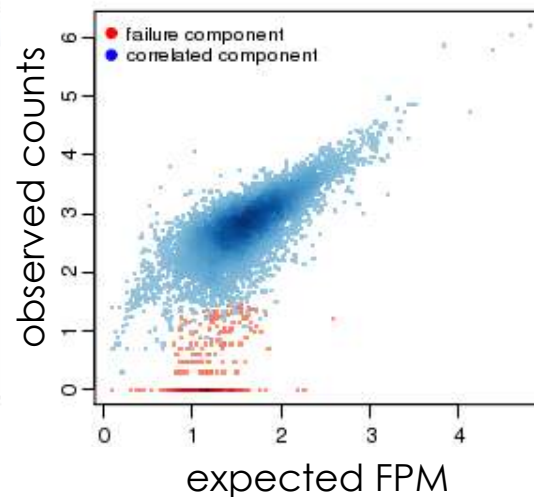
Modeling Noise of an Individual Cell

- Mixture: may **amplify** or **drop-out**, depending on expression level
- $count_i \sim \text{NegativeBinomial}(M_i) \mid count_i \sim \text{Poisson}()$
 - M_i – expected expression magnitude for gene i
(based on consensus of non-drop-out measurements within a group)
- Mixing between the two options depends on the magnitude itself
 - probability of drop-out is modeled using logistic regression

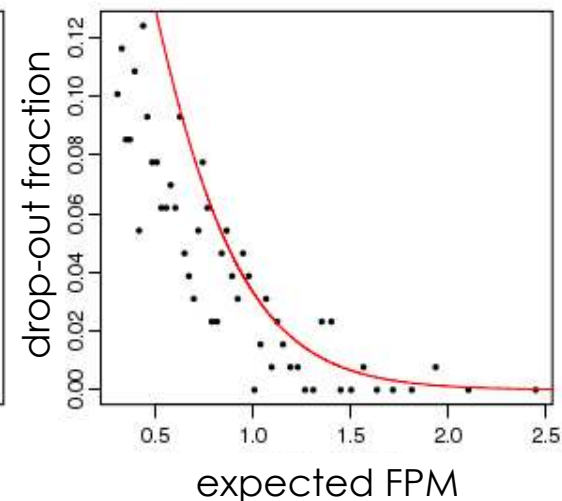
observed vs. expected



mixture fit

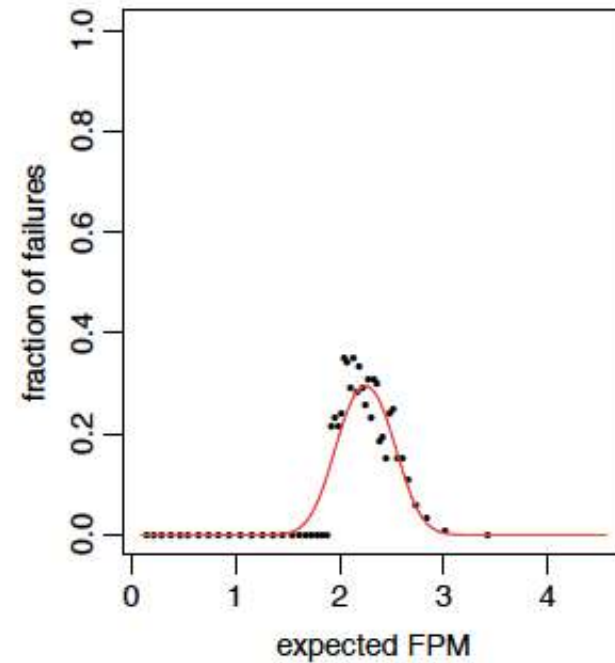
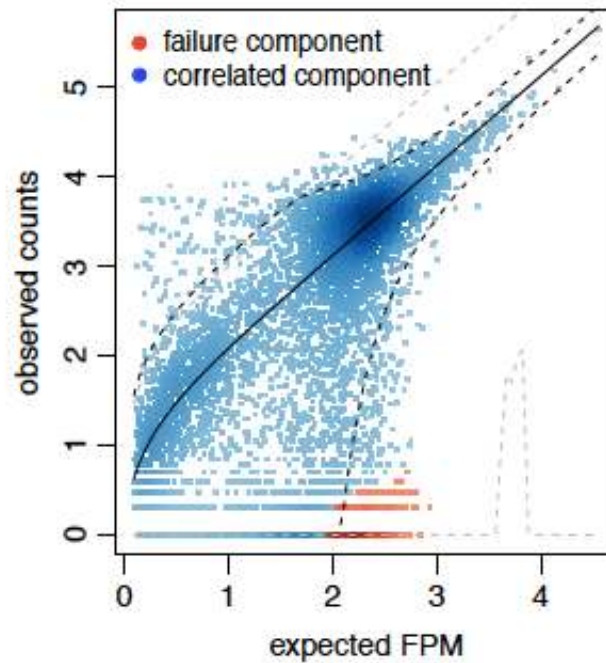
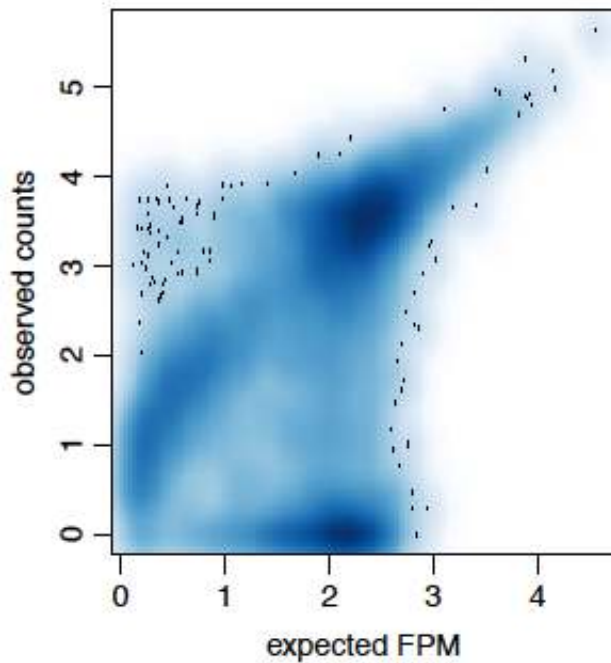


drop-out dependency



Modeling Noise of an Individual Cell

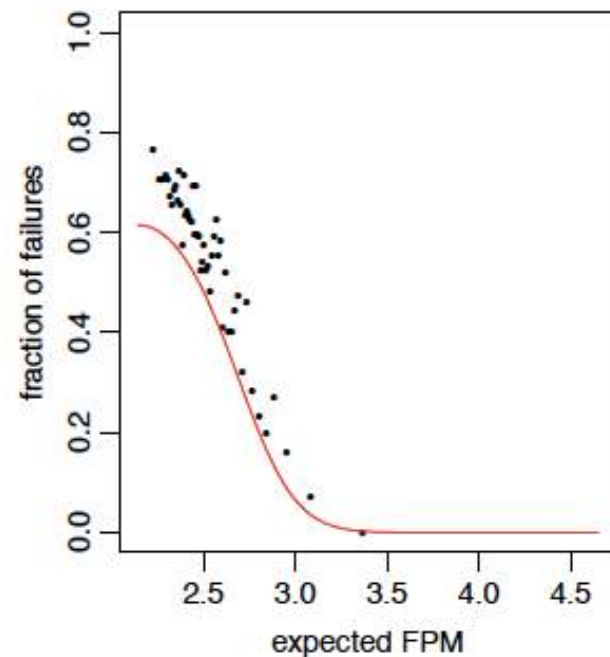
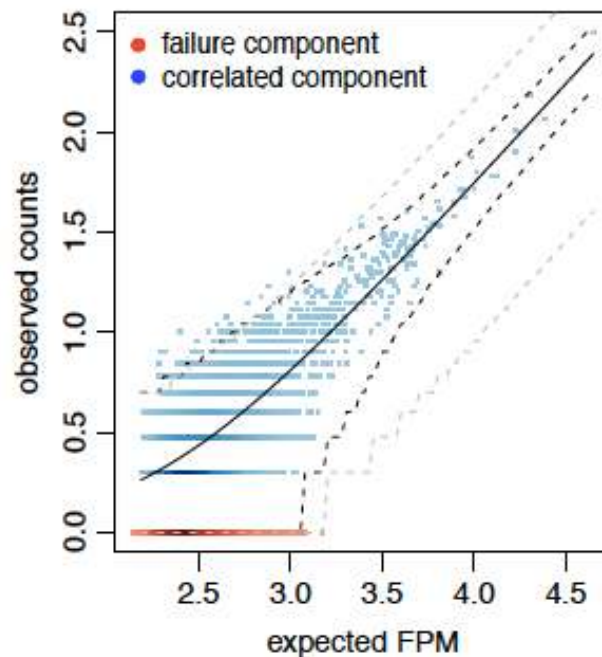
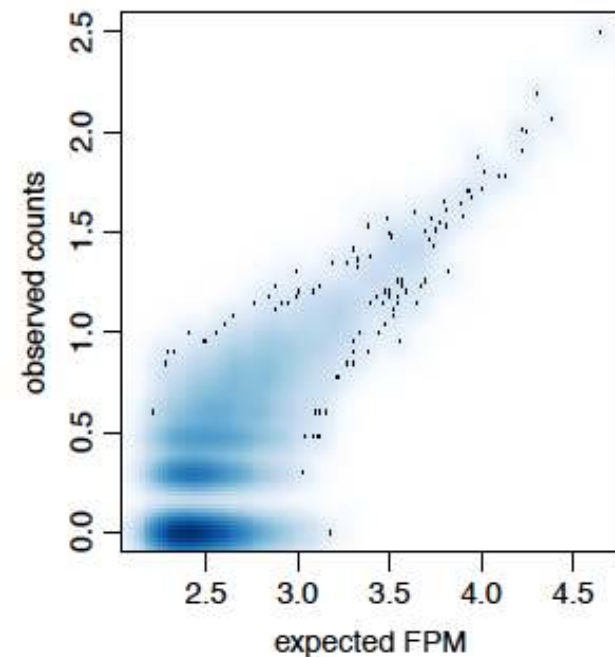
- Fluidigm C1
 - Generally high overdispersion
 - Compressed dynamic range
 - Exceptionally high amplification rate at low magnitudes



Modeling Noise of an Individual Cell

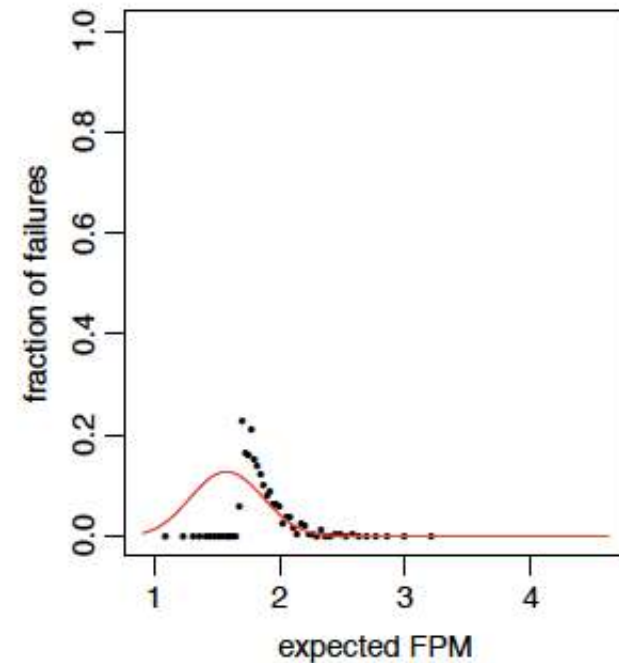
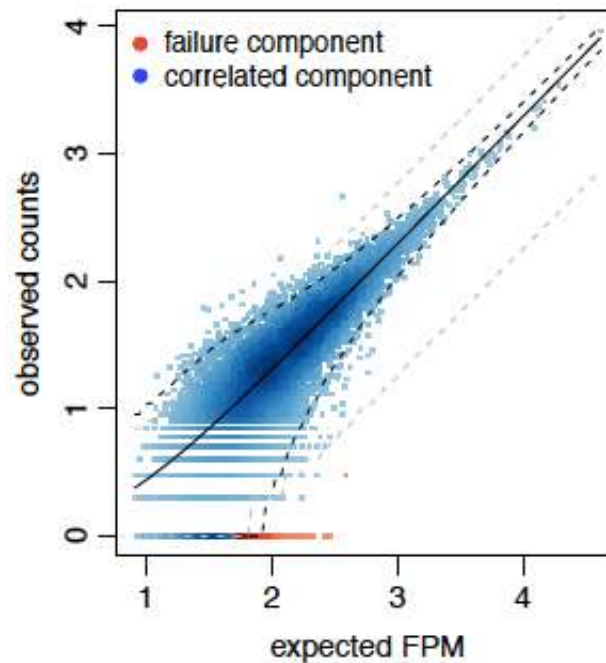
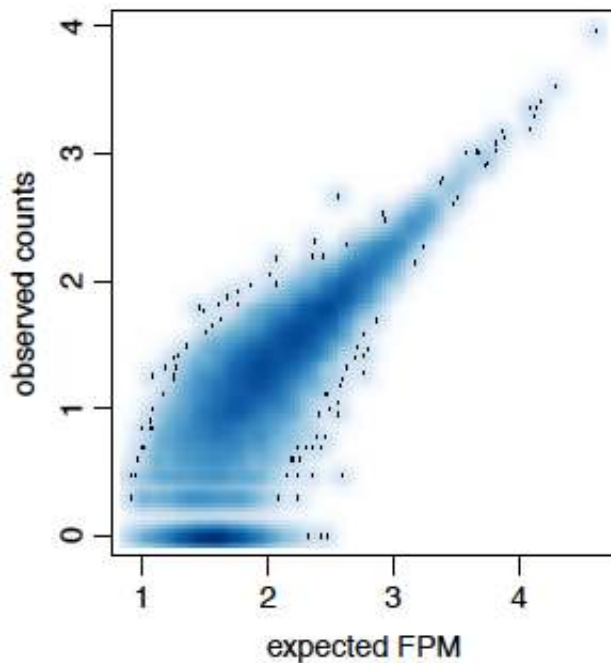
■ Indrop

- Relatively sparse: 20-50k UMIs/cell, 3-5K genes/cell
- Much lower overdispersion
- Elevated drop-out rates due to inefficient capture and RT
- ... but with many more cells (3K/batch, multiplexed)



Modeling Noise of an Individual Cell

- Optimized Smart-seq2
 - High coverage: 200k reads/cell
 - Wide dynamic range
 - Very low drop out rates
 - Very low overdispersion (comparable with UMI-based techniques)

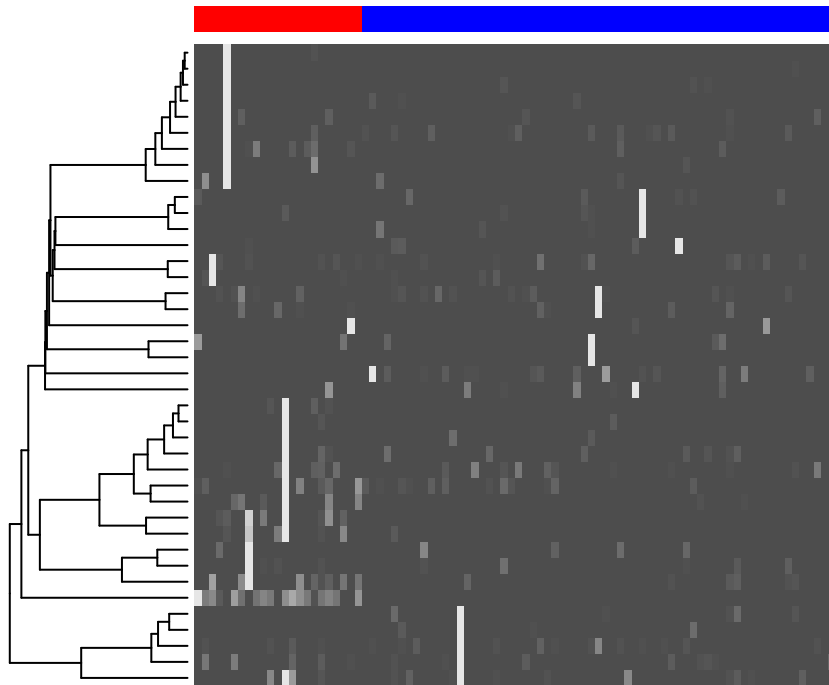


Robust inference of most variable genes



Astrocytes

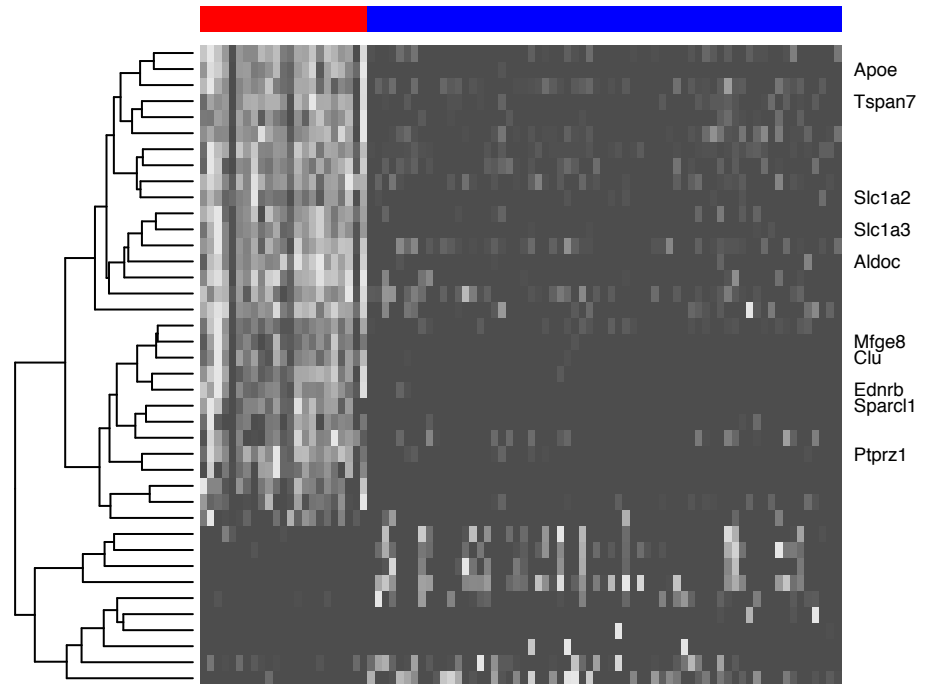
NPCs



Brennecke et al. Nat. Methods 2013

Astrocytes

NPCs



Cell-specific models (PAGODA)

Robust inference of most variable genes



Astrocytes

NPCs

Astrocytes

NPCs



RESEARCH ARTICLE

BASiCS: Bayesian Analysis of Single-Cell Sequencing Data

Catalina A. Vallejos^{1,2*}, John C. Marioni^{2*}, Sylvia Richardson^{1*}

¹ MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, United Kingdom, ² EMBL European Bioinformatics Institute, Cambridge, United Kingdom

* catalina@mrc-bsu.cam.ac.uk (CAV); marioni@ebi.ac.uk (JCM); sylvia.richardson@mrc-bsu.cam.ac.uk (SR)

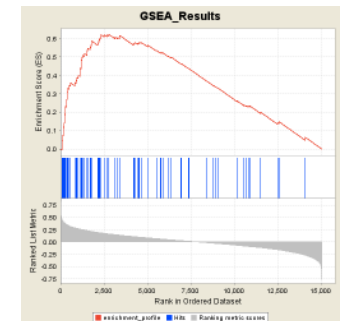
Brennecke et al. Nat. Methods 2013

Cell-specific models (PAGODA)

Transcriptional Heterogeneity: Improving Statistical Sensitivity



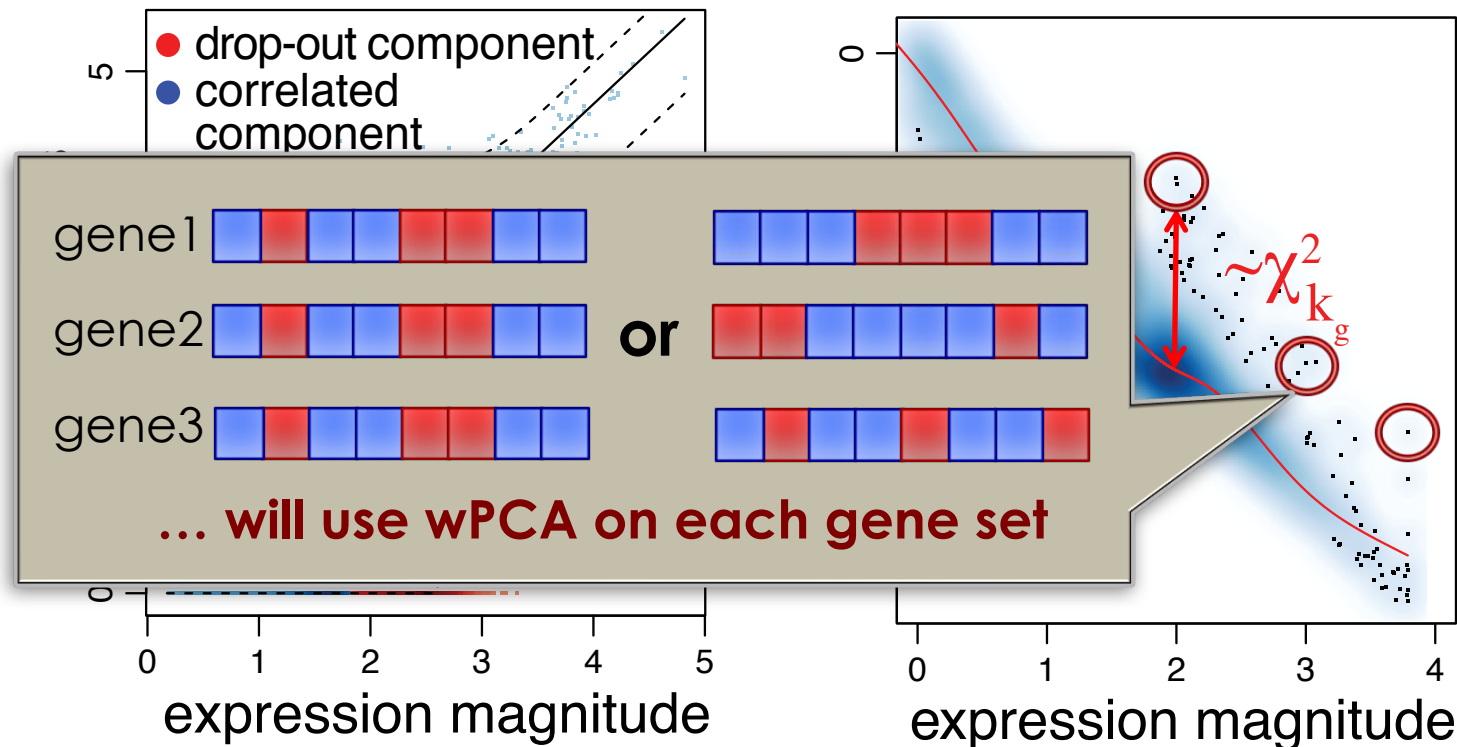
- Get a better handle on technical noise
 - Account for possible drop-out events
 - Estimate true biological variability of a gene
- Look for broader patterns of variability
 - Gene sets: annotated pathways, computationally derived sets
 - GO statistics, GSEA, widely used
 - Coordinated patterns of variability of genes linked to function/phenotype – a strong signal
 - Increases statistical power



Variance Normalization



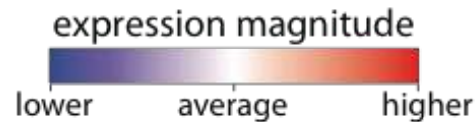
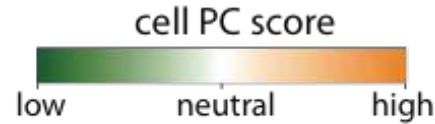
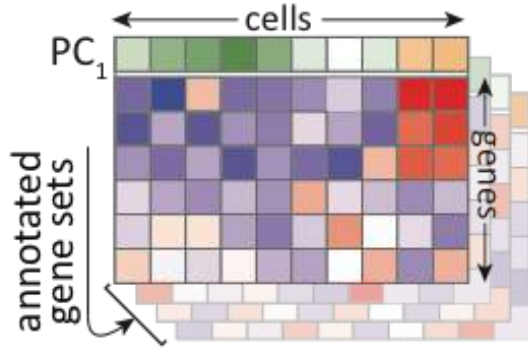
- Account for cell- and gene-specific uncertainty
 - Cell-specific error models. Expression-dependent size
 - NB/Poisson sample variance $\rightarrow \chi^2$ statistic
- Account for expression magnitude dependency
 - Adjusting with local regression fit



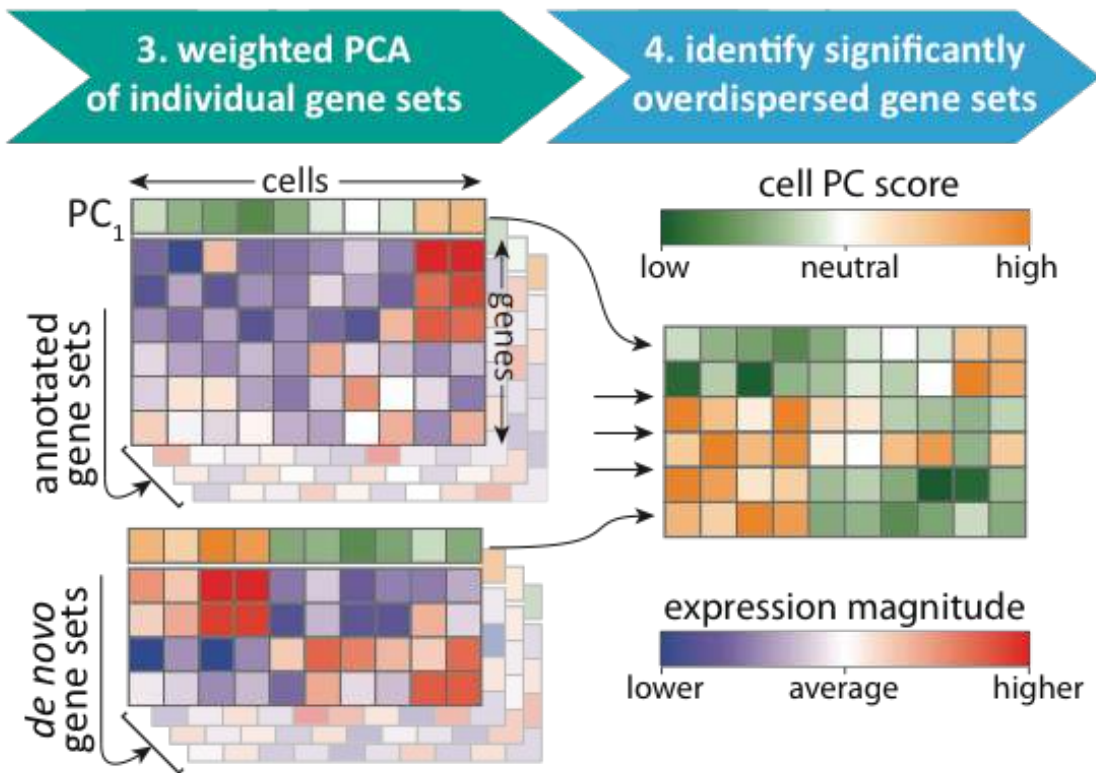
Heterogeneity Overview: Pathway Clustering



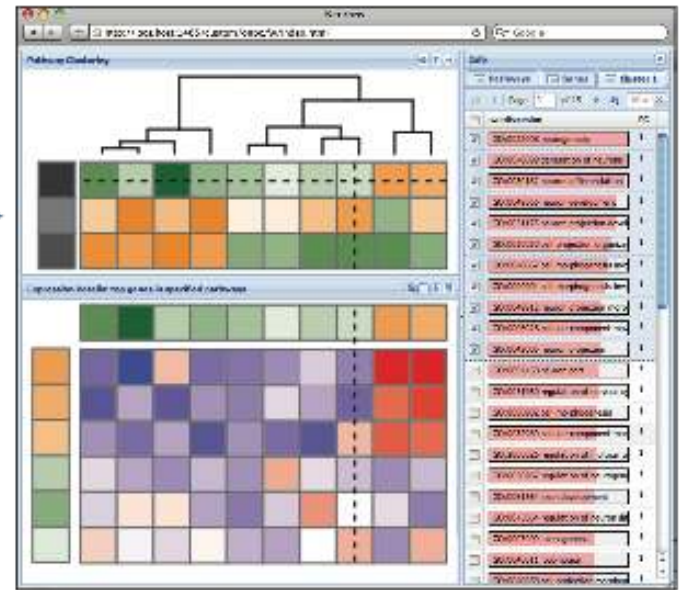
3. weighted PCA of individual gene sets



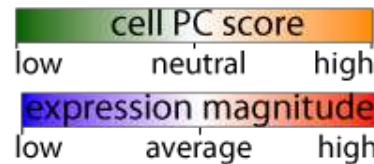
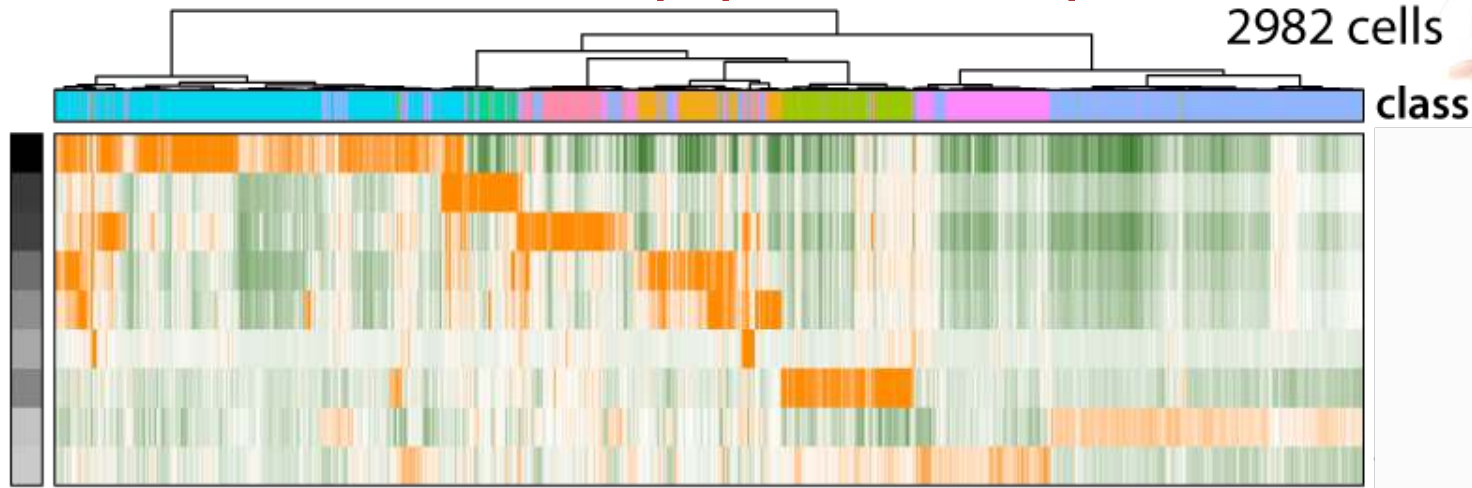
Heterogeneity Overview: Pathway Clustering



A photograph of a large display of apples, with red apples on the outside and green apples in the center. A small white sign with a red border is placed in the center of the green apples. The sign reads: "APPLANCIE WARNING!! PLEASE DO NOT HANDLE!! THANKS!!" with a small red apple icon and a heart symbol.



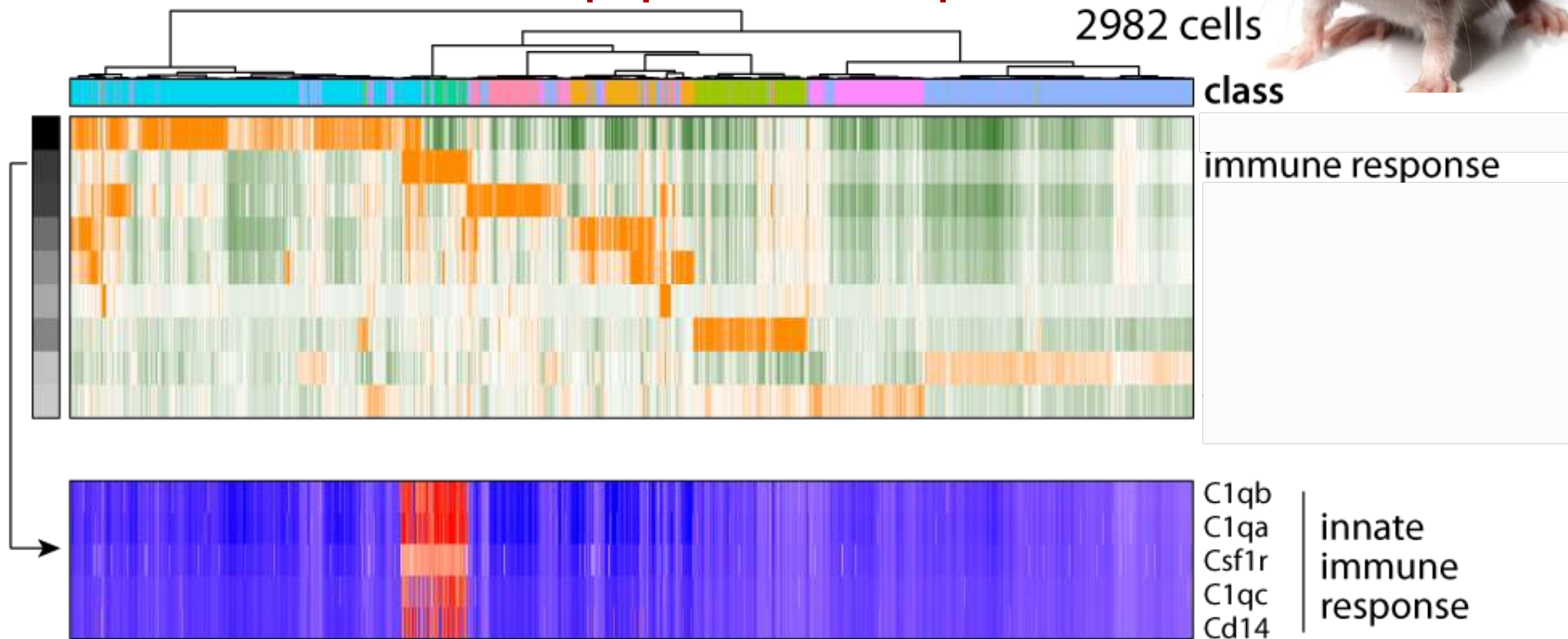
Heterogeneity in Mouse Cortex and Hippocampus



Heterogeneity in Mouse Cortex and Hippocampus



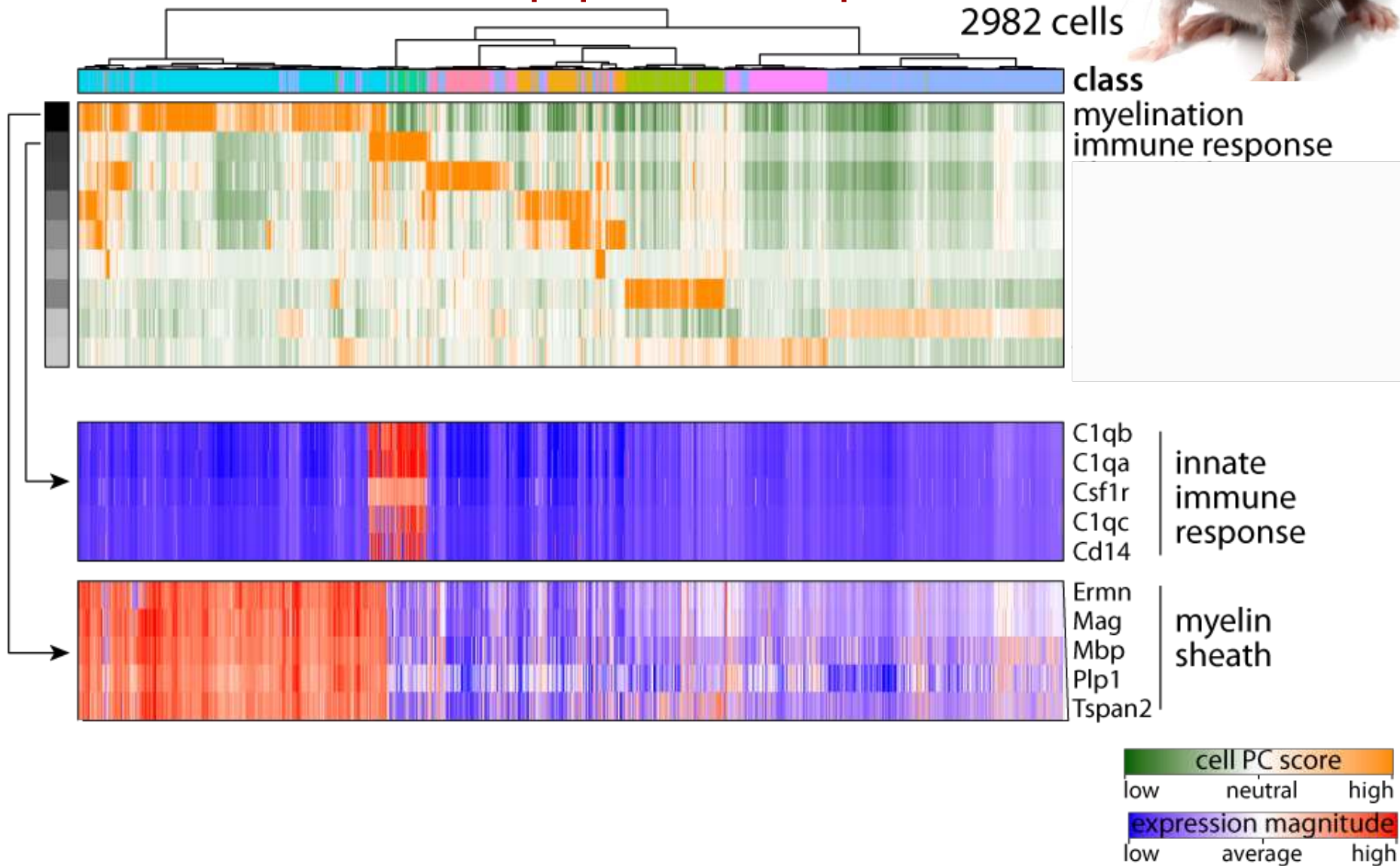
data from Zeisel et al., Science 2015



Heterogeneity in Mouse Cortex and Hippocampus



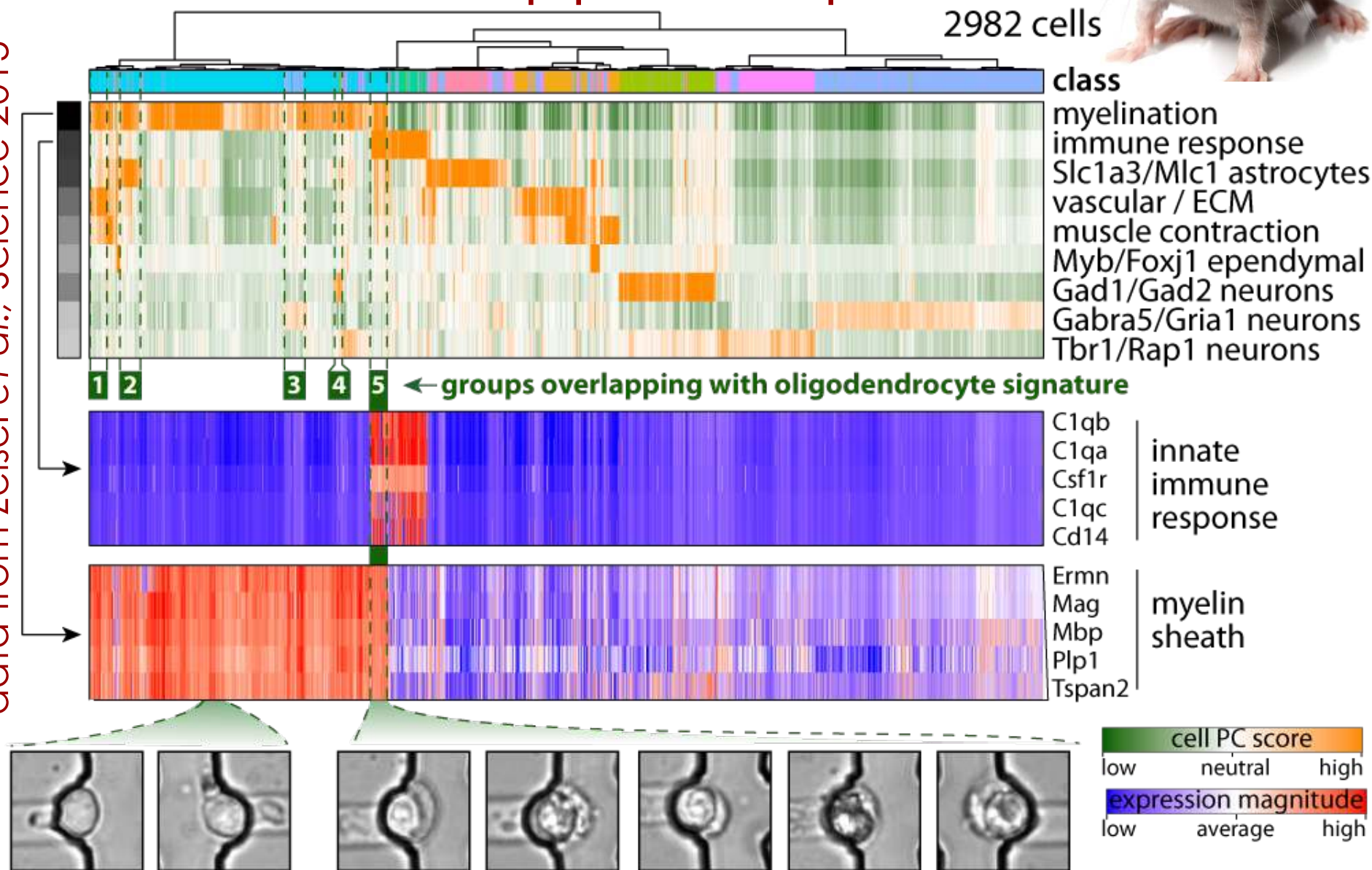
data from Zeisel et al., Science 2015



Heterogeneity in Mouse Cortex and Hippocampus



data from Zeisel et al., Science 2015

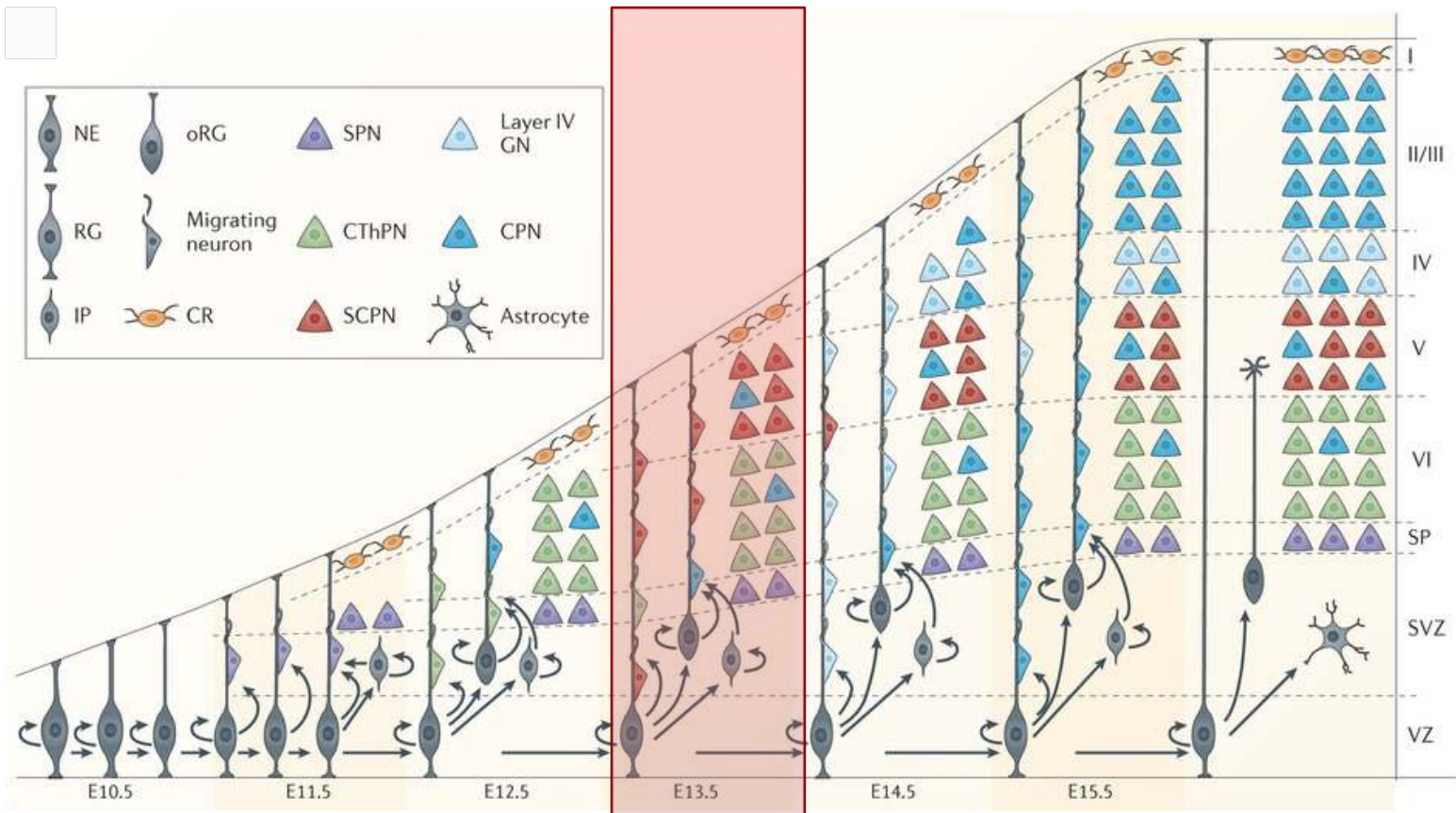


Cortical Neuron Progenitors



■ Formation of Cortical Layers

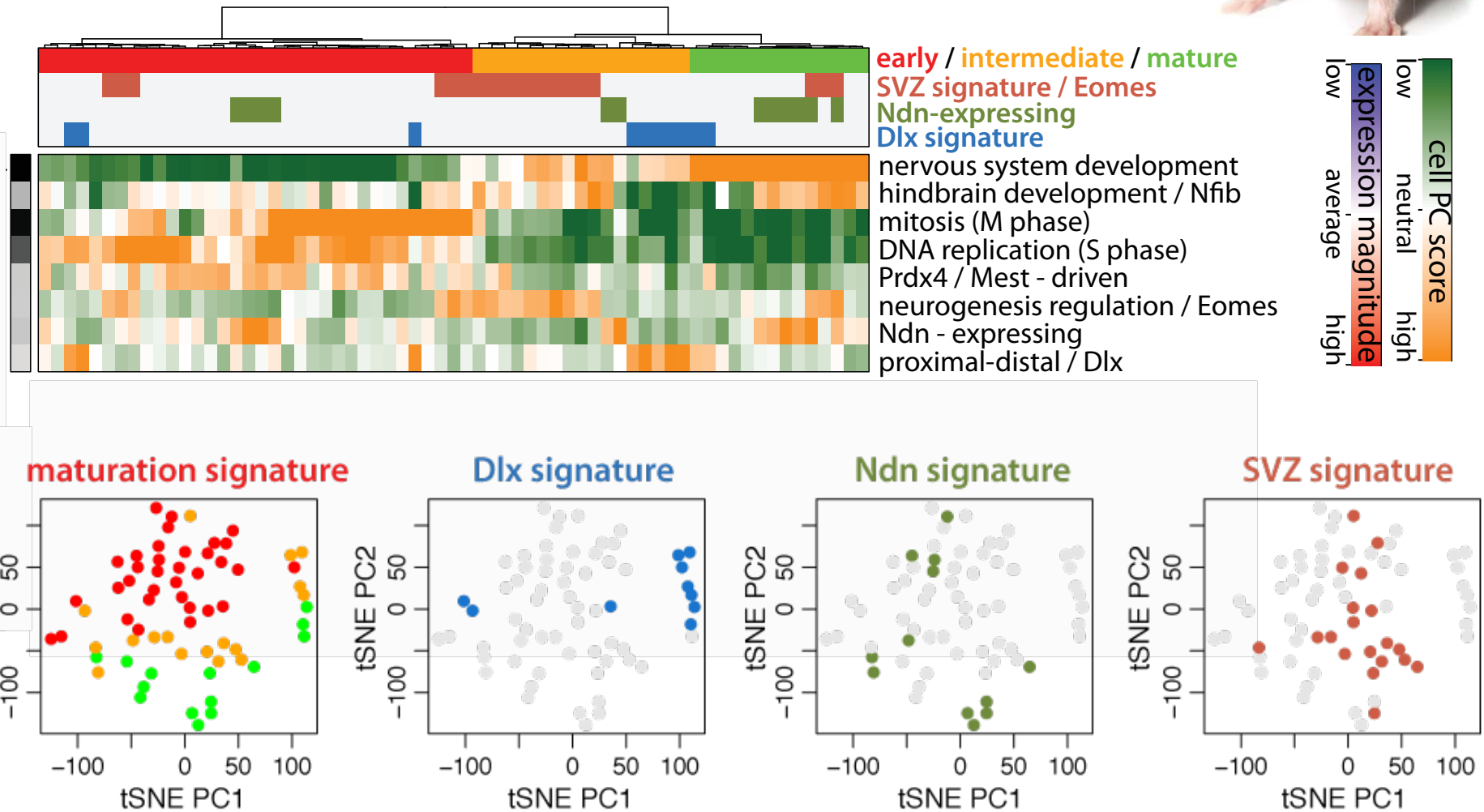
Greig et al. Nature Review Neuroscience 2013



■ Neuronal progenitor cells (NPCs)

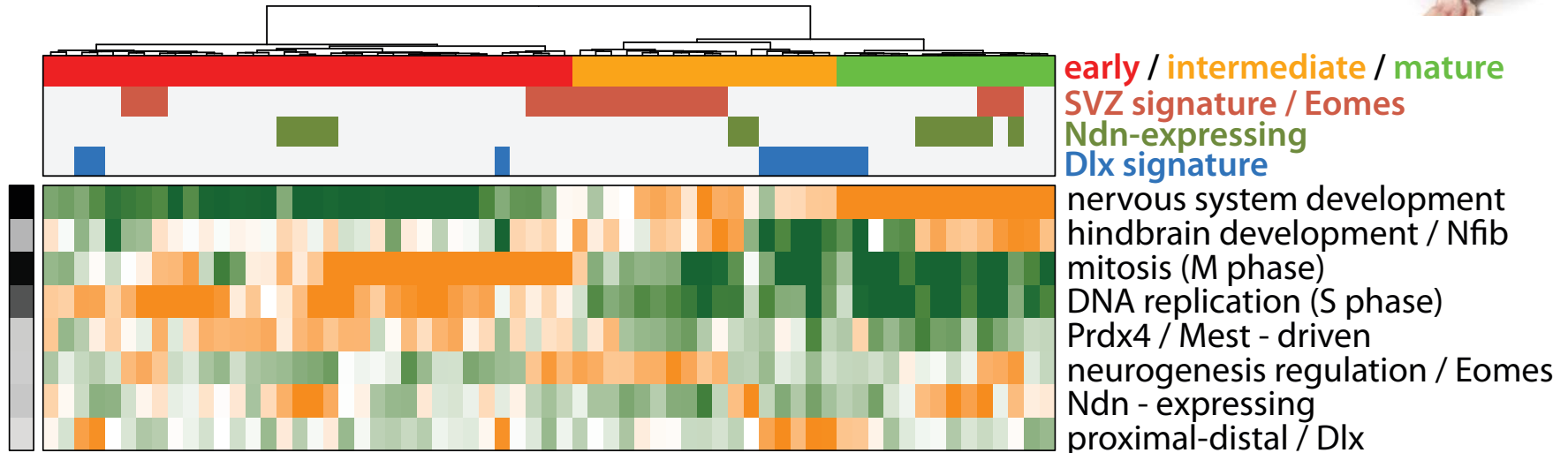
- Purified from E13.5. ~84 NPCs, 25 ASCs

More challenging cells: NPC



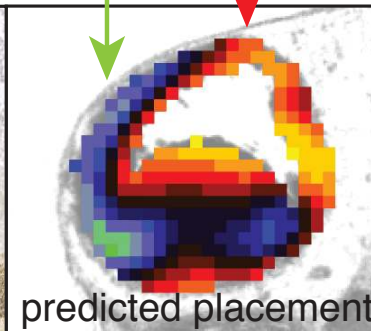
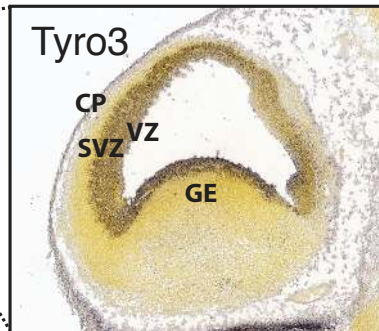
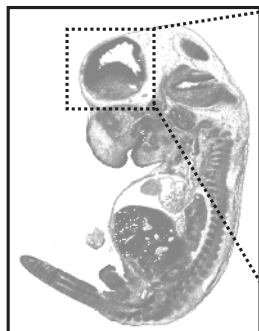
tSNE re-analysis

Cortical Neuron Progenitors



early

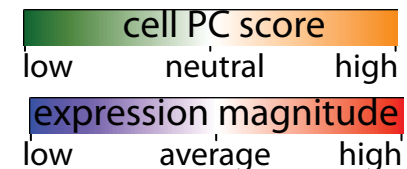
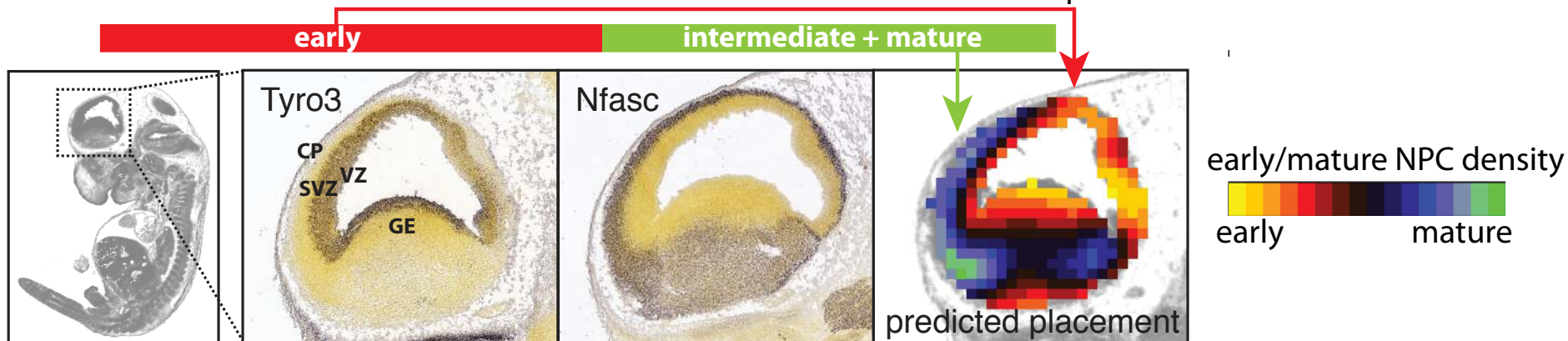
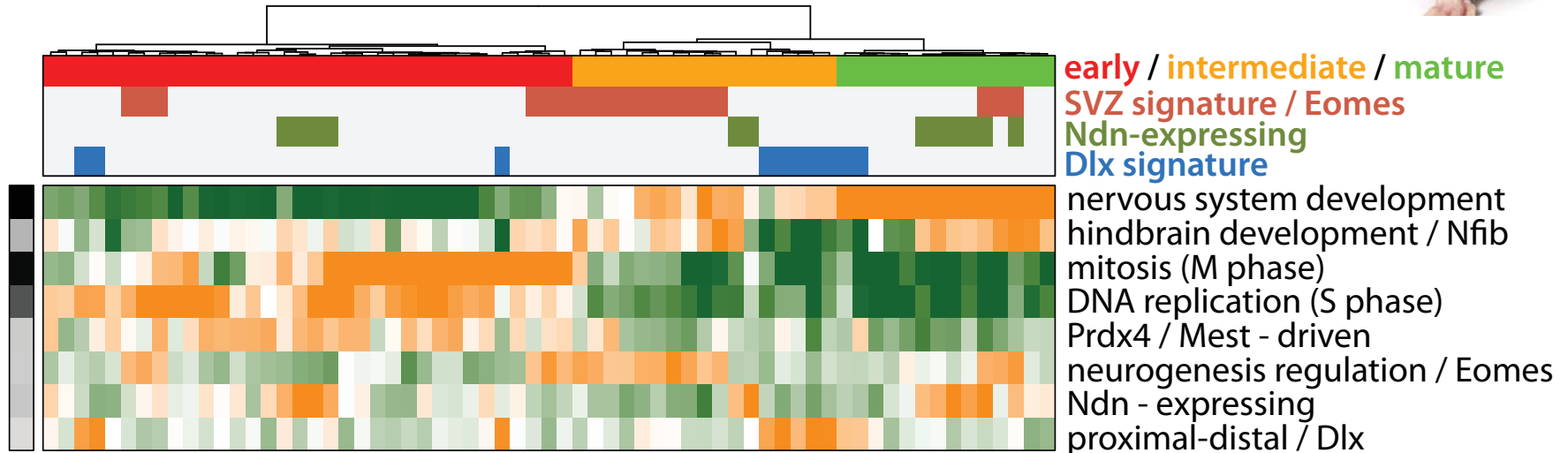
intermediate + mature

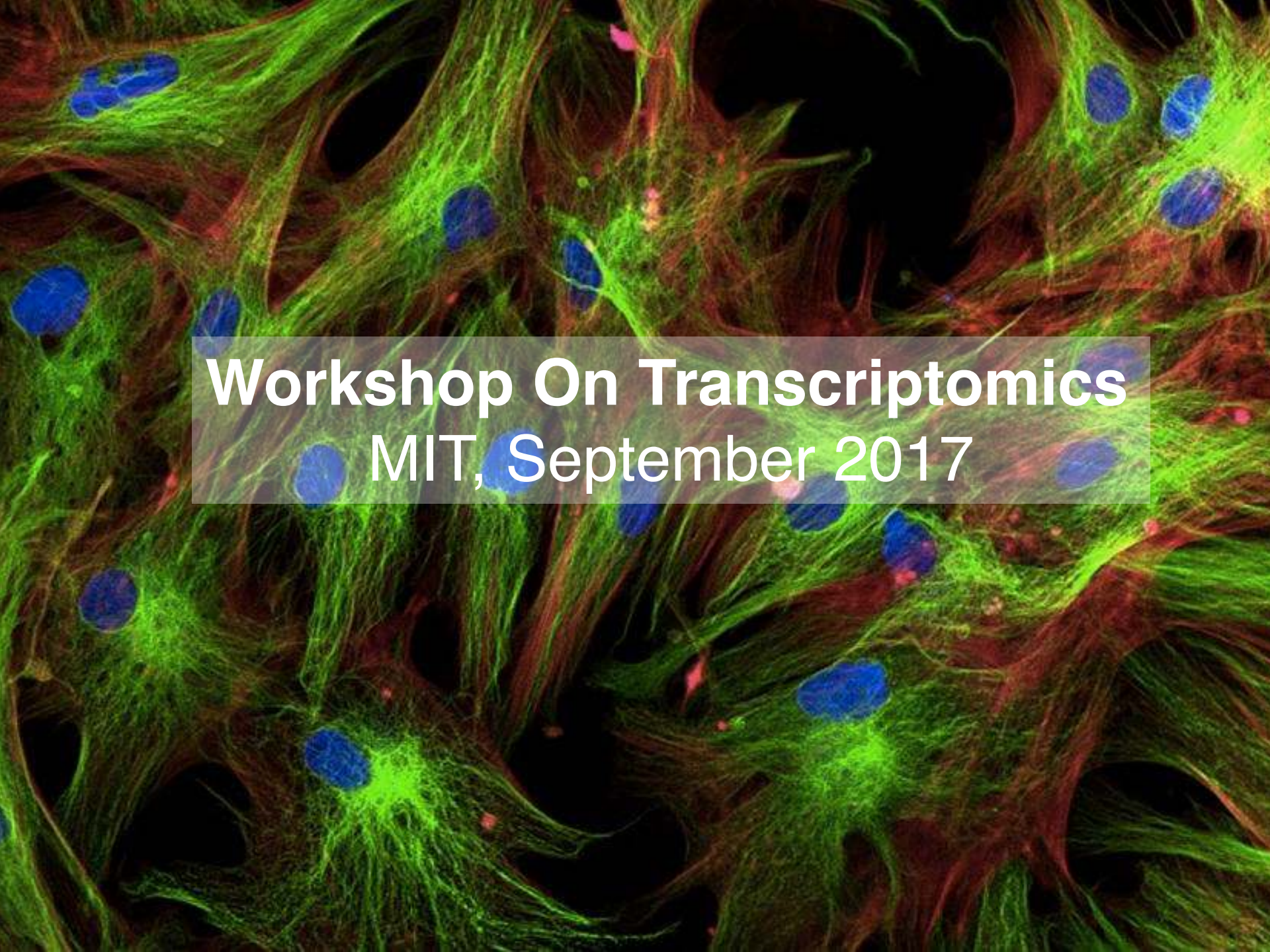


early/mature NPC density
early mature

cell PC score
low neutral high
expression magnitude
low average high

Cortical Neuron Progenitors



A fluorescence microscopy image showing a dense network of cells. The cytoskeleton is stained green, forming a complex web of filaments. The nuclei are stained blue, appearing as distinct, rounded structures within the cells. The background is black, highlighting the cellular structures.

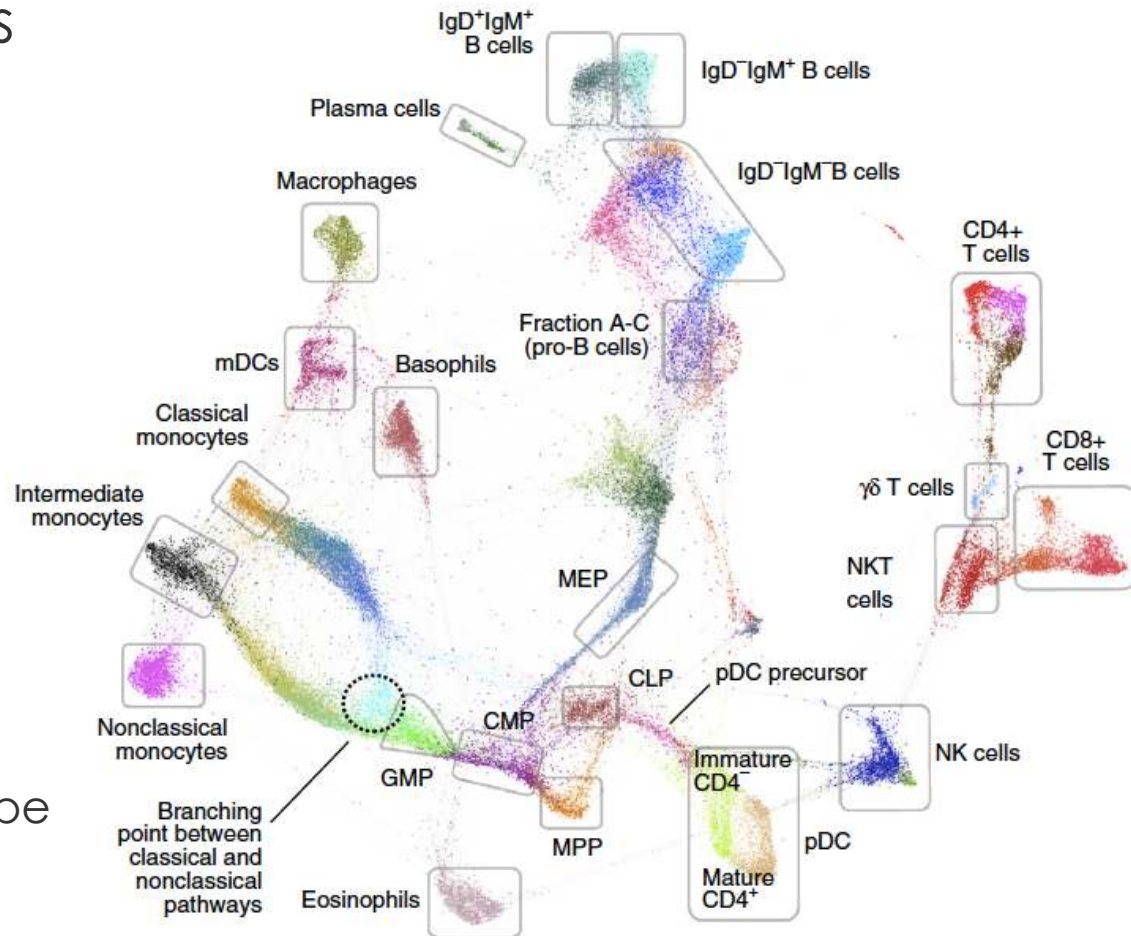
Workshop On Transcriptomics

MIT, September 2017

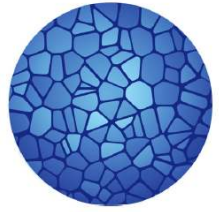
Cell Clusters: Units of Interpretation



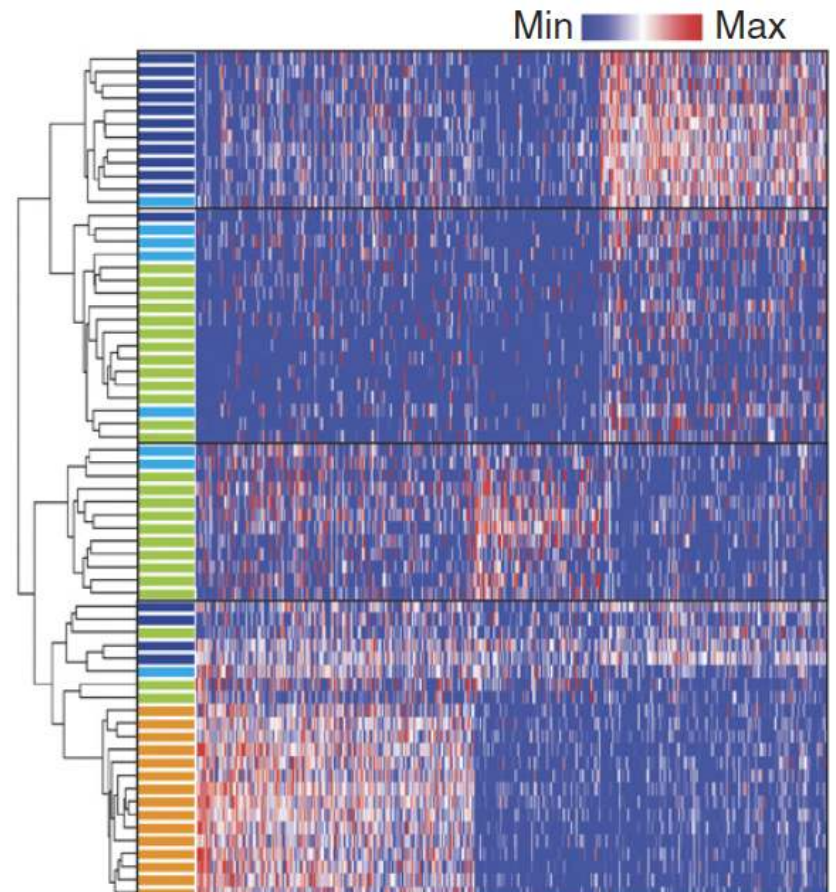
- Aim: Group similar cells
- Clusters are used for
 - Illustrating heterogeneity
 - Visualization, Navigation
 - Assigning interpretation
 - Estimating pooled signals
 - Differential expression
- Many clustering methods
 - Cluster assignments vary
 - Key clusters/groups tend to be stable



Clustering Approaches



- Direct clustering methods
 - Hierarchical clustering

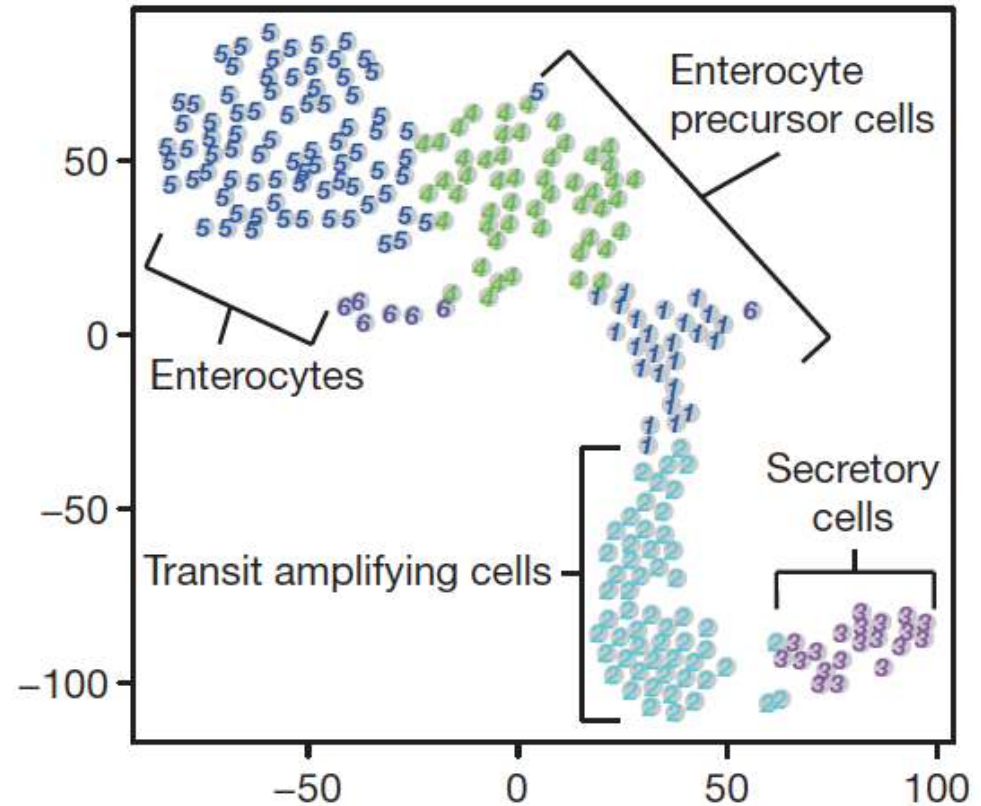


Pollen et al., Nat. Biotech '14

Clustering Approaches



- Direct clustering methods
 - Hierarchical clustering
 - K-means, PAM

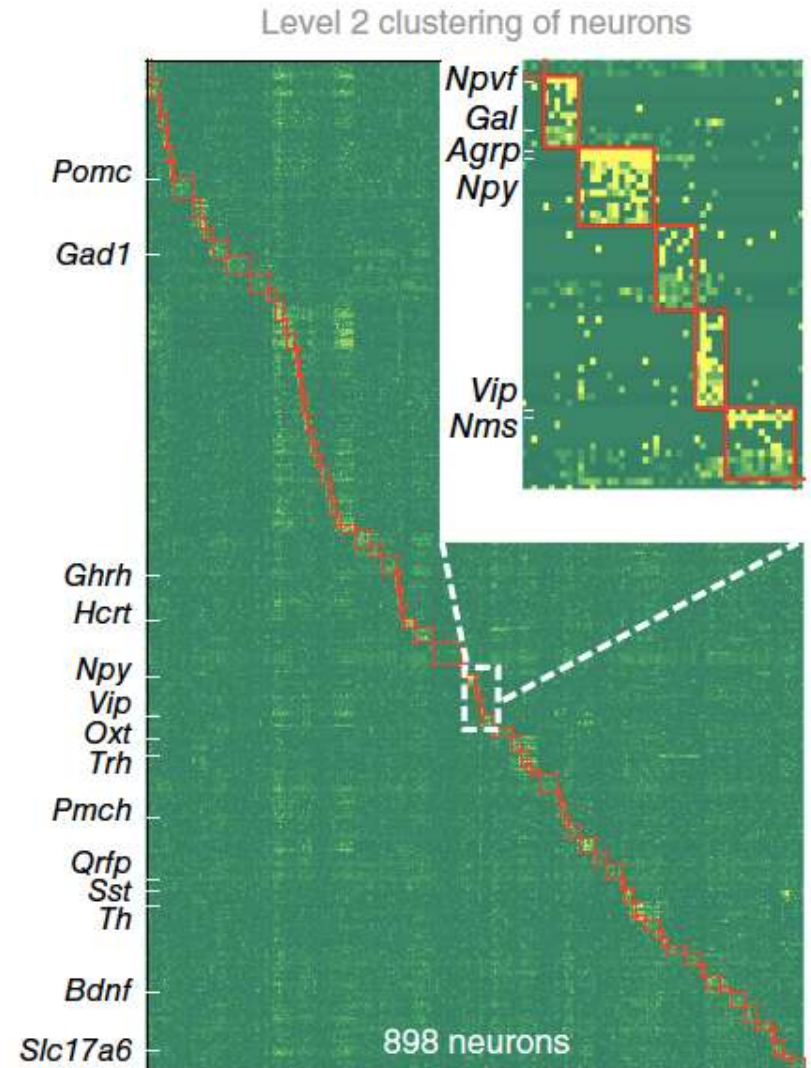


Grun et al., Nature '15

Clustering Approaches



- Direct clustering methods
 - Hierarchical clustering
 - K-means, PAM
 - BackSpin (Zeisel *et al.*, Science'15)

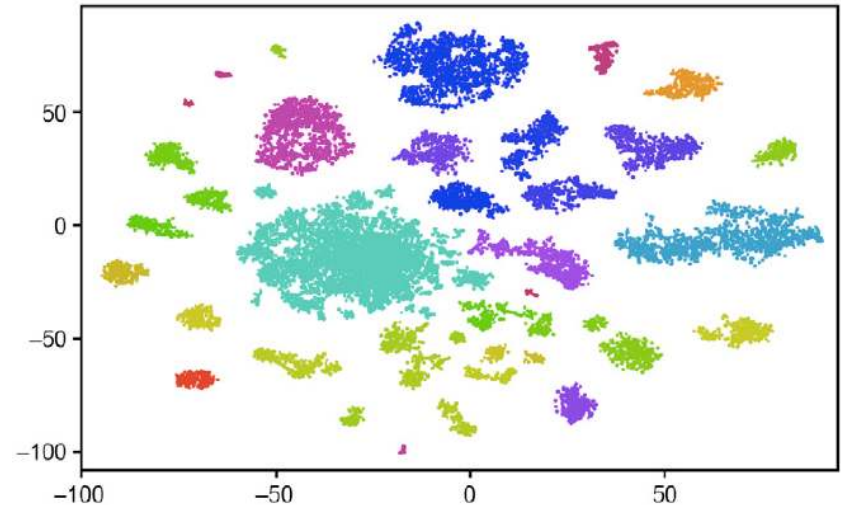


Romanov *et al.*, Nat Neuro '17

Clustering Approaches



- Direct clustering methods
 - Hierarchical clustering
 - K-means, PAM
 - BackSpin (Zeisel *et al.*, Science'15)
- Density clustering
 - Find cell density clumps
 - Embed cells into low dimensions

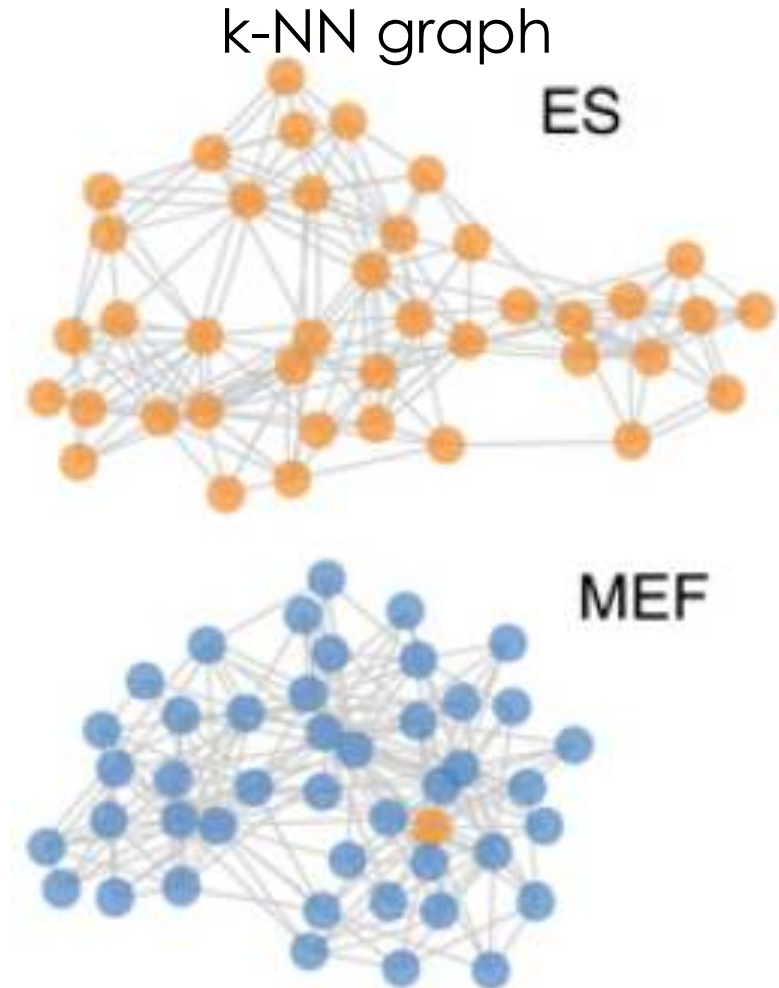


Macosko *et al.*, Cell'15

Clustering Approaches



- Direct clustering methods
 - Hierarchical clustering
 - K-means, PAM
 - BackSpin (Zeisel *et al.*, Science'15)
- Density clustering
 - Find cell density clumps
 - Embed cells into low dimensions
- Graph-based clustering
 - k-nearest neighbor graph
 - Community detection methods
 - Modularity, Edge betweenness, Laplacian eigenvectors, InfoMap
 - Phenograph, SLM (Seurat)

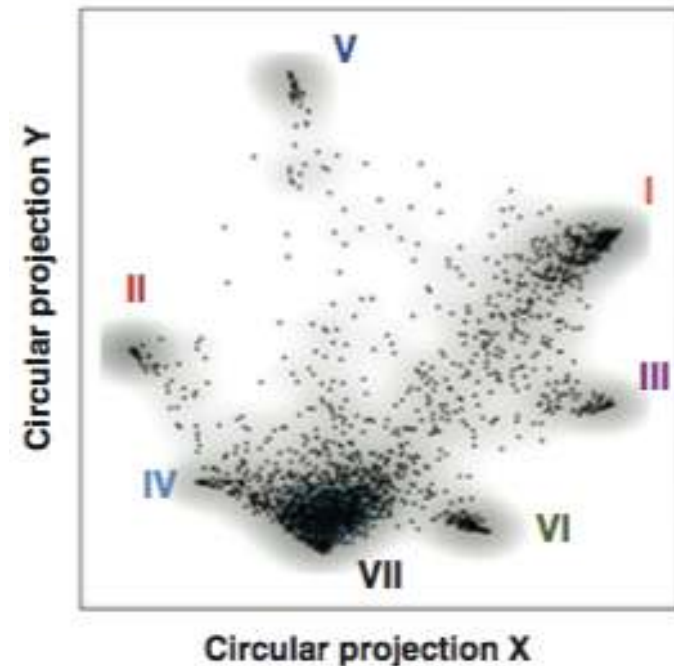
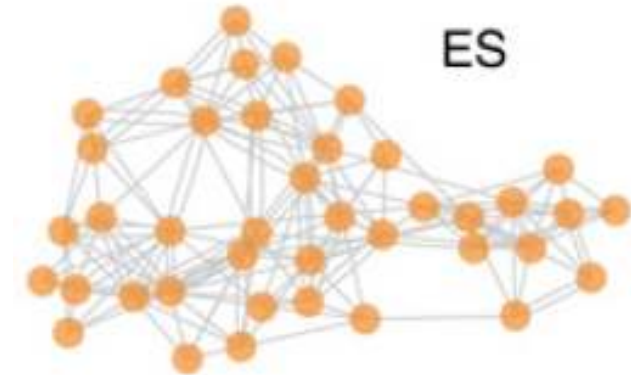


Islam *et al.*, Gen. Res'11

Clustering Approaches



- Traditional distance measures
 - Euclidian, L1, Correlation, Canberra, Jensen-Shannon
 - Transformed expression values
 - log scale
 - Reduced dimensions
 - Graph-based (scTDA)
 - Multi-scale distance (SIMLR)
 - Down-weighting of drop-outs
 - Restricted gene sets
- Model-based distances
 - Poisson
- Biological significance unclear



Islam et al., Gen. Res'11

Jaitin et al., Science'14

Clustering Approaches



Distance interpretation:

1. Deviation from equality

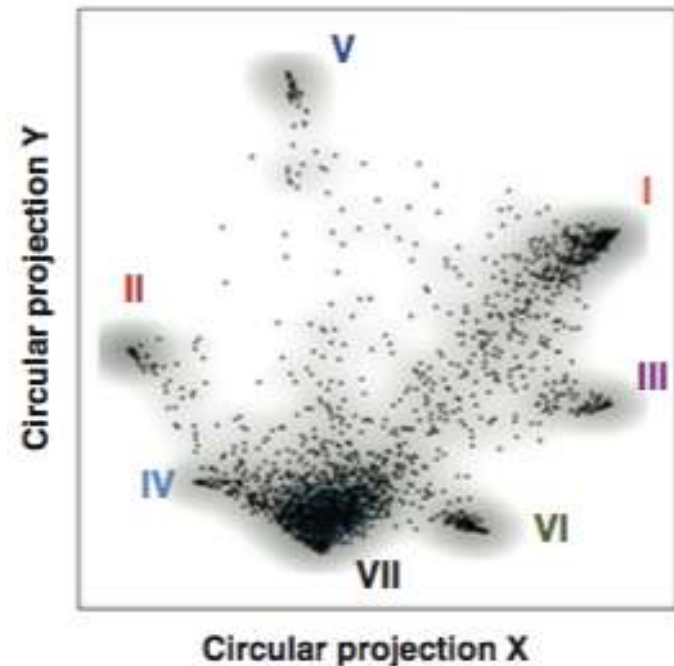
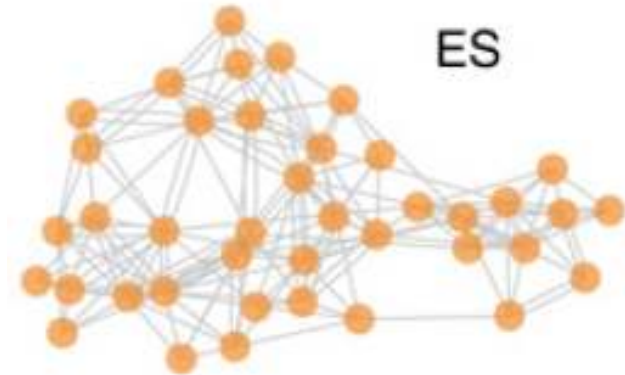
- e.g. Jensen-Shannon. Larger cells will appear more distant than cells with few molecules
- e.g. Poisson. Change in a highly-expressed gene can drive the overall likelihood
- Relative distances of other distinct clusters

2. Amount of transcriptional change

- e.g. L1. All genes are equal?
- Low-dimensional dangers
- Likelihood of transcriptional change
 - e.g. Weighting by observed variation
 - Stability, Assessing distant transitions

3. “Biological” distance

- Extent of phenotypic difference



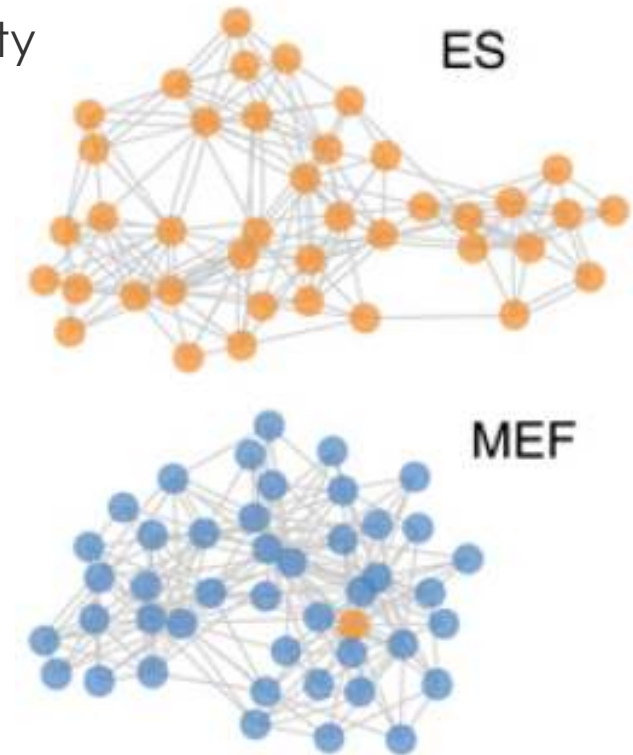
Islam et al., Gen. Res'11

Jaitin et al., Science'14

Challenge: Is that a real cluster?



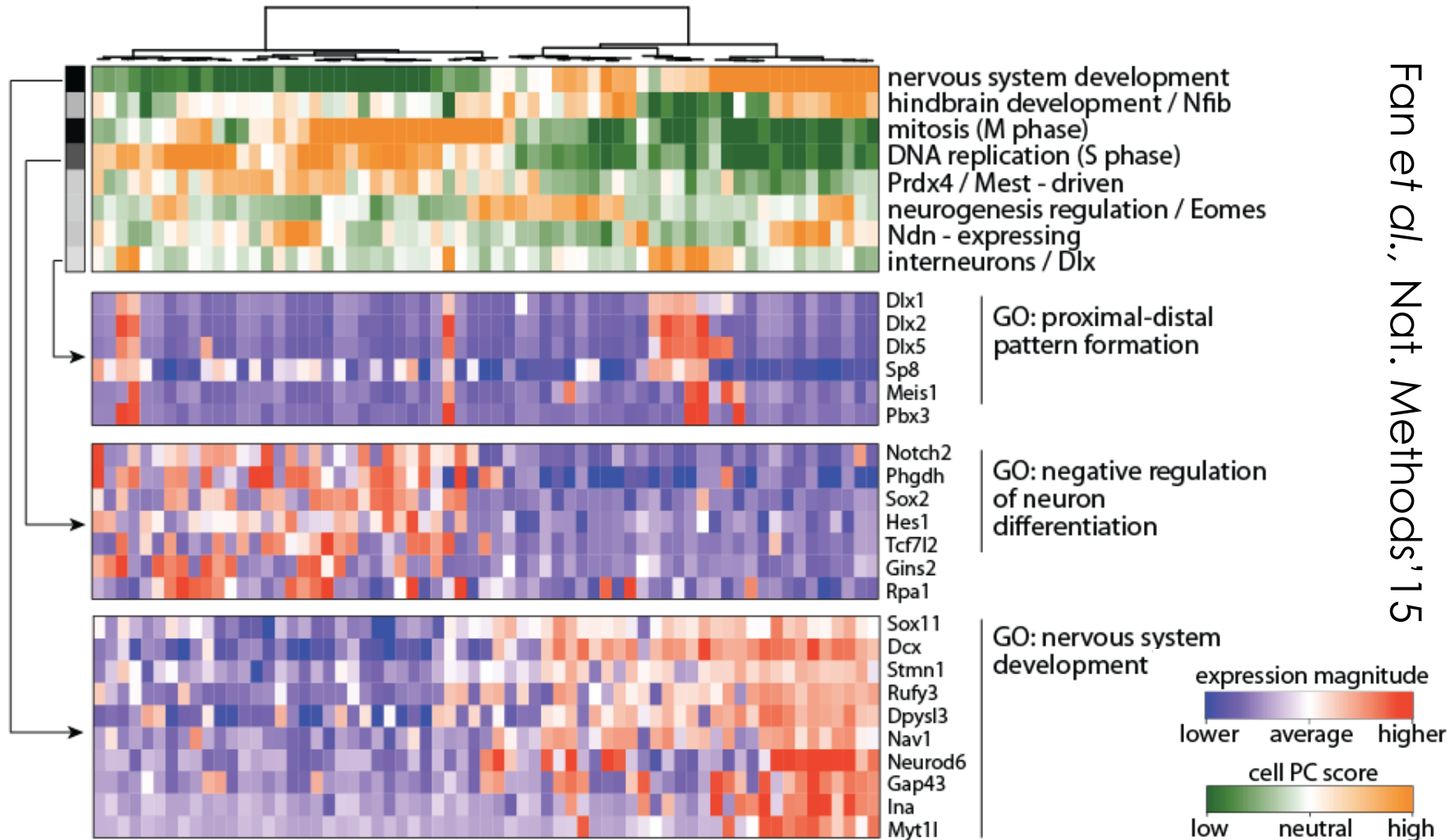
- Should a given subpopulation be split?
 - Cluster granularity varies between methods
 - Hierarchical clustering: improved stability
- Assessing statistical significance
 - Likelihood of observing the cluster
 - Under measurement noise
 - Cell subsampling
 - Across replicates
 - Robustness of expression signature
 - Gene subsampling
- Is there a distinct phenotype?



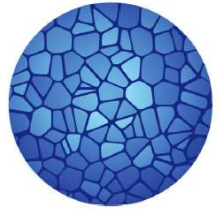
Challenge: Is that the whole story?



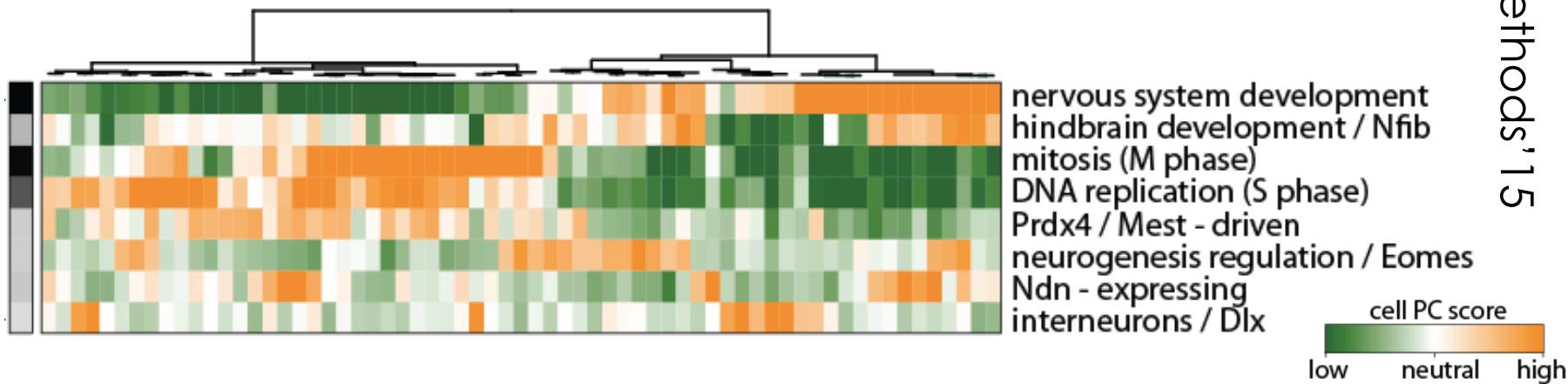
■ Cross-cutting classifications



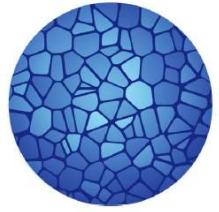
Challenge: Is that the whole story?



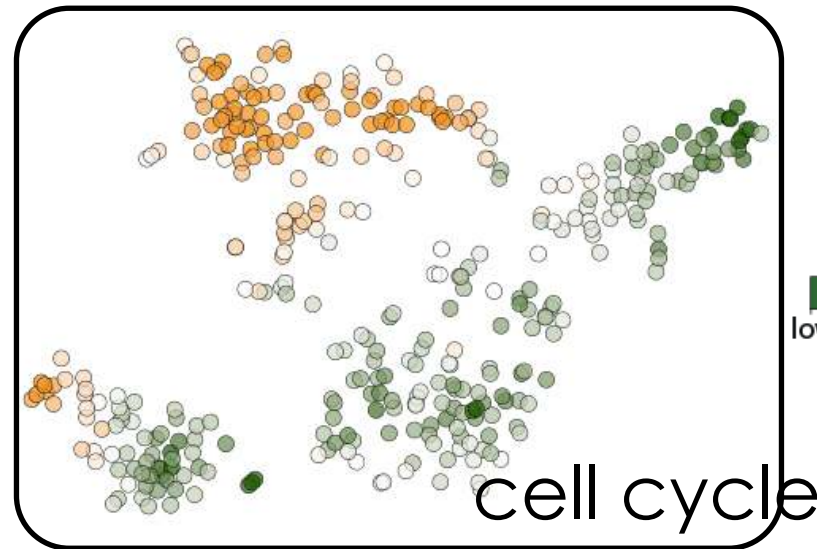
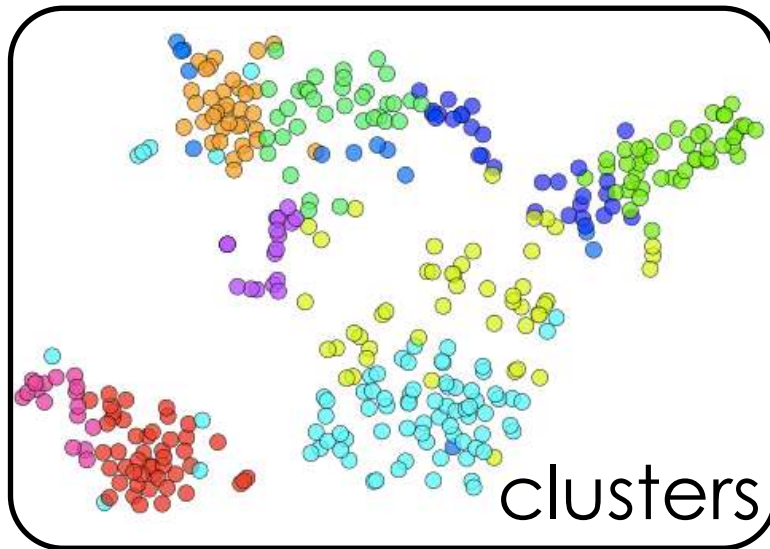
- Cross-cutting classifications
 - Cells have many physiologically-important properties
 - Clustering segments by some subset of properties
 - Limit -> by a combination of all properties
- How to identify and **visualize** cluster correspondence?



Challenge: Is that the whole story?



- Cross-cutting classifications
 - Cells have many physiologically-important properties
 - Clustering segments by some subset of properties
 - Limit -> by a combination of all properties
- How to identify and **visualize** cluster correspondence?
- Preferred configuration depends on a biological question



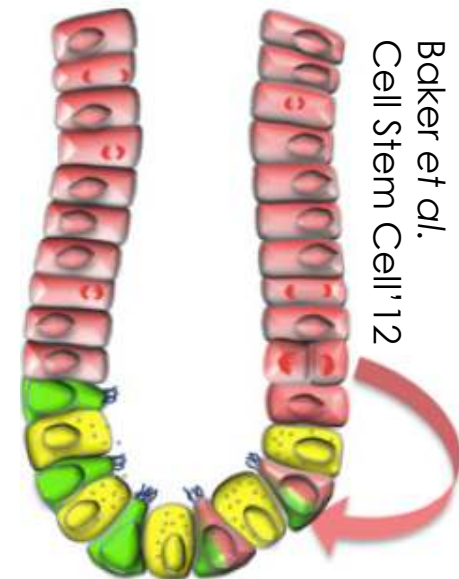
cell PC score
low neutral high

Cell Type vs. Cell State



“Cellular states span configuration space within which cells of a given type can move in a reversible manner under physiological conditions”

- Discrete / enumerable states - local maxima within the state configuration space
- Common processes can give rise to similar transcriptional states within different cell types (e.g. cell cycle phases)
- Types are separated by irreversible transitions*
 - Rare reversible events
- Intended as a guideline, not a strict definition



Fast Processing of Sparse Measurements

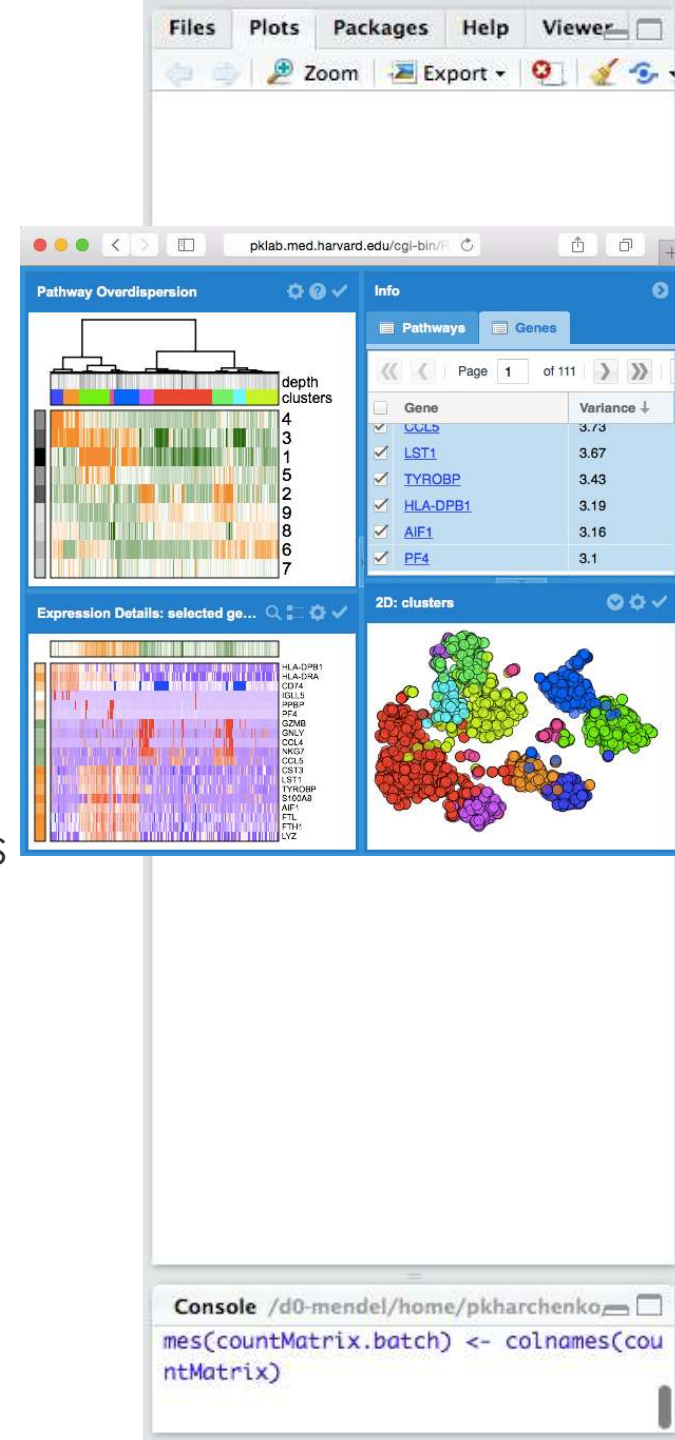


- Current single-cell datasets
 - 10^4 - 10^6 cells
 - With molecular barcoding
 - As few as a hundred molecules per cell

Fast Processing of Sparse Measurements

- Current single-cell datasets
 - 10^4 - 10^6 cells
 - With molecular barcoding
 - As few as a hundred molecules per cell
- Pagoda2
 - Optimized for large sparse measurements
 - Residualizing, Winsorization, corrections
 - Clustering estimation
 - PCA
 - Differential expression
 - Clustering aspects / states
 - Gene set/pathway analysis
 - Differential Expression

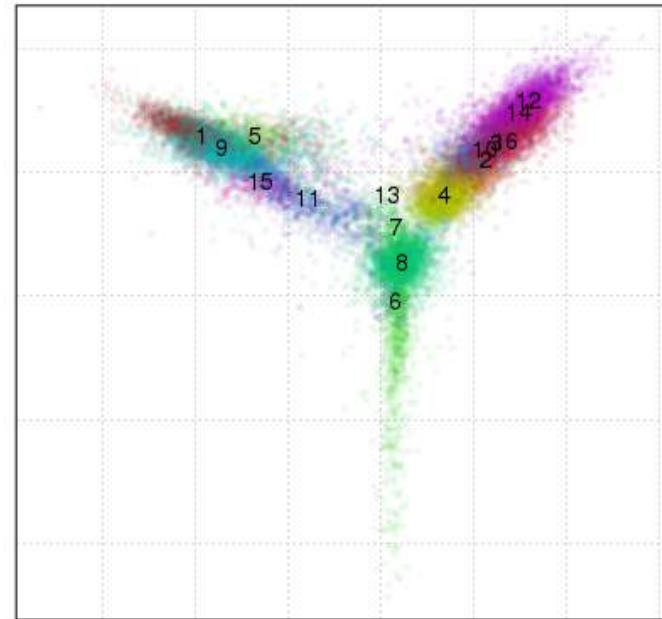
<http://github.com/hms-dbmi/pagoda2/>



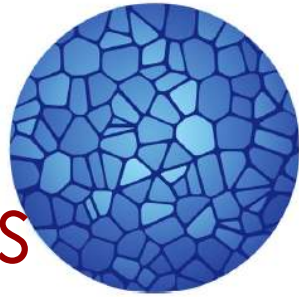
Fast Processing of Sparse Measurements: PCA



- Run PCA on a sparse matrix with:
 - $\sim 10^3$ - 10^4 variables
 - $\sim 10^5$ - 10^6 observations
- Lanczos algorithm
 - $x_{n+1} = Ax_n ; x_n / \|x_n\|$
 - Krylov subspaces
 - $\mathcal{K}_r(A, b) = \text{span} \{b, Ab, A^2b, \dots, A^{r-1}b\}$.
 - Orthogonalize Krylov space of A to obtain Arnoldi vectors, which are approximations of the largest eigenvectors
 - Restart with multiple random vectors b
 - Implemented by irlba package
 - 10 largest eigenvectors on 10^6 cells by 10^3 features:

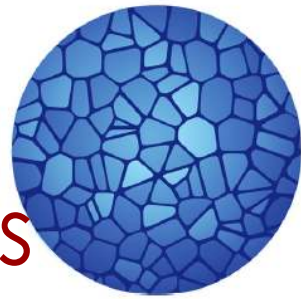


Fast Processing of Sparse Measurements: nearest neighbors

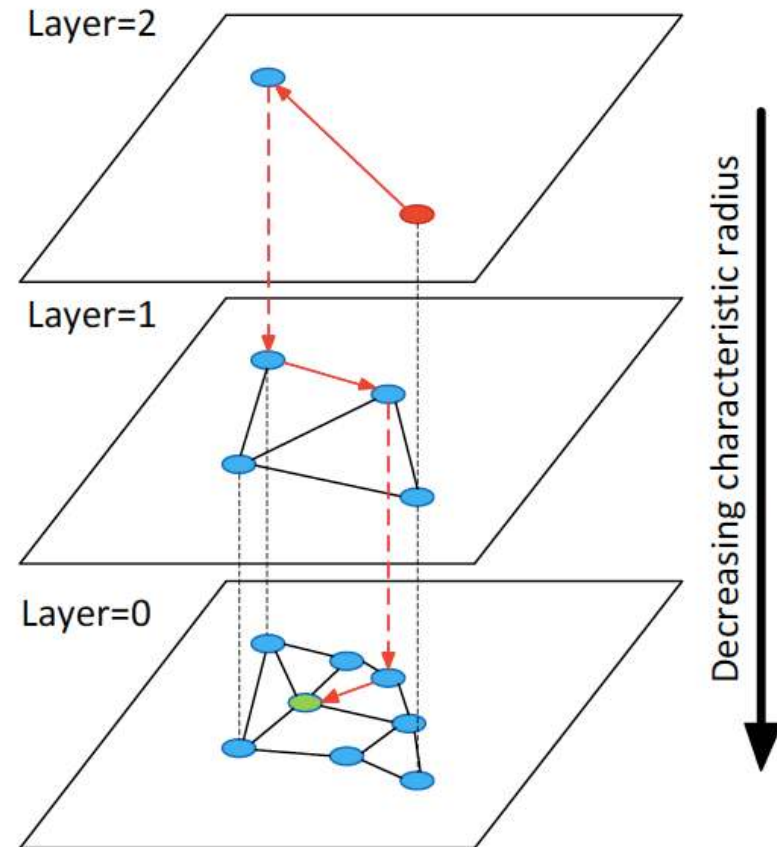


- Looking for nearest k (20-100) neighbors for $\sim 10^5$ - 10^6 cells
- A relatively common problem
 - Typically < 1000 variables
 - Approximate search is OK
 - Tree-based approaches (kd, PCA)
 - Locale-sensitive Hashing

Fast Processing of Sparse Measurements: nearest neighbors



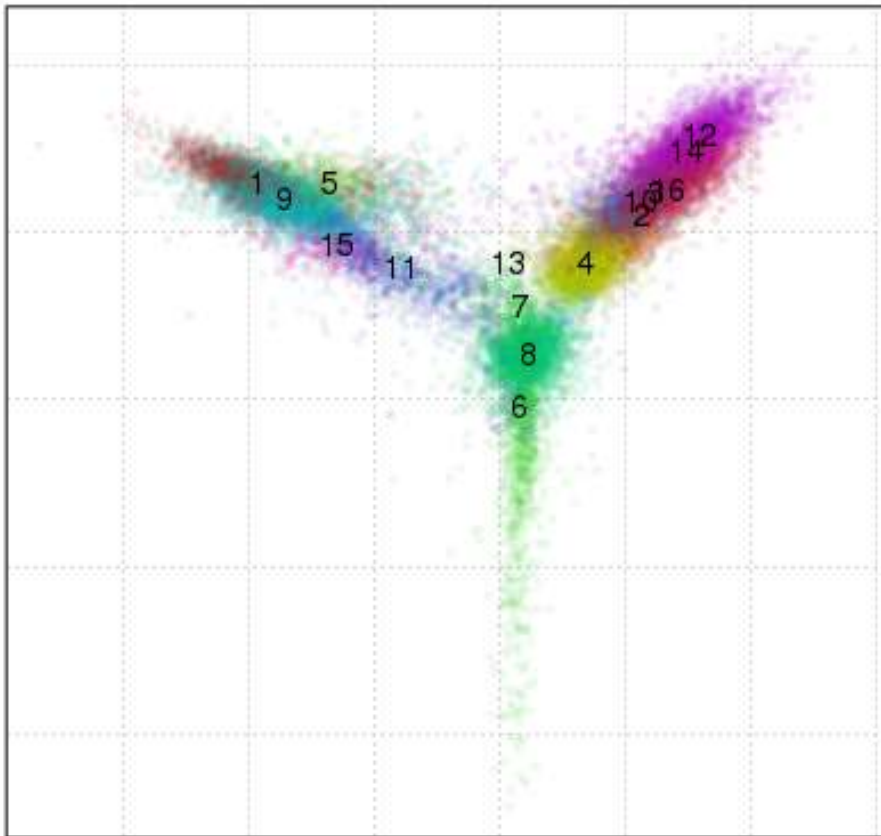
- Looking for nearest k (20-100) neighbors for $\sim 10^5$ - 10^6 cells
- A relatively common problem
 - Typically < 1000 variables
 - Approximate search is OK
 - Tree-based approaches (kd, PCA)
 - Locale-sensitive Hashing
- Navigable Small World Graphs
 - Malkov, Yashunin'16
 - Two-phase: graph insertion, search
 - Cell-cell correlation networks are great small-world graphs
 - Based on nmslib implementation



Fast Processing of Sparse Measurements: cluster, layout



- Interactive exploration is key to getting biological insights
- Effective visualization is important





Stochastic neighbor embedding (SNE)

- Convert the high-dimensional Euclidean distances between data points into conditional probabilities that represent similarities

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

Gaussian:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- probability that x_i would pick x_j as its neighbor conditional on neighboring being picked in proportion to their probability density under a Gaussian centered at x_i with variance σ_i



Stochastic neighbor embedding (SNE)

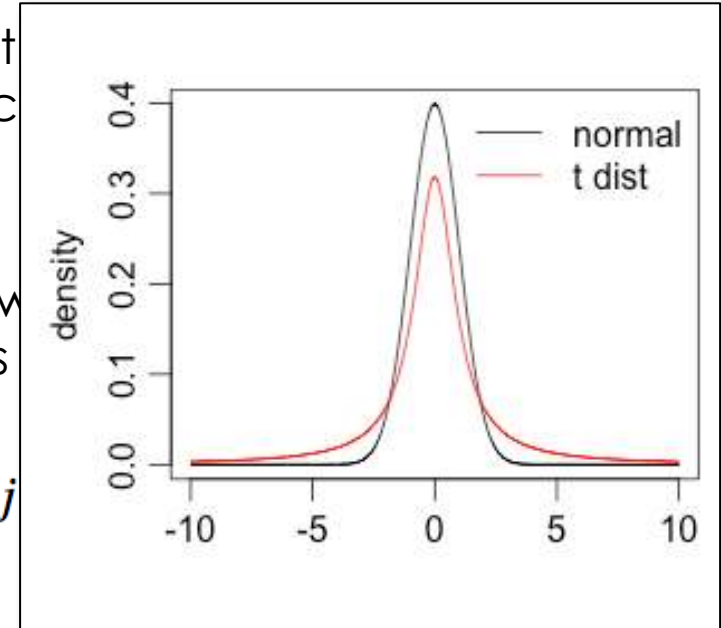
- Define a similar conditional probability / similarity metric for the low-dimensional counterparts y_i and y_j

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

t-SNE

- If the map points y_i and y_j correctly model the high-dimensional datapoints x_i and x_j , the $p_{j|i}$ and $q_{j|i}$ will be equal
- Optimize Kullback-Leibler divergences between probabilities $p_{j|i}$ and $q_{j|i}$ over all datapoints

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



Diffusion Maps



- Problems for PCA
 - Count values are noisy
 - Are not normally distributed
- Transformations of the count values
 - log
 - Imputation (MAGIC, scImpute)
 - Cell-cell similarity
- Diffusion probabilities

Diffusion maps for high-dimensional single-cell analysis of differentiation data

Laleh Haghverdi^{1,2}, Florian Buettner^{1,*} and Fabian J. Theis^{1,2,*}

“wavefunction”

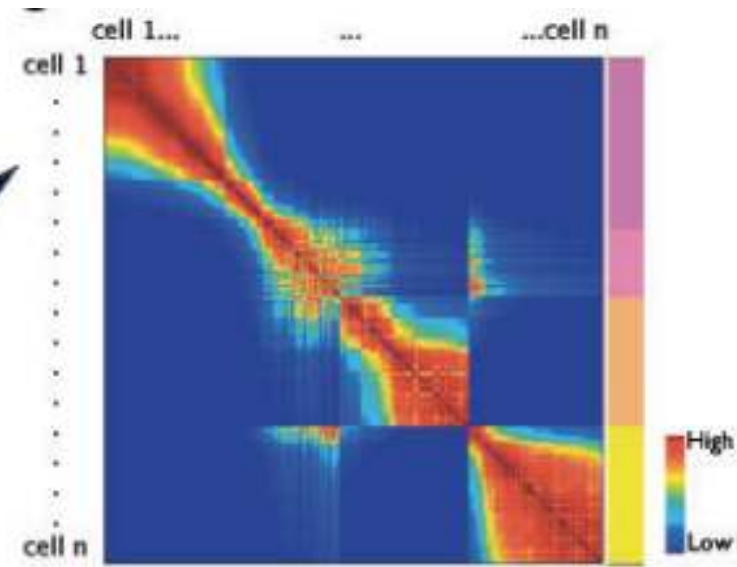
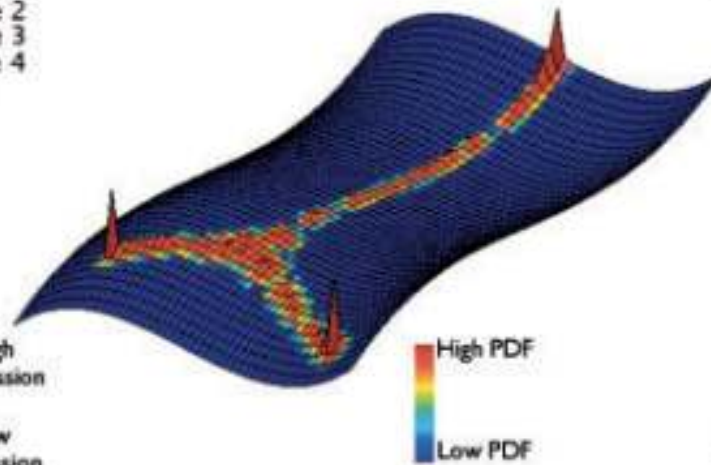
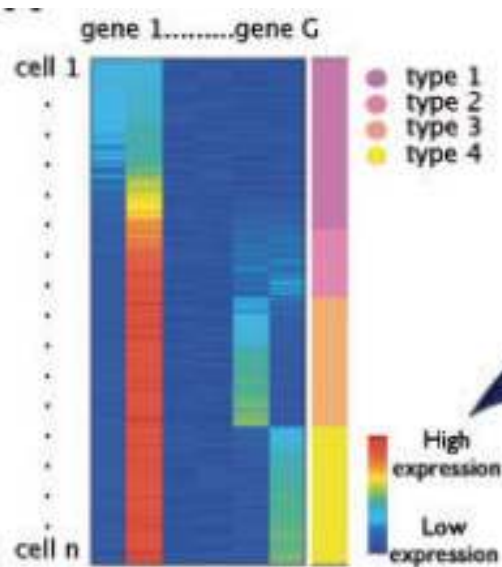
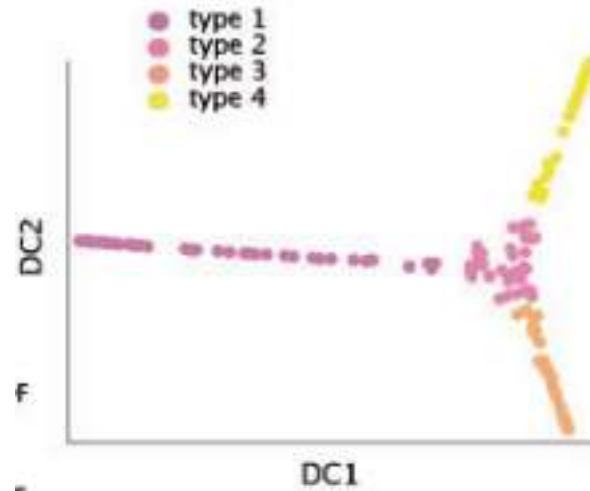
$$Y_x(x') = \left(\frac{2}{\pi\sigma^2}\right)^{1/4} \exp\left(-\frac{\|x' - x\|^2}{\sigma^2}\right)$$

“interference” -> diffusion p

$$P_{xy} = \frac{1}{Z(x)} \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

Diffusion Maps

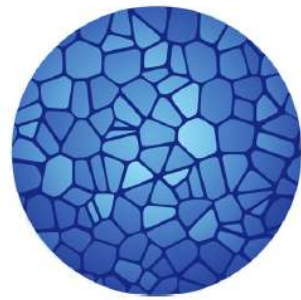
- Diffusion p
- PCA of the diffusion matrix
- Project on the first two PCs



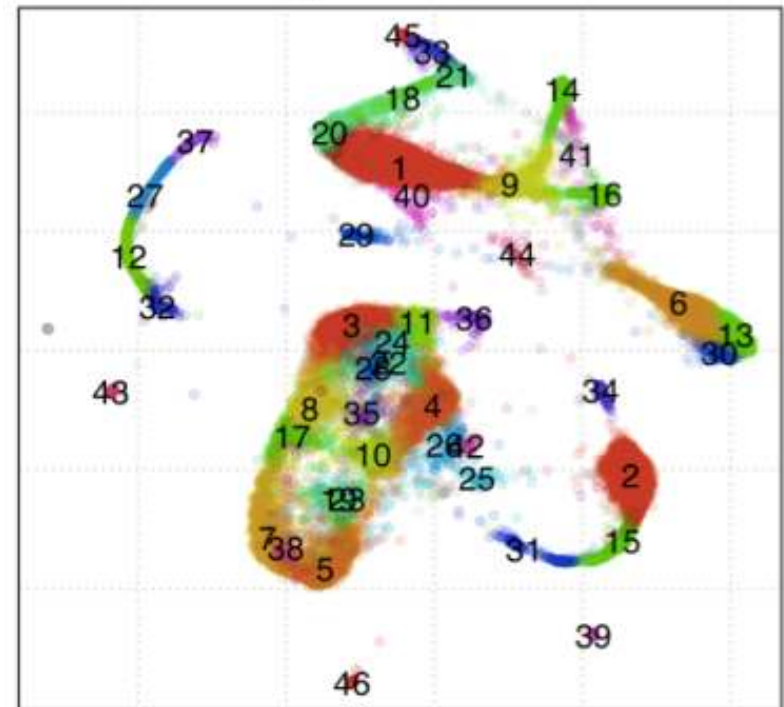
Diffusion maps for high-dimensional single-cell analysis of differentiation data

Laleh Haghverdi^{1,2}, Florian Buettner^{1,*†} and Fabian J. Theis^{1,2,*}

Fast Processing of Sparse Measurements: cluster, layout



- Interactive exploration is key to getting biological insights
- Effective visualization is important
- Popular embedding methods
 - PCA, MDS
 - tSNE
 - Graph layout



Cornell University
Library

arXiv.org > cs > arXiv:1602.00370

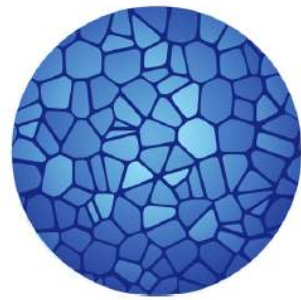
Computer Science > Learning

Visualizing Large-scale and High-dimensional Data

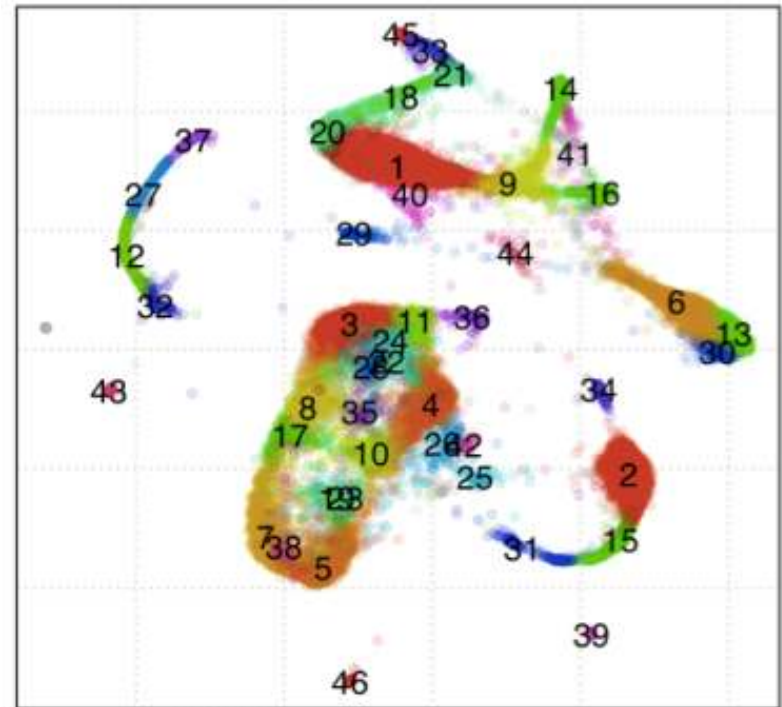
Jian Tang, Jingzhou Liu, Ming Zhang, Qiaozhu Mei

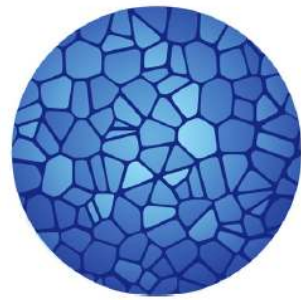
(Submitted on 1 Feb 2016 (v1), last revised 5 Apr 2016 (this version, v2))

Fast Processing of Sparse Measurements: cluster, layout



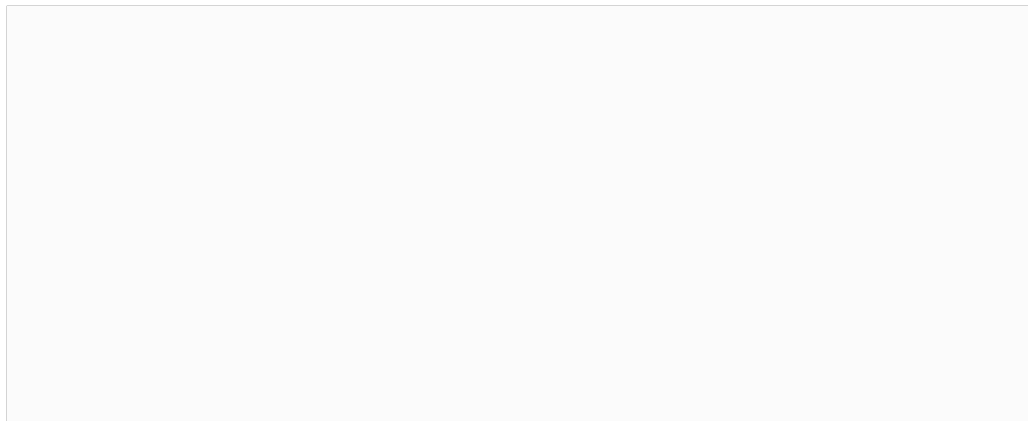
- Interactive exploration is key to getting biological insights
- Effective visualization is important
- Popular embedding methods
 - PCA, MDS
 - tSNE
 - Graph layout
- Clustering
 - Density clustering
 - Graph community clustering
 - Common problem
 - Multilevel community detection
 - Near-linear performance on large sparse graphs





Making use of reference data

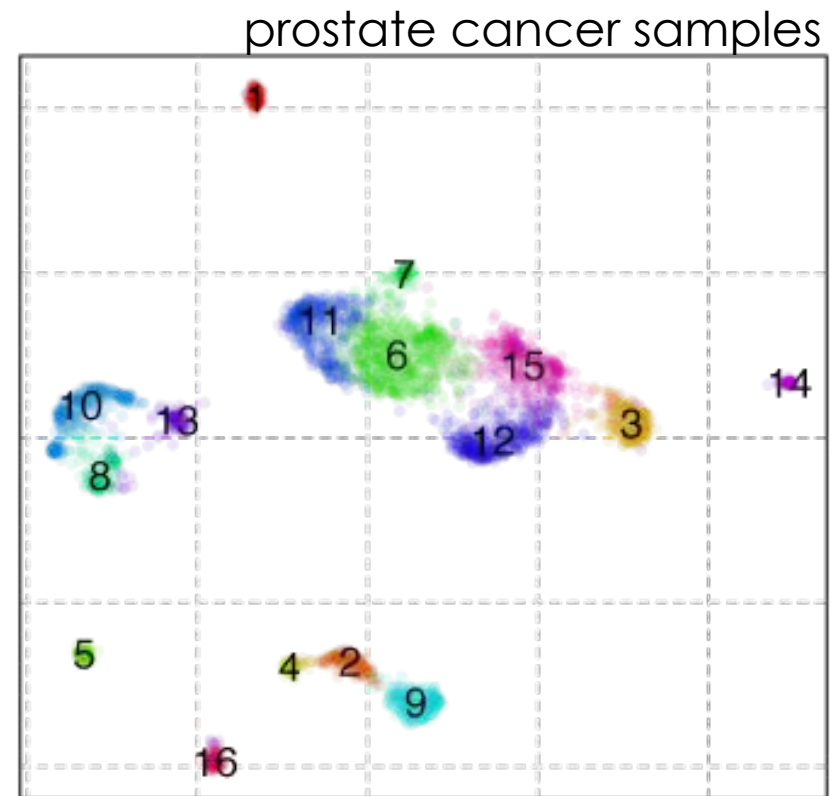
- Reference cell collections
 - Organism, tissue-specific data
- *De novo* vs. reference-based analysis
- Reference-augmented
 - Will capture both dataset-specific and reference-based signatures
 - Facilitates interpretation
 - Higher statistical power, stability



Making use of reference data



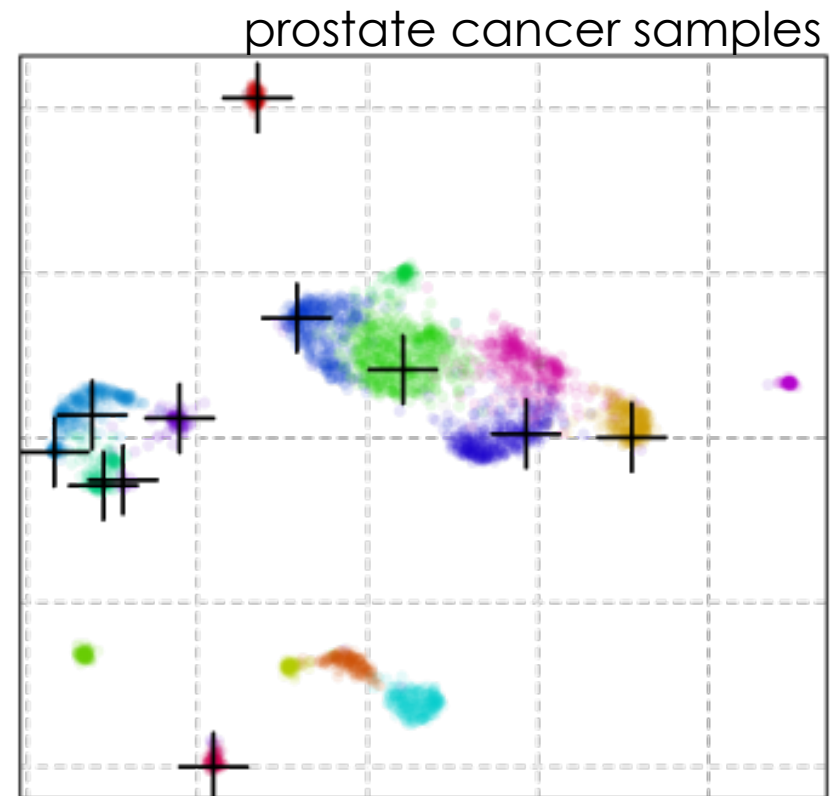
- Reference cell collections
 - Organism, tissue-specific data
- *De novo* vs. reference-based analysis
- Reference-augmented
 - Will capture both dataset-specific and reference-based signatures
 - Facilitates interpretation
 - Higher statistical power, stability



Making use of reference data



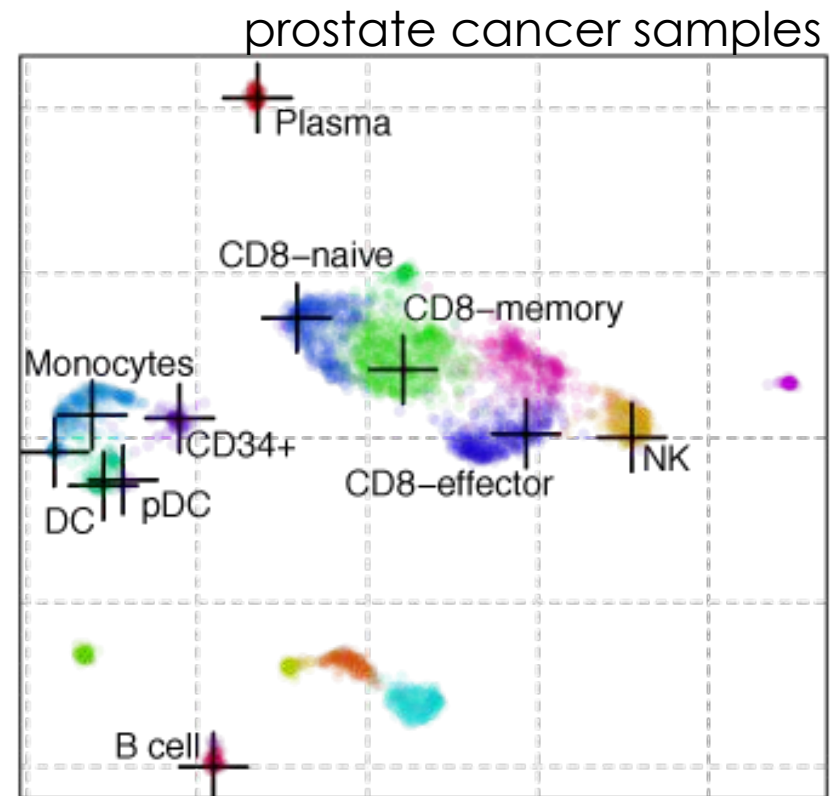
- Reference cell collections
 - Organism, tissue-specific data
- *De novo* vs. reference-based analysis
- Reference-augmented
 - Will capture both dataset-specific and reference-based signatures
 - Facilitates interpretation
 - Higher statistical power, stability



Making use of reference data



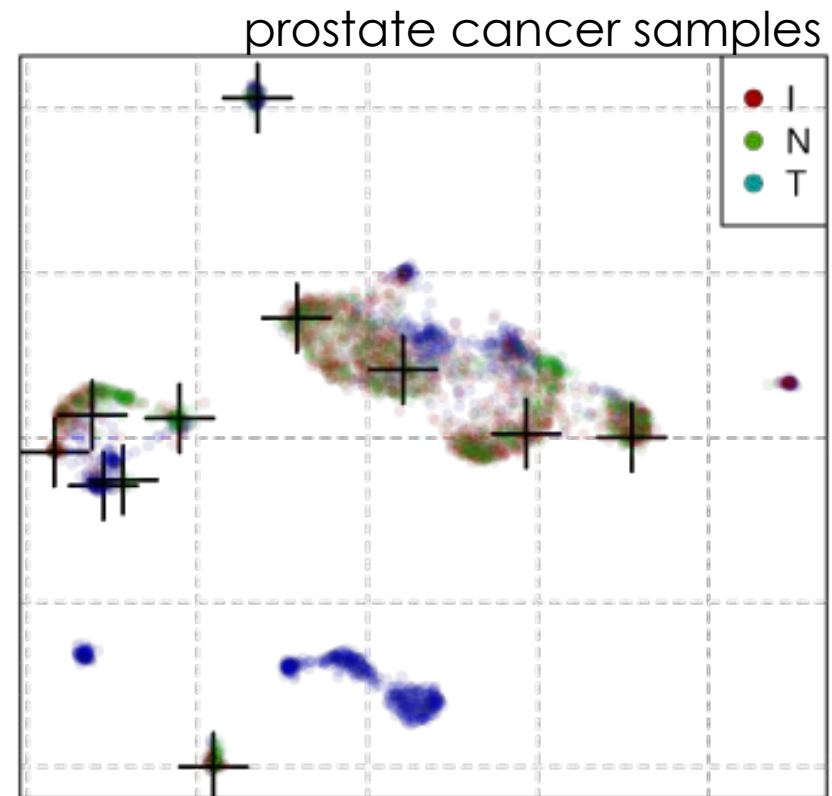
- Reference cell collections
 - Organism, tissue-specific data
- *De novo* vs. reference-based analysis
- Reference-augmented
 - Will capture both dataset-specific and reference-based signatures
 - Facilitates interpretation
 - Higher statistical power, stability



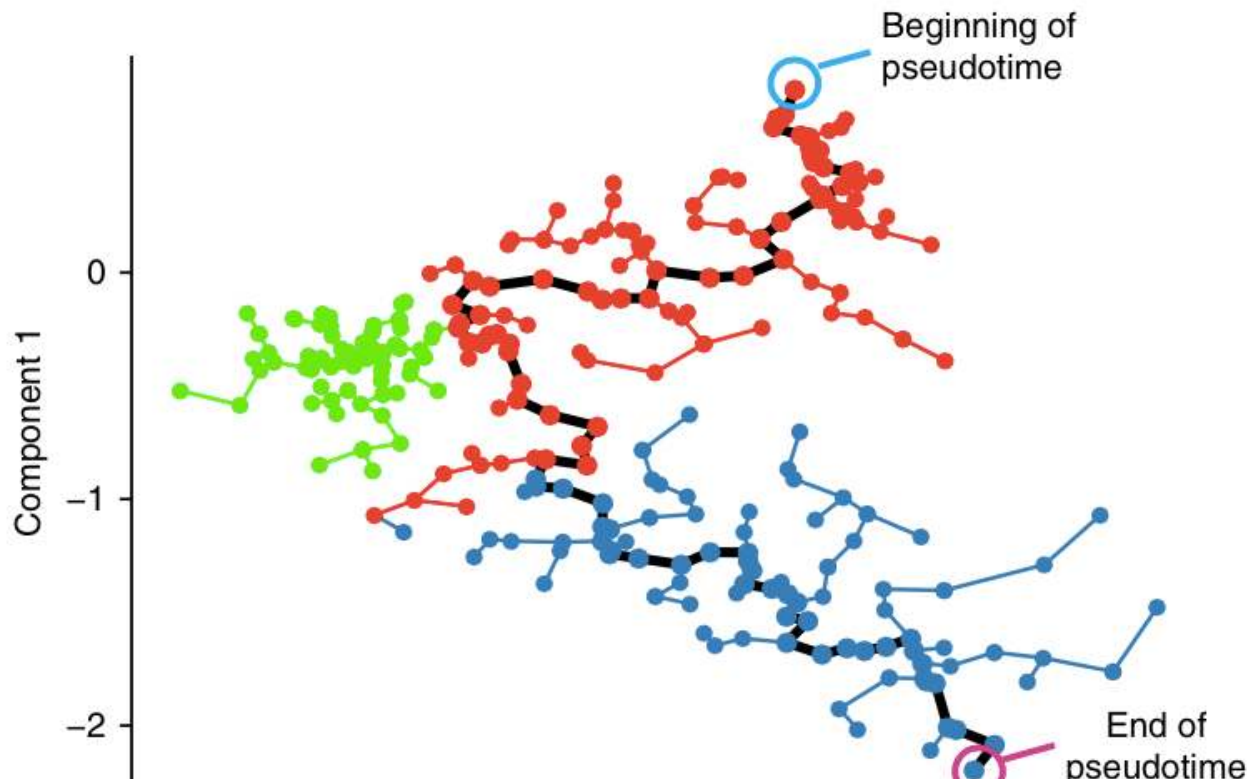
Making use of reference data



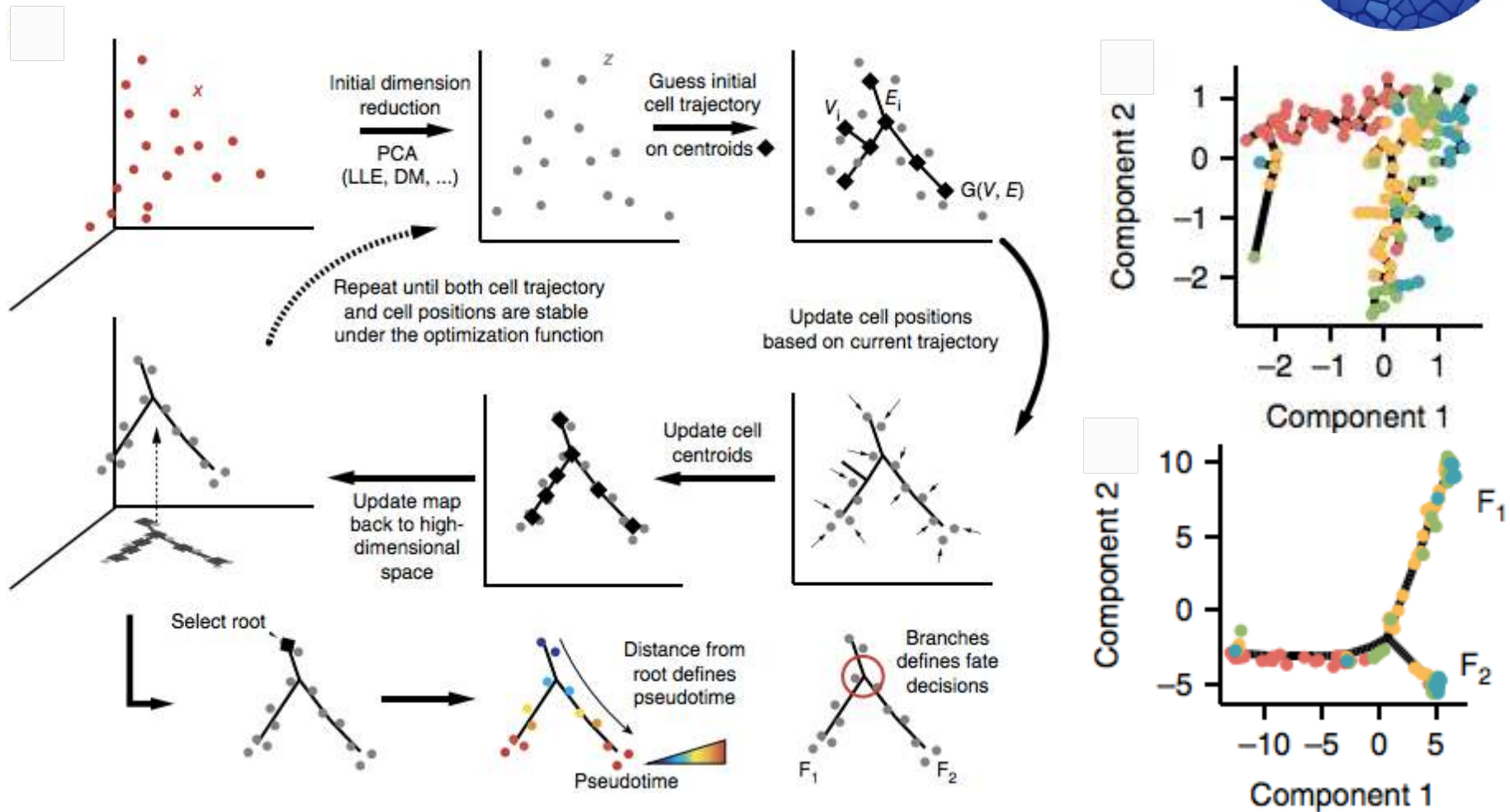
- Reference cell collections
 - Organism, tissue-specific data
- *De novo* vs. reference-based analysis
- Reference-augmented
 - Will capture both dataset-specific and reference-based signatures
 - Facilitates interpretation
 - Higher statistical power, stability



Trajectory Tracing: Monocle



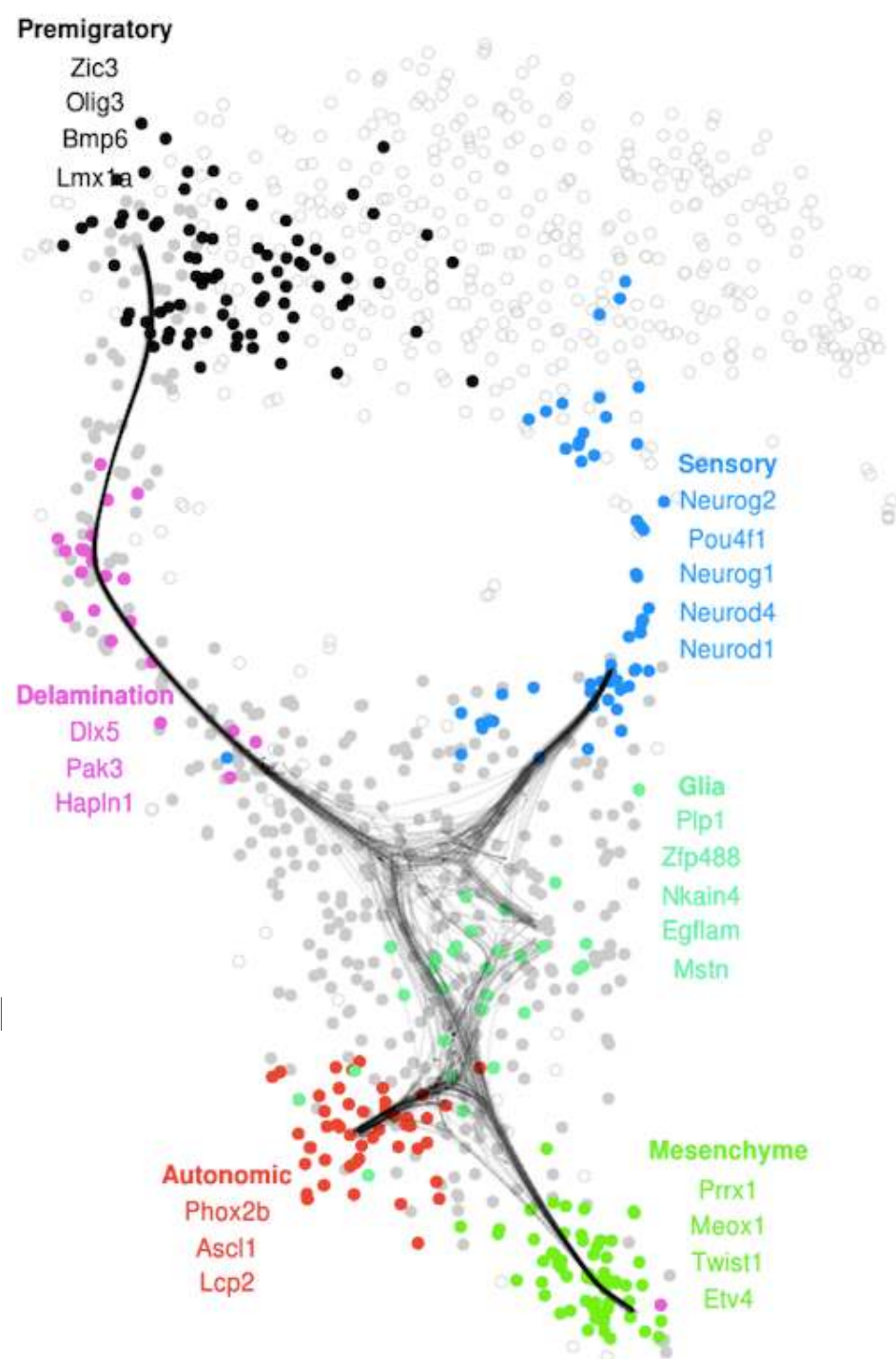
Trajectory Tracing: Monocle2



Qui *et al.*, Nat. Methods 2017

Trajectory Tracing

- Trying to reconstruct likely trajectory that the cells have taken
- Requires assumptions
 - Ergodicity
 - Continuity
 - Directionality
- Path uncertainty
 - How many decisions does a cell make?
 - Are they reversible?



Gene-gene correlations

