Comp790-166: Computational Biology

Lecture 6

February 4, 2021

Announcement

- Homework assigned today, available `https://github.com/stanleyn/Comp790-166-Comp-Bio/tree/main/Homework1`.
- Due to me by email on February 18, 2021 by 11:59pm
- You can ping me on slack or by email if you have questions.
- You can use the TeX template to write up your answers..... or not! Just submit PDF, please!

Today

- Unsupervised automated cell-population discovery with PhenoGraph
- Imputation for single-cell data (MAGIC)
- Visualization of single-cell data with PHATE

All of the Graph Knowledge Finally Applied....

We saw Spade last time. The purpose of PhenoGraph is also to define clusters of cells that recapitulates manual gating results

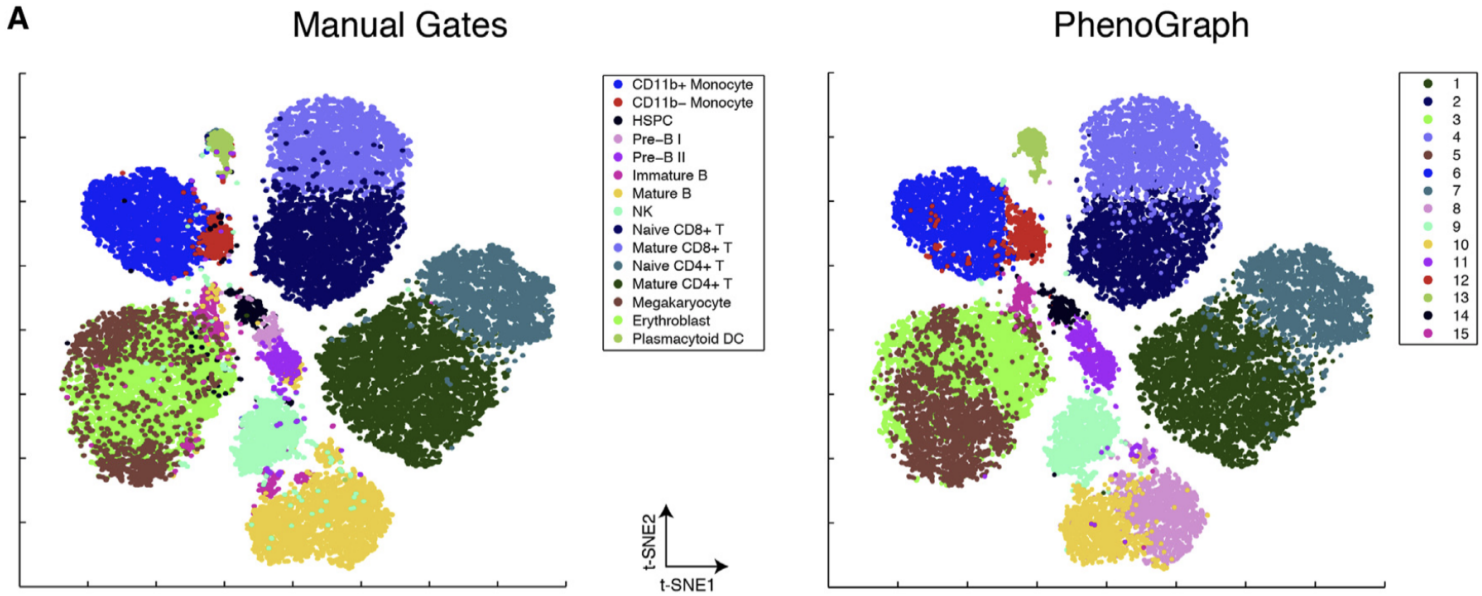


Figure: from Levine *et al.* Cell. 2015

A Direct Application of Louvain

- PhenoGraph is sold as being able to capture cell-populations of various sizes.
- A graph of cells is constructed for each samples
- Edges that exist are then turned into weighted edges, based on shared neighbors (a good contribution)
- The graph for each sample is partitioned with Louvain
- Clusters between samples are mapped with a metaclustering approach.

Communities Map to Cell-Populations

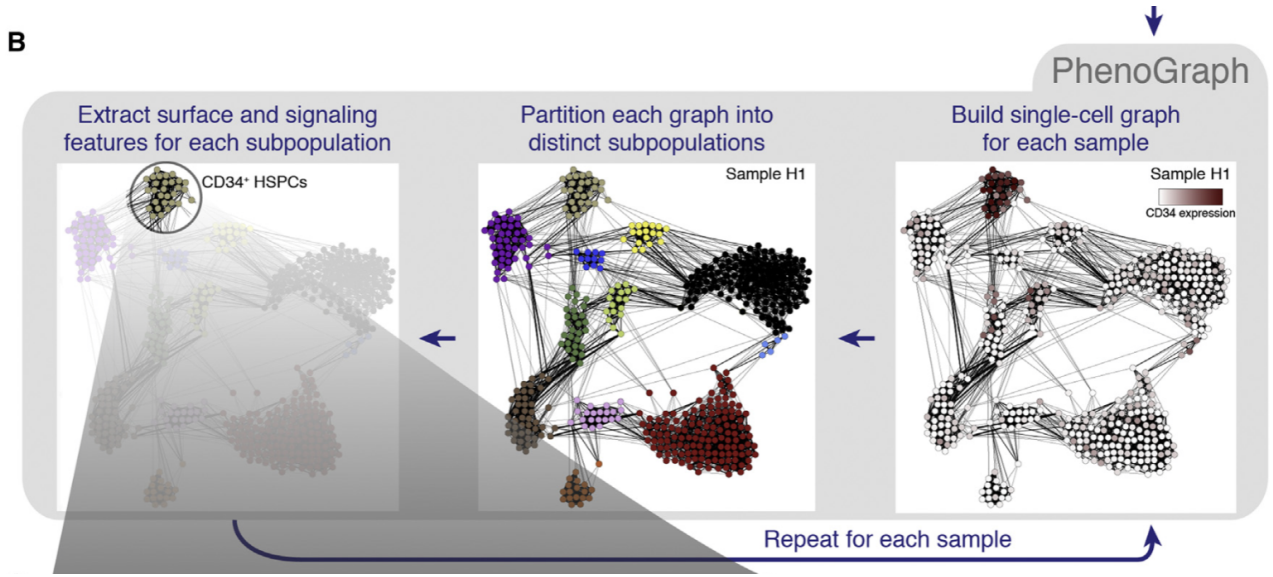


Figure: from Levine *et al.* Cell. 2015

Refining the kNN graph

Suppose an edge exists betweens nodes i and j. Then their edgeweight, w_{ij} is defined as the Jaccard Score of shared neighbors as,

$$W_{ij} = \frac{|v(i) \cap v(j)|}{|v(i) \cup v(j)|} \tag{1}$$

Here, $v(i)$ is the set of neighbors of node i. $|\cdot|$ is cardinality (number of neighbors)

Evaluating Similarity to Manual Gates

It's a bit reassuring to know that the performance (in terms of F-measure) is stable, regardless of the choice of k.

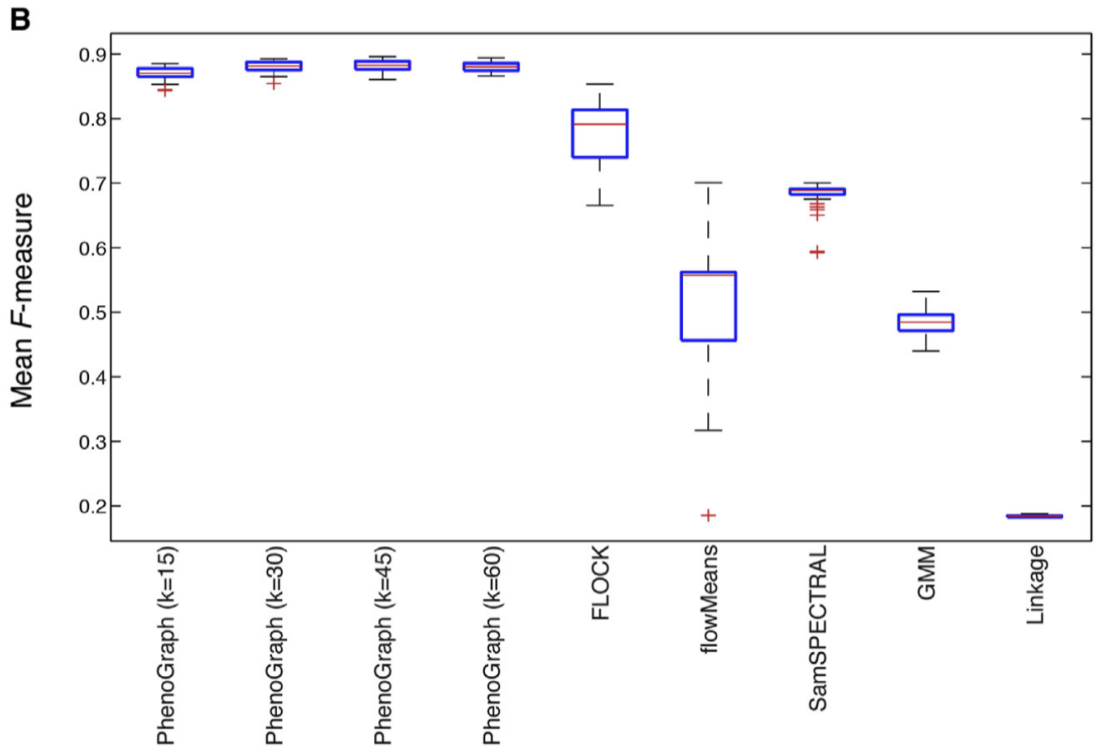


Figure: from Levine *et al.* Cell. 2015

Healthy vs AML with Cell Frequencies

Calculate the proportion of each sample's cells assigned to a particular metacluster. As you can see, there are differences in frequency between healthy (first few rows) and AML.

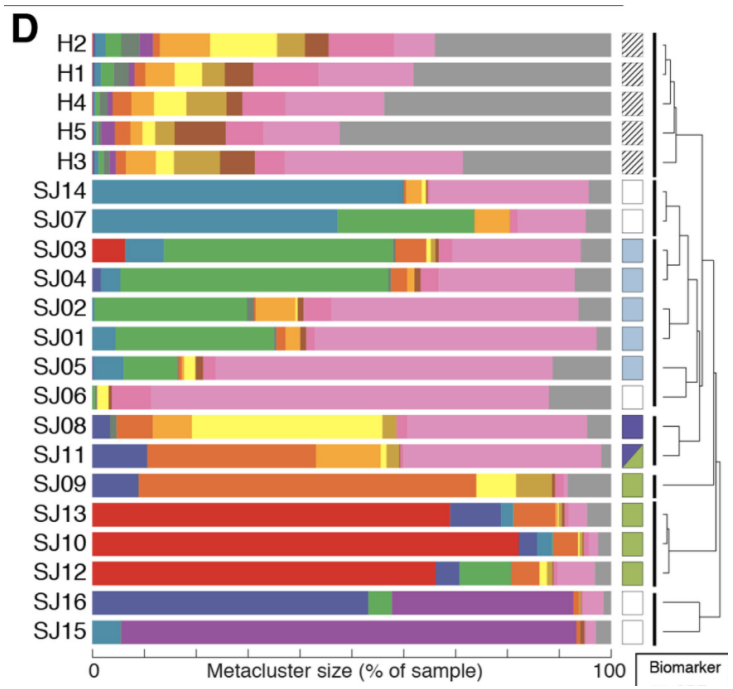


Figure: from Levine *et al.* Cell. 2015

You are ready for your first CyTOF dataset

If the data have signal, cell frequencies are a pretty powerful biologically interpretable feature that can allow us to classify patient samples.
To recap,

- Define per sample clusters
- Define metaclusters representing all samples
- Map cells from individual cells to metaclusters
- Compute frequency features
- You have now build a feature matrix that can be used for classification!

Example of a Powerful Frequency Feature

Frequency differences can be quite prominent in clinical settings. In this example, we are studying differences between patients who received steroid treatment after surgery, versus not.

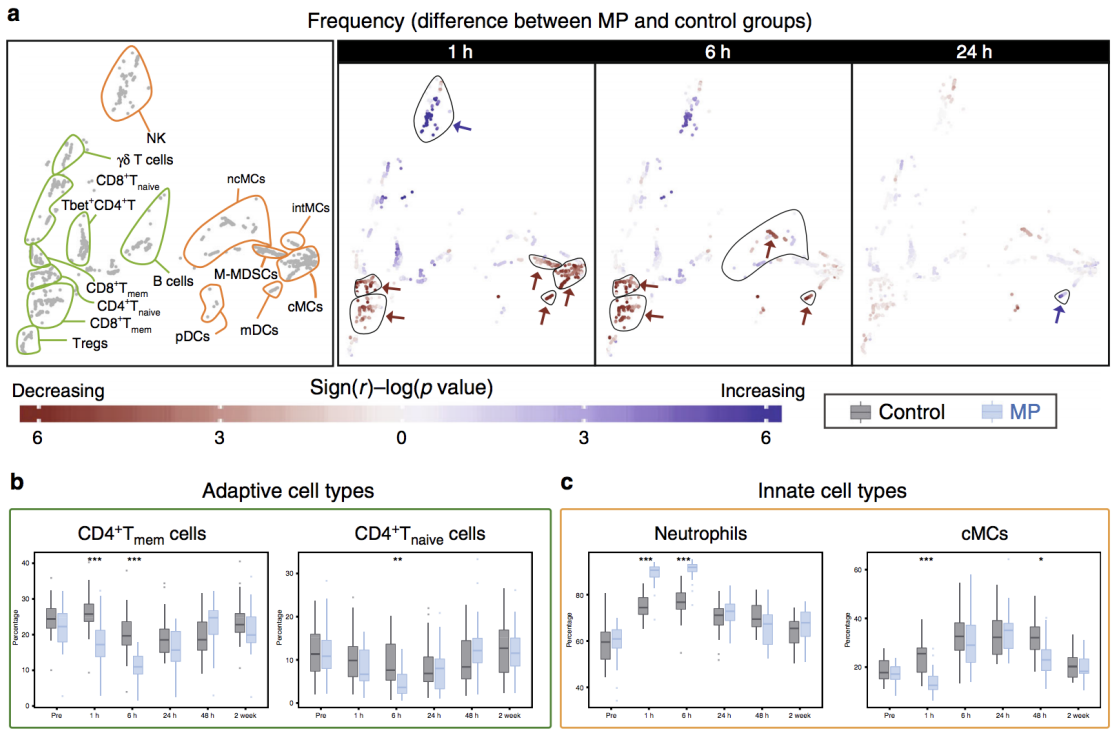


Figure: from Ganio *et al.* Nature Communications. 2019.

FlowSOM is another Contender

FlowSOM merges similar clusters through hierarchical clustering. So, clustering on the clusters....

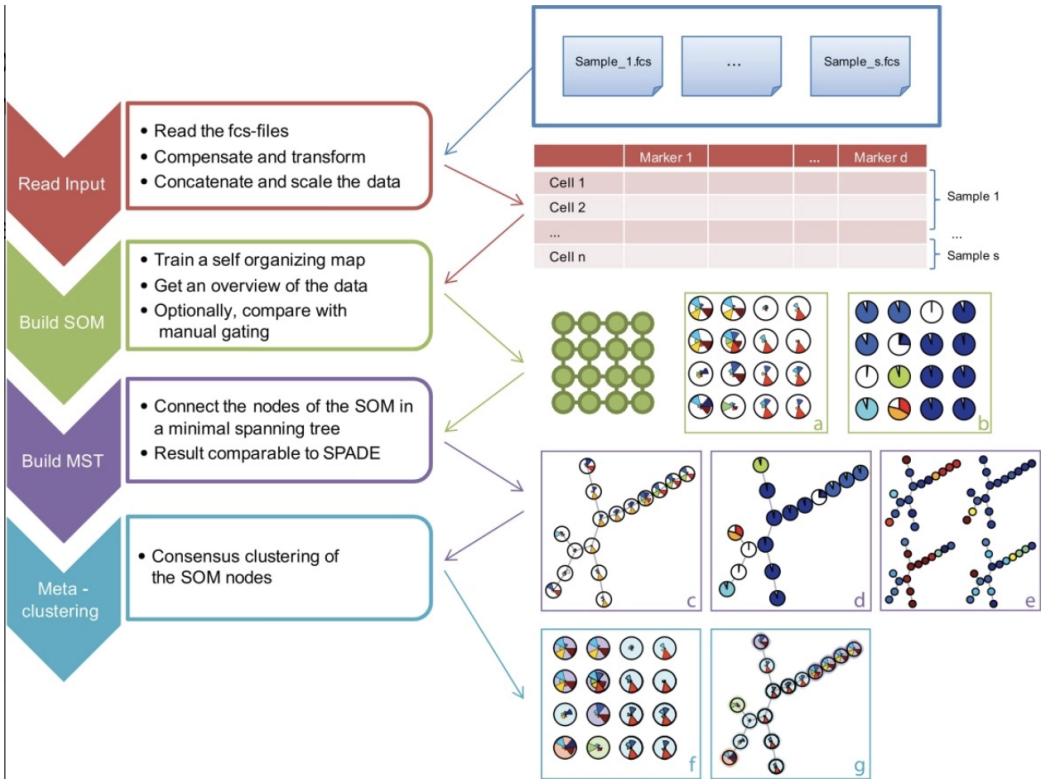


Figure: from Van Gassen *et al.* Cytometry A. 2015

Many Clustering Methods, Many Scores..

There are many possible scores (F1, NMI, Rand index, etc) that can be used to quantify the similarity between the true and predicted clustering assignments.

- If scoring clustering algorithms on single-cell data is interesting to you and you want to take into account all of the different ways this can be scored, I recommend the following brand new paper!
- Multi-Perspective Evaluation of Clustering Algorithms for Cytometry Data with Pareto Fronts `https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btab038/6122691?redirectedFrom=fulltext`

Missing Data

- We don't live in a perfect world where we always have all of the data (could be missing features, entire samples, etc.)
- Imputation refers to an approach to 'fill in the blanks' for the missing information, based on the data that is available.
- De-noising is also a form of imputation, where you estimate and remove noise.

Imputation for Single-Cell Data

- Especially in scRNAseq, the datasets can be quite sparse, meaning that lowly-expressed genes fail to be detected.
- We can imagine the same thing happening in mass cytometry, where not all of the signal is adequately detected.

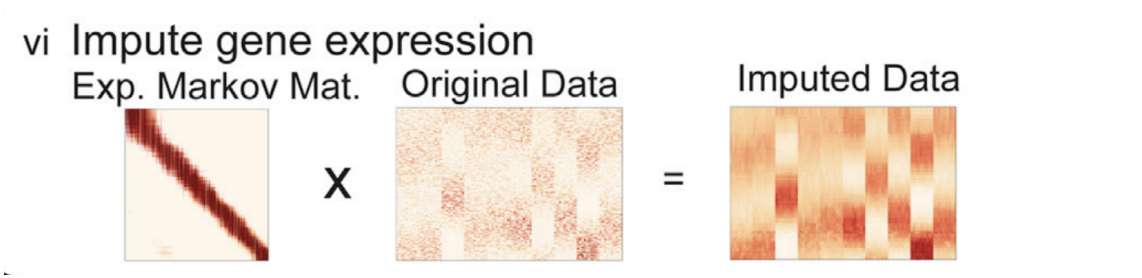


Figure: from van Dijk *et al.* Cell Systems. 2018

Accurately Representing Gene-Gene or Protein-Protein Interactions

The effect of lowly-expressed genes or proteins that are not adequately detected.

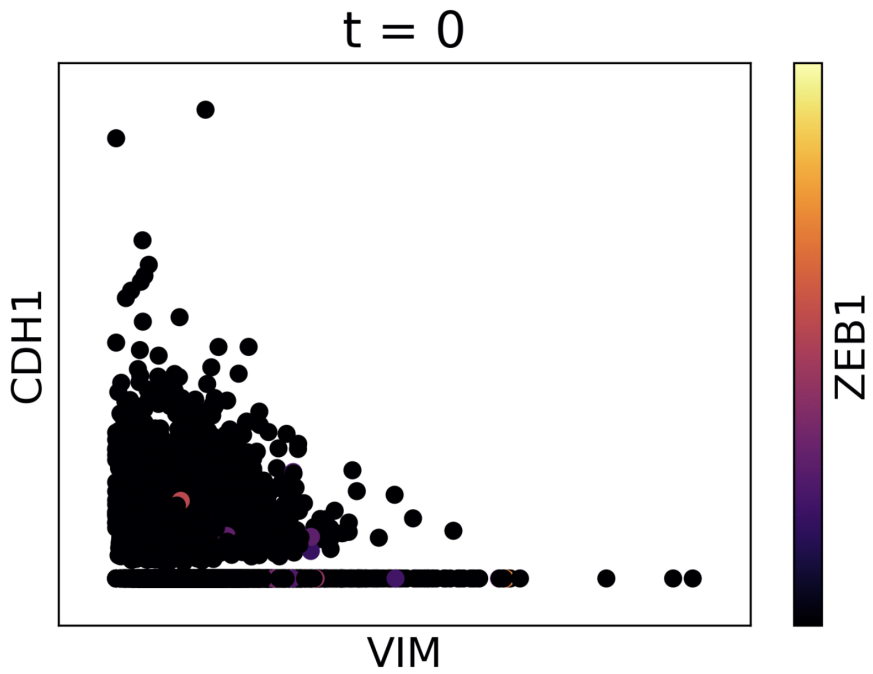
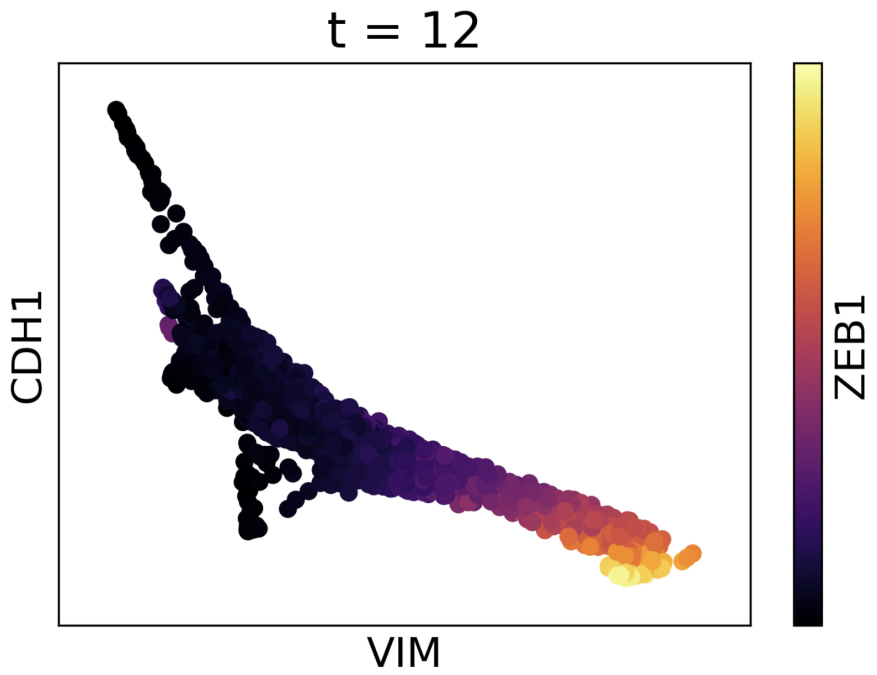


Figure: Note all of the 0 expression for CDH1

Estimating the Dropout, or Artificial Sparsity

If we can consider features of cells (in this case), that are otherwise similar to where dropout seems to have occurred, we can correct for the dropout and restore gene/gene or protein/protein interactions.



MAGIC Imputation for Single Cells

Here is an outline for the main idea.

- For a particular cell, find other cells that are most similar
- Construct a weighted affinity matrix between cells to more accurately capture between-cell similarities within a neighborhood
- Use affinities, and cell-cell neighborhood structure to correct the data

Step 1: Compute a Markov Affinity Matrix

Given the cell $\times$ feature matrix, $\mathbf{X}$, compute the cell-to-cell affinity matrix, $\mathbf{A}$ as follows,

$$A_{ij} = \exp(-(\text{Dist}(i,j)/\sigma)^2) \tag{2}$$

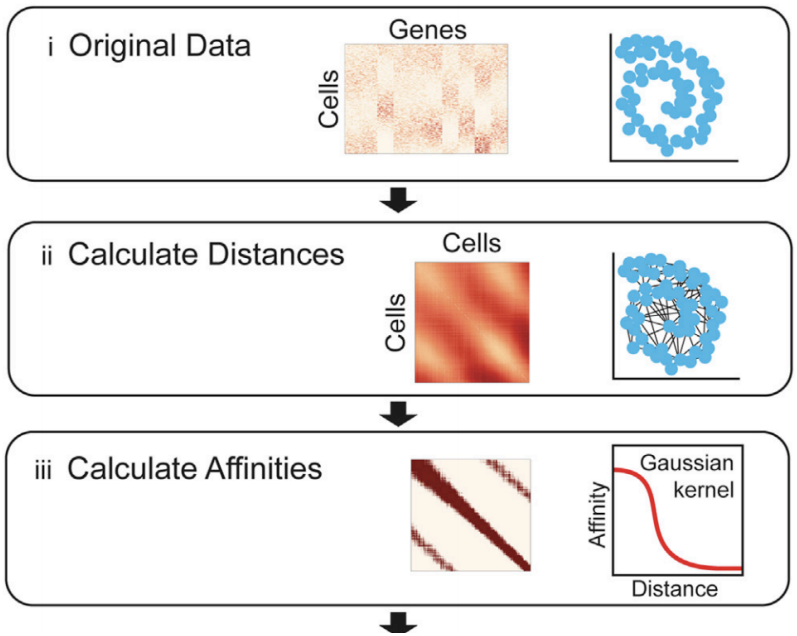


Figure: from van Dijk *et al.* Cell Systems. 2018

Notes on the σ parameter

In $A_{ij} = \exp(-(\text{Dist}(i,j)/\sigma)^2)$

- If σ is too small, the graph can become too disconnected.
- If σ is too large, cell-type resolution will be lost in the data and cell-types will be merged together.
- The solution is to adapt σ for each cell (σ_i) to take into account the density of its neighborhood.
- In particular, σ_i is the distance between i and its k*a*th nearest neighbor.

Affinity Matrix Cleaning

- The affinity matrix is not symmetric!
- This can quickly be made symmetric with,

$$\mathbf{A} = \mathbf{A} + \mathbf{A}^T \tag{3}$$

- The next step is in row-stochastic normalization that renders the affinity matrix into a Markov transition matrix, $\mathbf{M}$

$$M_{ij} = \frac{A_{ij}}{\sum_k A_{ik}} \tag{4}$$

Markov Affinity Based Graph Diffusion : Exponentiating **M**

- By powering the matrix ($\mathbf{M}^t$), you can refine between-cell similarities related to the neighborhood structure. In other words, increase weight between pairs that have many common neighbors.
- M_{ij}^t represents the probability that a random walk of length t will reach node j after starting from node i.
- As t increases, false similarity between cells that was based strongly on noise gets filtered out. So, spurious neighbors will be down-weighted.

Visualizing what is happening....

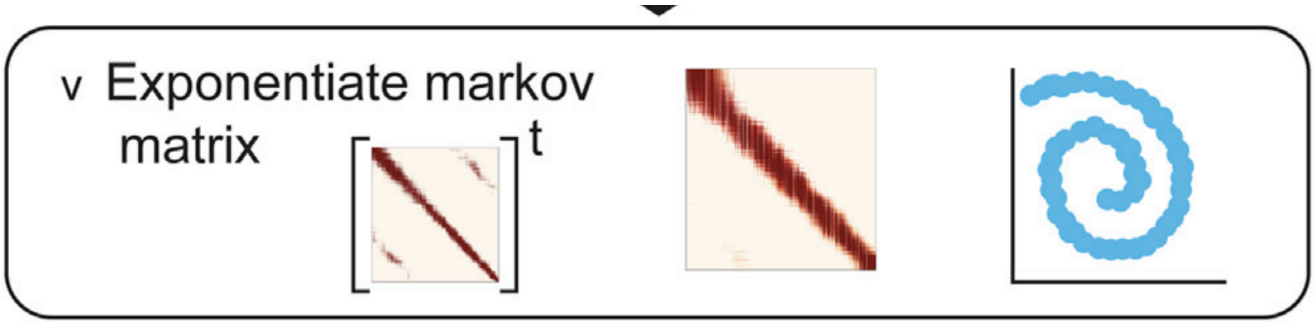


Figure: from van Dijk *et al.* Cell Systems. 2018

Imputation After Graph Diffusion

Finally, the original data $\mathbf{X}$ (of cells $\times$ features) is transformed as,

$$\mathbf{X}_{\text{impute}} = \mathbf{M}^T\mathbf{X} \tag{5}$$

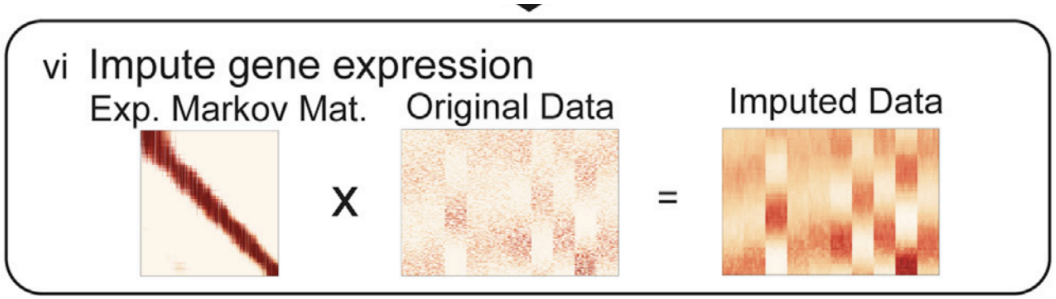


Figure: from van Dijk *et al.* Cell Systems. 2018

On a Practical Note, t

The t, or how you power the matrix is important to the results.

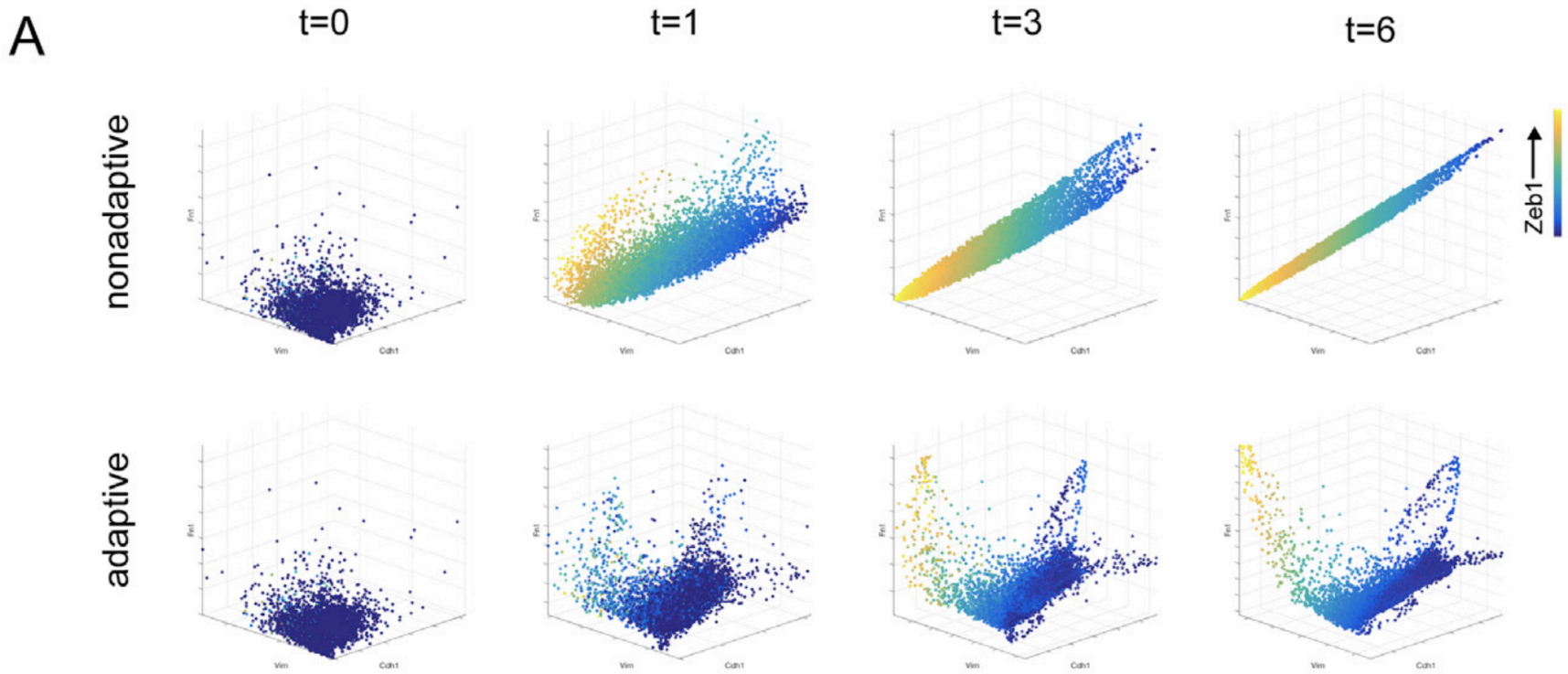


Figure: from van Dijk *et al.* Cell Systems. 2018. Adaptive means that you chose a σ_i for a particular cell, i.

Convergence of $\mathbf{X}_{\text{impute}}$ for different t

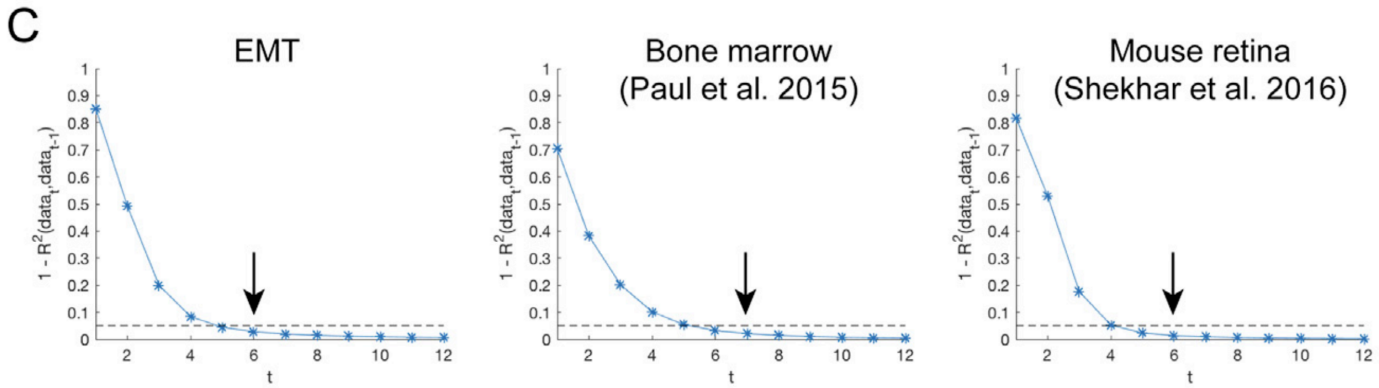


Figure: from van Dijk *et al.* Convergence of $\mathbf{X}_{\text{impute}}$ between t and $t-1$.

Effectiveness in Noise Reduction

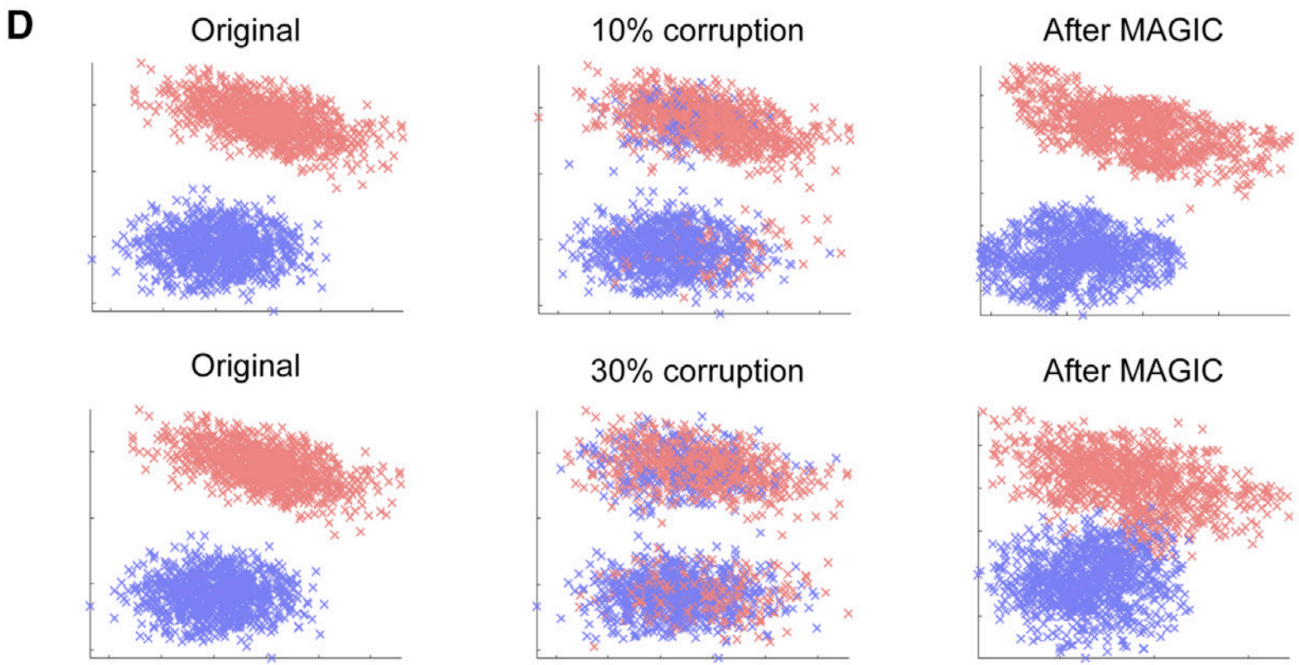


Figure: from van Dijk *et al.* MAGIC can restore noisy data, when features are artificially corrupted in a synthetic dataset.

Visualizing Single-Cell Data with PHATE

A Dimensionality Reduction Method for Single Cells

PHATE aims to preserve between-cell population similarities according to the way that cells differentiate.

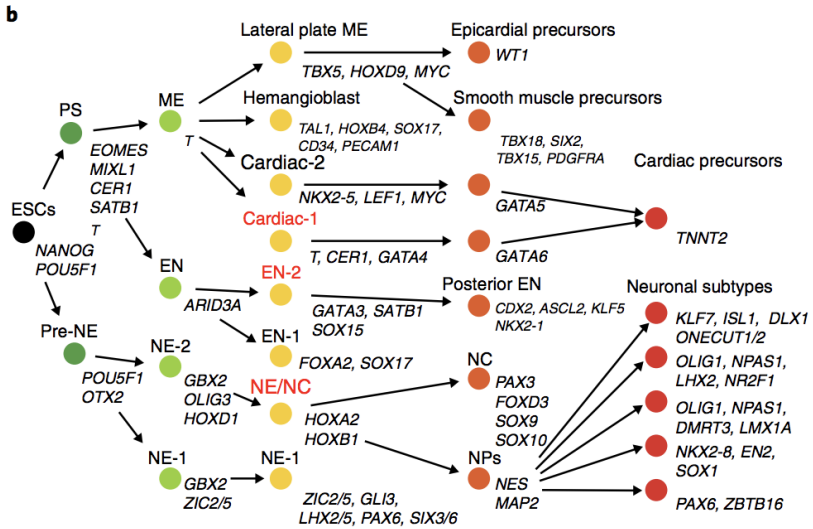


Figure: from Moon *et al.* Nature Biotechnology. 2019

PHATE Overview

Table 1 | General steps in the PHATE algorithm
Input: Data matrix, algorithm parameters (Methods)
Output: The PHATE visualization
(1) Compute the pairwise distances from the data matrix.
(2) Transform the distances to affinities to encode local information.
(3) Learn global relationships via the diffusion process.
(4) Encode the learned relationships using the potential distance.
(5) Embed the potential distance information into low dimensions for visualization.

Figure: from Moon *et al.* Nature Biotechnology. 2019

The First Few Steps are Identical to MAGIC

- **Step 1:** Cell $\times$ Cell distance matrix (based on Euclidean distance between cells)
- **Step 2:** Convert to Cell $\times$ Cell row-stochastic affinity matrix $\rightarrow$, $\mathbf{P}$, such that $\sum_j \mathbf{P}_{ij} = 1$.
- **Step 3:** Power $\mathbf{P} \rightarrow \mathbf{P}^t$.

A Complementary Method to Determine Optimal t

- The choice of t will determine how much noise you want to filter. Small eigenvalues of $\mathbf{P}$ correspond to noise, and will ultimately decrease towards 0.
- Let $[\eta(t)]_i = \lambda_i^t / \sum_{j=0}^{N-1} \lambda_j^t$ be the probability distribution defined by the non-negative eigenvalues of $\mathbf{P}^t$.

The von Neumann entropy $H(t)$ is then computed based on $[\eta(t)]_i$ and decreases as $t \to \infty$.

$$H(t) = -\sum_{i=1}^{N} [\eta(t)]_i \log([\eta(t)]_i) \tag{6}$$

Relationship Between t and $H(t)$

Higher t means less noise, but the implications of t are super powerful!

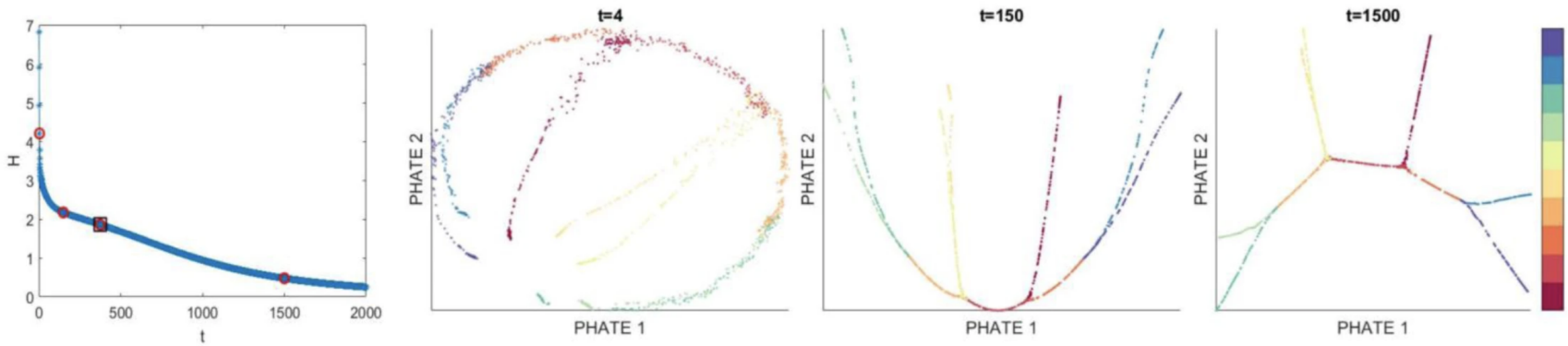


Figure: from Moon *et al.* Nature Biotechnology. 2019

Computing Potential Distances

- Potential distance is computed as a divergence between the associated diffusion probability distributions of the two cells to all other cells

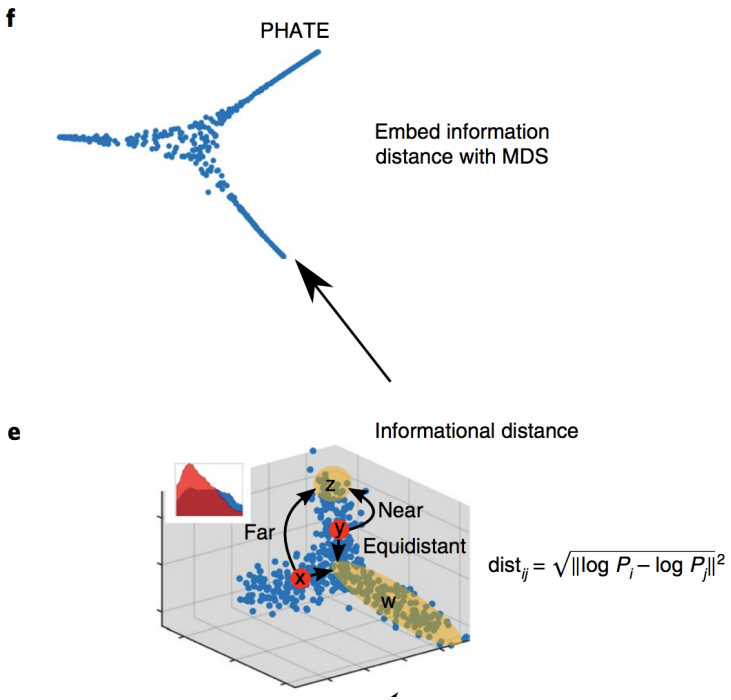


Figure: from Moon *et al.* Nature Biotechnology. 2019

Potential Distances, Written Formally

- $U_{\mathbf{x}}^t = -\log(p_{\mathbf{x}}^t)$
- Then $\text{PD}(\mathbf{x}, \mathbf{y}) = ||U_{\mathbf{x}}^t - U_{\mathbf{y}}^t||_2$

PHATE preserves branch structure

As expected tSNE is preserving local similarities, but not necessarily distances between groups of points.

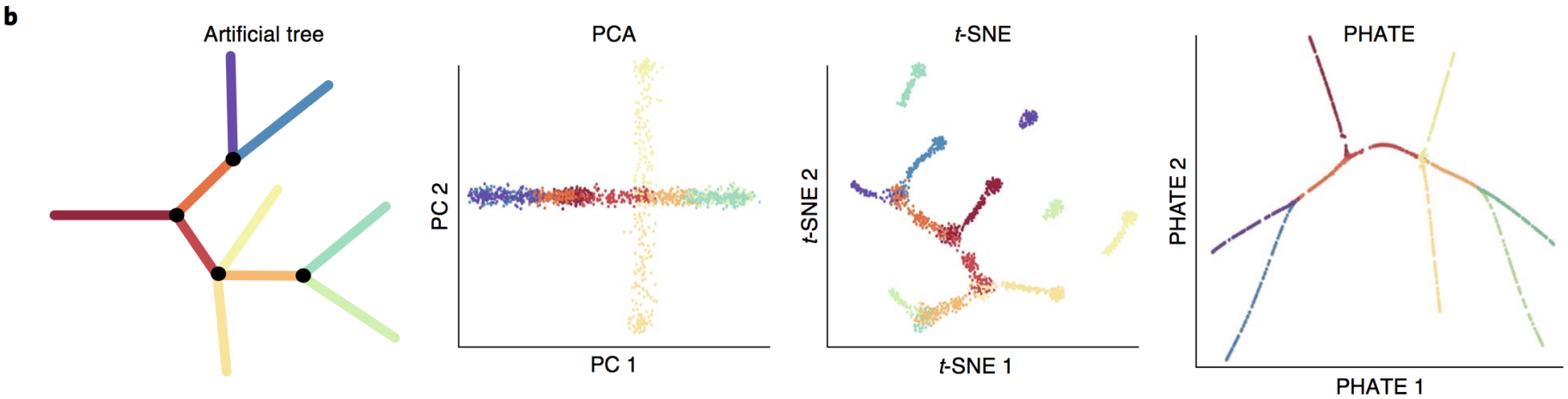


Figure: from Moon *et al.* Nature Biotechnology. 2019

And Preserves Cellular Differentiation Patterns!

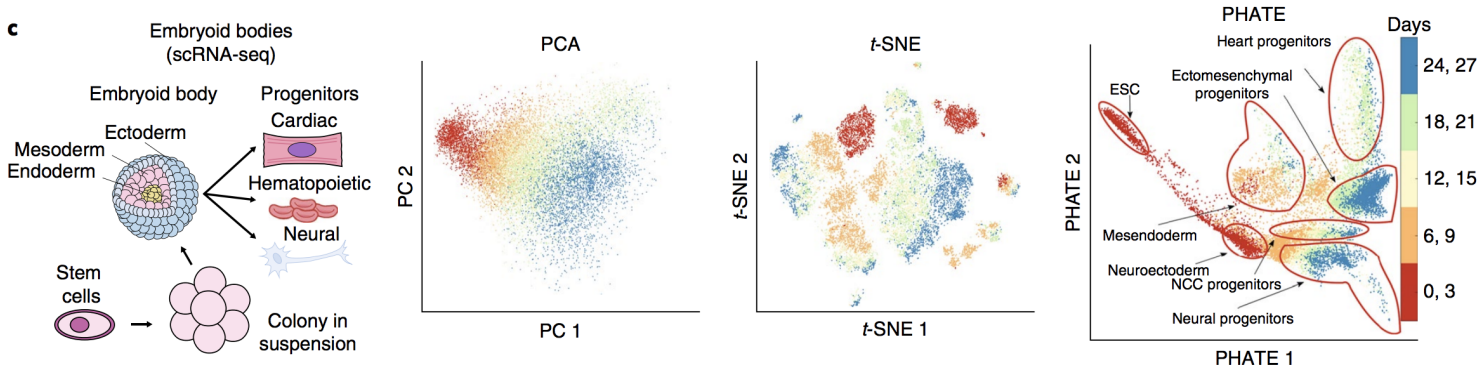


Figure: from Moon *et al.* Nature Biotechnology. 2019

Next Time...

- Synthetic Data Generation Informed by Geometry (Useful for validating algorithms)
- Linking single-cell data to external variables