Comp790-166: Computational Biology

Lecture 12

March 4, 2021

Announcements

- Please sign up for your project proposal presentation slot on the google doc! `https://docs.google.com/spreadsheets/d/1puFjmjadZuuVIy6nyJJPHJbucrW4-KH-xFd2P3HjIEI/edit?usp=sharing`
- Project proposals are due March 9, so in 1 week!

Today

- Finish up batch effect correction and multiple single-cell dataset integration (quickly)
- CyTOF Merge
- Start trajectory inference!
  - Diffusion Maps for analysis of differentiation
  - SLICER (UNC's own creation)

Another Approach : Harmony

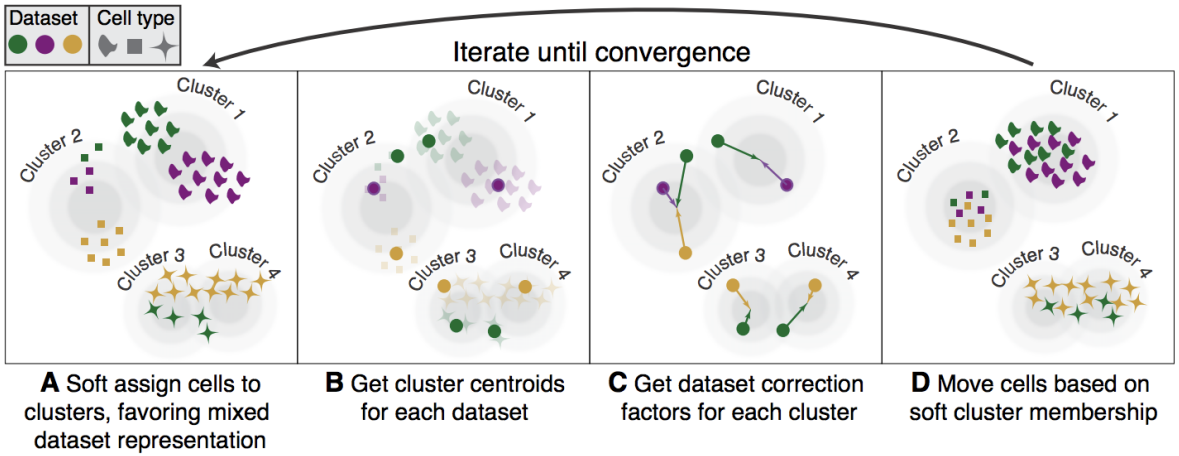


Figure: from `https://www.biorxiv.org/content/10.1101/461954v2.full.pdf`. Harmony uses a soft clustering approach and specifically favors clusters with representative members across all datasets. It also performs a dataset-specific correction per cluster.

Recap

- We just saw a couple of ways that we can combine multiple datasets
- These multiple datasets can correspond to **batches, technologies, or different tissue samples.**
- Conos tries to define a graph such that its structure (in this case, communities) contain cells that are mixed across datasets
- When integrating multiple datasets, we must be careful that cells are grouping together because of phenotype, not because of the function that might have been perturbed under a particular condition.

Another Kind of Dataset Integration Problem, Multiple Panels

Suppose you are trying to analyze a dataset with somewhat overlapping, but unique panels. This could happen for example if different sets of samples were profiled in different labs.

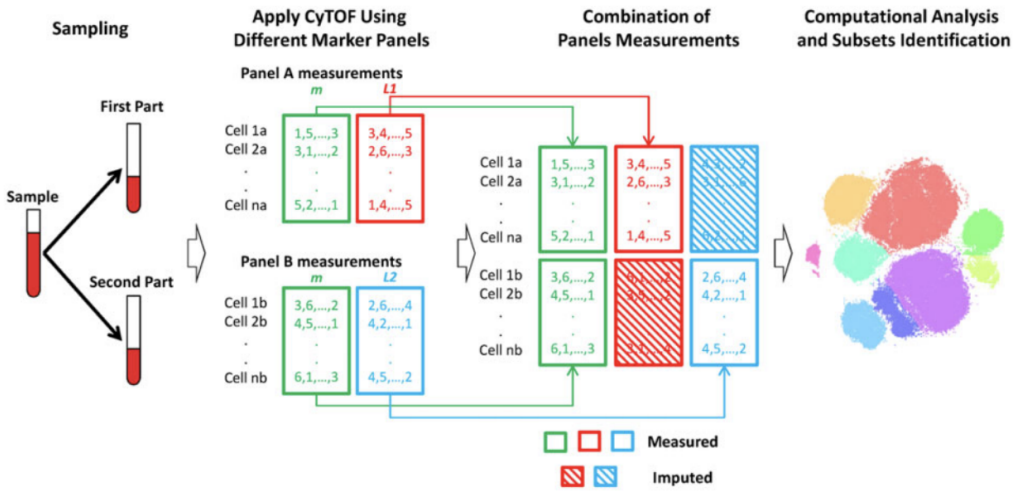


Figure: from Abdelaal *et al.* Bioinformatics. 2019.

Problem Formulation

- Imagine having two panels, where these panels share m markers.
- Assume that panel 1 has $L1$ unique markers and panel 2 has $L2$ unique markers.
- One goal is to determine an informative set of markers that overlap between the two datasets
- Another goal is to impute missing markers.

Imputation

- Starting in panel A, find the k most similar cells in panel B, using the set of m shared markers
- For each cell in panel A, impute missing values for the markers by finding the k most similar cells from panel B and computing missing features as the median of the corresponding features in these k cells.

Computing an Importance Score for Overlapping Markers

Among features overlapping between panels, an importance score for each feature can be calculated to help minimize redundancy across feature measured in panels etc. The importance score for marker p is computed using the first m PCs as,

$$i_p = \sum_{q=1}^{m} \beta_{pq}^2 \times \lambda_q \tag{1}$$

- β_{pq} is the loading of marker p to the qth PC
- λ_q is the variance explained by the qth PC.

You could use the top scored markers either in the design of a new panel or to do your imputation.

Evidence that Imputation is Working for Certain Cell-Populations

	Imputed data
CD4+T cells	0.78
CD8+T cells	0.79
B cells	0.83
CD3–CD7+cells	0.78
TCR$\gamma\delta$ cells	$0.77 \pm 8e{-5}$
Myeloid cells	$0.82 \pm 7e{-5}$
All cells	0.81

Figure: from Abdelaal *et al.* Bioinformatics. 2019.

Entering Pseudotime and Trajectory Inference

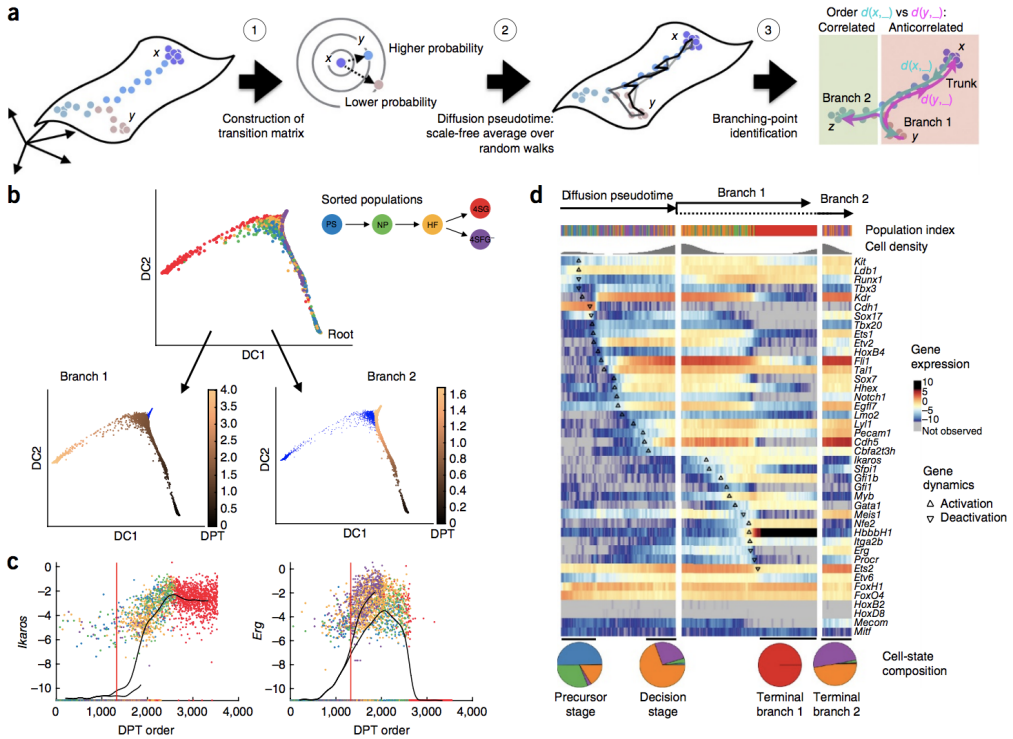


Figure: from Haghverdi *et al.* Nature Methods. 2016. We would like to establish a chronological ordering of cells that represents their differentiation. This will allow us to uncover gene programs associated with differentiation.

Welcome Back Diffusion Maps

Let's set up some notation as follows :

- Let Ω be the set of all measured cells
- Suppose that we have measured n total cells, each with G features, so a given $\mathbf{x}_i \in \mathbb{R}^G$.

Allow each cell to diffuse around its measured position as,

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

Here the σ^2 determines the scale over which each cell can randomly diffuse.

Modeling the Probability cell x transitioning to cell y

Define the transition matrix, **P** for all pairs of cells as,

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

Here $Z(\mathbf{x})$ is a partition function providing an estimate for the number of neighbors of **x** in a certain volume defined by σ, or the density of cells at that proximity.

Redefine a Density Normalized Transition Probability Matrix, $\tilde{\mathbf{P}}$

We are interested in the transition probabilities between cells and not the on-cell potentials imposed by local density. Therefore, the diagonal of $\tilde{\mathbf{P}}$ will be set to 0 and $\mathbf{y} = \mathbf{x}$ will be excluded from the sum in the partition function, $\tilde{Z}(\mathbf{x})$. Without some correction, it may look like cells within a very dense region have a small diffusion distance.

$$\tilde{P}_{\mathrm{xy}} = \frac{1}{\tilde{Z}(\mathbf{x})} \frac{\exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)}{Z(\mathbf{x})Z(\mathbf{y})} \quad , \tilde{P}_{\mathrm{xx}} = 0$$

$$\tilde{Z}(\mathbf{x}) = \sum_{\mathbf{y}\in\Omega/\mathbf{x}} \frac{\exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)}{Z(\mathbf{x})Z(\mathbf{y})}$$

Using Eigenvectors of $\tilde{\mathbf{P}}$

- $\tilde{\mathbf{P}}$ has n ordered eigenvalues, $\lambda_0 = 1 > \lambda_1 \geq \cdots \geq \lambda_{n-1}$
- Corresponding to these eigenvalues are eigenvectors $\boldsymbol{\Psi}_0 \ldots \boldsymbol{\Psi}_{n-1}$
- As we saw a few weeks ago, powering $\tilde{\mathbf{P}}$ to $\tilde{\mathbf{P}}^t$ represents the probability of transitioning between two cells with a walk of length t.

The diffusion distance between a pair of cells $\mathbf{x}$ and $\mathbf{y}$ can be written in terms of the eigenvectors of $\tilde{\mathbf{P}}$ as,

$$D_t^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n-1} \lambda_i^{2t} \left(\psi_i(\boldsymbol{x}) - \psi_i(\boldsymbol{y})\right)^2$$

Unpacking Diffusion Distance

$$D_t^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n-1} \lambda_i^{2t} \left(\psi_i(\boldsymbol{x}) - \psi_i(\boldsymbol{y})\right)^2$$

- The eigenvector to the largest eigenvalue, λ_0 is a constant vector, $\boldsymbol{\Psi}_0 = \mathbf{1}$. Therefore, it contributes 0.
- The eigenvalues of $\tilde{\mathbf{P}}$ determine the diffusion coefficients in the direction of the corresponding eigenvector
- After the first l prominent directions, the diffusion coefficients typically drop to a noise level.
- When you find the l such that there is a large difference between l and $l+1$ eigenvalues (an elbow), you can use the sum up to the l-th term as an approximation for diffusion distance. The first l eigenvectors correspond to the diffusion components.

What we have just defined, illustrated

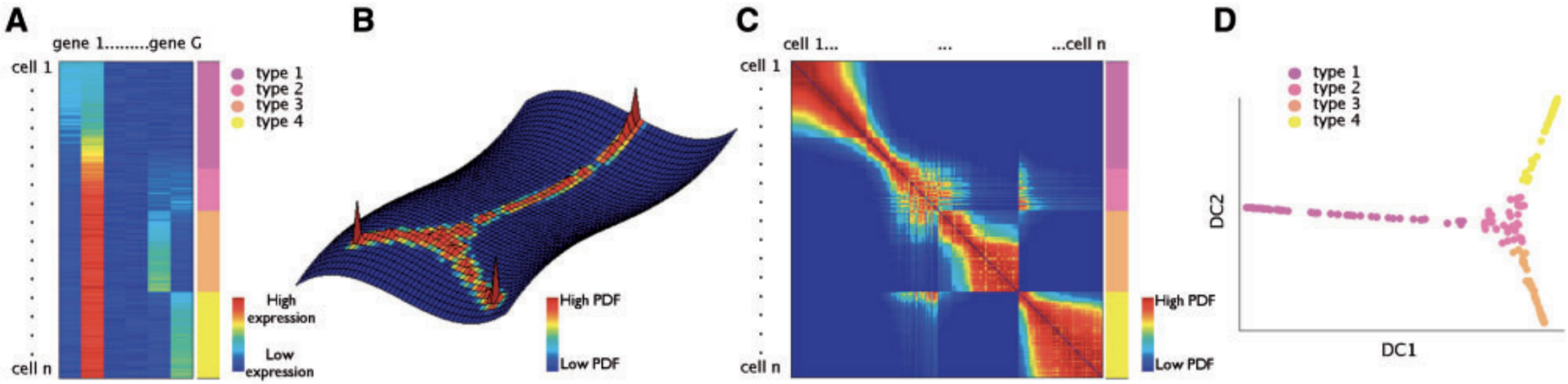


Figure: from Haghverdi *et al.* Bioinformatics. 2015. After projecting cells based on the first two diffusion components, it is still unclear what the time direction is.

Example Applied to A Dataset of Differentiating Cells

The diffusion map is capturing transitions between cell types that are not reflected in PCA or tSNE.

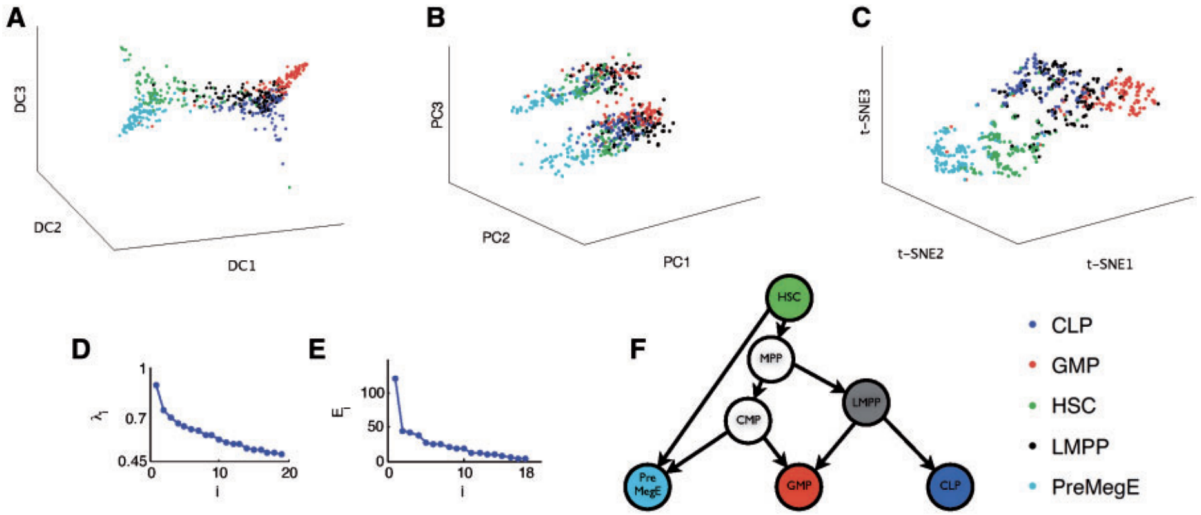


Figure: from Haghverdi *et al.* Bioinformatics. 2015.

This diffusion map approach was the beginning of thousands of people starting to think about cellular differentiation.....

Welcome SLICER

SLICER builds on and expands the very early Diffusion based techniques through the following

- Automatically select genes to use for building the trajectory (or in establishing the ordering between cells)
- Use locally linear embedding to capture non-linear relationships between gene expression levels and progression through a process
- Define ‘geodesic entropy‘ and use it to define branches
- Capture unique trajectory patterns such as bubbles.

SLICER Overview

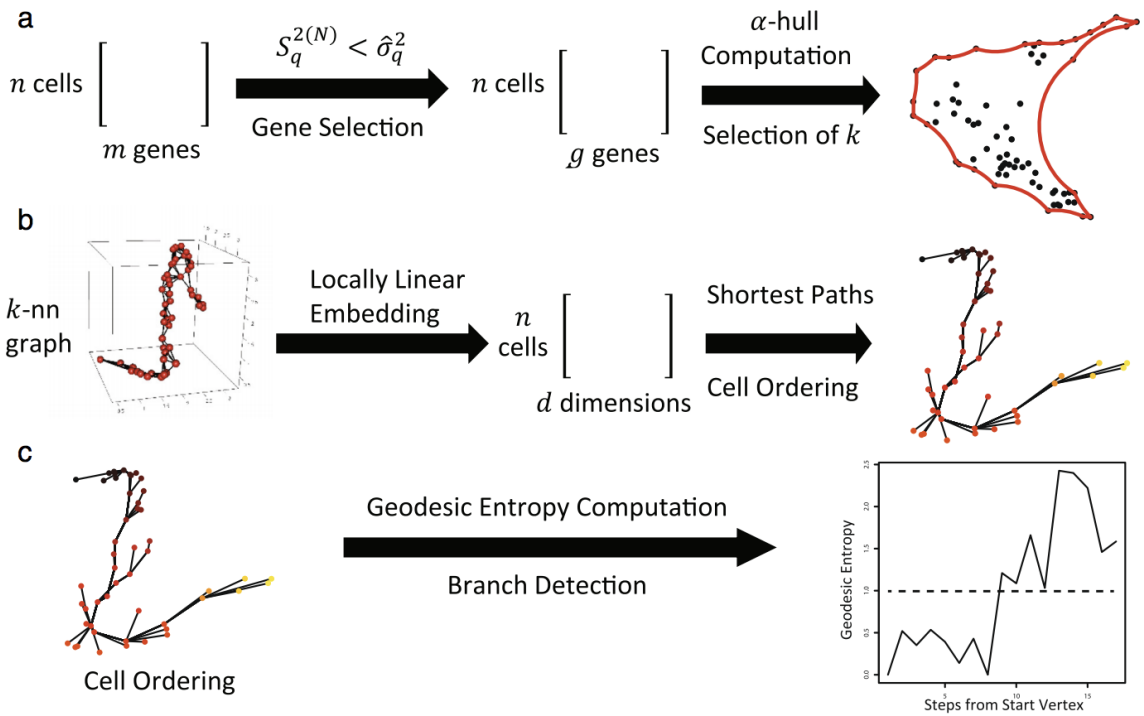


Figure: from Welch *et al.* Genome Biology. 2016

Step 1: Selecting Features to Use (Intuition)

Establishing some intuition about what makes a good ‘trajectory feature‘

- If a feature is involved in progression along a trajectory, expect gradual change in that feature along the trajectory
- A feature not involved should not fluctuate along the trajectory.
- In real life, we have no idea what is happening with this trajectory. Use similarity within neighborhoods to study ‘segments‘ of a trajectory.

Neighborhood Variance

Interesting features are those whose variance is greater than some level of neighborhood variance. Specifically, for the gth feature, we can compute its variance (σ_g^2) across samples and compare it (making sure it is at least as large as) to this defined neighborhood variance.
The neighborhood variance is defined as,

$$S_g^{2(N)} = \frac{1}{nk_c - 1} \sum_{i=1}^{n} \sum_{j=1}^{k_c} \left(e_{ig} - e_{N(i,j)g} \right)^2$$

- k_c is the number of nearest neighbors needed for each node for the graph to be connected.
- Each e_{ig} is representing the value of feature g in sample i.
- $e_{N(i,j)g}$ is representing the jth nearest neighbor in sample i.

Local Linear Embedding ($d = 2$)

Step 1: Find the weights (w_{ij}s) that can best reconstruct the original data (e.g. cell $\times$ feature),

$$W = \text{argmin}_W \sum_{i=1}^{n} \left| E_i - \sum_{j=1}^{k} w_{ij} E_j \right|_2^2$$

Step 2: Find optimal d-dimensional embedding, so in this case, L

$$L = \text{argmin}_L \sum_{i=1}^{n} \left| L_i - \sum_{j=1}^{k} w_{ij} L_j \right|_2^2$$

k-NN graph and shortest path

- Compute k-nearest neighbor graph between cells in terms of the LLE-determined coordinates.
- Specify a starting point (like a stem cell), and use a shortest path algorithm like Dijkstra to find the shortest path to some cell of interest.

Detecting Branches with Geodesic Entropy Measure

- Let $t_i = \{s = v_1, \ldots, v_k, \ldots, v_l = i\}$ be the shortest path from the starting point s to some cell, i.
- Denote the kth node on the shortest path from s to i by $t_i(k)$.
- Define f_{jk} as the number of paths passing through point j at distance k, $f_{jk} = \sum_i^n I[t_i(k) = j]$
- Then compute the fraction of all paths in S that pass though node j at distance k, $p_{jk} = \frac{f_{jk}}{\sum_{i=1}^n f_{ik}}$
- $H_k = -\sum_{i=1}^n p_{ik} \log_2 p_{ik} \rightarrow$ look at high entropy

SLICER Applied to Synthetic Data

Studying geodesic entropy over k. Higher entropy in terms of steps corresponds to the 'bubbles' in the data.

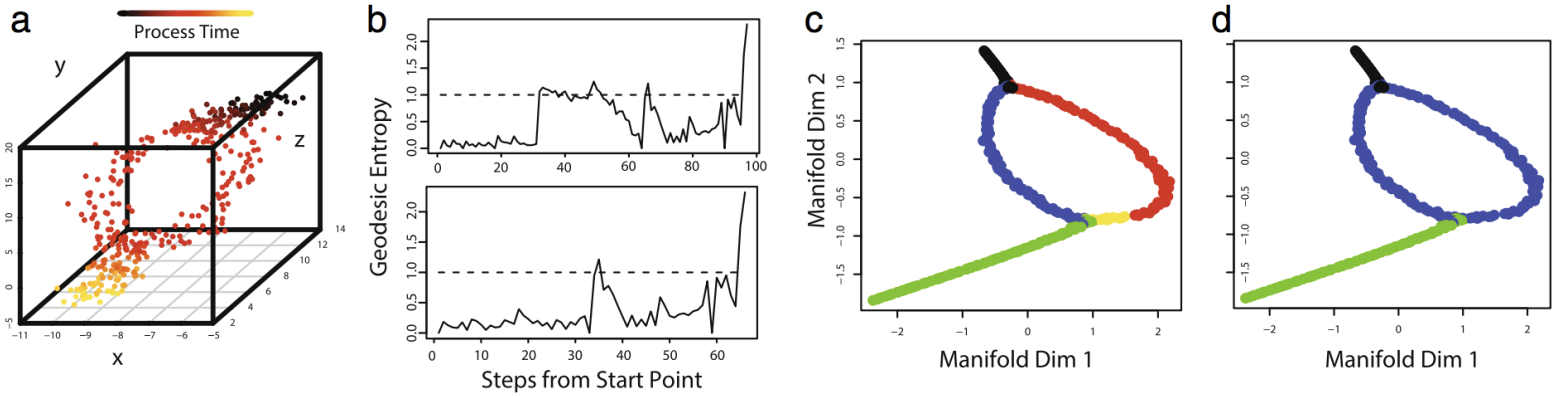


Figure: from Welch *et al.* Genome Biology. 2016.

Time Series Data

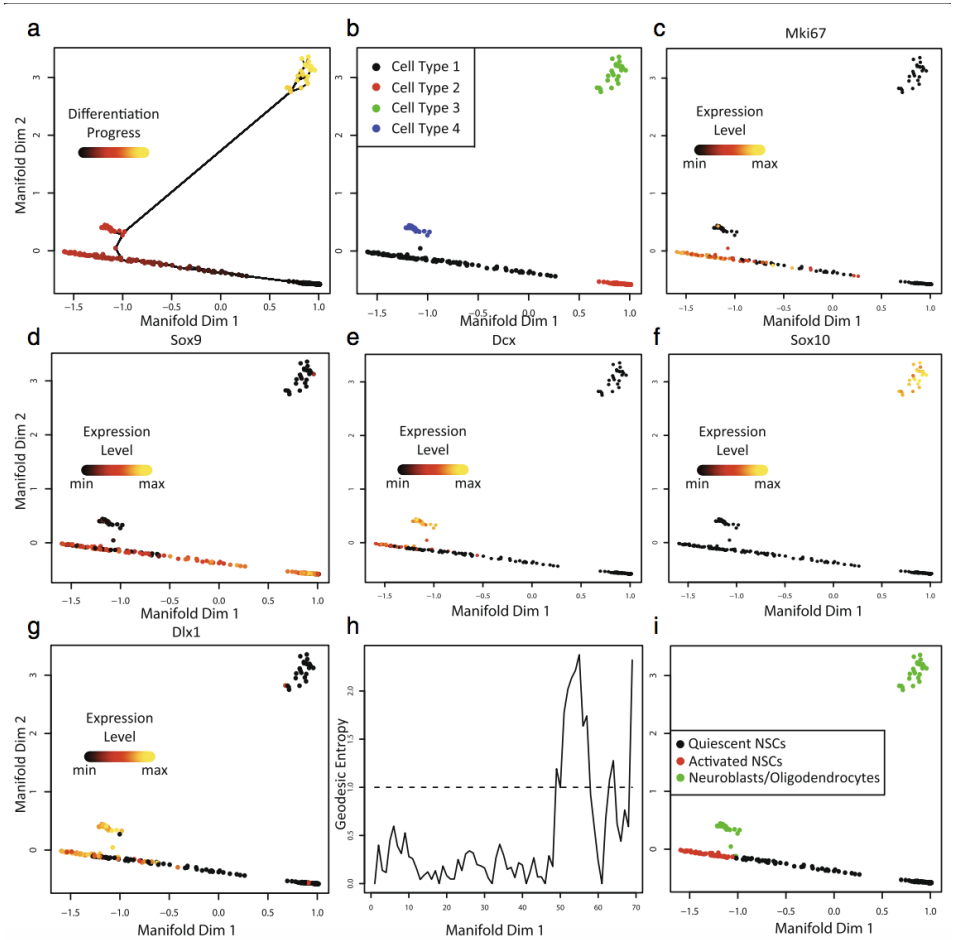
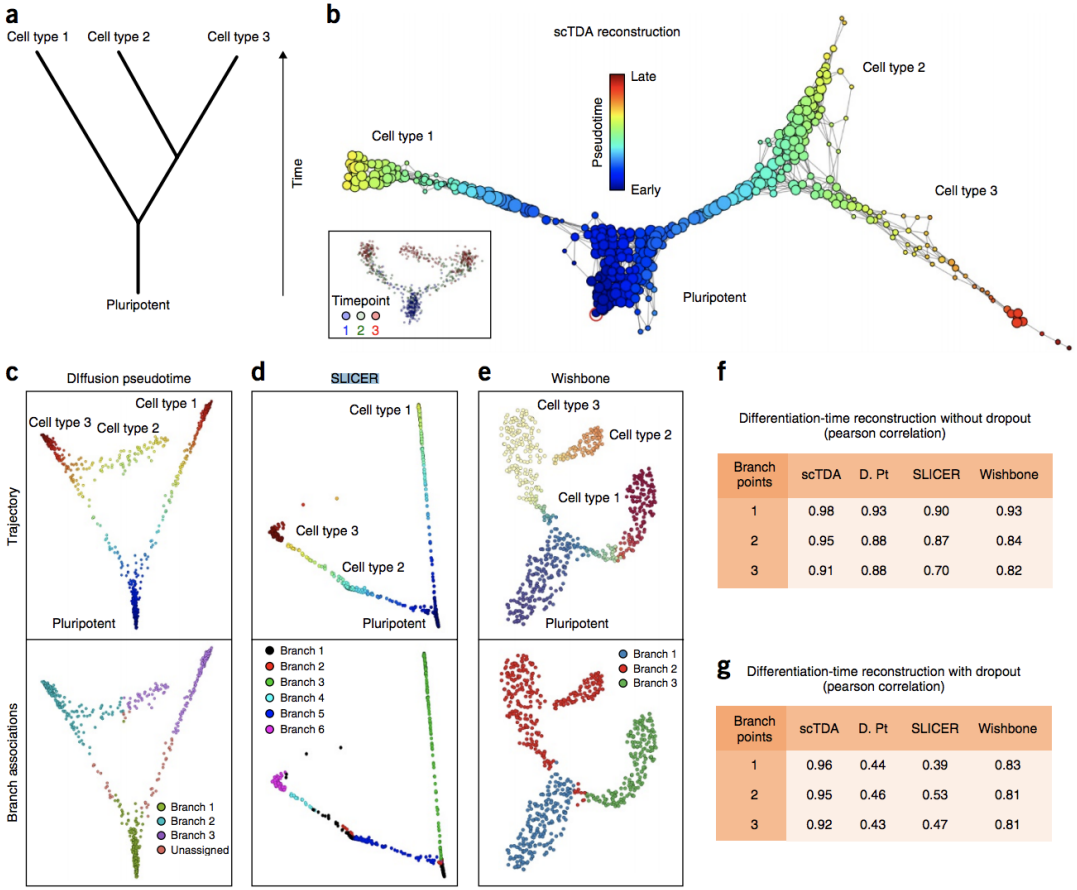


Figure: from Welch *et al.* Genome Biology. 2016.

SLICER Compared



Branch points	scTDA	D. Pt	SLICER	Wishbone
1	0.98	0.93	0.90	0.93
2	0.95	0.88	0.87	0.84
3	0.91	0.88	0.70	0.82

Branch points	scTDA	D. Pt	SLICER	Wishbone
1	0.96	0.44	0.39	0.83
2	0.95	0.46	0.53	0.81
3	0.92	0.43	0.47	0.81

Figure: from Rizvi *et al.* Nature Biotechnology. 2016.

Conclusion

- CyTOF Merge (simple imputation)
- Diffusion Maps as a first approach to capture some branching structure
- SLICER for LLE based embedding and paths from a fixed starting cell.

Next time...

- Benchmarking trajectory inference methods and how exactly you score all of these things (because there are so many of them).