STA 426: Diffusion maps for high-dimensional single-cell analysis of differentiation data

Giuachin Kreiliger, Samuel Pawel (GitHub: Kreile, SamCH93)

19.11.2018

Table of Contents

Motivation

Single Cell Variability and Differentiation

- Variability between cells and measurement error – Noise
- Differential expression of genes steering differentiation – Signal
- Continuous vs. abrupt, switch like gene expression change

Questions

- Which genes change in their expression due to differentiation?
- Given a single cell, what is its developmental stage?

$\rightarrow$ Retrieve the hidden temporal ordering of cells

Setup

- RNA-seq measurements from a cell population at one or multiple time points
- Visualize differentiation by aligning the cells along the differentiation trajectories

Statistical Model

Notation

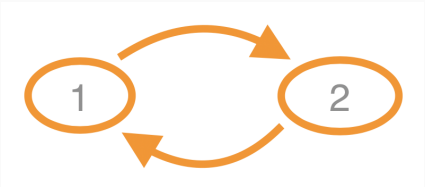
- Let n be the number of all cells and G the number of all genes measured
- $\mathbf{x}$ is the position of the cell in the multi-dimensional space $\mathbb{R}^G$
- Isotropic Gaussian wave function:

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

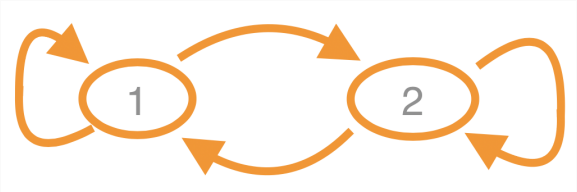
Markov Chain Transition Probability Matrix

Two state Markov chain:

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$



$$\begin{pmatrix} 0.3 & 0.7 \\ 0.2 & 0.8 \end{pmatrix}$$



Diffusion Transition Probability Matrix

- Interference between two cells $\mathbf{x}$ and $\mathbf{y}$:

$$\int_{-\infty}^{\infty} \mathbf{Y}_{\mathbf{x}}(\mathbf{x}')\mathbf{Y}_{\mathbf{y}}(\mathbf{x}')d\mathbf{x}' = \exp(-\frac{||\mathbf{y}-\mathbf{x}||^2}{2\sigma^2})$$

- Transition probability:

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

$$\text{where } Z(\mathbf{x}) = \sum_{i=1}^{n}\exp(-\frac{||\mathbf{y}_i-\mathbf{x}||^2}{2\sigma^2})$$

Diffusion Maps – Quantities

- Diffusion distance: Euclidian distance in the eigenvector space.
- Diffusion coefficients: Eigenvalues in the direction of their corresponding eigenvectors
- Assumption: Eigenvalues drop to a noise level after the first few components

Summary

- Similarly to the random fluctuation in real cells, we allow the cells to diffuse in $\mathbb{R}^G$
- Output: Main "diffusion" directions

Missing and Censored Values

- Decompose the kernel of the Gaussian wave into G components
- Replace missing/censored values with hypothetical distribution

Determination of the Gaussian Kernel Width

- The average dimensionality of the manifold is equal to

$$\langle d(\sigma) \rangle_x = \frac{\partial \langle \log(Z(\mathbf{x})) \rangle_x}{\partial \log(\sigma)}$$

- Choose σ such that $d(\sigma)\rangle_x$ is maximized
- Possibly multiple maxima

Comparison with Other Methods

Dimension Reduction Methods for Single-Cell Omics Data

Principal Component Analysis (PCA)

- Orthogonal transformation of the data
- Works best for linear data subspace
- Non-linear extensions (Kernel-PCA)

Dimension Reduction Methods for Single-Cell Omics Data

Principal Component Analysis (PCA)

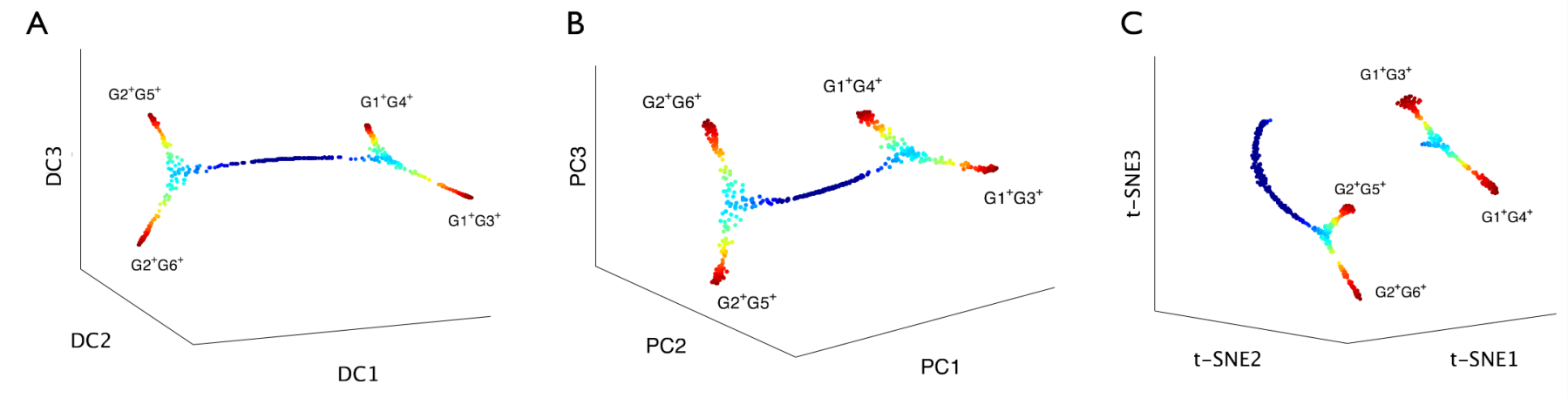
- Orthogonal transformation of the data
- Works best for linear data subspace
- Non-linear extensions (Kernel-PCA)

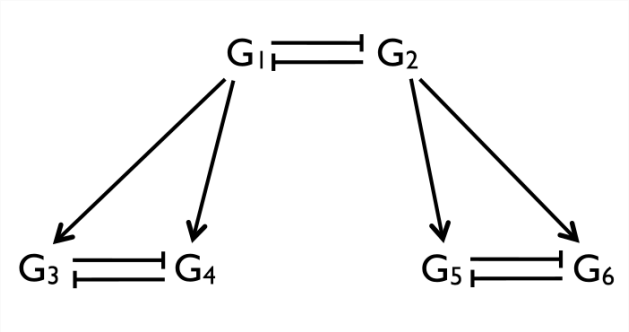
t-distributed Stochastic Neighbor Embedding (t-SNE)

- Probabilistic, non-linear method
- Very robust to noise / density heterogenities
- Tuning parameter perplexity

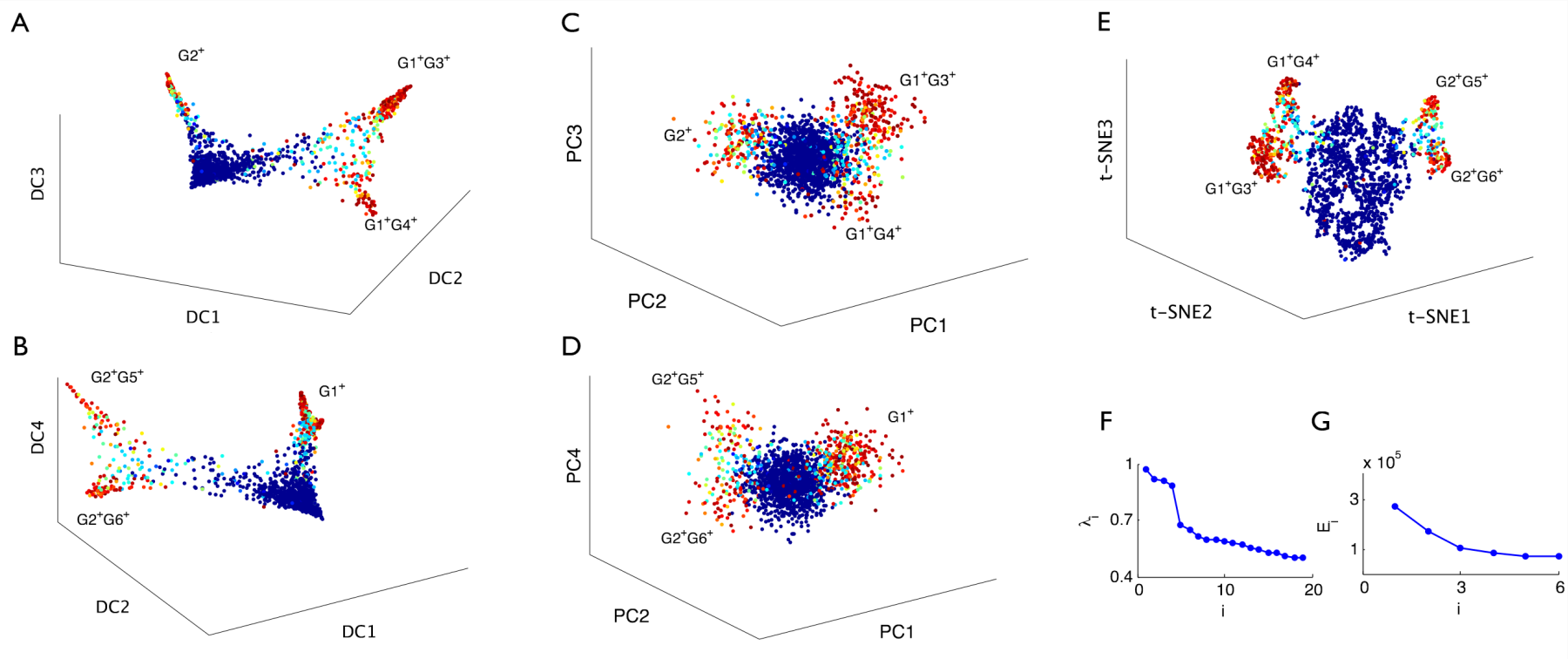
Many more methods exist!

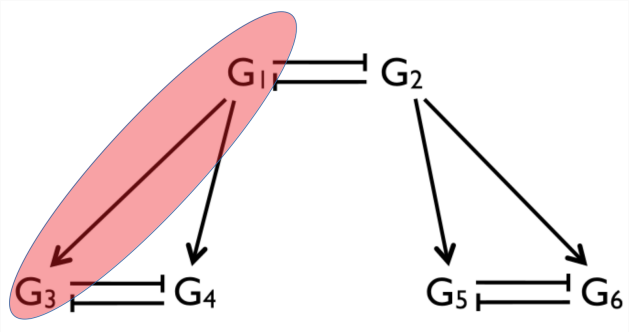
Simulated Data – Balanced



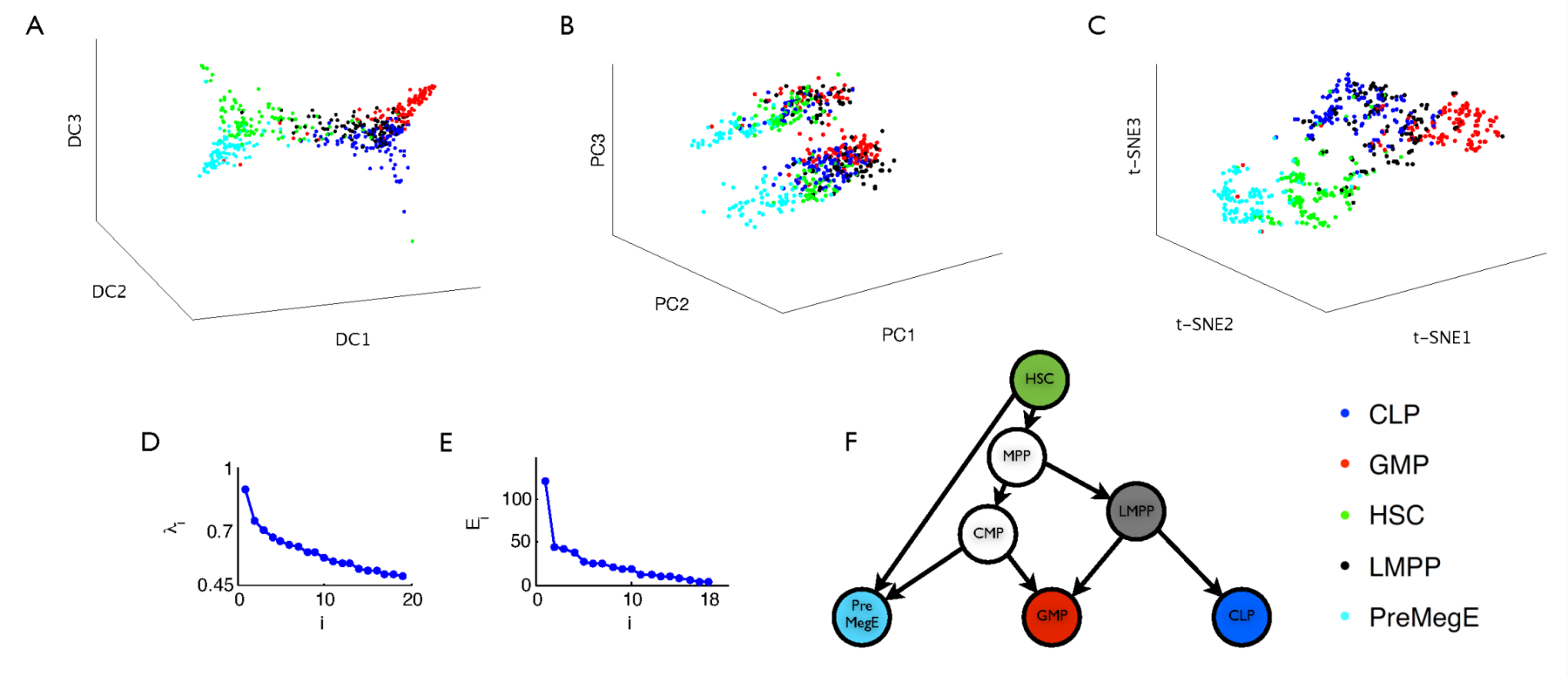


Simulated Data – Imbalanced + Noise





qPCR Data – Haematopoietic Stem Cells



Results

+ Diffusion maps outperform other methods in finding cell differentiation trajectories
+ Diffusion maps are robust to sampling density heterogenities and noise

Results

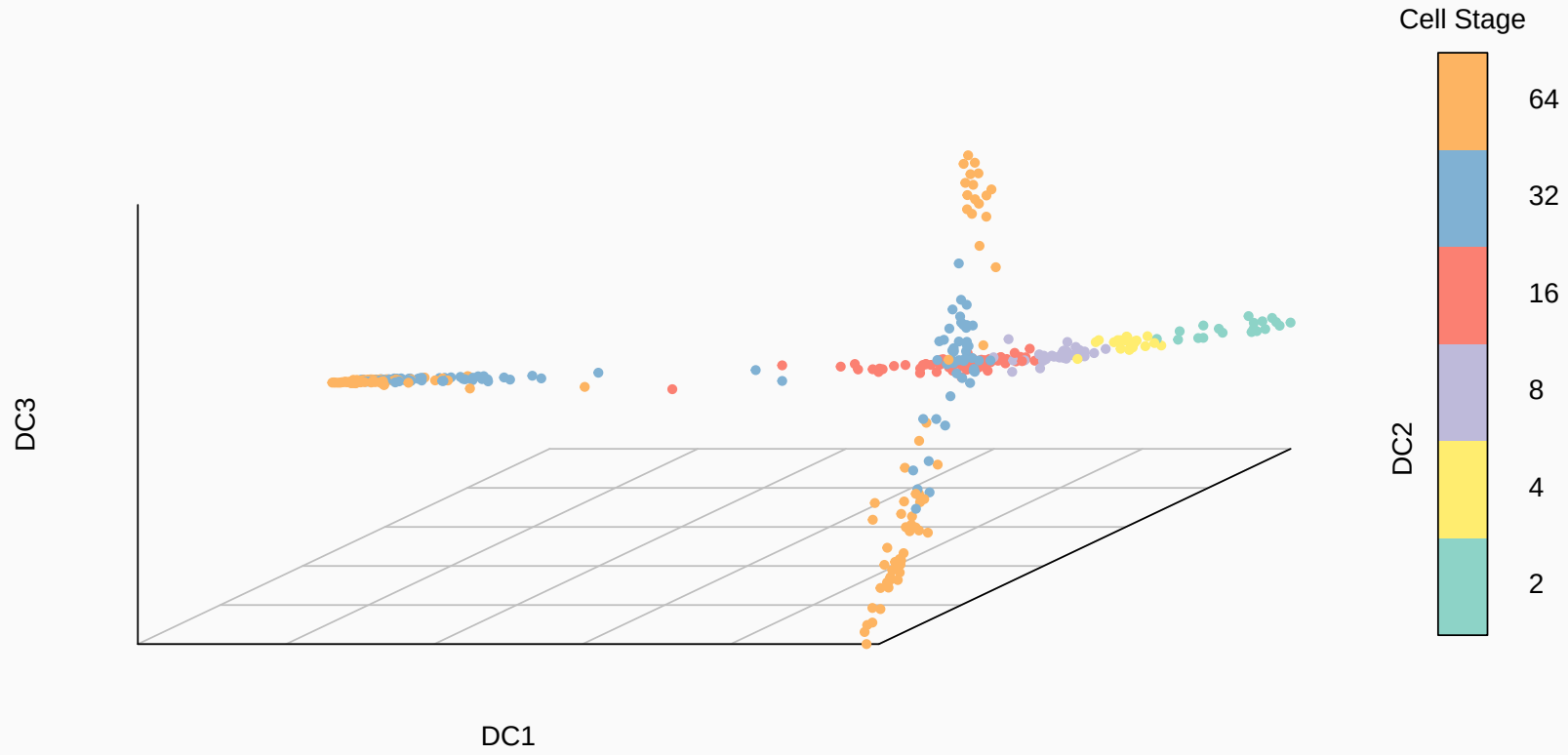
+ Diffusion maps outperform other methods in finding cell differentiation trajectories
+ Diffusion maps are robust to sampling density heterogenities and noise

- In all examples σ^2 was finetuned
- Unclear if finetuning perplexity parameter in t-SNE method would lead to similar results
- Number of diffusion components needed for visualization not known in advance

R Demonstration

destiny Package

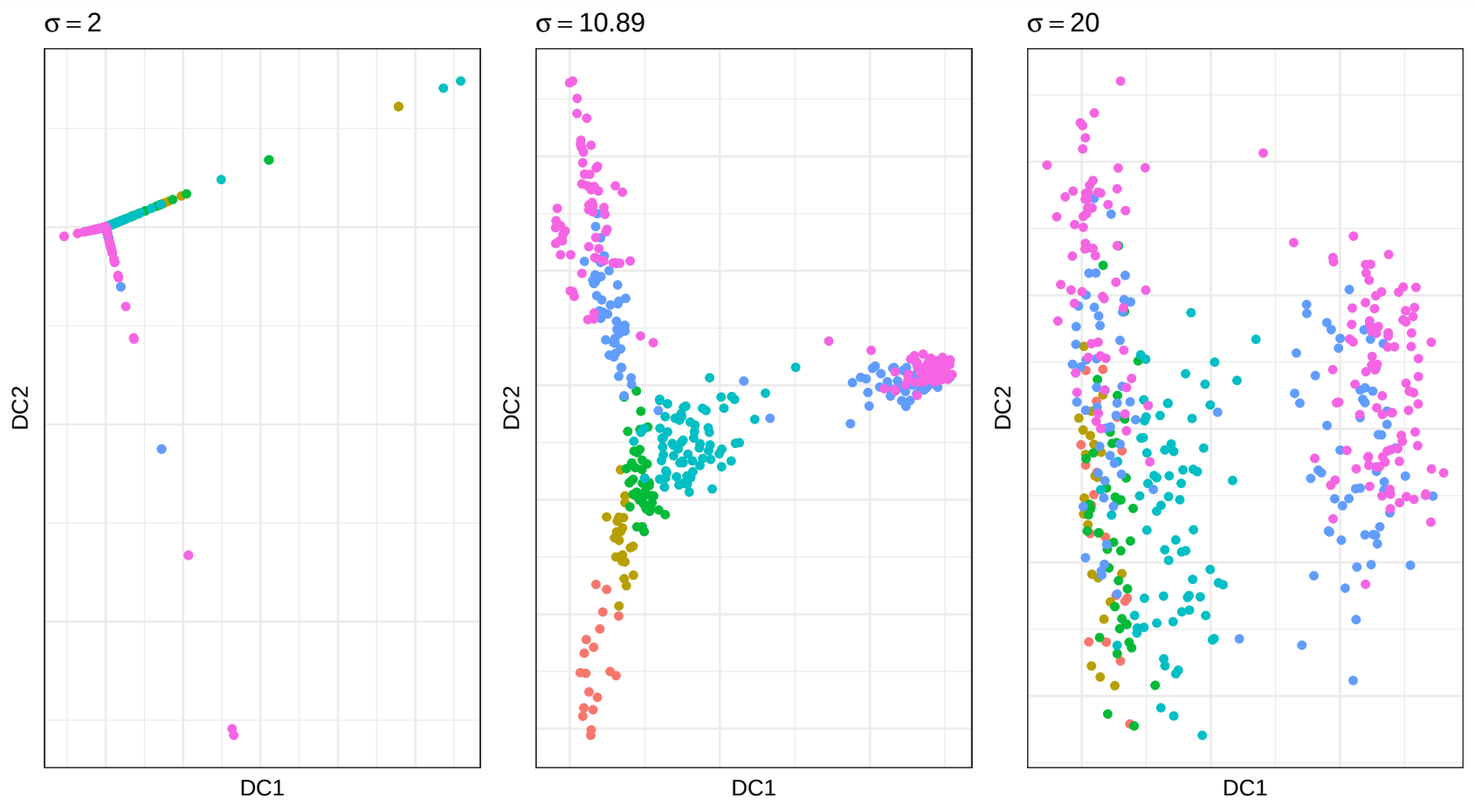
```
set.seed(12)
library(destiny)
data("guo_norm")
dm_guo <- DiffusionMap(guo_norm)
plot(dm_guo, pch = 20, col_by = "num_cells", legend_main = "Cell Stage")
```



Choice of σ^2

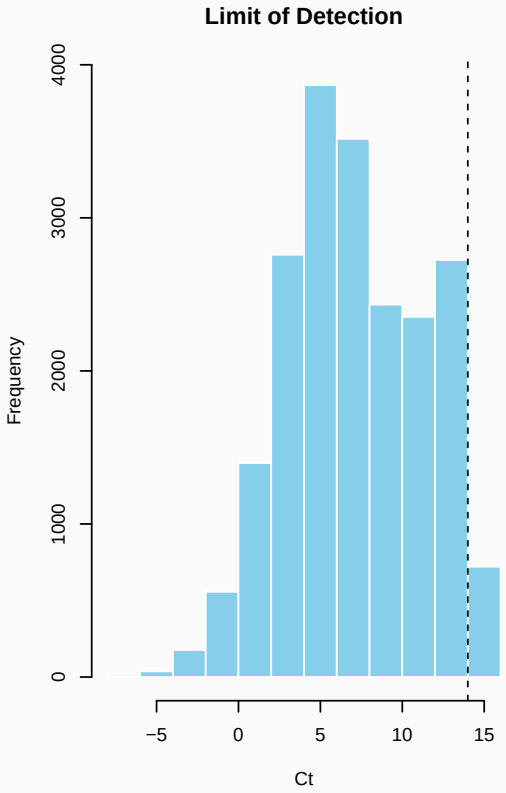
```
sigmas <- find_sigmas(guo_norm, verbose = FALSE)
optimal_sigma(sigmas)
```

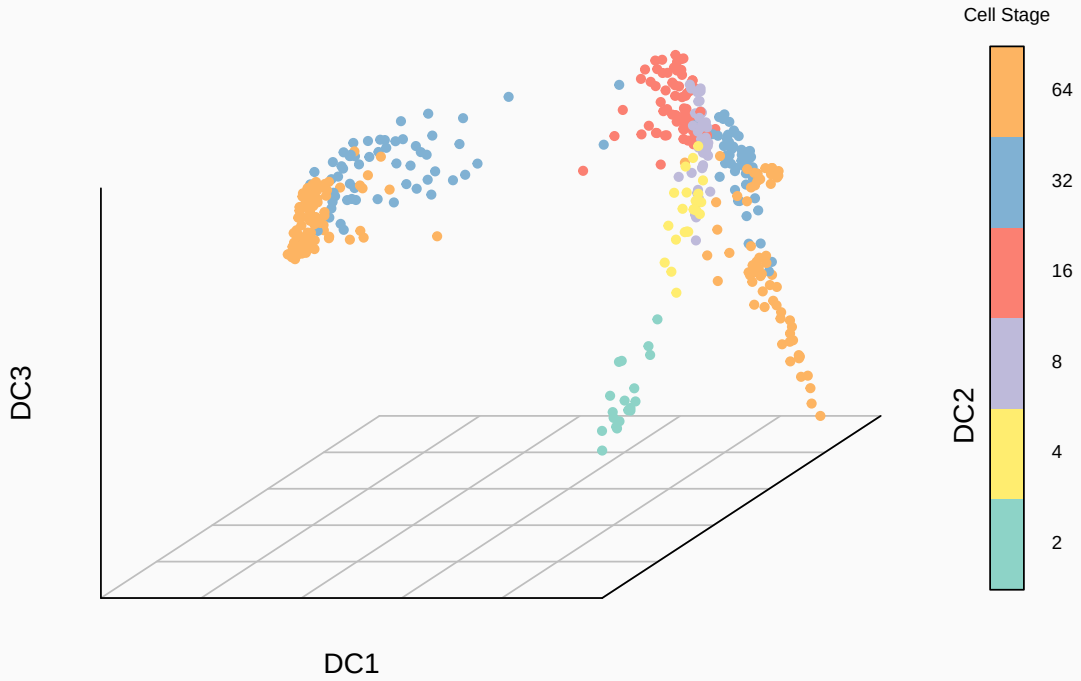
```
## [1] 10.8946
```



Censored/Missing Values

```
layout(matrix(c(1, 2, 2), ncol = 3))
exprs(guo_missing)[sample(length(exprs(guo_missing)), 100)] <- NA
dm_guo_missing <- DiffusionMap(guo_missing, verbose = FALSE,
                               censor_val = 14, censor_range = c(14, 40),
                               missing_range = c(-5, 10))
hist(exprs(guo_norm), main = "Limit of Detection", xlab = "Ct", col = "skyblue", border = "white")
abline(v = 14, lty = 2)
plot(dm_guo_missing, pch = 20, cex.symbols = 1.5, col_by = "num_cells", legend_main = "Cell Stage")
```





Conclusions

Discussion

+ Handle high noise levels, missing data, sampling heterogenitites
+ Diffusion distance is a biologically relevant distance metric
+ Capture nonlinear/complex differentiation dynamics

Discussion

+ Handle high noise levels, missing data, sampling heterogenitites
+ Diffusion distance is a biologically relevant distance metric
+ Capture nonlinear/complex differentiation dynamics

- Number of significant dimension not determinable in advance
- Finetuning of σ^2 required (but there is a proposed criterion)
- $n^2 \times G$ computation time (but there are approximate versions)

Discussion

+ Handle high noise levels, missing data, sampling heterogenitites
+ Diffusion distance is a biologically relevant distance metric
+ Capture nonlinear/complex differentiation dynamics

- Number of significant dimension not determinable in advance
- Finetuning of σ^2 required (but there is a proposed criterion)
- $n^2 \times G$ computation time (but there are approximate versions)

$\rightarrow$ Powerful dimension reduction tool for single cell differentiation data

References

Angerer, P., Haghverdi, L., Büttner, M., Theis, F., Marr, C., and Büttner, F. (2015). `destiny`: diffusion maps for large-scale single-cell data in R. *Bioinformatics*, 32(8):1241–1243.

Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, 21:5–30.

Haghverdi, L., Buettner, F., and Theis, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998.