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

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

19.11.2018

Overview

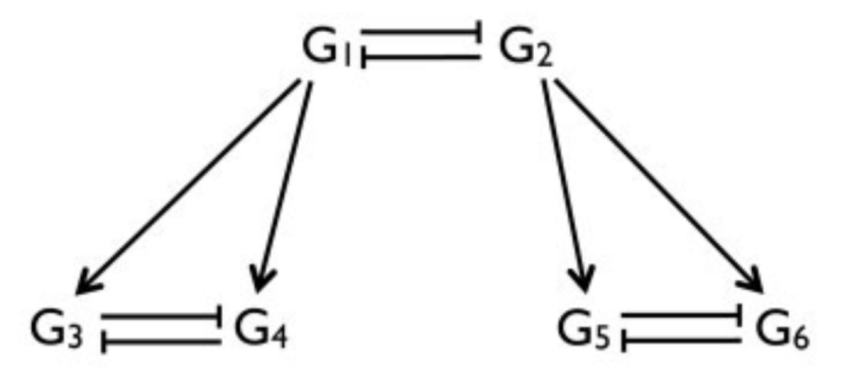
- Motivation - Statistical model - Comparison and evaluation

Introduction

Diffusion maps: A dimensionality reduction tool for non-linear data.

What is non-linear data?

Toy model



First low baseline expression

One gene starts to inhibit the other with equal probability -¿ Branching.

Biological Background

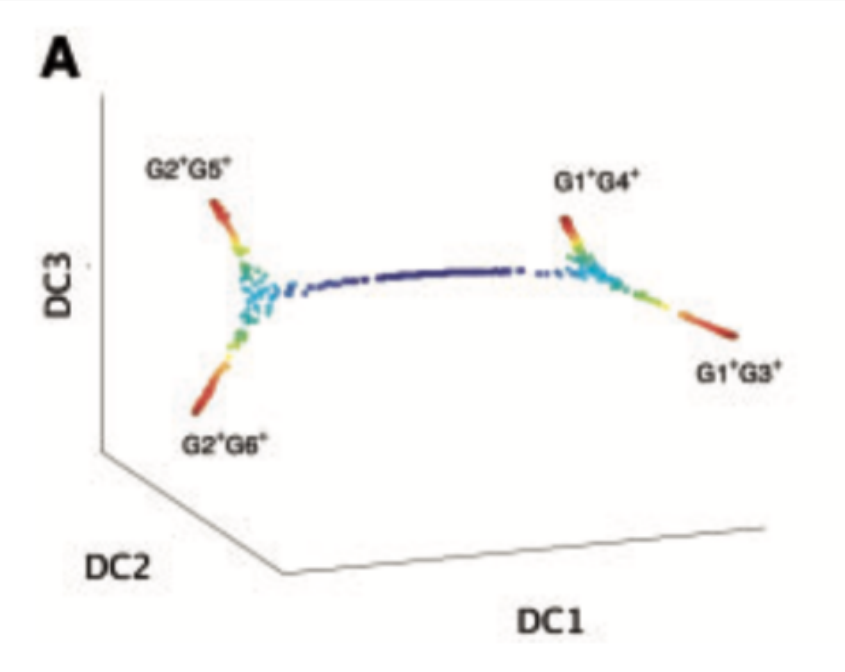
Single cell gene expression heterogeneity and sampling error - noise.

Differential expression of genes steering differentiation - signal.

Measurements of RNA expression profiles from large numbers of cells at different developmental stages.

Aim of the paper

Establishment a pseudo-temporal order of the cells.



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 $\mathbb{R}^G$

Core of the model

Isotropic gaussian wave function:

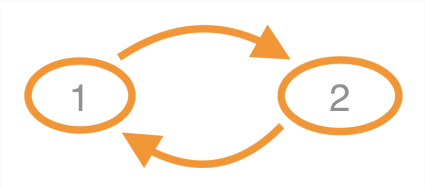
$$Y_{\mathrm{x}}(\boldsymbol{x}') = \left(\frac{2}{\pi\sigma^2}\right)^{1/4} \exp\left(-\frac{||\boldsymbol{x}' - \boldsymbol{x}||^2}{\sigma^2}\right)$$

And markov chain transition probability matrix P (ergodic for large enough σ).

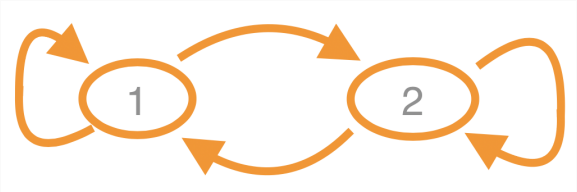
Markov chain transition probability matrix

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

Diffusion distance: euclidian distance in the eigenvector space.

Diffusion coefficients: eigenvalues in the direction of their corresponding eigenvectors

Dimensionality Reduction

Eigenvalues drop to a noise level after the first l components.

Approximation of the diffusion distance.

Summary

When each cell is allowed to "diffuse" randomly;

We can compute the probability that it will "diffuse" into another cell.

And show the direction in which most "diffusion" happens.

Missing Values

Replace the Gaussi of a missing Gene g' with a prior representing our assumption of the distribution of the missing gene.

Choice of a gaussian distribution may be preferred over uniform distribution due to computational reasons.

Determination of the Gaussian Kernel Width

$Z(x)$ partition function approximating the number of cells into which the cell x can diffuse.

Assume:

$$Z(\boldsymbol{x}) \propto \sigma^{d(\boldsymbol{x},\sigma)}$$

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)

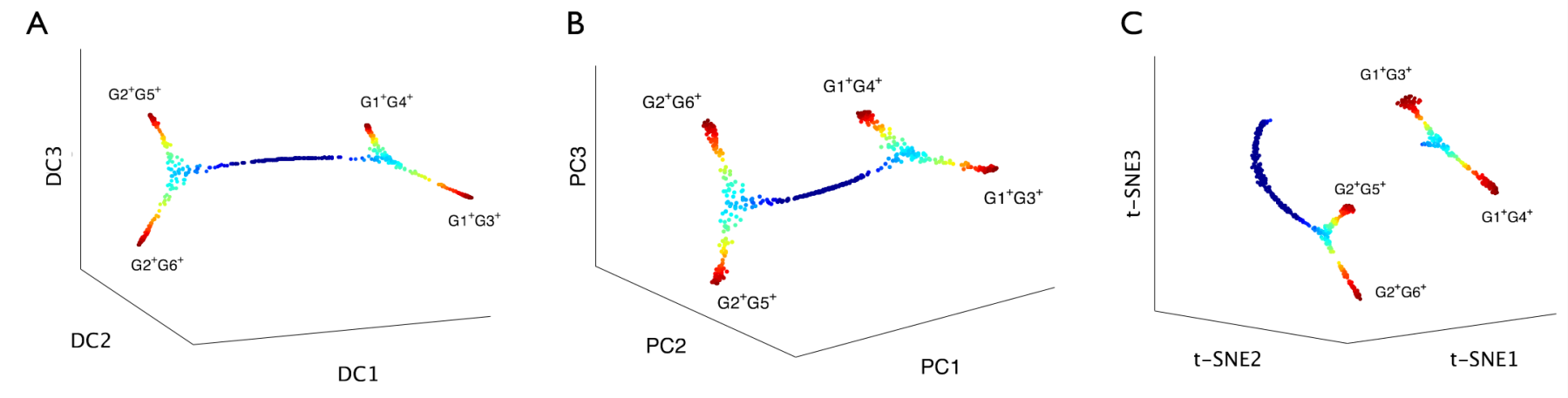
- Orthogonal transformation of the data
- Works best for linear data subspace
- Non-linear extensions (Kernel-PCA)

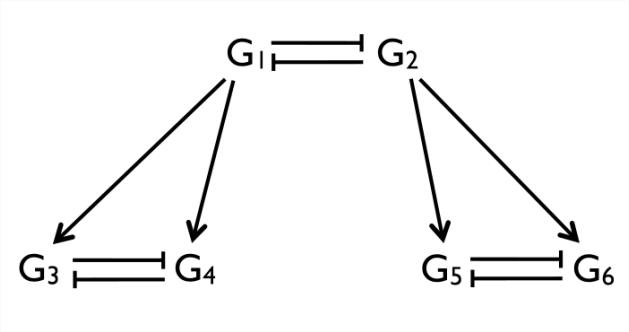
t-Distributed Stochastic Neighbor Embedding (t-SNE)

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

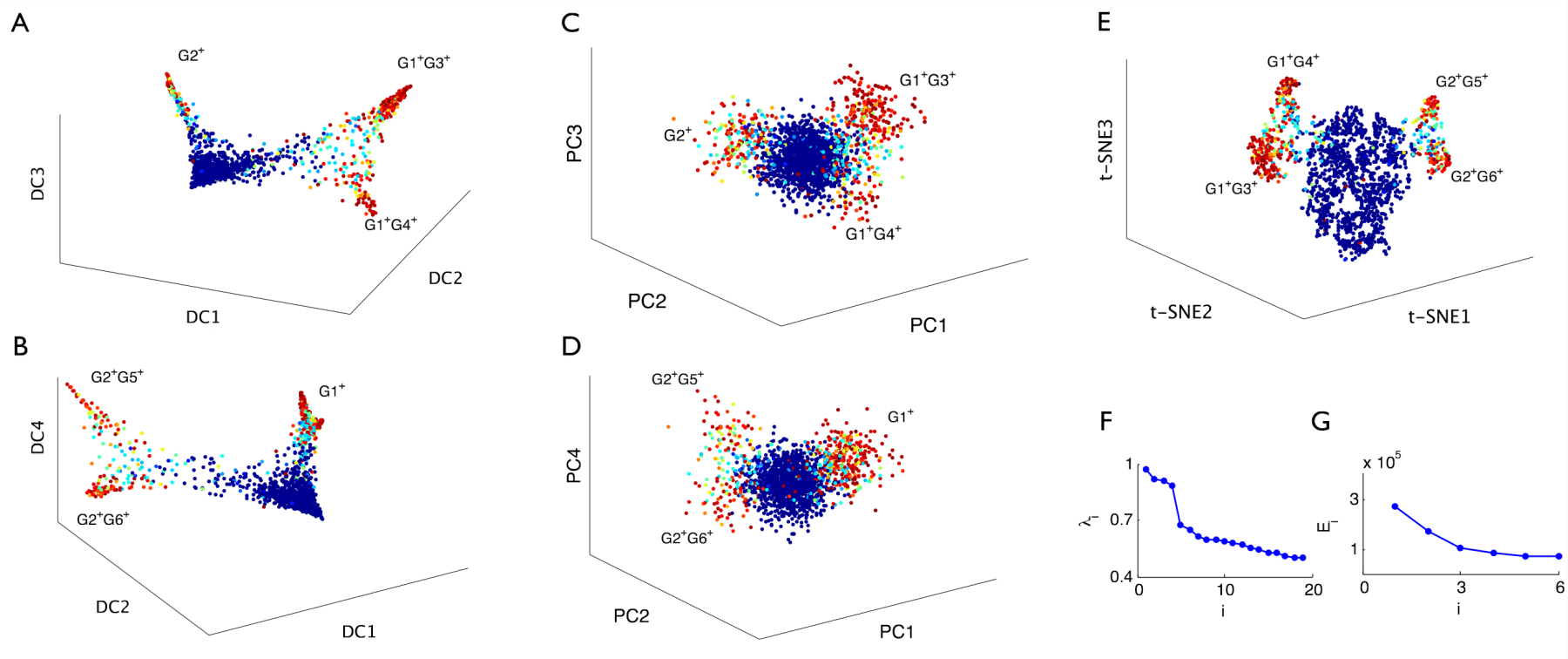
Many more methods exist!

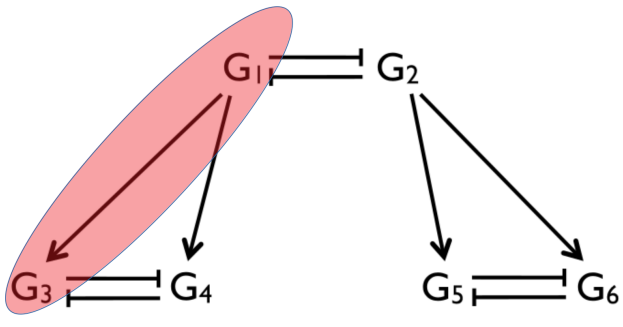
Simulated Data – Balanced



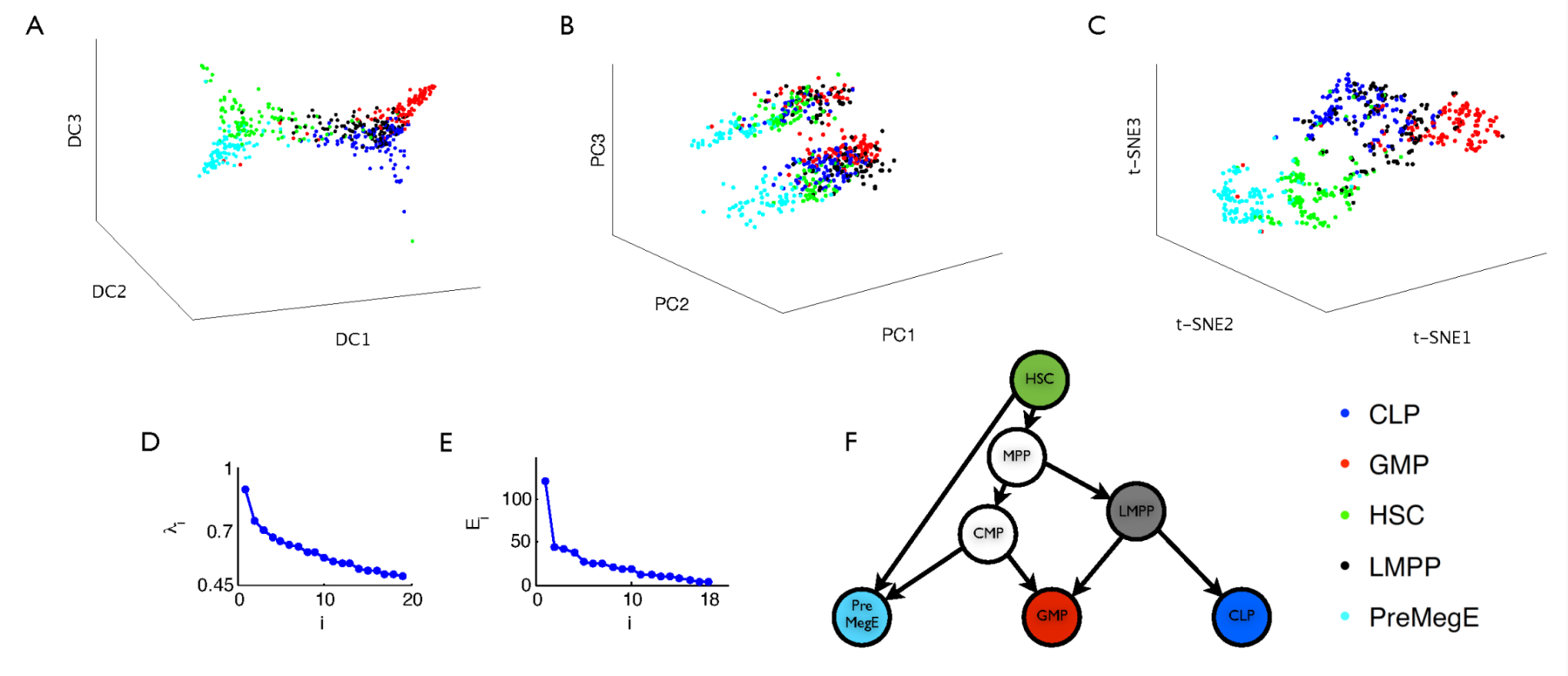


Simulated Data – Imbalanced + Noise





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

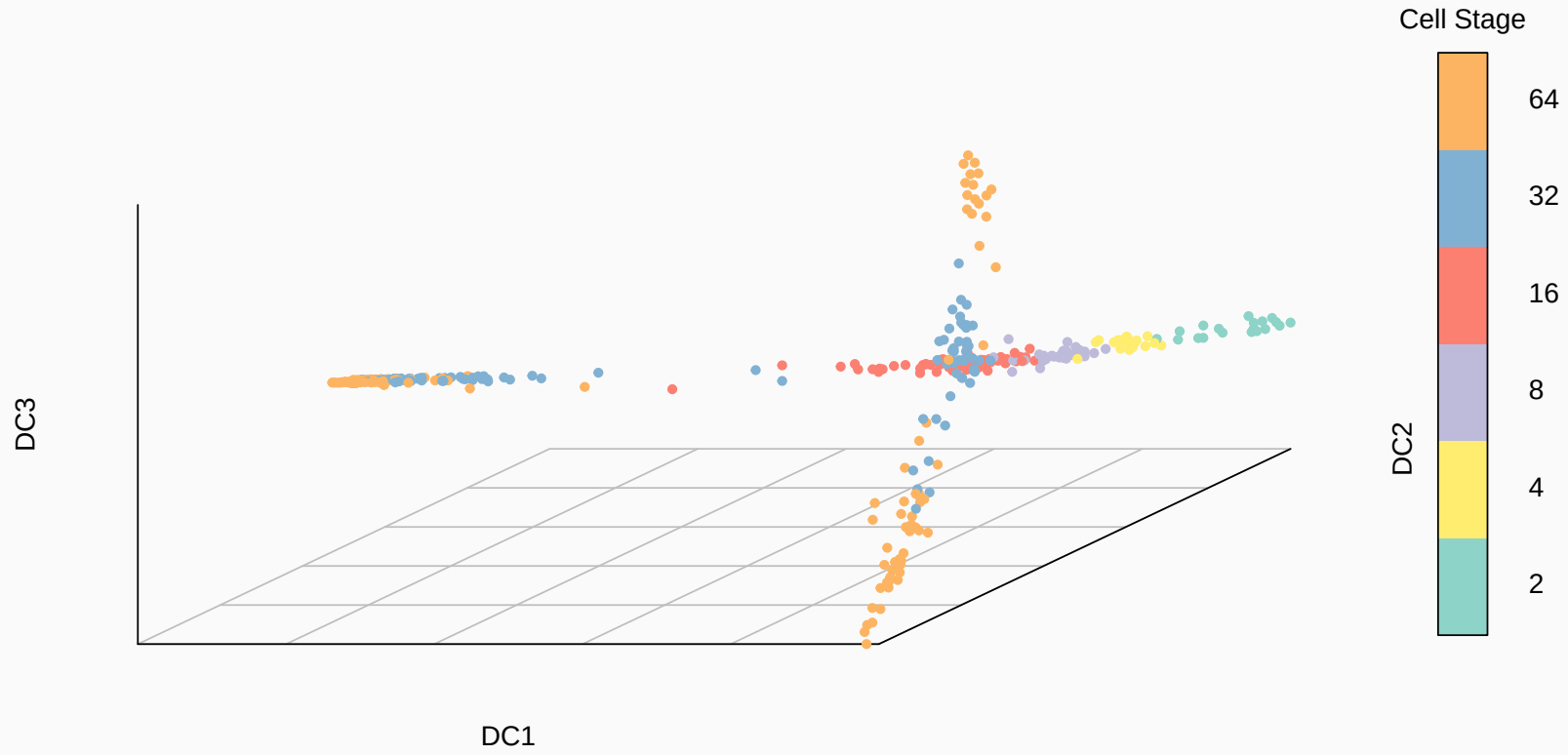
+ Diffusion maps outperform other methods in finding cell differentiation trajectories
+ Diffusion maps are robust to sampling density heterogenities and noise

- 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

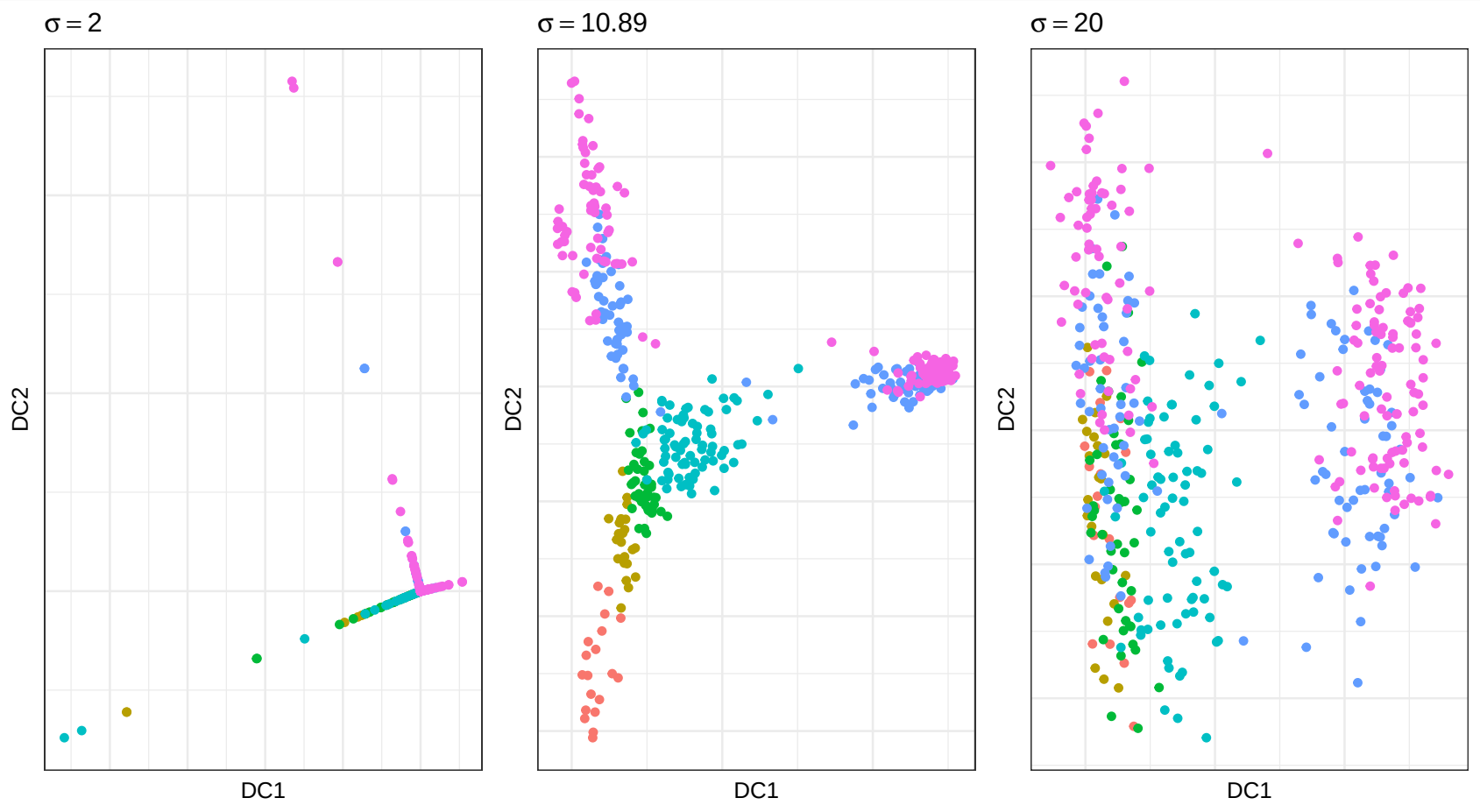
```
set.seed(12)
library(destiny)
data("guo_norm")
dm_guo <- DiffusionMap(guo_norm)
plot(dm_guo, pch = 20, col_by = "num_cells", legend_main = "Cell Stage")
```



Choice of σ^2

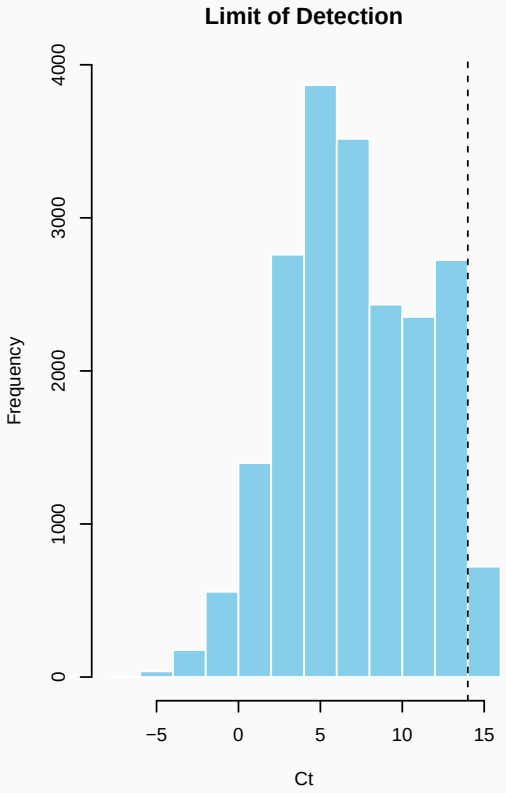
```
sigmas <- find_sigmas(guo_norm, verbose = FALSE)
optimal_sigma(sigmas)
```

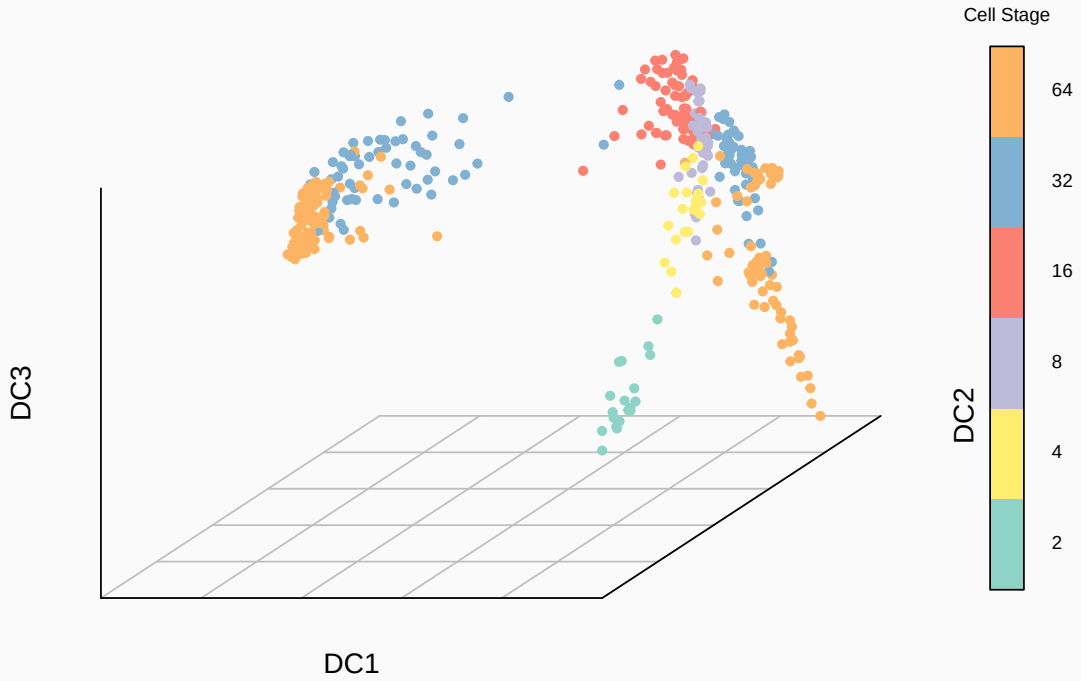
```
## [1] 10.8946
```



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")
```





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.