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

Giuachin Kreiliger, Samuel Pawel

19.11.2018

Introduction

Technical Details

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)

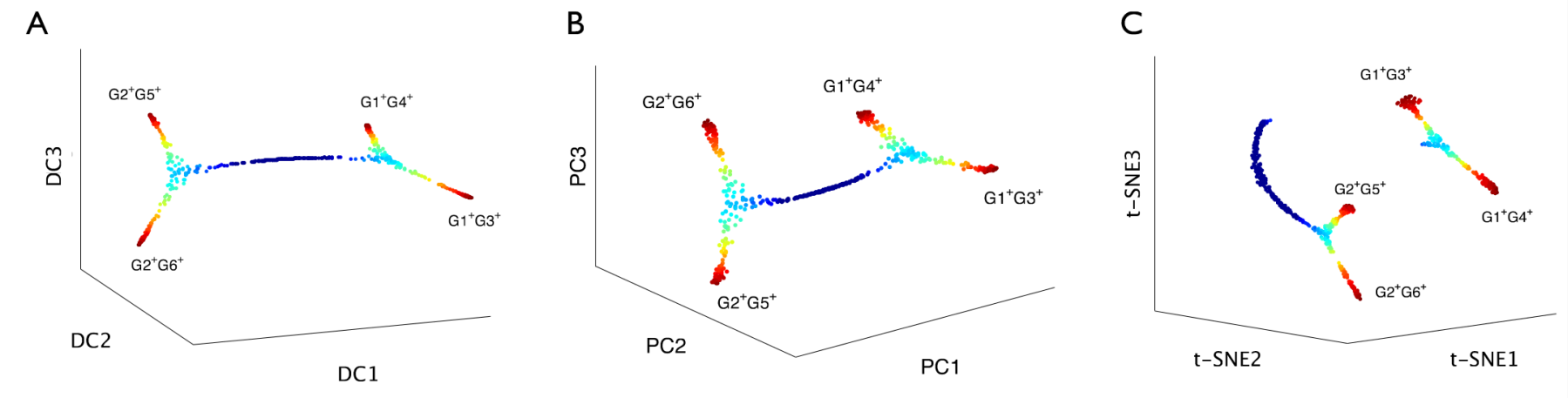
- Orthogonal transformation of the data
- Works best for linear data subspace
- Non-linear extensions (Kernel-PCA)

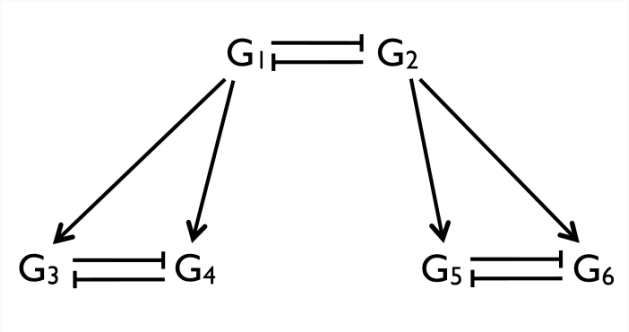
t-Distributed Stochastic Neighbor Embedding (t-SNE)

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

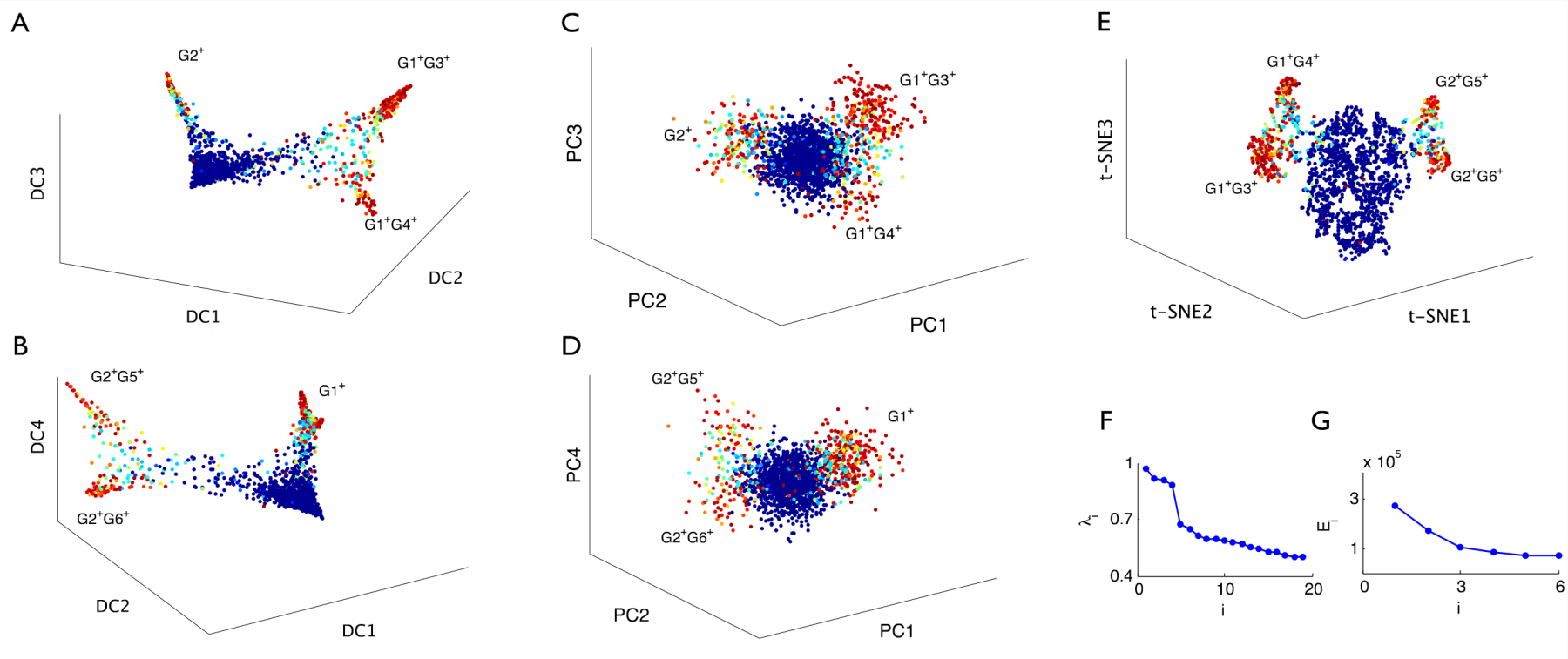
Many more methods exist!

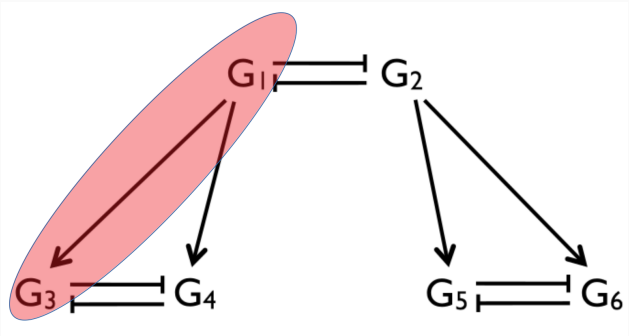
Simulated Data – Balanced



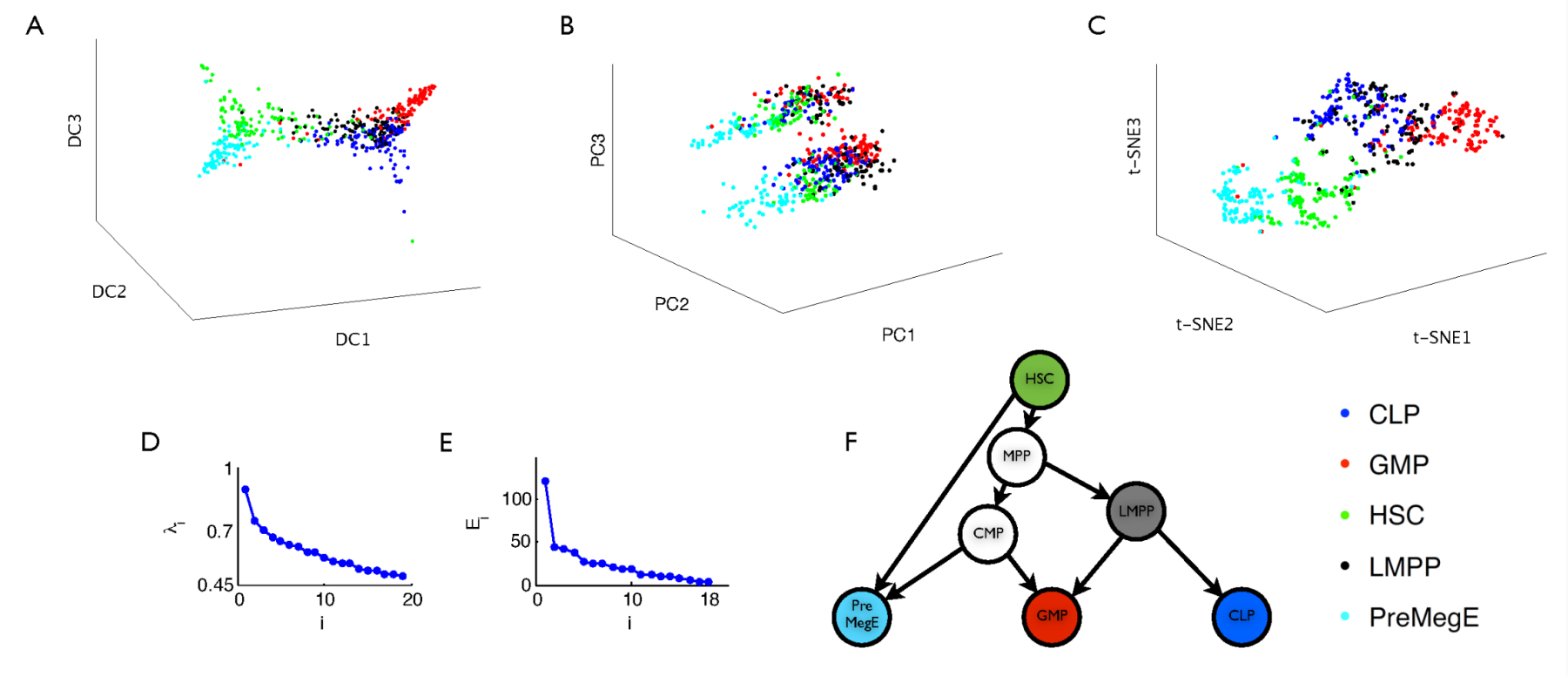


Simulated Data – Imbalanced

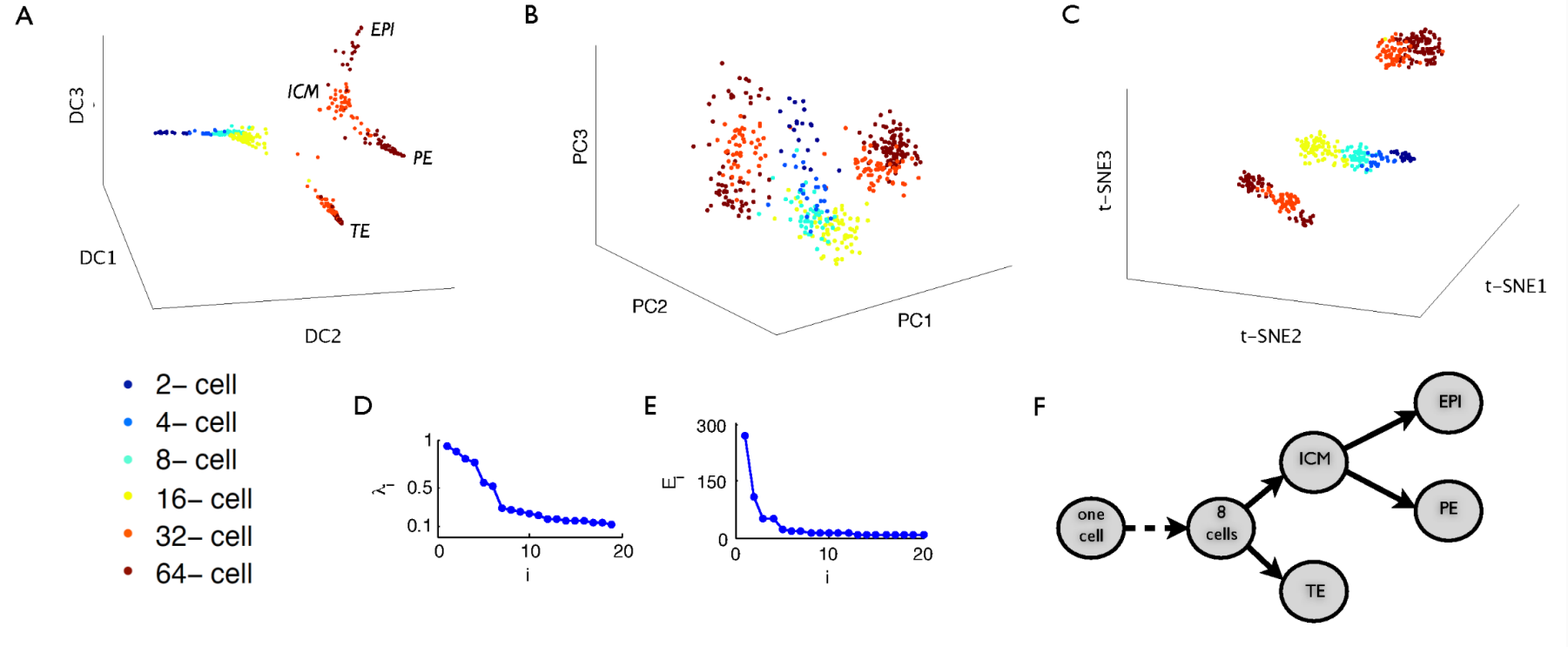




qPCR Data – Haematopoietic Stem Cells



qPCR Data – Mouse Embryonic Stem Cells



Results

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

Results

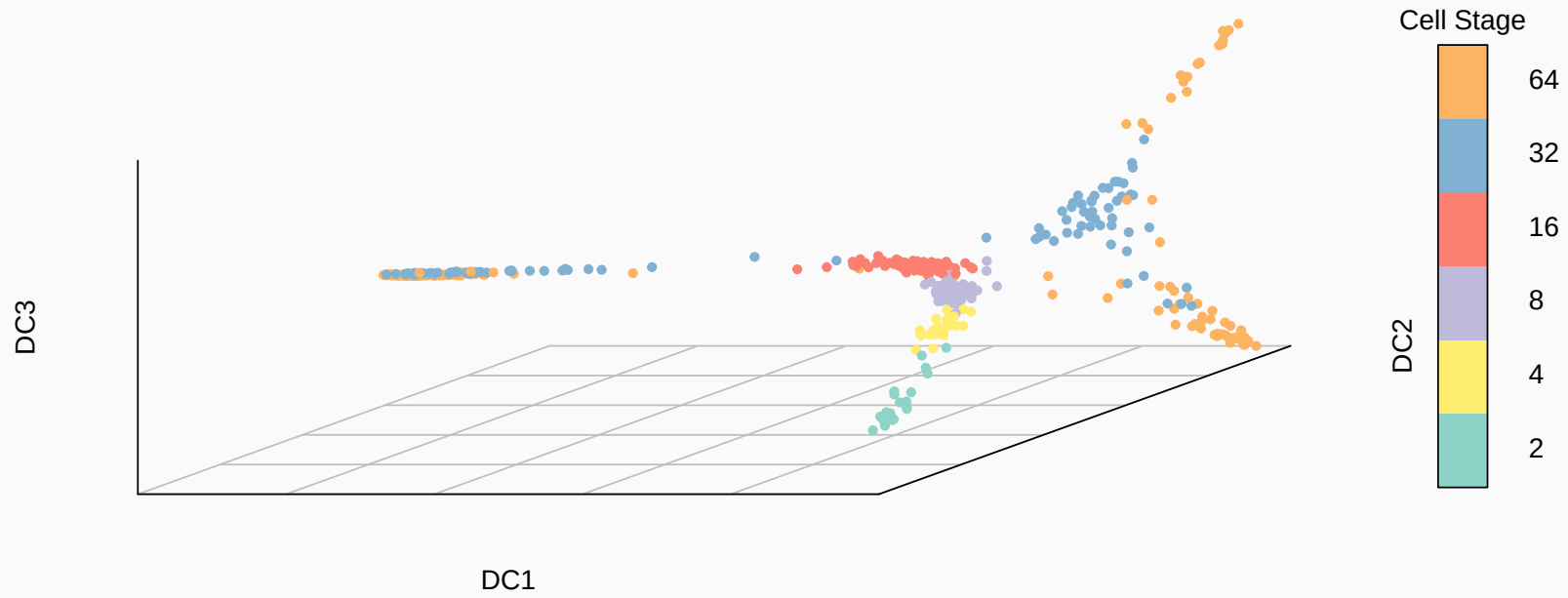
+ Diffusion maps outperform other methods in finding cell differentiation trajectories
+ Diffusion maps are robust to sampling density heterogenities and noise

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

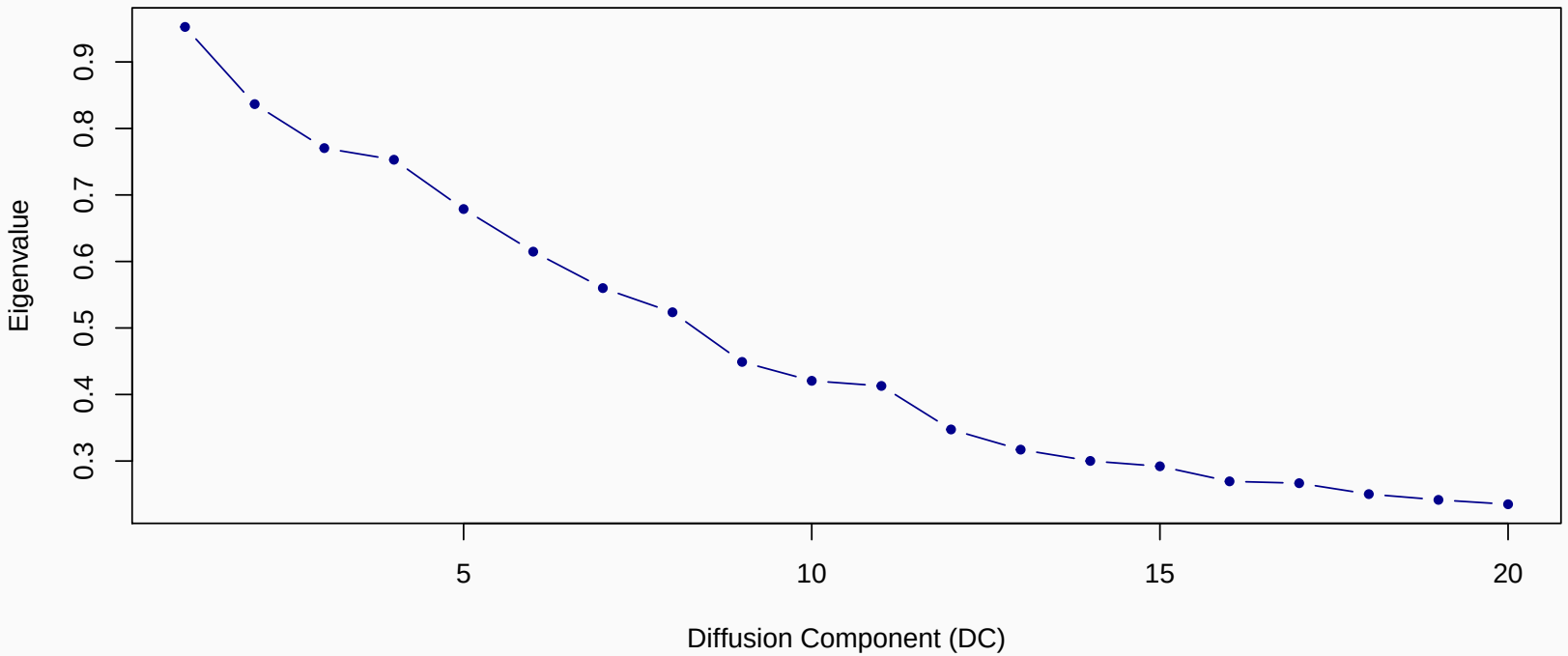
R Demonstration

destiny Package

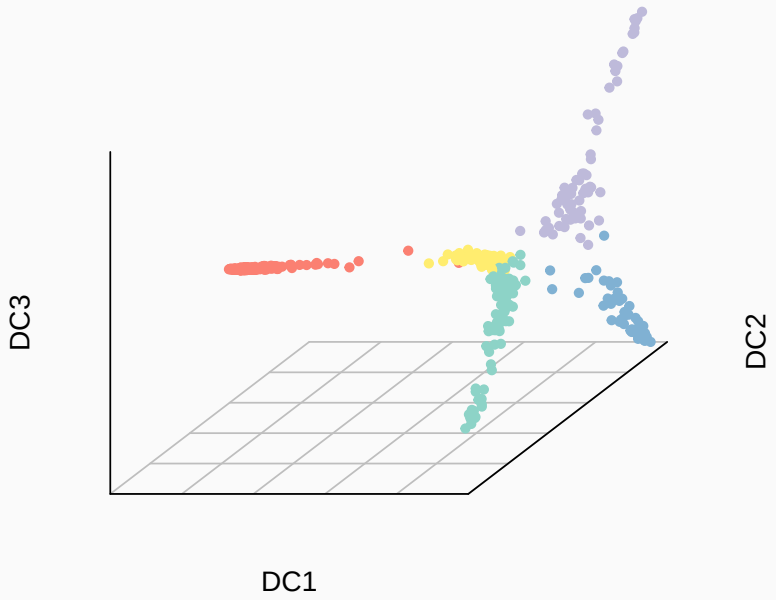
```
set.seed(12)
library(destiny)
data(guo)
dm_guo <- DiffusionMap(guo)
plot(dm_guo, pch = 20, col_by = "num_cells", legend_main = "Cell Stage")
```

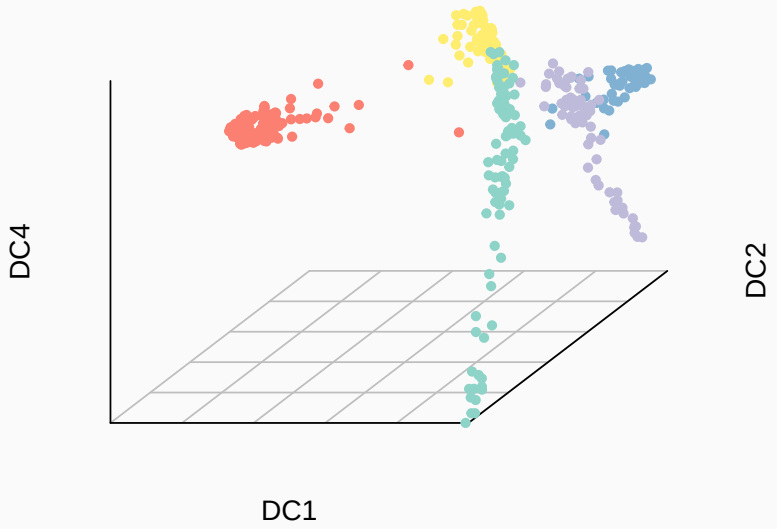


Eigenvalues of Diffusion Components



Plotting Several Diffusion Components

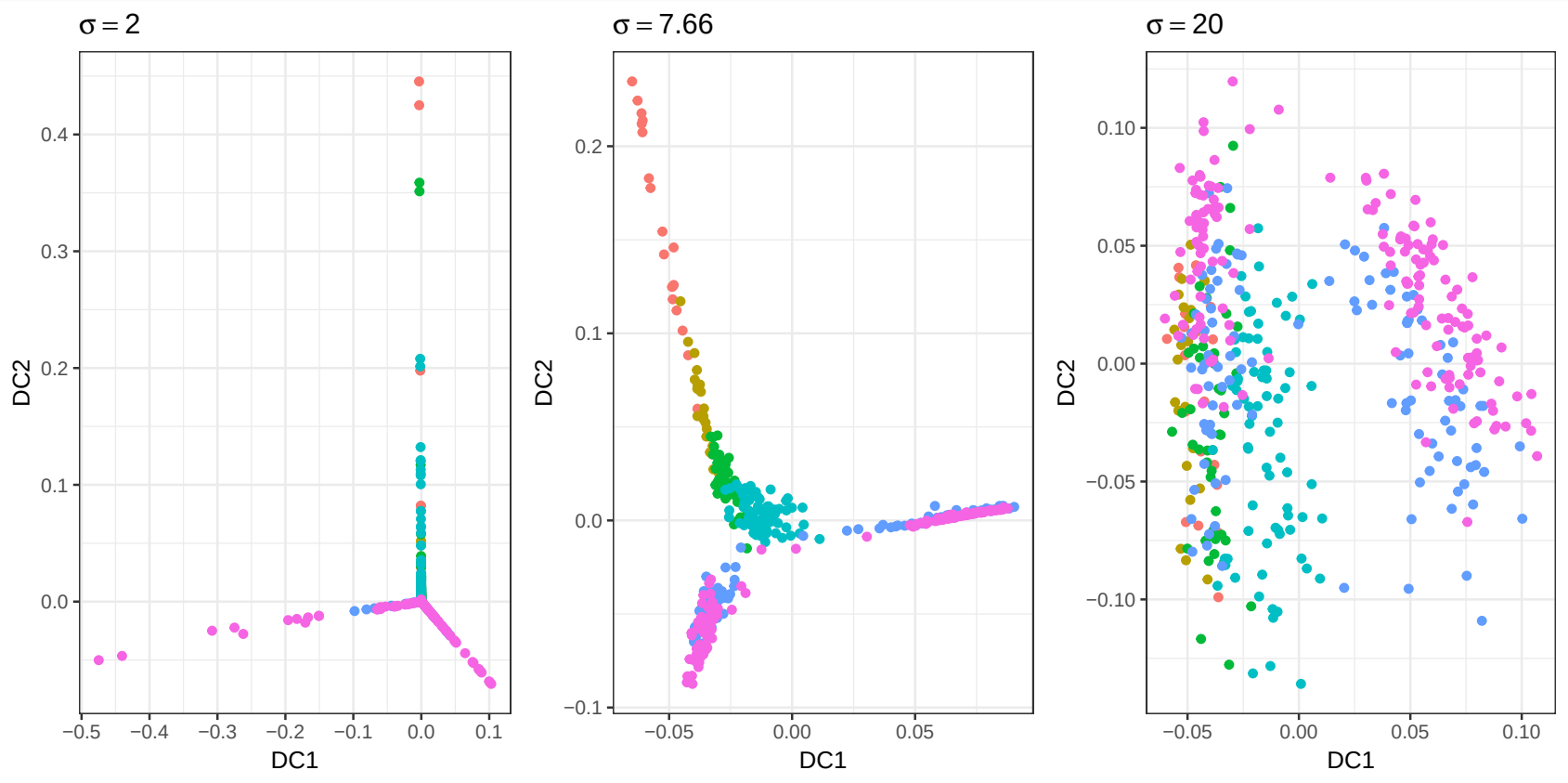




Choice of σ^2

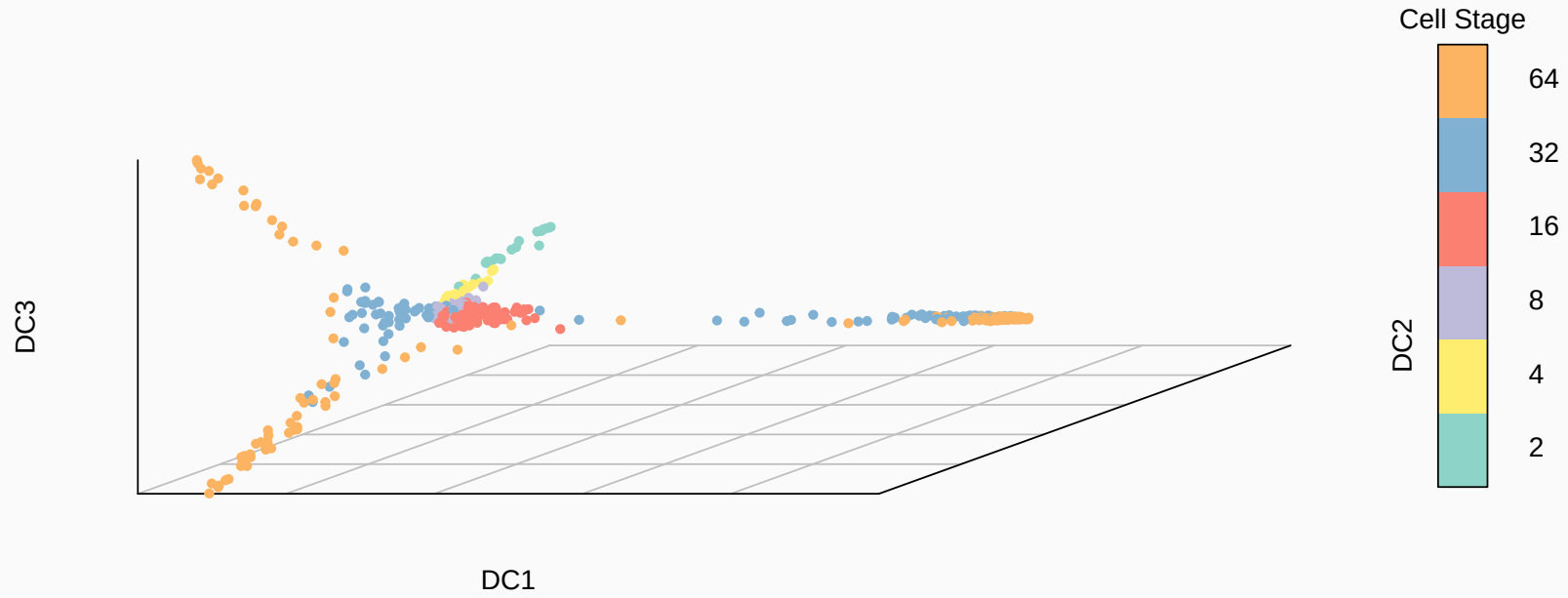
```
sigmas <- find_sigmas(guo, verbose = FALSE)
optimal_sigma(sigmas)
```

```
## [1] 7.663719
```



Censored/Missing Values

```
guo_missing <- guo
exprs(guo_missing)[sample(length(exprs(guo_missing)), 100)] <- NA
dm_guo_missing <- DiffusionMap(guo_missing, verbose = FALSE,
                               censor_val = 15, censor_range = c(15, 40),
                               missing_range = c(-5, 10))
plot(dm_guo_missing, pch = 20, 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.