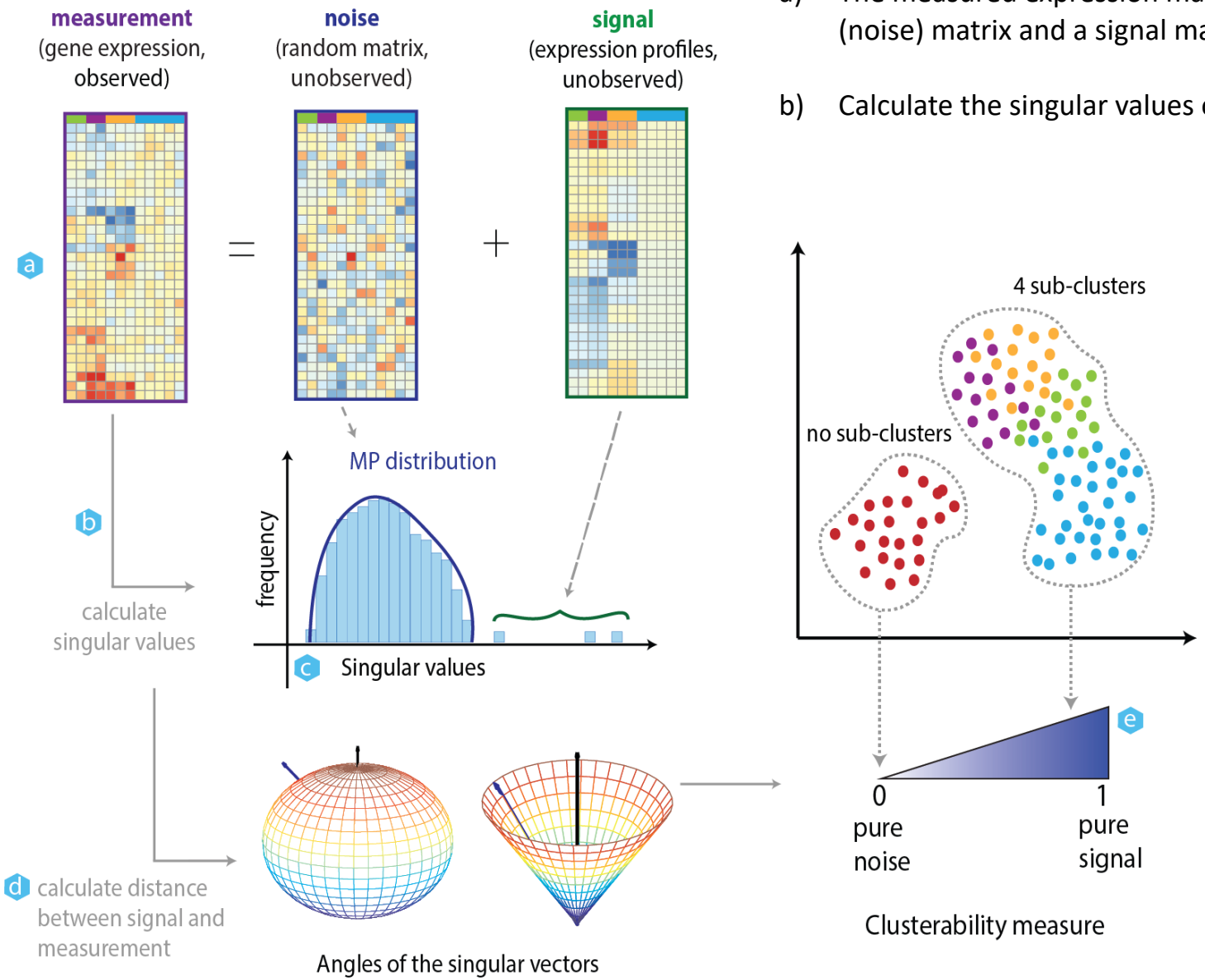
# A measure to determine clusterability in single cell RNA-seq data

**Maria Mircea**, Mazène Hochane, Diego Garlaschelli, Stefan Semrau

Universiteit Leiden




a) The measured expression matrix can be viewed as the sum of a random (noise) matrix and a signal matrix, which contains deterministic variations.

b) Calculate the singular values of the expression matrix.

c) The distribution of the singular values of the noise matrix corresponds to the Marchenko-Pastur (MP) distribution. The singular values that do not fall under the MP distribution (here 3 singular values) arise from the signal matrix and account for the biological variability.

d) The cosine of the angle between the singular vector of the expression matrix and the singular vector of the unobserved signal matrix is calculated. This is our measure for clusterability.

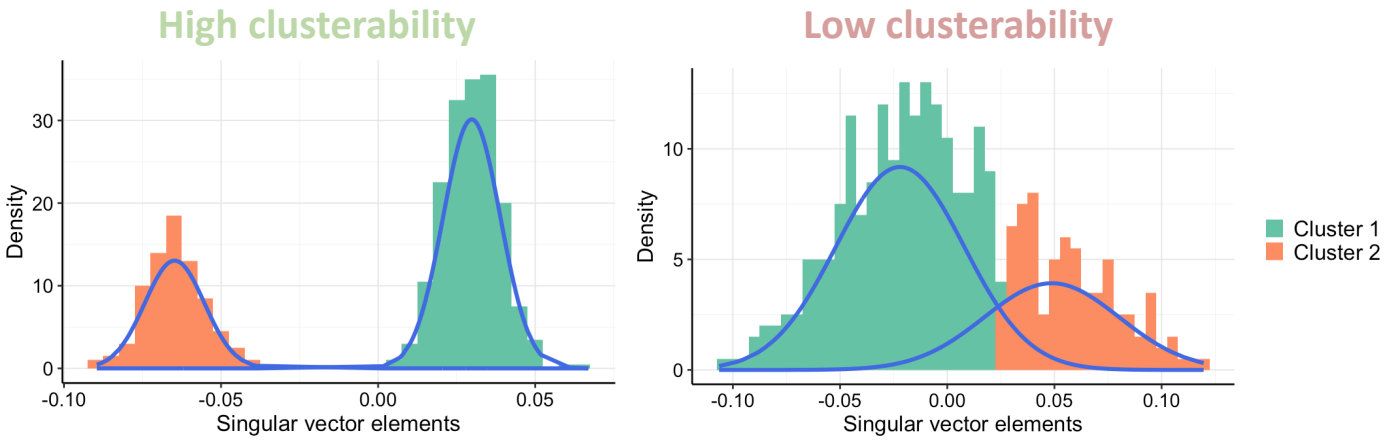
e) Our clusterability measure spans between 0 and 1. A value of 0, indicates that the cluster does not contain additional signal, thus, there is no need to sub-cluster. A value that is close to 1 indicates a higher likelihood to have relevant sub-structures in the data.

Semrau lab

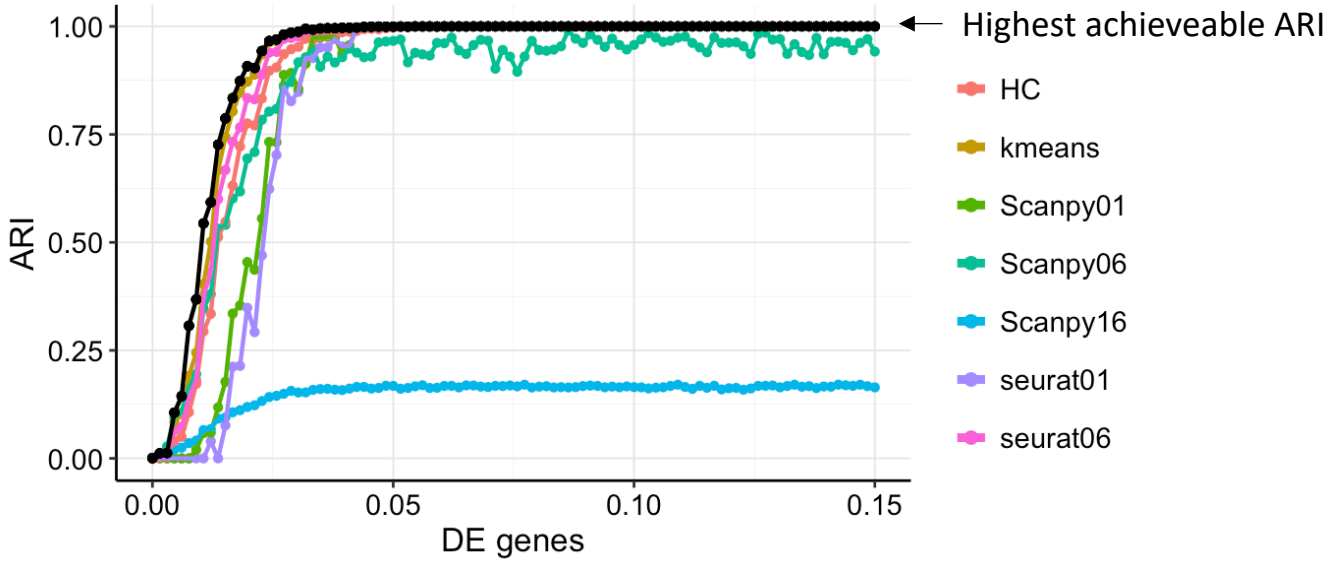
Quantitative single-cell biology

# Connection to the adjusted rand index (ARI)

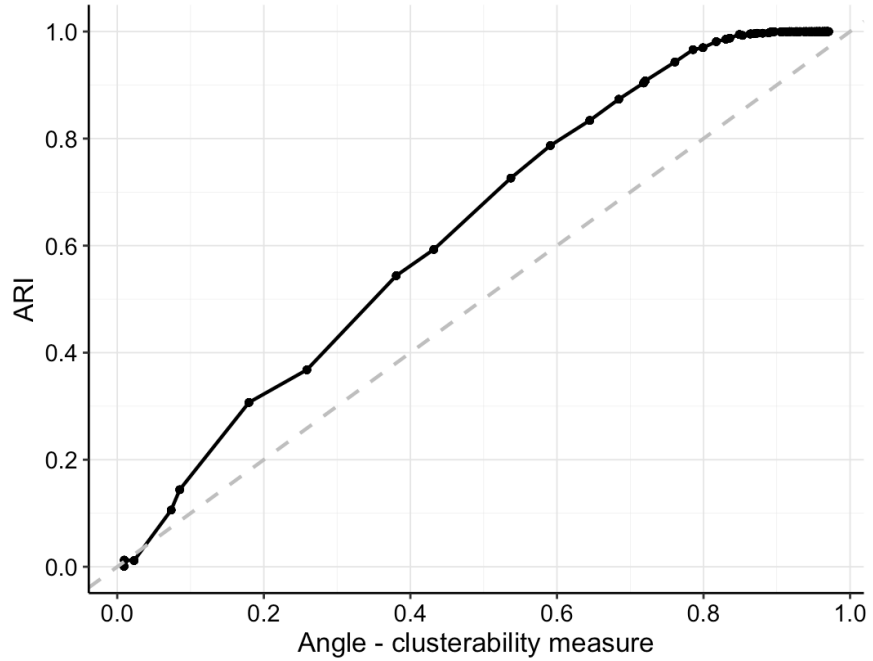
The ARI is a well known measure to determine the similarity between two clusterings. The aim is to calculate the highest achieveable ARI given a data set in order to establish an upper bound for its clusterability.




The ARI scales with the amount of variabilty and noise in the data.



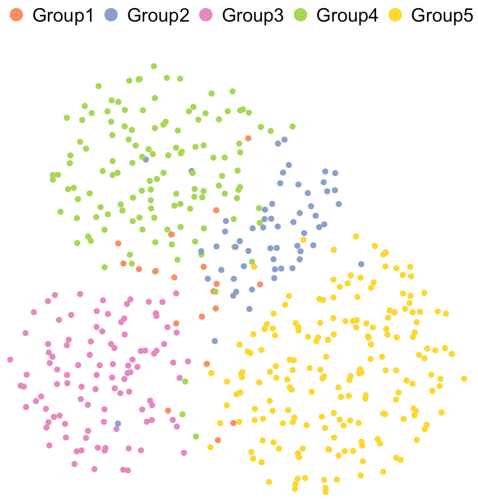

**Linear relationship between our clusterability measure and the highest achievable ARI**



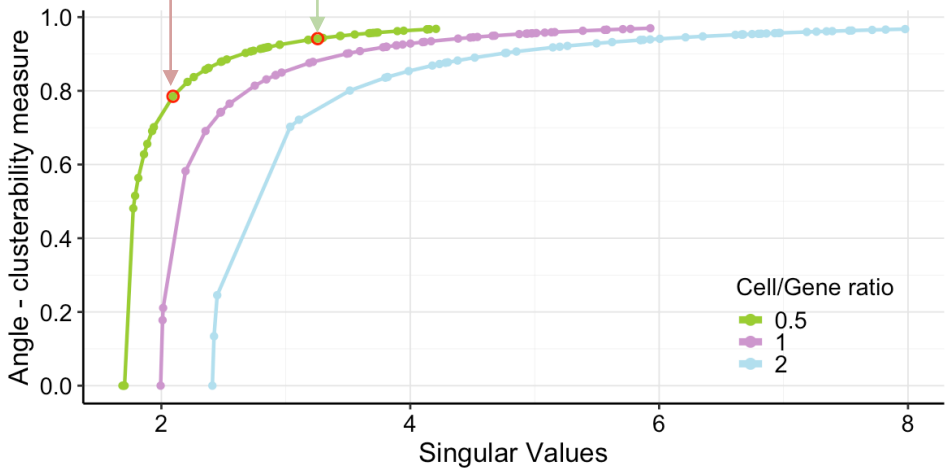

Our clusertability measure scales linearly with the achievable ARI. This indicates that it is a good approximation for the clusterability of a data set.

# Understanding the clusterability measure in simulated data

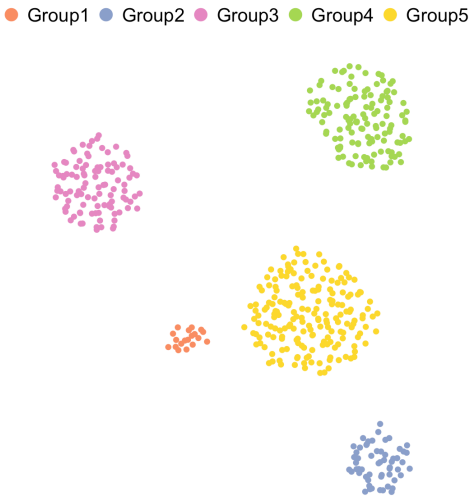
## Low signal




umap




## High signal

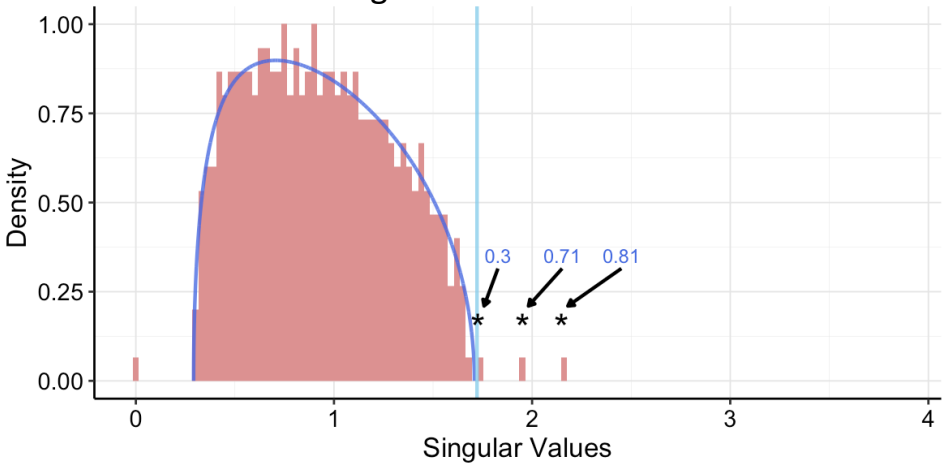



umap

The clusterability measure depends on the magnitude of the singular values as well as the cell to gene ratio.

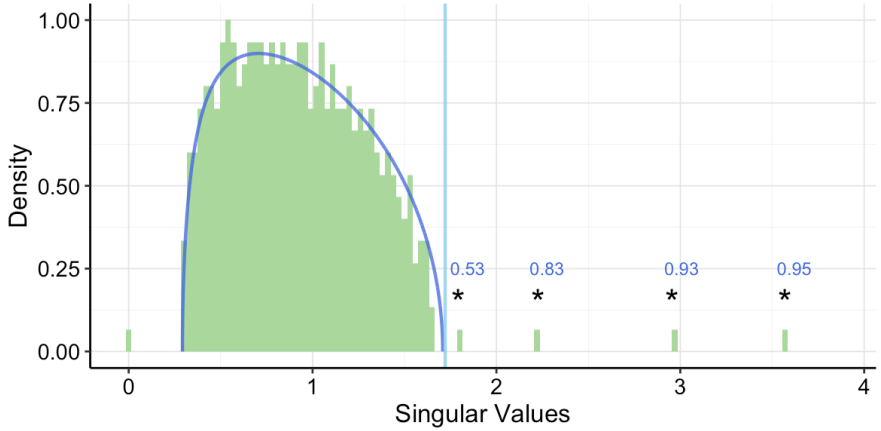
The clusterability measure scales with the amount of signal in the data. Examples show a data set with a low signal and high signal. The lower the signal, the lower are the values of the clusterability measure.

Distribution of singular values



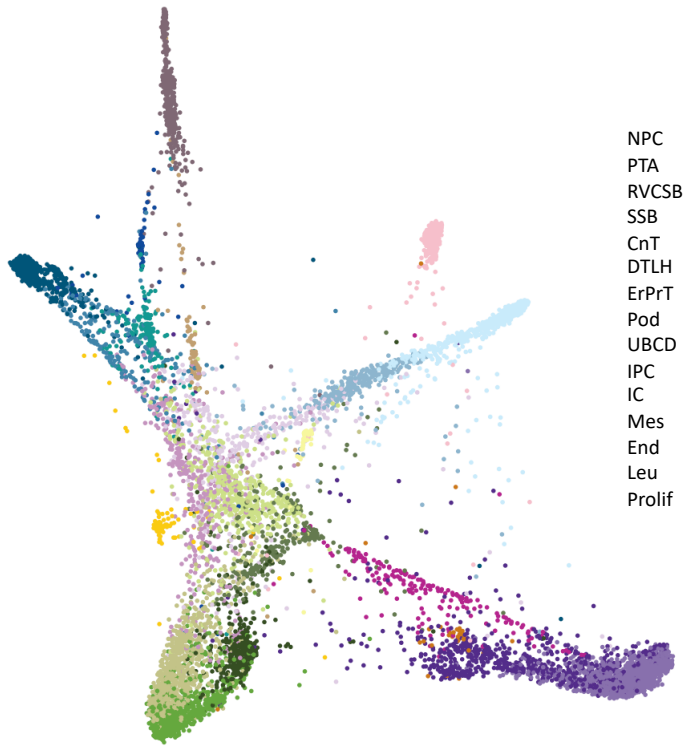

Importantly, in case of low signal data, some of the singular values can be shifted under the bulk of the MP distribution, which renders the distinction of the clusters impossible.

Distribution of singular values



# Application to human fetal kidney scRNA-seq data

We applied the clusterability measure on scRNA-seq data of human fetal kidney at week 16 of gestation in which 22 cell types have already been identified.



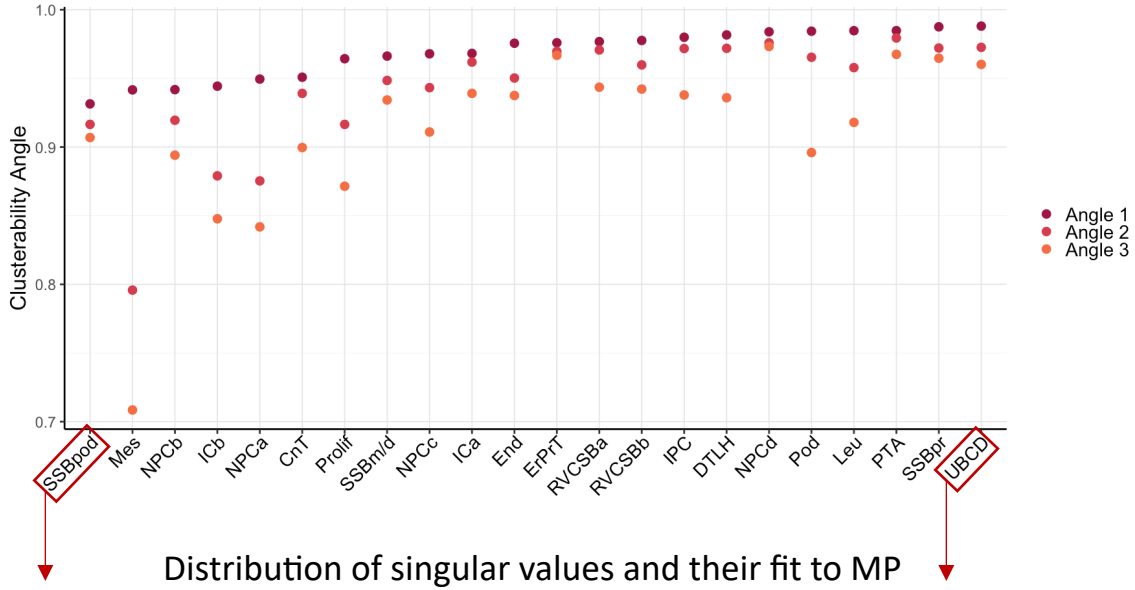
| | |
|---|---|
| NPC | Neprhon Progenitor cells |
| PTA | Pretubular aggregate |
| RVCSB | renal vesicle/comma-shaped body |
| SSB | S-shaped body |
| CnT | Connecting tubule |
| DTLH | Distal tubule/loop of Henle |
| ErPrT | Early proximal tubule |
| Pod | Podocytes |
| UBCD | Ureteric bud/collecting duct |
| IPC | Interstitial progenitor cells |
| IC | Interstitial cells |
| Mes | Mesangial cells |
| End | Endothelial cells |
| Leu | Leukocytes |
| Prolif | Proliferating cells |

paga

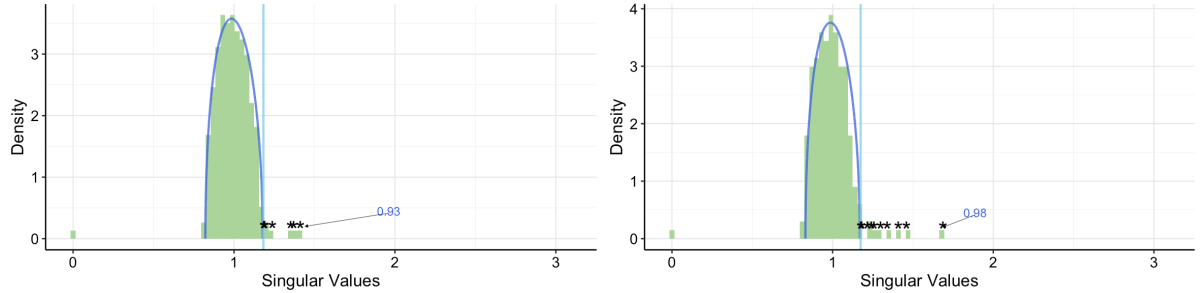

This measure is able to order the 22 cell types such that those having the highest values contain the most reliable sub-structures.

The lowest and highest measures were obtained for SSBpod and UBCD cells, respectively. This indicates UBCD has a high likelihood to contain more than one cell type or cell state, whereas SSBpod seems to consist of a purer cell population.

The values of the clusterability measure for the three highest singular values for each cluster.




Distribution of singular values and their fit to MP




umaps