UNIVERSITÄT DES SAARLANDES

Sample Solution: Project 1

Single-cell Bioinformatics

CHAIR FOR CLINICAL BIOINFORMATICS
SAARLAND UNIVERSITY

Integrative Cellular Biology & Bioinformatics

Pre-Processing

Which steps do you perform before and after merging and why?

Before:

- Filtering, Doublet-Removal, Feature Selection and Normalization
- Account for sample quality differences
- Determine thresholds and variables separately for each sample

After:

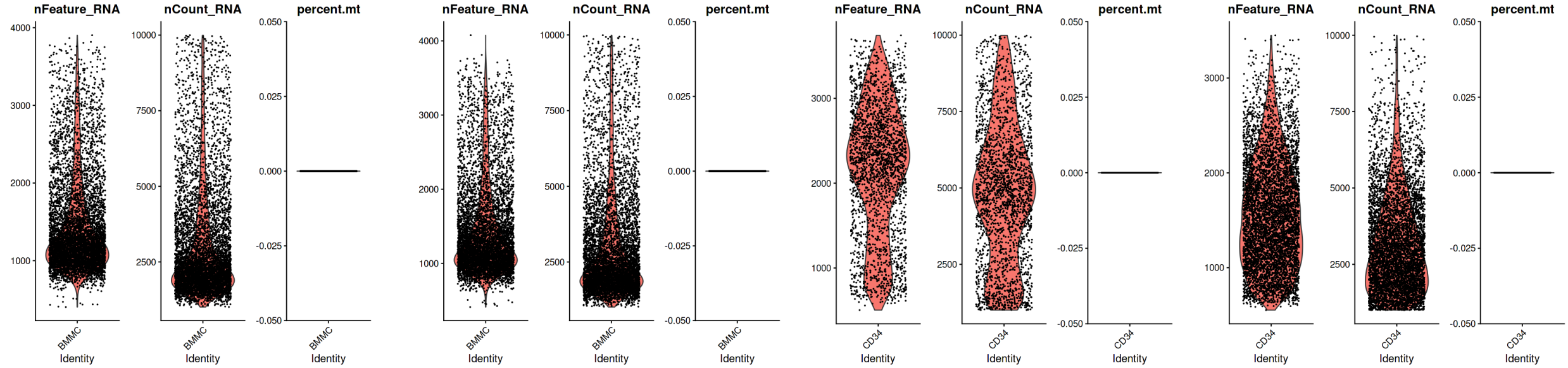
- Batch-correction (optional: Feature Selection and normalization)
- Correct between-sample differences

Pre-Processing

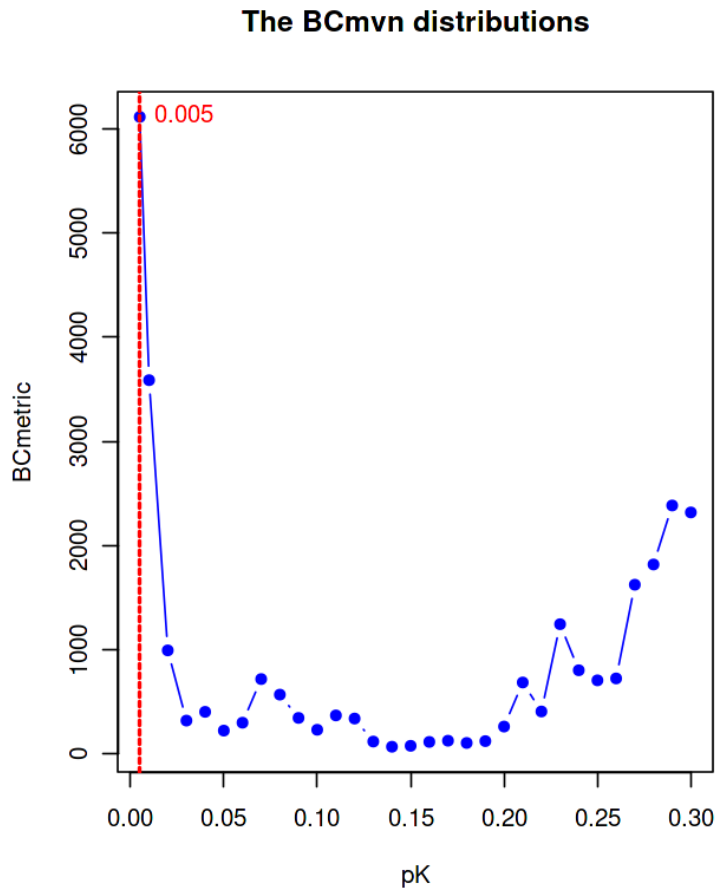
Name the parameters that have been used for filtering and argument how and why you have used them.

- Number of counts per barcode
- Number of genes found ber barcode
- Fraction o fmitochondrial genes per barcode

Filtering



Doublet-Removal (find pK)



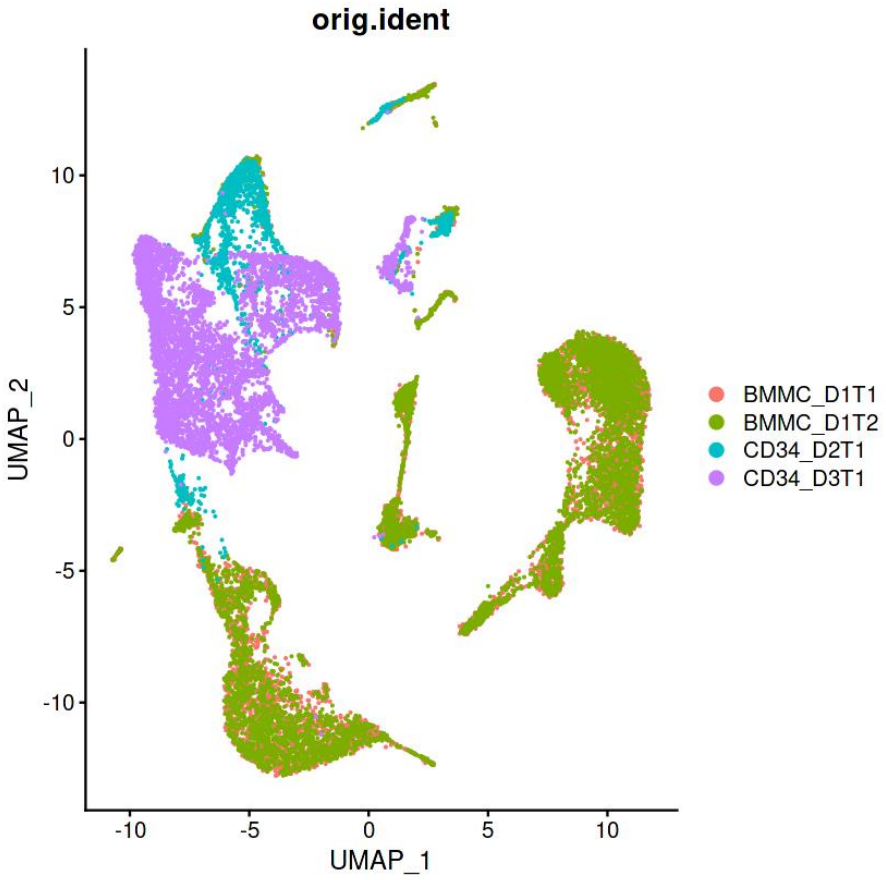
Batch-Correction

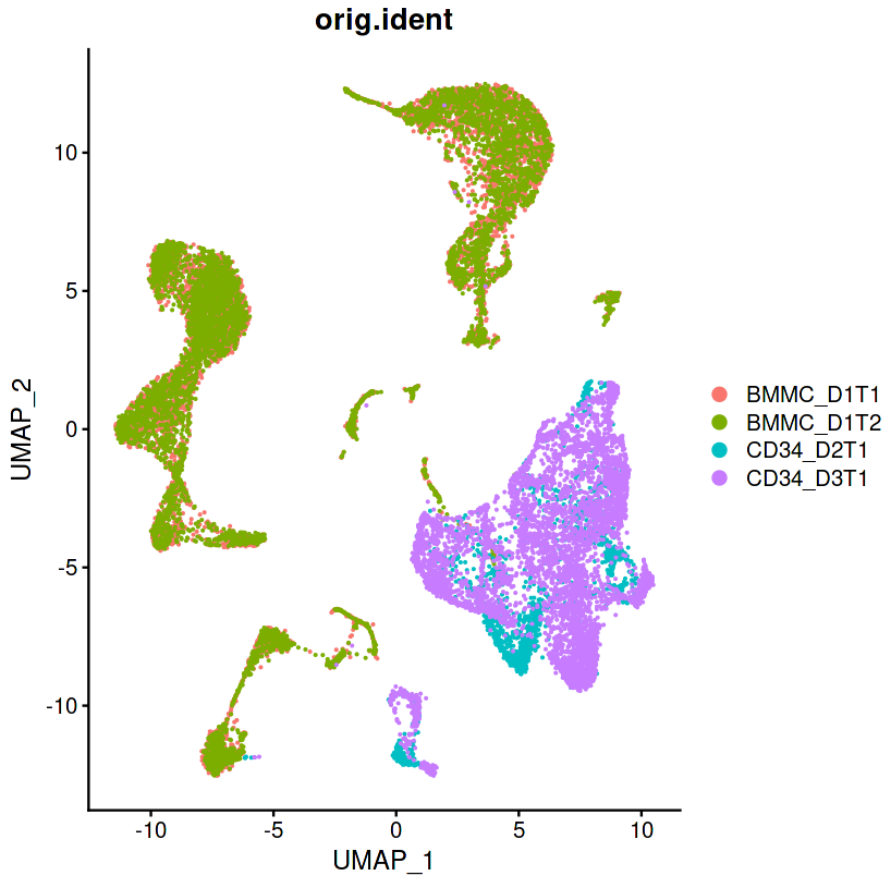
If you think a Batch-correction may be necessary, name the parameters and explain why a correction for this parameter may be necessary. Also show the plot you used to check it.

To check if you have to do batch-correction for a variable:

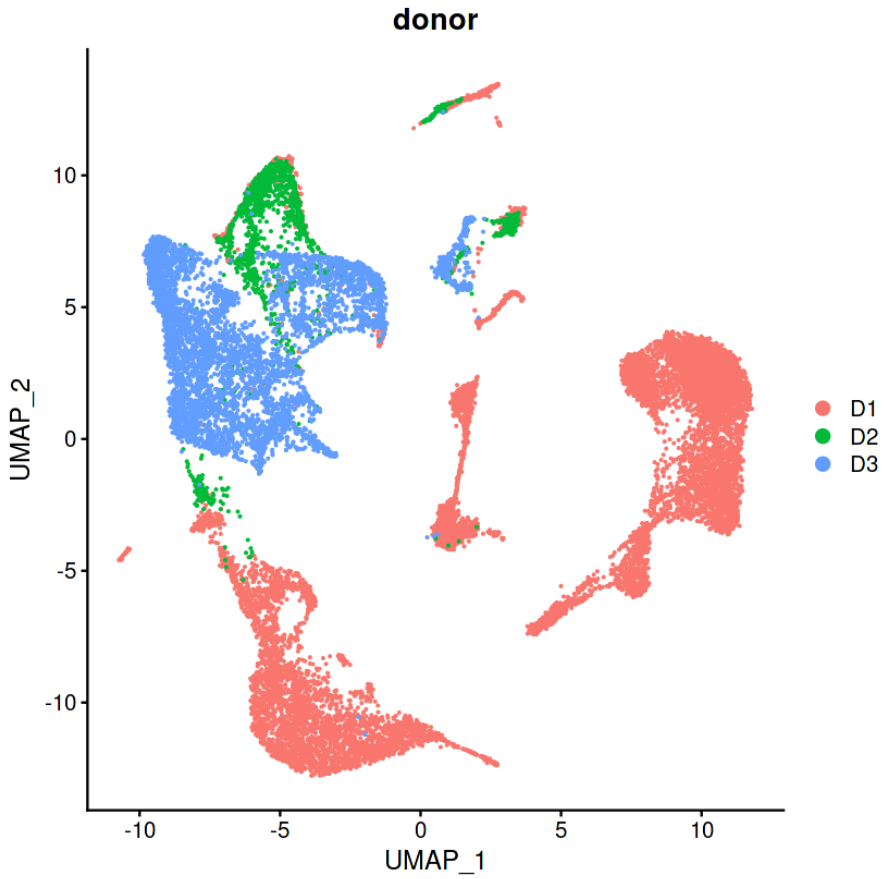
- Plot the distribution of the cells colored by this variable in the uncorrected data
- Plot the distribution of the cells colored by this variable in the data after applying batch correction
- Compare the two plots: If you see an improvement a batch-correction is necessary

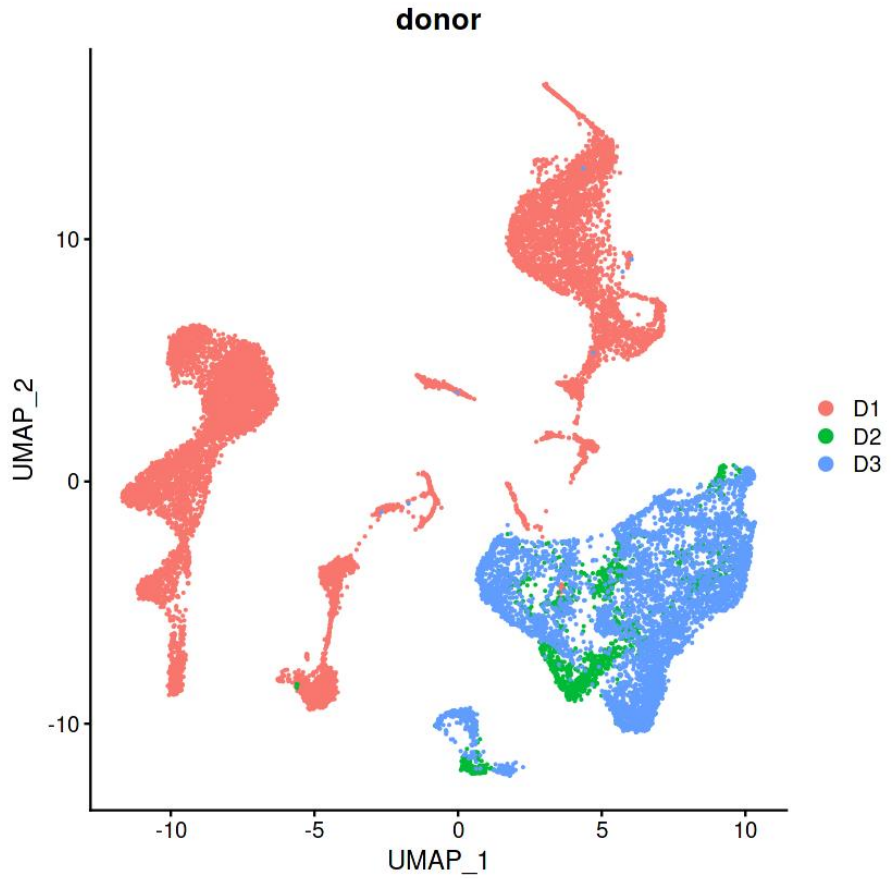
Batch-Effects (Sample)



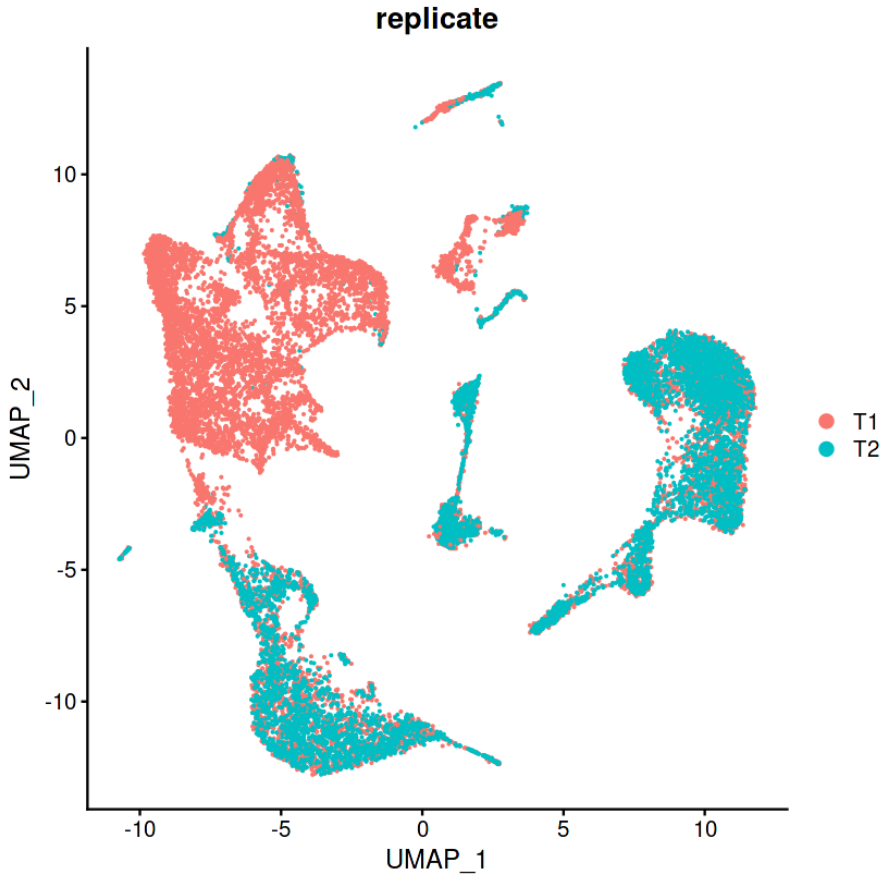


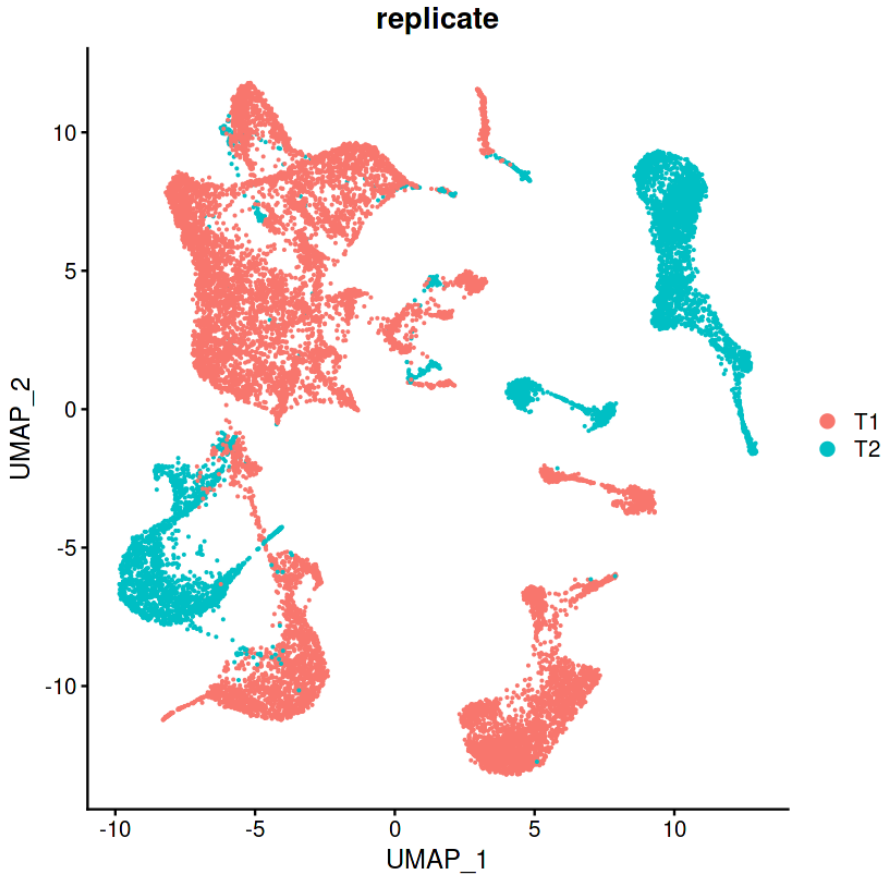
Batch-Effects (Donor)



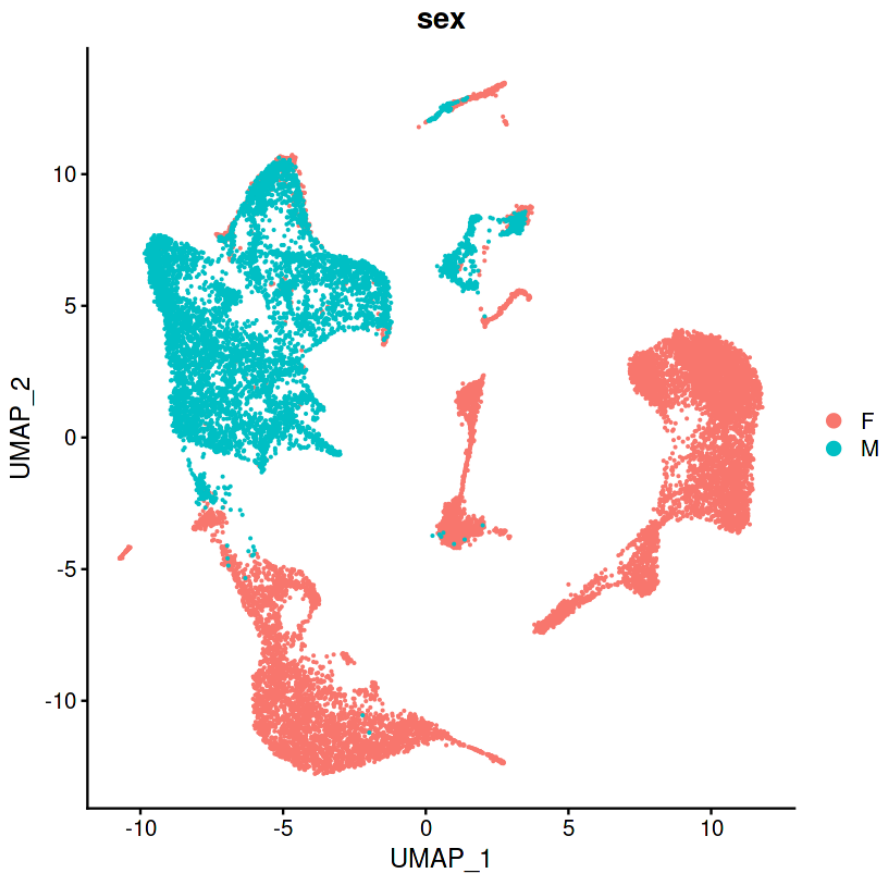


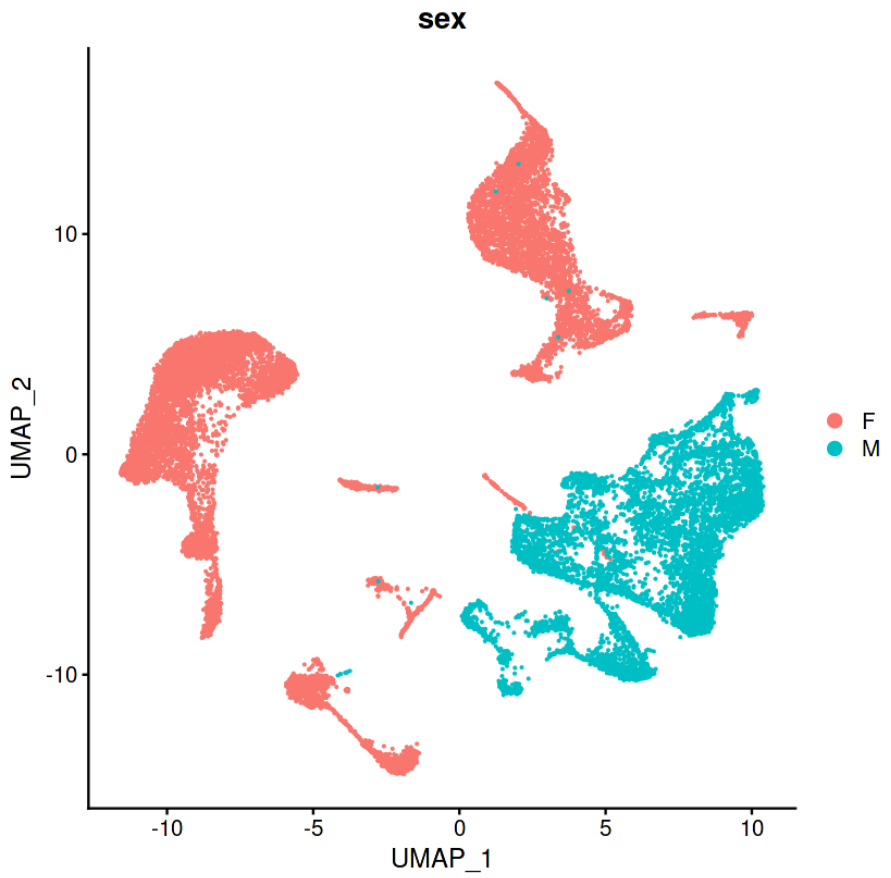
Batch-Effects (Replicate)



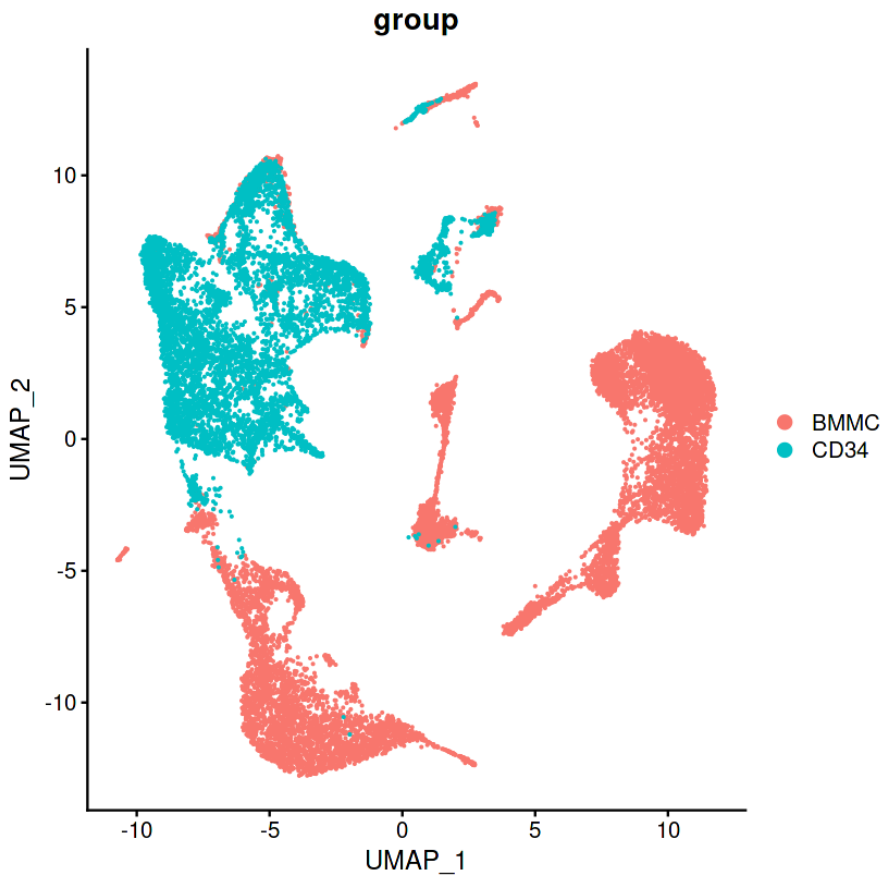


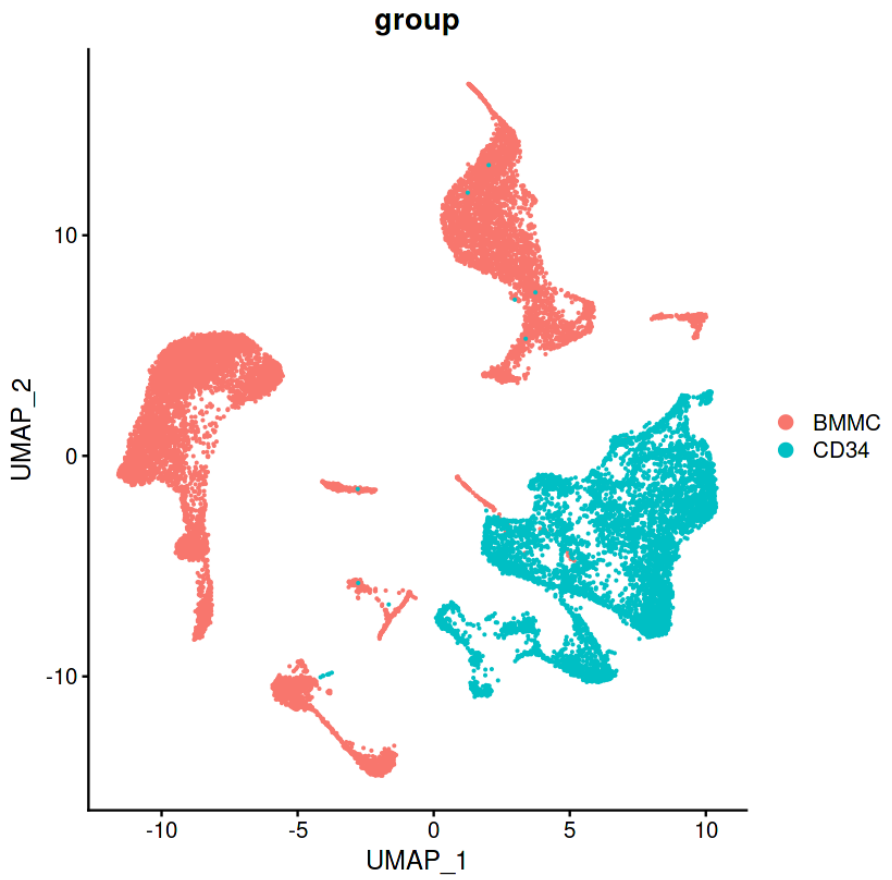
Batch-Effects (Sex)



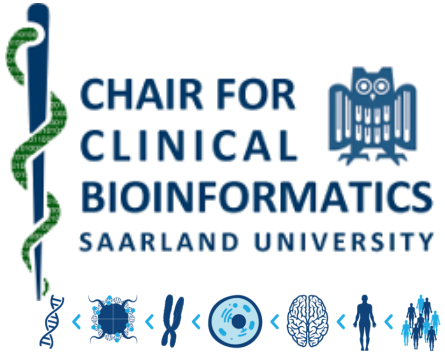


Batch-Effects (Group)

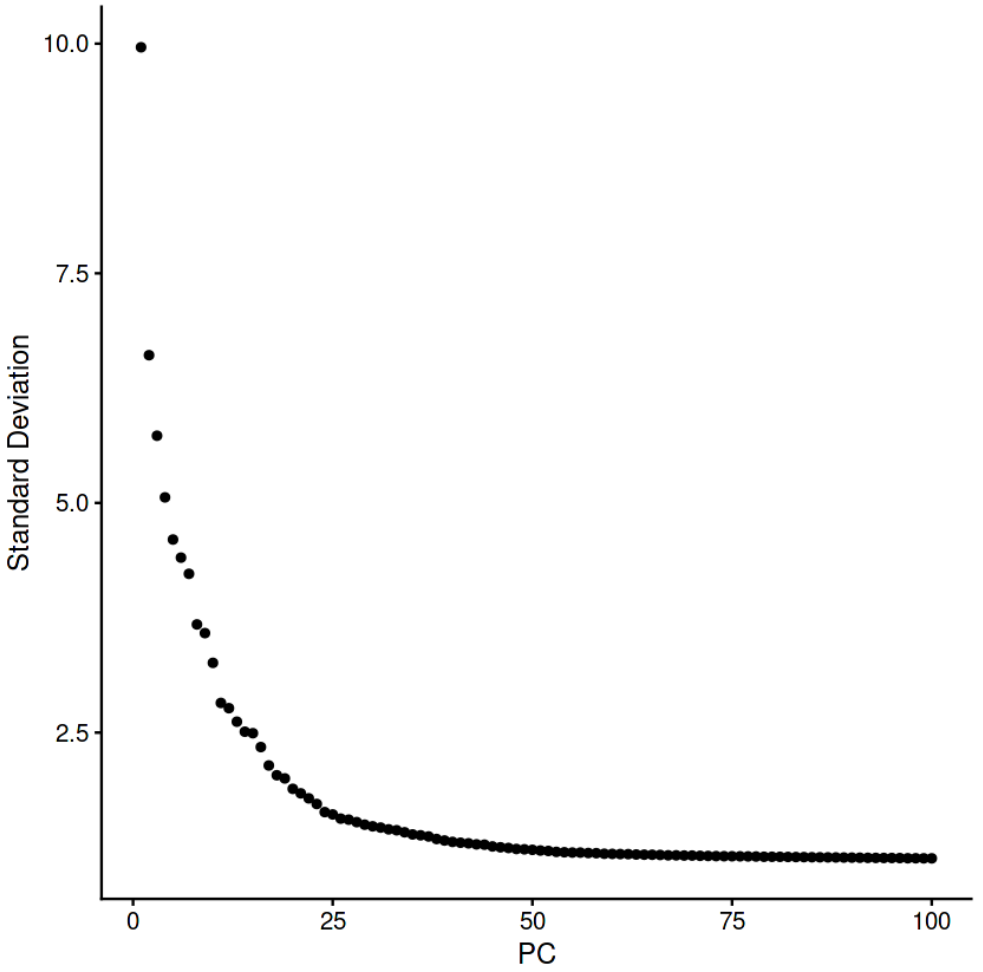




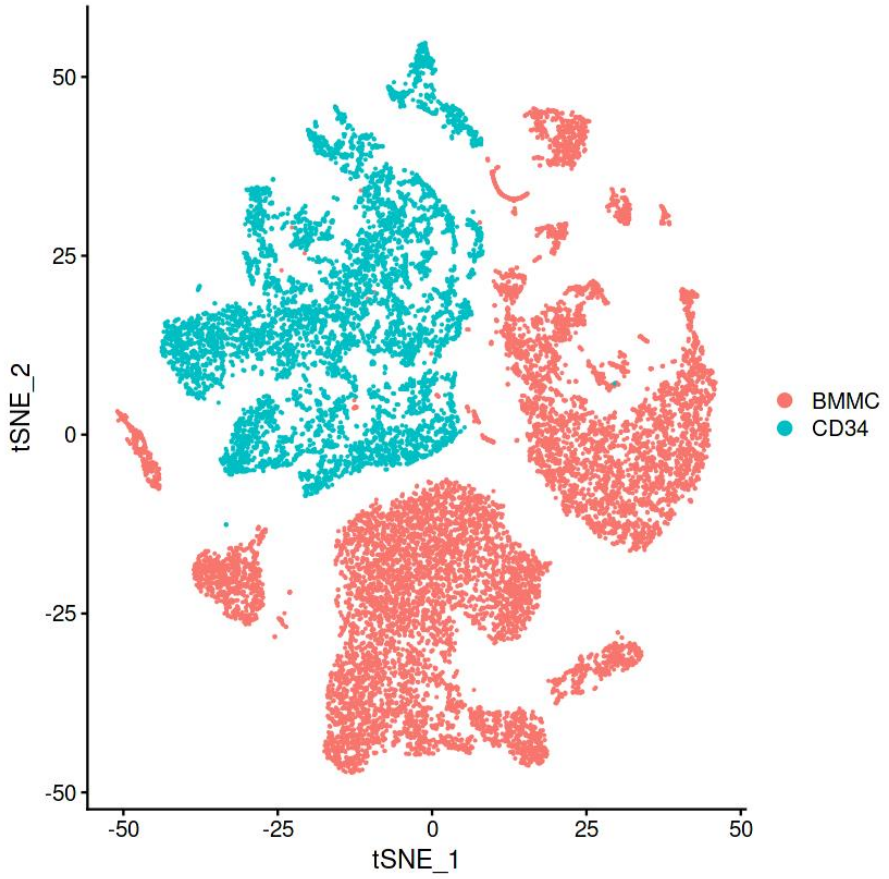
Dimensionality-Reduction and Clustering

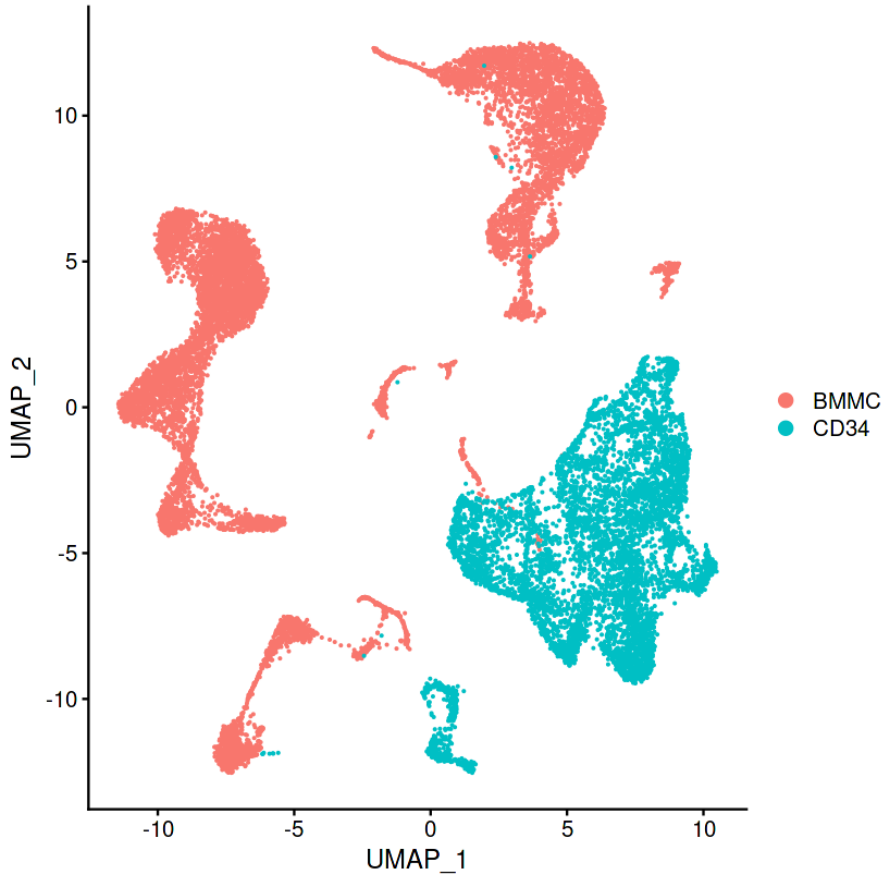


Dimensionality-Reduction



Dimensionality Reduction





Dimensionality-Reduction

Explain why we use a combination of PCA with tSNE/UMAP for clustering and not only one of the methods.

PCA:

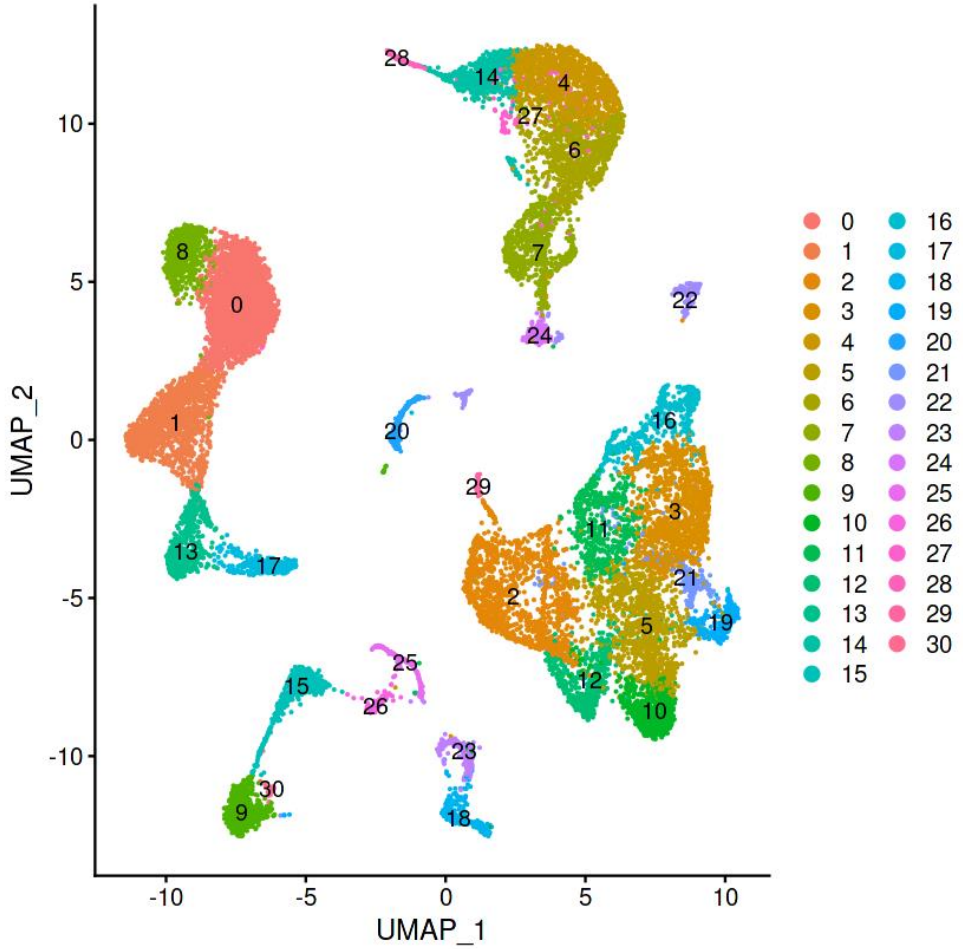
- Not as good for capturing the structure of the data in few dimensions
- Strong interpretability

UMAP and tSNA:

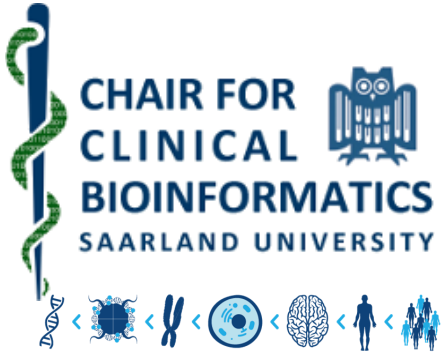
- Preserves the local (and global) structure of the data
- Not as good for interpretation

Use the lower-dimensional representation from PCA for further analysis and the 2-dimensional representation from UMAP or tSNE for visualization

Clustering



Cell-type Annotation



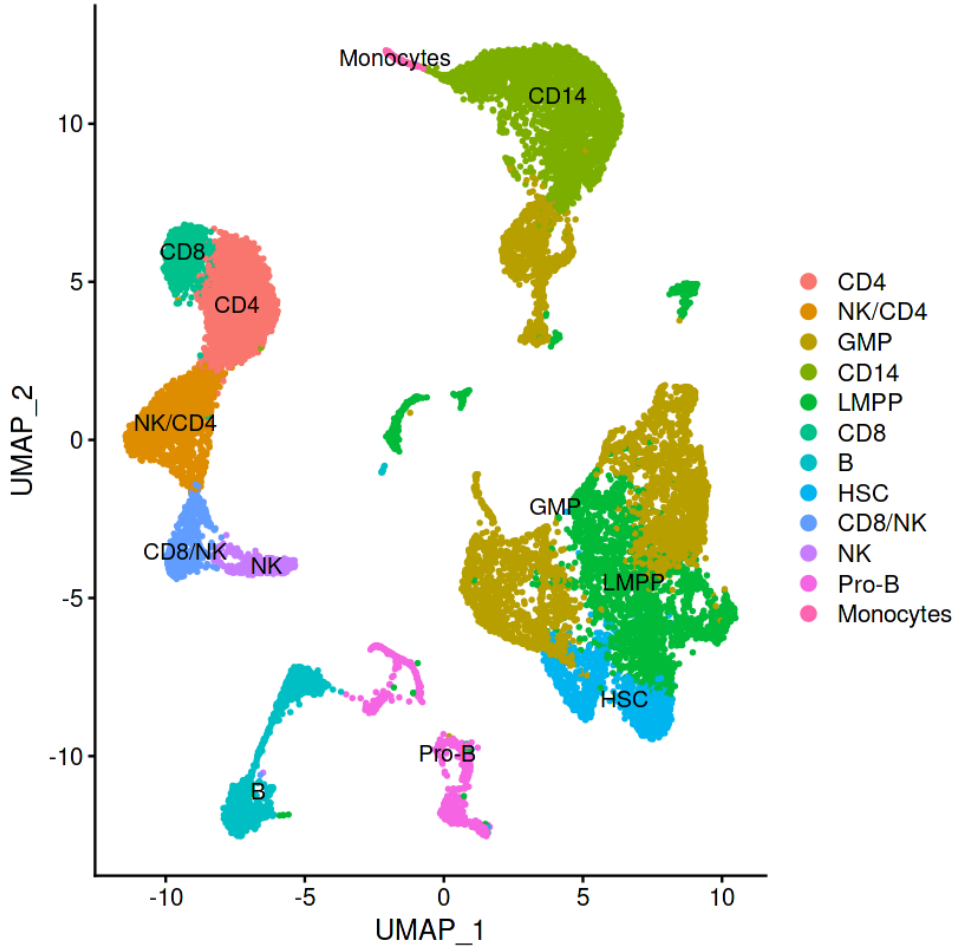
Differentially expressed genes

p_val <dbl>	avg_log2FC <dbl>	pct.1 <dbl>	pct.2 <dbl>	p_val_adj <dbl>	cluster <fct>	gene <chr>
0.000000e+00	2.212490	0.956	0.200	0.000000e+00	0	CD3E
0.000000e+00	2.076157	0.983	0.391	0.000000e+00	0	KLF2
0.000000e+00	2.719142	0.908	0.187	0.000000e+00	1	IL7R
0.000000e+00	2.572464	0.954	0.184	0.000000e+00	1	IL32
0.000000e+00	2.827043	0.588	0.023	0.000000e+00	2	HBD
1.723765e-138	4.675747	0.128	0.017	3.497002e-134	2	CLC
0.000000e+00	2.866129	0.888	0.202	0.000000e+00	3	MPO
0.000000e+00	2.327365	0.540	0.084	0.000000e+00	3	ELANE
0.000000e+00	3.739681	0.994	0.109	0.000000e+00	4	S100A12
0.000000e+00	3.683718	1.000	0.182	0.000000e+00	4	S100A9
0.000000e+00	2.670823	0.998	0.230	0.000000e+00	5	SPINK2
0.000000e+00	1.762673	0.698	0.118	0.000000e+00	5	C1QTNF4
0.000000e+00	3.356395	1.000	0.182	0.000000e+00	6	S100A8
0.000000e+00	3.202560	0.999	0.191	0.000000e+00	6	S100A9
0.000000e+00	3.636302	0.881	0.093	0.000000e+00	7	RETN
0.000000e+00	3.065553	1.000	0.366	0.000000e+00	7	LYZ

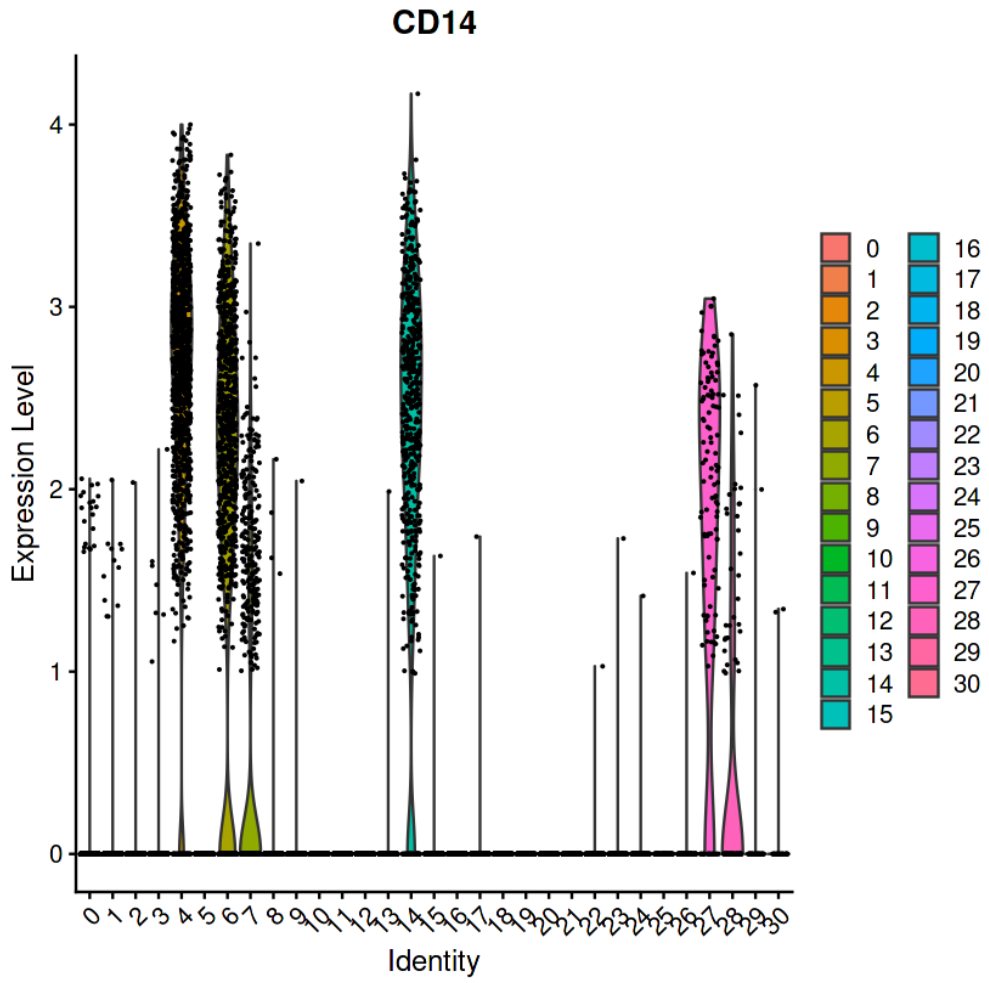
Marker-genes from Database

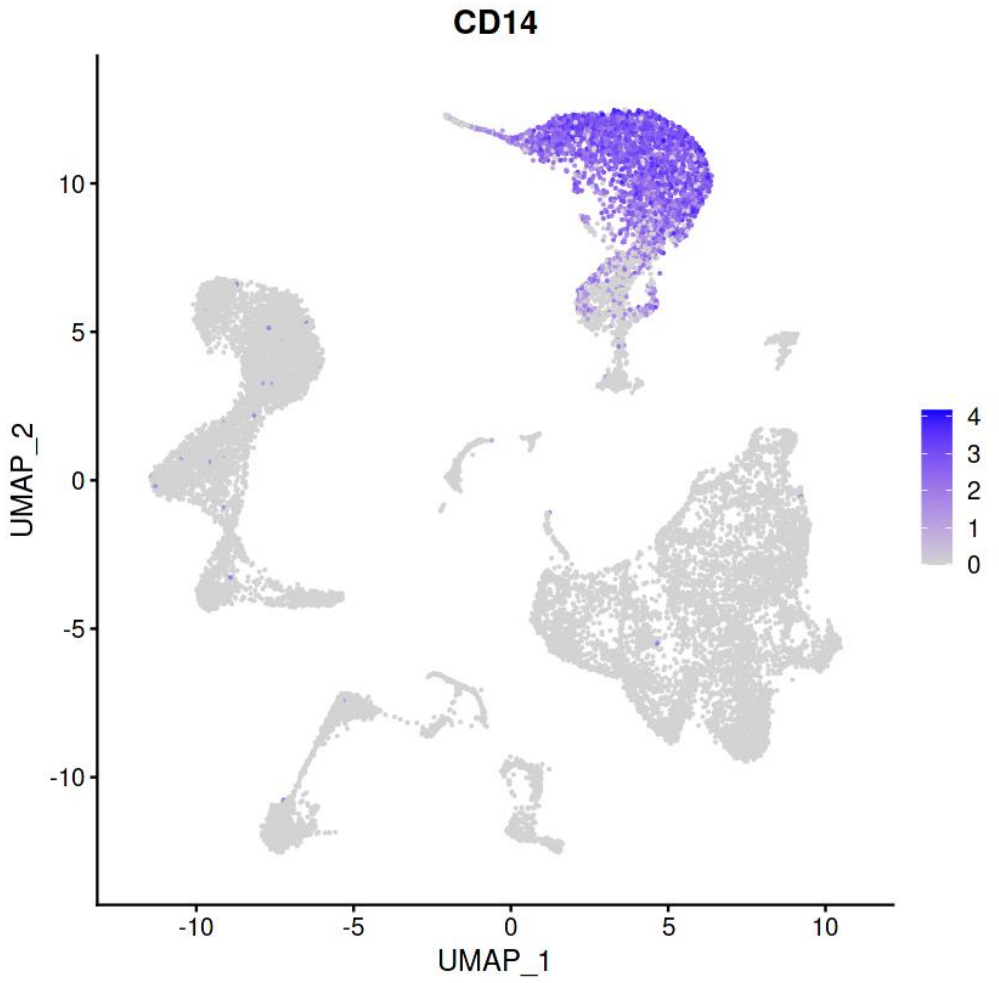
```
features_HSC <- c("CD34", "CD38")
features_LMPP <- c("CD38", "CD52", "CSF3R")
features_GMP <- c("ANXA1", "AP3S1", "APLP2")
features_CD14_Mono <- c("CD14")
features_CD16_Mono <- c("CD68", "S100A12")
features_B <- c("CD19", "CD38")
features_CD8N <- c("CD8A", "CD8B", "CD3D")
features_CD4N <- c("CD4", "CD3D")
features_NK <- c("NKG7", "KLRB1", "CD3D")
features_GMPN <- c("CD14", "APLP2")
features_Mono <- c("CX3CR1")
features_ProB <- c("CD19", "CD34")
```

Manual Annotation

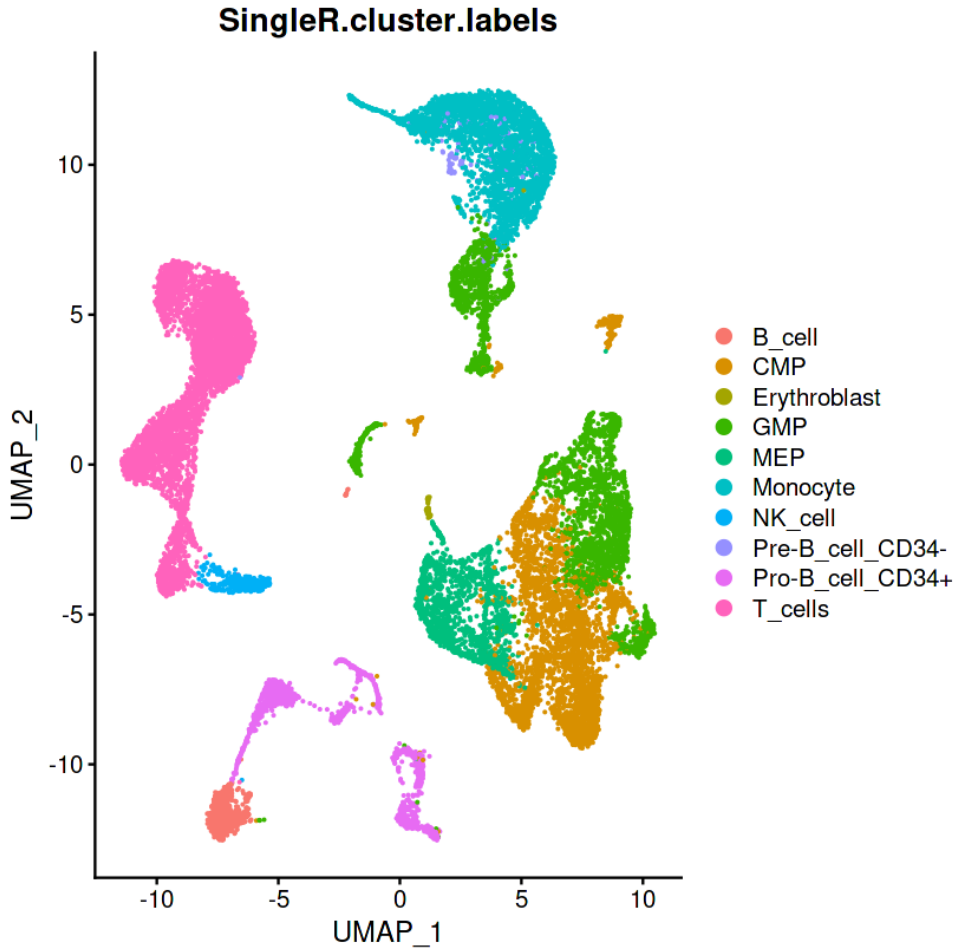


Marker-Genes

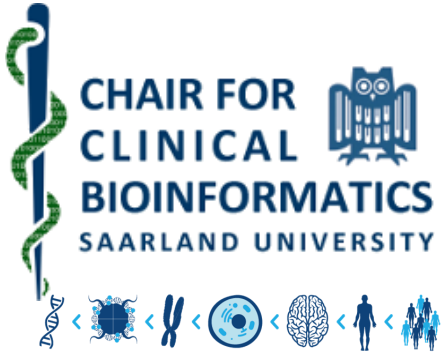




Automatic Annotation



Trajectory-Analysis



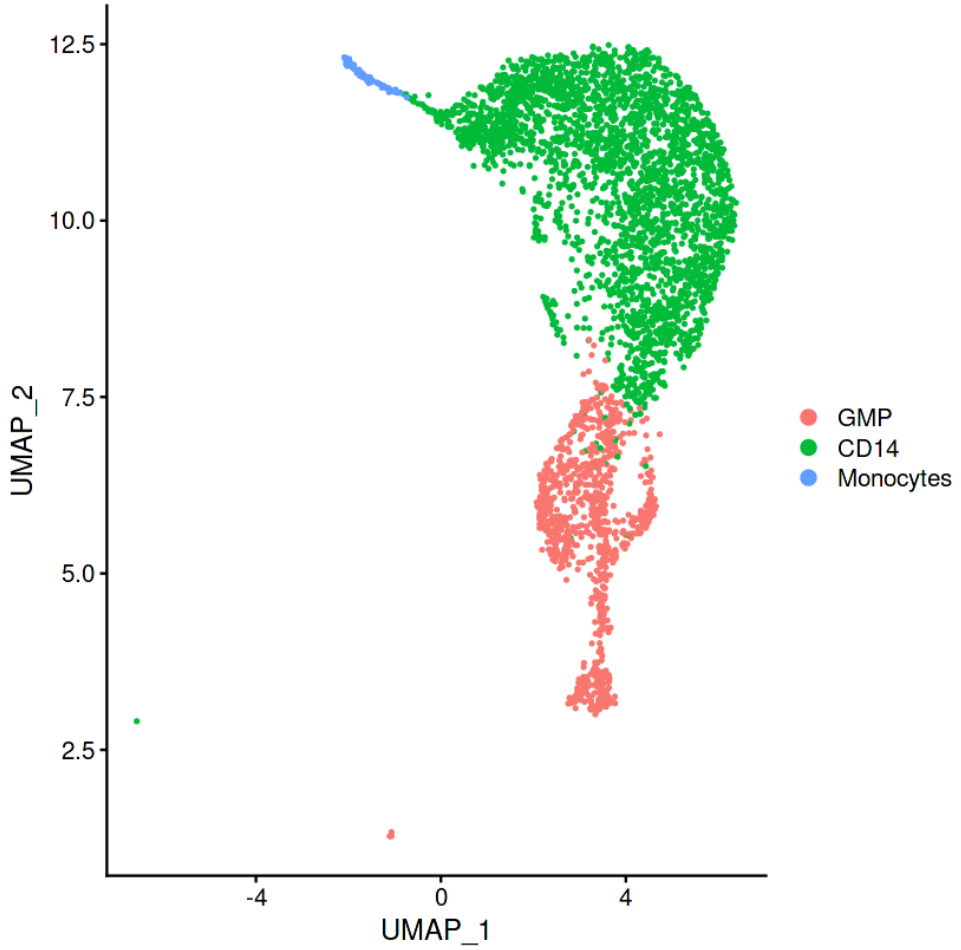
Integrative Cellular Biology
& Bioinformatics

Trajectory-Analysis

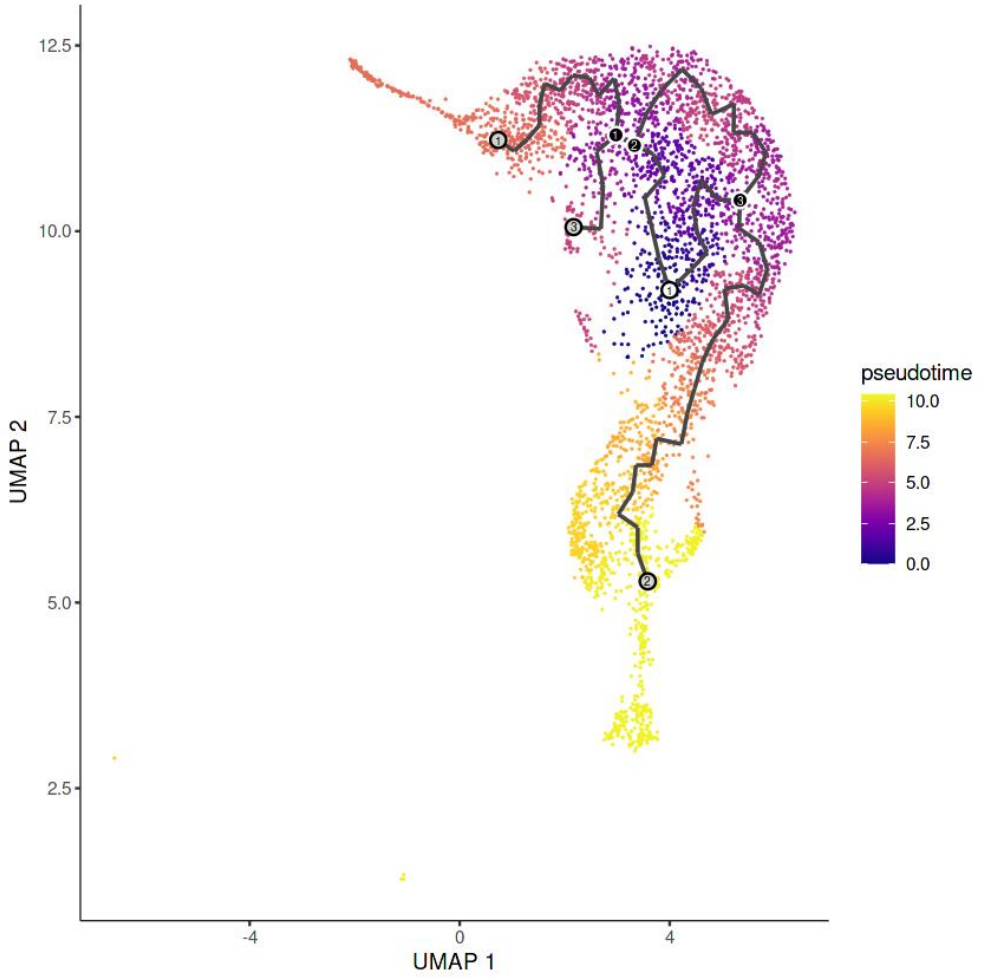
Why is this a good group to do trajectory analysis. Which other group do you think may be a good choice.

In general trajectory analysis is specially interesting for groups of cells that are distributed in the 2-dimensional space in a way that suggests a development from one cell type to another one.

Trajectory-Analysis (Subset)



Trajectory-Analysis (Manual Selection)



Trajectory-Analysis

Why is the selection of the root-nodes important for the algorithm?

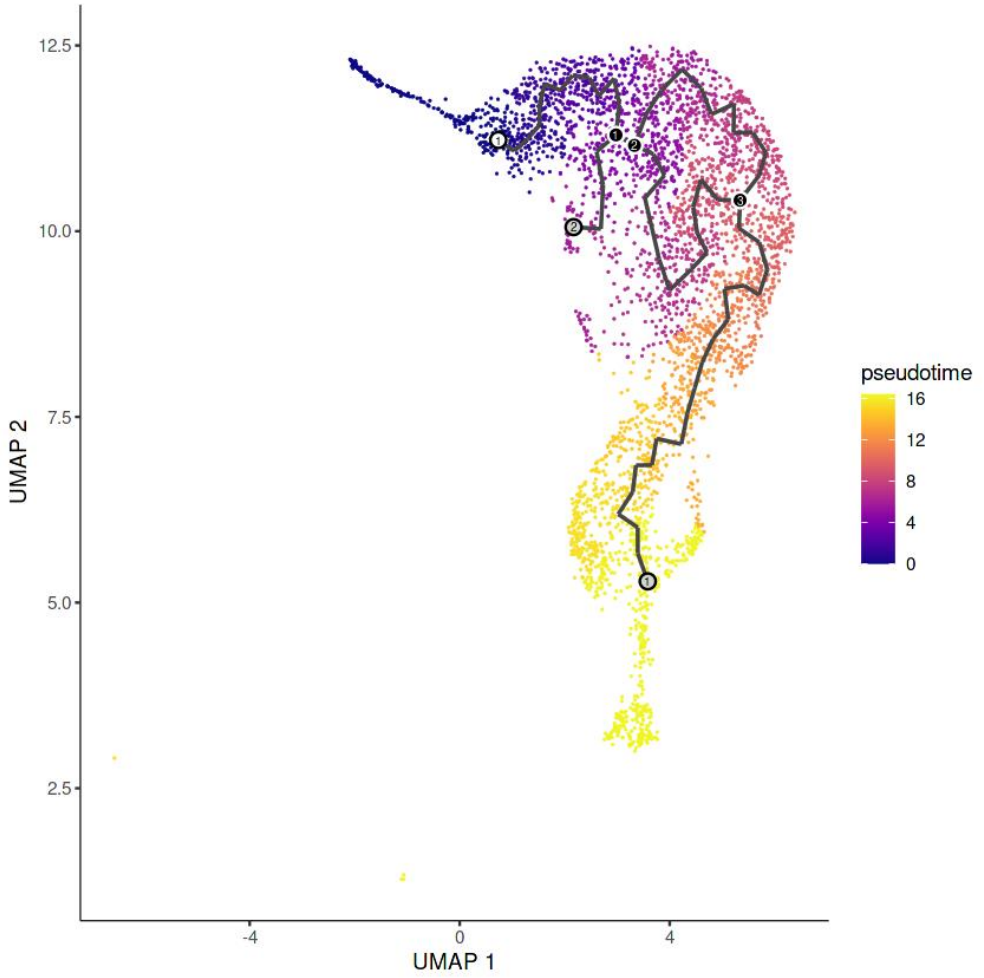
"In order to place the cells in order, we need to tell Monocle where the "beginning" of the biological process is. We do so by choosing regions of the graph that we mark as "roots" of the trajectory. In time series experiments, this can usually be accomplished by finding spots in the UMAP space that are occupied by cells from early time points“ (https://cole-trapnell-lab.github.io/monocle3/docs/introduction/, 26.11.2021)

Trajectory-Analysis

Which points are a good choice for root nodes of the analysis and why?

Find cells that represent early time points.

Trajectory-Analysis (Automatic Selection)



Trajectory-Analysis

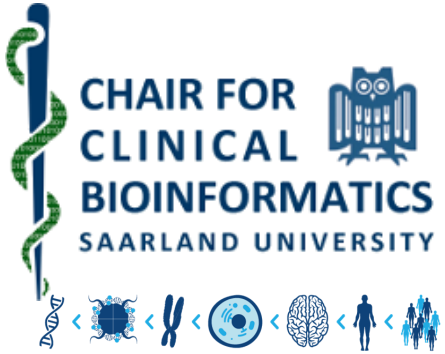
Did it improve the results? Explain why.

The automatic annotation may improve the result, especially when using this method without background knowledge about the cell population.

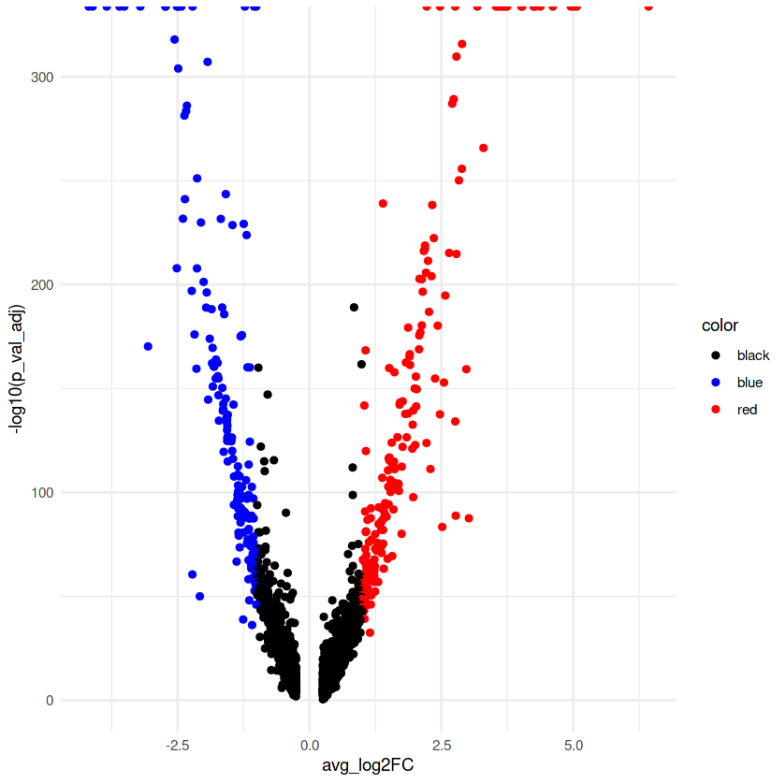
Choose one path in the trajectory and explain which cells are located on this path

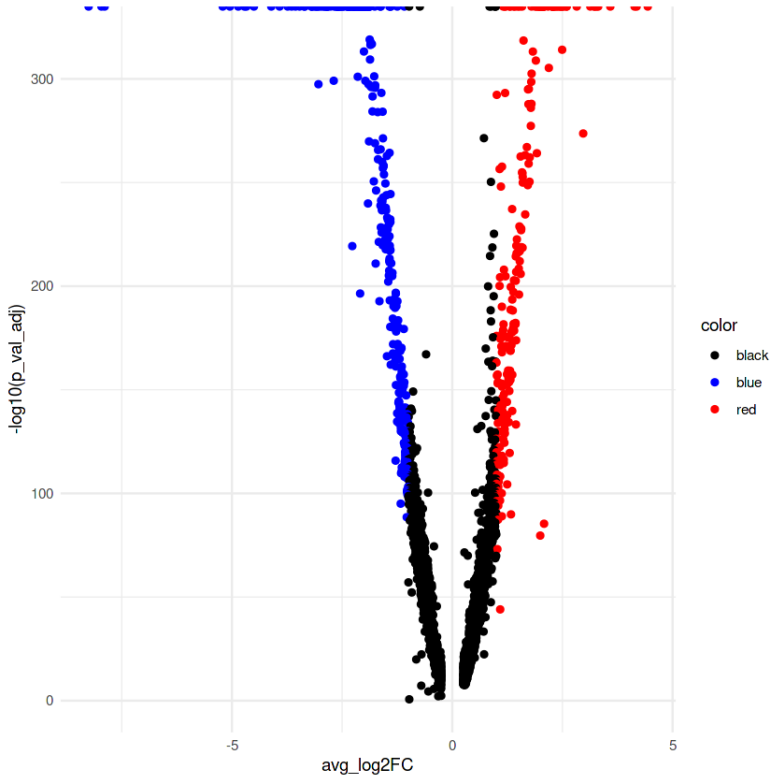
In this case the cells are ordered according to the pseudo-time during the diferentiation of cells into Monocytes and other cell-types.

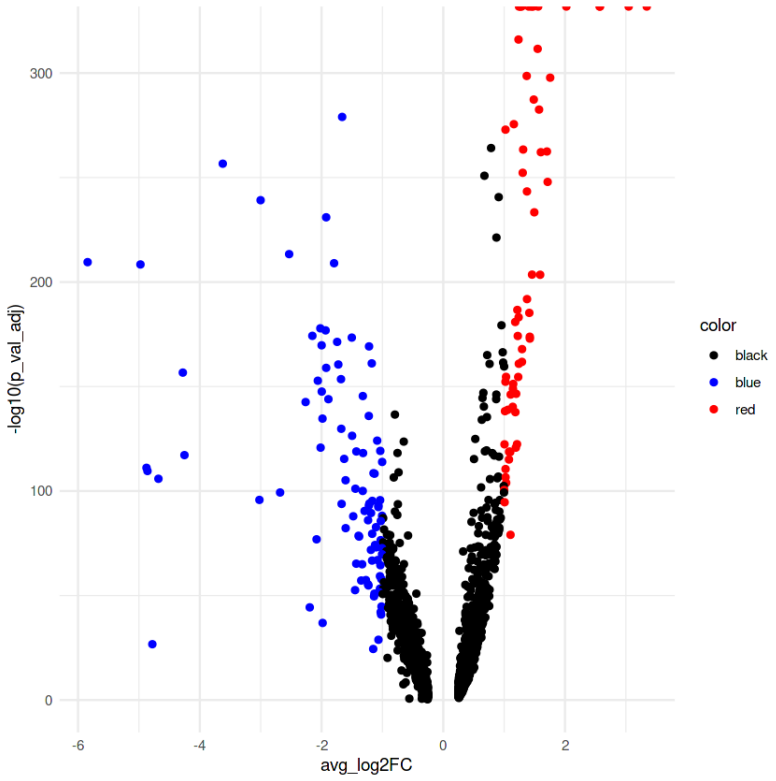
Differential Expression Analysis



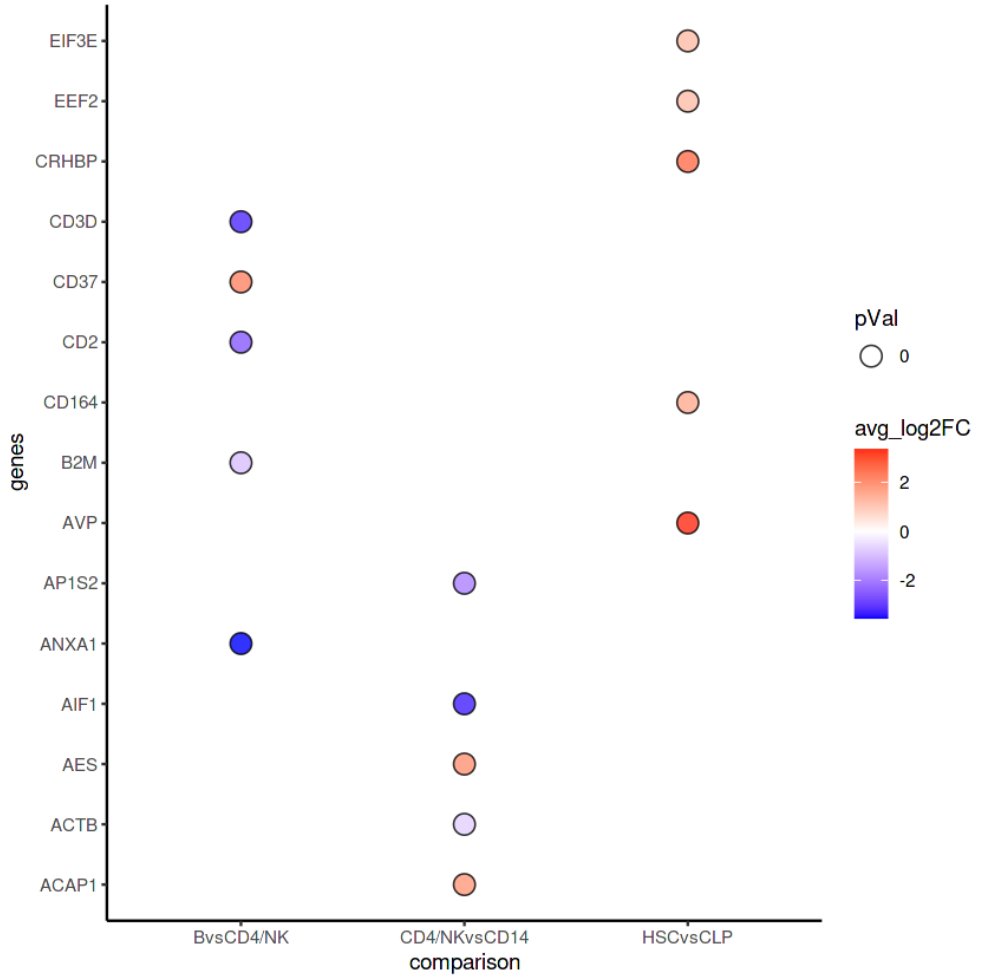
Differential Expression Analysis



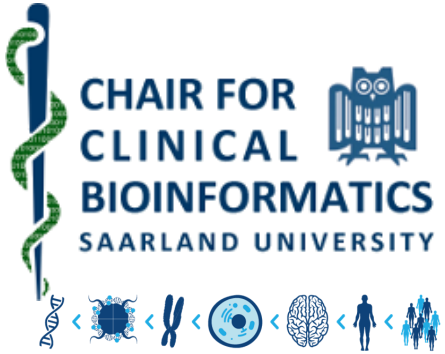




Differential Expression Analysis



Pathway Analysis

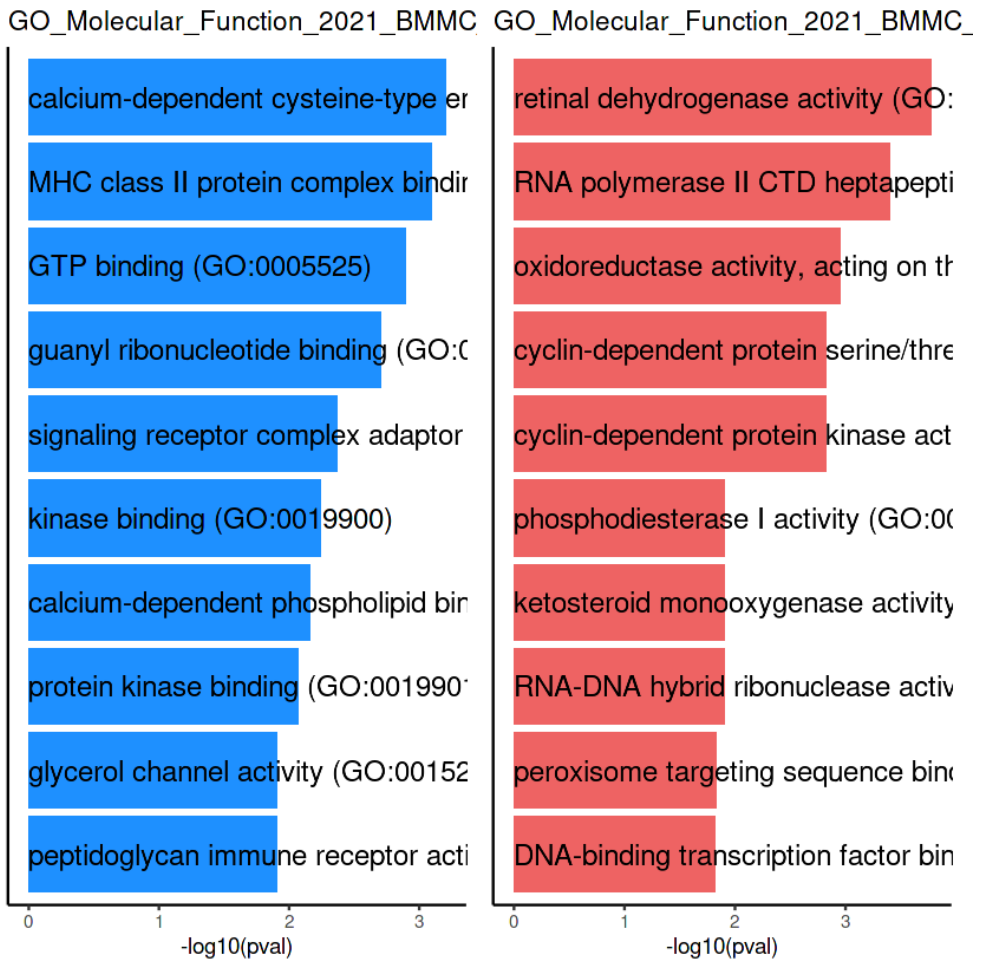


Integrative Cellular Biology
& Bioinformatics

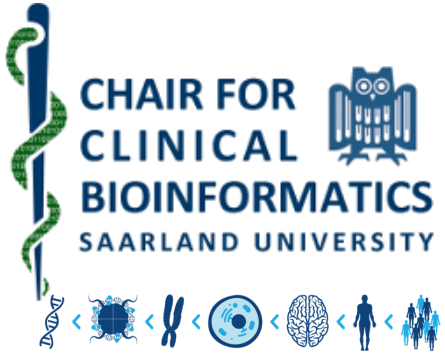
Pathway Analysis

	p_val <dbl>	avg_log2FC <dbl>	pct.1 <dbl>	pct.2 <dbl>	p_val_adj <dbl>
ABHD17A	0	1.2328896	0.379	0.110	0
ABTB1	0	0.9663742	0.254	0.046	0
AC002454.1	0	-1.1977517	0.008	0.265	0
ACAP1	0	1.0258501	0.461	0.260	0
ACOT13	0	-0.7829006	0.048	0.275	0
ACTB	0	2.5459642	0.999	0.960	0
ADA	0	-1.0745557	0.088	0.350	0
AES	0	2.3429315	0.781	0.265	0
AHCY	0	-0.7779489	0.108	0.353	0
AHI1	0	-0.7192482	0.028	0.210	0
AIF1	0	-0.8293217	0.415	0.772	0
AKAP9	0	-0.9242870	0.226	0.523	0
AKR1C3	0	-1.2435928	0.034	0.354	0
ALDH1A1	0	-0.9357699	0.049	0.254	0
ALDOA	0	1.0801005	0.820	0.587	0
ALOX5AP	0	1.0419015	0.393	0.115	0

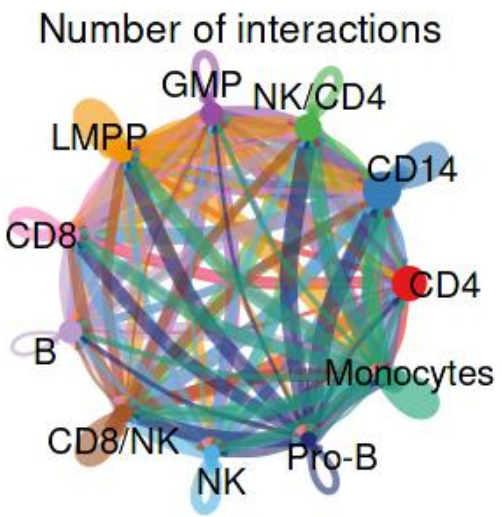
Pathway Analysis

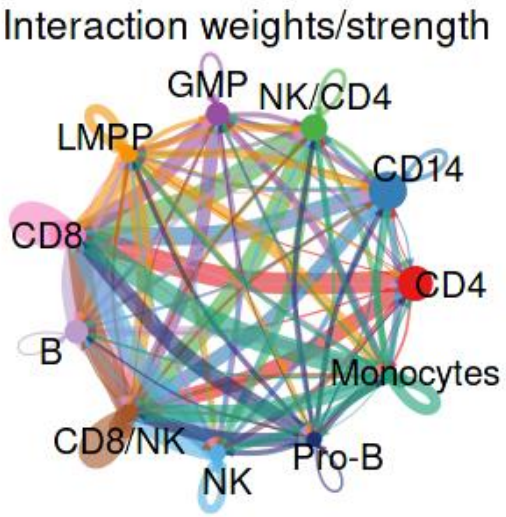


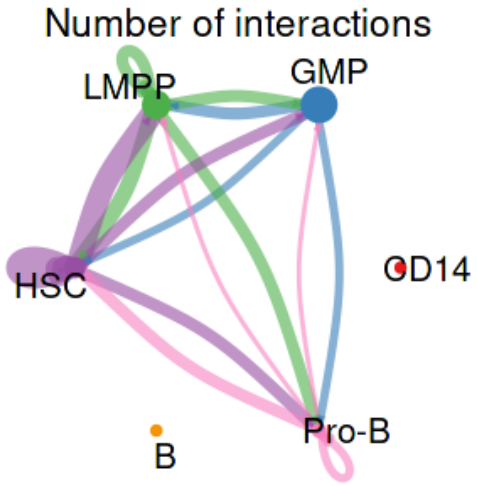
Cell-cell Communication

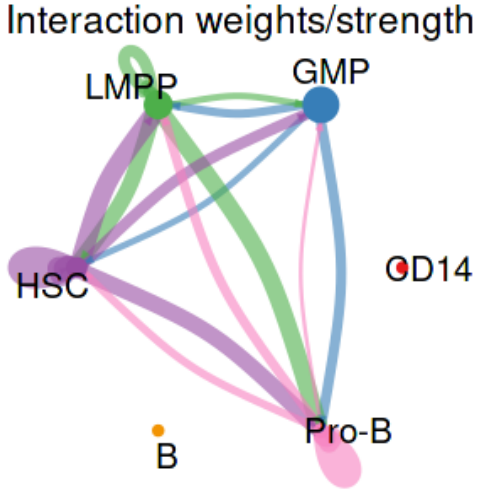


Cell-cell Communication









Cell-cell Communication