UNIVERSITÄT DES SAARLANDES

Sample-Solution: Project 2

Single-cell Bioinformatics

Integrative Cellular Biology
& Bioinformatics

Collect all samples into one data structure

How many cells does your project include?

11554

What is the median TSS-value and the median of the number of fragments?

Median of the number of fragments: 11377

What are the dimensions of your dataset?

23127 x 11554

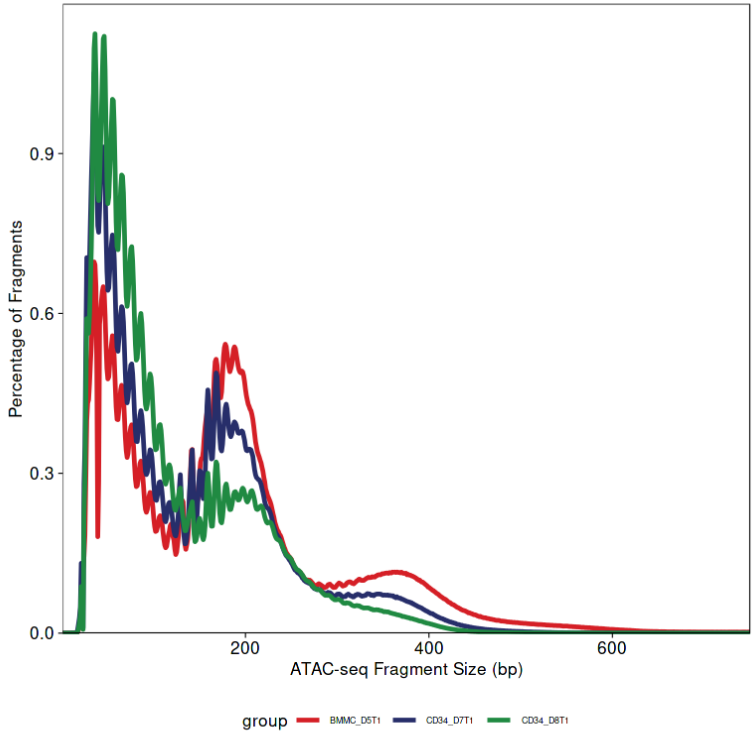
Quality control

How many cells do you have for each sample? Plot the fragment length distribution of all samples in a single plot.

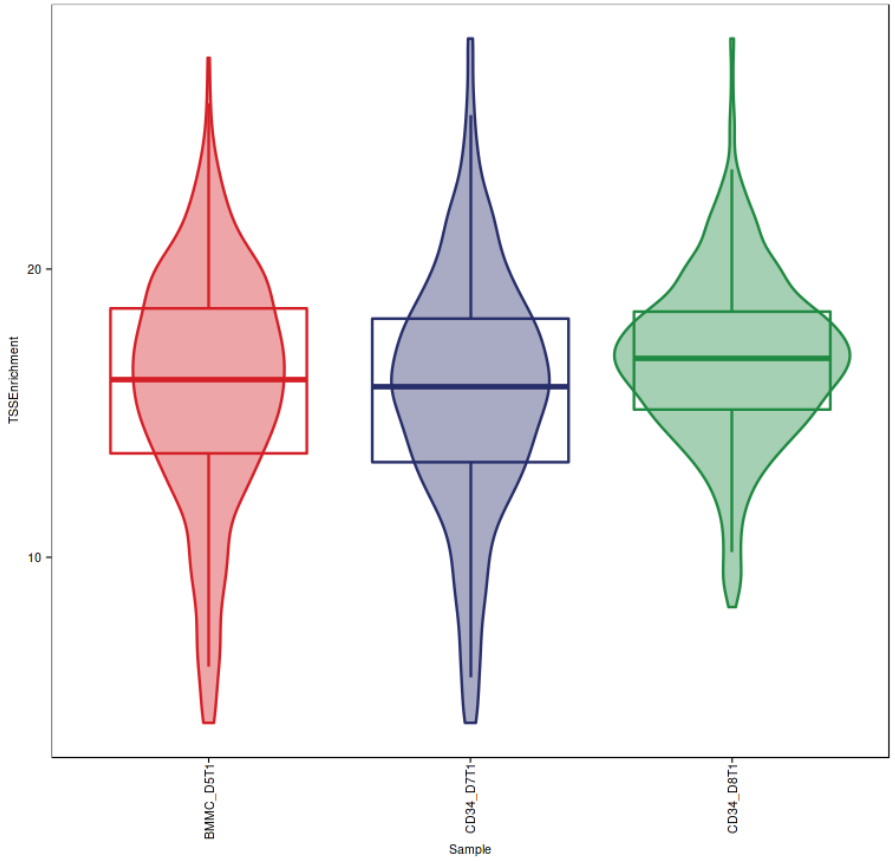
BMMC_D5T1: 4487

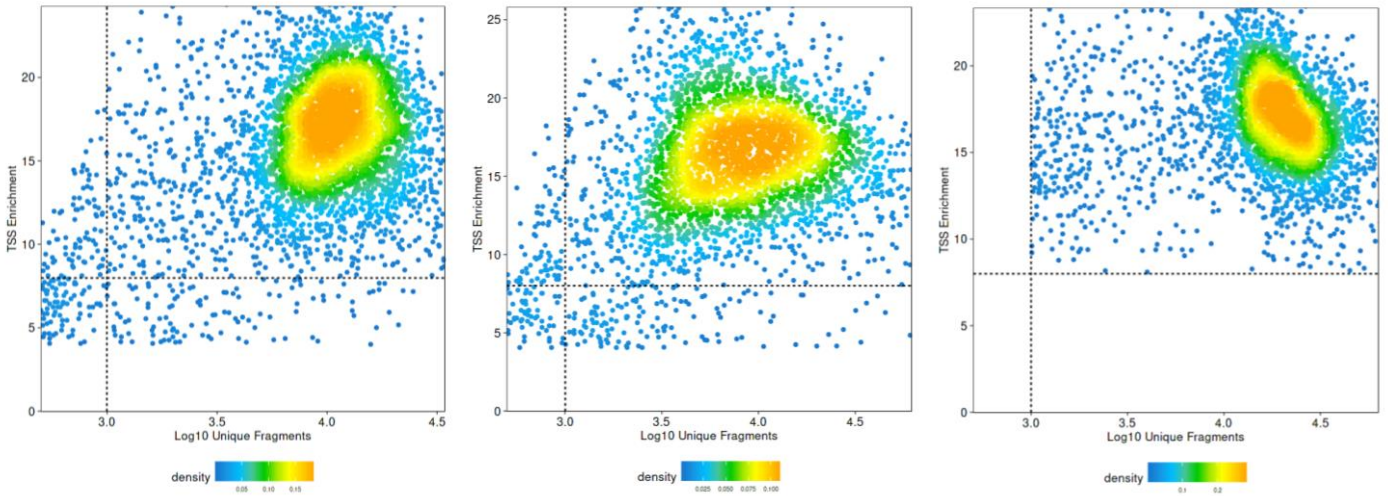
CD34_D7T1: 3575

CD34_D8T1: 3492



Quality control





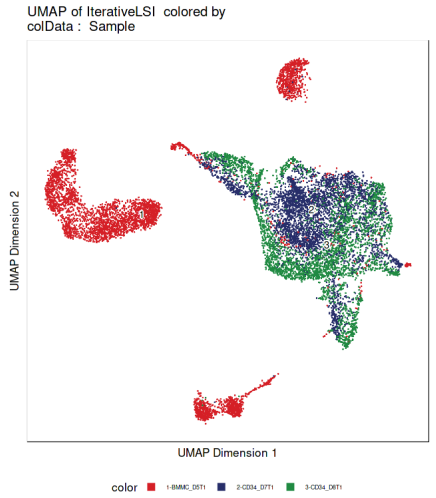
Inspect the plots and report your findings. Do you observe outlier samples/samples with low quality?
From the plots, I would rather choose a TSS score of 7 or 8, as there are still some outliers left. Additionally, I would set the number of Unique frements to 1000 or 2000.

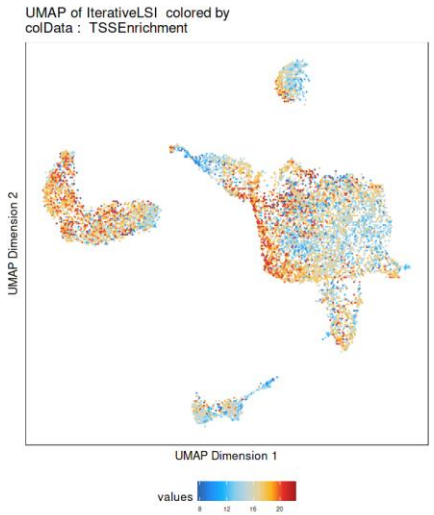
Dimensionality Reduction

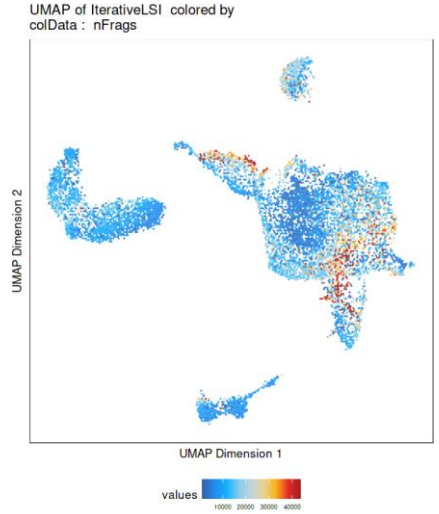
Why do we use LSI and not PCA for the dimensionality reduction of scATAC-data?

“If you were to perform a standard dimensionality reduction, like Principal Component Analysis, on this sparse insertion counts matrix and plot the top two principal components, you would not obtain the desired result because the sparsity causes high inter-cell similarity at all of the 0 positions.” (https://www.archrproject.com/bookdown/dimensionality-reduction-with-archr.html, 3.1.2022)

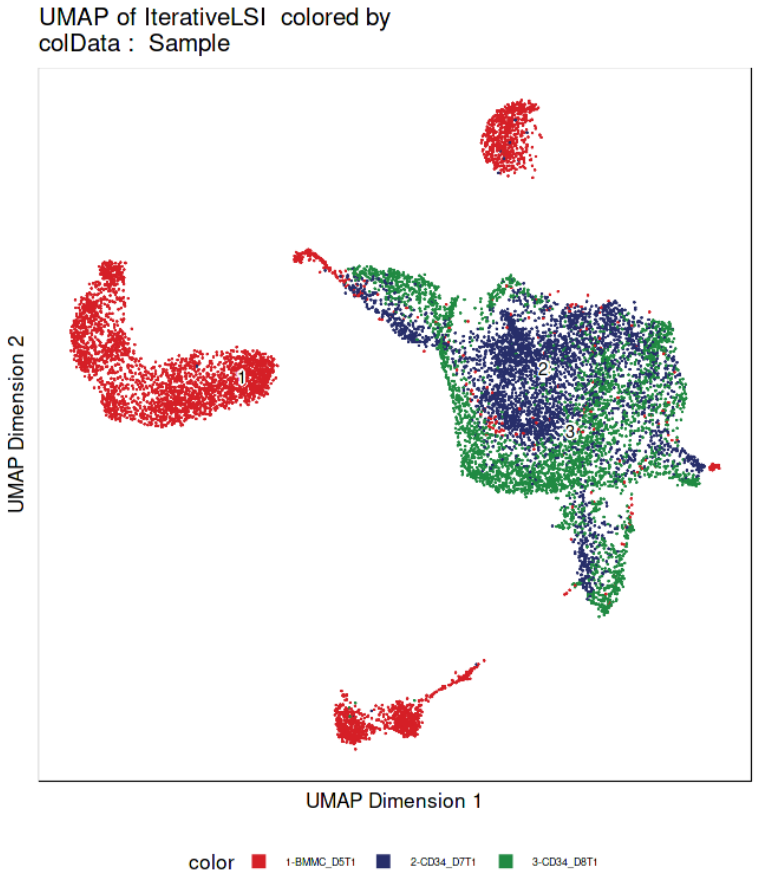
UMAP

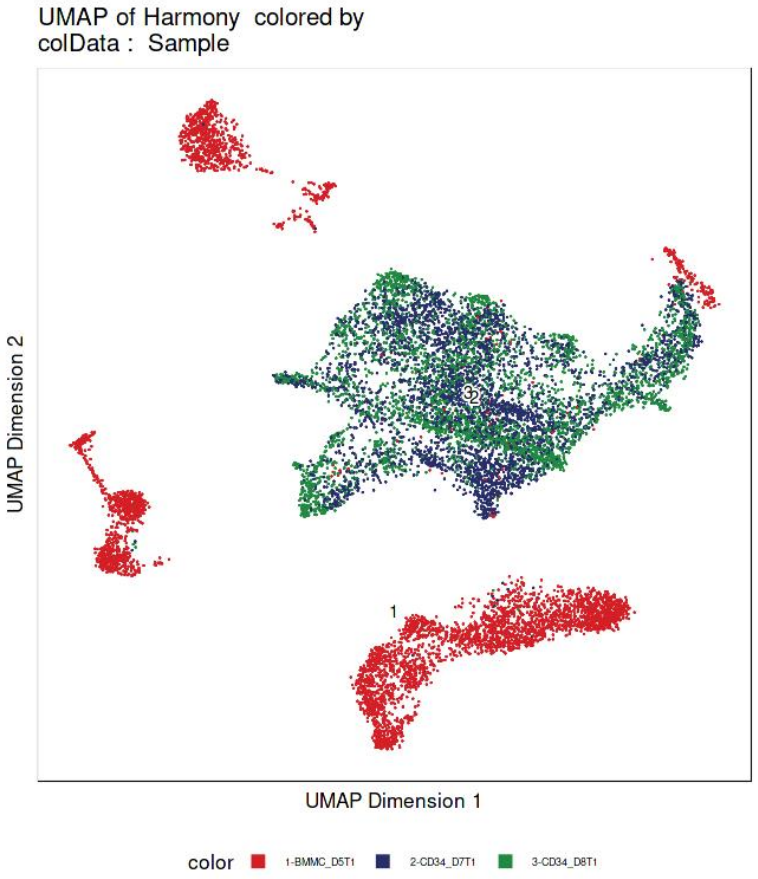




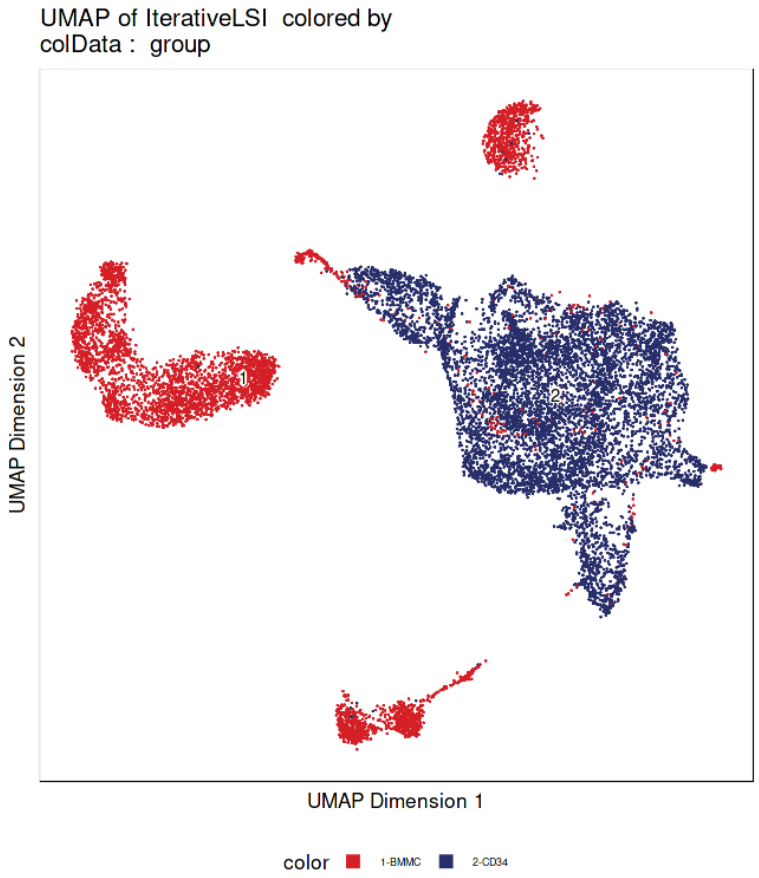


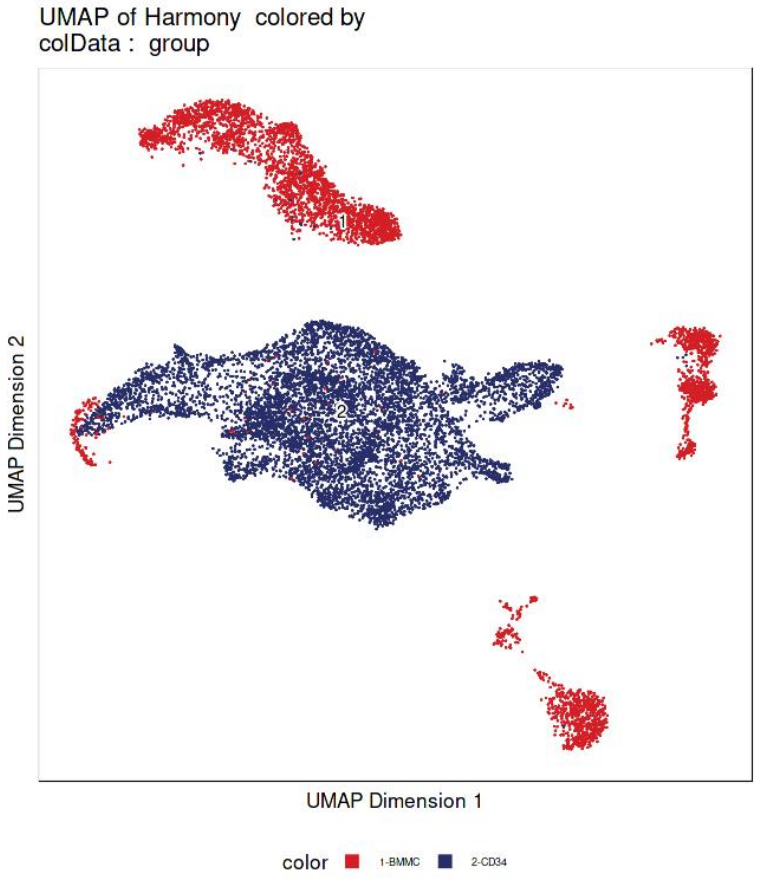
Dealing with batch-effects



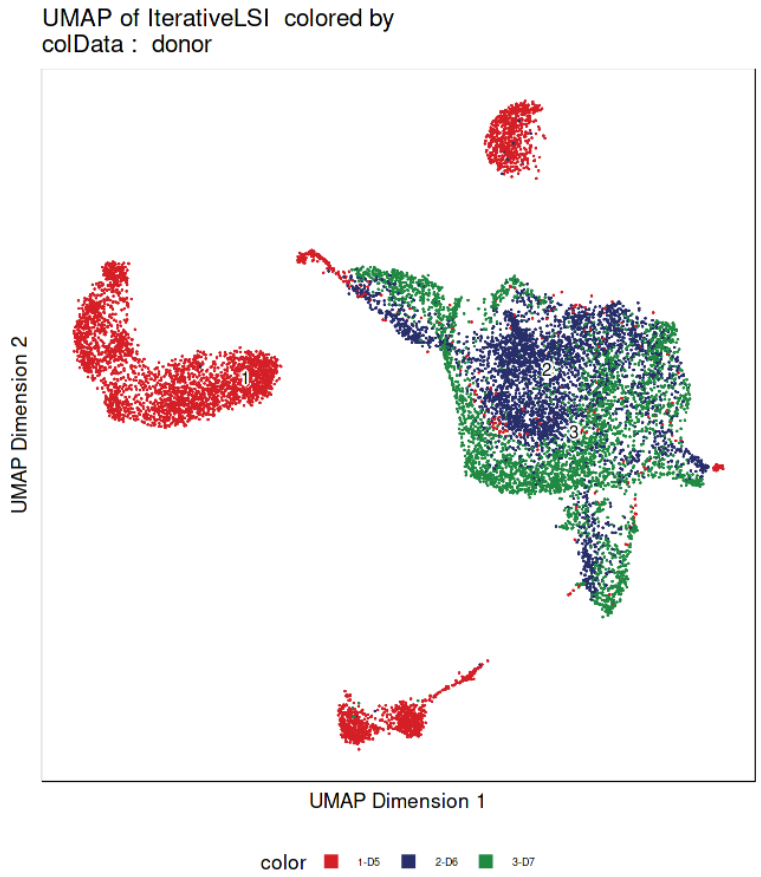


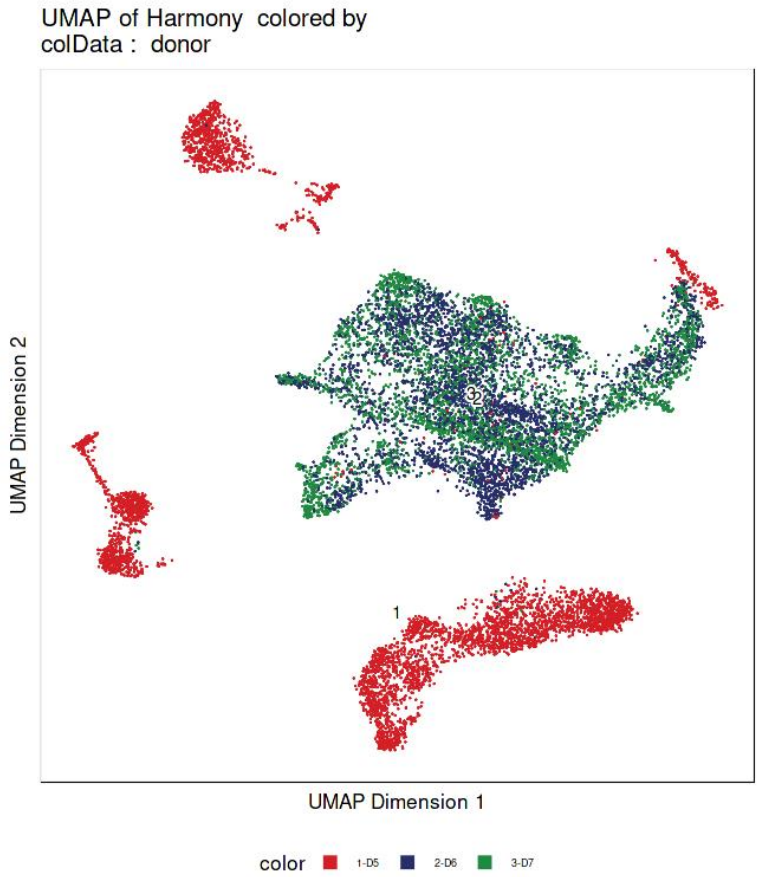
Dealing with batch-effects



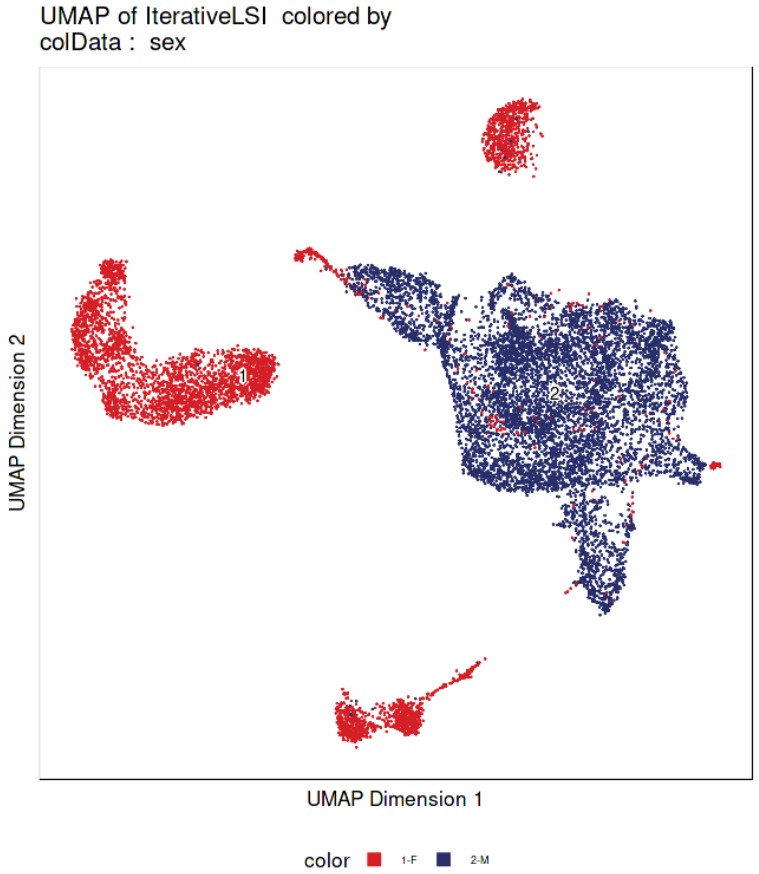


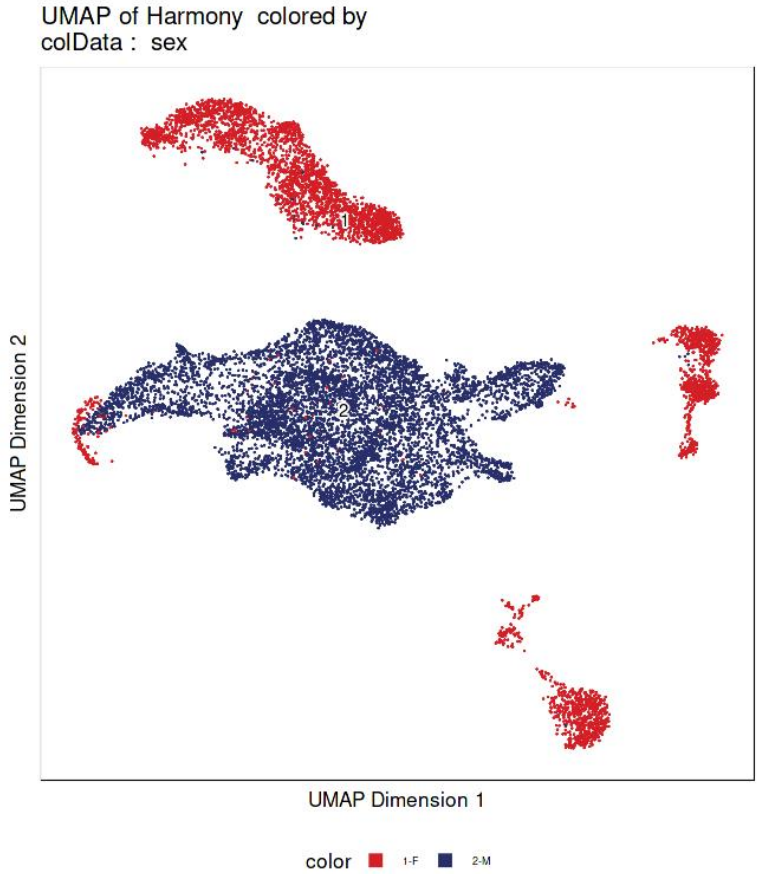
Dealing with batch-effects



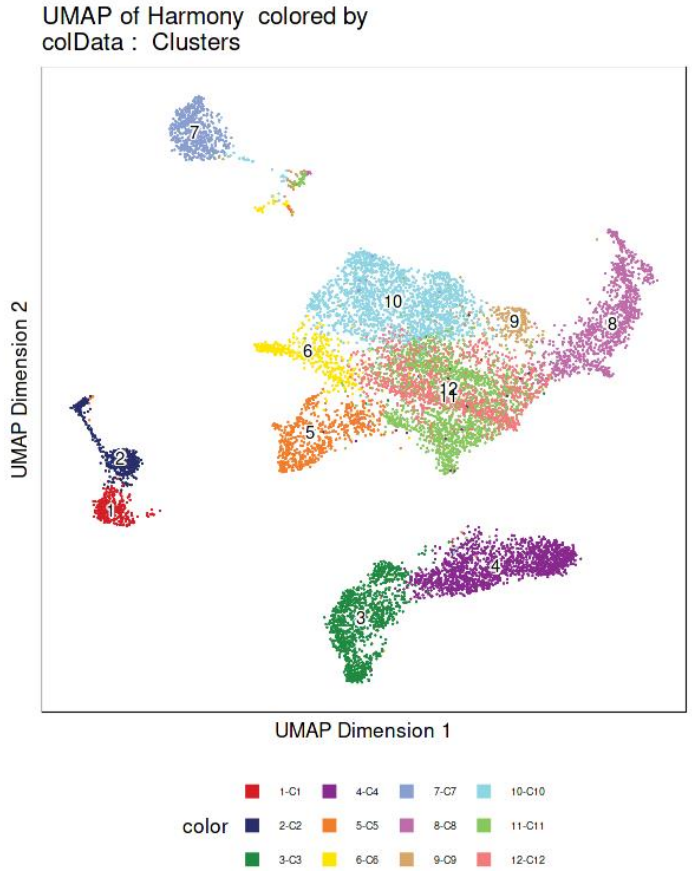


Dealing with batch-effects





Clustering



Clustering

- How many cells does each cluster contain?

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
383	480	1159	1539	687	472	613	1066	214	1841	1726	1374

- What are the sample proportions in each cluster?

	BMMC_D5T1	CD34_D7T1	CD34_D8T1
C3	1.00000000	0	0
C8	0.13977486	0.3611632270	0.499061914
C7	0.98042414	0.0146818923	0.004893964
C1	0.98433420	0.0078328982	0.007832898
C10	0.01955459	0.4894079305	0.491037480
C5	0.04366812	0.3333333333	0.622998544
C4	0.99935023	0.0006497726	.
C2	0.99583333	0.0020833333	0.002083333
C11	0.03012746	0.9455388181	0.024333720
C6	0.10169492	0.4724576271	0.425847458
C9	0.06542056	0.3411214953	0.593457944
C12	0.00363901	0.0858806405	0.910480349

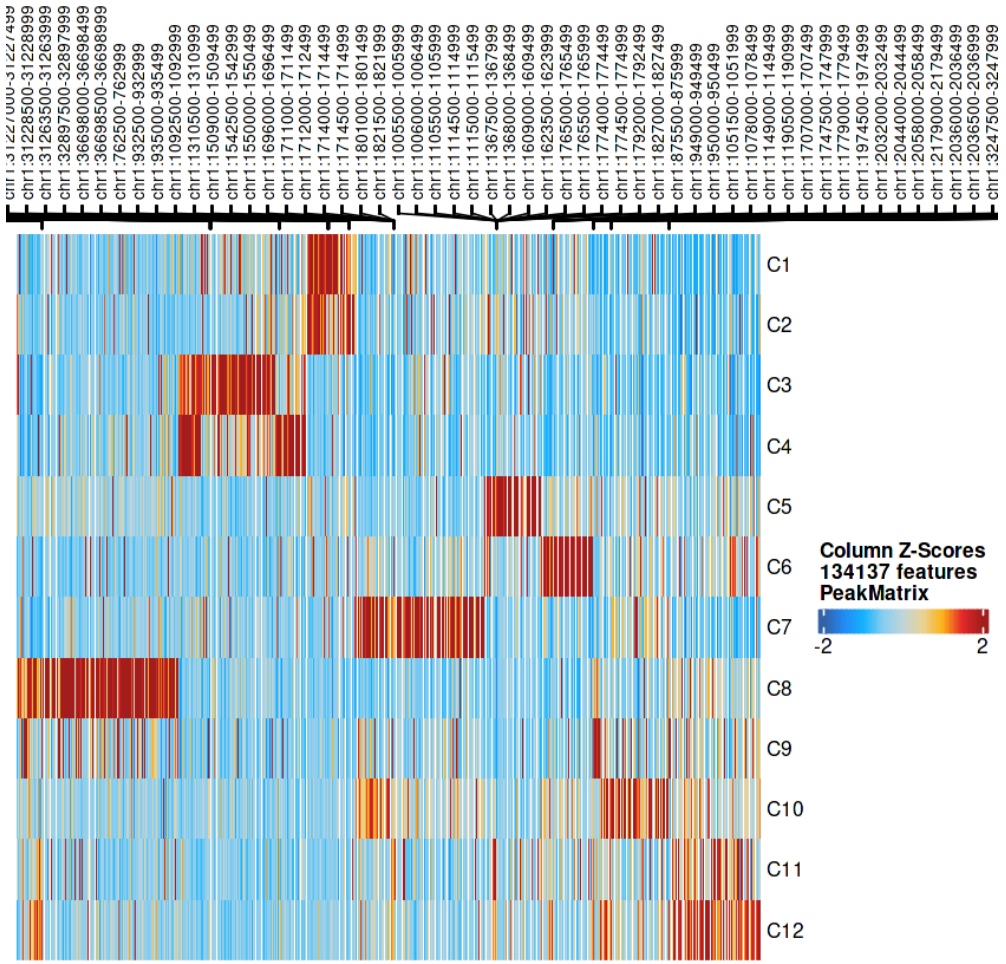
Peaks

Prepare the data for peak-calling. Explain what you did and why this is necessary.

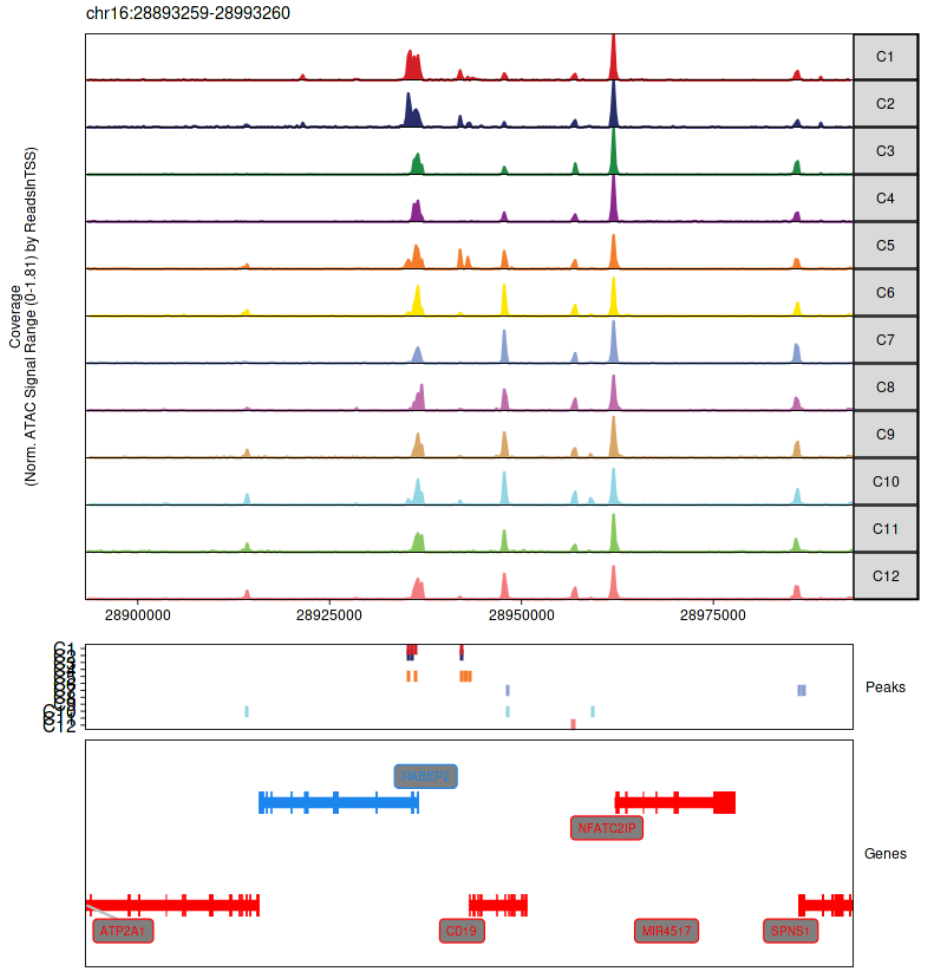
What should be done: pseudo-bulk analysis “data from each single cell is combined into a single pseudo sample”(https://www.archrproject.com/bookdown/calling-peaks-with-archr.html, 3.1.2022)

“Because per-cell scATAC-seq data is essentially binary (accessible or not accessible), we can not call peaks on an individual cell basis.” (https://www.archrproject.com/bookdown/calling-peaks-with-archr.html, 3.1.2022)

Cluster marker peaks

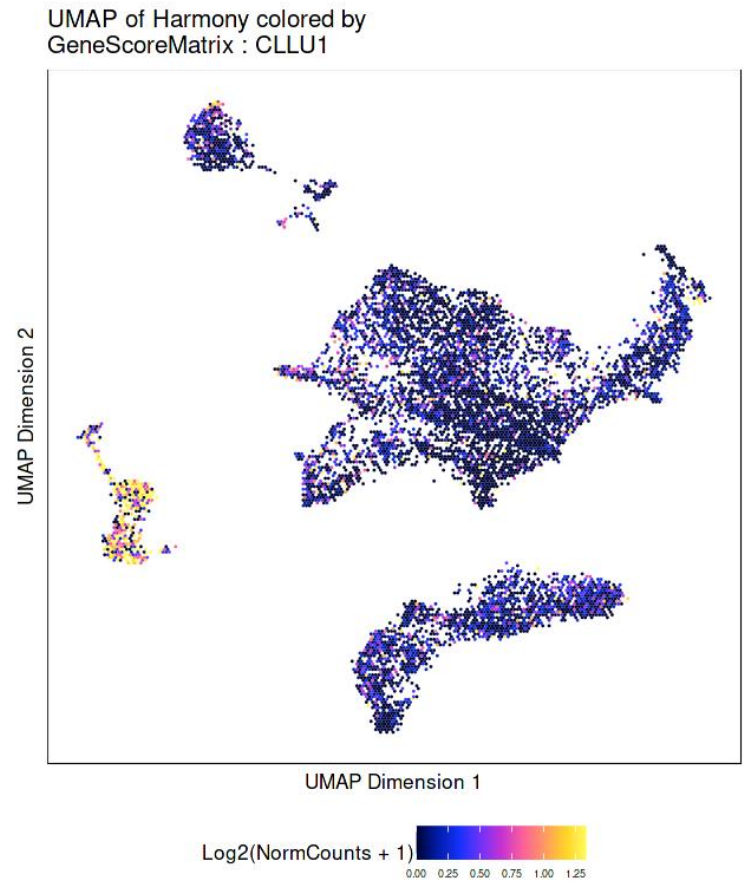


Cluster marker peaks

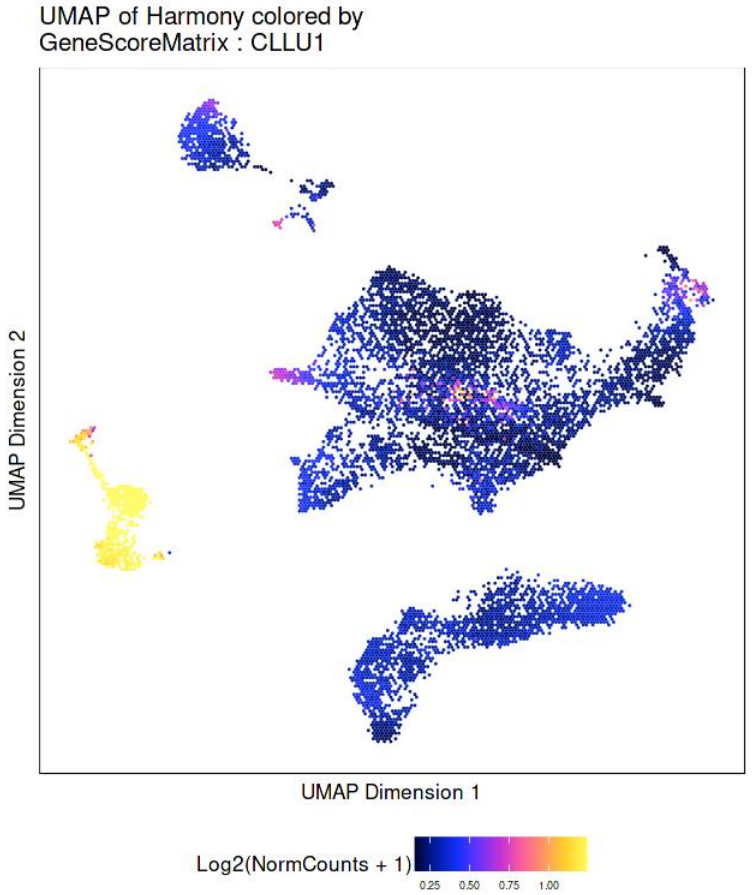


Using MAGIC

Without MAGIC



With MAGIC



Using MAGIC

What is the main idea behind MAGIC (1-2 sentences).

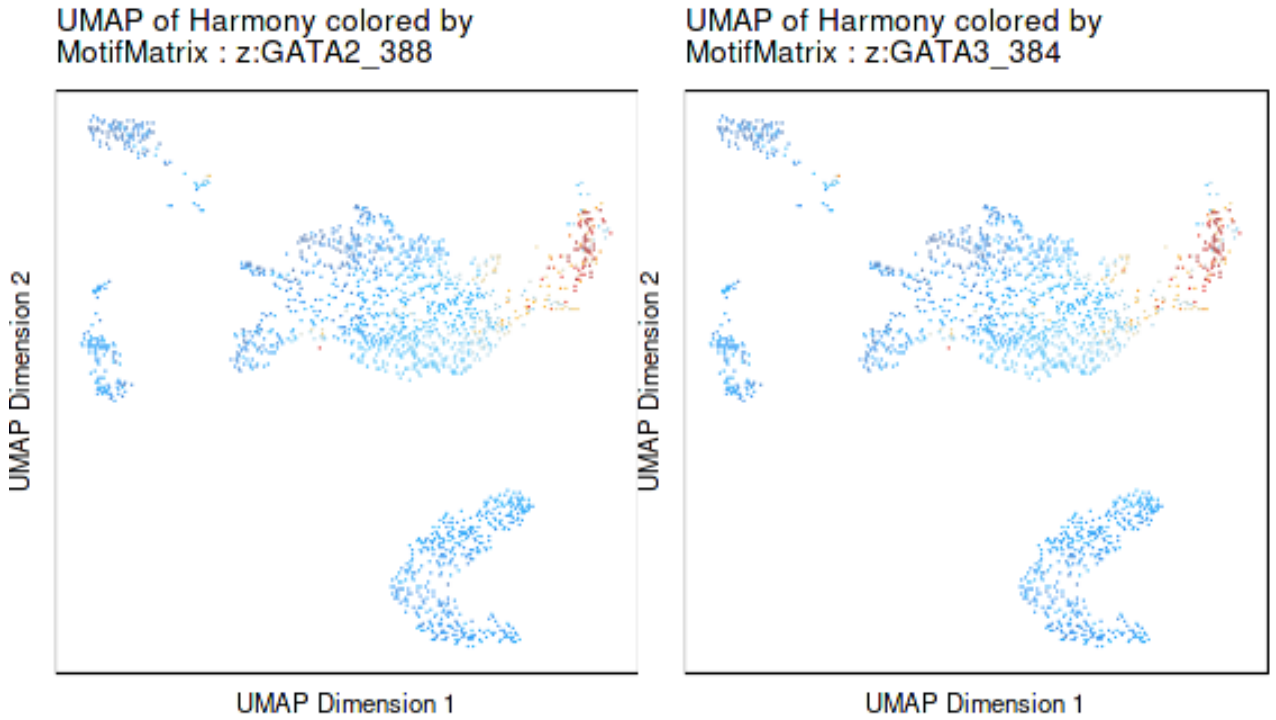
"Markov Affinity-based Graph Imputation of Cells (MAGIC) is an algorithm for denoising high-dimensional data most commonly applied to single-cell RNA sequencing data.
MAGIC learns the manifold data, using the resultant graph to smooth the features and restore the structure of the data."
(https://www.archrproject.com/bookdown/marker-genes-imputation-with-magic.html, 3.1.2022)

TF motif activity

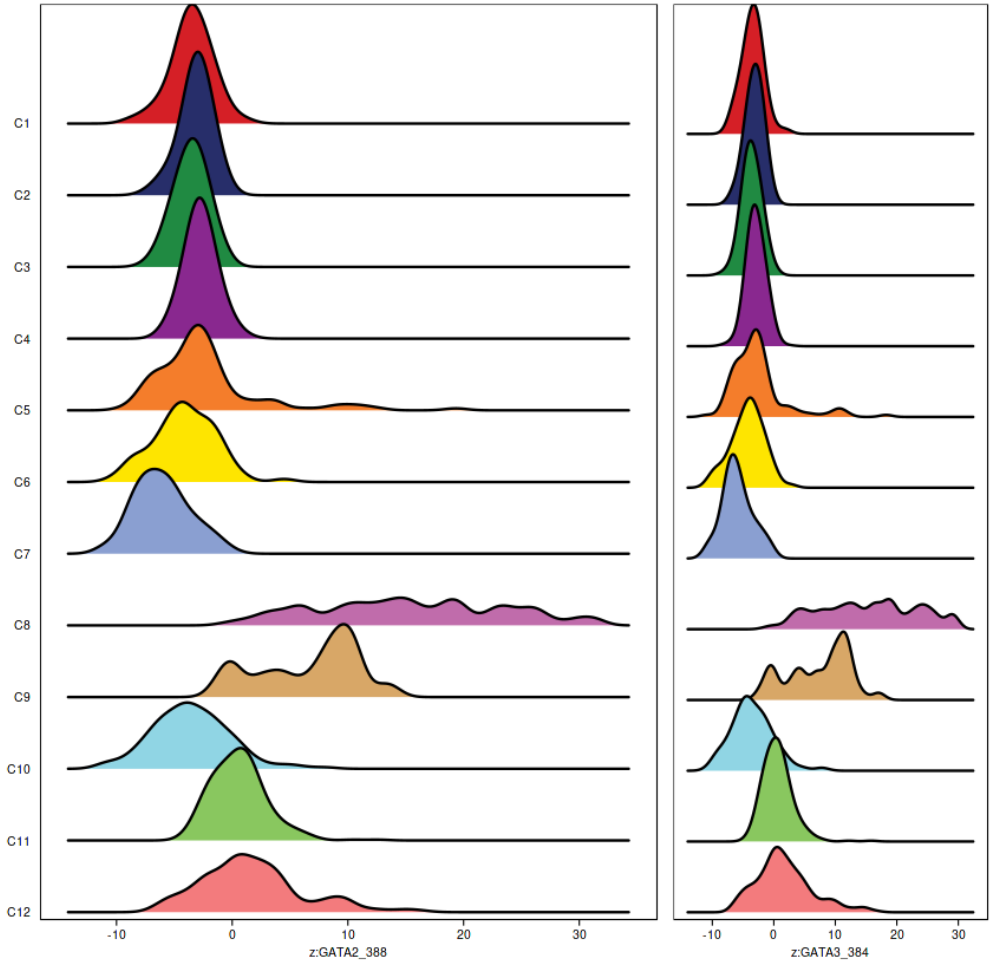
How was the annotation obtained?

“we can look for motifs that are enriched in peaks that are up or down in various cell types. To do this, we must first add these motif annotations to our ArchRProject. This effectively creates a binary matrix where the presence of a motif in each peak is indicated numerically.” (https://www.archrproject.com/bookdown/motif-enrichment-in-differential-peaks.html, 4.1.2022)

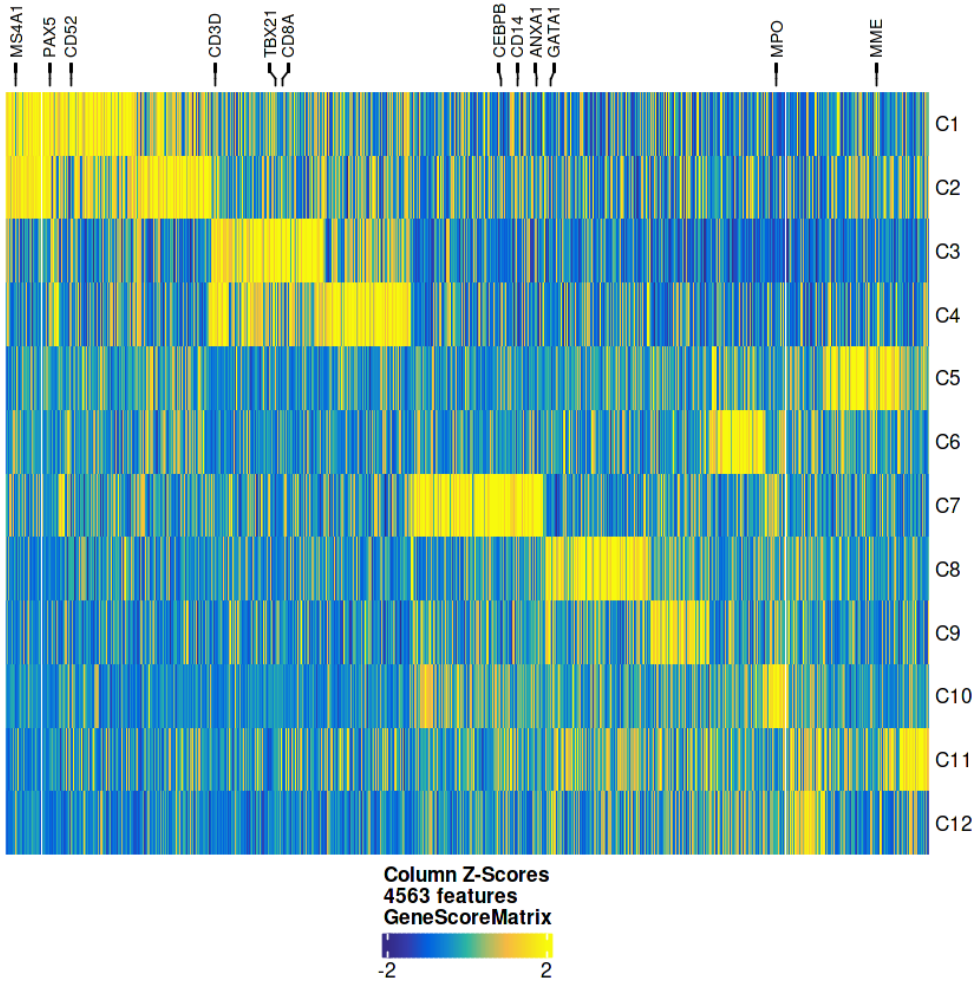
Plot UMAP embeddings for marker TFs

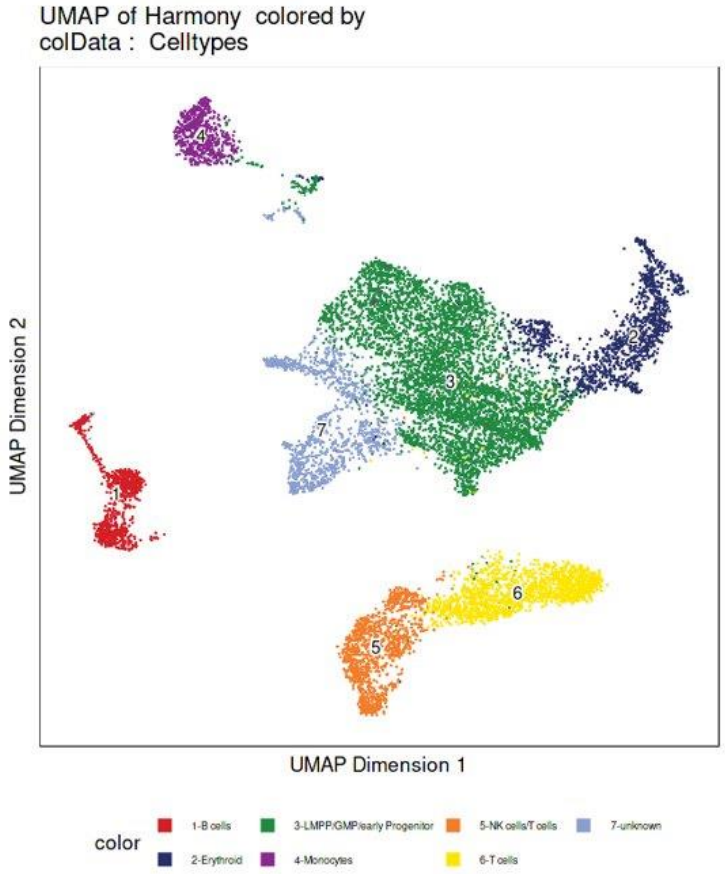


Motif activity

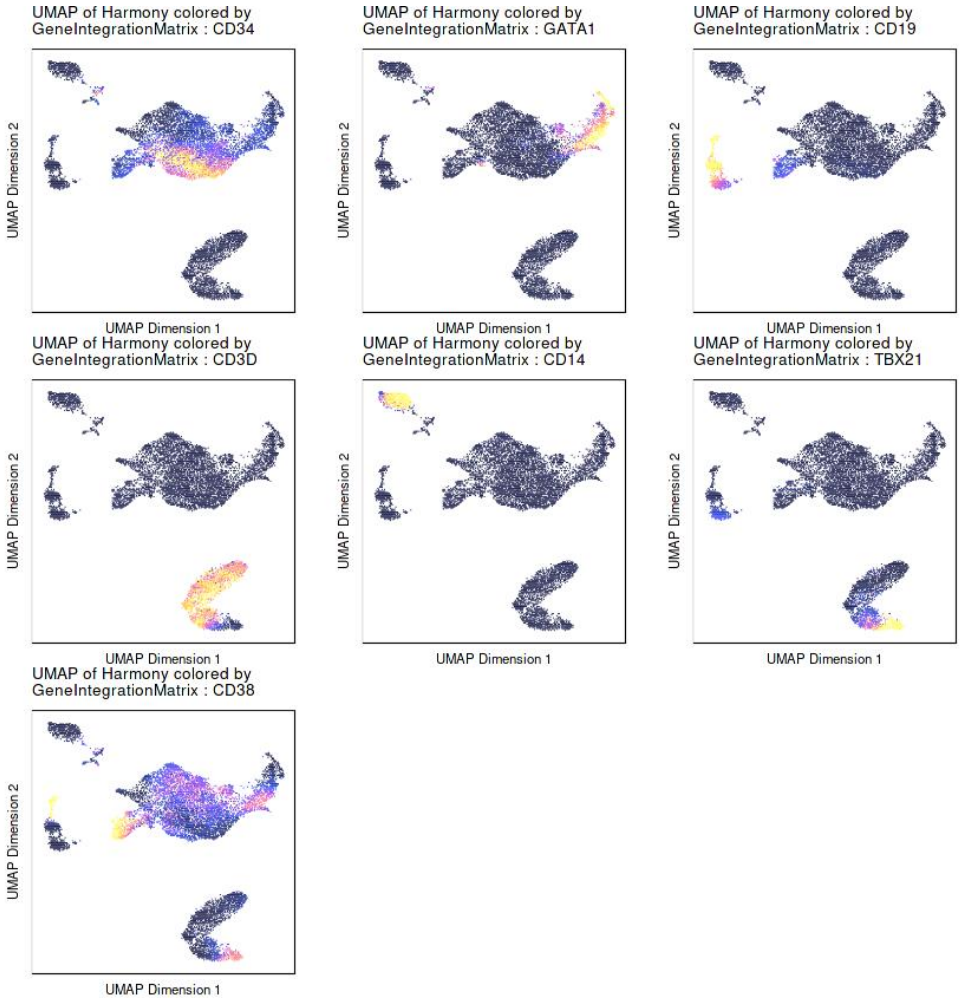


Assign cell-types to clusters

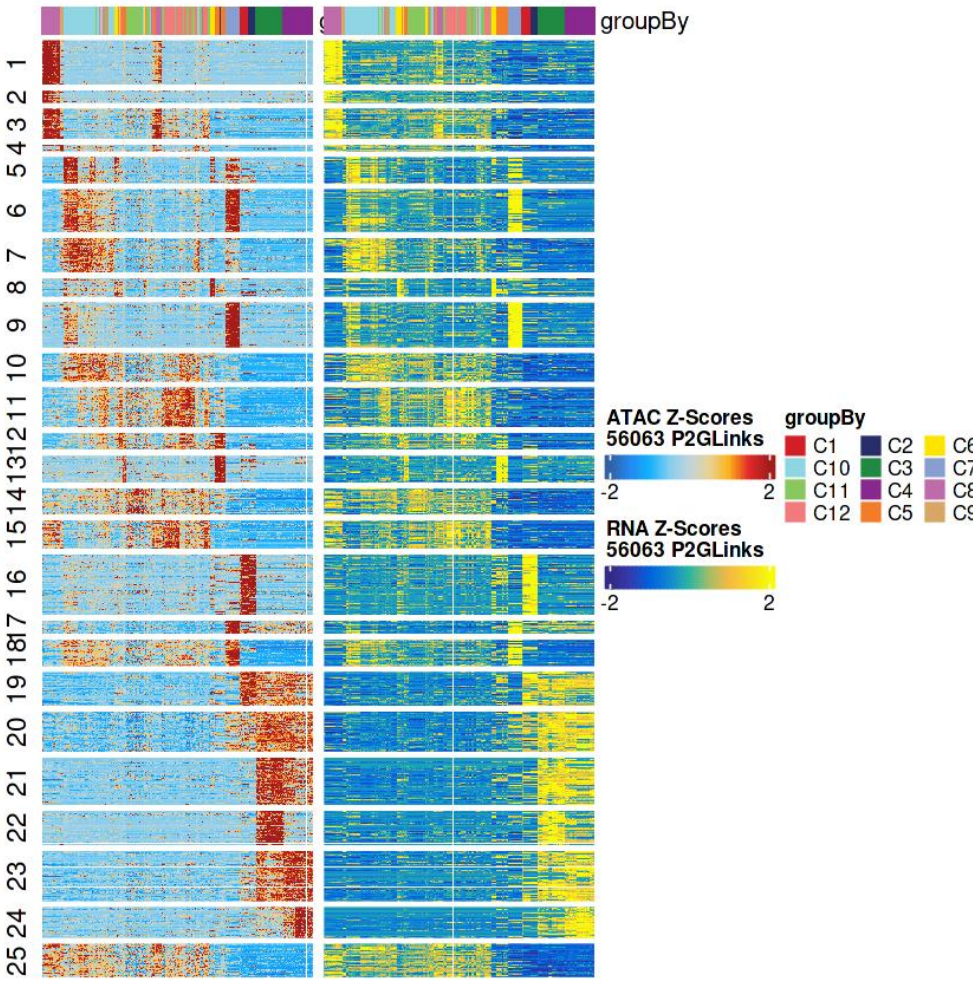


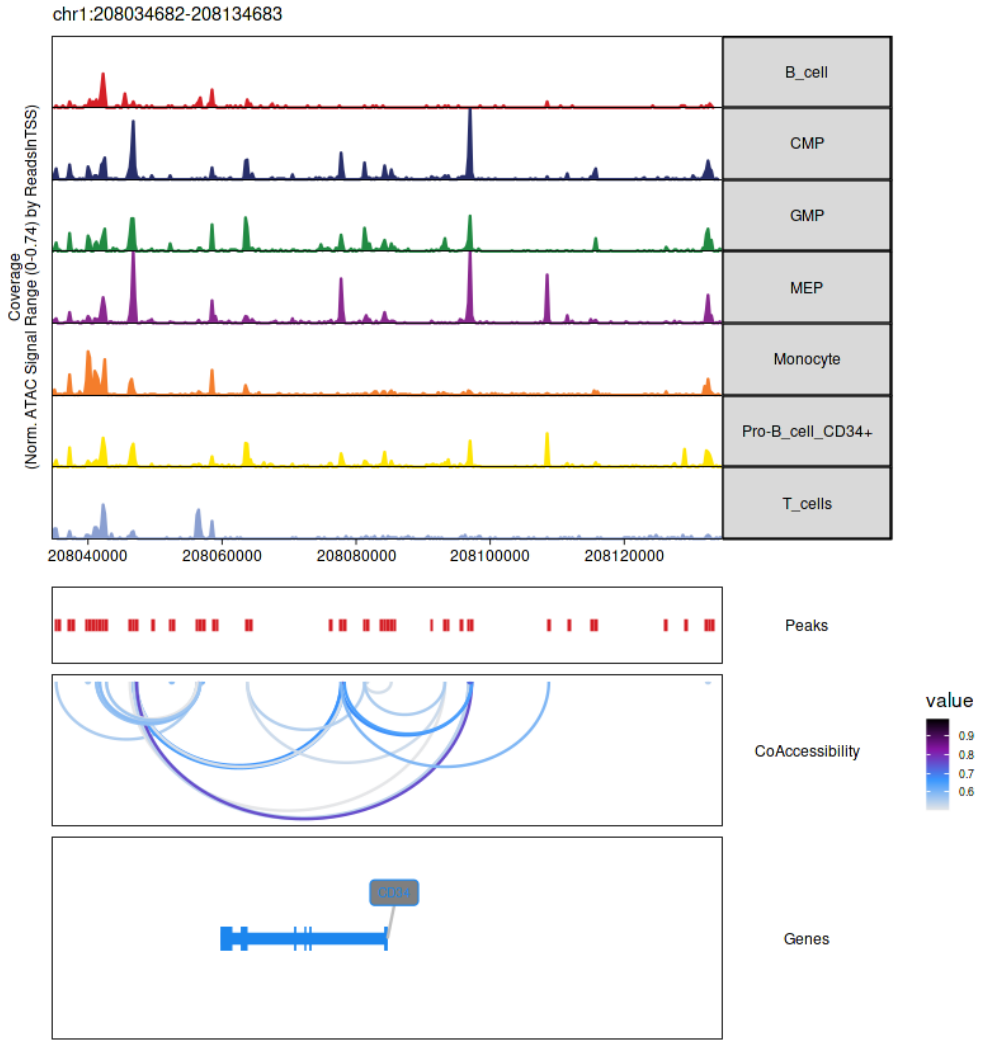


Data integration



Peak-gene lineage

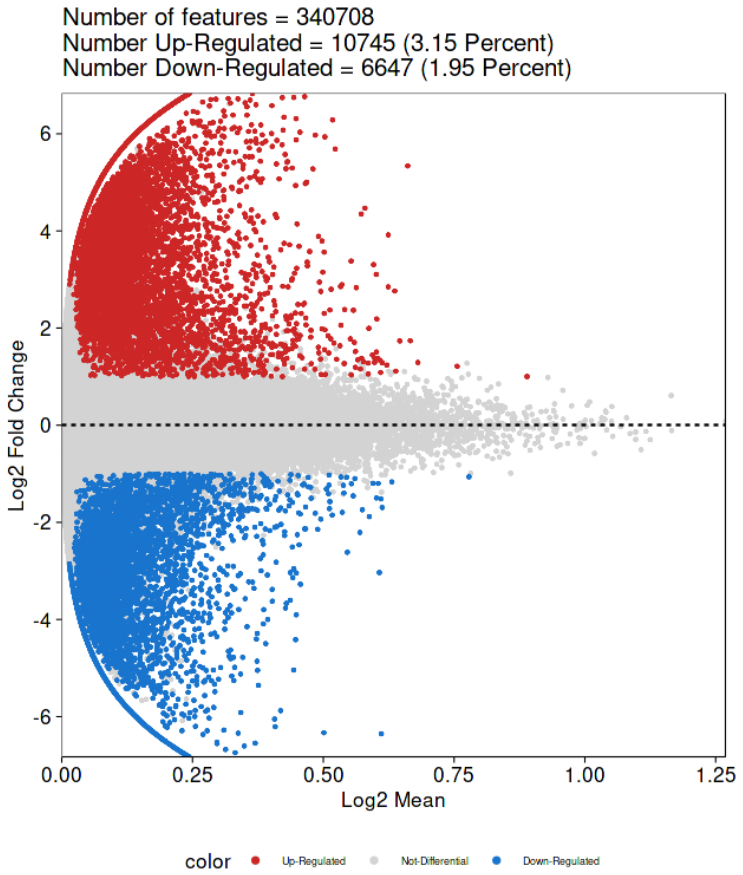


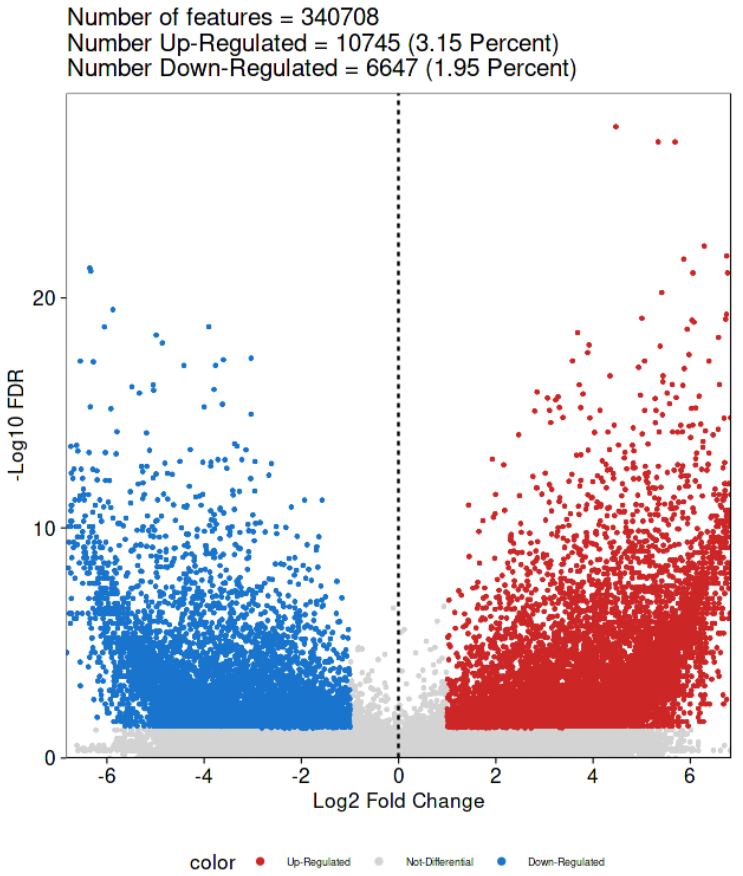


CHAIR FOR CLINICAL BIOINFORMATICS
SAARLAND UNIVERSITY

Integrative Cellular Biology & Bioinformatics

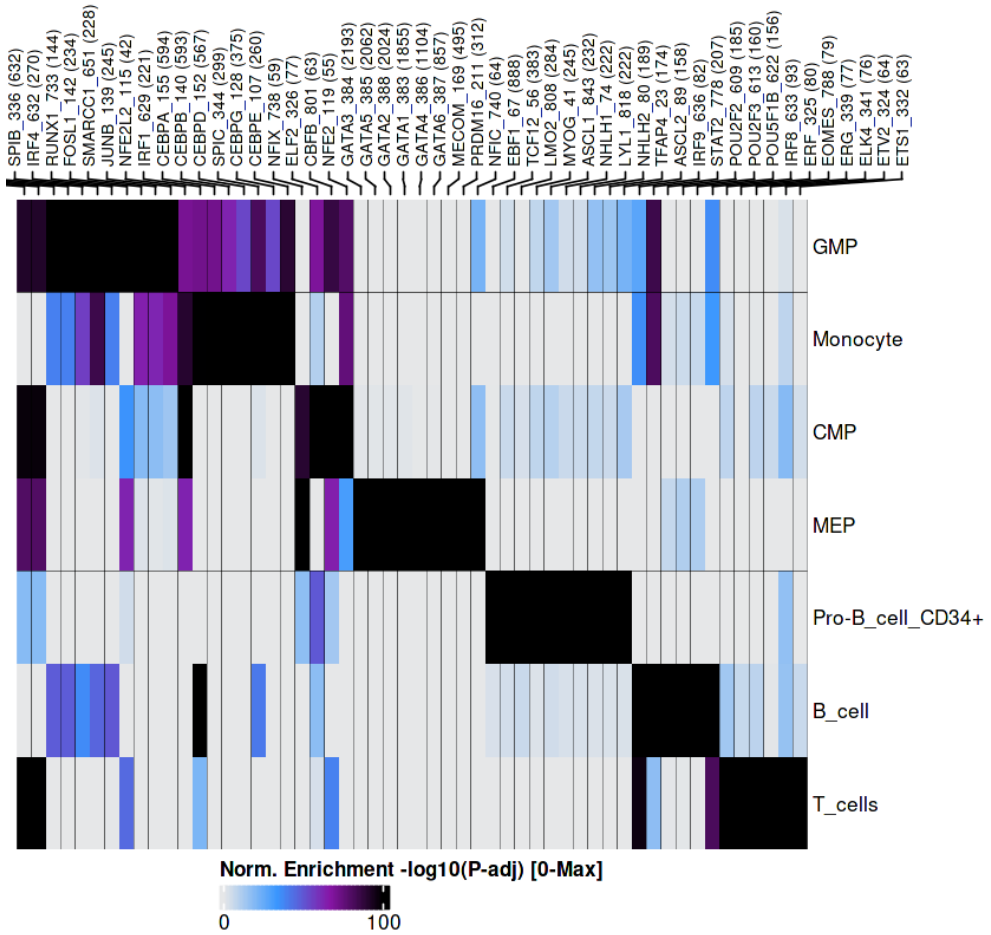
Differential peak accessibility





CHAIR FOR CLINICAL BIOINFORMATICS
SAARLAND UNIVERSITY
Integrative Cellular Biology & Bioinformatics

TF motif enrichment



TF footprinting

Explain the need to correct for Tn5 bias.

“. . . the insertion sequence bias of the Tn5 transposase [. . .] can lead to misclassification of TF footprints” (https://www.archrproject.com/bookdown/normalization-of-footprints-for-tn5-bias.html, 3.1.2022)

Co-accessibility