Vilniaus universitetas

Doktoranto Karolio Šablausko ataskaita už 2022/2023 mokslo metų antrąjį pusmetį

Darbo vadovas: prof. Audronė Jakaitienė

Vilniaus universitetas

Disertacijos pavadinimas: Characterization of genetic changes using deep neural networks

Doktorantūros pradžios ir pabaigos metai: 2022 – 2026

Studijų metai: 1.

Vilniaus universitetas

1 lentelė: Doktorantūros studijų planas

Studijų metai	Egzaminai	
	Planas	Įvykdyta
I (2022/2023)	1	1
II (2023/2024)	2	0
III (2024/2025)	1	0
IV (2025/2026)	0	0
Iš viso:	4	0

Vilniaus universitetas

Studijų metai	Dalyvavimas konferencijose				Publikacijos					
	Tarptautinėse		Nacionalinėse		Su citavimo rodikliu			Be citavimo rodiklio		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė	Planas	Įvykdyta	Būklė
I (2022/ 2023)	0	0	0	0	0	0		0	0	
II (2023/ 2024)	1	0	0	0	0	0		0	0	
III (2024/ 2025)	1	0	0	0	1	0		0	0	
IV (2025/ 2026)	1	0	0	0	1	0		0	0	
Iš viso:	3	0	0	0	2	0		0	0	

Vilniaus universitetas

Publikacijos

Priimta spaudai (2023-09-16):

British Journal of Haematology

IF: 8.615

TITLE: Immunogenicity and clinical effectiveness of BNT162b2 booster against SARS-CoV-2 Omicron in patients with haematological malignancies: a national prospective cohort study

SHORT TITLE: Booster dose against Omicron in haematological patients

KEYWORDS: BNT162b2, booster, Omicron, haematologic patients, protection

Lukas Kevličius[1, 2*], Karolis Šablauskas[1, 2*] Kazimieras Maneikis[1,2], Dovilė Juozapaitė[1], Ugnė Ringelevičiūtė[1], Vilmantė Vaitekėnaitė[1, 2], Birutė Davainienė[1, 2], Guoda Daukėlaitė[1, 2], Dominika Vasilevska[2], Mindaugas Stoškus[1], Ieva Narkevičiūtė[1], Violeta Sivickienė[1], Kęstutis Rudaitis[1], Birutė Masaitytė – Vitlipienė[1], Mantas Minkauskas[1], Daniel Naumovas[1], Tumas Beinortas[3, 4#], Laimonas Griškevičius[1, 2#]

Vilniaus universitetas

Publikacijos

Pateikta spaudai (2023-08-31)

Nature Communications

IF: 16.6

Title

Comprehensive reanalysis of CNVs in unsolved rare disease patients results in new diagnoses

Authors

German Demidov*, Burcu Yaldiz*, José Garcia-Pelaez*, Elke de Boer*, Nika Schuermans*, Liedewei Van de Vonde*, Ida Paramonov, Lennart Johansson, Francesco Musacchia, Elisa Benetti, Gemma Bullich, Karolis Sablauskas, Sergi Beltran, Christian Gilissen, Alex Hoischen, Stephan Ossowski, Richarda M. de Voer, Katja Lohmann, Carla Oliveira, Ana Topf, Lisenka E.L.M. Vissers, Steven Laurie

Vilniaus universitetas

1 Thesis design

1.1 Thesis aim

- To contribute to the advancement of deep learning techniques for the analysis of next generation sequencing data, with a focus on single cell RNA sequencing (scRNA-seq) data.

1.2 Thesis objectives

- **Data preprocessing and feature engineering**: create efficient data preprocessing pipeline suited for scRNA-seq data, including normalization and batch effect correction to prepare the data for further analysis.
- **Deep differential expression analysis**: develop and implement deep learning-based approach for identifying differentially expressed genes and pathways.
- **Evaluation and benchmarking**: conduct extensive benchmarking and cross-validation experiments to assess the performance and generalizability of deep learning models, comparing them to traditional methods.
- **Biological case study**: apply said techniques in the interpretation of biological data gathered during a biomedical study.

Data structure

Vilniaus universitetas

	GENE_1	GENE_2	GENE_3	…	GENE_N
CELL_1	10	0	15		0
CELL_2	0	12	8		9
CELL_3	9	0	0		0
…					
CELL_M	8	7	1		0

Single cell data is represented in M x N matrix

	barcode	library	total_counts	pct_counts_mito	patient	part	timepoint	response	n
55	AACAGCGAAGCACTGT	AC3_2	1705	2.05279	AC	2	3	RD	
66	AACAGCGAAGCTCGTA	AC3_2	618	2.42718	AC	2	3	RD	
97	AACAGCGAATGACCTG	AC3_2	1655	0.060423	AC	2	3	RD	
134	AACAGCGACATGAGTC	AC3_2	606	3.13531	AC	2	3	RD	
135	AACAGCGACATGCACA	AC3_2	509	2.75049	AC	2	3	RD	

Associated information (e.g. timepoint, patient, response) about each barcode / cell is saved in a separate dataframe.

Cell counts by patient by timepoint

Vilniaus universitetas

		Timepoint			
	patient	1	2	3	4
0	AC	3707	4932	5360	7486
1	AG	3171	6170	3575	3643
2	AA	4700	4453	4681	1457
3	AL	4649	3429	2884	5124
4	AB	1673	3264	2249	3287
5	AM	3885	2320	961	4845
6	V	777	681	547	605
7	AD	5014	938	3510	5336
8	AI	4691	3210	6033	10919
9	N	3288	3262	1700	1881
10	AH	7985	4449	0	4543
11	AE	8190	1372	2234	3126
12	AK	7069	16162	2933	0
13	S	1320	754	265	209
14	Z	3864	1853	3687	4581
15	T	484	441	550	336
16	P	1226	799	1053	0

Analysis steps:

1. Calculate normalized Z scores:
    1. Normalize values per cell
    2. Log-scale values
    3. Scale values
2. Select highly informative genes (37k -> 3k):
    1. Remove known biological artefacts
    2. Select variable genes
3. Principal component analysis (PCA)
4. Integrate using Harmony (reduces batch effect)
5. Compute neighborhood graph
6. Dimensionality reduction:
    1. UMAP
    2. Force directed layout
7. Find doublets and remove them
8. Repeat 1-6 without doublets
9. Find clusters / cell types

Vilniaus universitetas

Before filtering doublets: (223 782, 37 733)

After filtering doublets: (222 932, 37 733)

37,733

222,782

	GENE_1	GENE_2	GENE_3	…	GENE_N
CELL_1	10	0	15		0
CELL_2	0	12	8		9
CELL_3	9	0	0		0
…					
CELL_M	8	7	1		0

Vilniaus universitetas

Analysis steps:

1. **Calculate normalized Z scores:**
    1. **Normalize values per cell**
    2. **Log-scale values**
    3. **Scale values**

	GENE_1	GENE_2	GENE_3	…	GENE_N
CELL_1	10	0	15		0

SUM=10,000

- Linear -> logarithmic effect
- Scale data to unit variance and zero mean.

Analysis steps:

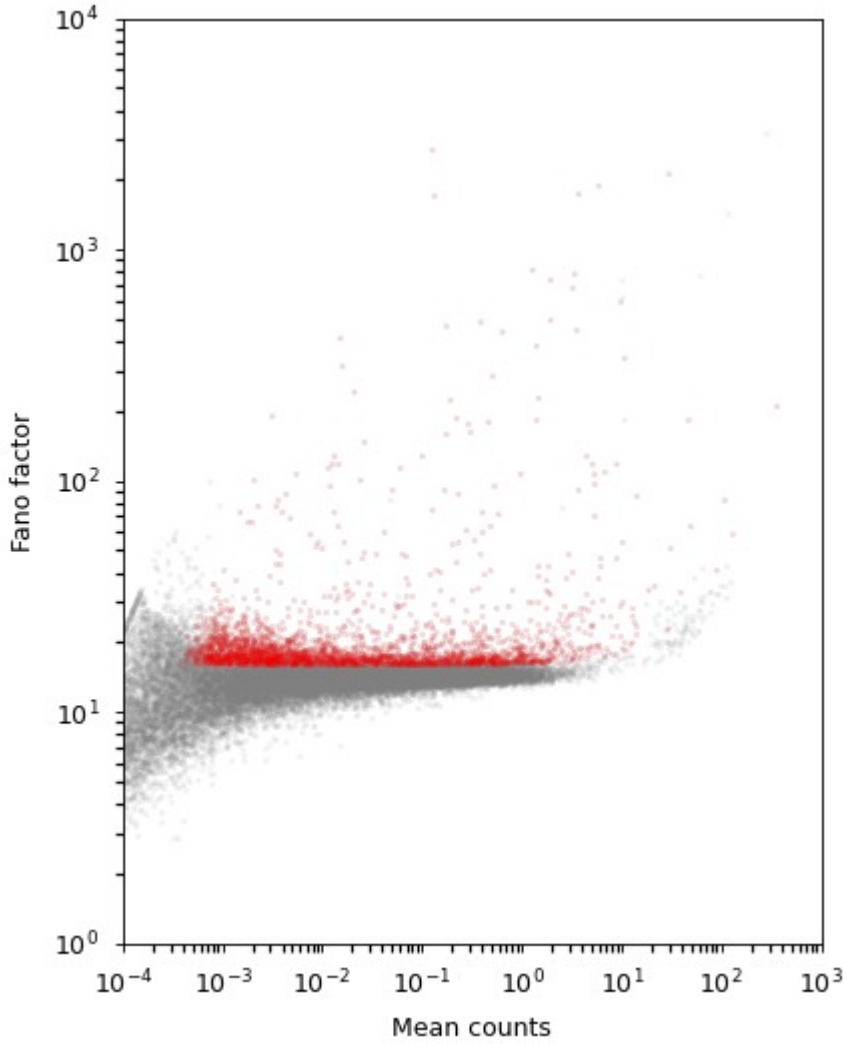
Select highly informative genes (37k -> 3k):

- Remove known biological artefacts
- Select variable genes

Remove known biological artefacts:

- Exclude mitochondrial, ribosomal and hemoglobin genes

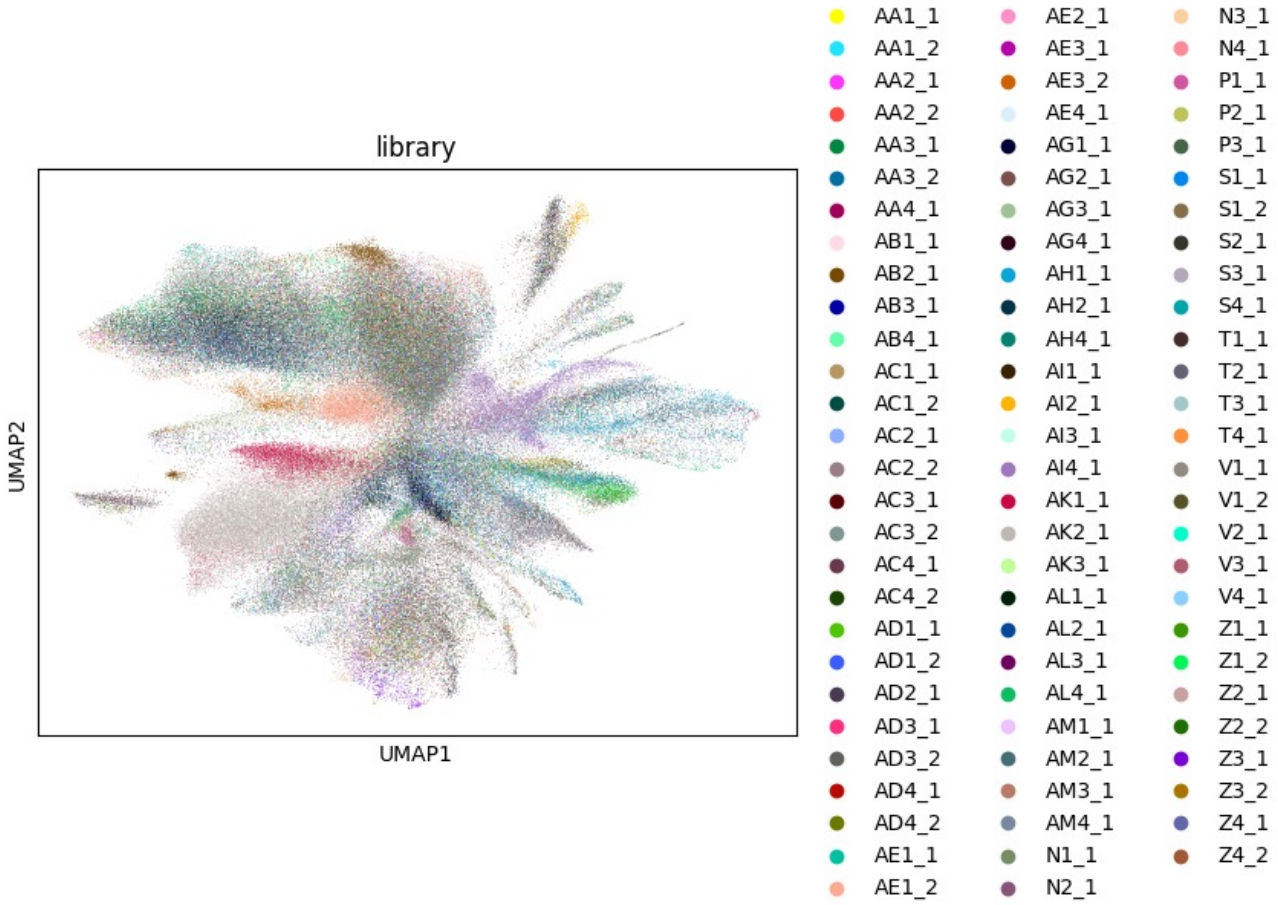
Find and select the most variable genes



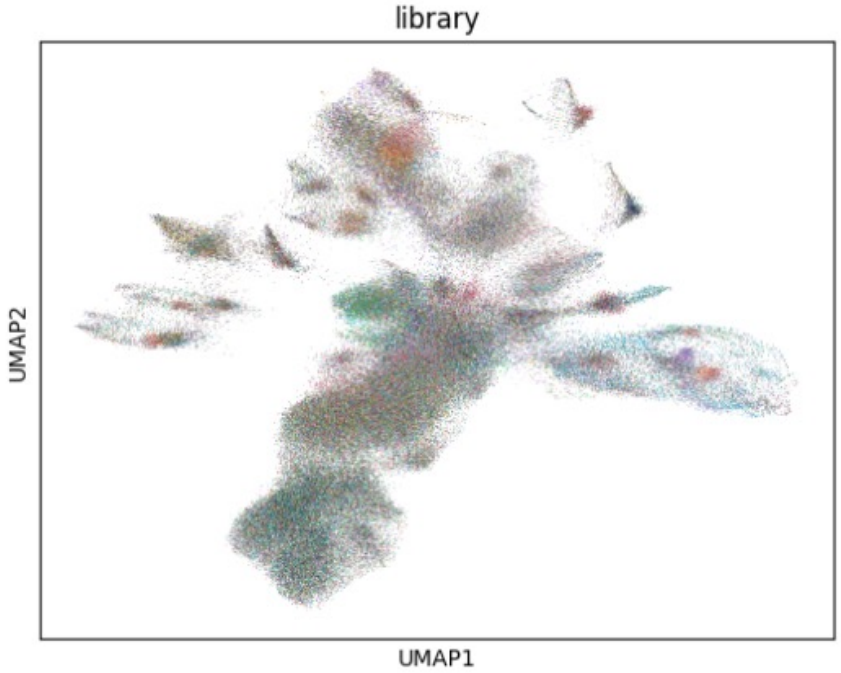
Principal component analysis (PCA)
Integrate using Harmony (reduces batch

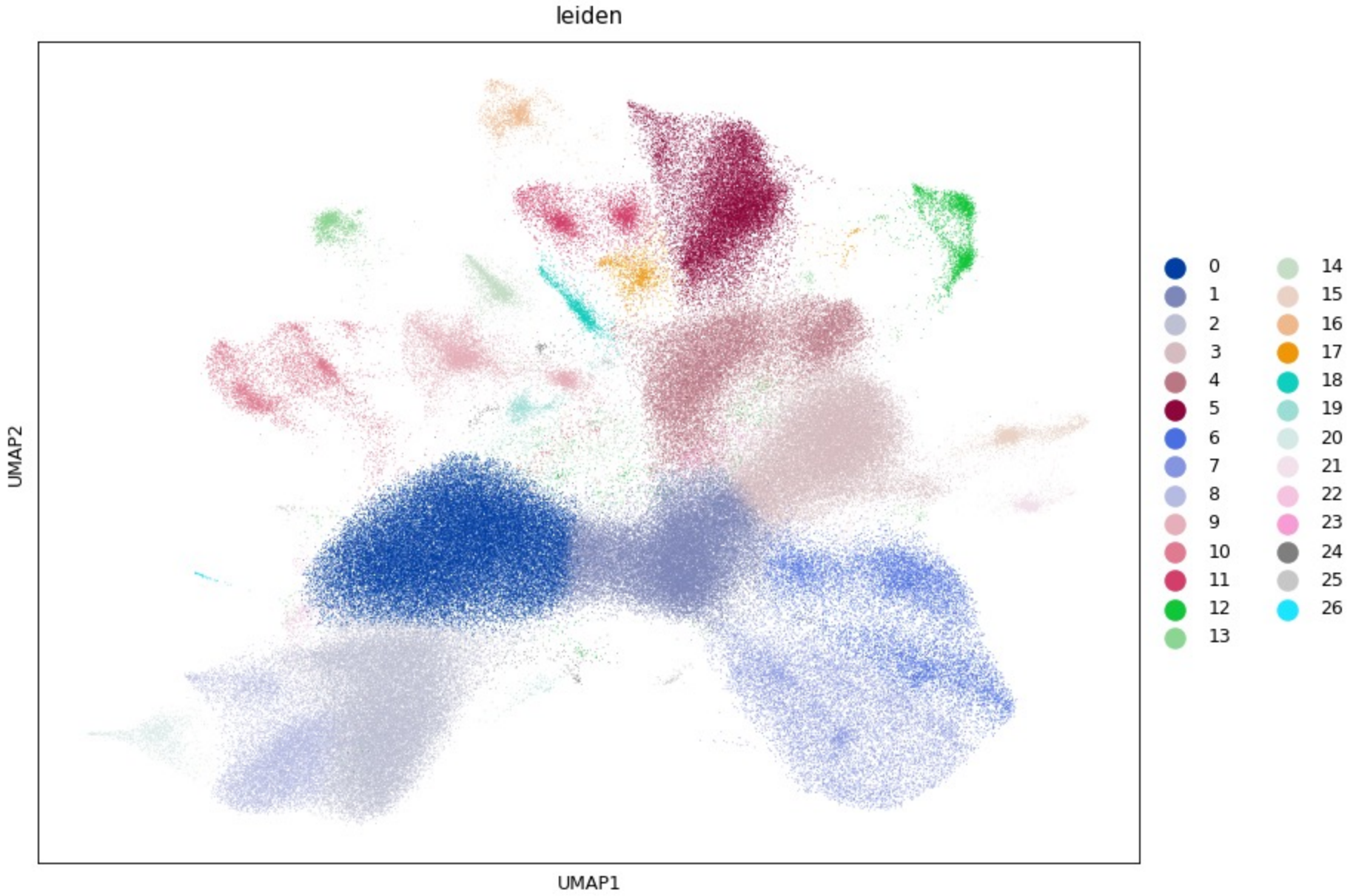
Vilniaus universitetas

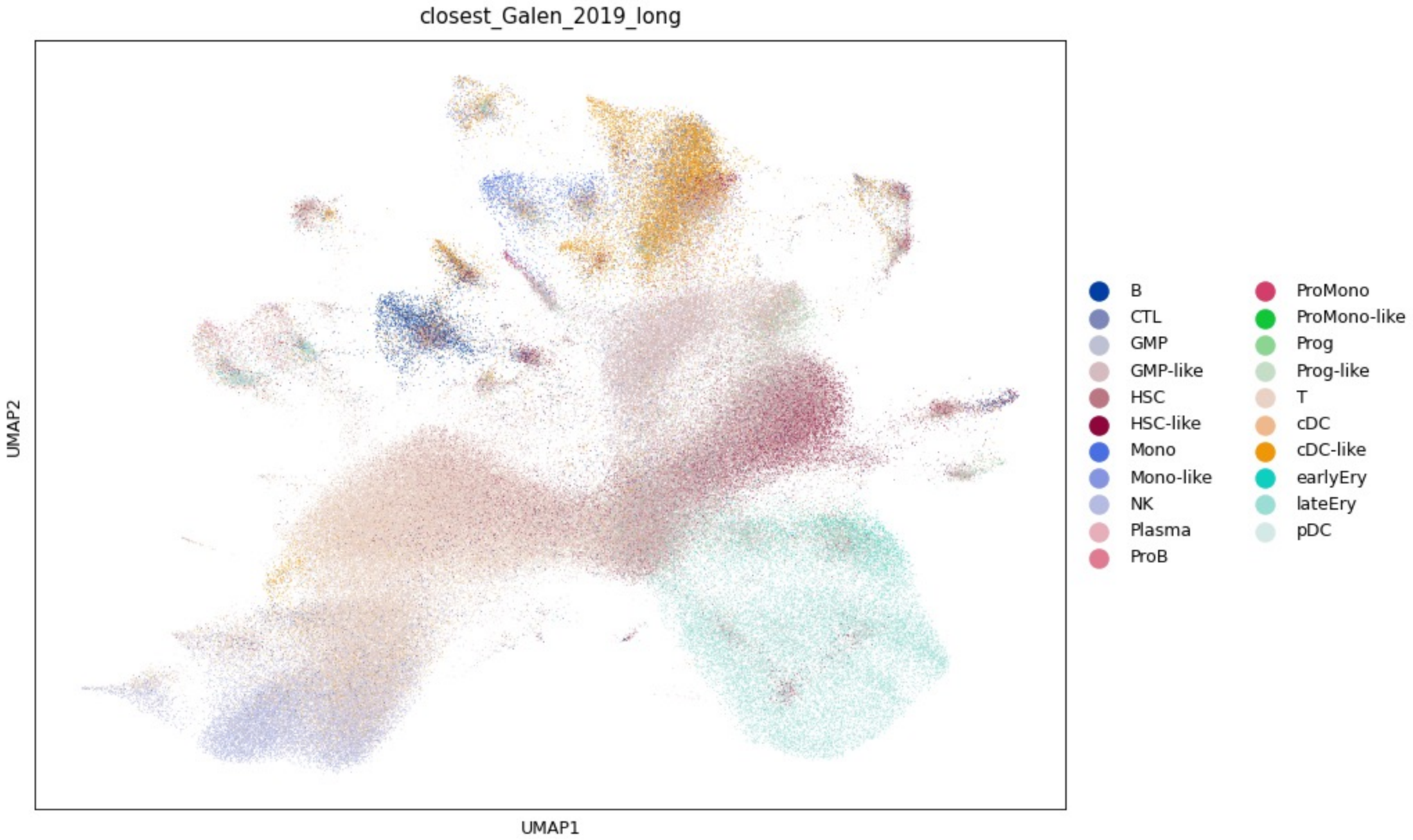
Before Harmony integrate



After Harmony integrate



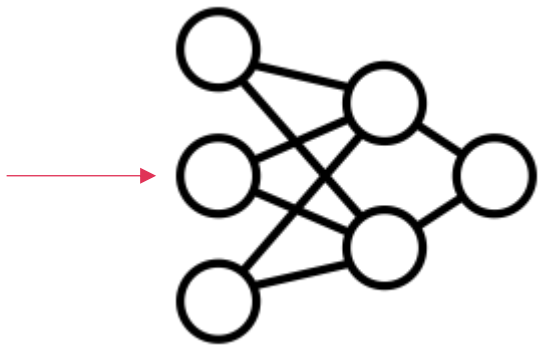

leiden
UMAP2
UMAP1
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26


closest_Galen_2019_long
UMAP2
UMAP1
B
CTL
GMP
GMP-like
HSC
HSC-like
Mono
Mono-like
NK
Plasma
ProB
ProMono
ProMono-like
Prog
Prog-like
T
cDC
cDC-like
earlyEry
lateEry
pDC

	barcode	library	total_counts	pct_counts_mito	patient	part	timepoint	response	n
55	AACAGCGAAGCACTGT	AC3_2	1705	2.05279	AC	2	3	RD	
66	AACAGCGAAGCTCGTA	AC3_2	618	2.42718	AC	2	3	RD	
97	AACAGCGAATGACCTG	AC3_2	1655	0.060423	AC	2	3	RD	
134	AACAGCGACATGAGTC	AC3_2	606	3.13531	AC	2	3	RD	
135	AACAGCGACATGCACA	AC3_2	509	2.75049	AC	2	3	RD	

Vilniaus universitetas

	GENE_1	GENE_2	GENE_3	…	GENE_N
CELL_1	10	0	15		0



REFRACTORY

SENSITIVE

CR vs RD timepoint 1 only
Rapidly dividing cells only

niaus
iversitetas