ITMO UNIVERSITY

CT Lab
ITMO UNIVERSITY

Single-cell Navigator: visualizing scRNA-seq data

Konstantin Zaitsev, ITMO University

August 27th, 2021. Tomsk / My Hotel Room.

Visualizing scRNA-seq data

Main goals:

- To make hypothesis generations easier
- Remove “man-in-the-middle”

Extra goals:

- Fast
- Responsive

Visualizing scRNA-seq data

https://artyomovlab.wustl.edu/scn/

(still in production, so feedback is very welcome)

Let’s open the dataset

- Go to https://artyomovlab.wustl.edu/scn/

scNavigator: beta

scNavigator: beta

Single-cell Navigator is an open-source project dedicated to processing and visualization of single-cell RNA-seq data

Below we have a large collection of datasets and tools to play with:

- Large collection of automatically processed datasets. We processed almost every scRNA-seq dataset from GEO Omnibus database. We make it available for you in our browser.
- Collection of curated datasets. Curated dataset are those that we process by hand. These will include datasets from Human Cell Atlas (HCA), Tabula Muris and some of the datasets that we generated in our lab.
- You can search for cell type specific gene signatures! When we processed all the public scRNA-seq datasets we also calculated all the markers of all the clusters in all these datasets. Just put a list of genes and we will tell you which cluster in which dataset it looks like.
- If you were provided with secret dataset token, you can use it at the very right of this page

Enter a secret token below:

Secret token

Go!

All scRNA-seq datasets | Curated datasets | Gene signature search

Name	Description	Organism	# of cells	Exte...
GSE101901/SRS2384613	Single cell sequencing of hippocampus tissues in traumatic brain injury	Mus Musculus	8878	
GSE103976/SRS2523512	Detecting Activated Cell Populations Using Single-Cell RNA-Seq	Mus Musculus	6488	
GSE129730/SRS4617144	Single cell RNA-seq shows cellular heterogeneity and lineage expansion in a mouse model of SHH-driven medulloblastoma support resistance to SHH inhibitor therapy	Mus Musculus	4552	
GSE103983/SRS2523775	Single-cell RNA-seq (Drop-seq) of MGE, CGE and LGE of E13.5 (MGE) and E14.5 (CGE, LGE) mouse embryos	Mus Musculus	11704	
GSE93374/SRS1913127	A Molecular Census of Arcuate Hypothalamus and Median Eminence Cell Types	Mus Musculus	61225	
GSE103983/SRS2523784	Single-cell RNA-seq (Drop-seq) of MGE, CGE and LGE of E13.5 (MGE) and E14.5 (CGE, LGE) mouse embryos	Mus Musculus	709	
GSE137007/SRS5355828	Proliferation-competent Tcf1+ CD8 T-cells in dysfunctional populations are CD4 T-cell help independent	Mus Musculus	434	
GSE106960/SRS2690039	The single cell RNA seq of pulmonary alveolar epithelial cells	Mus Musculus	2683	
GSE113111/SRS3165512	sc-RNA sequencing of skeletal muscle macrophages during T. gondii infection and injury	Mus Musculus	6625	
GSE129730/SRS4617149	Single cell RNA-seq shows cellular heterogeneity and lineage expansion in a mouse model of SHH-driven medulloblastoma support resistance to SHH inhibitor therapy	Mus Musculus	5110	

Previous | Page 1 of 35 | 10 rows | Next

Let’s open the dataset

- Go to https://artyomovlab.wustl.edu/scn/
- Search for 10x
- And click on the dataset

scNavigator: beta

Single-cell Navigator is an open-source project dedicated to processing and visualization of single-cell RNA-seq data

Below we have a large collection of datasets and tools to play with:

- Large collection of automatically processed datasets. We processed almost every scRNA-seq dataset from GEO Omnibus database. We make it available for you in our browser.
- Collection of curated datasets. Curated dataset are those that we process by hand. These will include datasets from Human Cell Atlas (HCA), Tabula Muris and some of the datase
- You can search for cell type specific gene signatures! When we processed all the public scRNA-seq datasets we also calculated all the markers of all the clusters in all these datase you which cluster in which dataset it looks like.
- If you were provided with secret dataset token, you can use it at the very right of this page

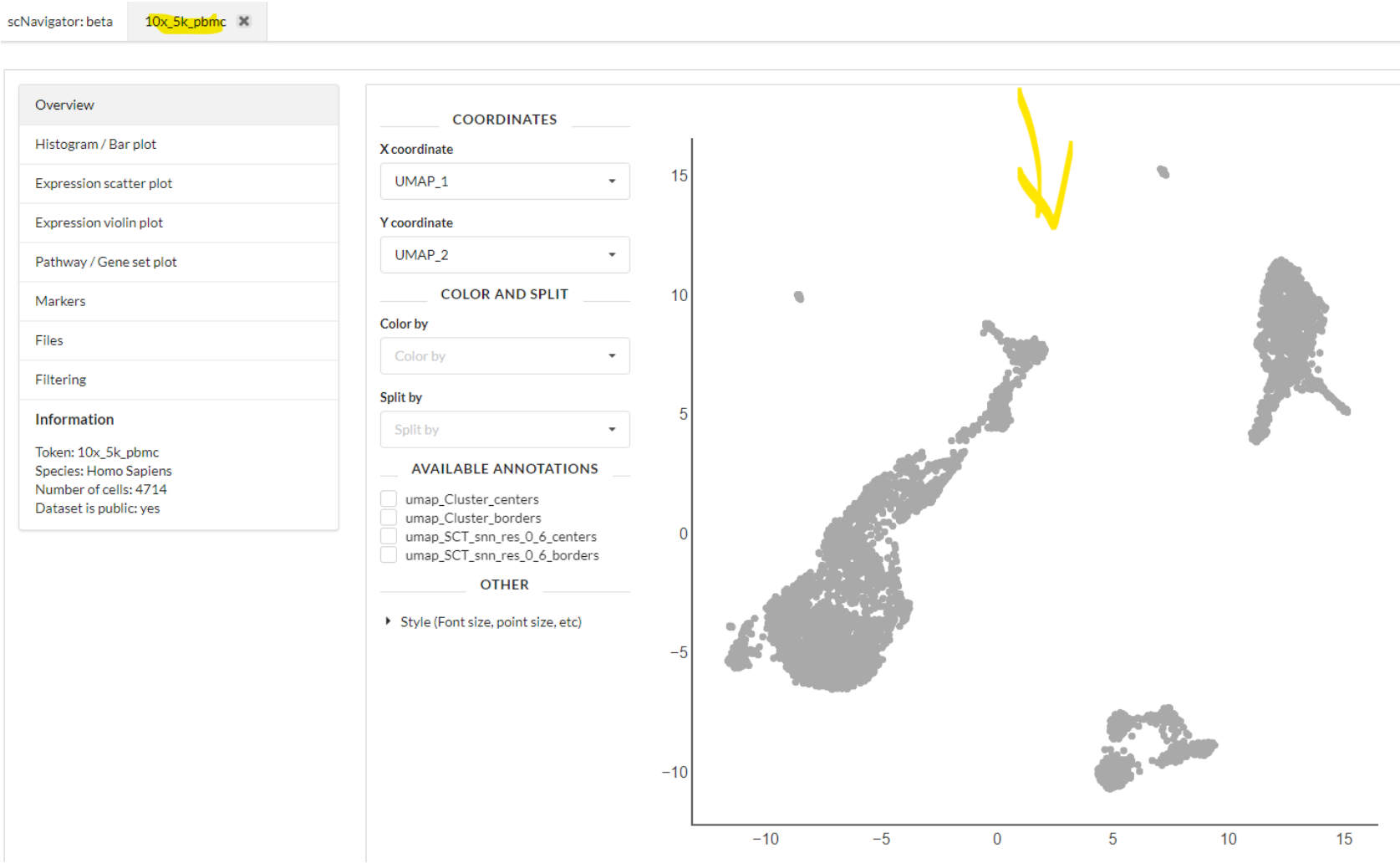
All scRNA-seq datasets | Curated datasets | Gene signature search

Name	Description
10x	
HCA_ColonImmune10XSS2V	Distinct microbial and immune niches of the human colon
tabula_muris_senis_all_cells_	Tabula muris senis: A single-cell transcriptomic atlas characterizes ageing tissues in the mouse [10xv2]
10x_5k_pbmc	5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor (Next GEM)

If you have any problem finding dataset

- Just go to https://artyomovlab.wustl.edu/scn/?token=10x_5k_pbmc

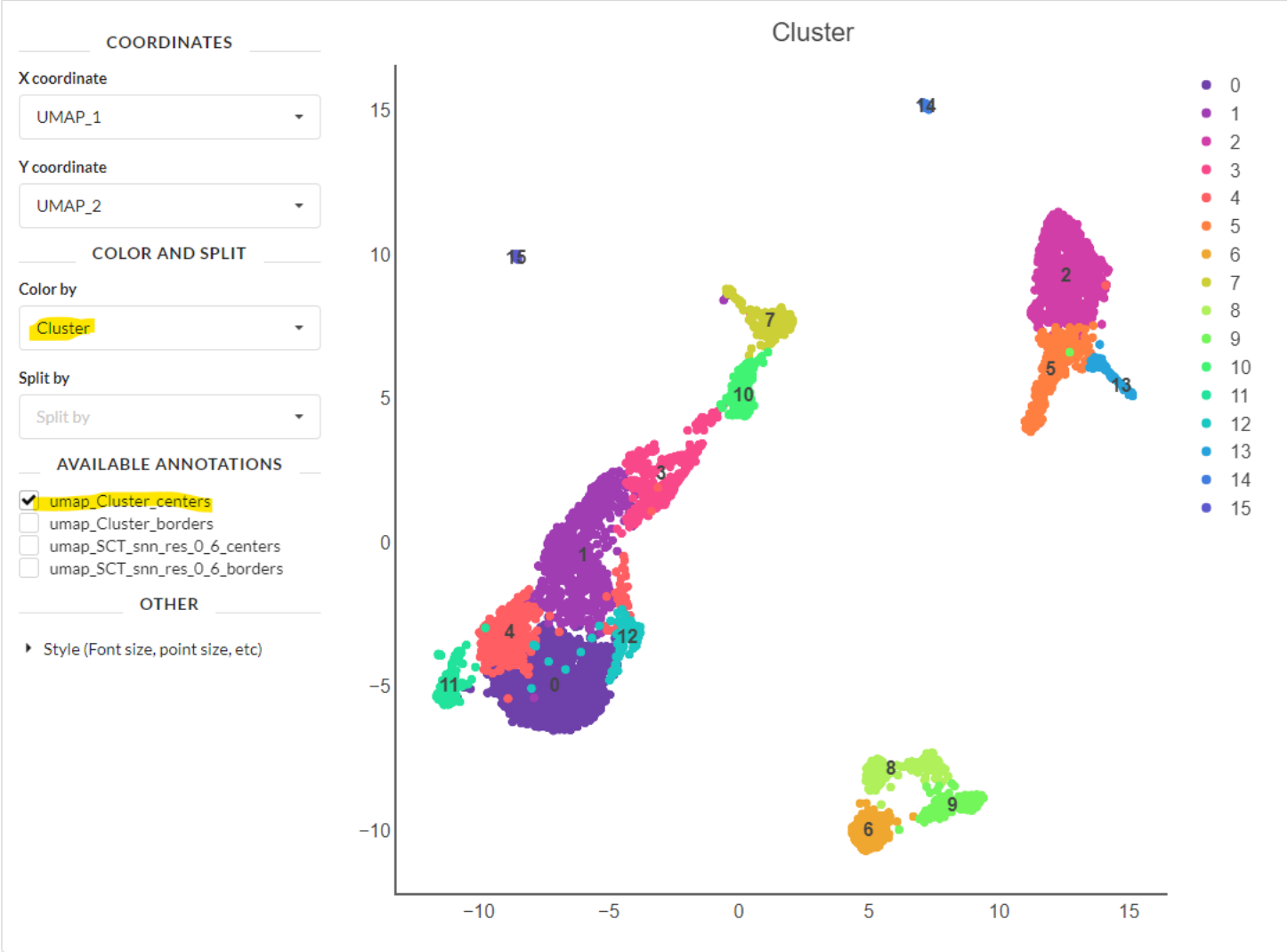
Result should look like that



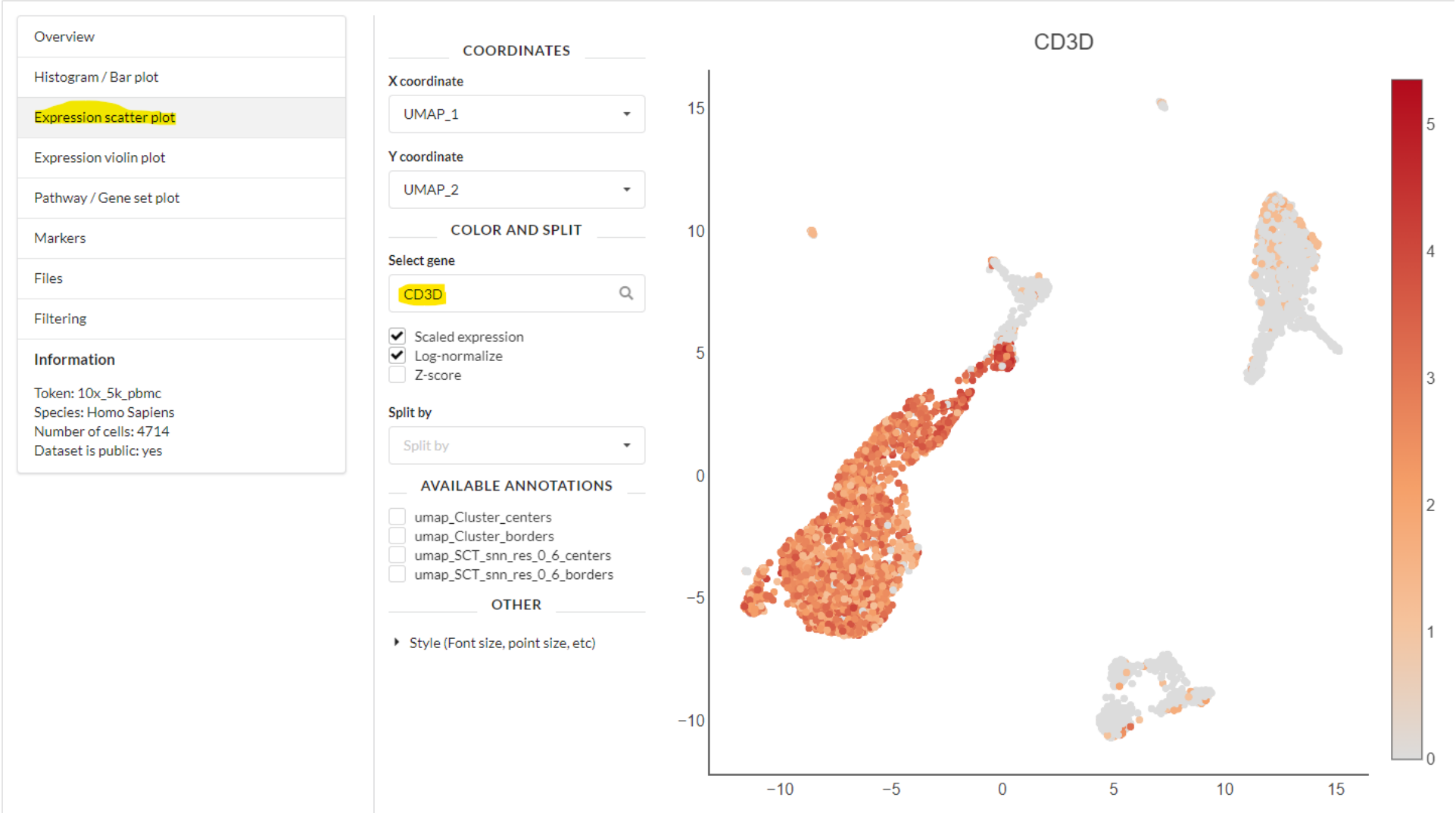
We can color the cells

- Cluster
- Number of UMIs
- Number of genes detected

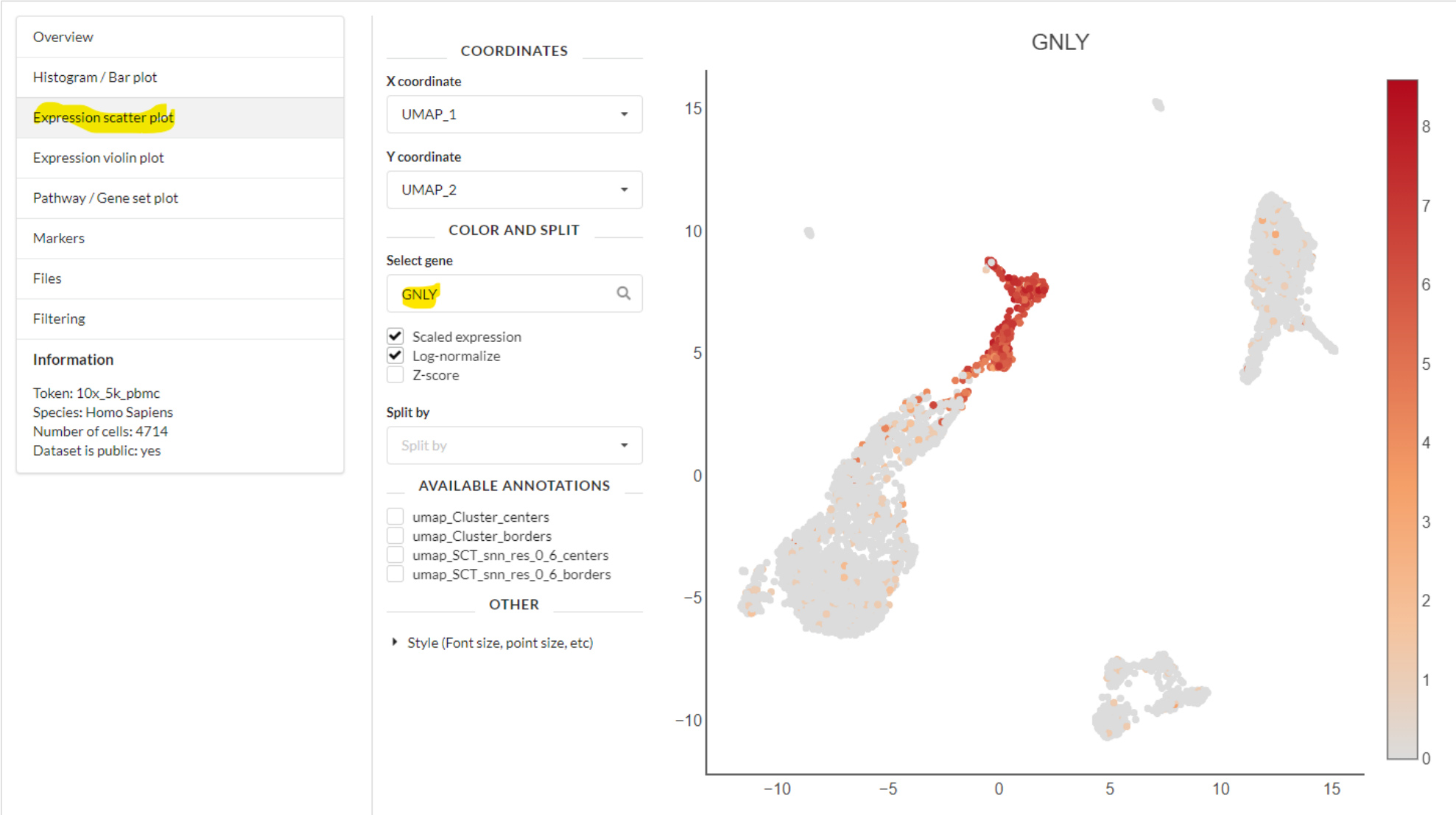
- umap_Cluster_centers



Expression of CD3d



Or you can go for any of your favorite genes



Expression scatter plot

- Expression scatter plot shows gene expression **in each cell**
- We can see that expression of some genes is localized with clusters

Violin plot

Overview

Histogram / Bar plot

Expression scatter plot

Expression violin plot

Pathway / Gene set plot

Markers

Files

Filtering

Information

Token: 10x_5k_pbmc
Species: Homo Sapiens
Number of cells: 4714
Dataset is public: yes

COORDINATES

X coordinate

Cluster

COLOR AND SPLIT

Select gene

CD79A

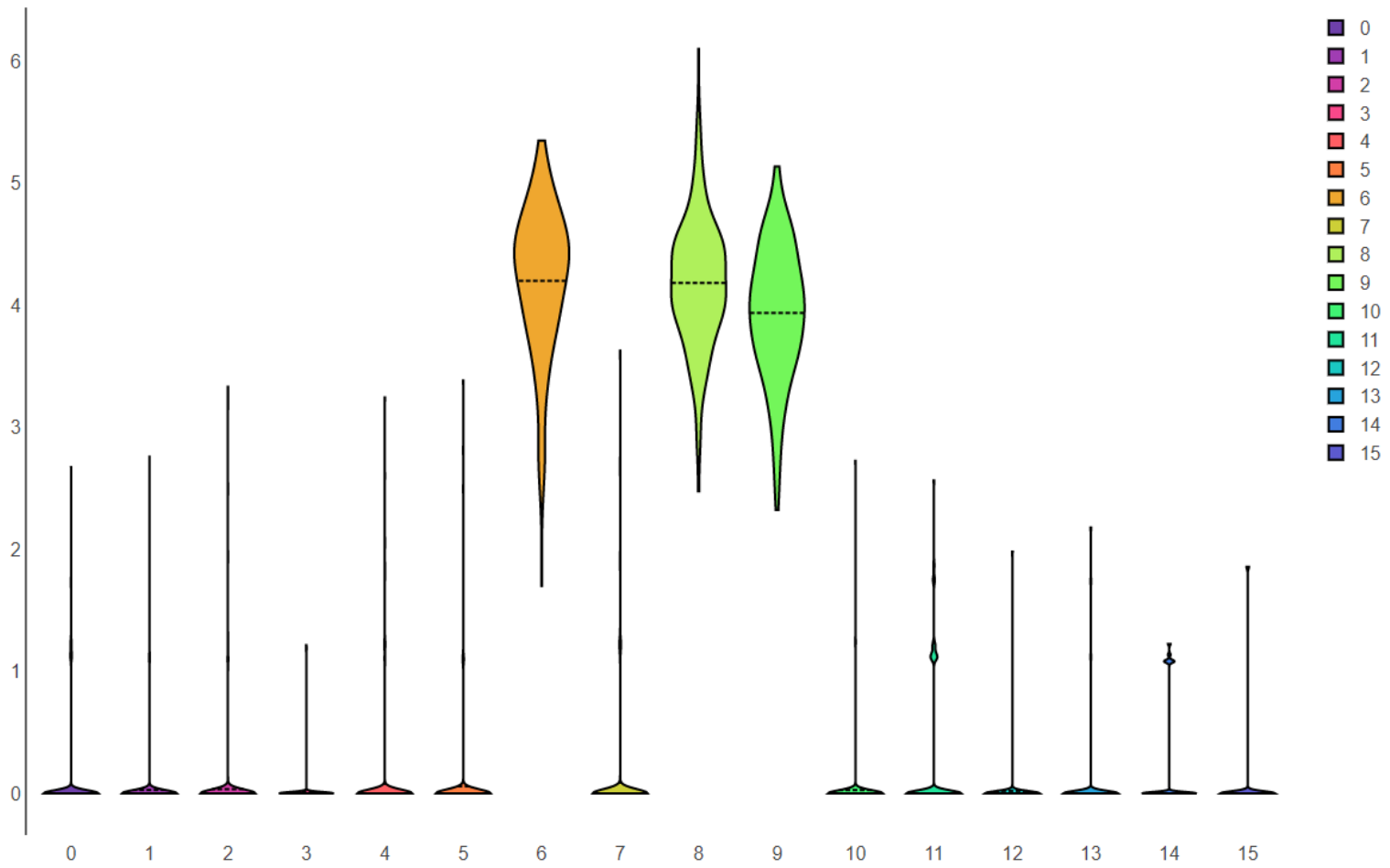
- [x] Scaled expression
- [x] Log-normalize
- [ ] Z-score

Split by

Split by

OTHER

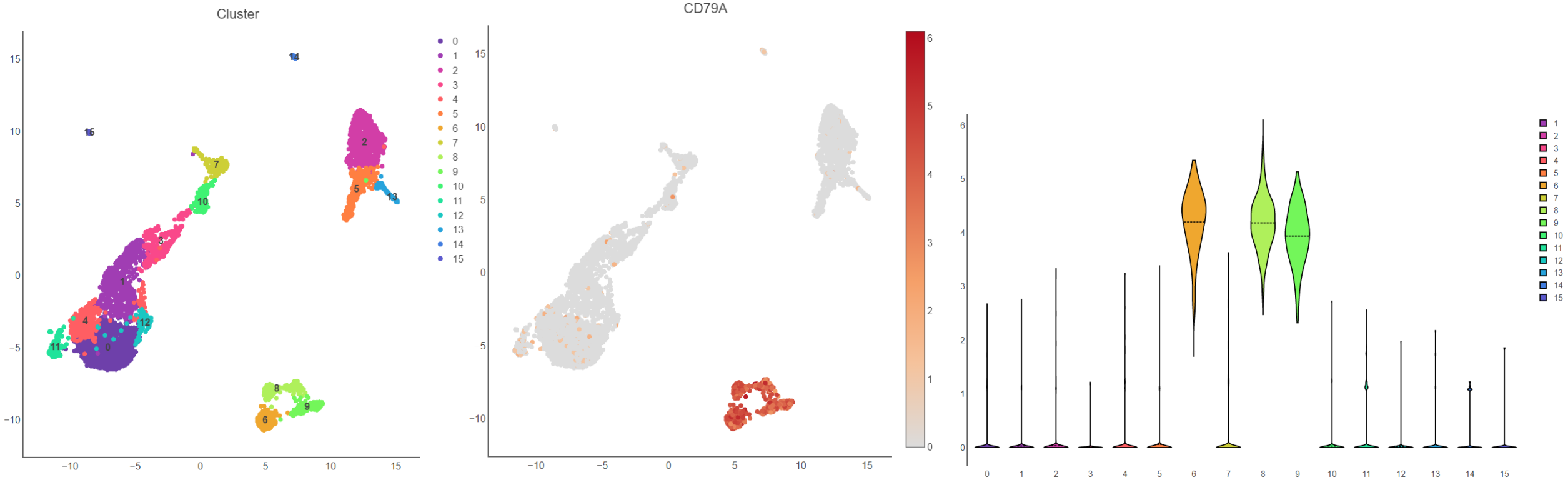
Style (Font size, point size, etc)



Violin plot

- Violin plot shows **distribution** of gene expression within several groups of cells (in our case groups are clusters)
- Higher the violin – higher the expression in the group

Cd79a:
expression scatter and expression violin



Markers

- Usually we run differential expression to identify cluster markers
- You can compare a cluster against all the other clusters and identify genes that have higher expression than in the other clusters

Markers tab

- Overview
- Histogram / Bar plot
- Expression scatter plot
- Expression violin plot
- Pathway / Gene set plot
- Markers
- Files
- Filtering

Information

Token: 10x_5k_pbmc
Species: Homo Sapiens
Number of cells: 4714
Dataset is public: yes

Choose the table

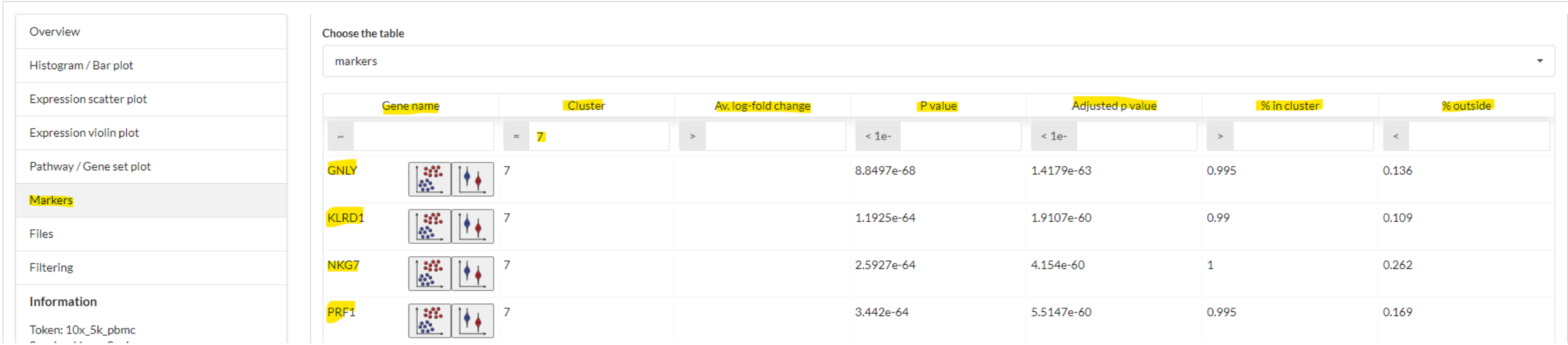
markers

Gene name	Cluster	Av. log-fold change	P value	Adjusted p value	% in cluster	% outside
~	=	>	< 1e-	< 1e-	>	<
RPL30	0		2.3882e-45	3.8263e-41	1	0.999
RPL32	0		6.5972e-44	1.057e-39	1	0.999
RPL22	0		2.3363e-43	3.7432e-39	1	0.999
RPS3A	0		1.9711e-42	3.1581e-38	1	0.999
RPS25	0		4.0714e-42	6.5231e-38	1	0.999
RPS15A	0		7.755e-42	1.2425e-37	1	1
TPT1	0		1.261e-41	2.0204e-37	1	1
RPL11	0		3.99e-41	6.3928e-37	1	0.999
RPL34	0		5.7283e-41	9.1779e-37	1	0.999
RPS14	0		1.5747e-40	2.523e-36	1	0.999

Previous Page 1 of 987 10 rows Next

Download current table

Markers tab: what’s the cluster 7?



- GNLY – gene name
- Cluster 7 – we are checking results for cluster 7 vs other clusters
- Average log-fold change: average difference between expression of GNLY in cluster 7 and in other clusters
- P value (we test difference between average expression of this gene inside and outside cluster 7)
- P adjusted – adjusted p value for multiple hypothesis
- % in and outside of the cluster – in how many cells GNLY is detected in cluster 7 and in other clusters

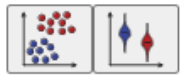
Markers tab: what’s the cluster 7?

- You have two buttons next to the gene name

1) First will open gene expression on scatter plot
2) Second will open gene expression on violin plot

Choose the table

markers

Gene name	Cluster	Av. log-fold change	P value	Adjusted p value	% in cluster	% outside
~	= 7	>	< 1e-	< 1e-	>	<
GNLY 	7		8.8497e-68	1.4179e-63	0.995	0.136
KLRD1 	7		1.1925e-64	1.9107e-60	0.99	0.109

Now let's play with it

- I want you to check out any other genes

Public datasets

- We try to process many other public datasets trying to make them available to scientific community
- You can always go back to the main tab (top left corner)

Public datasets

scNavigator: beta | 10x_5k_pbmc

scNavigator: beta

Single-cell Navigator is an open-source project dedicated to processing and visualization of single-cell RNA-seq data

Below we have a large collection of datasets and tools to play with:

- Large collection of automatically processed datasets. We processed almost every scRNA-seq dataset from GEO Omnibus database. We make it available for you in our browser.
- Collection of curated datasets. Curated dataset are those that we process by hand. These will include datasets from Human Cell Atlas (HCA), Tabula Muris and some of the datasets that we generated in our lab.
- You can search for cell type specific gene signatures! When we processed all the public scRNA-seq datasets we also calculated all the markers of all the clusters in all these datasets. Just put a list of genes and we will tell you which cluster in which dataset it looks like.
- If you were provided with secret dataset token, you can use it at the very right of this page

Enter a secret token below:

Secret token

Go!

All scRNA-seq datasets | Curated datasets | Gene signature search

Name	Description	Organism	# of cells	Exte...
GSE101901/SRS2384613	Single cell sequencing of hippocampus tissues in traumatic brain injury	Mus Musculus	8878	
GSE103976/SRS2523512	Detecting Activated Cell Populations Using Single-Cell RNA-Seq	Mus Musculus	6488	
GSE129730/SRS4617144	Single cell RNA-seq shows cellular heterogeneity and lineage expansion in a mouse model of SHH-driven medulloblastoma support resistance to SHH inhibitor therapy	Mus Musculus	4552	
GSE103983/SRS2523775	Single-cell RNA-seq (Drop-seq) of MGE, CGE and LGE of E13.5 (MGE) and E14.5 (CGE, LGE) mouse embryos	Mus Musculus	11704	
GSE93374/SRS1913127	A Molecular Census of Arcuate Hypothalamus and Median Eminence Cell Types	Mus Musculus	61225	
GSE103983/SRS2523784	Single-cell RNA-seq (Drop-seq) of MGE, CGE and LGE of E13.5 (MGE) and E14.5 (CGE, LGE) mouse embryos	Mus Musculus	709	
GSE137007/SRS5355828	Proliferation-competent Tcf1+ CD8 T-cells in dysfunctional populations are CD4 T-cell help independent	Mus Musculus	434	
GSE106960/SRS2690039	The single cell RNA seq of pulmonary alveolar epithelial cells	Mus Musculus	2683	
GSE113111/SRS3165512	sc-RNA sequencing of skeletal muscle macrophages during T. gondii infection and injury	Mus Musculus	6625	
GSE129730/SRS4617149	Single cell RNA-seq shows cellular heterogeneity and lineage expansion in a mouse model of SHH-driven medulloblastoma support resistance to SHH inhibitor therapy	Mus Musculus	5110	

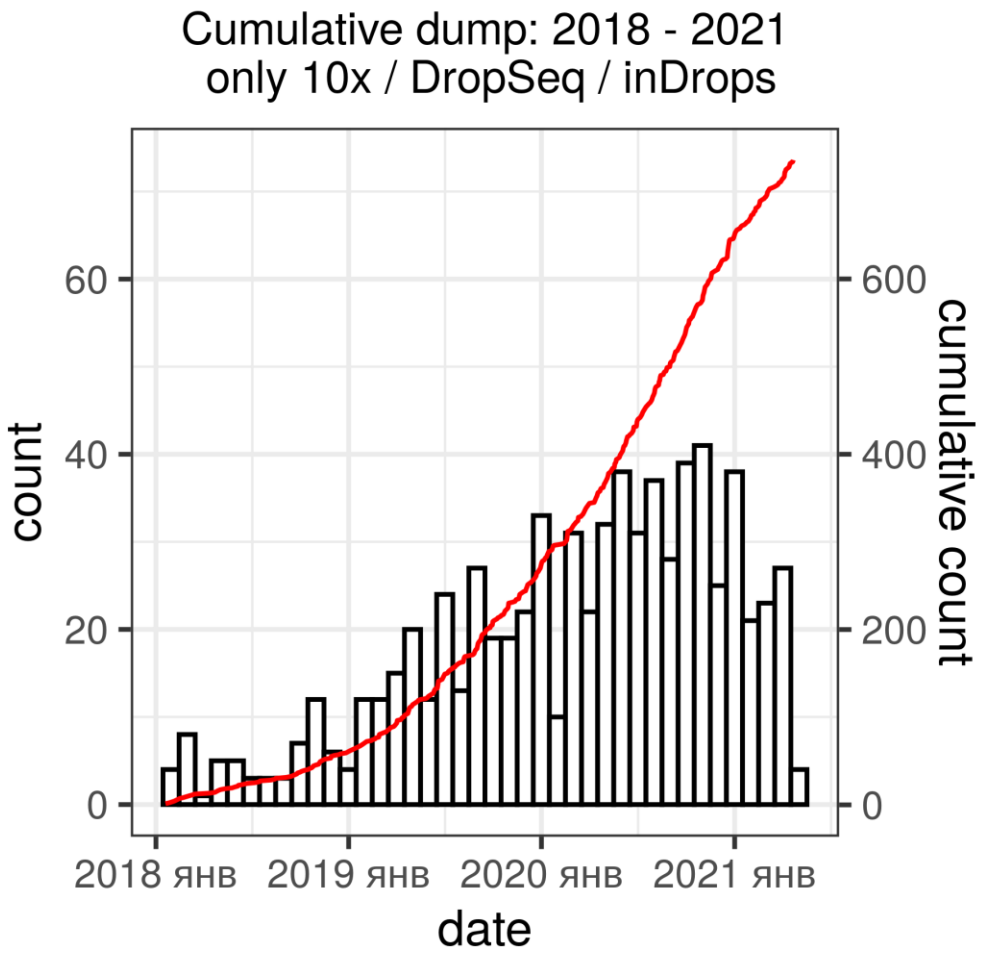
Previous | Page 1 of 35 | 10 rows | Next

Public datasets

- Two main sources
  - NCBI GEO (Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/)
  - EMBL EBI (European Bioinformatics Institute, part of EMBL, https://www.ebi.ac.uk/)

Increasing number of public datasets

- Number of public GSE dataset which were qualified as scRNA-seq
- Histogram and cumulative count



Public scRNA-seq datasets

Most of the scRNA-seq datasets are available at NCBI GEO (or SRA)

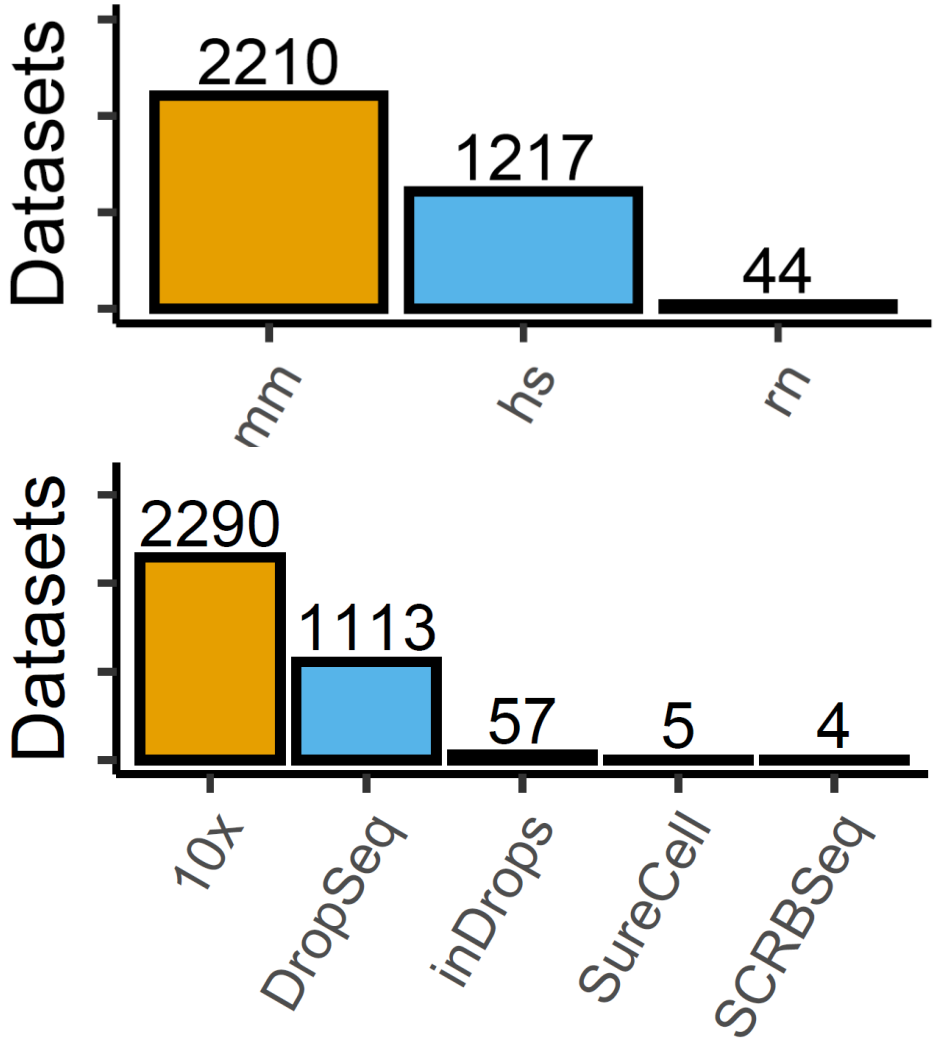
Problems are:

- Different technologies used to perform experiment (10x, DropSeq, SmartSeq2, C1 Fluidigm etc)
- Different pipelines were used to analyze
- Different formats in which data is kept

Most of the dataset processing was done by Maria Firuleva

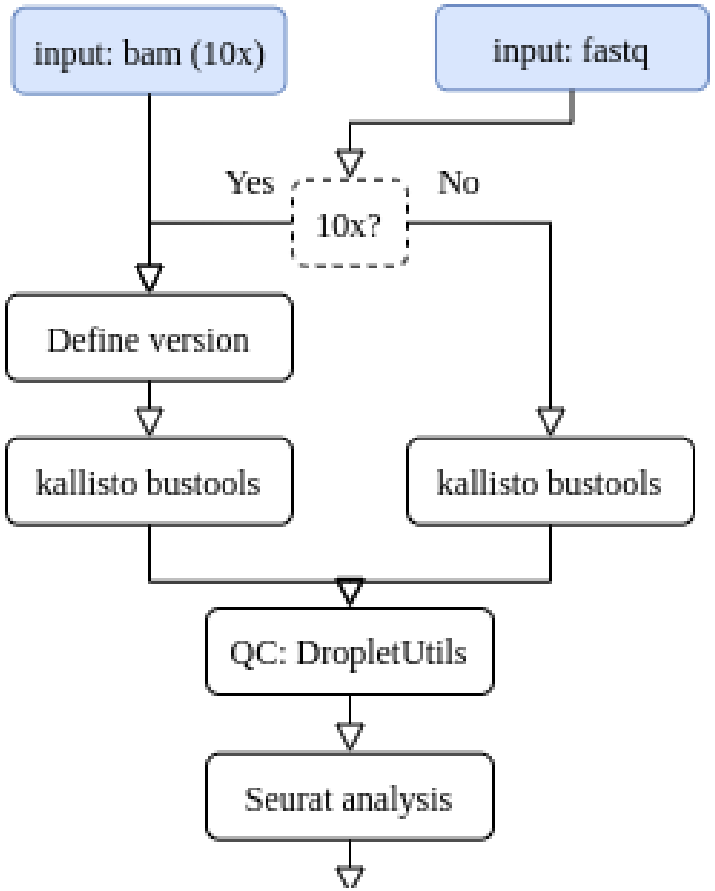
Current database snapshot

- Total ~3500 single-cell samples (as of two weeks ago)
- Some merged GSEs are available
- We will troubleshoot unprocessed datasets
- We tested the pipeline for bulk-like scRNA-seq dataset, will start those soon



How do we process single-cell RNA-seq

- Determine chemistry version
- Kallisto Bustools (from both reads and bams)
- EmptyDrops to remove noise
- Seurat analysis
- SCNPrep



Seurat analysis

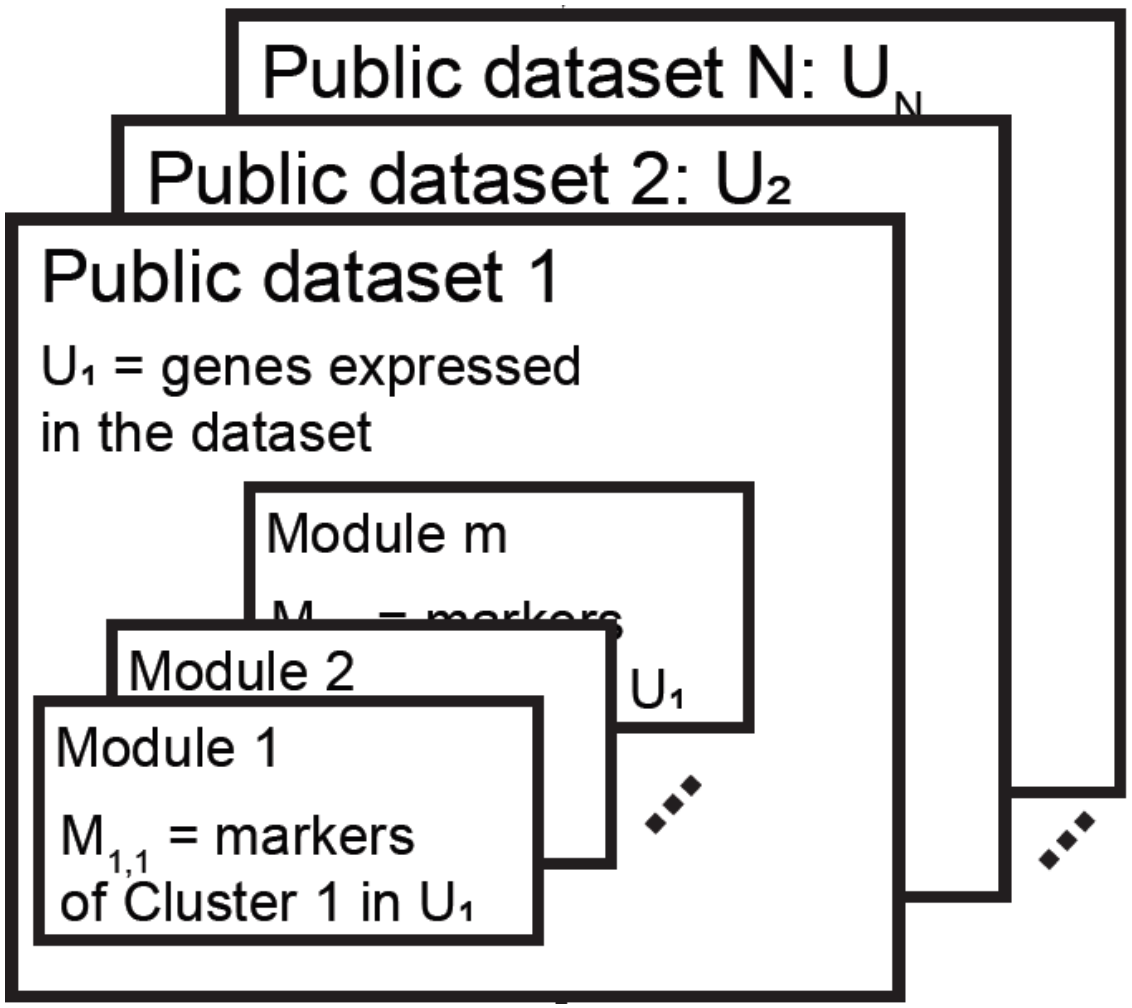
Seurat is an R package for analysis of single-cell RNA-seq data

- Some more QC: removing cells with high mito-content
- Normalization
- (if dataset consists of multiple samples) merging samples together
- PCA
- Dimensionality reductions: both tSNE and UMAP
- Clustering
- Markers identification

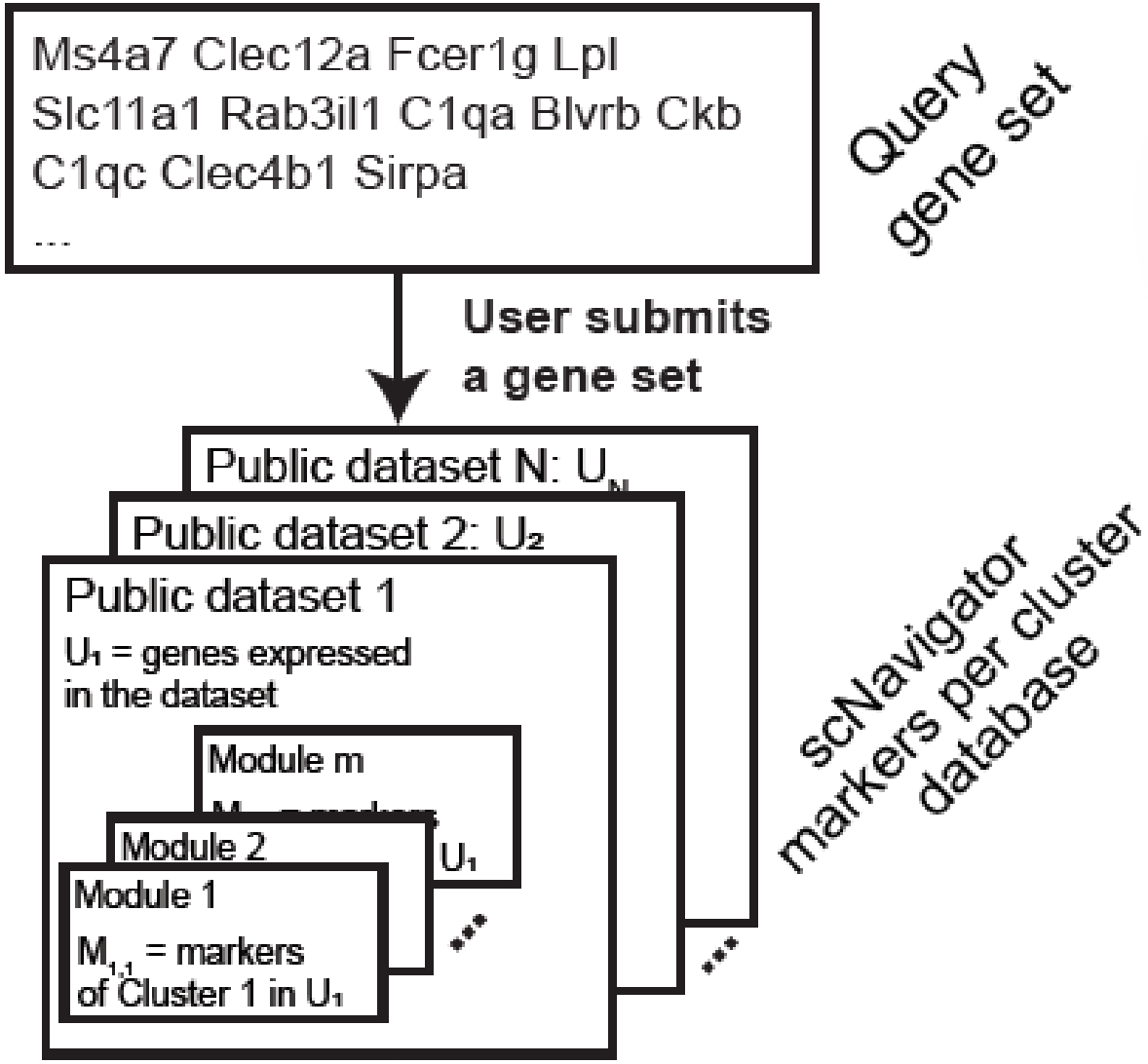
GeneQuery inspired gene signature search

- Having this large database of public scRNA-seq datasets we wanted to implement gene signature search
- Given with a list of gene, we can match it against markers of all the clusters present in our database

Database of all the markers in all the single-cell clusters

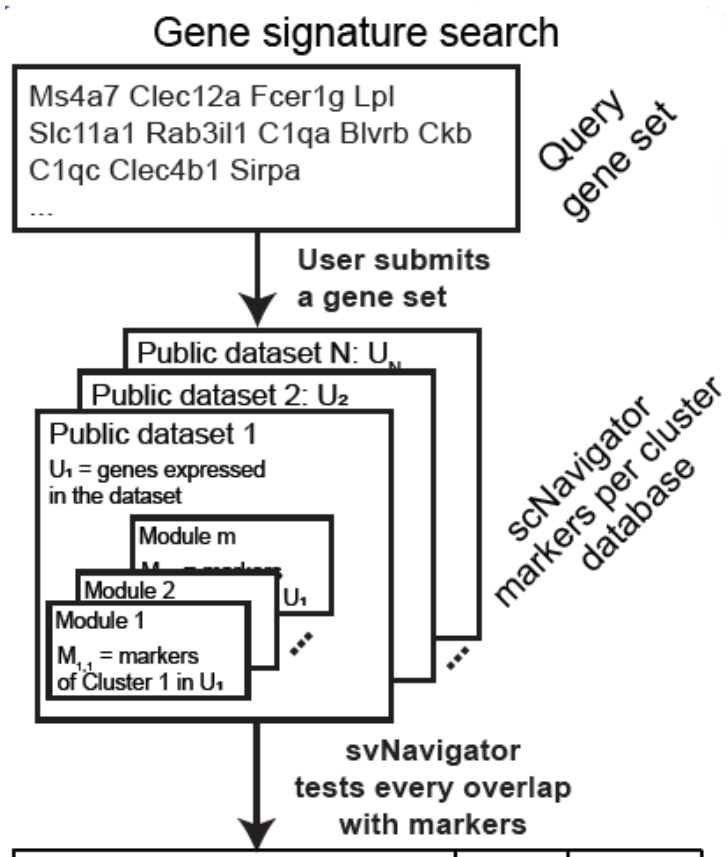


Gene signature search



Gene Signature Search

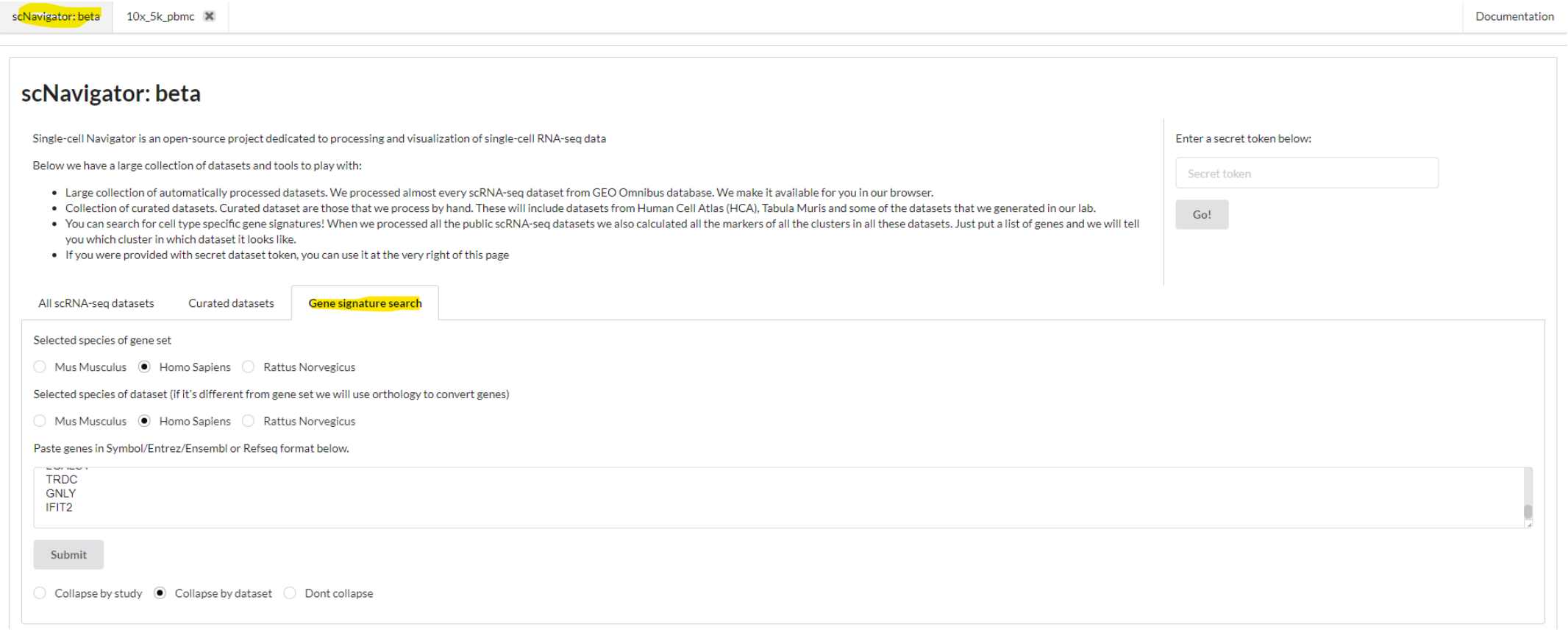
- User submits a gene set
- We compare gene set against markers of all the populations
- Find datasets and clusters where signature is expressed



Dataset	Cluster	adj. p value
10X single-cell RNA Seq profiling of immune infiltrate from Adar1 KO and Wildtype B16 Melanoma	1	9.62E-52
A single-cell atlas of mouse brain macrophages reveals unique transcriptional identities shaped by ontogeny and tissue environment.	3	6.10E-37
Single cell RNA sequencing analysis of mouse resident and infiltrating immune cells during acute kidney injury and fibrosis	1	8.31E-36
Single cell sequencing of mouse syngeneic tumor models	5	1.58E-31
...	...	...

Example: top 50 genes

- You can select top 50 genes



Example: top 50 genes

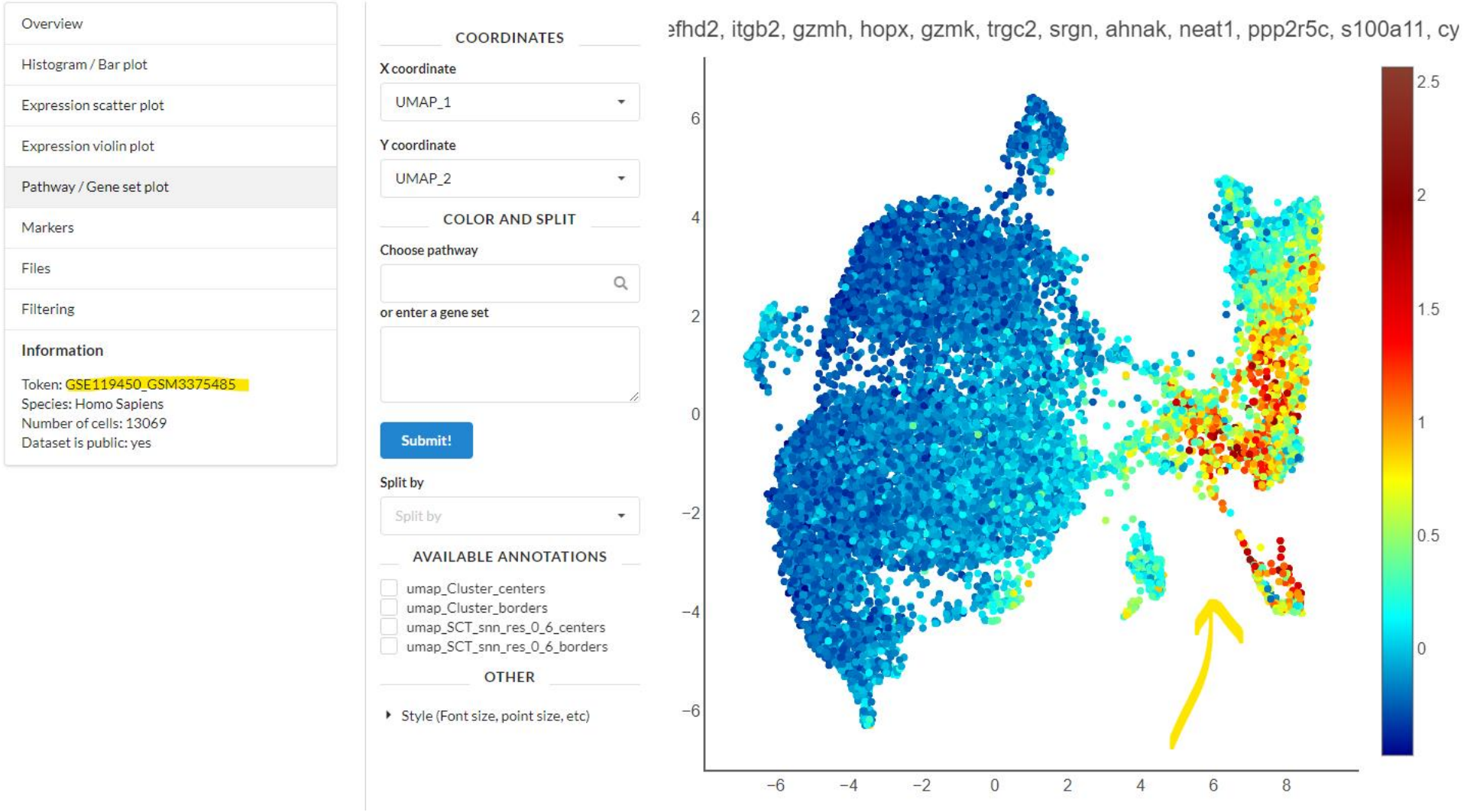
- We get a lot of results

Name	Show Enrichment	Exte...	Title	Adjusted p v...	Mod...	Inte...
~			~	< 1e-		
GSE120221_GSM3396181	Show enrichment		Human Bone Marrow Assessment by Single Cell RNA Sequencing, Mass Cytometry and Flow Cytometry [scRNA]	6.20e-65	80	33
GSE119450_GSM3375485	Show enrichment		CROP-Seq in Primary Human T Cells	8.13e-62	72	31
GSE128223_GSM3667471	Show enrichment		Single cell RNAseq of human TCRVdelta 1 and TCRVdelta 2 gammadelta T lymphocytes purified from healthy adults blood	9.74e-62	138	36
GSE128066_GSM3665016	Show enrichment		A Bayesian mixture model for clustering droplet-based single cell transcriptomic data from population studies	1.43e-57	134	34
GSE153765_GSM4653906	Show enrichment		Targeting CD38 with Daratumumab in refractory Systemic Lupus Erythematosus	8.80e-57	121	33
GSE157829_GSM4775590	Show enrichment		An Atlas of Immune Cell Exhaustion in HIV-Infected Individuals Revealed by Single-Cell Transcriptomics	7.56e-56	82	30
GSE128066_GSM3665019	Show enrichment		A Bayesian mixture model for clustering droplet-based single cell transcriptomic data from population studies	9.41e-56	66	29
GSE157829_GSM4775594	Show enrichment		An Atlas of Immune Cell Exhaustion in HIV-Infected Individuals Revealed by Single-Cell Transcriptomics	2.44e-55	60	28
GSE130157_GSM3733113	Show enrichment		Signaling state and abundance of circulating immune cell subpopulations determine cancer response to immunotherapy	5.57e-55	117	32
GSE132338_GSM4512385	Show enrichment		Characterization of the immunologic impact of sarcoidosis in peripheral blood mononuclear cells via single-cell RNA-seq	4.92e-54	77	30

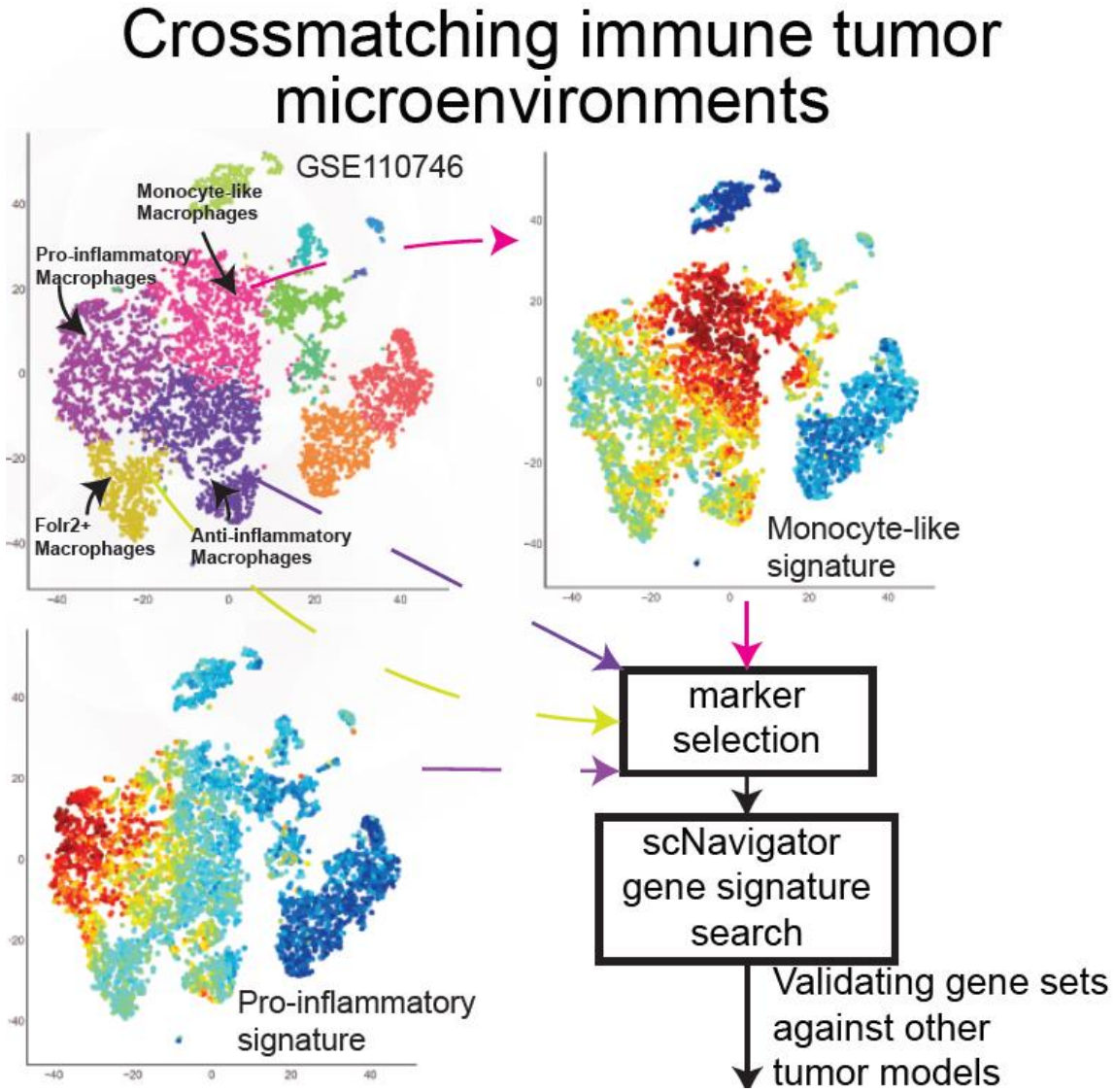
Previous Page 1 of 52 10 rows Next

Press the button

We can see enrichment of these genes in other datasets

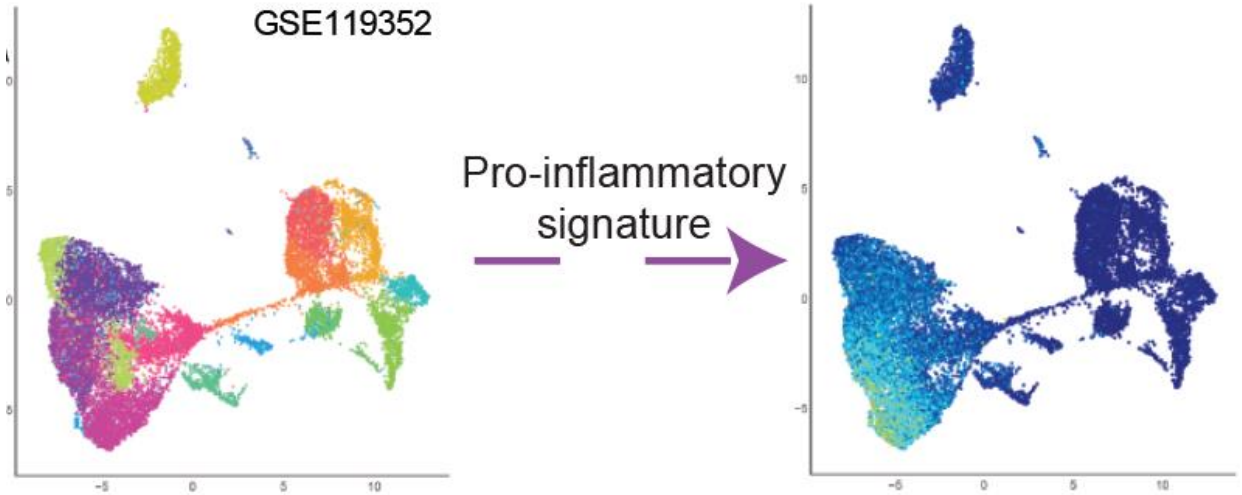


Study case: tumor microenvironments



Signature of macrophage populations were well-conserved in different mouse tumor models

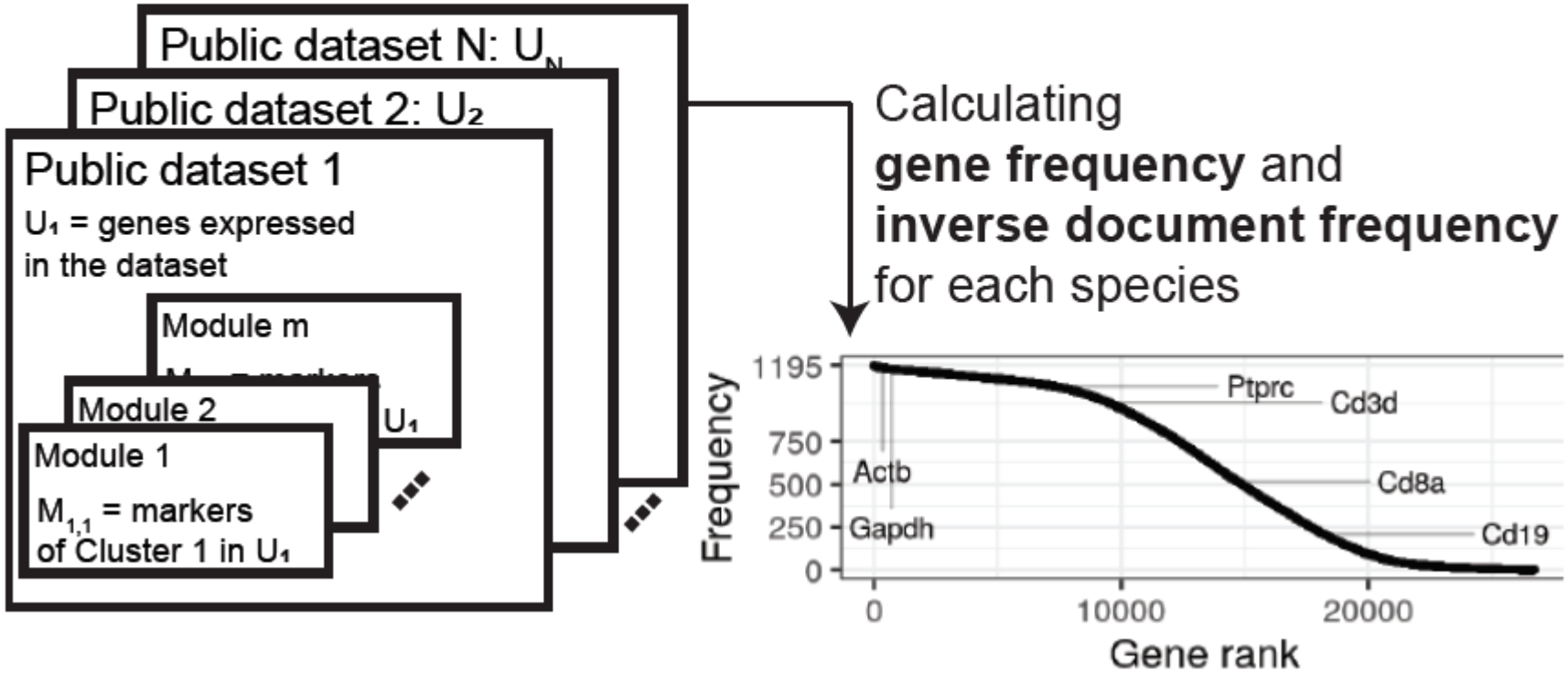
		Anti-inflammatory	Pro-inflammatory	Monocyte-like	Fol2r+
GSE110746	B16 tumor model (this dataset)	7.68E-52	4.36E-48	8.01E-46	2.33E-46
GSE119352	D42m1 tumor model	7.41E-29	3.22E-33	2.41E-45	4.48E-16
GSE112865	MC38 tumor model	2.90E-25	2.33E-18	2.54E-26	1.68E-07
GSE121861	LL2/SA1 tumor models	1.59E-31	1.82E-15	8.56E-32	1.01E-10



What else can we do with it

- We obtained the markers for all the populations and datasets
- We can figure out the similarity between these populations

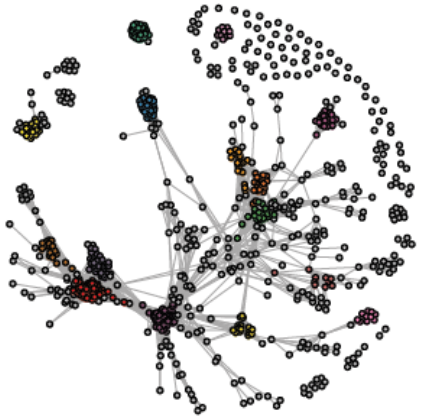
Not all the genes are interesting



Building similarity networks

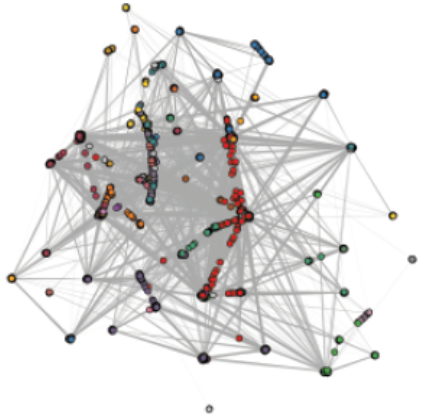
Each dataset and each cluster is an **IDF vector** now

Cosine similarity of dataset IDF vectors



revaling samples that were sorted/taken from same/similar tissue

Cosine similarity of cluster IDF vectors



revaling cell-type specific gene signature across all the datasets

Conclusion

- We hope that single-cell navigator will make interpretation of scRNA-seq data easier
- https://artyomovlab.wustl.edu/scn/

- We try to get there as much datasets as we can
- If you want to use SCN for your private data:
  - You can just e-mail me kzaitsev@itmo.ru and I will give you a private link to your data
  - Wait until it gets published (ETA?), you will be able to host SCN locally, or for your department

Extra gene signatures from the study case

- Anti-inflammatory: Ms4a7 Clec12a Fcer1g Lpl Slc11a1 Rab3il1 C1qa Blvrb Ckb C1qc Clec4b1 Sirpa Fcgr4 Grn Pycard C1qb Adgre1 Ctsc Cd72 Clec4a1 Hexa Aif1 Clec4a2 Lst1 Slamf9 Lgmn AF251705 Nr1h3 Cd300e Ctsb
- Pro-inflammatory: Arg1 Adam8 Ninj1 Mmp12 Basp1 Slc2a1 Hilpda Cstb Il1rn Clec4d Il7r Ndrg1 Hmox1 Ftl1 Cd36 Lgals3 Fabp5 Cxcl2 Plin2 Emp1 Rgcc Bnip3 Egln3 Thbs1 Fth1 Ctsl Spp1 Card19 Ero1l Fabp4
- Monocyte-like: Il1b Gm9733 Btg2 Ccr2 Zbp1 Plbd1 Ifitm3 H2-DMa Ly6i Plac8 Spi1 Osm Ms4a6c Samhd1 Cybb Lyz2 Naaa Fos Ms4a4c H2-DMb1 Hp Prdx5 Junb Cd74 Tgfbi Ly6c2 Slamf8 Klra2 Zfp36 Scimp
- Folr2+: Sepp1 Trf Cd163 Apoe Mrc1 Fxyd2 Fcgrt Igf1 Ccl24 Folr2 Itm2b Igfbp4 F13a1 Ednrb Tmem37 Gas6 Ltc4s Glul Cbr2 C4b Wfdc17 Pltp Lyve1 Cd209f Clec10a Npl Pf4 Timp2 Rnase4 C1qc