ITMO UNIVERSITY

CT Lab
ITMO UNIVERSITY

Analysis of scRNA-seq data

Konstantin “Kostya” Zaitsev, ITMO University

Systems biology workshop, Nice, Sep 22th

Basic steps to analysis of scRNA-seq

- Filtering out “bad” barcodes
- Normalizing expression levels
- Visualization (tSNE plots)
- Clustering
- **Cellular subset annotation**

Annotation

- Cell subset annotation is one of the most important steps
- We can look at several immunological expression markers to identify cell subsets
- Let’s open https://artyomovlab.wustl.edu/sce/?token=PBMC_10k

Immunological markers?

We got plenty of those:

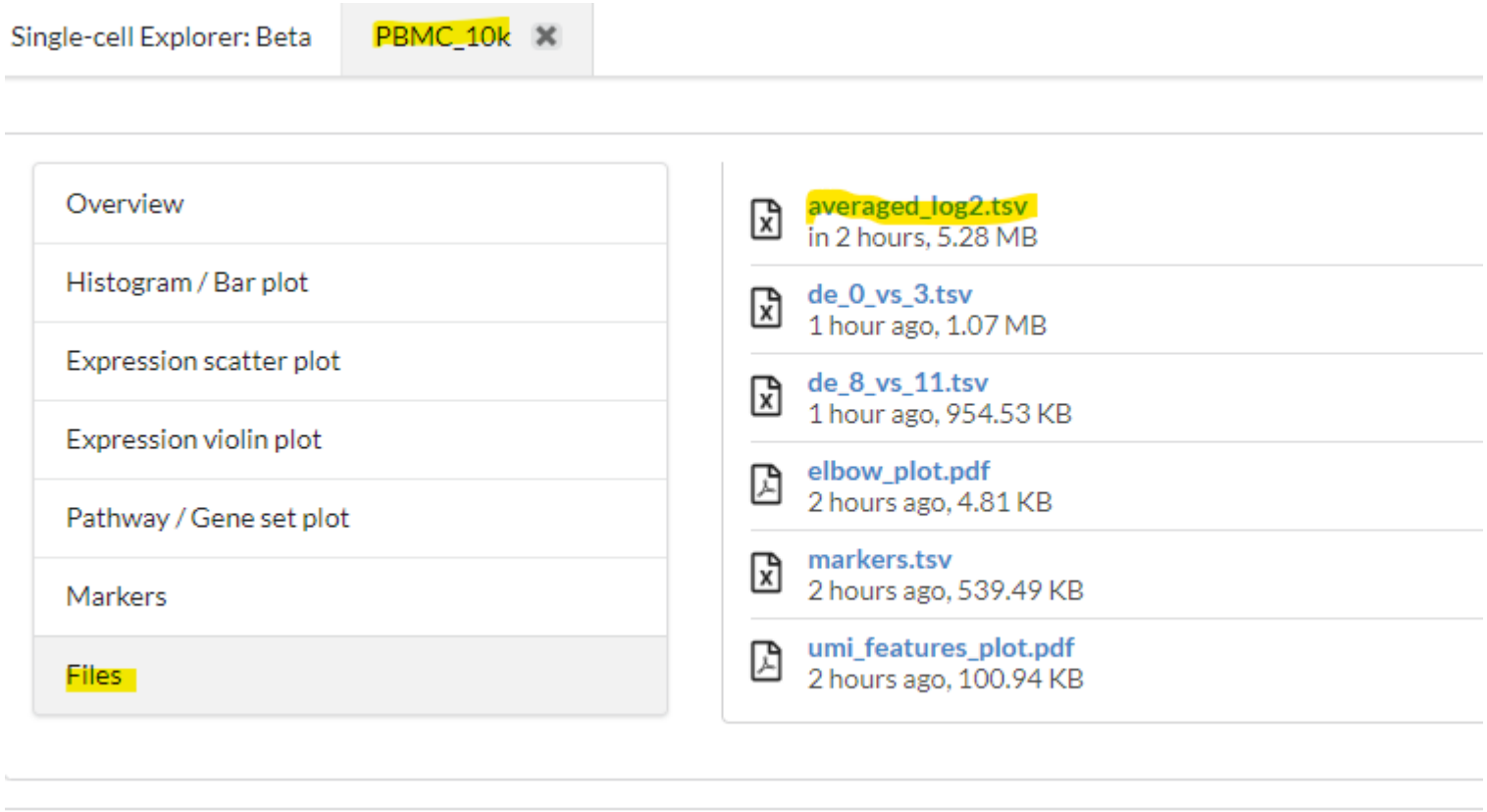
CD19 CD79A CD79B CD14 CD3E GNLY PRF1 FCGR3A SELL CCR7 ITGAX ITGAM HLA-DRA CD8A CD8B CD4 FLT3

Averaged expression

- Sometimes going though all the genes is impractical
- We would like to look at these genes at the same time
- We can average expression of these genes in clusters and use Phantasus to visualize expression of these genes

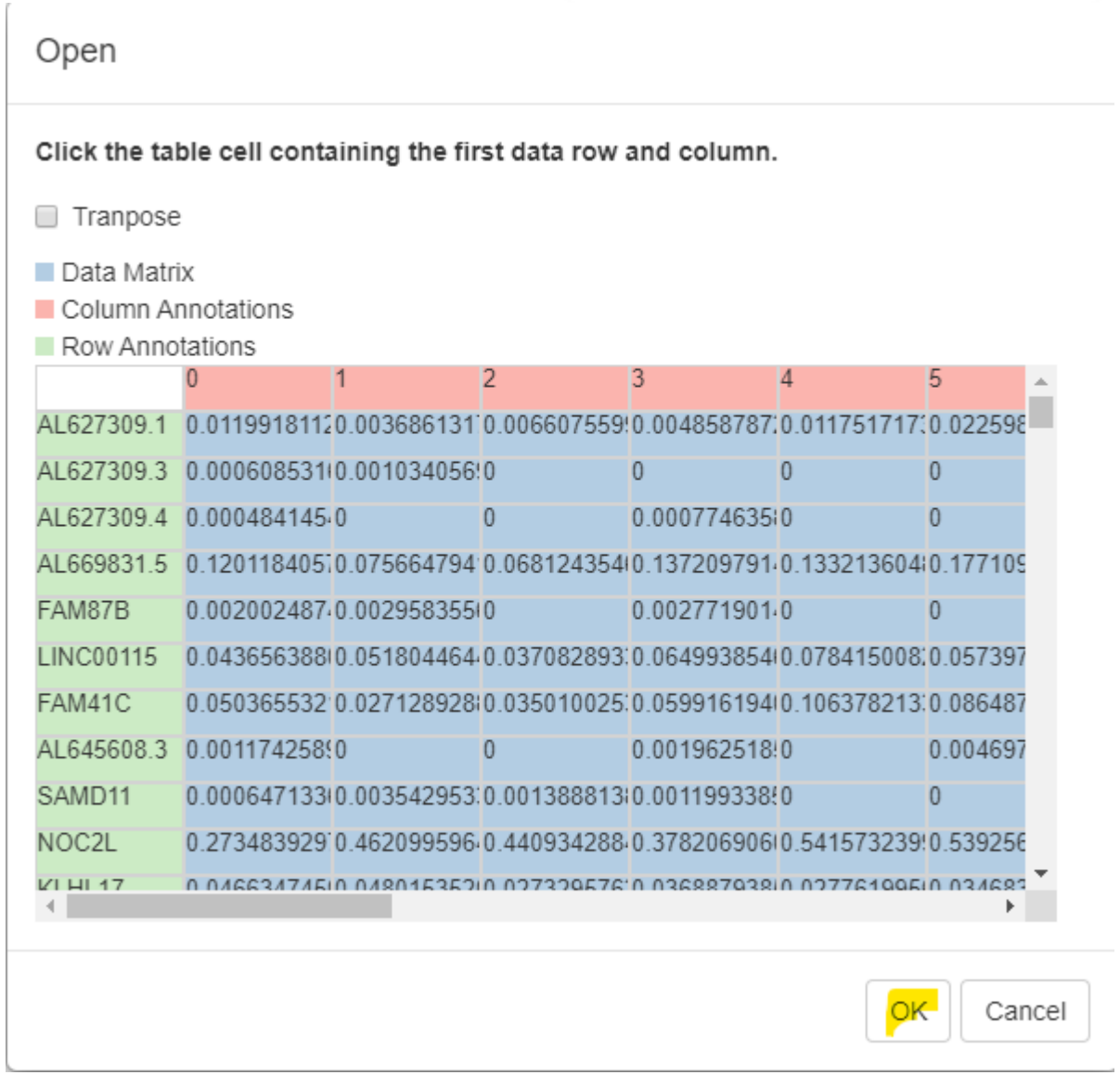
Averaged expression

- Let’s first download the averaged expression table



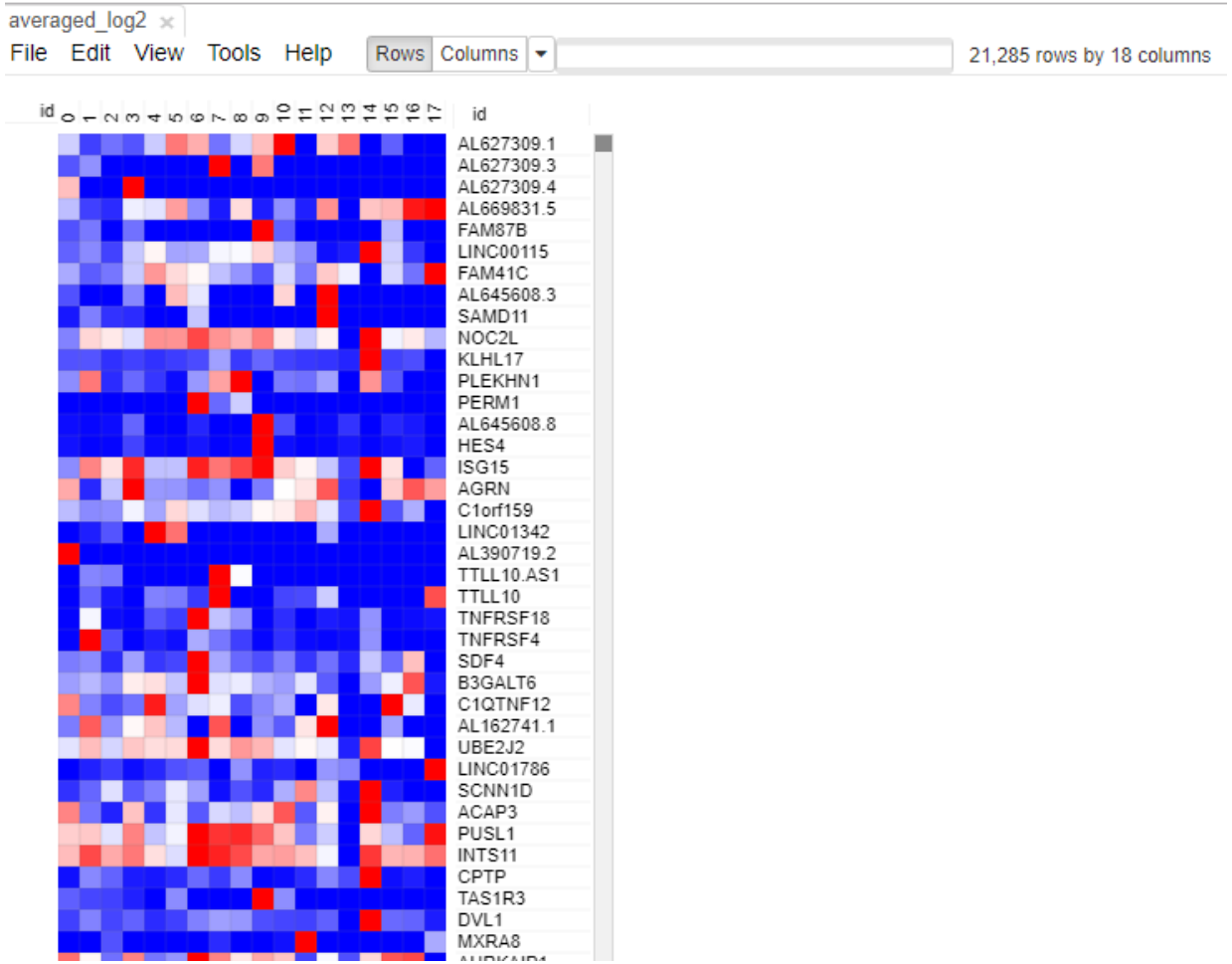
Phantasus

- Let’s open averaged dataset in phantasus (http://ctlab.itmo.ru/phantasus/)



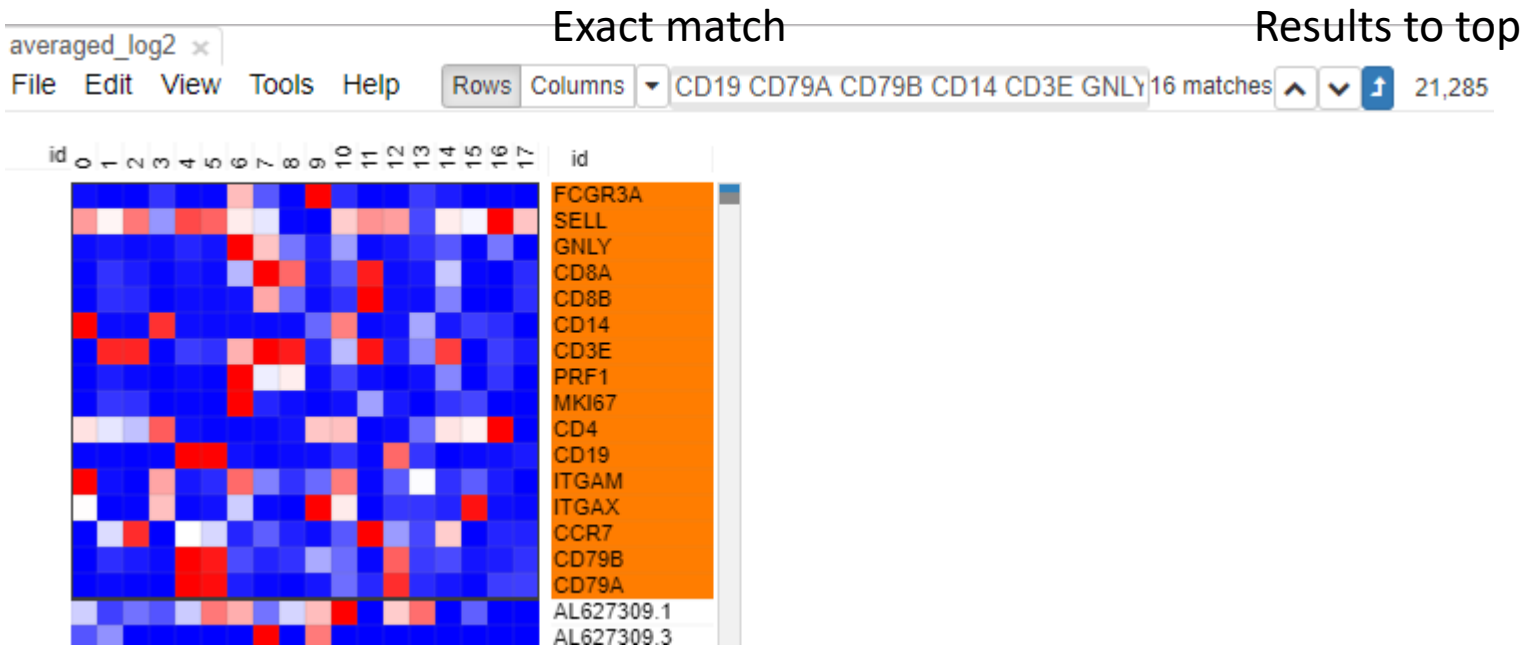
Result should look like that

- 18 columns (clusters)
- 21 285 rows (genes)



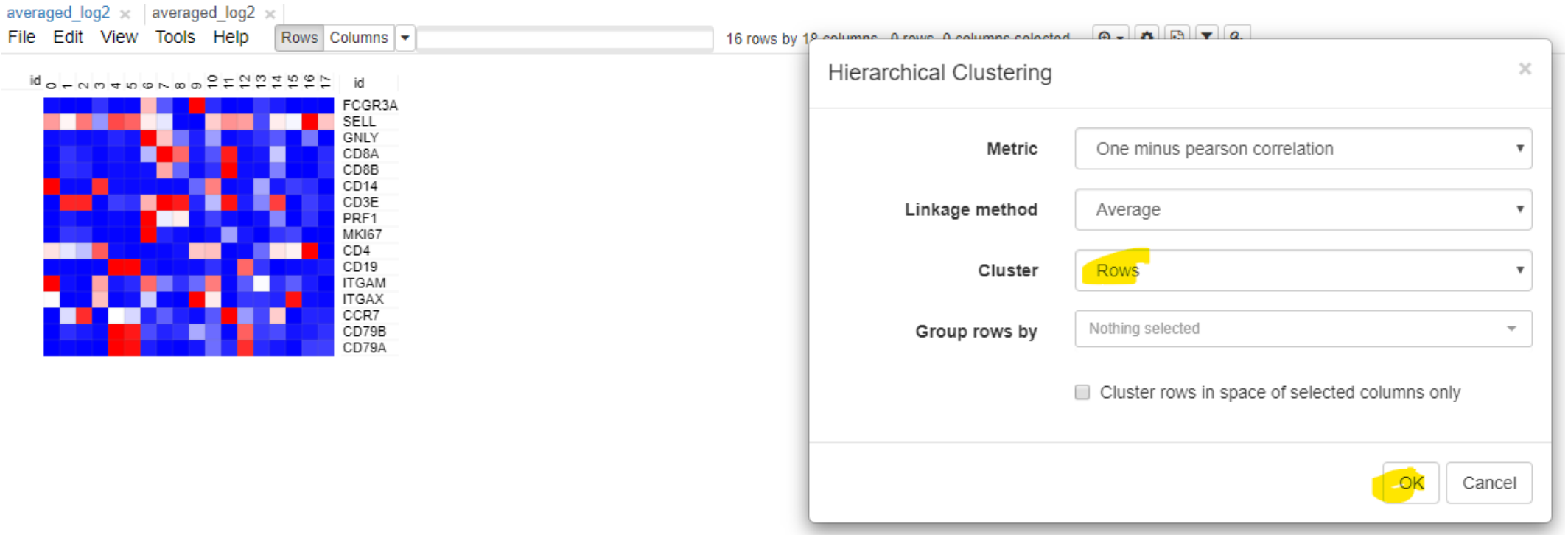
Let’s checkout expression of known markers

- CD19 CD79A CD79B CD14 CD3E GNLY PRF1 FCGR3A SELL CCR7 ITGAX ITGAM HLA-DRA CD8A CD8B CD4 MKI67



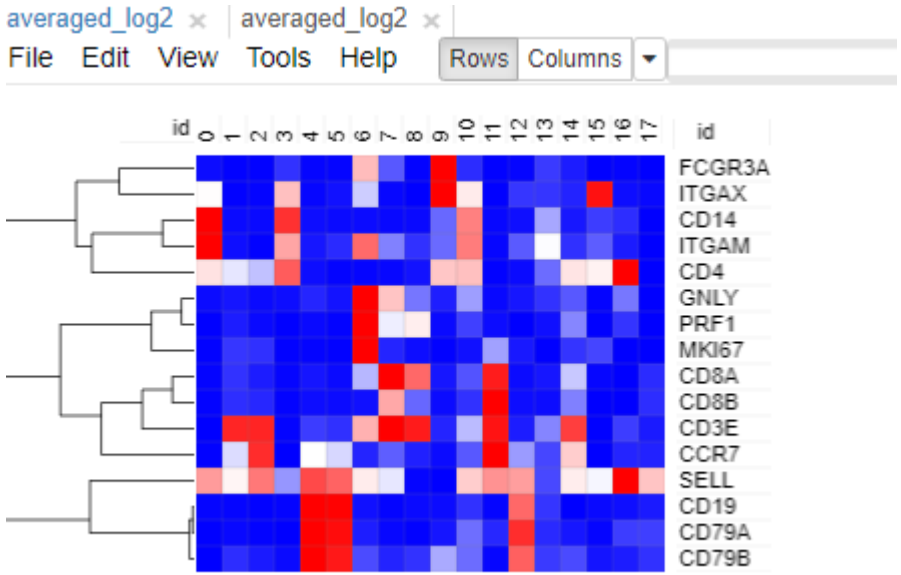
Let’s checkout expression of known markers

- After genes are selected (tools -> new heat map)
- Then tools-> hierarchical clustering -> Cluster (rows)



Let’s checkout expression of known markers

- After genes are selected (tools -> new heat map)
- Then tools-> hierarchical clustering -> Cluster (rows)

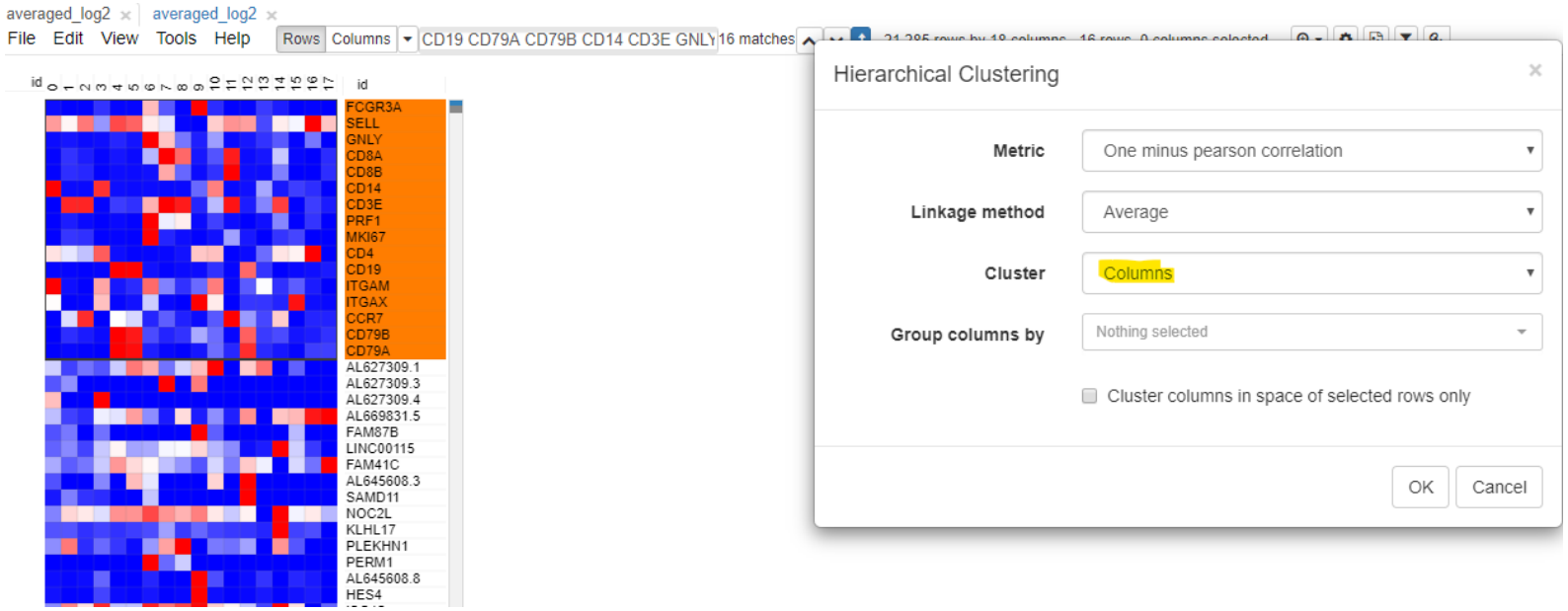


Save this heatmap

- Saving heatmaps is a good thing
- File -> Save Image (Ctrl+S) -> Choose Filename -> Choose format (I prefer svg, svg can be open in browser) -> positive feedback

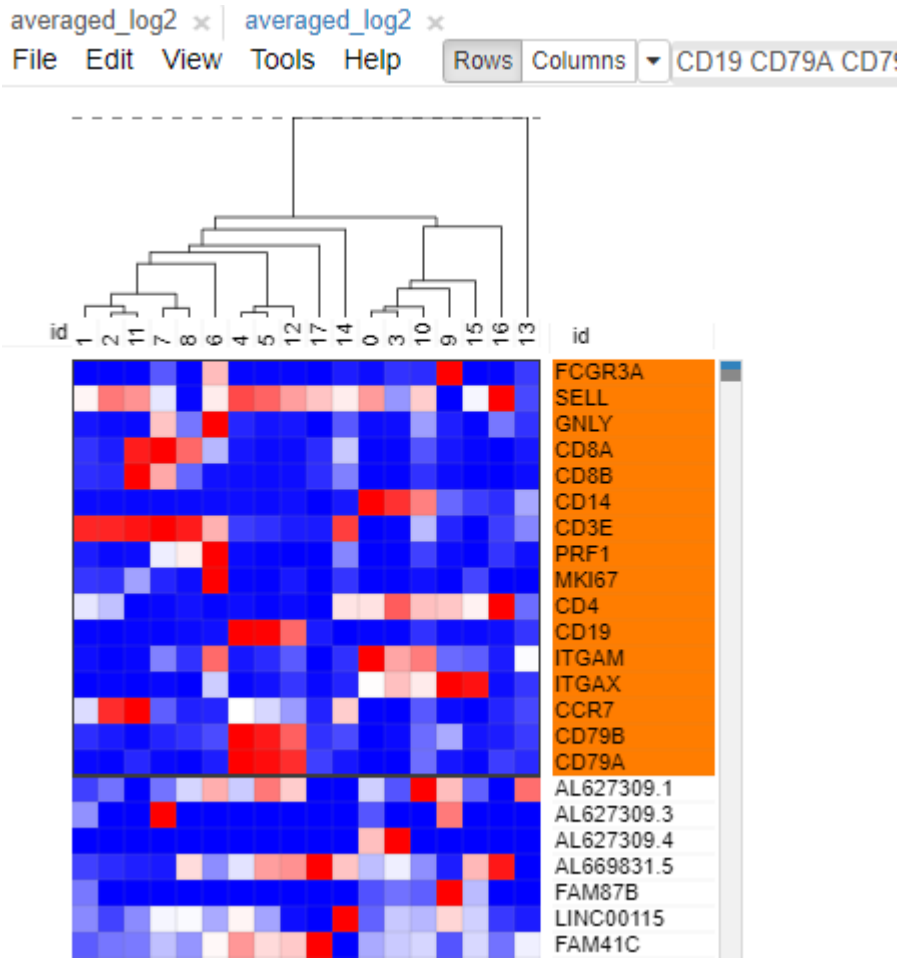
Come back to all gene and cluster columns

- Come back to the tab with all the genes
- Tools -> hierarchical clustering -> Cluster: columns



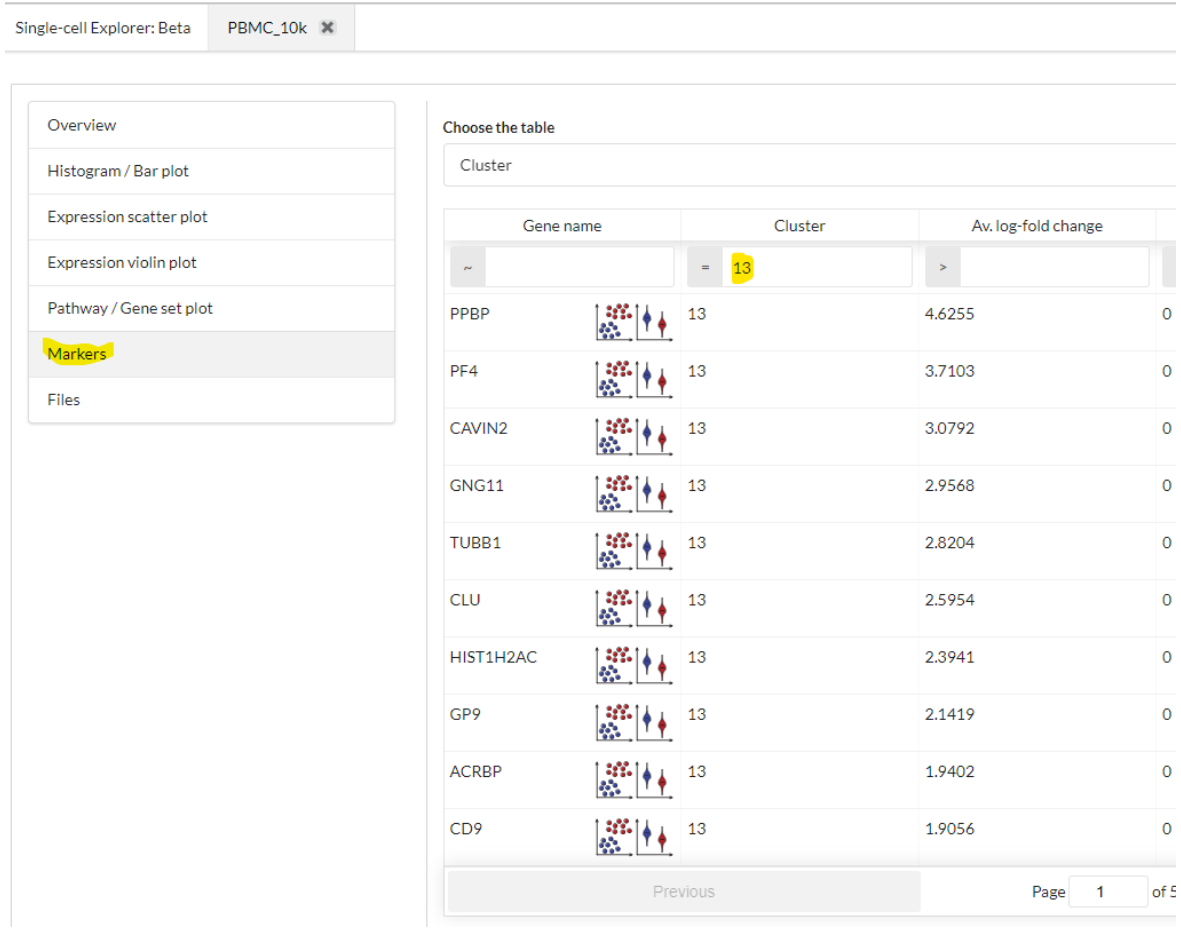
Come back to all gene and cluster columns

- Come back to the tab with all the genes
- Tools -> hierarchical clustering -> Cluster: columns
- Question: what is cluster 13??



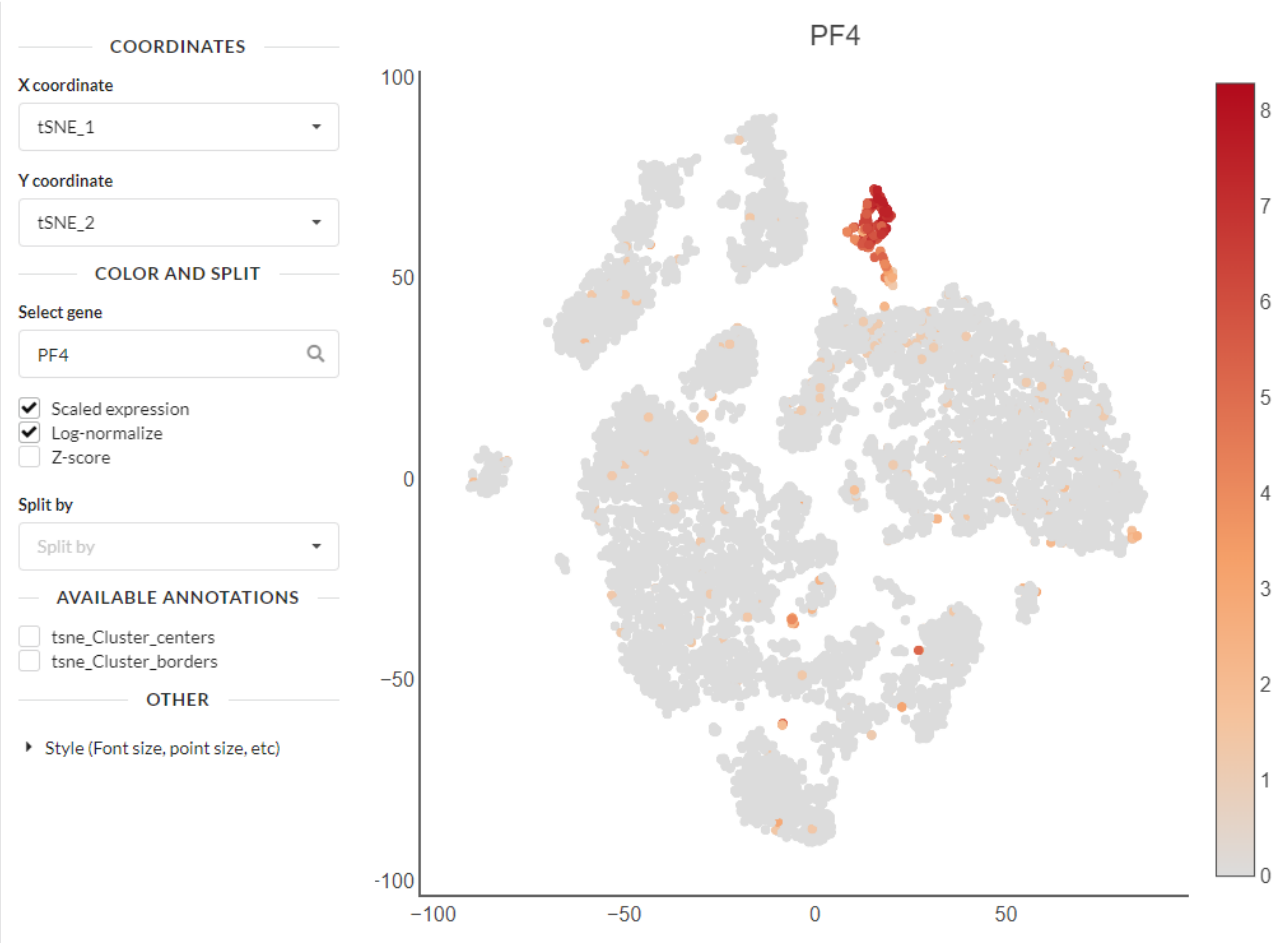
Back to SCE

- Let’s go back to SCE and try to figure out what’s cluster 13
- Open markers tab



Back to SCE

- PF4 is for “platelet factor 4”: most likely just contamination



Summarizing: annotation

- Clusters 1, 2, 11, 7, 8, 14: T cells
  - 7, 8, 11: CD8 T cells
  - 1, 2: CD4 T cells
- Clusters 4, 5, 12: B cells
- Cluster 0, 3, 9, 10: Monocytes
- Cluster 6: NK cells
- Cluster 13: Platelets
- Cluster 15: cDC1, cDC2
- Cluster 16: pDC
- Cluster 17: hematopoietic

scRNA-seq usual way

- Preprocessing
- tSNE / UMAP visualization
- Clustering
- Annotating clusters:
  - Checked known markers
  - Identified markers automatically and looked at them
- Asking scientific questions

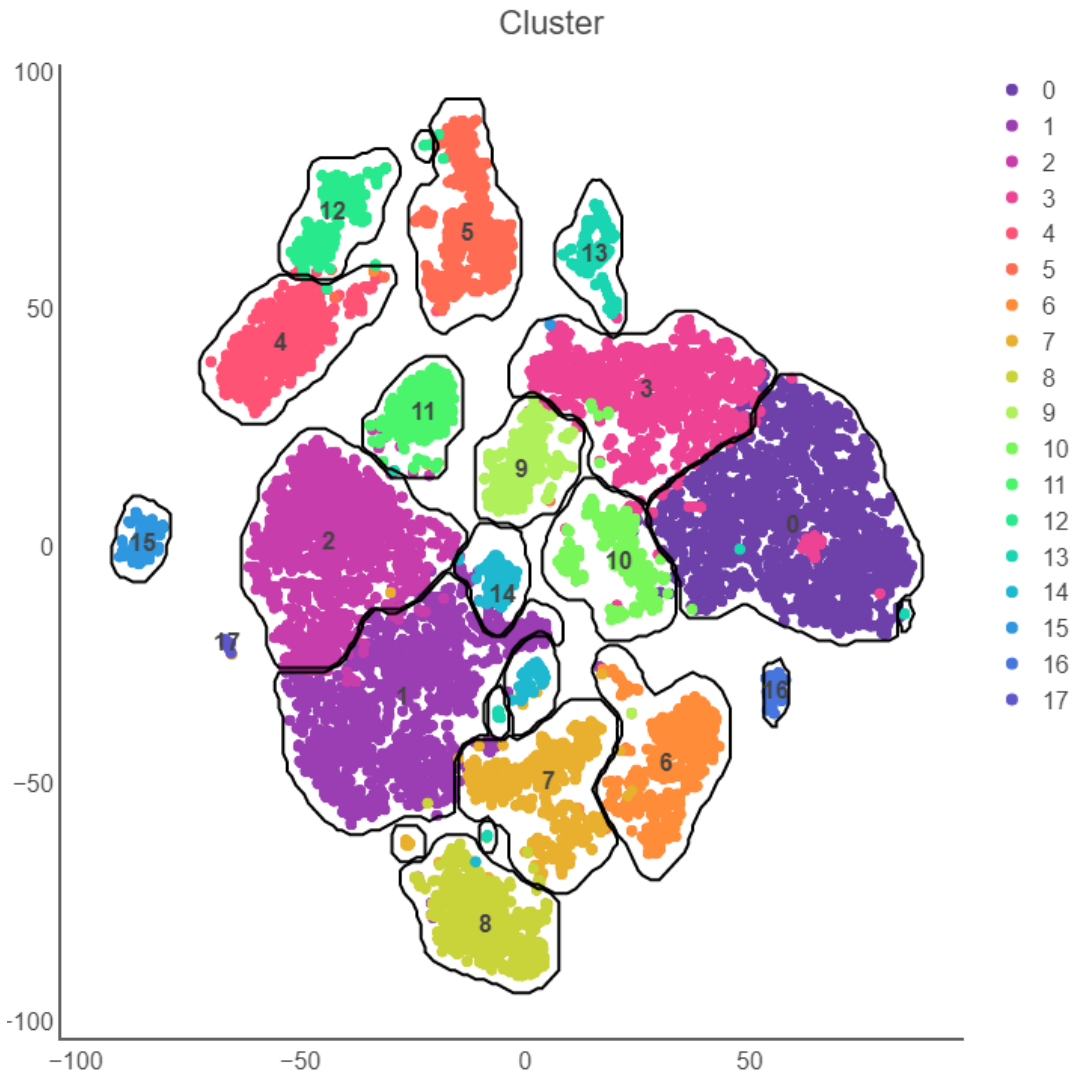
Asking scientific questions

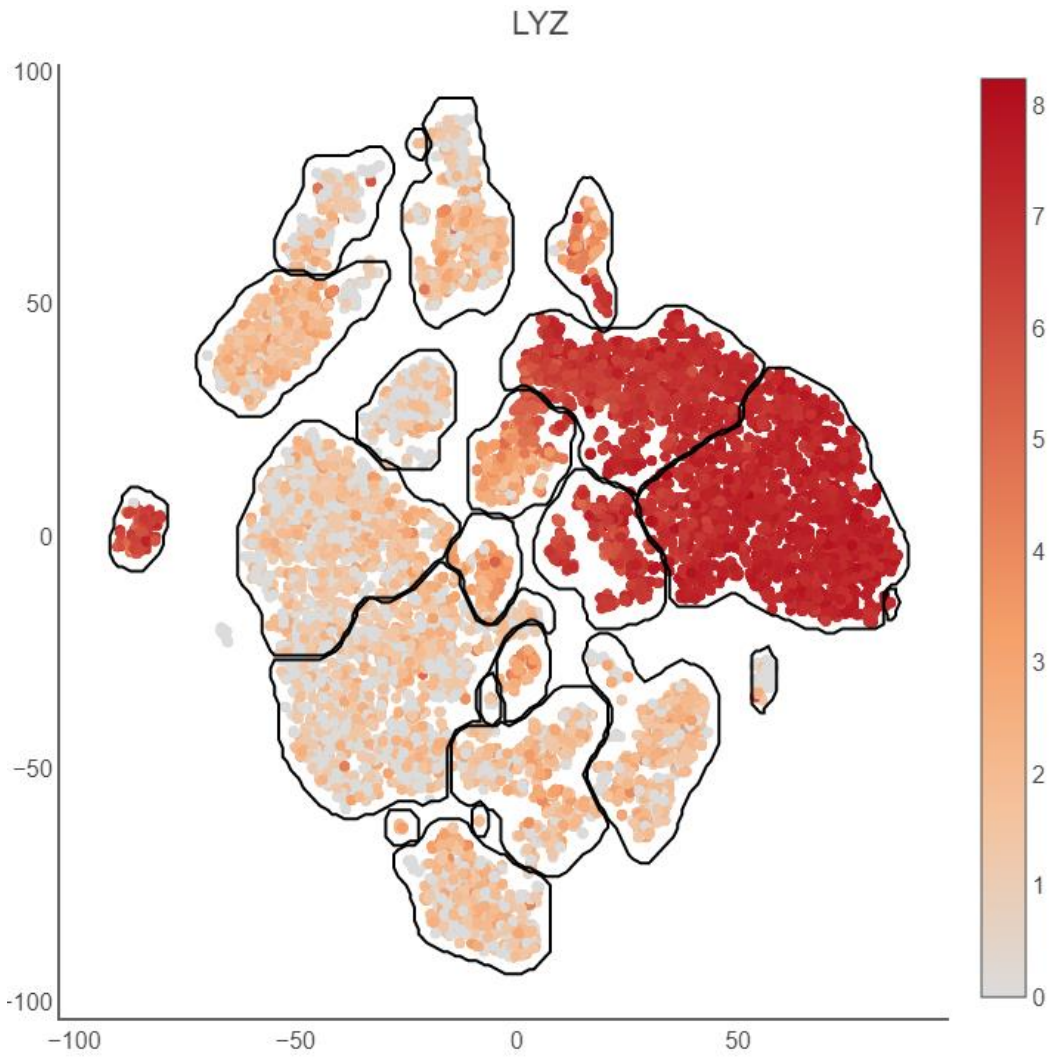
- Once you figured out (annotated) cellular subpopulations you can start asking scientific questions
- Clusters 3 and 0 are both monocytes, what’s the difference between them?

Differential expression

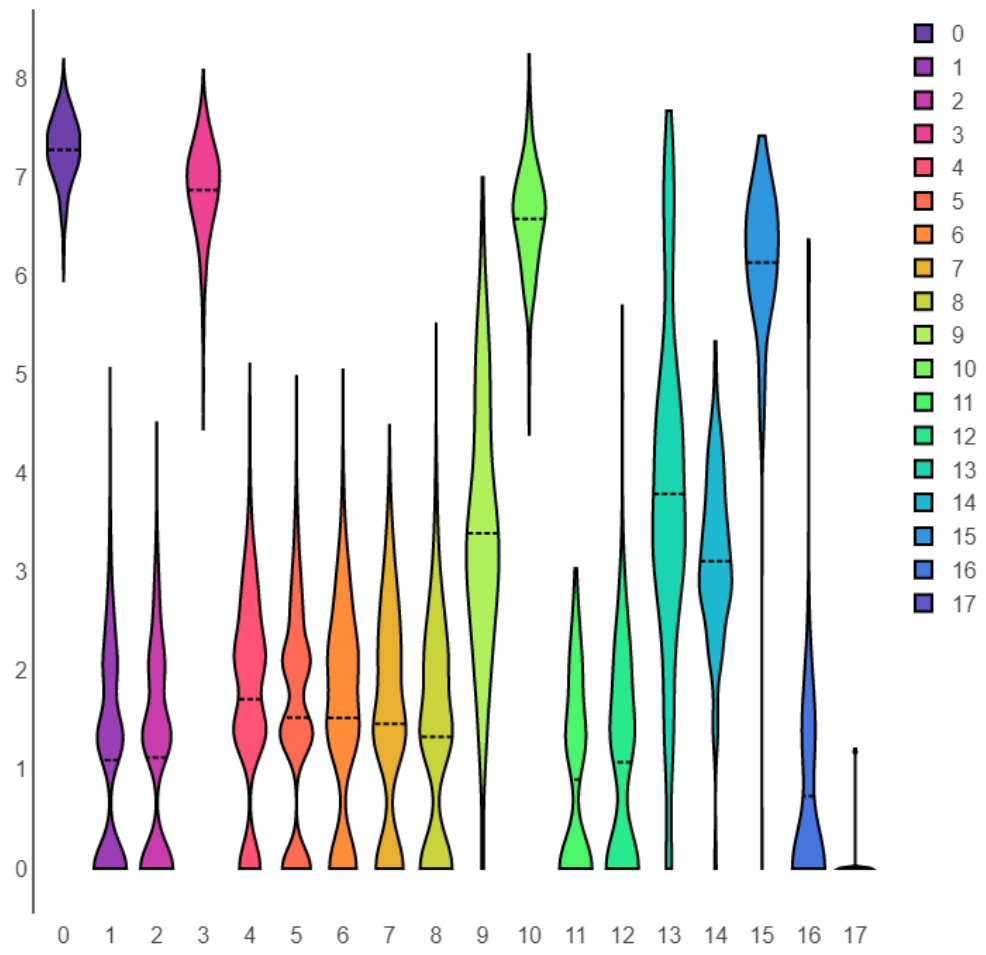
- In bulk RNA-seq we compared groups of several samples with each other
- In single-cell RNA-seq we will compare cell groups against each other:
  - **One cluster against the other**
  - One cluster against all the other clusters (marker identification)
  - One condition against the other (almost bulk RNA-seq)
  - Same cell type in different conditions

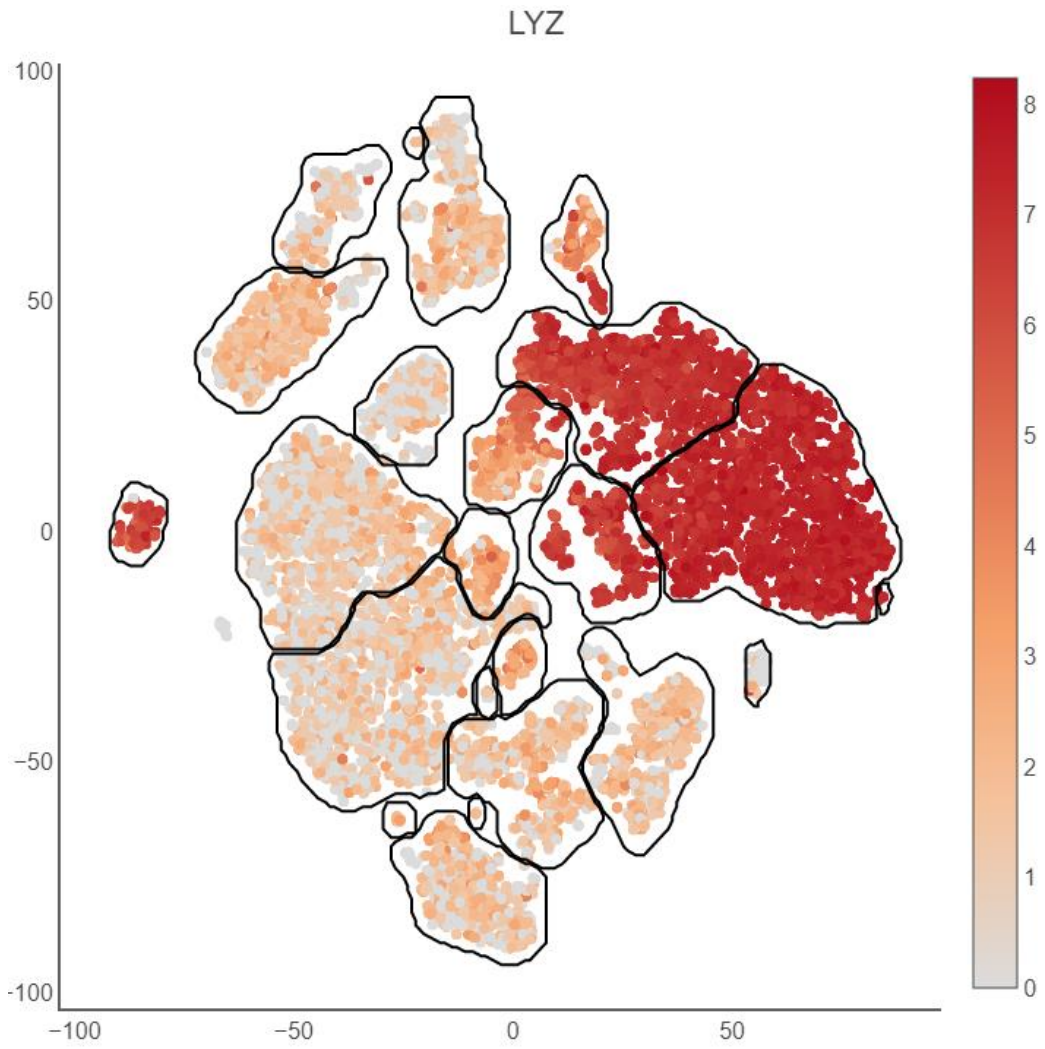
Let’s look at one gene first





Let’s look at one gene first





Finak *et al. Genome Biology* (2015) 16:278
DOI 10.1186/s13059-015-0844-5

Genome Biology

METHOD

Open Access

MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data

Greg Finak[1†], Andrew McDavid[1†], Masanao Yajima[1†], Jingyuan Deng[1], Vivian Gersuk[2], Alex K. Shalek[3,4,5,6], Chloe K. Slichter[1], Hannah W. Miller[1], M. Juliana McElrath[1], Martin Prlic[1], Peter S. Linsley[2] and Raphael Gottardo[1,7*]

Comparing similar populations

- We can compare clusters 0 and 3 to figure out what is different between these clusters
- The generated table with results will contain several important fields
- Download de_0_vs_3.tsv and open it in excel
- And sort it by log fold change

Diff expression results

	A	B	C	D	E	F	G
1		p_val	avg_logFC	pct.1	pct.2	p_val_adj	
2	S100A8	0	1.109038	1	0.994	0	
3	S100A12	0	1.007759	1	0.933	0	
4	S100A9	8.69E-305	0.754482	1	1	1.66E-300	
5	SLC2A3	1.26E-116	0.675191	0.759	0.377	2.41E-112	
6	CYP1B1	8.98E-113	0.604021	0.747	0.386	1.71E-108	
7	PLBD1	3.75E-117	0.602147	0.89	0.761	7.16E-113	
8	ALOX5AP	1.20E-119	0.60193	0.539	0.148	2.29E-115	
9	SELL	2.81E-136	0.576161	0.774	0.377	5.36E-132	
10	VCAN	1.73E-160	0.504749	0.998	0.958	3.29E-156	
11	RGS2	5.86E-72	0.462258	0.941	0.826	1.12E-67	
12	VNN2	1.85E-95	0.421641	0.595	0.232	3.54E-91	
13	RBP7	2.74E-66	0.420781	0.735	0.512	5.23E-62	
14	PADI4	8.89E-109	0.416832	0.489	0.119	1.70E-104	
15	HMGB2	5.58E-61	0.399081	0.757	0.581	1.06E-56	

- avg_logFC – average log fold change
- p_val – p value (bad)
- p_val_adj – p value adjusted for multiple hypothesis (good)
- pct.1 – % of cells in first group (cluster 0) that have non-zero expression values of gene
- pct.2 – % of cells in second group (cluster 3) that have non-zero expression values of gene

Diff expression results (sorted other way)

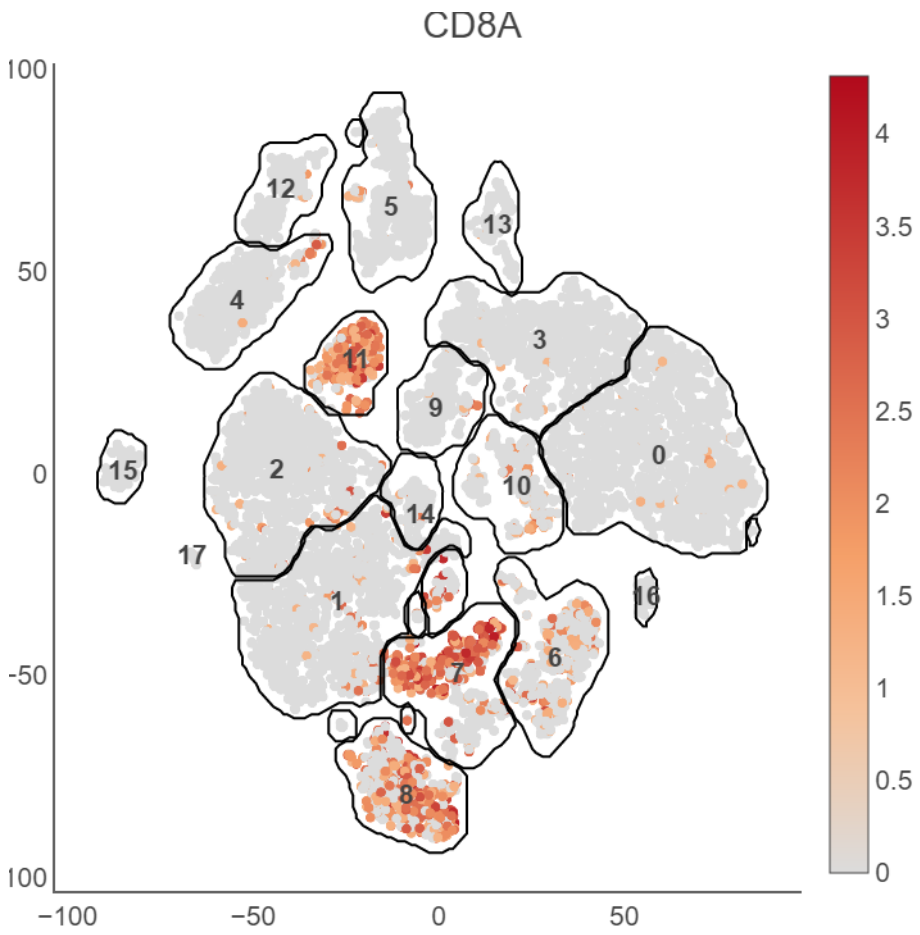
- avg_logFC – average log fold change
- p_val – p value (bad)
- p_val_adj – p value adjusted for multiple hypothesis (good)
- pct.1 – % of cells in first group (cluster 0) that have non-zero expression values of gene
- pct.2 – % of cells in second group (cluster 3) that have non-zero expression values of gene

	A	B	C	D	E	F	G	H
1		p_val	avg_logFC	pct.1	pct.2	p_val_adj		
2	HLA.DPA1	0	-1.35769	0.792	0.984	0		
3	HLA.DPB1	0	-1.31893	0.688	0.969	0		
4	HLA.DRA	0	-1.06345	0.983	0.999	0		
5	HLA.DRB1	0	-1.0302	0.885	0.991	0		
6	HLA.DQB1	4.11E-210	-0.85772	0.332	0.788	7.84E-206		
7	CD74	0	-0.84914	0.991	0.999	0		
8	HLA.DRB5	8.46E-201	-0.70942	0.394	0.829	1.61E-196		
9	HLA.DMA	9.97E-207	-0.69584	0.478	0.877	1.90E-202		
10	HLA.DQA2	9.32E-169	-0.6718	0.2	0.644	1.78E-164		
11	HLA.DQA1	3.37E-144	-0.6574	0.157	0.567	6.43E-140		
12	HLA.DMB	6.77E-168	-0.59253	0.577	0.892	1.29E-163		
13	LGALS2	3.03E-153	-0.58006	0.722	0.949	5.77E-149		
14	CPVL	1.74E-184	-0.54994	0.868	0.995	3.33E-180		
15	MARCKS	1.72E-57	-0.47508	0.597	0.782	3.28E-53		
16	ISG15	4.66E-41	-0.45947	0.217	0.397	8.90E-37		
17	LY6E	4.06E-58	-0.45549	0.149	0.388	7.74E-54		
18	LIPA	8.44E-89	-0.44742	0.587	0.81	1.61E-84		
19	CLEC10A	1.88E-83	-0.42527	0.047	0.29	3.58E-79		
20	IL1B	1.05E-32	-0.42242	0.446	0.631	2.00E-28		
21	PSME2	2.36E-70	-0.41914	0.44	0.712	4.51E-66		

Asking scientific questions

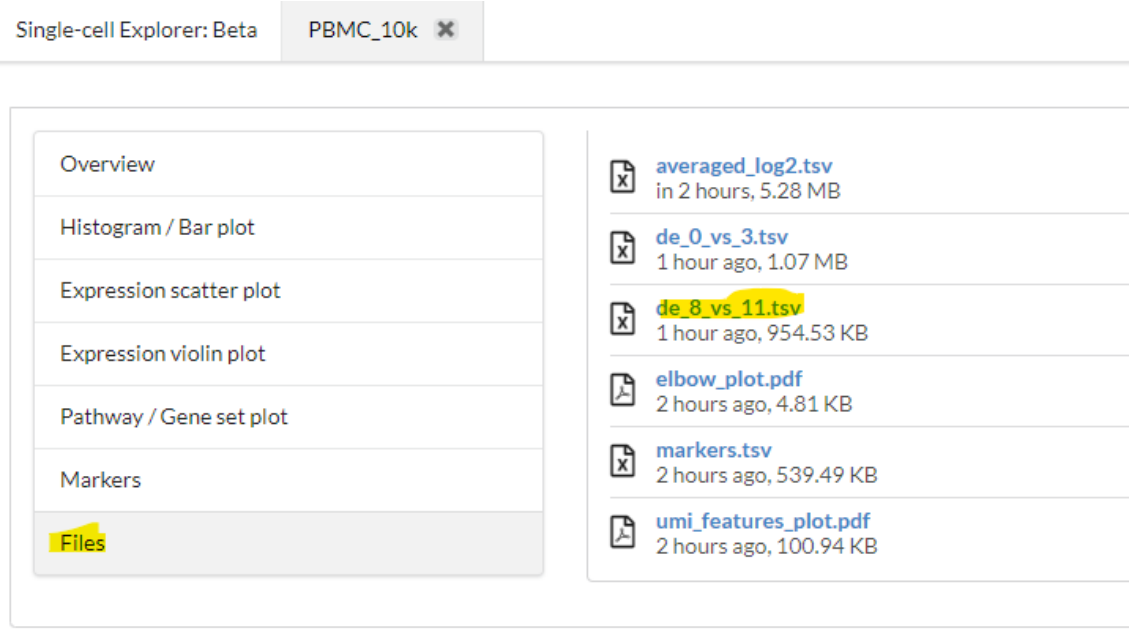
- Cluster 3 is monocytes with higher expression of MHC class II
- Cluster 0 is just CD14+ monocytes

Let’s ask another question



Let’s ask another question

- Two cd8+ t cell clusters: cluster 8 vs cluster 11
- We will try to look at some genes first and then do pathway analysis to figure out the difference between those
- Let’s download the DE list and have a look



Looking at the genes by eye: sorted descending

- Cytotoxic markers
- A lot of cytotoxic markers :)

A1

	A	B	C	D	E	F	G
1		p_val	avg_logFC	pct.1	pct.2	p_val_adj	
2	KLRB1	2.87E-304	2.894641	1	0.046	5.47E-300	
3	NKG7	6.86E-239	2.128671	0.982	0.098	1.31E-234	
4	CCL5	8.05E-204	1.989348	0.961	0.058	1.54E-199	
5	S100A4	5.25E-291	1.86399	1	0.69	1.00E-286	
6	GZMA	3.13E-213	1.833921	0.961	0.04	5.97E-209	
7	GZMK	1.43E-208	1.82362	0.953	0.021	2.73E-204	
8	KLRG1	4.06E-173	1.278354	0.917	0.052	7.75E-169	
9	CST7	2.73E-156	1.195073	0.844	0.015	5.21E-152	
10	ANXA1	5.02E-135	1.191004	0.917	0.193	9.57E-131	
11	PRF1	5.56E-121	1.101231	0.798	0.064	1.06E-116	
12	CLIC1	4.44E-122	1.056059	0.935	0.353	8.47E-118	
13	HOPX	4.46E-135	1.033727	0.776	0.009	8.51E-131	
14	MYO1F	5.14E-127	1.013486	0.808	0.052	9.81E-123	
15	TRGC2	2.74E-99	1.010431	0.74	0.067	5.22E-95	
16	NCR3	2.28E-106	0.993501	0.751	0.058	4.35E-102	
17	LYAR	5.46E-102	0.974212	0.842	0.181	1.04E-97	
18	S100A6	1.94E-142	0.970113	0.997	0.902	3.71E-138	
19	IL32	2.83E-122	0.904538	0.992	0.951	5.41E-118	
20	IL7R	1.20E-78	0.859007	0.987	0.868	2.29E-74	
21	PHACTR2	2.45E-86	0.806925	0.672	0.052	4.68E-82	

Looking at the genes by eye: sorted ascending

- Some markers
- CCR7, SELL?

	A	B	C	D	E	F	G
1		p_val	avg_logFC	pct.1	pct.2	p_val_adj	
2	CD8B	2.95E-175	-1.33658	0.208	0.966	5.63E-171	
3	CCR7	5.77E-126	-0.9418	0.05	0.77	1.10E-121	
4	SELL	4.72E-99	-0.87466	0.14	0.813	9.02E-95	
5	LEF1	1.01E-116	-0.82483	0.034	0.715	1.93E-112	
6	AIF1	2.84E-91	-0.80244	0.127	0.742	5.41E-87	
7	LDHB	3.37E-124	-0.80214	0.839	0.994	6.43E-120	
8	TRABD2A	1.96E-104	-0.79409	0.098	0.764	3.73E-100	
9	RGS10	1.33E-92	-0.76075	0.403	0.908	2.53E-88	
10	FYB1	1.37E-79	-0.75901	0.592	0.96	2.62E-75	
11	LINC02446	3.59E-106	-0.72738	0.002	0.577	6.86E-102	
12	ACTN1	3.52E-95	-0.71595	0.046	0.656	6.71E-91	
13	TMSB10	3.38E-127	-0.70631	1	1	6.45E-123	
14	FOXP1	4.37E-64	-0.69861	0.415	0.89	8.33E-60	
15	PASK	3.98E-72	-0.65459	0.046	0.552	7.59E-68	
16	PIK3IP1	4.18E-57	-0.64525	0.372	0.819	7.98E-53	
17	NELL2	1.79E-67	-0.61723	0.109	0.629	3.41E-63	
18	MAL	3.84E-80	-0.60564	0.021	0.531	7.32E-76	
19	SERINC5	6.59E-78	-0.58531	0.033	0.558	1.26E-73	
20	TCF7	4.90E-56	-0.5777	0.59	0.929	9.35E-52	
21	APBA2	2.48E-82	-0.5739	0.018	0.534	4.73E-78	
22	LDLRAP1	3.86E-54	-0.53299	0.159	0.647	7.36E-50	
23	RPS6	9.37E-122	-0.50551	1	1	1.79E-117	

Let’s look at the pathways

- We believe that transcriptional changes do not come at random and are driven by different pathways
- Computationally speaking, pathway is just a set of genes

Hypothesis

- We kinda know that cluster 8 are effector Cd8 T cells?
- Cluster 11 are naïve/memory Cd8 t cells?

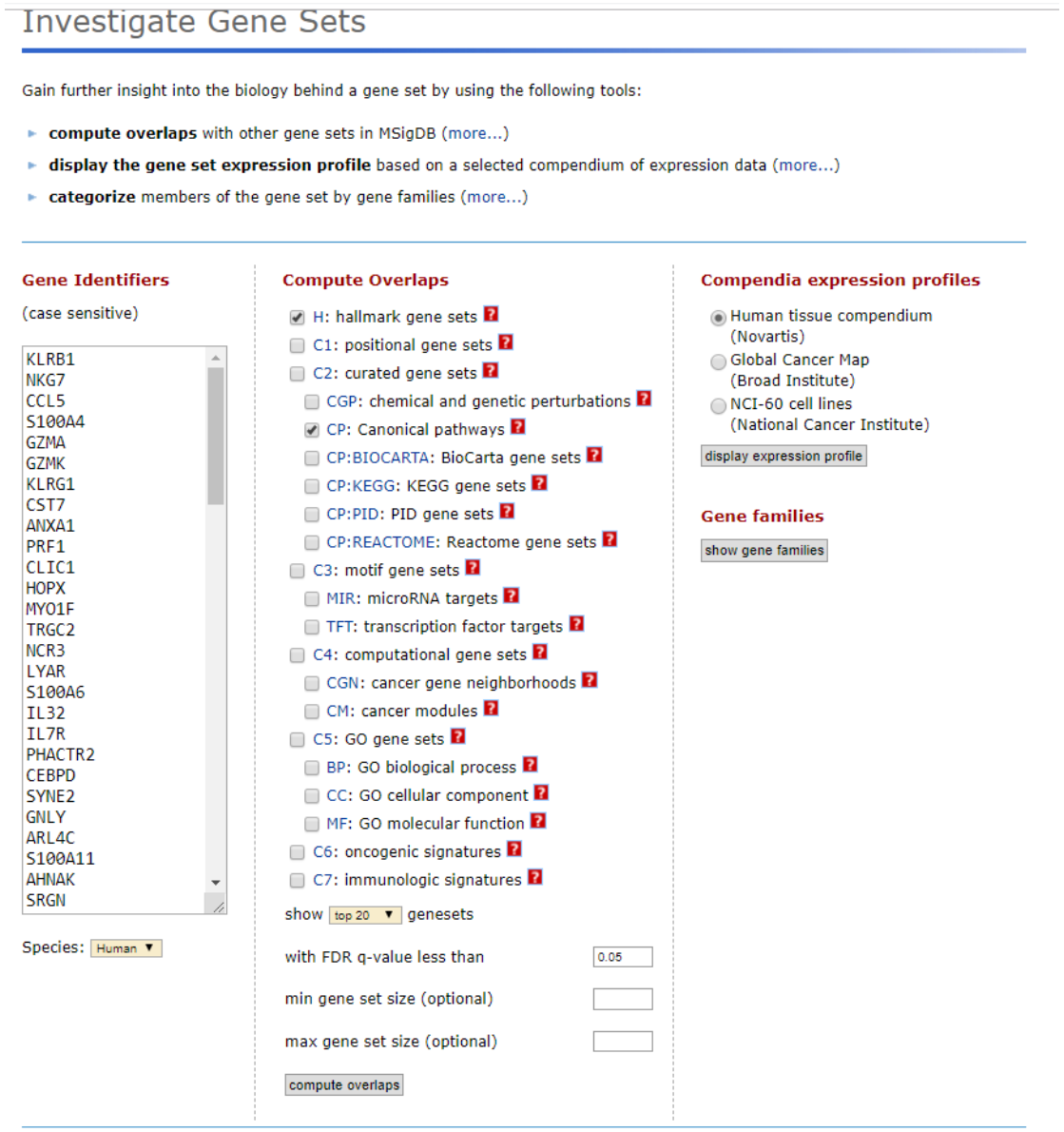
- Can we look at the pathways to get more information?

msigdb

- Let’s open in excel de_8_vs_11.tsv
- Let’s select top 100 genes upregulated in activated T cells
- Let’s search for the pathways
- http://software.broadinstitute.org/gsea/msigdb/annotate.jsp

msigdb

http://software.broadinstitute.org/gsea/msigdb/annotate.jsp



msigdb

Compute Overlaps for Selected Genes

Converted 99 submitted identifiers into 95 entrez genes. click here for details.

Collections	# Overlaps Shown	# Gene Sets in Collections	# Genes in Comparison (n)	# Genes in Universe (N)
CP, H	20	2249	95	38055

Click the gene set name to see the gene set page. Click the number of genes [in brackets] to download the list of genes.

Color bar shading from light green to black, where lighter colors indicate more significant FDR q-values (< 0.05) and black indicates less significant FDR q-values (>= 0.05).

Save to: Excel

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_NE_SYSTEM [856]	Cytokine Signaling in Immune system	22		$1.87\ e^{-16}$	$4.2\ e^{-13}$
REACTOME_ADAPTIVE_IMMUNE_SYSTEM [811]	Adaptive Immune System	20		$1.13\ e^{-14}$	$1.27\ e^{-11}$
REACTOME_SIGNALING_BY_INTERLEUKINS [631]	Signaling by Interleukins	18		$2.4\ e^{-14}$	$1.8\ e^{-11}$
REACTOME_RESPONSE_TO_ELEVATED_PLATELET LET_CYTOSOLIC_CA2PLUS [132]	Response to elevated platelet cytosolic Ca2+	11		$3.49\ e^{-14}$	$1.96\ e^{-11}$
HALLMARK_ALLOGRAFT_REJECTION [200]	Genes up-regulated during transplant rejection.	12		$1.19\ e^{-13}$	$4.45\ e^{-11}$
HALLMARK_COMPLEMENT [200]	Genes encoding components of the complement system, which is part of the innate immune system.	12		$1.19\ e^{-13}$	$4.45\ e^{-11}$
REACTOME_HEMOSTASIS [674]	Hemostasis	17		$9.68\ e^{-13}$	$3.11\ e^{-10}$
REACTOME_PLATELET_ACTIVATION_SIGNALING ING_AND_AGGREGATION [260]	Platelet activation, signaling and aggregation	11		$5.77\ e^{-11}$	$1.62\ e^{-8}$
HALLMARK_TNFA_SIGNALING_VIA_NFKB [200]	Genes regulated by NF-kB in response to TNF [GeneID=7124].	10		$8.79\ e^{-11}$	$2.2\ e^{-8}$
REACTOME_IMMUNOREGULATORY_INTERACTIONS ONS_BETWEEN_A_LYMPHOID_AND_A_NON_LYMPH MPHOID_CELL [186]	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	9		$1.07\ e^{-9}$	$2.41\ e^{-7}$
HALLMARK_IL2_STAT5_SIGNALING [200]	Genes up-regulated by STAT5	9		$2.03\ e^{-9}$	$4.15\ e^{-7}$

GeneQuery

- Let’s take the same 100 genes and ask GeneQuery for similar datasets
- https://artyomovlab.wustl.edu/genequery/searcher/

GeneQuery

GeneQuery$^{\alpha}$

Database species: Homo Sapiens Mus Musculus Rattus Norvegicus

Query species: Homo Sapiens Mus Musculus Rattus Norvegicus

Gene list (separated by newline/whitespace/tab)

```
HLA.C
GPR171
CDC42EP3
ITGB2
PYHIN1
LST1
PPIB
H3F3B
GAPDH
UBC
P4HB
CD40LG
CASP1
RHOC
GYG1
CELF2
MYL12A
CXXC5
PFN1
```

Search

Run example

GeneQuery

#	Experiment title	Module	$\log_{10}(\text{adj.}p_{value})$	Overlap	GSE	GMT
1	Gene expressions of CD4+ T cells in each developmental stages	3	-30.88	53/558	GSE61697	
2	Nave-like Yellow-Fever specific CD8 T cells and reference CD8 T cell subsets in humans	3	-27.84	44/399	GSE65804	
3	Gene Expression of Circulating B Lymphocytes for Smoking-related Osteoporosis in Postmenopausal Females	7	-19.02	30/231	GSE13850	
4	Clinical implications of gene dosage and gene expression patterns in diploid breast carcinoma (transcriptomic profiling)	2	-17.56	37/506	GSE20462	
5	Clinical implications of gene dosage and gene expression patterns in diploid breast carcinoma	2	-17.56	37/506	GSE20486	
6	Peripheral blood mononuclear cell gene expression in chronic obstructive pulmonary disease	6	-16.76	29/194	GSE42057	
7	Prognostic value of gene signatures and proliferation in lymph node negative breast cancer	3	-16.70	30/334	GSE46563	
8	Gene Expression of Circulating B Lymphocytes for Osteoporosis	4	-16.57	34/389	GSE7429	

#	Experiment title	Module	$\log_{10}(\text{adj.}p_{value})$	Overlap	GSE	GMT
1	Gene expressions of CD4+ T cells in each developmental stages	3	-30.88	53/558	GSE61697	
2	Naye-like Yellow-Fever specific CD8 T cells and reference CD8 T cell subsets in humans	3	-27.84	44/399	GSE65804	

NCBI

GEO Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display**

Not logged in | Login

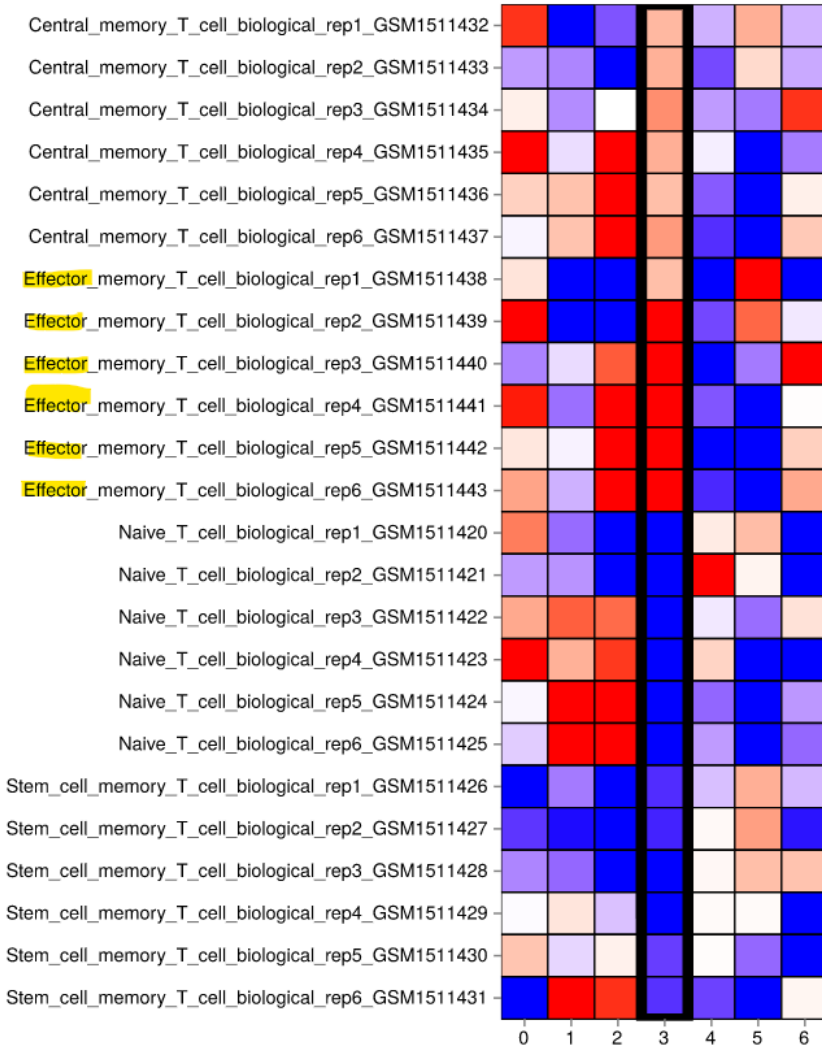
Scope: Self ▼ Format: HTML ▼ Amount: Quick ▼ GEO accession: GSE61697 GO

Series GSE61697 Query DataSets for GSE61697

Status	Public on May 12, 2015
Title	Gene expressions of CD4+ T cells in each developmental stages
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	The development of T cells has been characterized as taking place over three stages: naïve (Tn), central memory (Tcm), and effector memory (Tem) cells. Recently, stem cell memory T cells (Tscm) were found to be the least-developed memory subset. We performed detailed analysis of the gene expression of human CD4+ T cells with clear distinction of the Tn, Tscm, Tcm, and Tem stages.
Overall design	We sorted Tn, Tscm, Tcm, and Tem CD4+ T cells from the peripheral blood of six healthy volunteers to see the differences of gene expression between each developmental stage.
Contributor(s)	Takeshita M, Takeuchi T
Citation(s)	Takeshita M, Suzuki K, Kassai Y, Takiguchi M et al. Polarization diversity of human CD4+ stem cell memory T cells. *Clin Immunol* 2015 Jul;159(1):107-17. PMID: 25931384

#	Experiment title	Module	$\log_{10}(\text{adj.}p_{value})$	Overlap	GSE	GMT
1	Gene expressions of CD4+ T cells in each developmental stages	3	-30.88	53/558	GSE61697	
2	Naye-like Yellow-Fever specific CD8 T cells and reference CD8 T cell subsets in humans	3	-27.84	44/399	GSE65804	

- These are indeed effector CD8 T cells



Averaged pathway expression

- Every cell has very limited coverage in UMIs
- Even abundant transcripts might be hard to detect
- Expression of “one gene” might be not representative
- Averaged expression of gene set is much more robust
- Cd19 Cd79a Cd79b

Averaged pathway expression

- What is average Z score
- We first normalize the gene expression (z score, standard score)

$$Z = \frac{X - \mu}{\sigma}$$

  where μ is mean value and σ is the standard deviation
- Then we calculate averaged expression z score
- Cd19 Cd79a Cd79b

Single-cell Explorer: Beta | PBMC_10k

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

COORDINATES

X coordinate

tSNE_1

Y coordinate

tSNE_2

COLOR AND SPLIT

Choose pathway

or enter a gene set

Cd19 Cd79A cd79b

Submit!

Split by

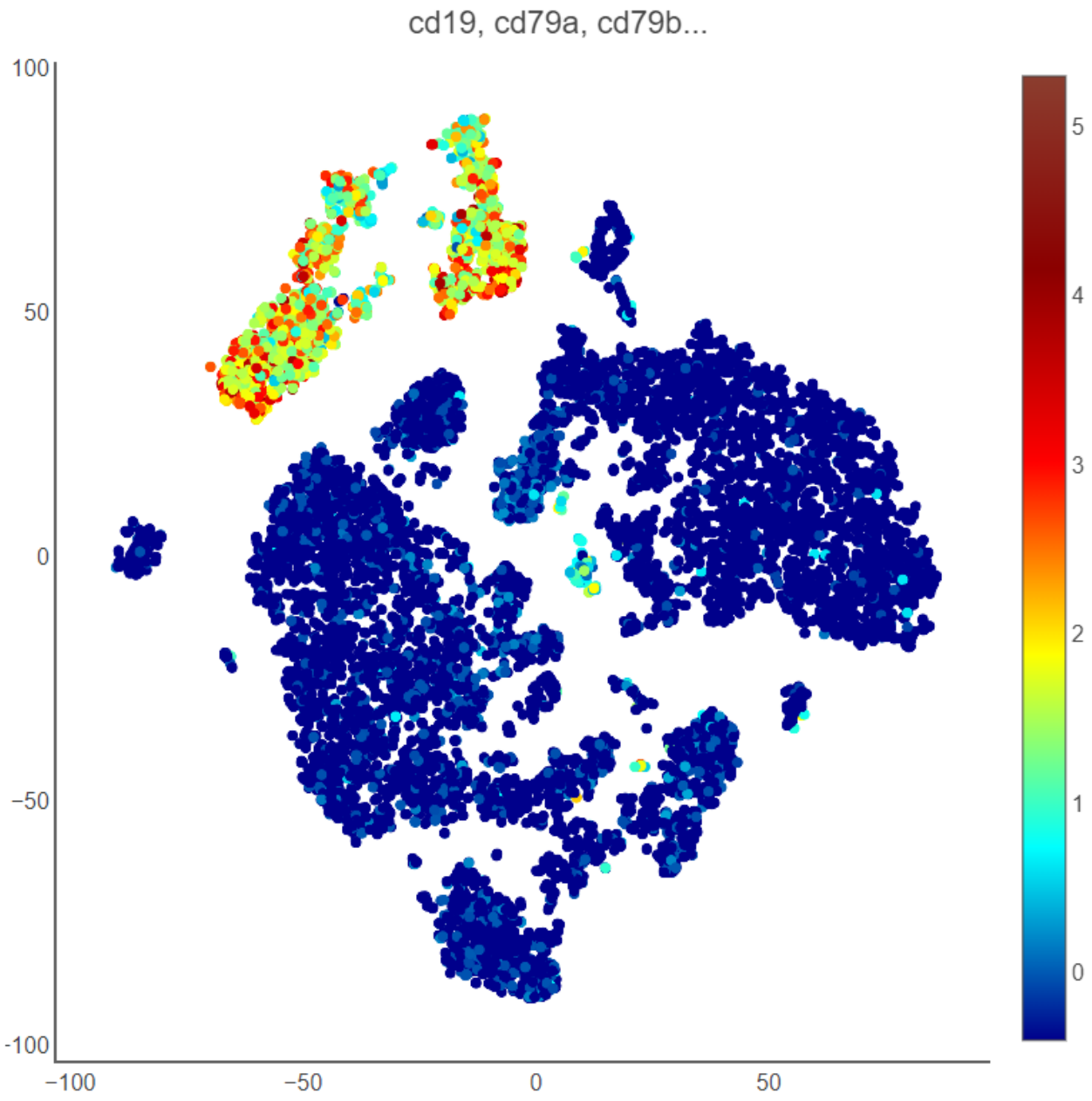
Split by

AVAILABLE ANNOTATIONS

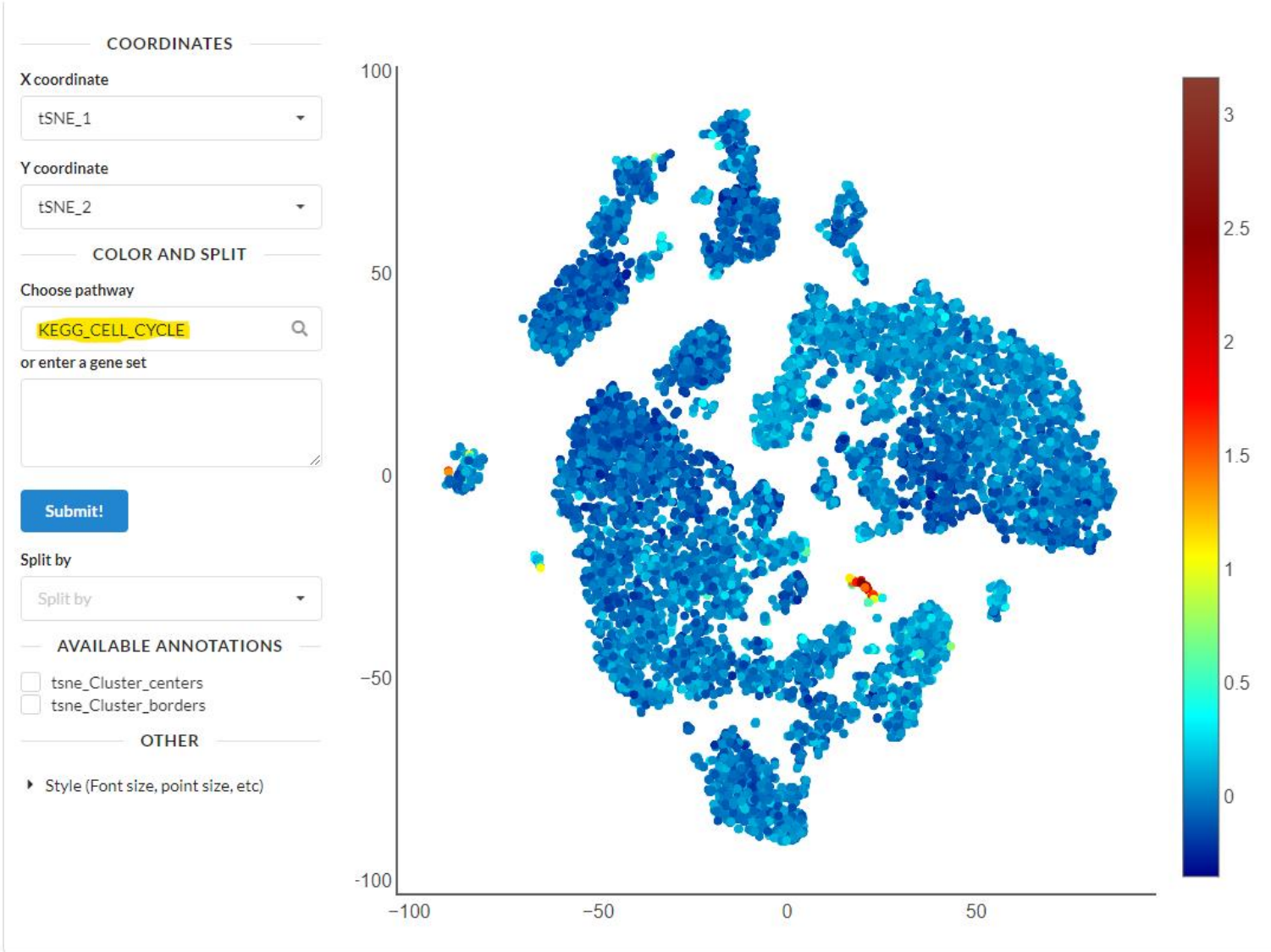
- tsne_Cluster_centers
- tsne_Cluster_borders

OTHER

Style (Font size, point size, etc)



But we can look at whole pathways



Summing up

- Single-cell RNA-seq datasets provide us with a lot of information
- Pathway analysis, phenotype searching (genequery) and other techniques enhance our ability to generate better hypothesis
- A lot of similar phenotypes are already present in scRNA-seq data, one just has to carefully evaluate that

Questions?