ITMO UNIVERSITY

CT Lab
ITMO UNIVERSITY

Advanced annotation

Konstantin Zaitsev

December 3rd, 2020. Tomsk / Saint-Petersburg

Our setup

- Address is the same https://ctlab.itmo.ru/rstudio-sbNN/
- Folder scrna-seq
- File advanced-annotation.R

Lets first load the object

```
library(Seurat)
library(Matrix)
library(MAST)
library(ggplot2)
library(dplyr)
library(fgsea)

seurat <- readRDS("blood_seurat.rds")
```

Calculating averaged expression

```
average <- AverageExpression(seurat)$SCT
averageLog <- log2(as.matrix(average) + 1)
colnames(averageLog) <- paste0("Cluster ", colnames(average))
write.table(averageLog, "average_log.tsv", sep="\t", col.names=NA, quote=F)
```

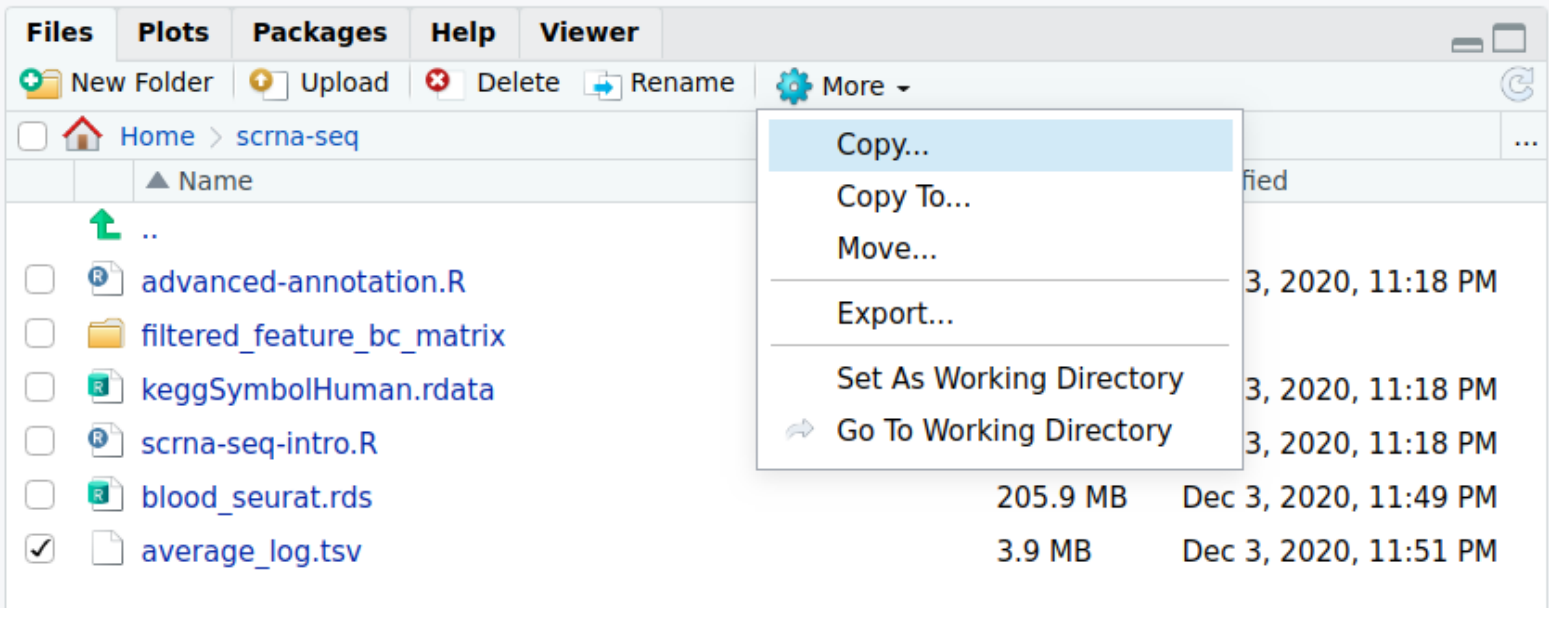
Phantasus

- Phantasus that you used yesterday for bulk RNA-seq can be used for single-cell
- We will look at averaged expression within the clusters
- https://ctlab.itmo.ru/phantasus/

Feedback is welcome!

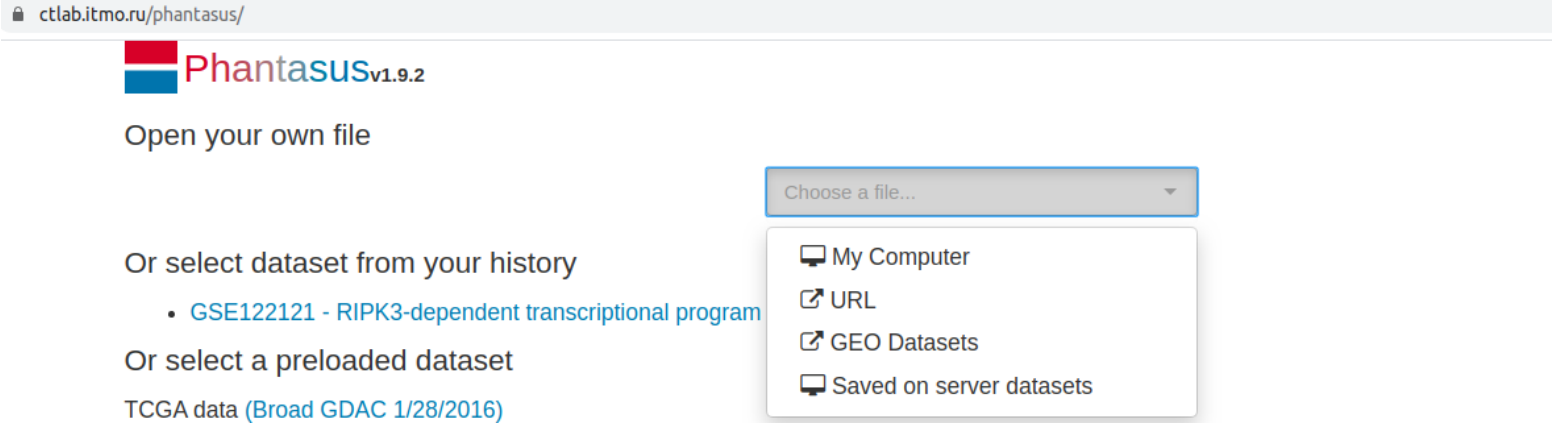
Lets do it

- Download average_log.tsv -> Open it in phantasus
- More -> Export

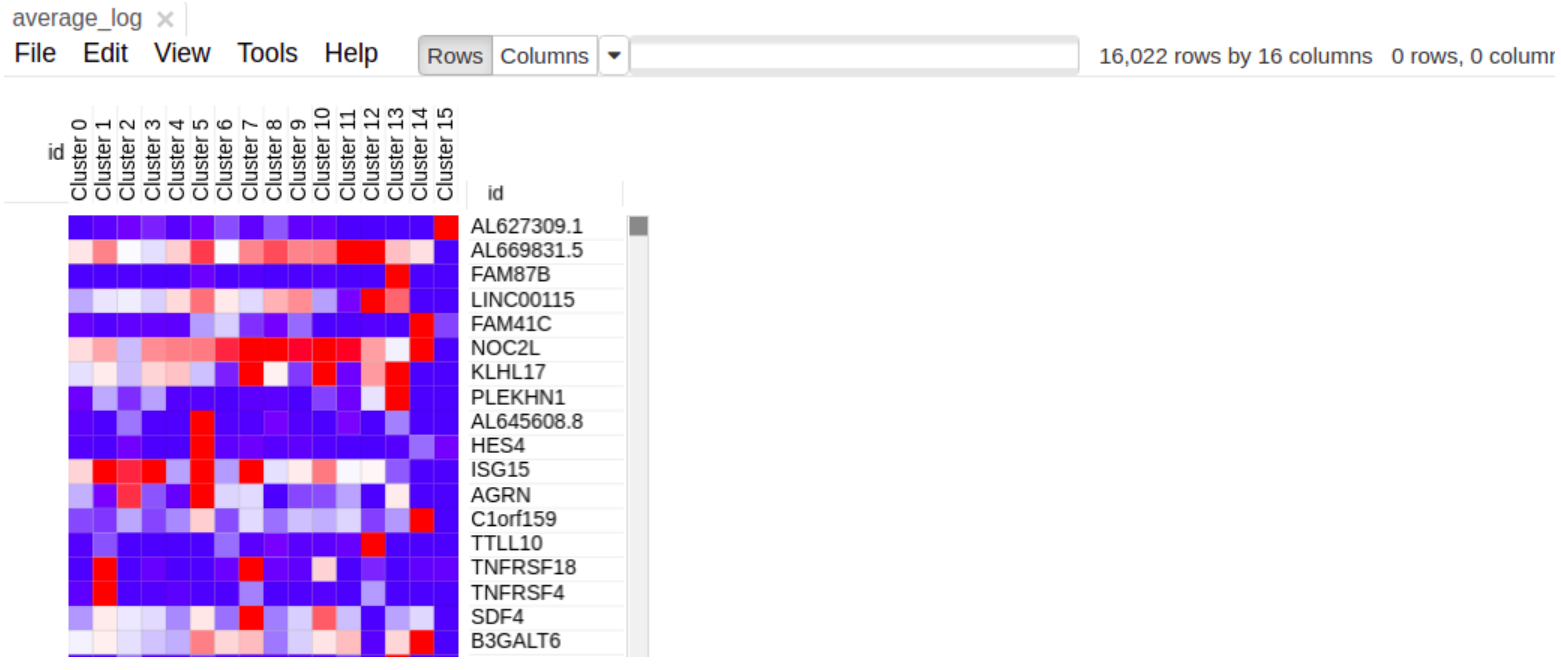


Lets do it

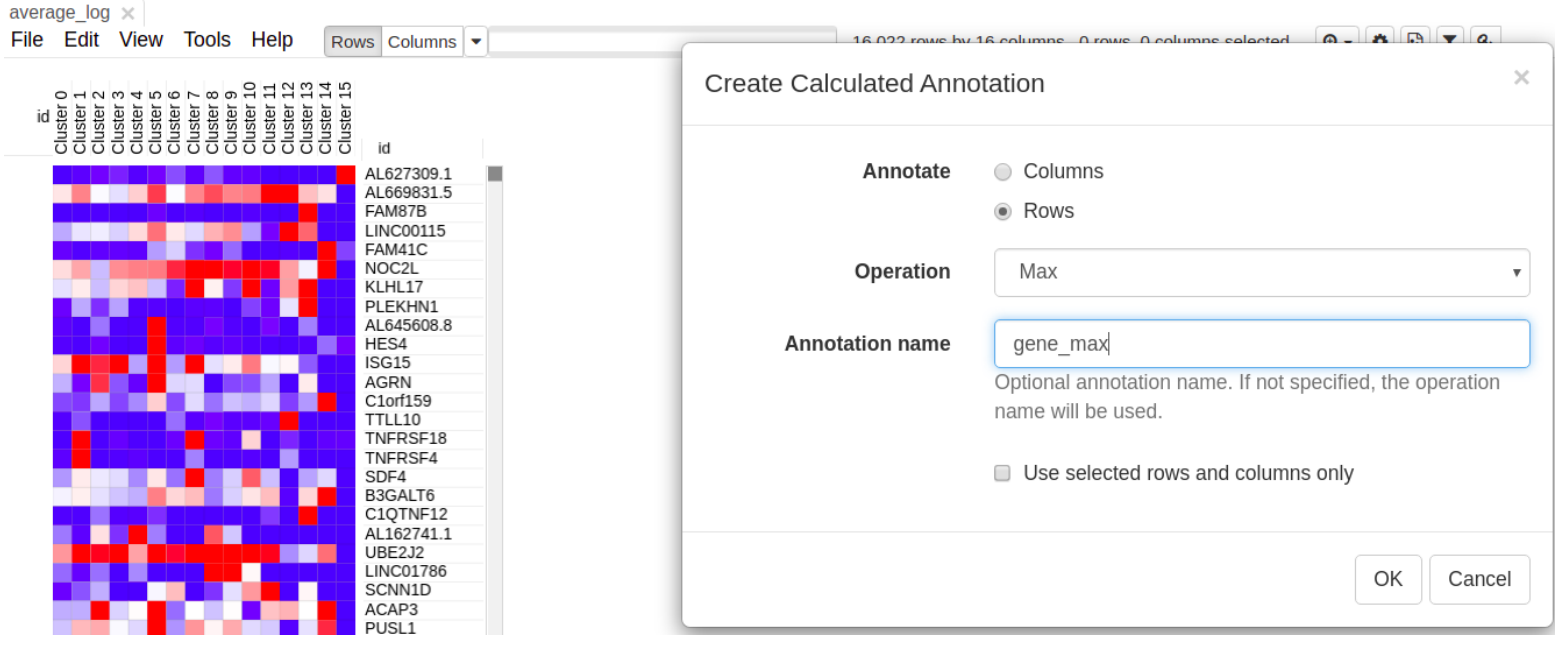
- Download average_log.tsv -> Open it in phantasus (https://ctlab.itmo.ru/phantasus/)
- Open dataset -> My computer -> average_log.tsv



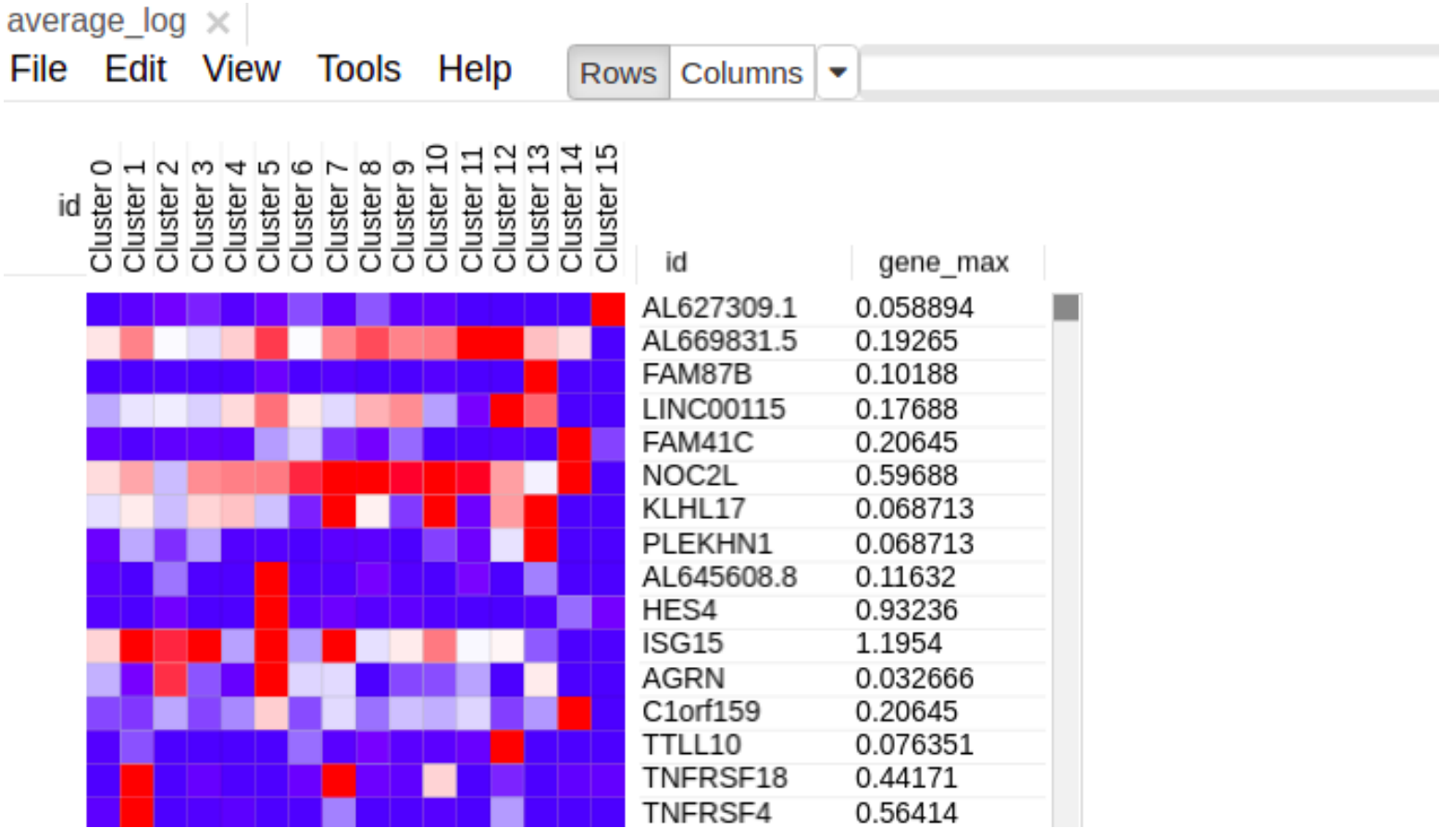
Lets open averaged table in phantasus



Tools -> create calculated annotation

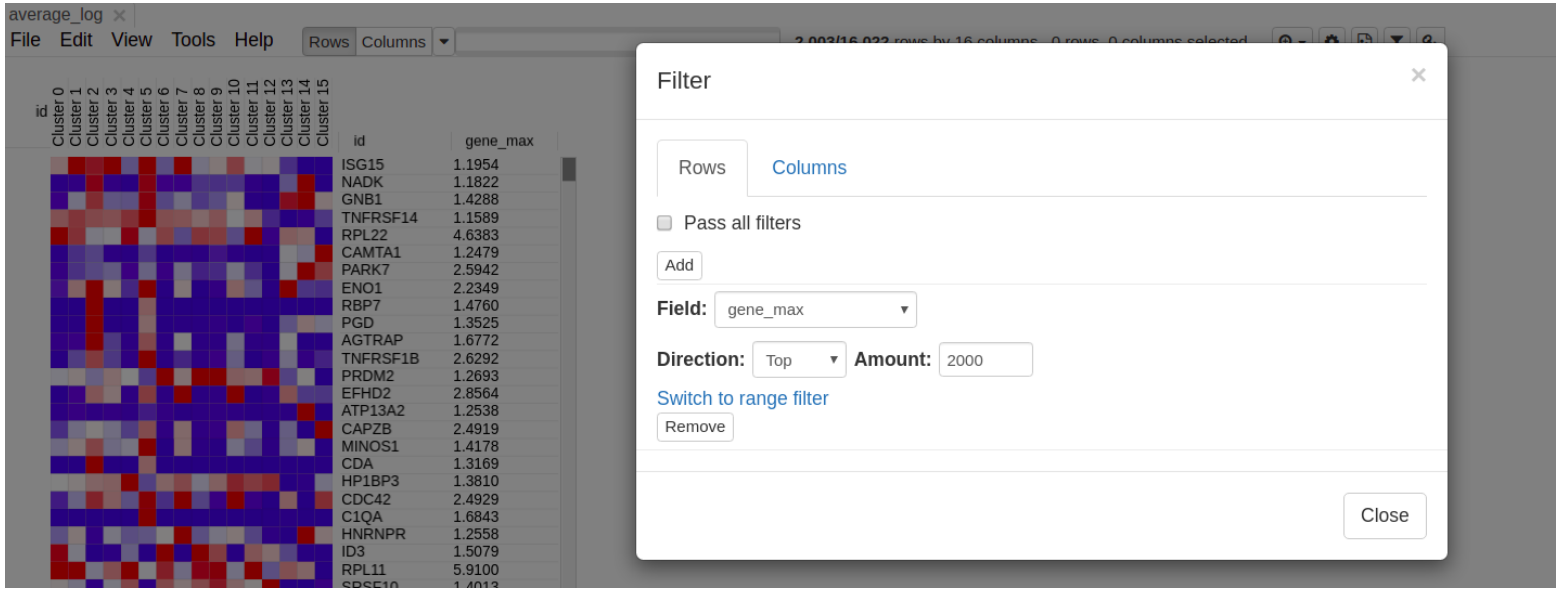


Tools -> create calculated annotation

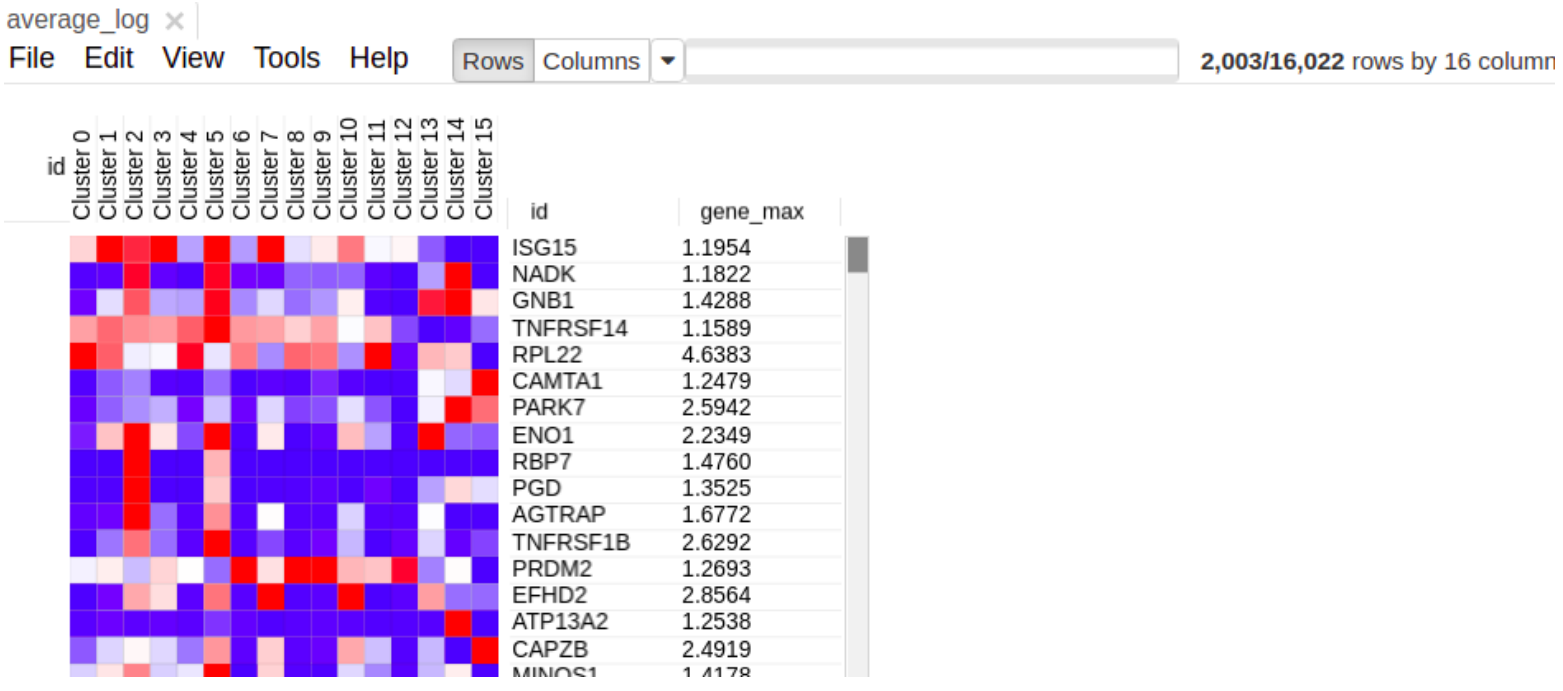


Filter out some genes

- Lets filter genes by average expression
- Tools -> Filter (Add, field = gene_max, switch to top, amount = 2000, close)

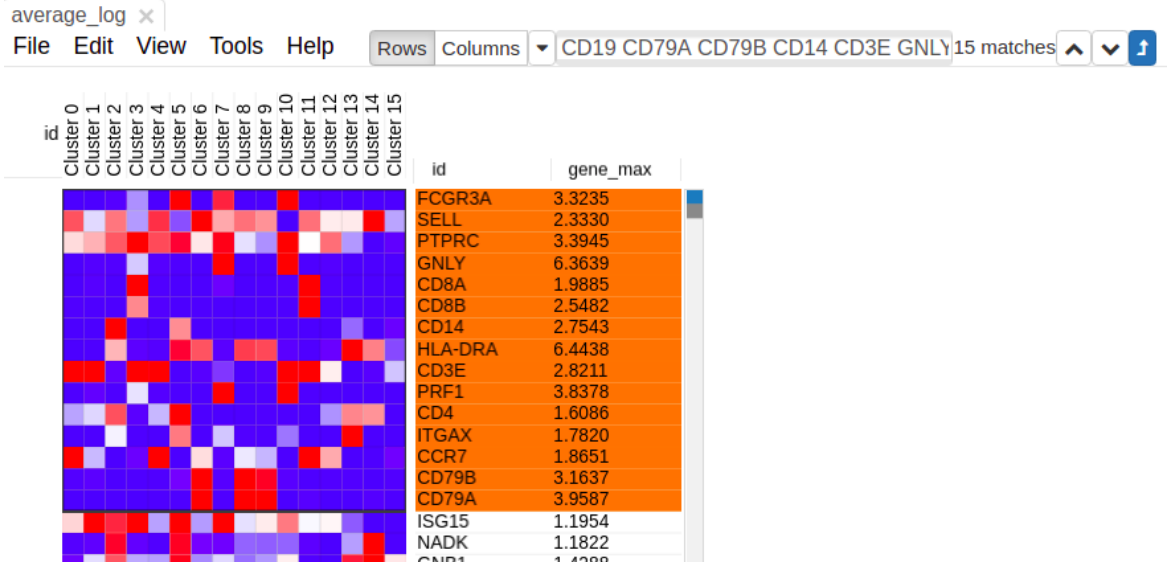


Filtered matrix looks like this



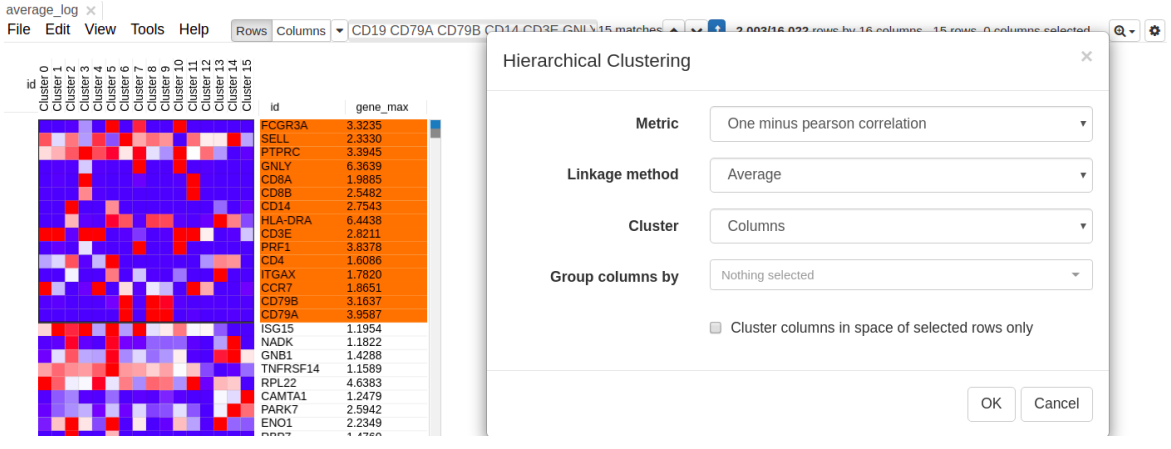
Lets look at some immunological markers

- Lets search for these genes: CD19 CD79A CD79B CD14 CD3E GNLY PRF1 FCGR3A SELL CCR7 ITGAX ITGAM HLA-DRA CD8A CD8B CD4 PTPRC



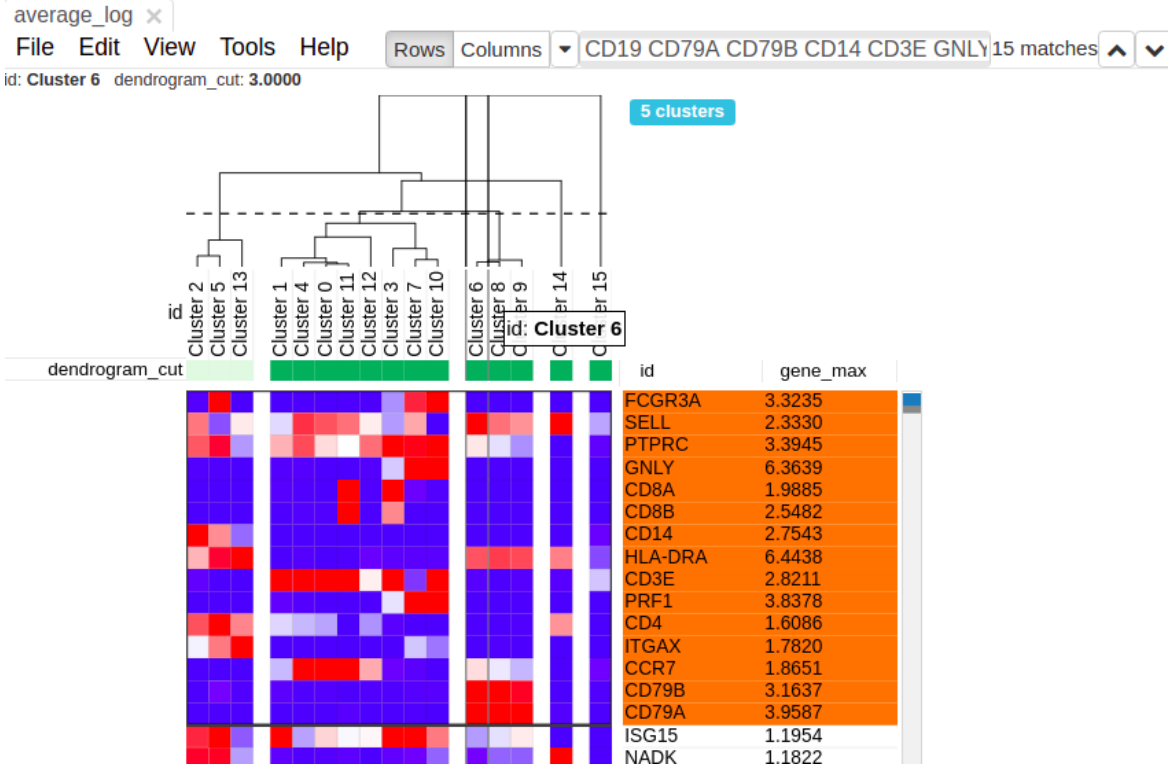
Let's cluster

- Then tools -> clustering -> hierarchical clustering -> Cluster (columns)



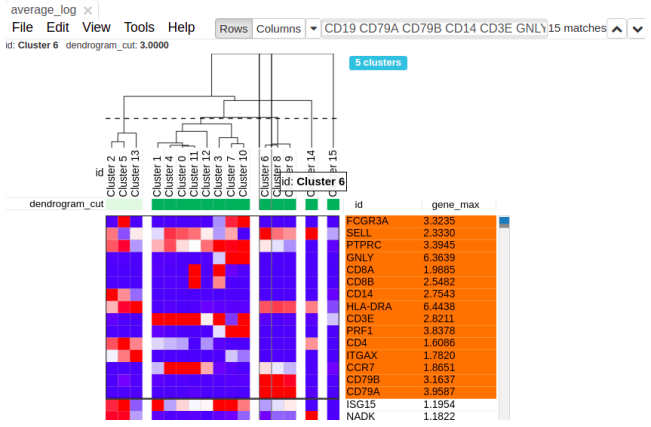
Now we can tell "who is who"

- You can adjust the height of the clustering



Cell lineage defines similarity of clusters

- Clusters 2, 5 are CD14+ monocytes (based on CD14 expression), and cluster 13 are CD16 (FCGR3a expression). Cluster 2, 5 and 13 are from myeloid cell lineage (3 clusters on the left)
- Clusters 6, 8 and 9 are B cell based on CD79 expression (3 clusters in the middle)
- Clusters 0, 1, 3, 4, 7, 10, 11, 12 are T cells and NK cells (CD3 and cytotoxic markers)
- Clusters 14 and 15 are some sort of outliers



Saving heatmaps

- Create new heatmap only of selected genes (Ctrl + X)
- Saving heatmaps is a good thing
- File -> Save Image (Ctrl + S) -> Choose Filename -> Choose format (I prefer svg, svg can be open in browser) -> hooray

While this heatmap is not something you will necessarily put in the paper, but it is ok for supplement or any kind of presentation where you present single-cell RNA-seq data

Differential expression

In bulk RNA-seq we compared groups of several samples (same cell type, same condition, same treatment) between each other. In single-cell RNA-seq we will compare cell groups against each other:

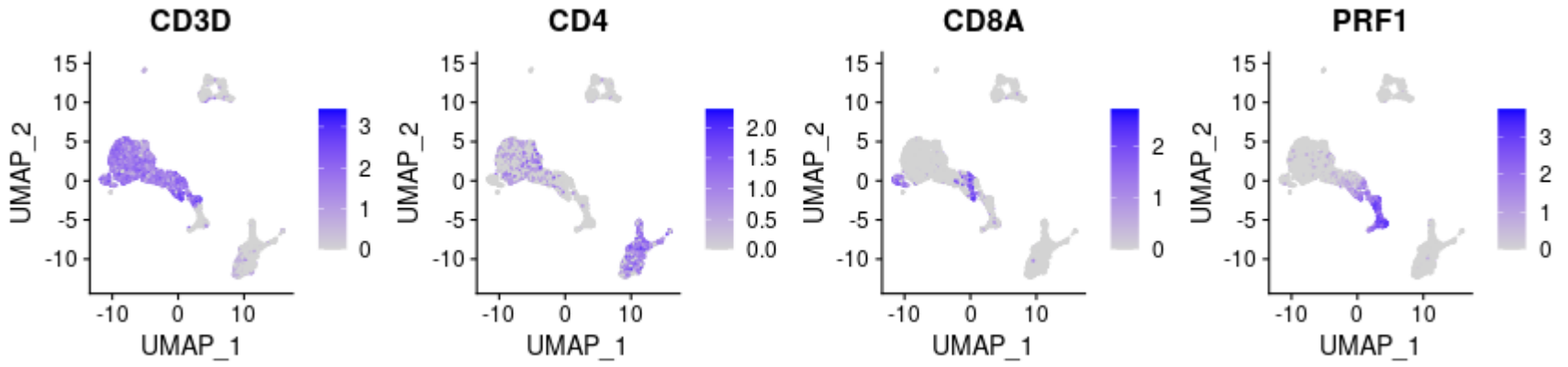
- One cluster against the other
- One cluster against all the other clusters (marker identification)
- One condition against the other (almost bulk RNA-seq)
- Same cell type in different conditions

Comparison of T cells

- Based on the previous investigation we have 2 clusters of CD8 T cells: 3 and 11, which are close to each other
- Lets figure out what's the difference

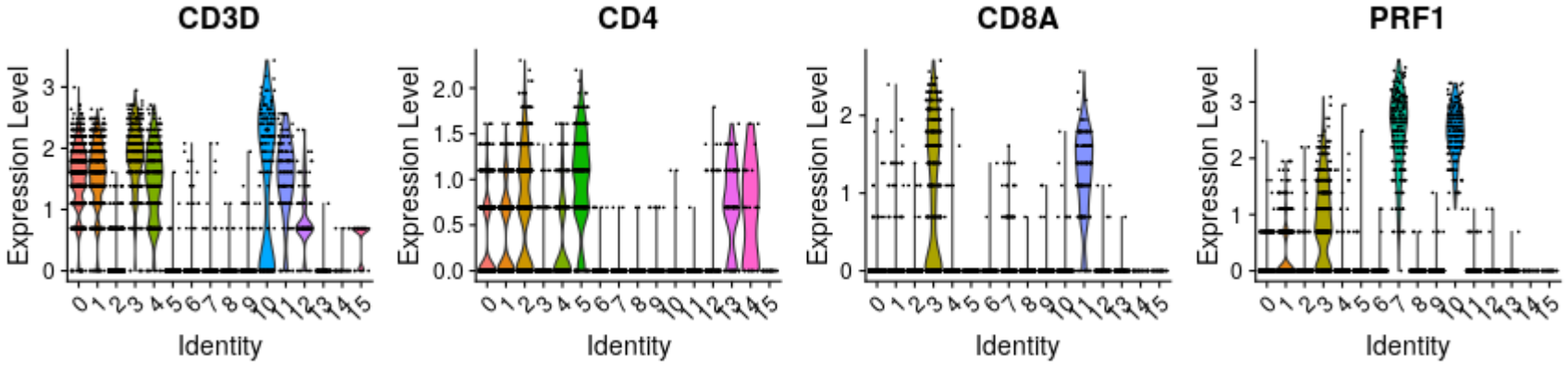
Comparison of T cells

```
FeaturePlot(seurat, features=c("CD3D", "CD4", "CD8A", "PRF1"), ncol = 4)
```



Comparison of T cells

```
VlnPlot(seurat, features=c("CD3D", "CD4", "CD8A", "PRF1"), ncol = 4, pt.size = 0.02)
```



Comparison of T cells

- We will compare population using differential expression
- This will generate a table with many important fields

MAST test

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*]

Abstract

Single-cell transcriptomics reveals gene expression heterogeneity but suffers from stochastic dropout and characteristic bimodal expression distributions in which expression is either strongly non-zero or non-detectable. We propose a two-part, generalized linear model for such bimodal data that parameterizes both of these features. We argue that the cellular detection rate, the fraction of genes expressed in a cell, should be adjusted for as a source of nuisance variation. Our model provides gene set enrichment analysis tailored to single-cell data. It provides insights into how networks of co-expressed genes evolve across an experimental treatment. MAST is available at https://github.com/RGLab/MAST.

Keywords: Bimodality, Cellular detection rate, Co-expression, Empirical Bayes, Generalized linear model, Gene set enrichment analysis

Differential expression

```
de_03_vs_11 <- FindMarkers(
  seurat, assay="SCT", ident.1 = 3, ident.2 = 11,
  test="MAST", logfc.threshold = 0, min.pct = 0
)
write.table(de_03_vs_11, "de_03_vs_11.tsv", sep="\t", col.names=NA, quote=F)
topGenes <- head(rownames(de_03_vs_11))
```

Differential expression

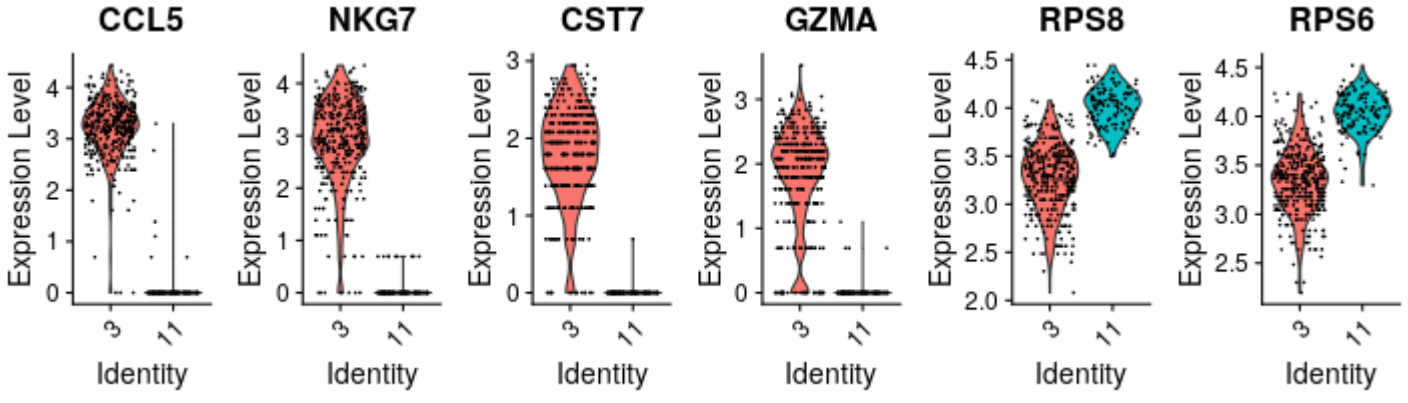
```
head(de_03_vs_11)
```

```
##              p_val  avg_logFC pct.1 pct.2     p_val_adj
## CCL5 1.115608e-113  3.0218880 0.992 0.048 1.787427e-109
## NKG7 1.339048e-107  3.0873554 0.978 0.071 2.145422e-103
## CST7  1.276064e-98  1.9549712 0.958 0.008  2.044509e-94
## GZMA  1.080751e-82  2.0257942 0.914 0.032  1.731579e-78
## RPS8  4.496728e-76 -0.6880436 1.000 1.000  7.204657e-72
## RPS6  1.542867e-75 -0.6595797 1.000 1.000  2.471982e-71
```

- avg_logFC - average log fold change
- p_val - p value (bad)
- p_val_adj - p value adjusted for multiple hypothesis (good)
- pct.1 - % of cell in the first group (cluster 3) that have non-zero expression values of gene
- pct.2 - % of cell in the first group (cluster 11) that have non-zero expression values of gene

Differential expression: visualized

```
VlnPlot(seurat, topGenes, pt.size = 0.02, idents=c(3, 11), ncol=6)
```



Differential expression

In single-cell RNA-seq we will compare cell groups against each other:

- One cluster against the other (we just did it)
- One cluster against all the other clusters (marker identification) (we did it in the first part)

- One condition against the other (almost bulk RNA-seq)
- Same cell type in different conditions

Cd8 T cell investigation

- We got two clusters, run DE and know whats different
- What's next?

Pathway enrichment

By marker expression we know:

- Cluster 3 is (activated ?) Cd8 T cells
- Cluster 11 is (naïve/memory ?) Cd8 T cells

Is there a pathway that drive these transcriptional changes?

Is there a set of differentially expressed genes between these two groups?

Let's save top genes

```
de_03_vs_11$gene <- rownames(de_03_vs_11)

top50 <- de_03_vs_11 %>% top_n(50, avg_logFC) %>% pull(gene)
top200 <- de_03_vs_11 %>% top_n(200, avg_logFC) %>% pull(gene)
bottom50 <- de_03_vs_11 %>% top_n(50, -avg_logFC) %>% pull(gene)
bottom200 <- de_03_vs_11 %>% top_n(200, -avg_logFC) %>% pull(gene)

writeLines(top50, "top_50.txt")
writeLines(top200, "top_200.txt")
writeLines(bottom50, "bottom_50.txt")
writeLines(bottom200, "bottom_200.txt")
```

msigdb

- Lets open top50.txt
- Lets search for the pathways
- http://software.broadinstitute.org/gsea/msigdb/annotate.jsp

msigdb

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

To cite your use of this page, please reference this website and also the MSigDB (see Citing MSigDB on the main MSigDB page)

Input Gene Identifiers

(case sensitive)

```
TRGC2
SRGN
AHNAK
NEAT1
PPP2R5C
S100A11
CYTOR
CCL4
ZEB2
SYNE2
CTSW
CD74
HLA-DRB1
HLA-DPB1
KLF6
KLRB1
IFNG
FGFBP2
TRGC1
GZMB
CMC1
PMAIP1
LGALS1
TRDC
GNLY
IFIT2
```

Species: Human

Compute Overlaps

[about the MSigDB collections]

- [x] H: hallmark gene sets
- [ ] C1: positional gene sets
- [ ] C2: curated gene sets
  - [ ] CGP: chemical and genetic perturbations
  - [ ] CP: canonical pathways
    - [ ] CP:BIOCARTA: BioCarta gene sets
    - [ ] CP:KEGG: KEGG gene sets
    - [ ] CP:PID: PID gene sets
    - [ ] CP:REACTOME: Reactome gene sets
    - [ ] CP:WIKIPATHWAYS: WikiPathways gene sets
- [ ] C3: regulatory target gene sets
  - [ ] MIR: microRNA targets
    - [ ] MIR:MIR_Legacy: legacy microRNA targets
    - [ ] MIR:MIRDB: MIRDB microRNA targets
  - [ ] TFT: all transcription factor targets
    - [ ] TFT:GTRD: GTRD transcription factor targets
    - [ ] TFT:TFT_Legacy: legacy transcription factor targets
- [ ] C4: computational gene sets
  - [ ] CGN: cancer gene neighborhoods
  - [ ] CM: cancer modules
- [ ] C5: ontology gene sets
  - [ ] GO: Gene Ontology

Compendia Expression Profiles

- (•) GTEx compendium
- ( ) Human tissue compendium (Novartis)
- ( ) Global Cancer Map (Broad Institute)
- ( ) NCI-60 cell lines (National Cancer Institute)

display expression profile

Gene Families

show gene families

NDEx Biological Network Repository

query NDEx

msigdb results

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

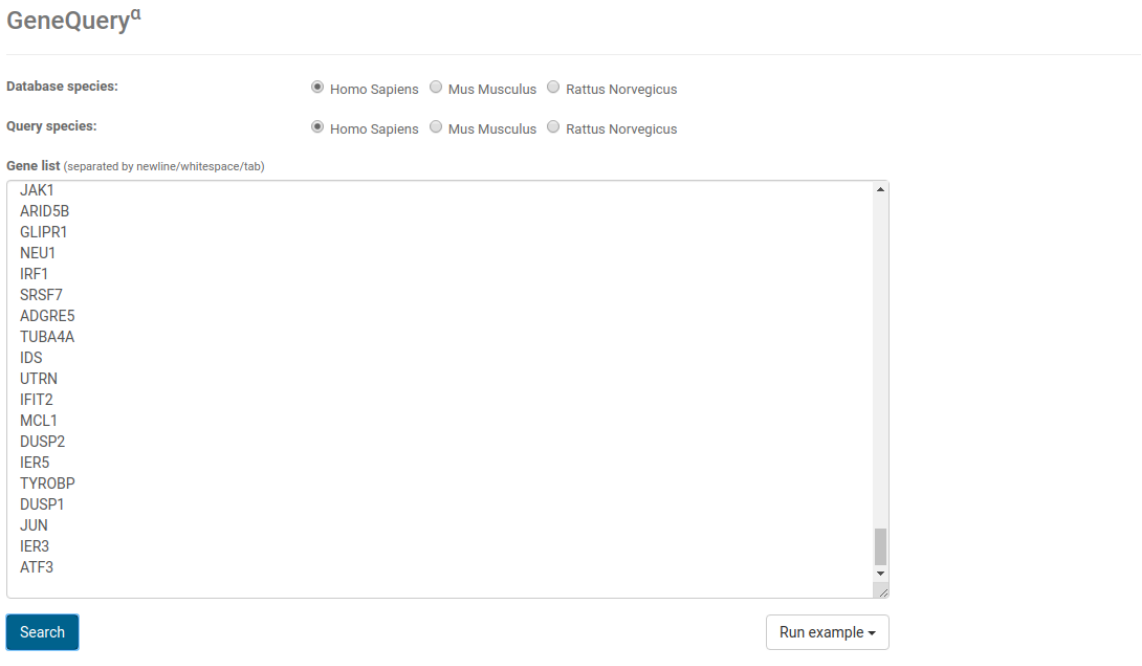
Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
HALLMARK_ALLOGRAFT_REJECTION [200]	Genes up-regulated during transplant rejection.	10		$6.6\ e^{-14}$	$3.3\ e^{-12}$
HALLMARK_COMPLEMENT [200]	Genes encoding components of the complement system, which is part of the innate immune system.	6		$1.9\ e^{-7}$	$4.74\ e^{-6}$
HALLMARK_IL2_STAT5_SIGNALING [199]	Genes up-regulated by STAT5 in response to IL2 stimulation.	5		$5.08\ e^{-6}$	$5.2\ e^{-5}$
HALLMARK_INTERFERON_GAMMA_RESPONSE [200]	Genes up-regulated in response to IFNG [GeneID=3458].	5		$5.2\ e^{-6}$	$5.2\ e^{-5}$
HALLMARK_TNFA_SIGNALING_VIA_NFKB [200]	Genes regulated by NF-kB in response to TNF [GeneID=7124].	5		$5.2\ e^{-6}$	$5.2\ e^{-5}$
HALLMARK_APOPTOSIS [161]	Genes mediating programmed cell death (apoptosis) by activation of caspases.	3		$1.09\ e^{-3}$	$9.05\ e^{-3}$
HALLMARK_HYPOXIA [200]	Genes up-regulated in response to low oxygen levels (hypoxia).	3		$2.02\ e^{-3}$	$1.44\ e^{-2}$
HALLMARK_INTERFERON_ALPHA_RESPONSE [97]	Genes up-regulated in response to alpha interferon proteins.	2		$6.59\ e^{-3}$	$4.12\ e^{-2}$

GeneQuery

- Lets open top 200 genes upregulated in activated T cells
- Lets search for hits in GeneQuery
- http://artyomovlab.wustl.edu/genequery/searcher/

GeneQuery

- http://artyomovlab.wustl.edu/genequery/searcher/



GeneQuery

- http://artyomovlab.wustl.edu/genequery/searcher/

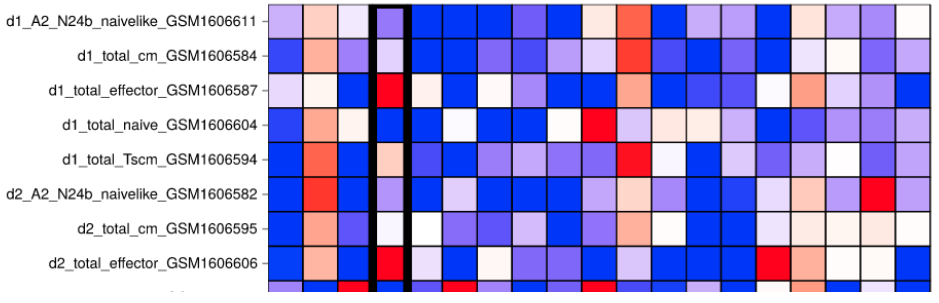
#	Experiment title	Module	$\log_{10}(\text{adj.}p_{\text{value}})$	Overlap	GSE	GMT
1	Nave-like Yellow-Fever specific CD8 T cells and reference CD8 T cell subsets in humans	3	-60.92	92/399	GSE65804	
2	Peripheral blood mononuclear cell gene expression in chronic obstructive pulmonary disease	6	-49.51	66/194	GSE42057	
3	MicroRNA regulate immune pathways in T-cells in multiple sclerosis (MS)	4	-49.09	78/307	GSE43592	
4	Comparison of transcriptional profiles of CD4+ and CD8+ T cells from HIV-infected pateints and uninfected control group	5	-45.61	81/422	GSE6740	
5	Phenotype, Function and Gene Expression Profiles of PD-1 high CD8 T cells in Healthy Human Adults	6	-45.29	79/344	GSE26495	
6	Distinct, non-overlapping gene panels of peripheral blood gene expression predict response to infliximab therapy in rheumatoid arthritis and Crohn's disease	10	-43.25	58/171	GSE42296	
7	Identification and characterization of human Natural Killer (NK) lineage restricted progenitors	2	-41.67	139/1581	GSE60448	
8	Absence of significant overlap in transcriptional patterns between operationally tolerant liver and kidney recipients	10	-40.23	59/186	GSE22707	
9	Gene expressions of CD4+ T cells in each developmental stages	3	-39.76	84/558	GSE61697	
10	Lack of effect in desensitization with intravenous immunoglobulin and rituximab in highly-sensitized patients	8	-39.13	49/126	GSE31729	

GeneQuery

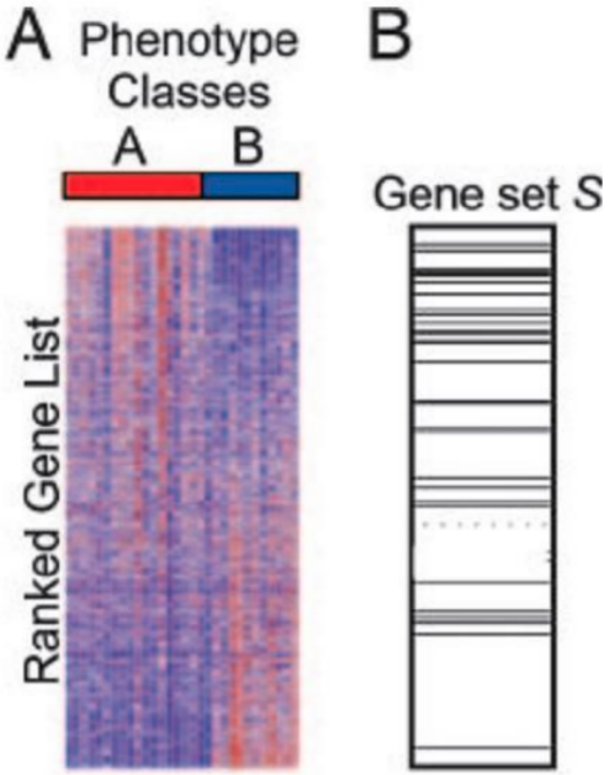
- http://artyomovlab.wustl.edu/genequery/searcher/

GeneQuery

- http://artyomovlab.wustl.edu/genequery/searcher/



Pathway enrichment



Enrichment score function

→ ES(S)

↓ Empirical null distribution from random sets

P-value & Normalized Enrichment Score (NES)

Partially taken from: https://www.pnas.org/content/102/43/15545

FGSEA

```
load("keggSymbolHuman.rdata")

ranks <- de_03_vs_11$avg_logFC
names(ranks) <- rownames(de_03_vs_11)
fgseaRes <- fgsea(pathways = keggSymbolHuman,
                  stats = ranks,
                  minSize=15,
                  maxSize=500,
                  nperm=100000)
```

```
## Warning in fgsea(pathways = keggSymbolHuman, stats = ranks, minSize = 15, : You
## are trying to run fgseaSimple. It is recommended to use fgseaMultilevel. To run
## fgseaMultilevel, you need to remove the nperm argument in the fgsea function
## call.
```

```
## Warning in preparePathwaysAndStats(pathways, stats, minSize, maxSize, gseaParam, : There are ties
## The order of those tied genes will be arbitrary, which may produce unexpected results.
```

FGSEA

```
head(fgseaRes)
```

```
##                                                pathway      pval      padj
## 1:    Glycolysis / Gluconeogenesis - Homo sapiens (human) 0.2362735 0.4231303
## 2:       Citrate cycle (TCA cycle) - Homo sapiens (human) 0.5767097 0.7123113
## 3:       Pentose phosphate pathway - Homo sapiens (human) 0.3108722 0.4981228
## 4: Fructose and mannose metabolism - Homo sapiens (human) 0.5910692 0.7197909
## 5:            Galactose metabolism - Homo sapiens (human) 0.3624148 0.5642140
## 6:           Fatty acid elongation - Homo sapiens (human) 0.8069547 0.8747035
##            ES        NES nMoreExtreme size                          leadingEdge
## 1:  0.5021257  1.1830853        15822   43   GAPDH,PGAM1,GALM,LDHA,ENO1,PKM,...
## 2:  0.4206890  0.9168580        36379   26   IDH2,OGDH,SDHB,SDHA,MDH1,IDH3A,...
## 3: -0.4649284 -1.1164831        11625   24    TKT,PGD,RPIA,ALDOC,PRPS1,FBP1,...
## 4:  0.4090891  0.9068844        37763   29 PFKP,TPI1,TSTA3,GMPPB,PMM2,GMPPA,...
## 5:  0.5092343  1.0886743        22601   23               GALM,B4GALT1,PFKP,GLB1
## 6: -0.3196092 -0.7543024        30631   22       HADHA,PPT1,HSD17B12,TECR,HACD1
```

Using fgsea

```
topPathwaysUp <- fgseaRes[ES > 0 & padj < 0.01, ][head(order(pval), n=10), pathway]
topPathwaysDown <- fgseaRes[ES < 0 & padj < 0.01, ][head(order(pval), n=10), pathway]
topPathways <- c(topPathwaysUp, rev(topPathwaysDown))
```

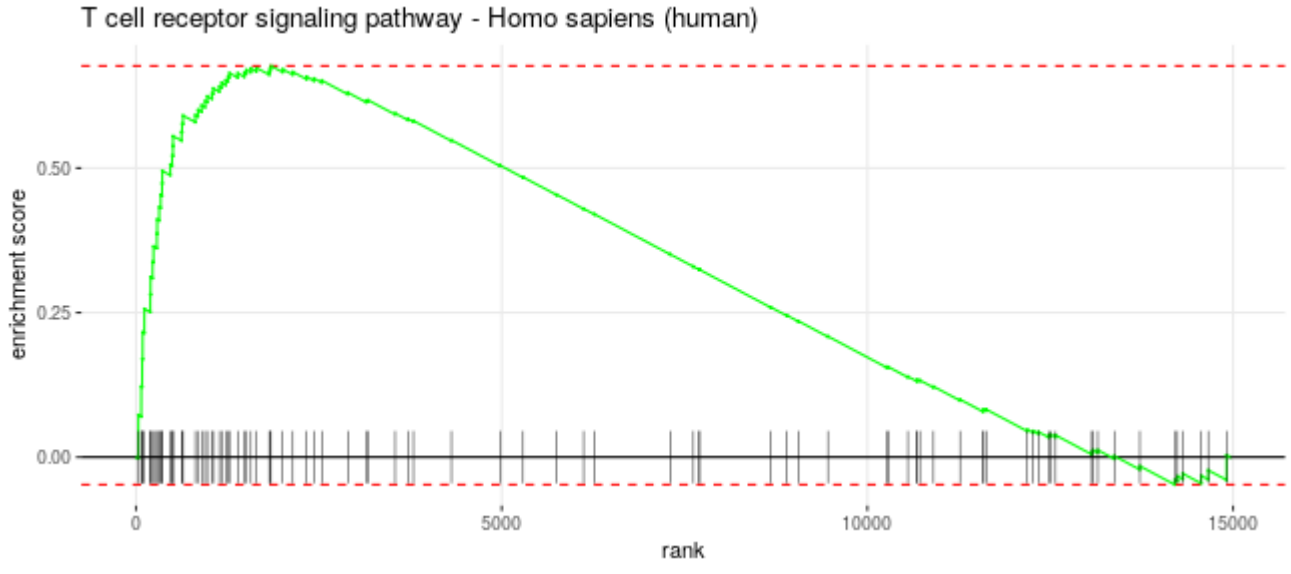
Using fgsea

```
plotGseaTable(keggSymbolHuman[topPathways], ranks, fgseaRes,
              gseaParam = 0.2, colwidths = c(5, 1, 0.8, 0.8, 0.8))
```

Pathway	Gene ranks	NES	pval	padj
Epstein-Barr virus infection - Homo sapiens (human)		1.86	1.2e-05	2.8e-04
Herpes simplex infection - Homo sapiens (human)		1.92	1.2e-05	2.8e-04
Regulation of actin cytoskeleton - Homo sapiens (human)		1.95	1.2e-05	2.8e-04
Influenza A - Homo sapiens (human)		1.86	1.3e-05	2.8e-04
Tuberculosis - Homo sapiens (human)		1.94	1.3e-05	2.8e-04
NOD-like receptor signaling pathway - Homo sapiens (human)		1.88	1.3e-05	2.8e-04
Apoptosis - Homo sapiens (human)		1.89	1.3e-05	2.8e-04
Phagosome - Homo sapiens (human)		2.05	1.3e-05	2.8e-04
Natural killer cell mediated cytotoxicity - Homo sapiens (human)		2.08	1.3e-05	2.8e-04
Th1 and Th2 cell differentiation - Homo sapiens (human)		1.98	1.4e-05	2.8e-04
Ribosome - Homo sapiens (human)		-2.91	4.6e-05	6.1e-04

Using fgsea

```
plotEnrichment(keggSymbolHuman[["T cell receptor signaling pathway - Homo sapiens (human)
ranks) + labs(title="T cell receptor signaling pathway - Homo sapiens (human)")
```



Summary

- We have many ways to annotate gene sets, if it's hard to annotate by markers
- Differential expression is one of key ways to do that
- Once we have differential expression results we have many ways to annotate transcriptional differences with the pathways

Questions?