ITMO UNIVERSITY

CT Lab
ITMO UNIVERSITY

Exploring gene expression datasets

Alexey Sergushichev

Sep 23, Nice

About the module

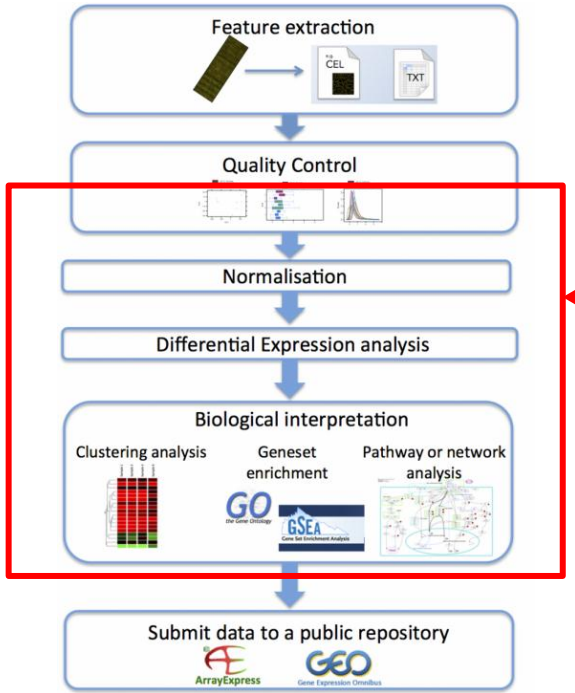
- We will cover the basic analysis of gene expression matrices
  - No working with raw data
- The focus is on being able to do a quick analysis, not the perfect one

- Materials and slides are available at Google Drive

Outline

- **Exploring gene expression datasets**
- Simple analysis methods
- Working with public datasets

Overall gene expression pipeline

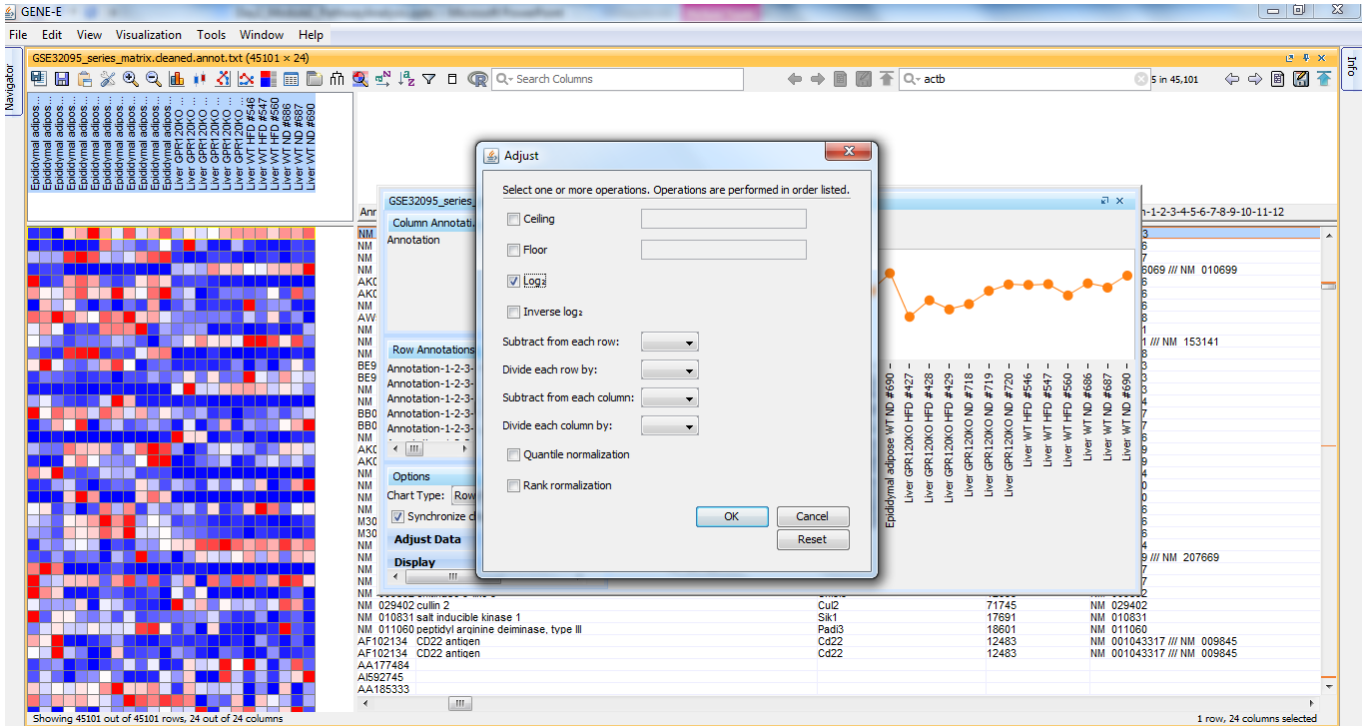


← This workshop

https://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/analysis-microarray-data

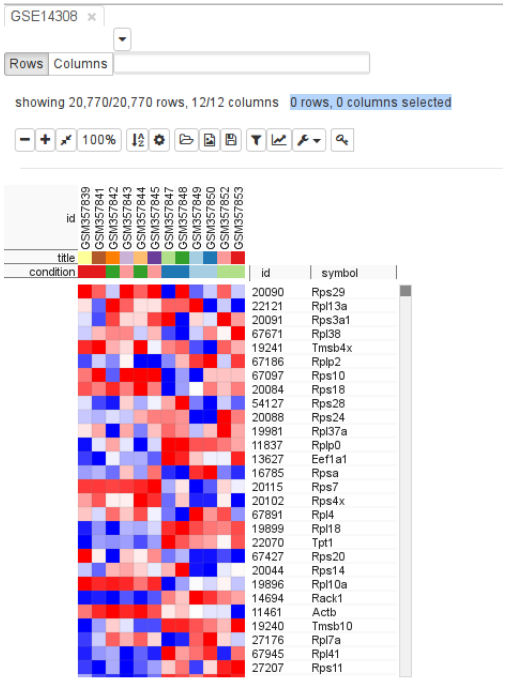
ITMO UNIVERSITY

GENE-E: software for working with gene expression data (obsolete)



Morpheus – heatmap visualization software replacing GENE-E

- Developed at Broad Institute (by Joshua Gould)
- Works in browser
- Fully client-side application
  - data is not sent to server!
- Open source
- Limited functionality



https://software.broadinstitute.org/morpheus/

Phantasus – Morpheus integrated with R environment

- An extension developed by Daria Zenkova & Vlad Kamenev at ITMO University
- Server-side application -> requires internet access
  - unless installed locally
- Can be easily extended to support different R/Bioconductor packages
- Free and open-source

- Feedback is welcome!

https://github.com/ctlab/phantasus

Phantasus can be accessed in multiple ways

Online:

- https://ctlab.itmo.ru/phantasus/
- https://artyomovlab.wustl.edu/phantasus/

It can be installed locally from Bioconductor

- http://bioconductor.org/packages/phantasus

As a docker image:

- https://hub.docker.com/r/dzenkova/phantasus

Where datasets are coming from?

- From papers!

LETTER

doi:10.1038/nature13152

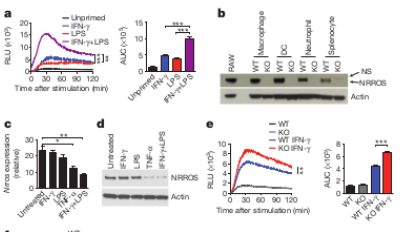
NRROS negatively regulates reactive oxygen species during host defence and autoimmunity

Rajkumar Noubade[1]†, Kit Wong[1], Naruhisa Ota[1], Sascha Rutz[1], Celine Eidenschenk[1], Patricia A. Valdez[1]†, Jiabing Ding[1], Ivan Peng[1], Andrew Sebrell[2], Patrick Caplazi[3], Jason DeVoss[1], Robert H. Soriano[4], Tao Sai[2], Rongze Lu[1], Zora Modrusan[4], Jason Hackney[5] & Wenjun Ouyang[1]

Reactive oxygen species (ROS) produced by phagocytes are essential for host defence against bacterial and fungal infections. Individuals with defective ROS production machinery develop chronic granulomatous disease[1,2]. Conversely, excessive ROS can cause collateral tissue damage during inflammatory processes and therefore needs to be tightly regulated. Here we describe a protein, we termed negative regulator of ROS (NRROS), which limits ROS generation by phagocytes during inflammatory responses. NRROS expression in phagocytes can be repressed by inflammatory signals. NRROS-deficient phagocytes produce increased ROS upon inflammatory challenges, and mice lacking NRROS in their phagocytes show enhanced bactericidal activity against *Escherichia coli* and *Listeria monocytogenes*. Conversely, these mice develop severe experimental autoimmune encephalomyelitis owing to oxidative tissue damage in the central nervous system. Mechanistically, NRROS is localized to the endoplasmic reticulum, where it directly interacts with nascent NOX2 (also known as gp91^{phox} and encoded by *Cybb*) monomer, one of the membrane-bound subunits of the NADPH oxidase complex, and facilitates the degradation of NOX2 through the endoplasmic-reticulum-associated degradation pathway. Thus, NRROS provides a hitherto undefined mechanism for regulating ROS prodution—one that enables phagocytes to produce higher amounts of ROS, if required to control invading pathogens, while minimizing unwanted collateral tissue damage.

In response to microorganisms and inflammatory stimuli, professional phagocytes can generate ROS either within mitochondria or through a process named oxidative burst mediated by the NADPH oxidase 2 (NOX2) complex[1–3]. Although many regulatory factors for (Fig. 1b and Extended Data Fig. 1d, e). Interestingly, priming with a combination of IFN-γ and LPS or tumour necrosis factor (TNF)-α alone markedly repressed *Nrros* messenger RNA and protein expression in wild-type BMDMs (Fig. 1c, d).

To reveal the biological functions of NRROS, we generated NRROS-specific antibody and NRROS-deficient mice (Extended Data Fig. 1f–j). At 6 weeks of age, all mice were viable and immune organs and leukocyte subsets were indistinguishable from those of wild-type mice (Extended Data Table 1 and data not shown). However, significantly augmented ROS production was observed from NRROS-deficient primary BMDMs upon zymosan stimulation after priming for 24 h with either IFN-γ (Fig. 1e) or LPS (Fig. 1f). These observations were confirmed in a variety of phagocytes, under several priming and activation



There is a mention of microarray

tion in phagocytes. Gene expression analysis by microarray under these conditions identified a previously uncharacterized gene, EMSMUSG 00000052384, which we named *Nrros* (negative regulator of ROS, previously known as *Lrrc33*) that was markedly downregulated upon priming with a combination of IFN-γ and LPS (Extended Data Fig. 1a). The

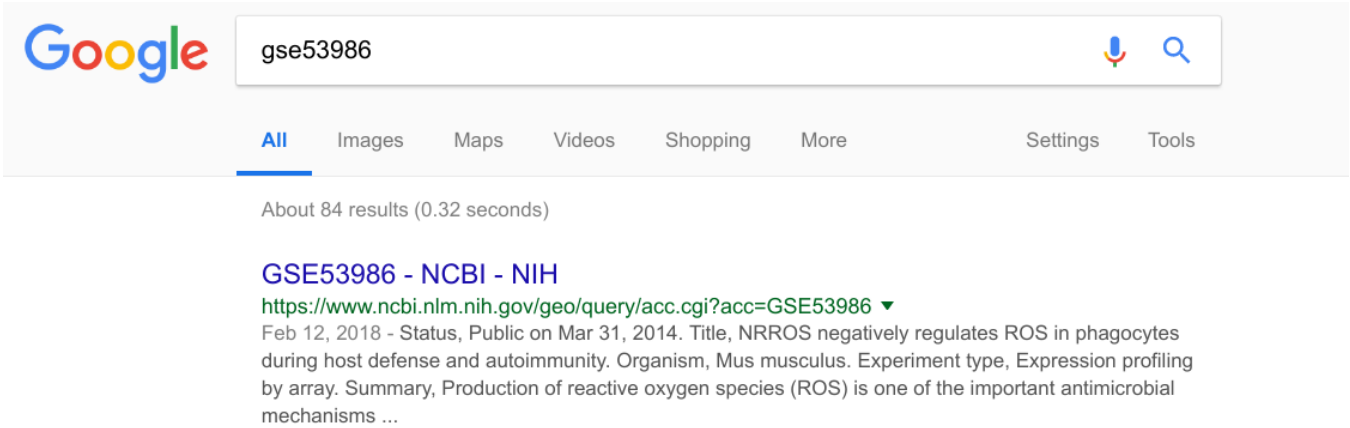
The data should be available from somewhere!

Check the methods section

“accession number GSE53986”

Microarray analysis. Statistical analyses of microarray data were performed using the R programming language (http://r-project.org). Microarray data were normalized using the RMA method[27]. Data were prefiltered to remove probes that were not mapped to an annotated Entrez gene. We also filtered our data to retain only a single probe per gene, selecting the probe with the highest variance, if multiple probes were found for the gene[28]. For differential expression analysis, the limma R package was used[29]. We modelled the synergistic regulation of gene expression by the combined IFN-γ and LPS treatment as an interaction term in our linear model. This model will identify changes that are significantly different from the sum of the individual treatments. Multiple test correction was done using the method of Benjamini and Hochberg[30]. Genes were considered significantly different if they changed more than 1.4-fold at a false discovery rate of 0.05. Genes were further filtered for immune-cell-specific expression using the gene sets defined by the Immune Response In Silico (IRIS) project[31]. As the IRIS-defined gene sets were derived from human immune cells, we mapped the human genes to mouse orthologues using the HomoloGene database[32]. Genes from all IRIS-defined categories were included in the analysis. Data were submitted to the NCBI (accession number GSE53986).

Let’s google that



Let's look at GSE53986

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53986

Series GSE53986	Query DataSets for GSE53986
Status	Public on Mar 31, 2014
Title	NRROS negatively regulates ROS in phagocytes during host defense and autoimmunity
Organism	Mus musculus
Experiment type	Expression profiling by array
Summary	Production of reactive oxygen species (ROS) is one of the important antimicrobial mechanisms of phagocytic cells. Enhanced oxidative burst requires these cells to be primed with agents such as IFNg and LPS with a synergistic effect of these agents on the level of the burst. However, excessive ROS generation will lead to tissue damage and has been implicated in a variety of inflammatory and autoimmune disease. Therefore, this process needs to be tightly regulated. In order to understand the genes regulating this process, we will treat bone marrow derived macrophages with above mentioned priming agents and study the gene expression. We used microarrays to determine the changes in gene expression that occur in bone marrow derived macrophages after treatment with IFNg, LPS, or a combination of IFNg and LPS
Overall design	Four condition experiment; Biological replicates: four replicates per condition
Contributor(s)	Noubade R, Wong K, Ota N, Rutz S, Eidenschenk C, Ding J, Valdez PA, Peng I, Sebrell A, Caplazi P, DeVoss J, Soriano RH, Modrusan Z, Hackney JA, Sai T, Ouyang W
Citation(s)	Noubade R, Wong K, Ota N, Rutz S et al. NRROS negatively regulates reactive oxygen species during host defence and autoimmunity. *Nature* 2014 May 8;509(7499):235-9. PMID: 24739962

Samples from GSE53986

Samples (16)
Less...

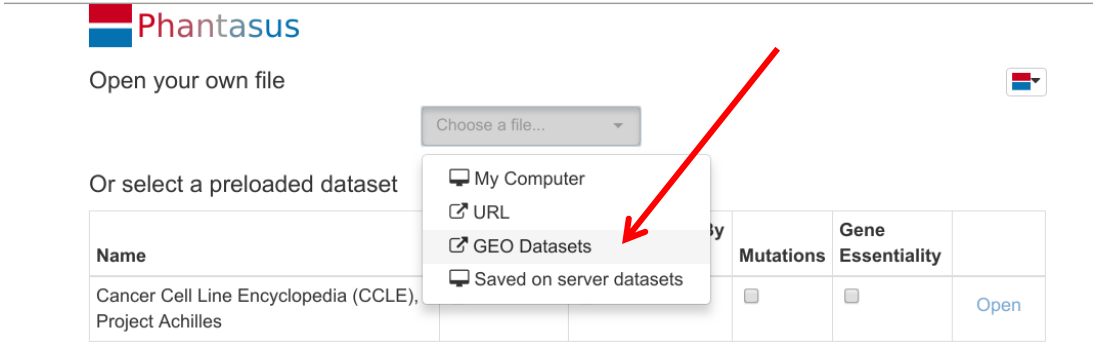
GSM1304836	BMDM, untreated, 1
GSM1304837	BMDM, untreated, 2
GSM1304838	BMDM, untreated, 3
GSM1304839	BMDM, untreated, 4
GSM1304840	BMDM, IFNg, 1
GSM1304841	BMDM, IFNg, 2
GSM1304842	BMDM, IFNg, 3
GSM1304843	BMDM, IFNg, 4
GSM1304844	BMDM, LPS, 1
GSM1304845	BMDM, LPS, 2
GSM1304846	BMDM, LPS, 3
GSM1304847	BMDM, LPS, 4
GSM1304848	BMDM, IFNg+LPS, 1
GSM1304849	BMDM, IFNg+LPS, 2
GSM1304850	BMDM, IFNg+LPS, 3
GSM1304851	BMDM, IFNg+LPS, 4

A lot of datasets can be found at GEO

(will come back to this later)

Let’s explore this dataset

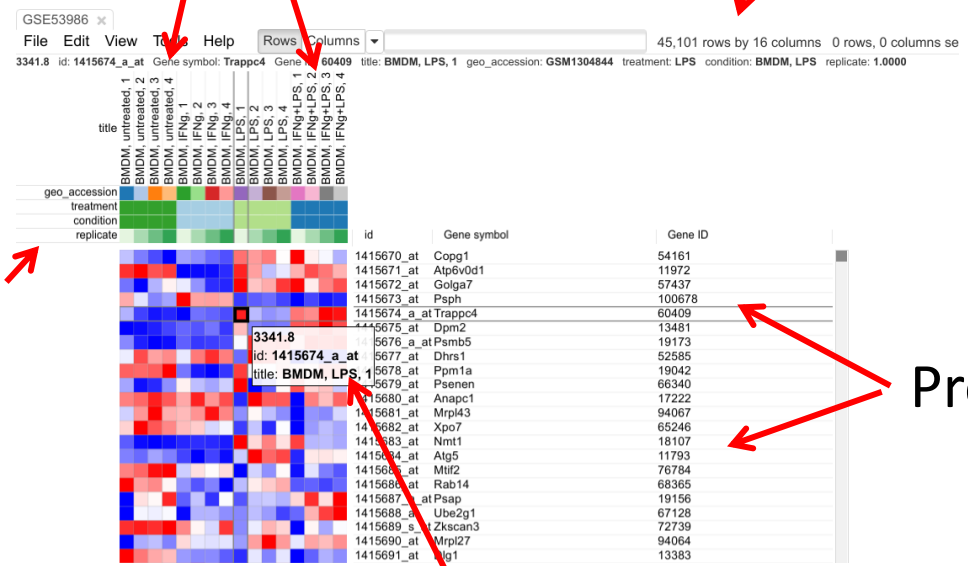
- Open https://ctlab.itmo.ru/phantasus/ or
- Open https://artyomovlab.wustl.edu/phantasus/
- Load dataset into phantasus:
  - Choose a file/GEO Datasets/GSE53986



Interface overview

Samples

Dataset dimension



Sample annotations (right click for context menu)

Probes/genes

Expression value (color scheme is relative)

Exploring individual genes

1. Enter “Acod1”

2. Click to scroll to the next hit

Row profile chart

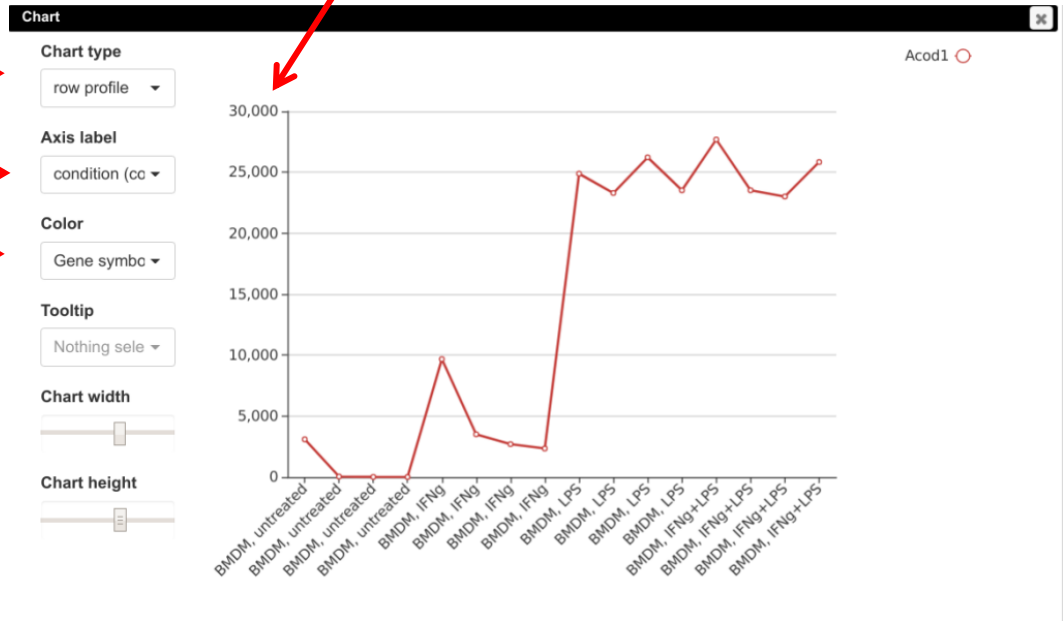
- Select all columns and Acod1 row
- Tools/Plots/Chart

Data is in linear scale!

row profile →

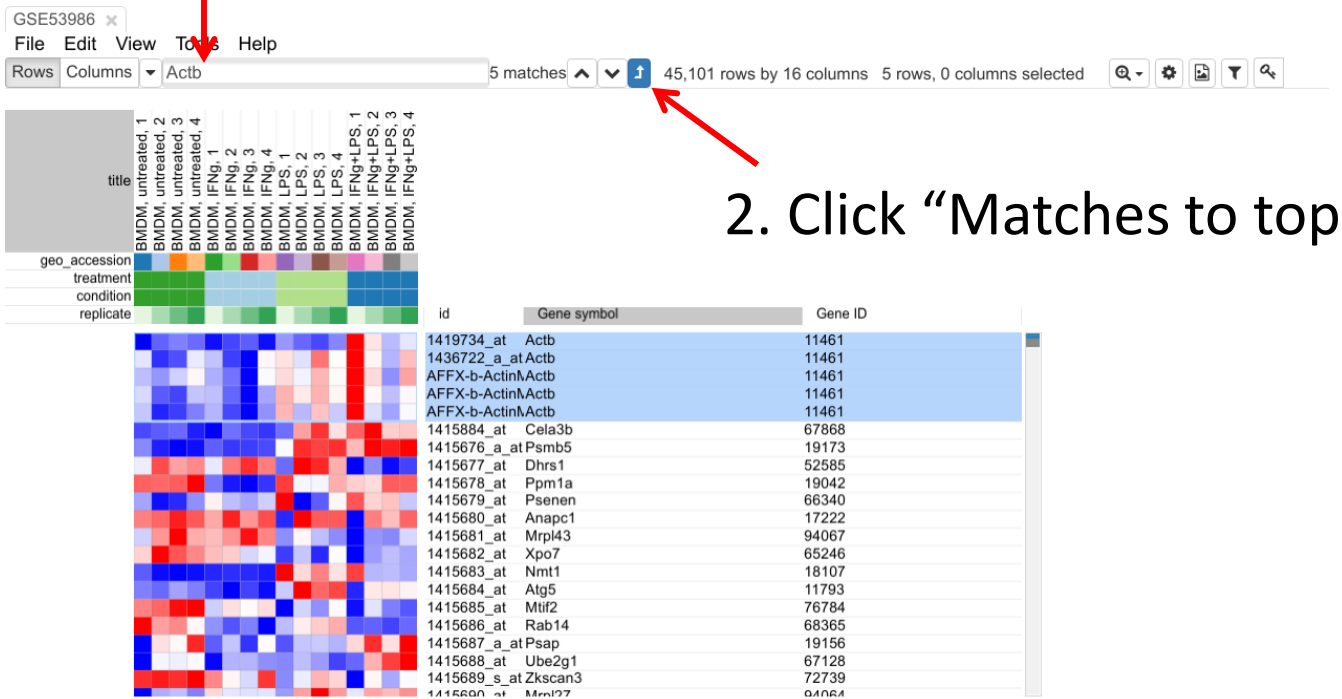
condition →

gene symbol →



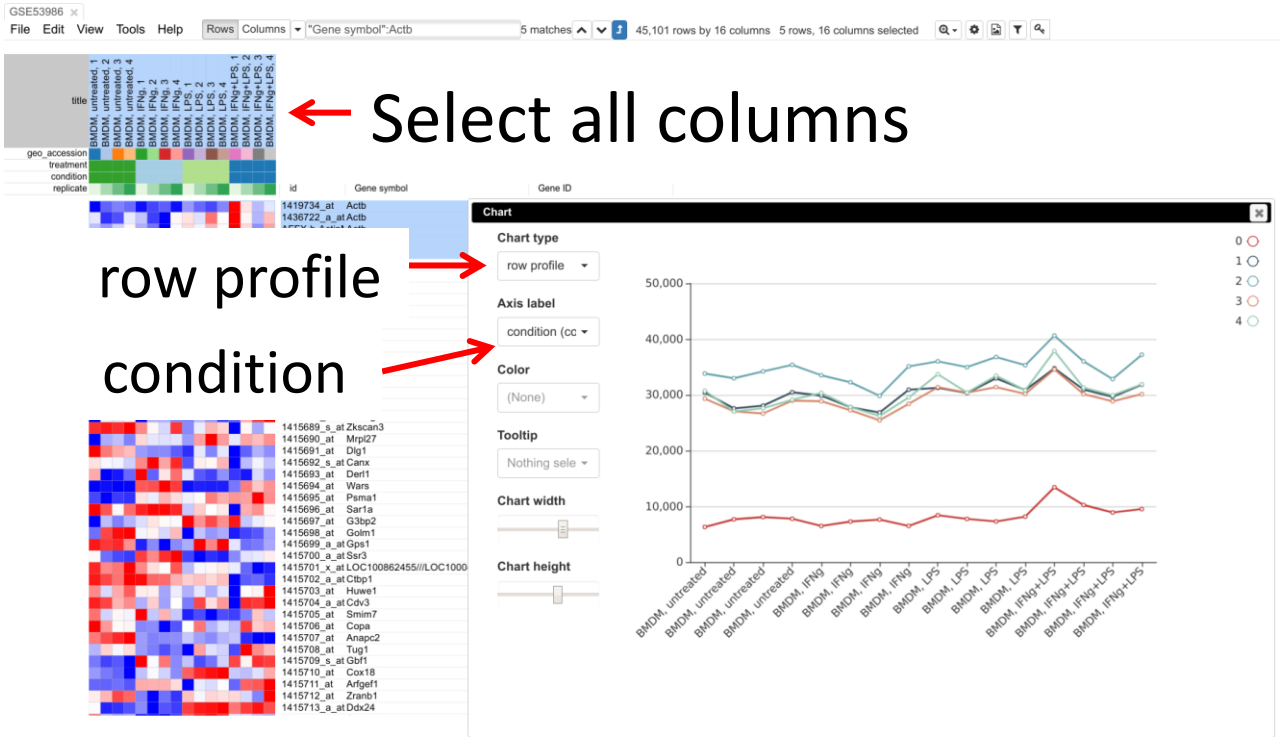
Let’s look at Actb as a control

1. Enter “Actb”



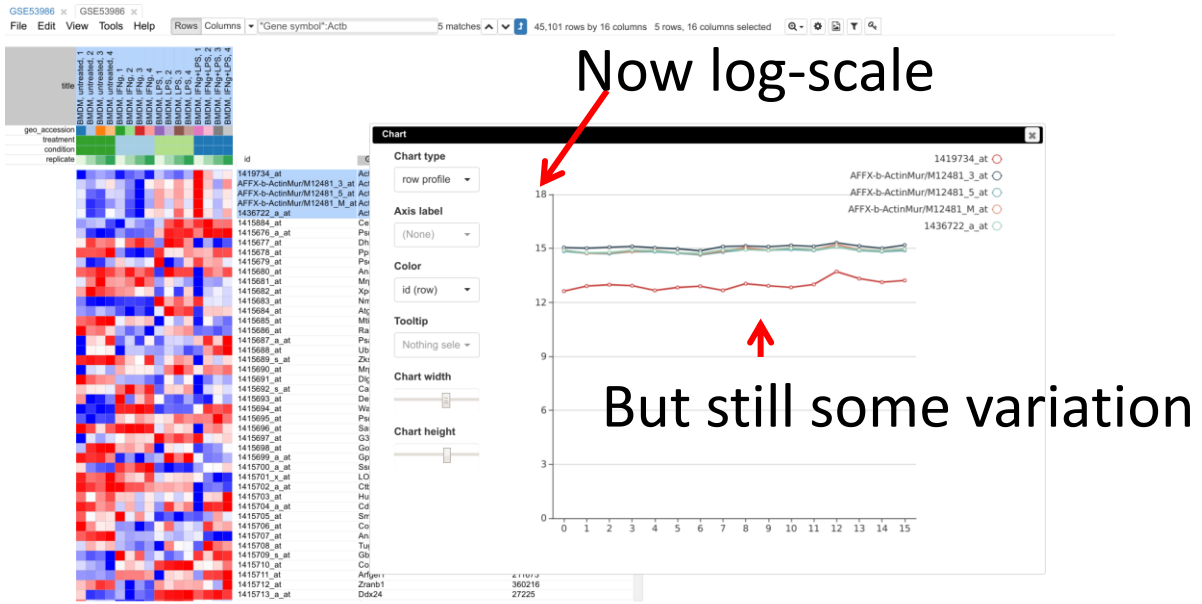
2. Click “Matches to top”

Actb expression chart: high variation (but in a linear scale)



Log 2 normalization

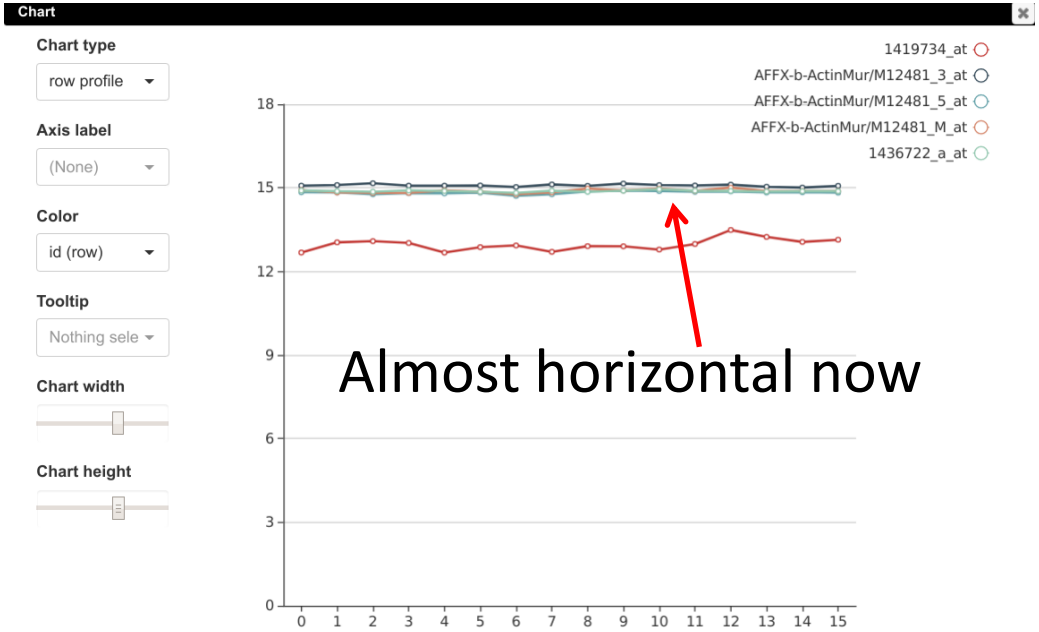
- Close the chart window
- Tools/Adjust, check “Log 2”
- Redo the plot



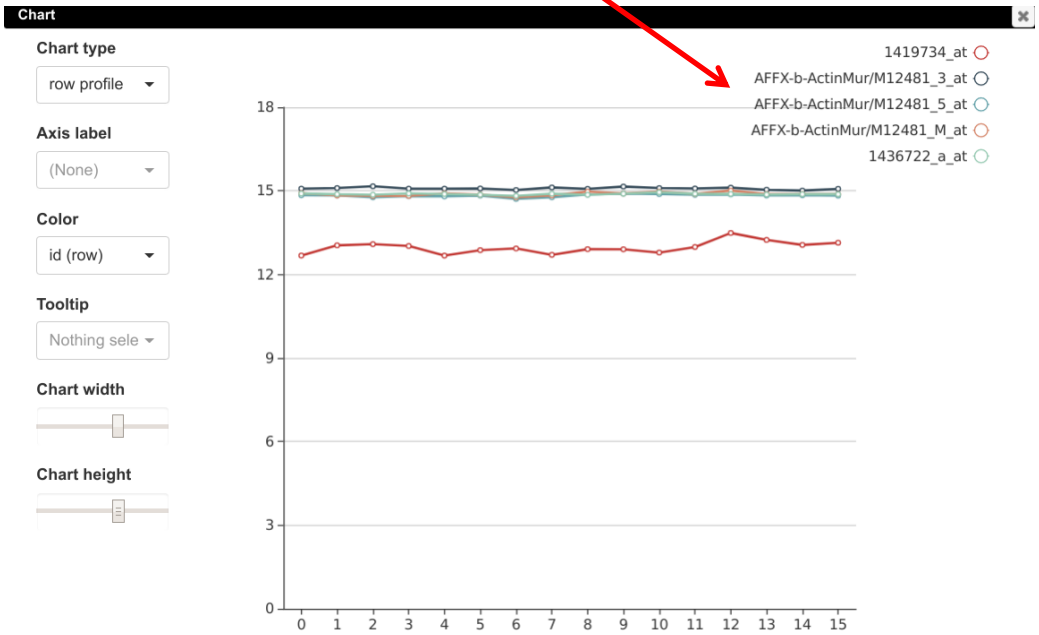
Quantile normalization

- Close the chart window
- Tools/Adjust, check “quantile”
- Redo the plot

- Log2 and quantile can be done in one step
- Don’t do Log2 twice, twice quantile is OK

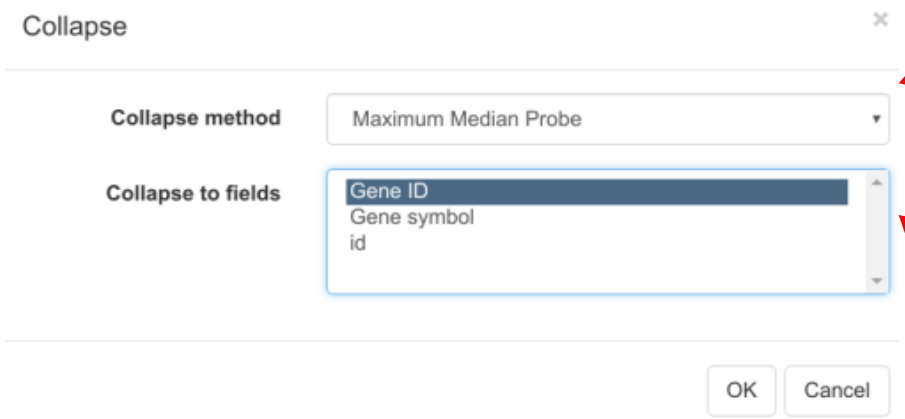


There are multiple probes per gene in microarrays



Collapsing duplicated probes to genes: keeping only one probe per gene

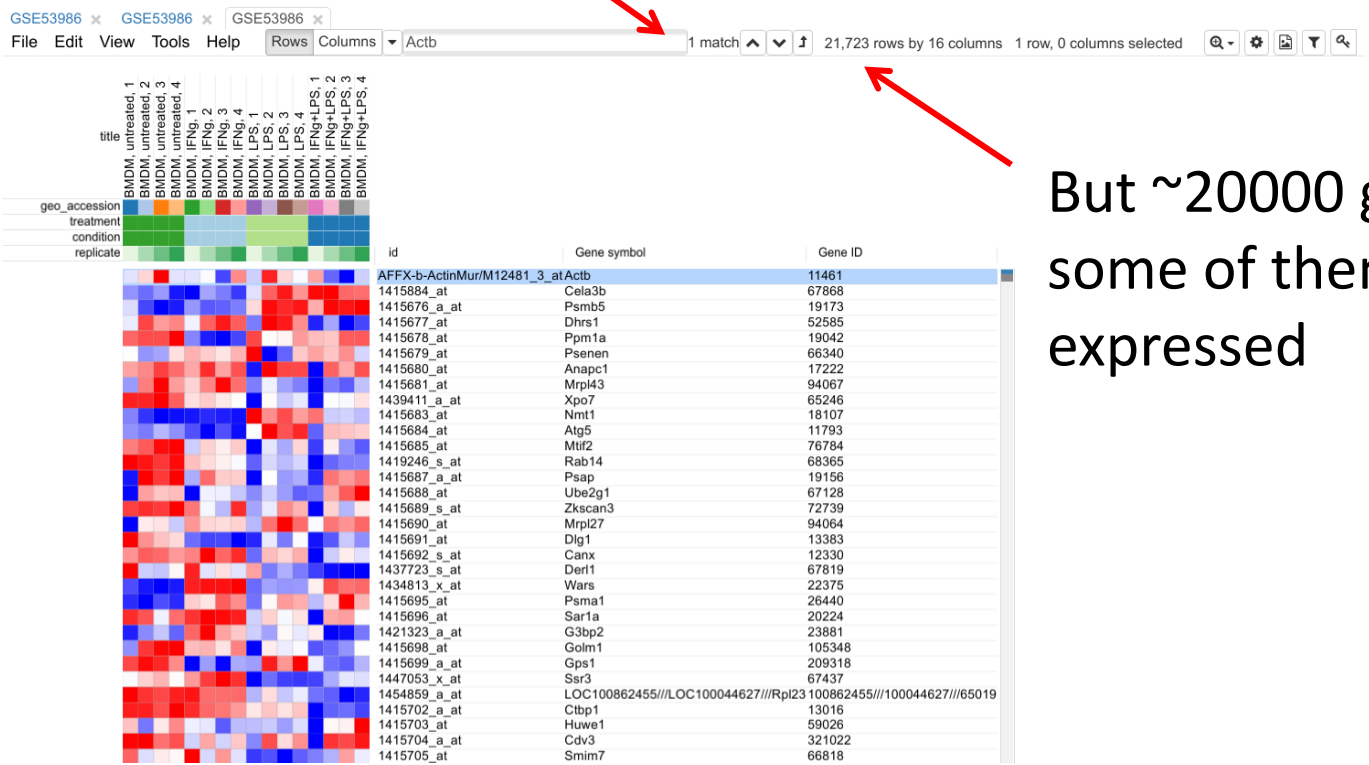
- Tools/Collapse



method = maximum median probe

Grouping by Gene ID

No more duplicates

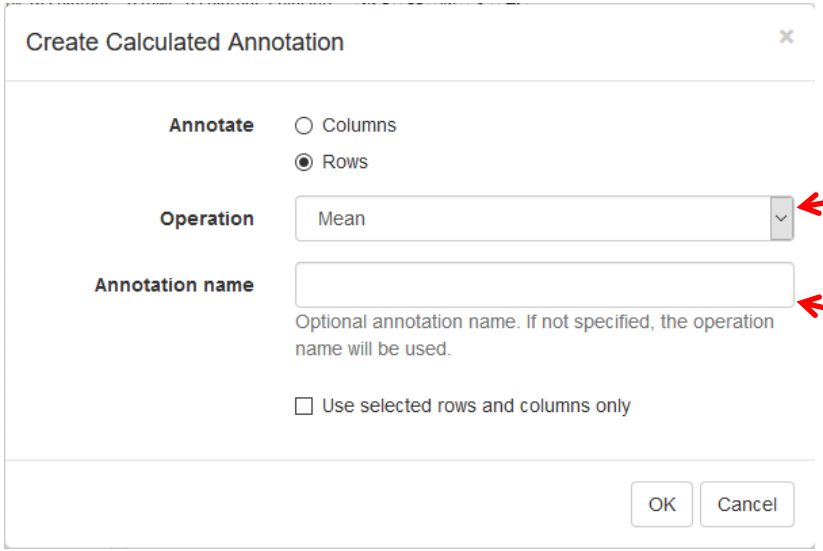


id	Gene symbol	Gene ID
AFFX-b-ActinMur/M12481_3_at	Actb	11461
1415884_at	Cela3b	67868
1415676_a_at	Psmb5	19173
1415677_at	Dhrs1	52585
1415678_at	Ppm1a	19042
1415679_at	Psenen	66340
1415680_at	Anapc1	17222
1415681_at	Mrpl43	94067
1439411_a_at	Xpo7	65246
1415683_at	Nmt1	18107
1415684_at	Atg5	11793
1415685_at	Mtif2	76784
1419246_s_at	Rab14	68365
1415687_a_at	Psap	19156
1415688_at	Ube2g1	67128
1415689_s_at	Zkscan3	72739
1415690_at	Mrpl27	94064
1415691_at	Dlg1	13383
1415692_s_at	Canx	12330
1437723_s_at	Derl1	67819
1434813_x_at	Wars	22375
1415695_at	Psma1	26440
1415696_at	Sar1a	20224
1421323_a_at	G3bp2	23881
1415698_at	Golm1	105348
1415699_a_at	Gps1	209318
1447053_x_at	Ssr3	67437
1454859_a_at	LOC100862455///LOC100044627///Rpl23	100862455///100044627///65019
1415702_a_at	Ctbp1	13016
1415703_at	Huwe1	59026
1415704_a_at	Cdv3	321022
1415705_at	Smim7	66818

But ~20000 genes, some of them are not expressed

Filtering lowly expressed genes: calculating mean expression

- Tools/Create Calculated Annotation



Operation: Mean

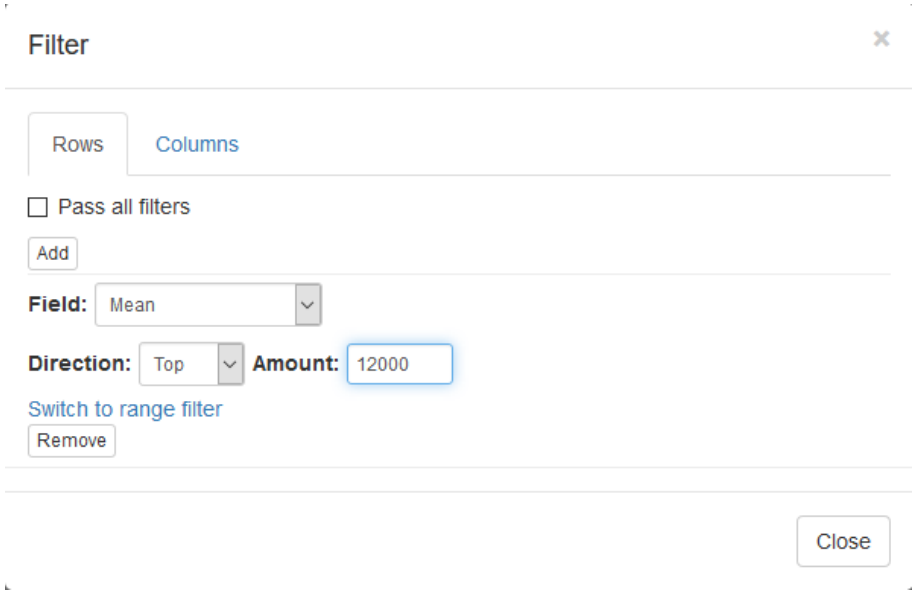
Optional name
(e.g. “mean_expression”)

Filtering lowly expressed genes: calculating mean expression result

id	Gene symbol	Gene ID	Mean
1452703_at	Ahcyl2	74340	11.278
1419731_at	Cyp2b19	13090	4.0218
1419732_at	Rtn4r	65079	6.0851
1419733_at	Glipr1l2	67537	2.3634
AFFX-b-ActinM	Actb	11461	15.087
1419735_at	Csn3	12994	5.1191
1419736_a_at	Eif1ax	66235	11.662
1419737_a_at	Ldha	16828	13.926
1419738_a_at	Tpm2	22004	7.9156

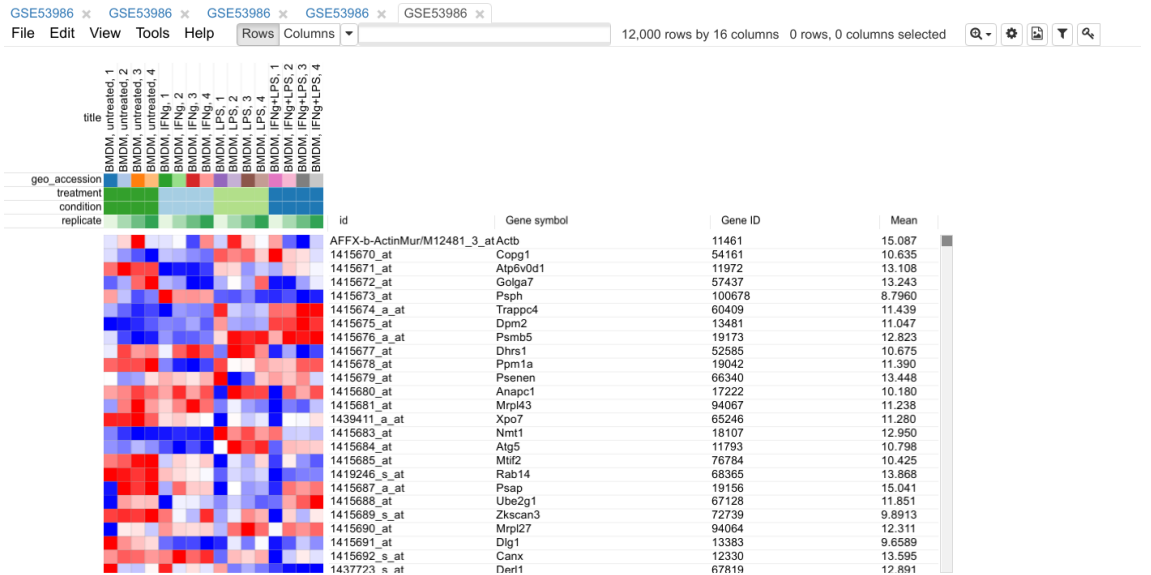
Filtering lowly expressed genes: keeping only top 12000 genes

- Tools/Filter
- Add
- Field <- Mean
- Switch to top filter
- N <- 12000



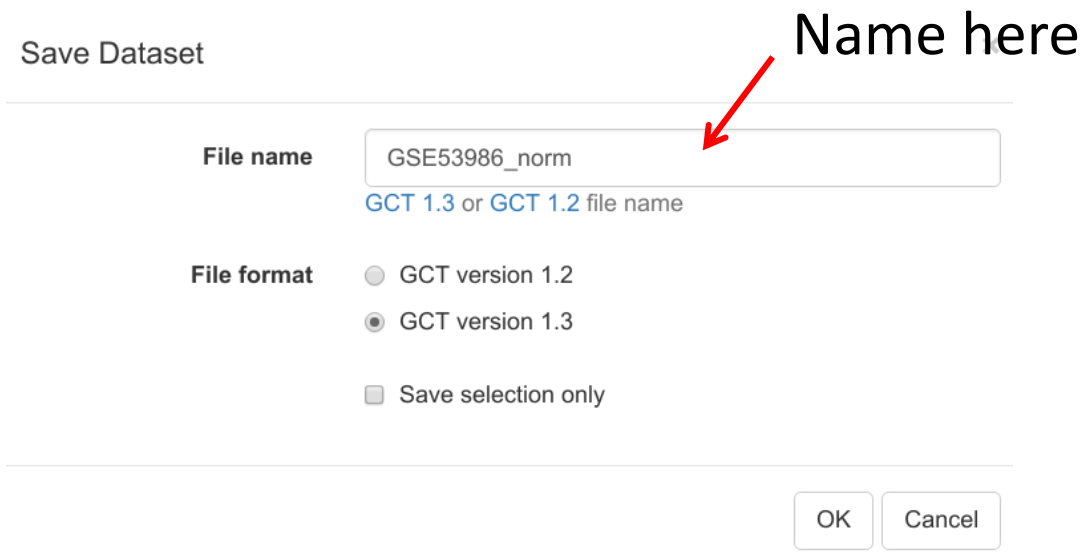
Filtering lowly expressed genes: creating new dataset

- Select all genes (click on any gene and Ctrl+A)
- Hit Ctrl-X to create new dataset (or Tools/New Heat Map)



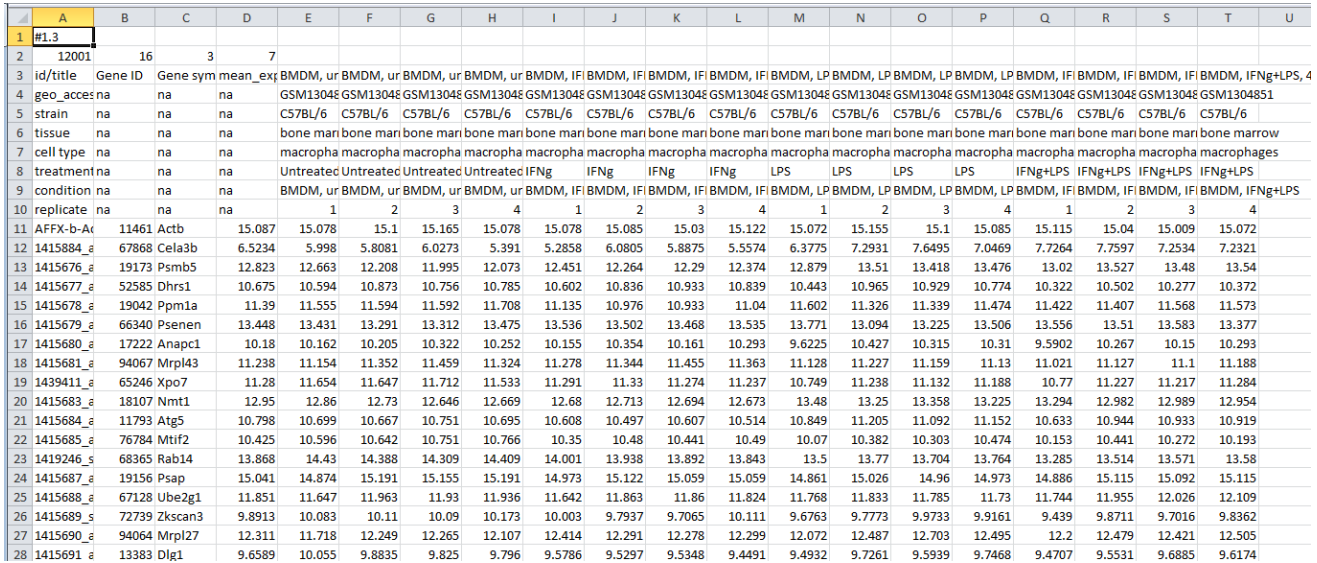
Saving dataset

- File/Save Dataset



Let’s look at what we got

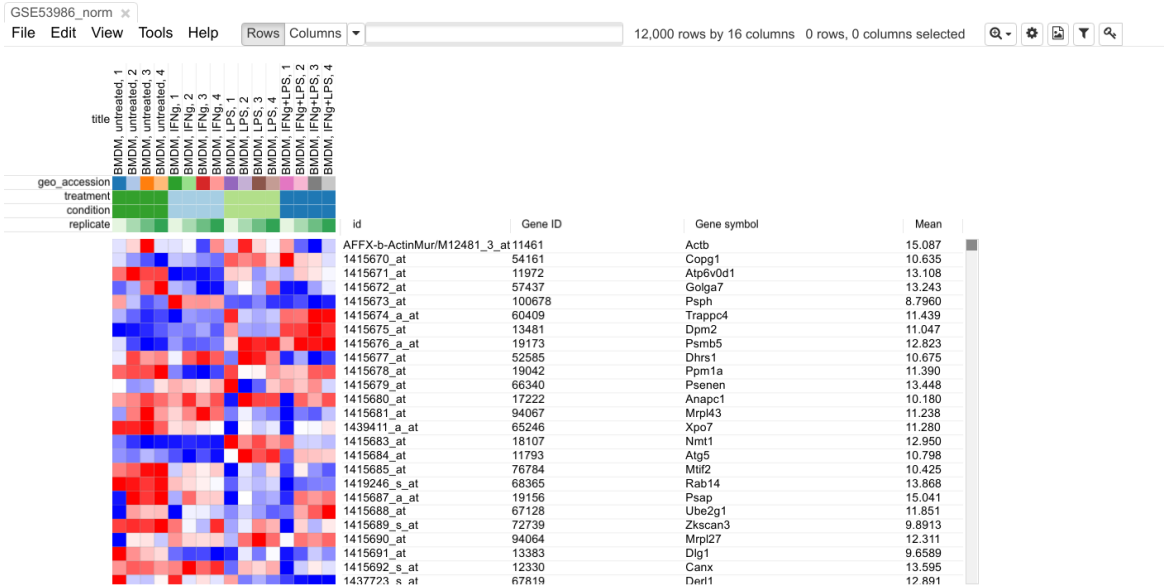
- Open gct file in Excel/Calc/Notepad



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	#1.3																				
2	12001	16	3	7																	
3	id/title	Gene ID	Gene sym	mean_exp	BMDM, ur	BMDM, ur	BMDM, ur	BMDM, ur	BMDM, IF	BMDM, IF	BMDM, IF	BMDM, IF	BMDM, LP	BMDM, LP	BMDM, LP	BMDM, LP	BMDM, IF	BMDM, IF	BMDM, IF	BMDM, IFNg+LPS, 4	
4	geo_acces	na	na	na	GSM1304	GSM1304	GSM1304	GSM1304	GSM1304	GSM1304	GSM1304	GSM1304	GSM1304	GSM1304	GSM1304	GSM1304	GSM1304	GSM1304	GSM1304	GSM1304851	
5	strain	na	na	na	C57BL/6	C57BL/6	C57BL/6	C57BL/6	C57BL/6	C57BL/6	C57BL/6	C57BL/6	C57BL/6	C57BL/6	C57BL/6	C57BL/6	C57BL/6	C57BL/6	C57BL/6	C57BL/6	
6	tissue	na	na	na	bone mari	bone mari	bone mari	bone mari	bone mari	bone mari	bone mari	bone mari	bone mari	bone mari	bone mari	bone mari	bone mari	bone mari	bone mari	bone marrow	
7	cell type	na	na	na	macropha	macropha	macropha	macropha	macropha	macropha	macropha	macropha	macropha	macropha	macropha	macropha	macropha	macropha	macropha	macrophages	
8	treatment	na	na	na	Untreated	Untreated	Untreated	Untreated	IFNg	IFNg	IFNg	IFNg	LPS	LPS	LPS	LPS	IFNg+LPS	IFNg+LPS	IFNg+LPS	IFNg+LPS	
9	condition	na	na	na	BMDM, ur	BMDM, ur	BMDM, ur	BMDM, ur	BMDM, IF	BMDM, IF	BMDM, IF	BMDM, IF	BMDM, LP	BMDM, LP	BMDM, LP	BMDM, LP	BMDM, IF	BMDM, IF	BMDM, IF	BMDM, IFNg+LPS	
10	replicate	na	na	na	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	
11	AFFX-b-Ac	11461	Actb	15.087	15.078	15.1	15.165	15.078	15.078	15.085	15.03	15.122	15.072	15.155	15.1	15.085	15.115	15.04	15.009	15.072	
12	1415884_a	67868	Cela3b	6.5234	5.998	5.8081	6.0273	5.391	5.2858	6.0805	5.8875	5.5574	6.3775	7.2931	7.6495	7.0469	7.7264	7.7597	7.2534	7.2321	
13	1415676_a	19173	Psmb5	12.823	12.663	12.208	11.995	12.073	12.451	12.264	12.29	12.374	12.879	13.51	13.418	13.476	13.02	13.527	13.48	13.54	
14	1415677_a	52585	Dhrs1	10.675	10.594	10.873	10.756	10.785	10.602	10.836	10.933	10.839	10.443	10.965	10.929	10.774	10.322	10.502	10.277	10.372	
15	1415678_a	19042	Ppm1a	11.39	11.555	11.594	11.592	11.708	11.135	10.976	10.933	11.04	11.602	11.326	11.339	11.474	11.422	11.407	11.568	11.573	
16	1415679_a	66340	Psenen	13.448	13.431	13.291	13.312	13.475	13.536	13.502	13.468	13.535	13.771	13.094	13.225	13.506	13.556	13.51	13.583	13.377	
17	1415680_a	17222	Anapc1	10.18	10.162	10.205	10.322	10.252	10.155	10.354	10.161	10.293	9.6225	10.427	10.315	10.31	9.5902	10.267	10.15	10.293	
18	1415681_a	94067	Mrpl43	11.238	11.154	11.352	11.459	11.324	11.278	11.344	11.455	11.363	11.128	11.227	11.159	11.13	11.021	11.127	11.1	11.188	
19	1439411_a	65246	Xpo7	11.28	11.654	11.647	11.712	11.533	11.291	11.33	11.274	11.237	10.749	11.238	11.132	11.188	10.77	11.227	11.217	11.284	
20	1415683_a	18107	Nmt1	12.95	12.86	12.73	12.646	12.669	12.68	12.713	12.694	12.673	13.48	13.25	13.358	13.225	13.294	12.982	12.989	12.954	
21	1415684_a	11793	Atg5	10.798	10.699	10.667	10.751	10.695	10.608	10.497	10.607	10.514	10.849	11.205	11.092	11.152	10.633	10.944	10.933	10.919	
22	1415685_a	76784	Mtif2	10.425	10.596	10.642	10.751	10.766	10.35	10.48	10.441	10.49	10.07	10.382	10.303	10.474	10.153	10.441	10.272	10.193	
23	1419246_s	68365	Rab14	13.868	14.43	14.388	14.309	14.409	14.001	13.938	13.892	13.843	13.5	13.77	13.704	13.764	13.285	13.514	13.571	13.58	
24	1415687_a	19156	Psap	15.041	14.874	15.191	15.155	15.191	14.973	15.122	15.059	15.059	14.861	15.026	14.96	14.973	14.886	15.115	15.092	15.115	
25	1415688_a	67128	Ube2g1	11.851	11.647	11.963	11.93	11.936	11.642	11.863	11.86	11.824	11.768	11.833	11.785	11.73	11.744	11.955	12.026	12.109	
26	1415689_s	72739	Zkscan3	9.8913	10.083	10.11	10.09	10.173	10.003	9.7937	9.7065	10.111	9.6763	9.7773	9.9733	9.9161	9.439	9.8711	9.7016	9.8362	
27	1415690_a	94064	Mrpl27	12.311	11.718	12.249	12.265	12.107	12.414	12.291	12.278	12.299	12.072	12.487	12.703	12.495	12.2	12.479	12.421	12.505	
28	1415691_a	13383	Dlg1	9.6589	10.055	9.8835	9.825	9.796	9.5786	9.5297	9.5348	9.4491	9.4932	9.7261	9.5939	9.7468	9.4707	9.5531	9.6885	9.6174	

Loading a gct file in Phantasus

- File/Open, choose GSE53986_norm.gct



https://artyomovlab.wustl.edu/publications/supp_materials/SBW2018-StL/GSE53986_norm.gct

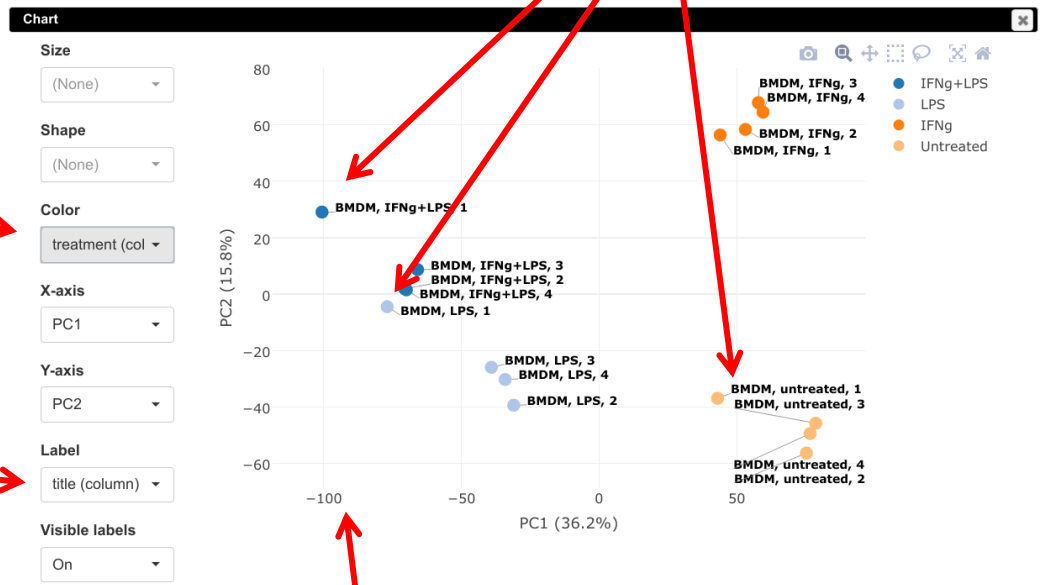
Exploring dataset: principal component analysis (PCA) plot

- Tools/Plots/PCA plot

color <- treatment

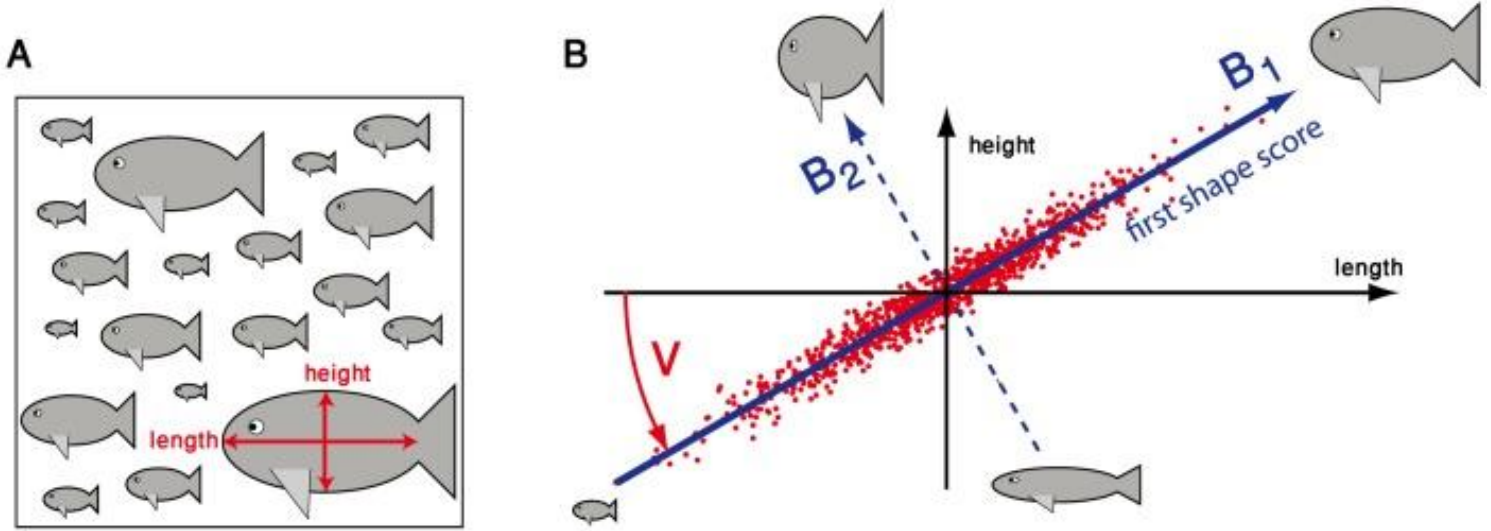
label <- title

Outliers!



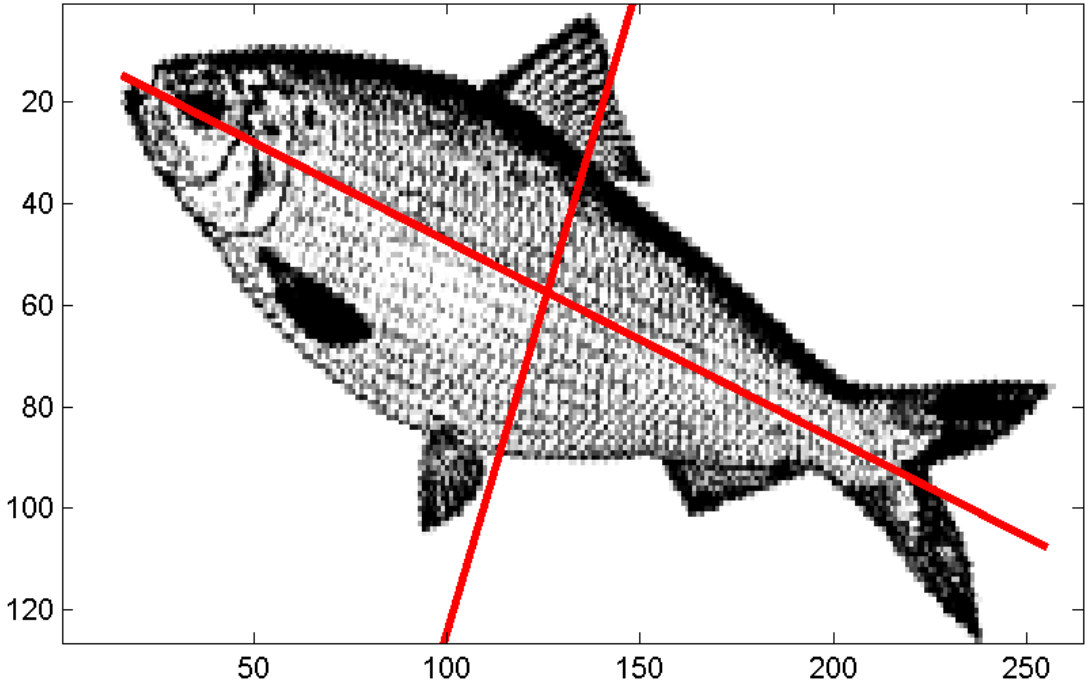
Scale should be ~10-100, not 1000000

What is PCA?



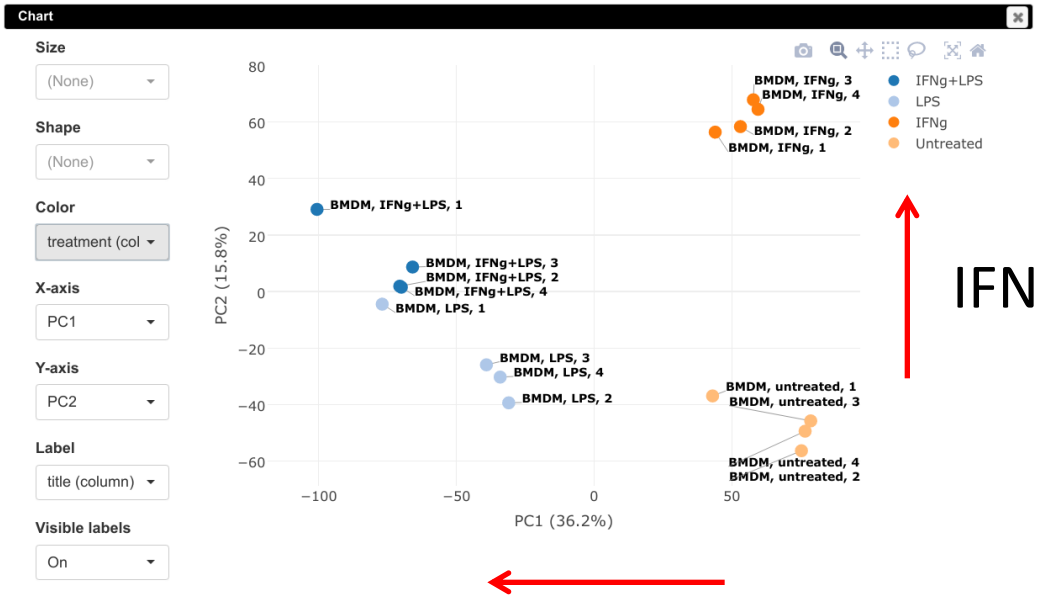
https://www.ncbi.nlm.nih.gov/pubmed/25426857

PCA of the fish



https://commons.wikimedia.org/wiki/File:PCA_fish.png

Exploring dataset: principal components can be meaningful

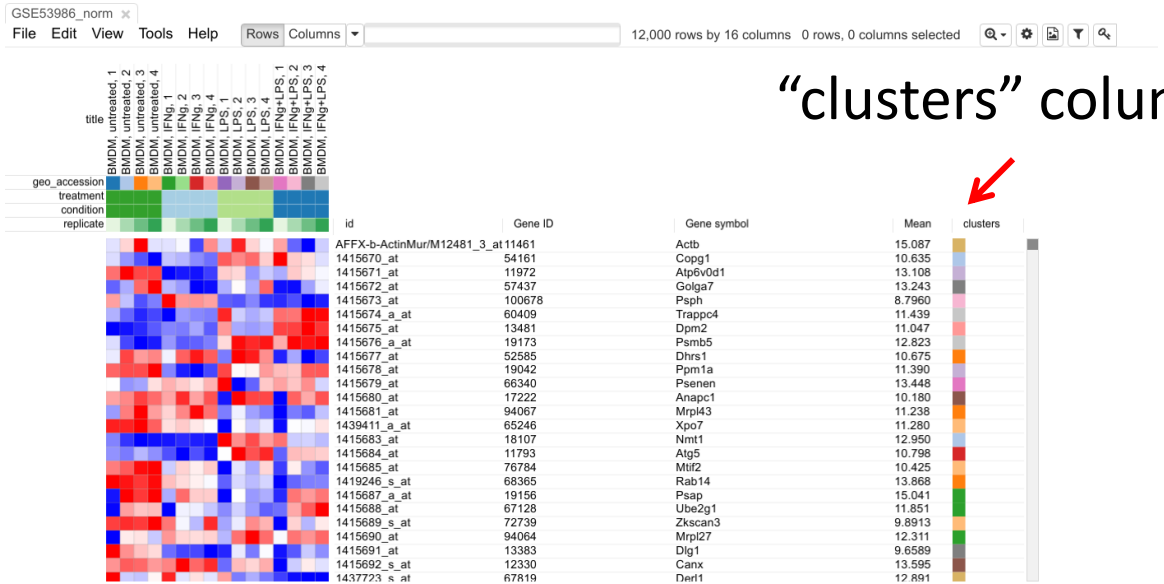


IFNg response

LPS response

Exploring dataset: k-means

- Tools/Clustering/k-means
- Number of cluster = 16



“clusters” column appeared

Exploring dataset: k-means, bird's eye view

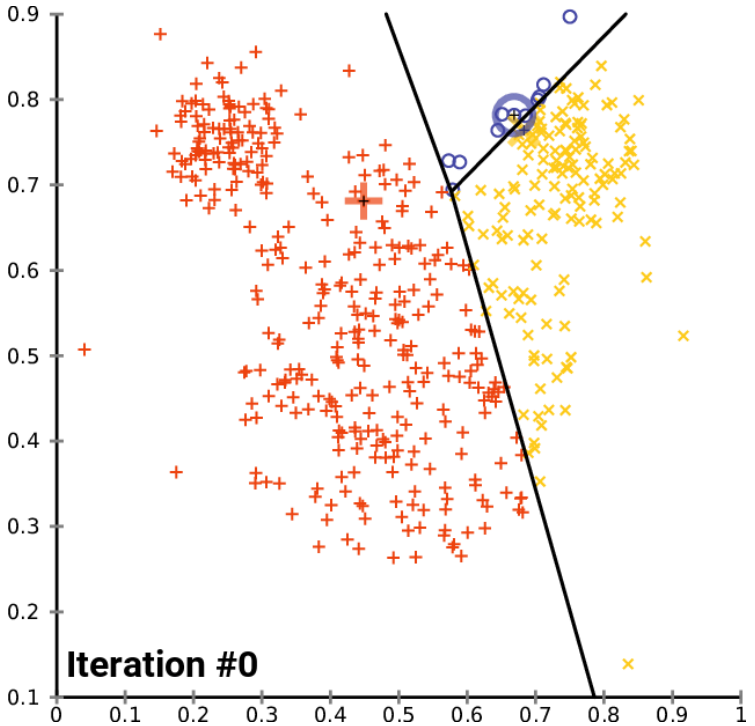
Our outliers

Select "Fit To Window"

Sort by "clusters"

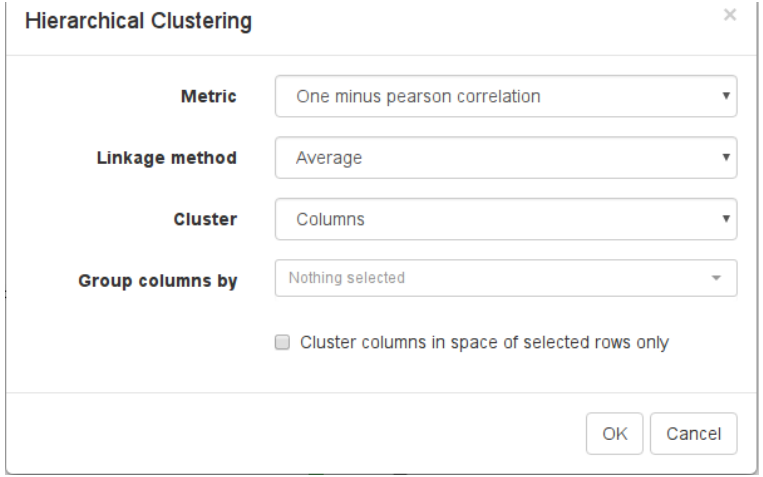
How k-means clustering works

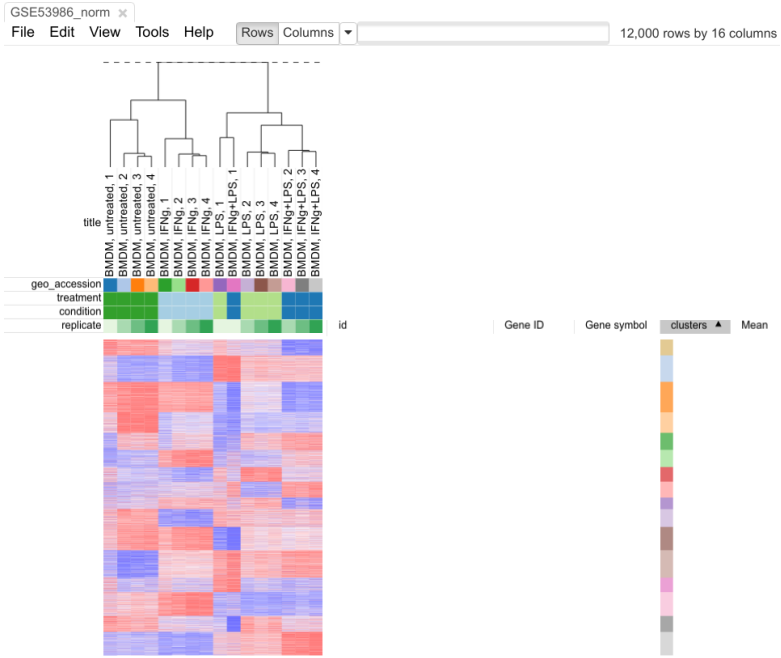
- Select k *random* centers
- Assign each gene to the closes cluster center
- Refine center
- Repeat until convergence



Exploring dataset: hierarchical clustering

- Tools/Hierarchical clustering
- Metric <- 1 - pearson correlation

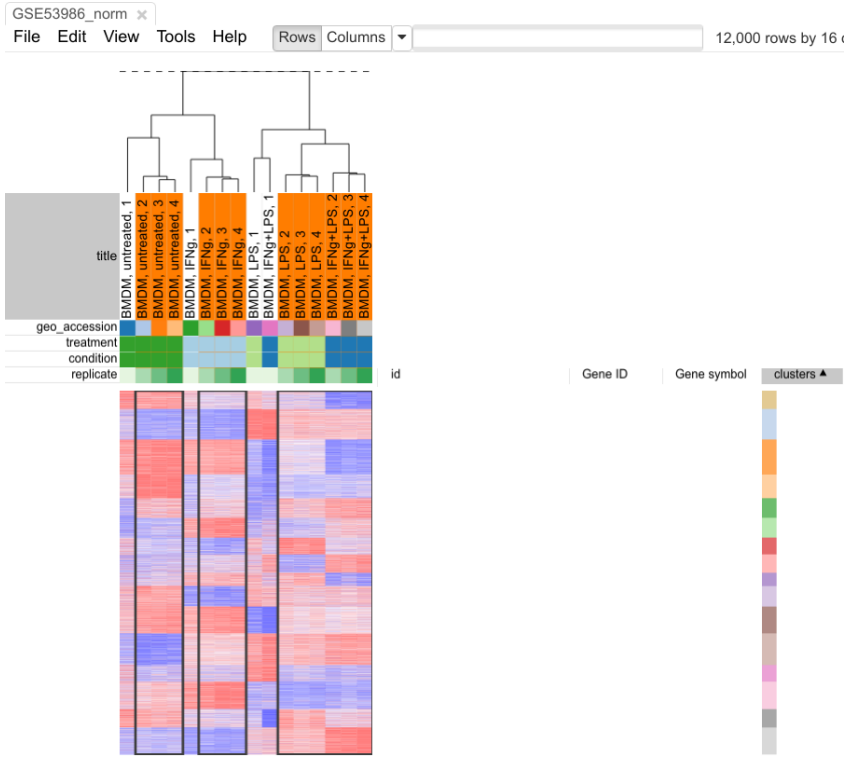




Filtering outliers

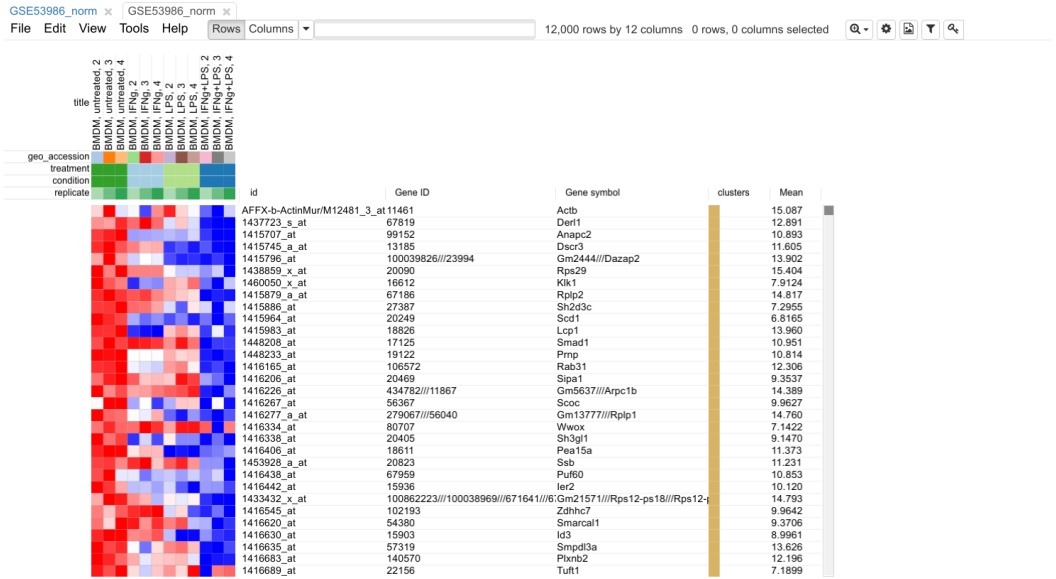
- Select good samples
- Tools/New heatmap (Ctrl-X)

- Very bad outlier should be removed at the start of the analysis, before normalization



Saving filtered dataset

- File/Save dataset
- Name like GSE53986_filtered.gct



Summary

- Check expression scale (should be log2)
- Data should be normalized
- Do a quality check by looking at markers, PCA, k-means and hierarchical clustering
- Save processed datasets

- Next: doing a differential expression analysis