General Introduction to RNA-seq analysis:

The power of an unbiased approach to RNA expression analysis

Florencia Schlamp, PhD

February 24th, 2020

Bioinformatics Lecture Series

- Today!
- March 23rd - Inside the Black Box: The steps of RNA-seq data processing, data exploration, and data analysis
- April 27th - Intro to Machine Learning (guest lecturer from Cornell)
- TBD (May/June) - Analysis consideration when designing an RNA-seq experiment
- August 24th - Biodatabases (guest lecturer from Cornell)

Aims for today's lecture

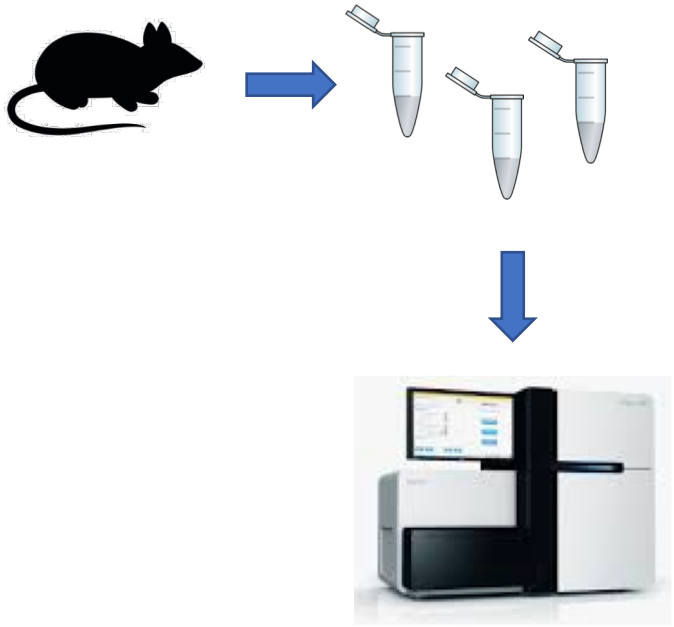
- General overview of what can you get out of an RNA-seq experiment
- Focus on downstream analyses
  - Starting up with table of gene counts and table with differential expression results (log fold change and p-values)
- Emphasis on data visualization

RNA sequencing

- Goal: profile transcription in different samples
- How: measuring mRNA levels of genes at point of sample collection
- Measuring technique. True power lays in how it's used (treatments, knockouts, time course, different organs/tissues, etc.)

More on experimental design TBD

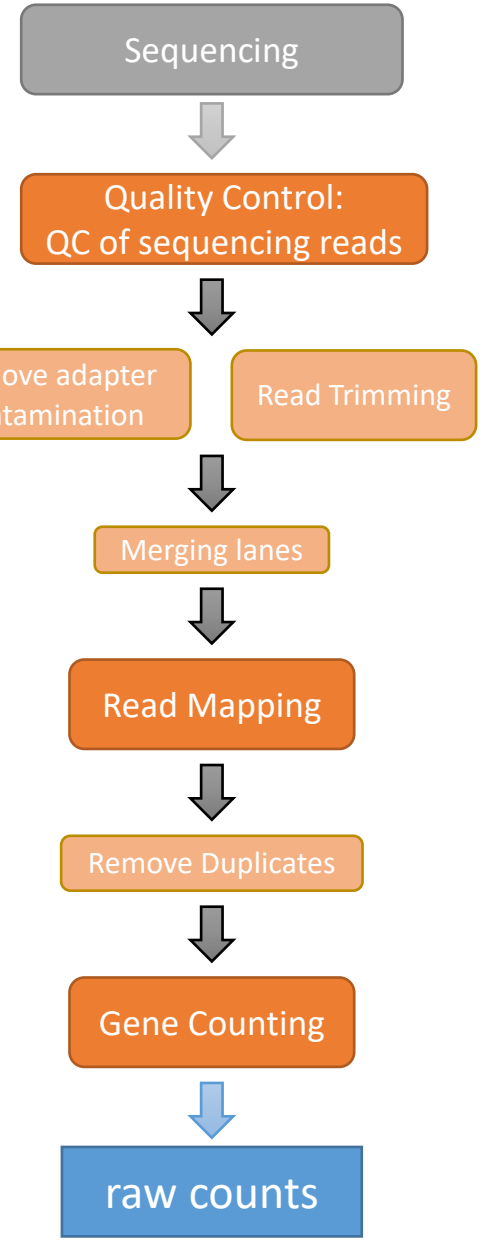
RNA sequencing



Data generation

- Sequencing methods vary (read length, coverage, price, speed, accuracy, etc.)

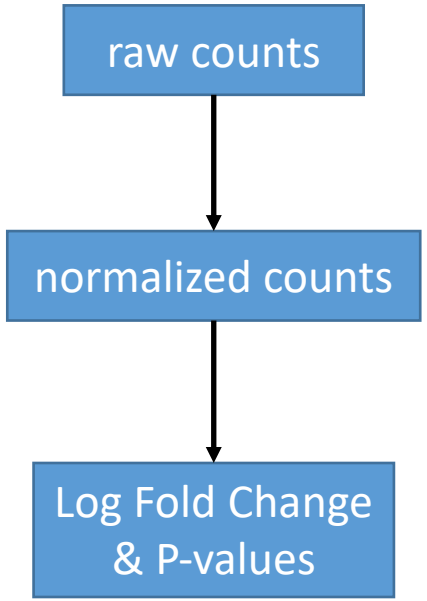
More on design considerations TBD



Data processing

More on data processing on March 23rd

Data Types



More on analysis pipeline on March 24th

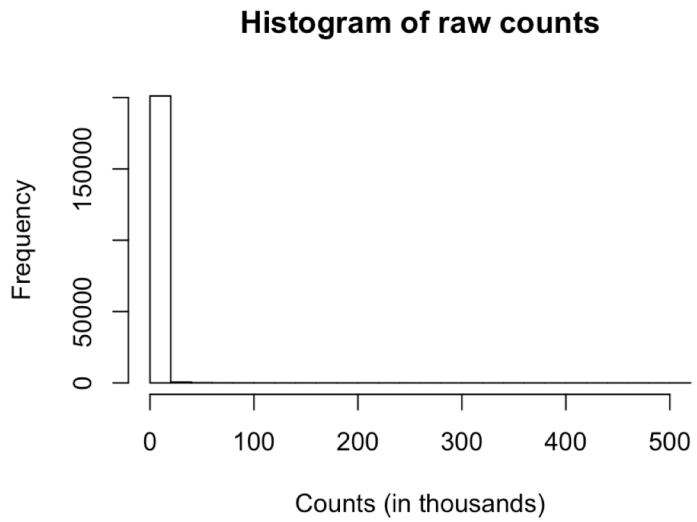
Data Types

raw counts

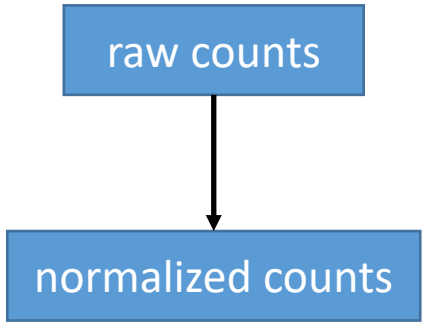
one value per gene per sample

	C1	T1	C2	T2
ENSG00000000003	723	486	904	445
ENSG00000000005	0	0	0	0
ENSG00000000419	467	523	616	371
ENSG00000000457	347	258	364	237
ENSG00000000460	96	81	73	66
ENSG00000000938	0	0	1	0
ENSG00000000971	3413	3916	6000	4308
ENSG00000001036	2328	1714	2640	1381
ENSG00000001084	670	372	692	448

```
> range(raw_counts)
[1]      0 510107
```



Data Types



raw data needs to be adjusted to account for factors that prevent direct comparison of expression measures

More on normalization on March 24th

one value per gene per sample

Data Types

raw counts

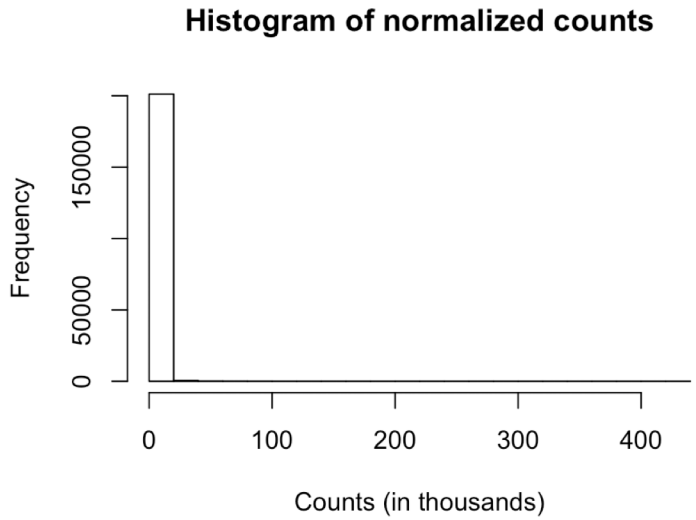
↓

normalized counts

1 normalized

	C1	T1	C2	T2
ENSG00000000003	757.025831	554.971221	7.679000e+02	628.241871
ENSG00000000419	488.977957	597.222116	5.232593e+02	523.770189
ENSG00000000457	363.330516	294.614352	3.091987e+02	334.591738
ENSG00000000460	100.517953	92.495203	6.200963e+01	93.177446
ENSG00000000938	0.000000	0.000000	8.494469e-01	0.000000
ENSG00000000971	3573.622628	4471.743418	5.096682e+03	6081.946027
ENSG00000001036	2437.560351	1957.244182	2.242540e+03	1949.667471
ENSG00000001084	701.531544	424.792786	5.878173e+02	632.477210

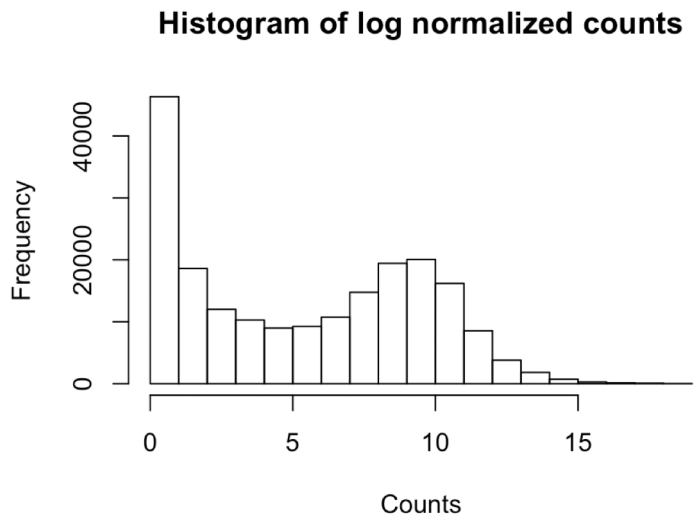
```
> range(norm_counts)
[1]      0.0 433308.8
```



2 normalized + log transformed

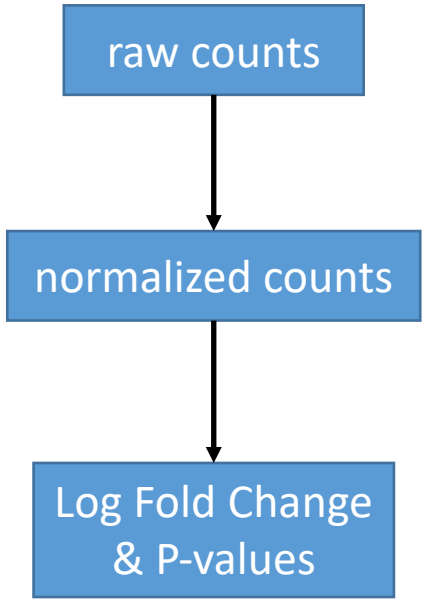
	C1	T1	C2	T2
ENSG00000000003	9.566103	9.118866	9.5866522	9.297471
ENSG00000000419	8.936573	9.224537	9.0341368	9.035542
ENSG00000000457	8.509104	8.207573	8.2770488	8.390563
ENSG00000000460	6.665591	6.546820	5.9775003	6.557310
ENSG00000000938	0.000000	0.000000	0.8870939	0.000000
ENSG00000000971	11.803575	12.126944	12.3156255	12.570554
ENSG00000001036	11.251814	10.935345	11.1315611	10.929752
ENSG00000001084	9.456419	8.734008	9.2016762	9.307149

```
> range(norm_log_counts)
[1]  0.00000 18.72504
```



More on normalization on March 24th

Data Types



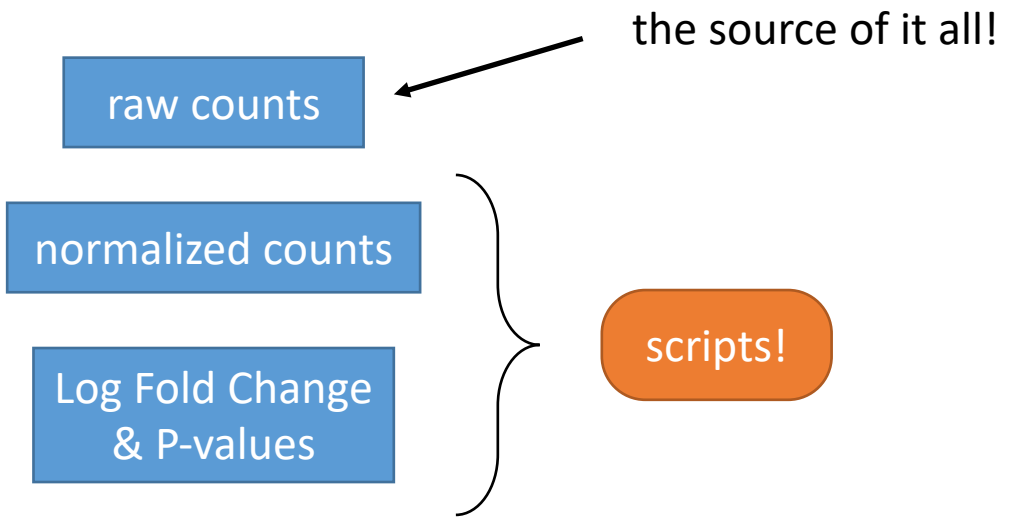
one value per gene per **comparison**

```
> DESeq2_results
log2 fold change (MLE): dex treated vs control
Wald test p-value: dex treated vs control
DataFrame with 25258 rows and 6 columns
                     baseMean     log2FoldChange             pvalue              padj
                    <numeric>          <numeric>          <numeric>         <numeric>
ENSG00000000003  747.194195359907  -0.35070302068658 0.0371174658432818 0.164762454716952
ENSG00000000419  520.134160051965  0.206107766417862 0.0414026263001157 0.177910328224659
ENSG00000000457  322.664843927049 0.0245269479387466  0.865810560623564  0.96221383733165
ENSG00000000460   87.682625164828  -0.14714204922212  0.566969065257939 0.818631924409481
ENSG00000000938 0.319166568913118  -1.73228897394308  0.620002884826012                NA
```

More on analysis pipeline on March 24th

Rigor and reproducibility note

- Files you should have / save:



Normalized counts (after data transformation)

- Normalized counts per gene per sample (information on each replicate!)

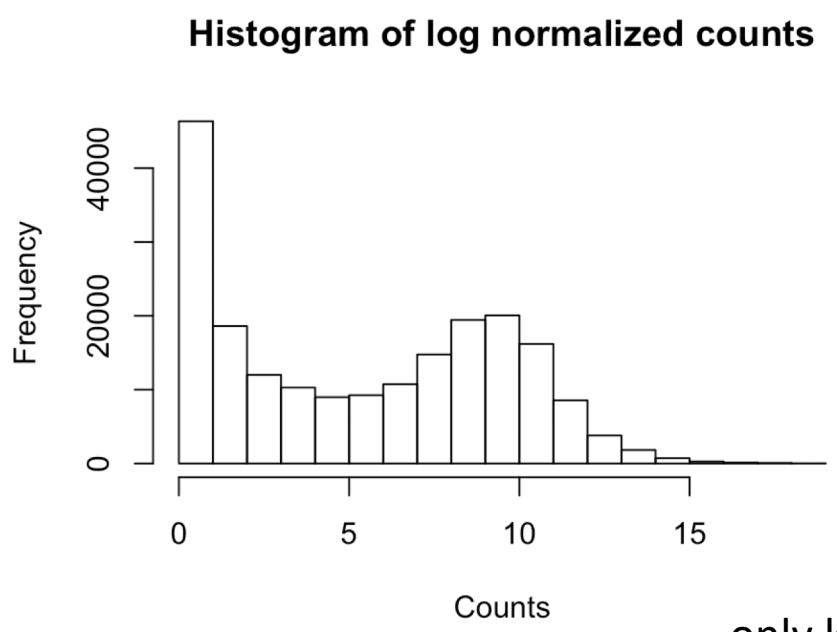
	C1	T1	C2	T2	C3	T3	C4	T4
ENSG00000000003	9.566103	9.118866	9.5866522	9.297471	10.0119186	9.5778244	9.825015	9.257247
ENSG00000000419	8.936573	9.224537	9.0341368	9.035542	9.0059121	9.0884190	8.875766	9.010804
ENSG00000000457	8.509104	8.207573	8.2770488	8.390563	8.1362475	8.2853502	8.538994	8.360729
ENSG00000000460	6.665591	6.546820	5.9775003	6.557310	6.7146725	6.0530697	6.853741	6.245083
ENSG00000000938	0.000000	0.000000	0.8870939	0.000000	1.4663602	0.0000000	0.000000	0.000000
ENSG00000000971	11.803575	12.126944	12.3156255	12.570554	12.4677387	12.8652019	12.467964	12.946479
ENSG00000001036	11.251814	10.935345	11.1315611	10.929752	10.8991347	10.6208872	11.256177	10.819798
ENSG00000001084	9.456419	8.734008	9.2016762	9.307149	9.6607891	9.1355797	9.709670	9.438524

Normalized counts (after data transformation)

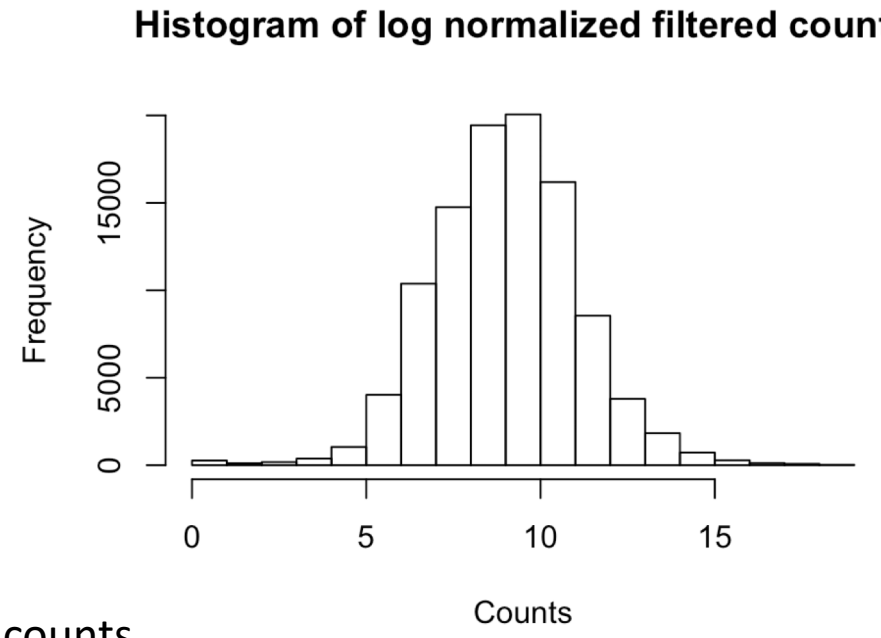
- Quality Control
  - filter out genes with low counts

HOW?

- Histograms (distribution of counts)



only keep genes with 6 counts
in at least 2 samples



March 24th

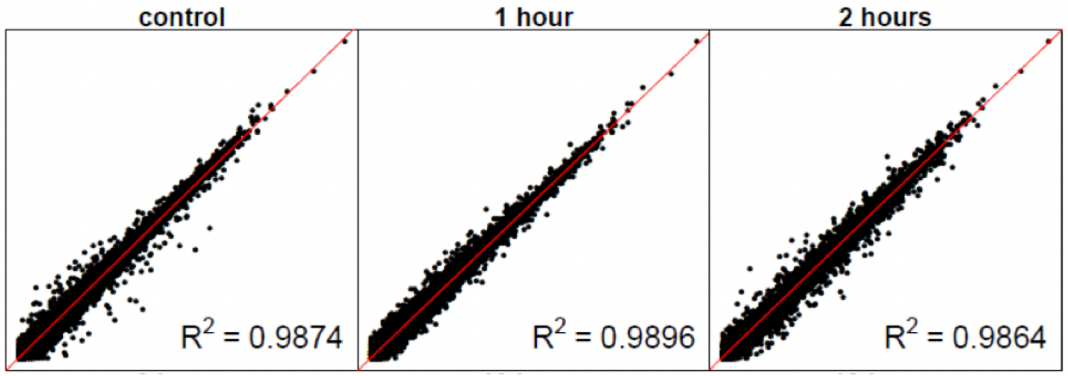
Normalized counts (after data transformation)

- Quality Control
  - Variation within and between groups
    - good replicates
    - differentiation between treatments

HOW?

- Pairwise scatterplots between replicates

Replicate A vs B



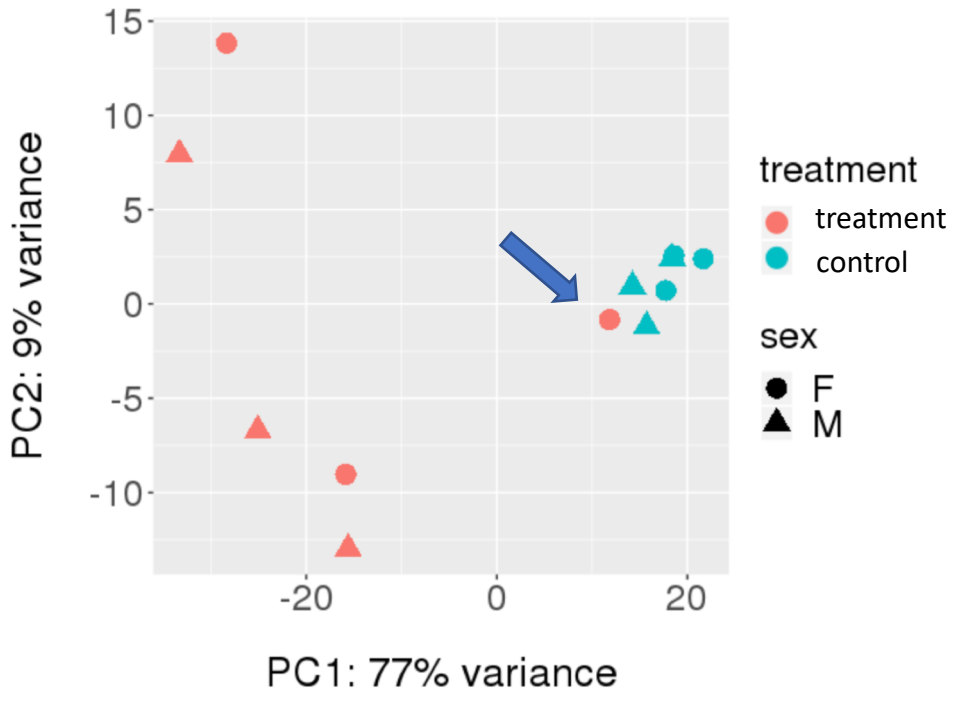
[Schlamp *et al.* 2019]

Normalized counts (after data transformation)

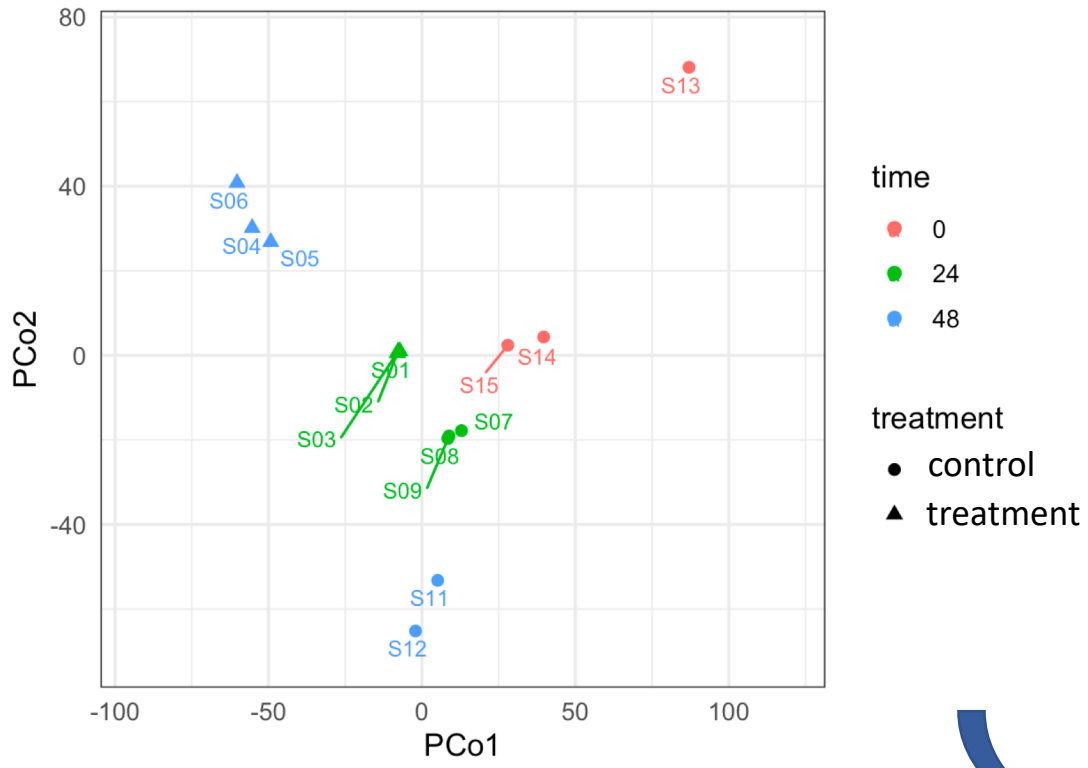
- Quality Control
  - Variation within and between groups
    - good replicates
    - differentiation between treatments

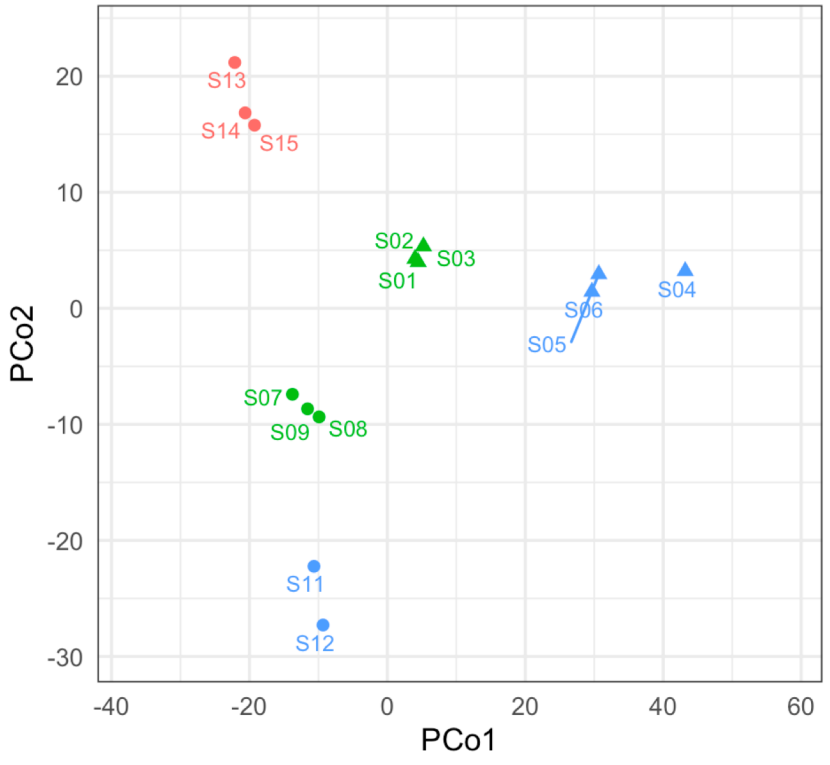
HOW?

- Pairwise scatterplots between replicates
- Dimension reduction plots (PCA, MDS)



Example: improving clustering after removing genes with low counts





Only keep genes with 70 raw counts
in at least one sample

(29,656 to 11,343 genes)

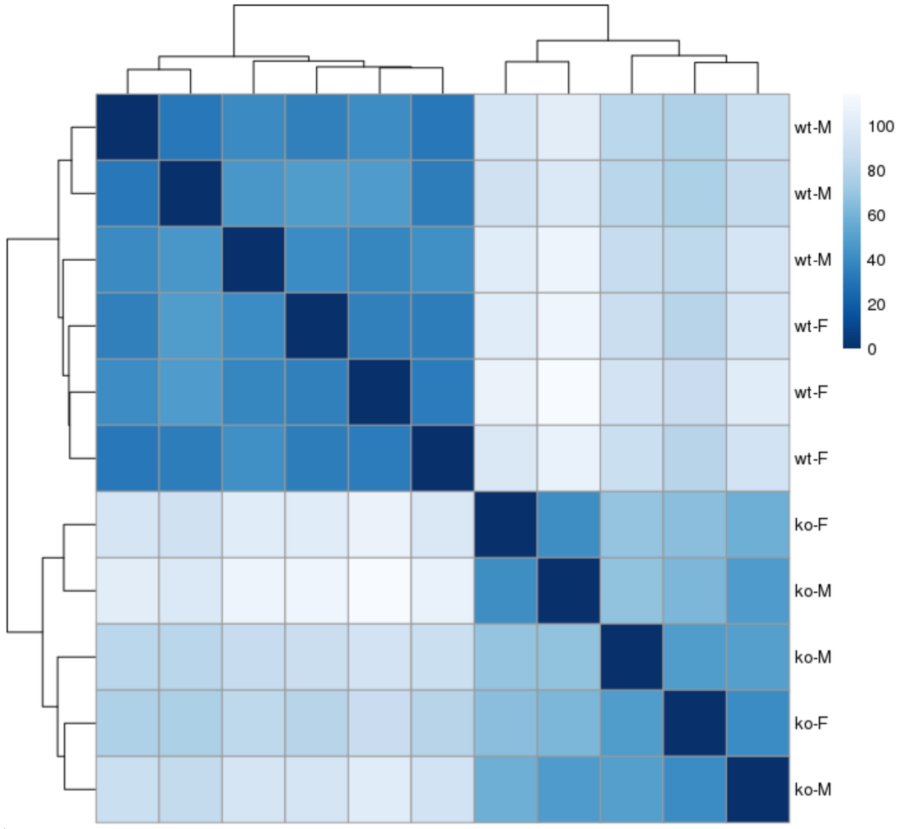
March 24th

Normalized counts (after data transformation)

- Quality Control
  - Variation within and between groups
    - good replicates
    - differentiation between treatments

HOW?

- Pairwise scatterplots between replicates
- Dimension reduction plots (PCA, MDS)
- Distance heatmaps, dendrograms

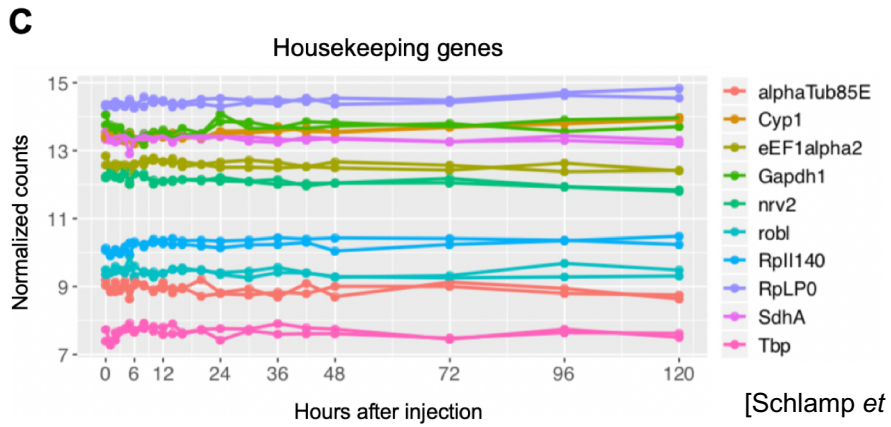


Normalized counts (after data transformation)

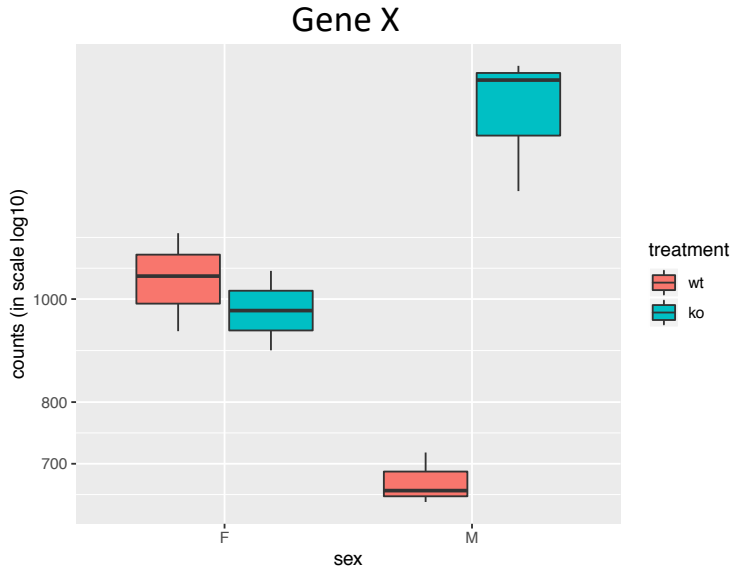
- Quality Control
- Confirm expectations
  - Genes with expectations
    - Up (induced genes)
    - Down (knockouts / knockdowns)
    - No change (housekeeping genes)

HOW?

- Line plots
- Box plots



[Schlamp *et al.* 2019]

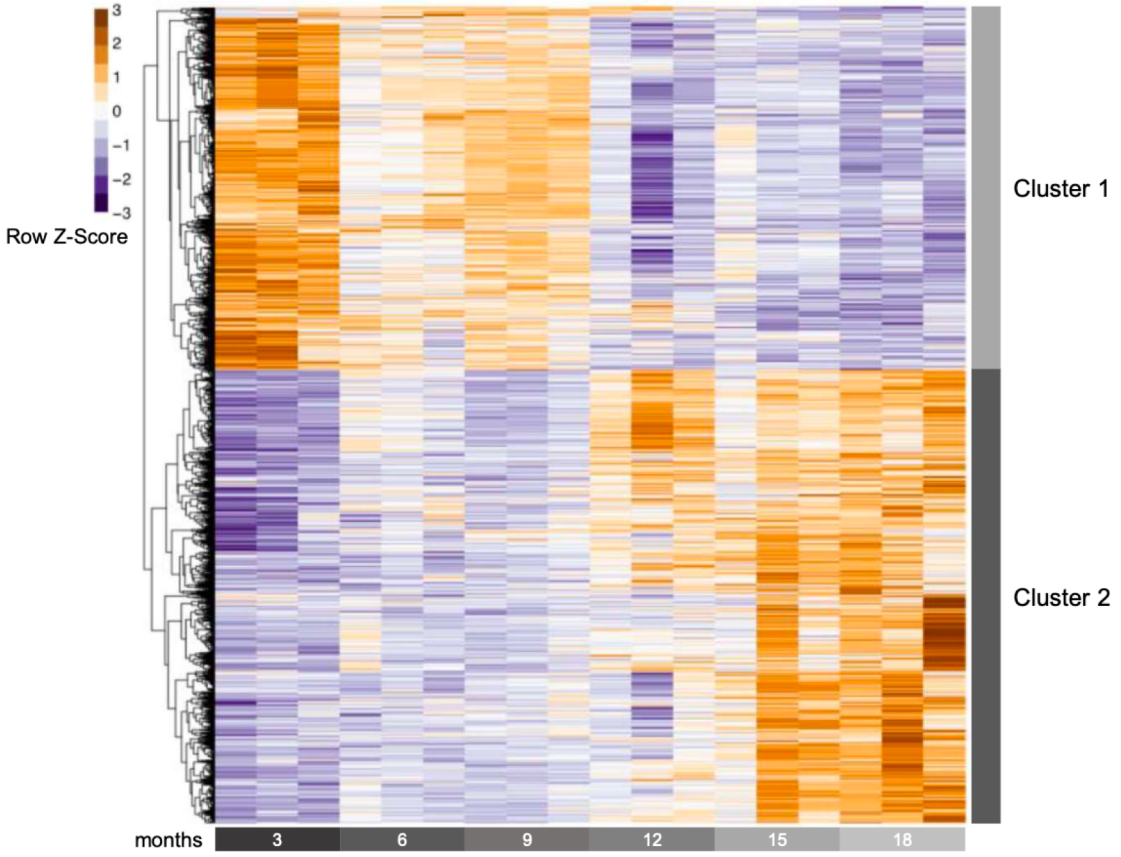


Normalized counts (after data transformation)

- Quality Control
- Confirm expectations
- Explore general dynamics

HOW?

- Row Z-score scaled heatmaps + Clustering



[Zhang *et al.* 2020]

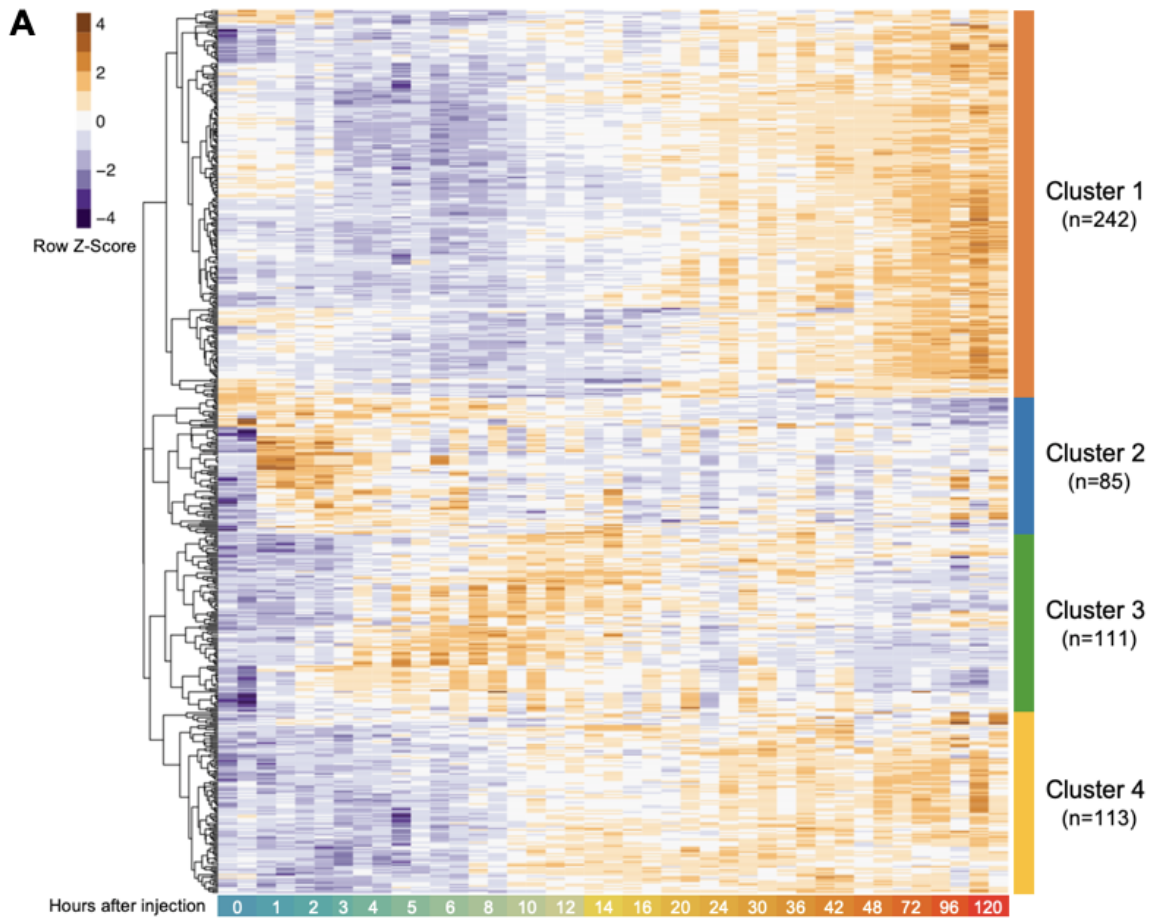
Row Z-score scaling will always show min to max PER GENE

Normalized counts (after data transformation)

- Quality Control
- Confirm expectations
- Explore general dynamics

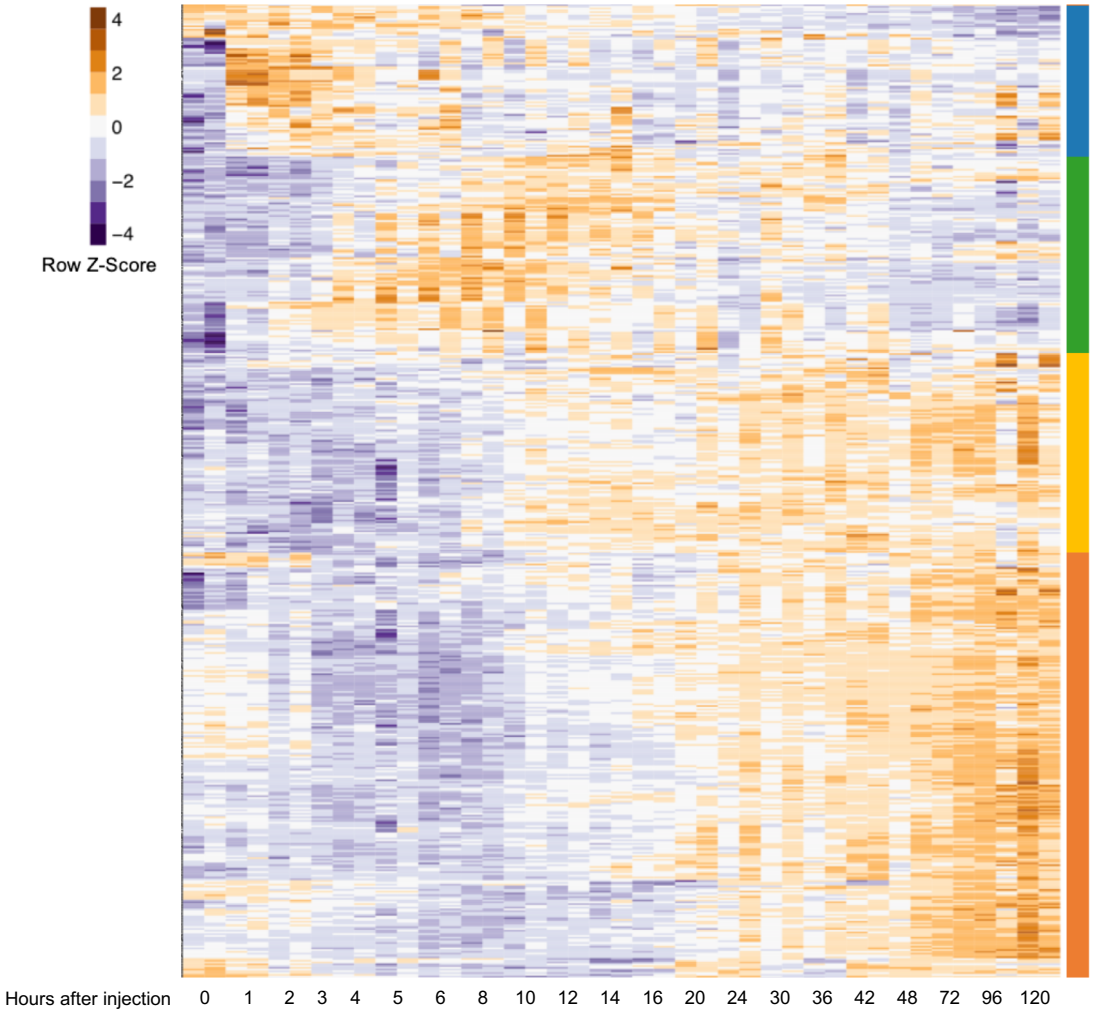
HOW?

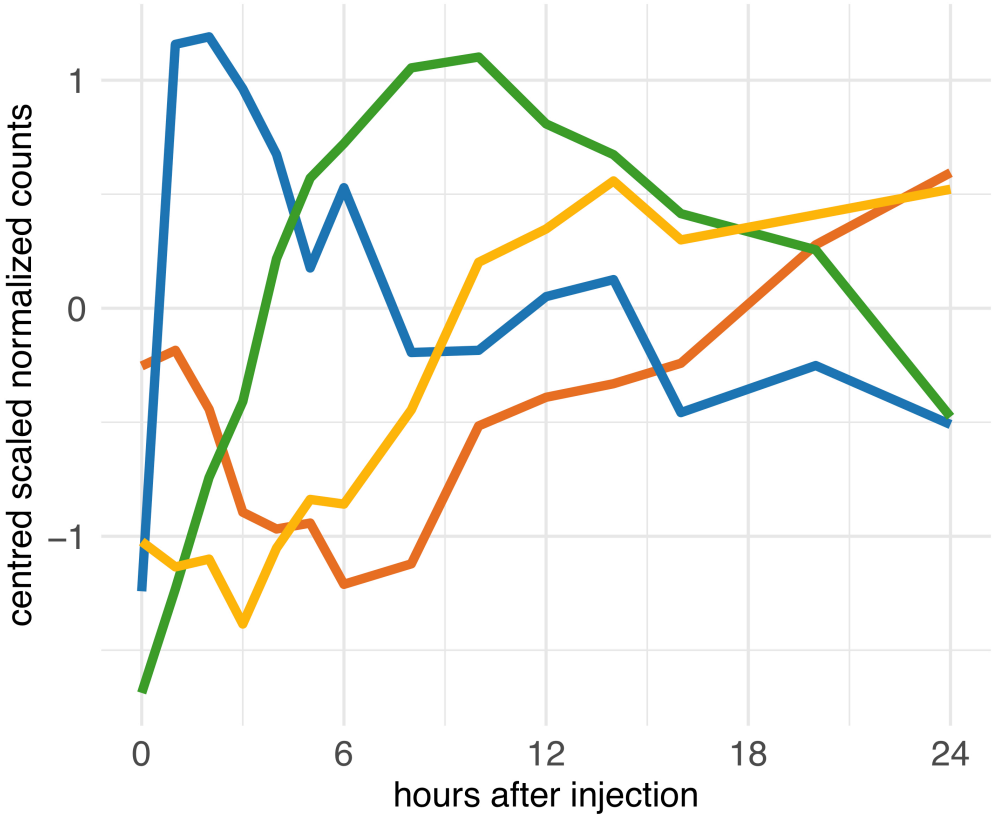
- Row Z-score scaled heatmaps + Clustering



[Schlamp *et al.* 2019]

Example: Summary of global dynamics using clusters





[Schlamp *et al*. 2019]

Differential Expression

- **Differential expression analysis** means taking the normalized read count data and performing statistical **analysis** to discover quantitative changes in **expression** levels between experimental groups.

More on analysis pipeline on March 24th

Log Fold Change and P-values (after statistical testing / differential expression)

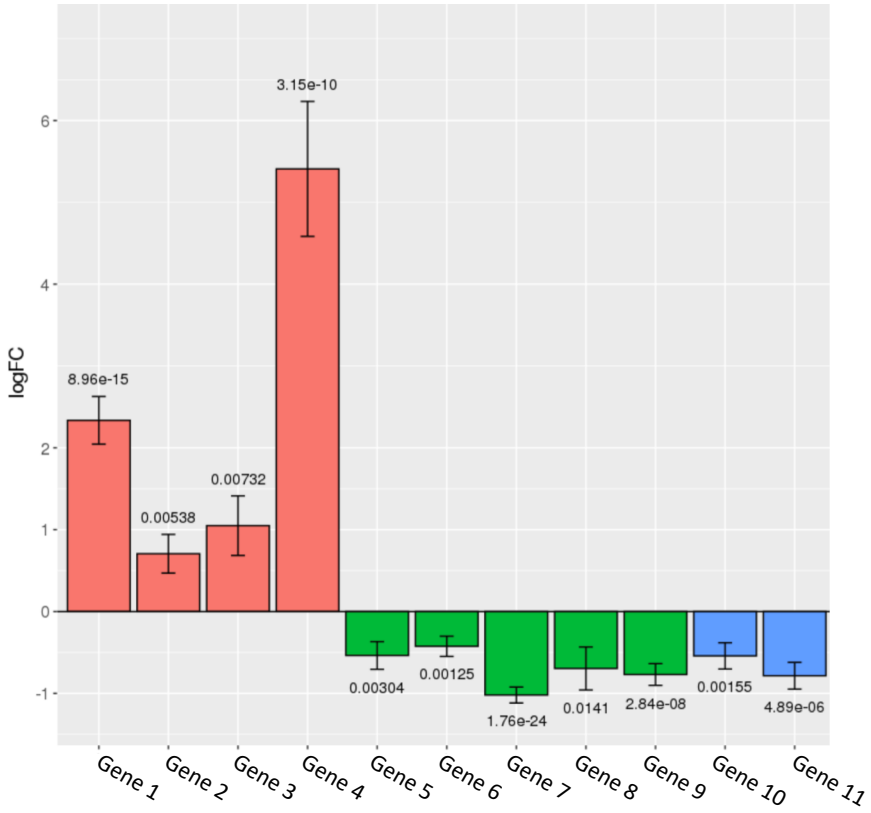
- LogFC and p-values per gene per comparison

```
> DESeq2_results
log2 fold change (MLE): dex treated vs control
Wald test p-value: dex treated vs control
DataFrame with 25258 rows and 6 columns
                     baseMean     log2FoldChange             lfcSE                stat              pvalue               padj
                    <numeric>          <numeric>         <numeric>           <numeric>           <numeric>          <numeric>
ENSG00000000003  747.194195359907  -0.35070302068658 0.168245681332529  -2.08446967499531 0.0371174658432818 0.164762454716952
ENSG00000000419  520.134160051965  0.206107766417862 0.101059218008052   2.03947517584631 0.0414026263001157 0.177910328224659
ENSG00000000457  322.664843927049 0.0245269479387466 0.145145067649248  0.168982303952742  0.865810560623564  0.96221383733165
ENSG00000000460   87.682625164828  -0.14714204922212 0.257007253994673  -0.57252099672319  0.566969065257939 0.818631924409481
ENSG00000000938 0.319166568913118  -1.73228897394308  3.49360097648095 -0.495846258804286  0.620002884826012                NA
```

Log Fold Change and P-values (after statistical testing / differential expression)

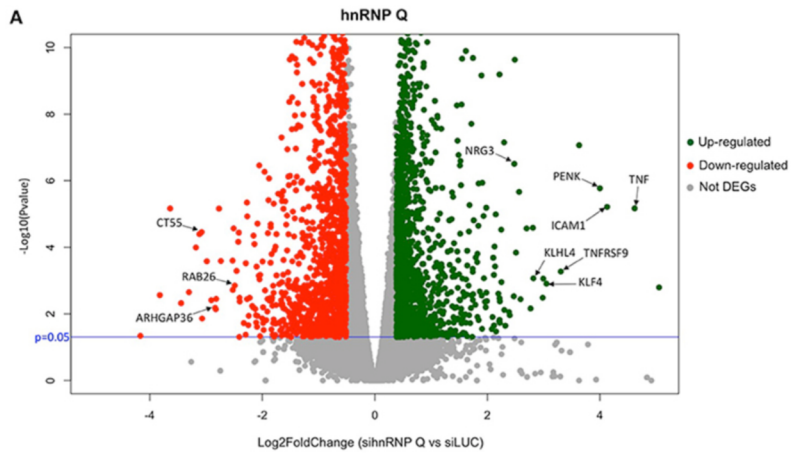
- Bar plots of specific genes ← Most common plots
  - Height of bar represents logFC
  - Error bars (hopefully)
  - Significance labeled
    - Discrete: stars to represent cutoff

      (* = 0.1, ** = 0.05, *** = 0.01)
    - Continuous: number

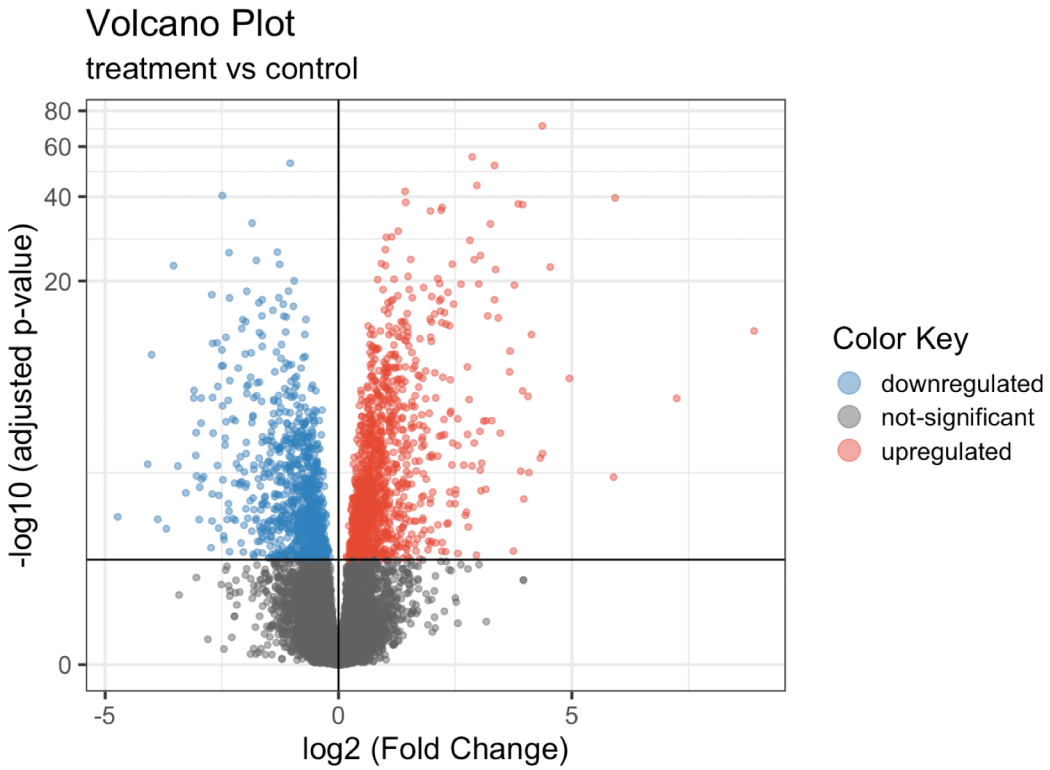


Log Fold Change and P-values

- Volcano plots
  - Significance as a continuous variable in Y axis
  - Discrete cutoffs can be labeled for both p-value and logFC
  - Genes of interest can be labeled
  - Can be interactive!



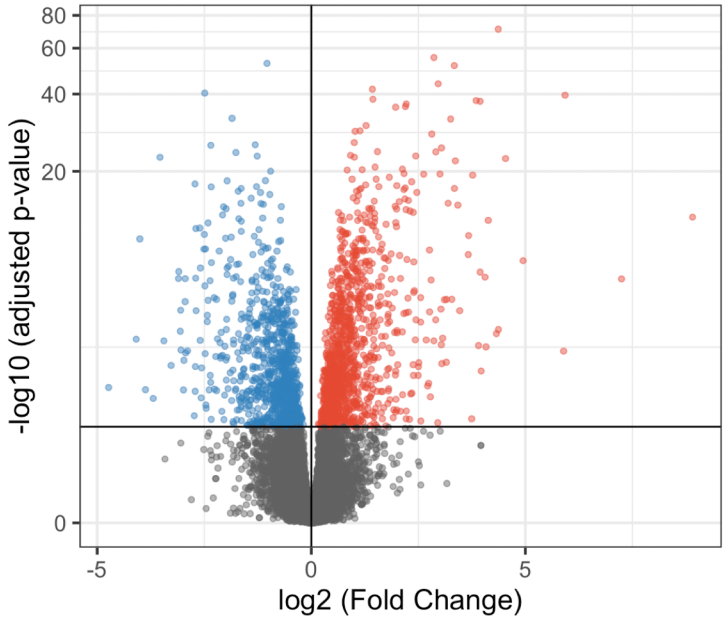
[Cappelli *et al.* 2018]



Note: be aware of distributions when choosing cutoffs

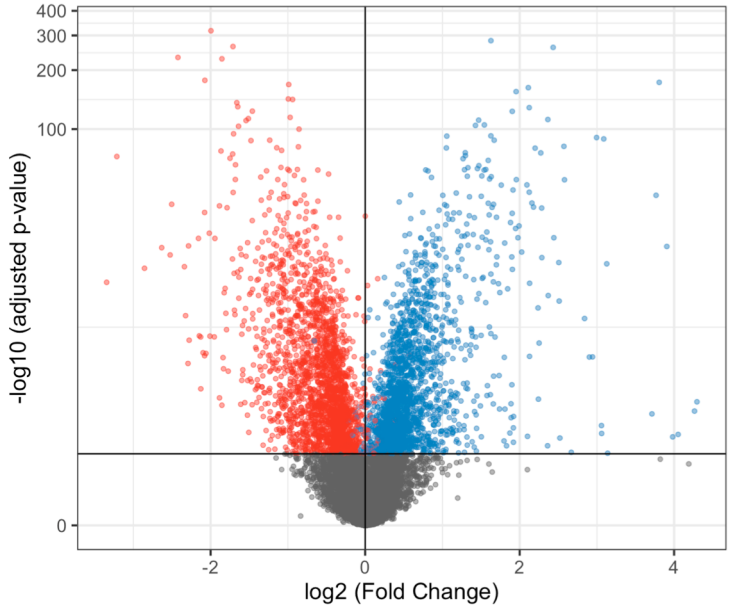
- range of p-values and logFC can be different across studies / comparisons

Study 1: larger logFCs

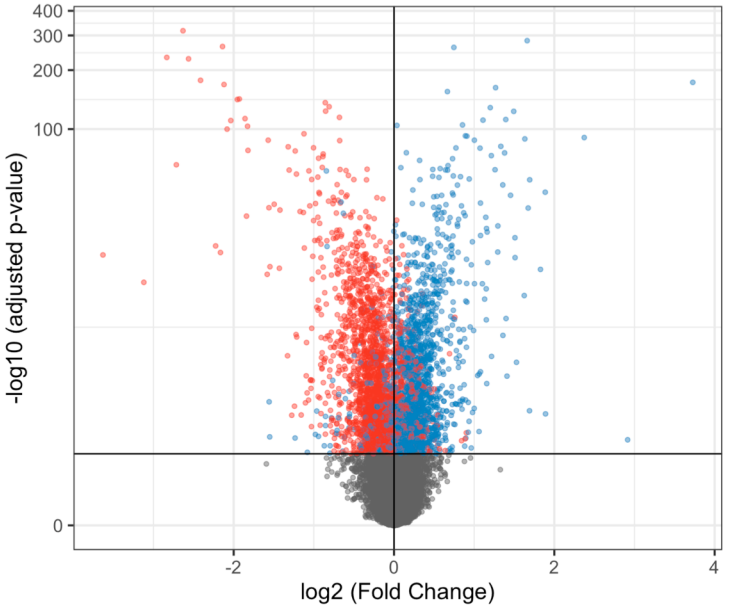


Study 2: more significant adjusted p-values

Comparison 1

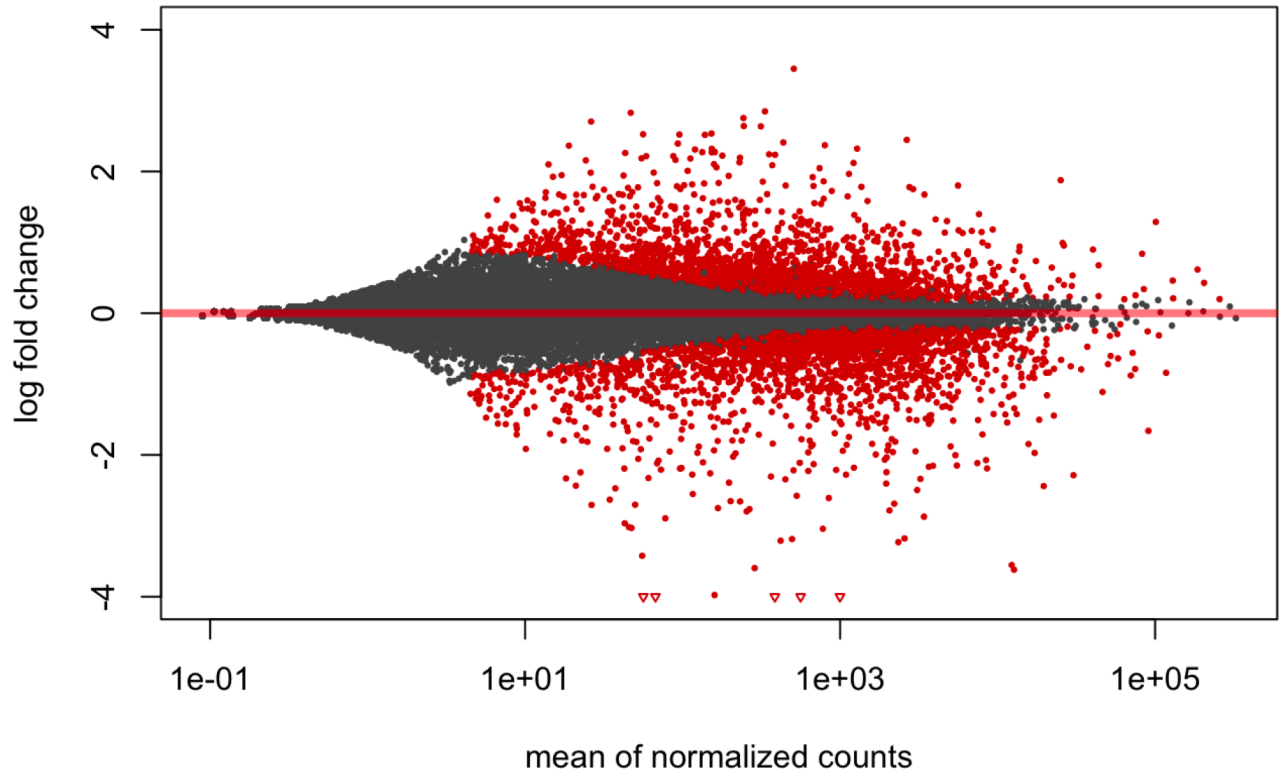


Comparison 2



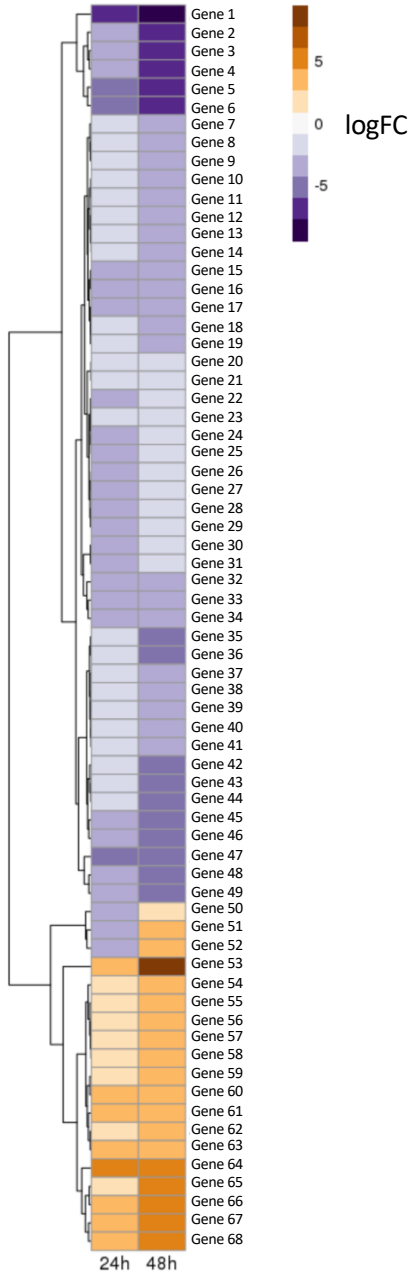
Using **both** data types to subset genes

- MA-plots
  - Significance as discrete variable (genes above cutoff in red)
  - LogFC vs mean of normalized counts



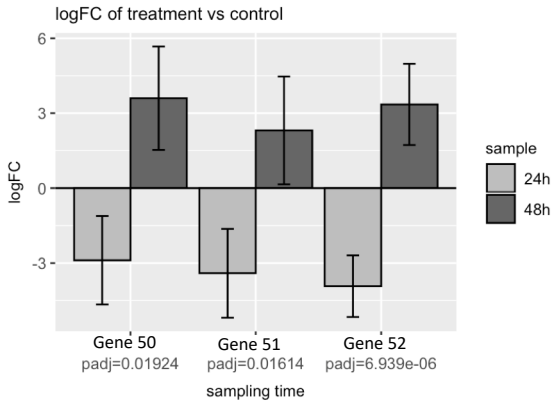
Log Fold Change and P-values

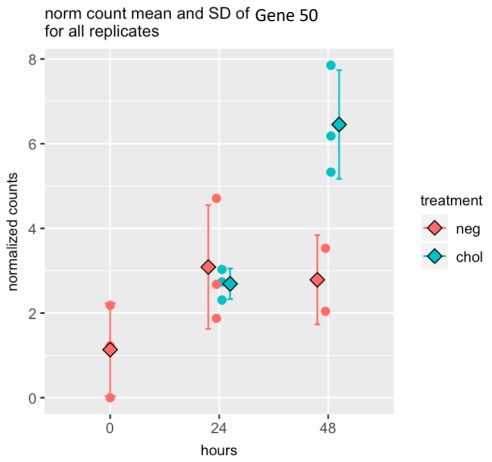
- Heatmap of log fold changes
  - Filter by p-value first
  - No information on replicates

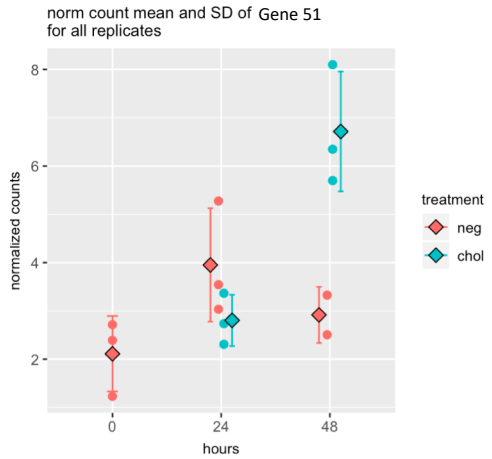


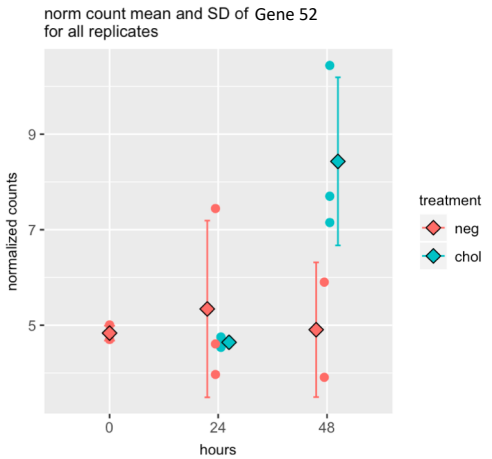
Using **both** data types to be certain of results

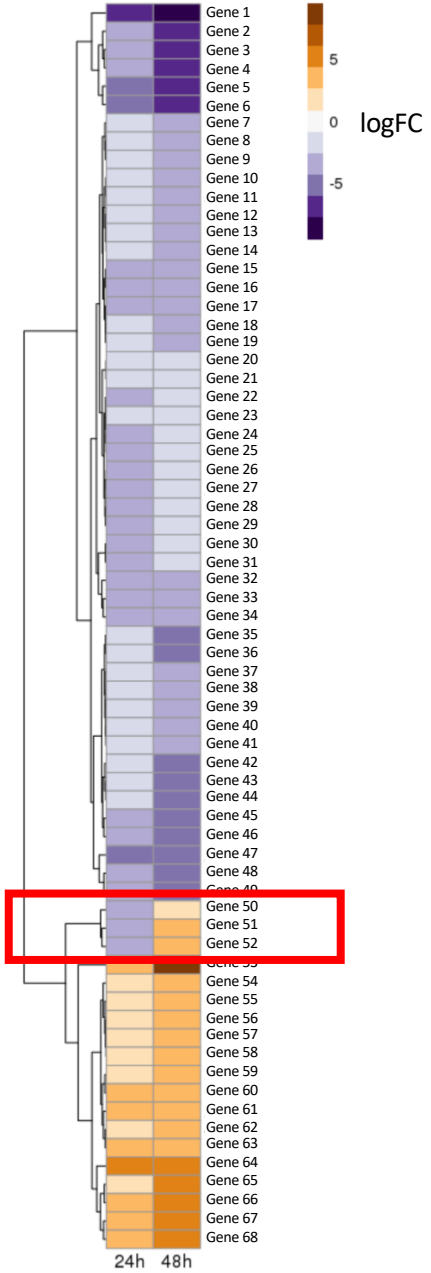
- Plots of normalized counts and logFC









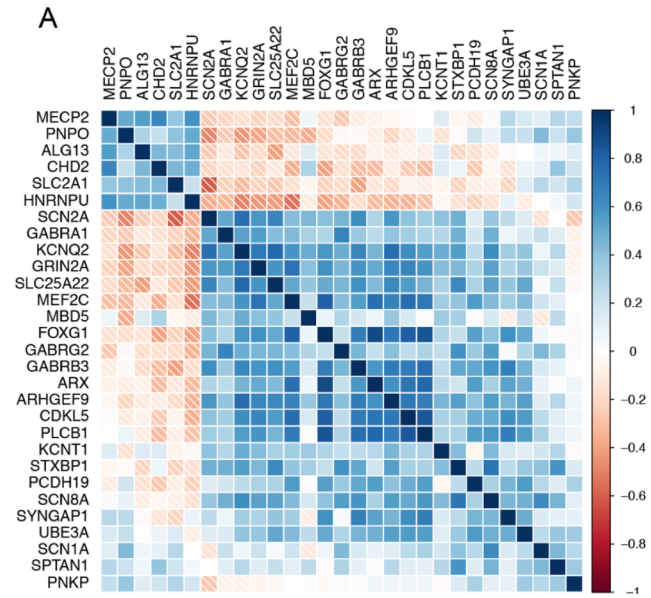


Summary so far

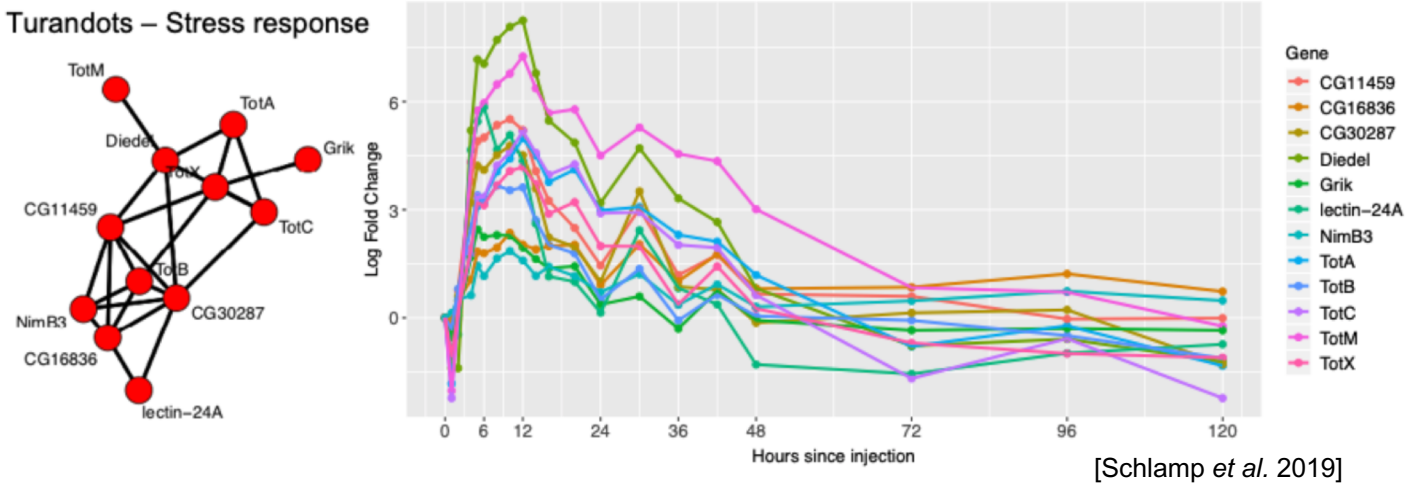
- Normalized counts
  - Pairwise scatterplots
  - Dimension reduction plots (PCA/MDS)
  - Distance heatmaps, dendrograms, clustering
  - Row Z-score heatmaps
  - Line plots, box plots
- Differential expression results
  - Bar plots
  - Volcano plots
  - Log fold change heatmaps
  - MA-plots

Gene co-expression: correlations & clustering

- Find how genes are correlated with each other (distance metrics)
- Cluster genes based on this metric
- Plot as heatmap
- Plot as undirected networks, confirm expression behavior in individual genes



[Oliver *et al.* 2014]



[Schlamp *et al.* 2019]

Functional enrichment

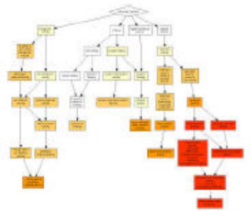
- Input:
  - subset of genes
  - database that assigns functions to genes
- Online resources
  - No coding /programming skills necessary
  - Databases are updated constantly (technically)
  - Limited reproducibility ☹

GENE ONTOLOGY
Unifying Biology

KEGG
Kyoto Encyclopedia of Genes and Genomes

PANTHER
Classification System

GORILLA



Gene Ontology enRIchment anaLysis and visuaLizAtion tool

DAVID BIOINFORMATICS DATABASE

reactome

Enrichr

INGENUITY
PATHWAY ANALYSIS

Metascape
A Gene Annotation & Analysis Resource

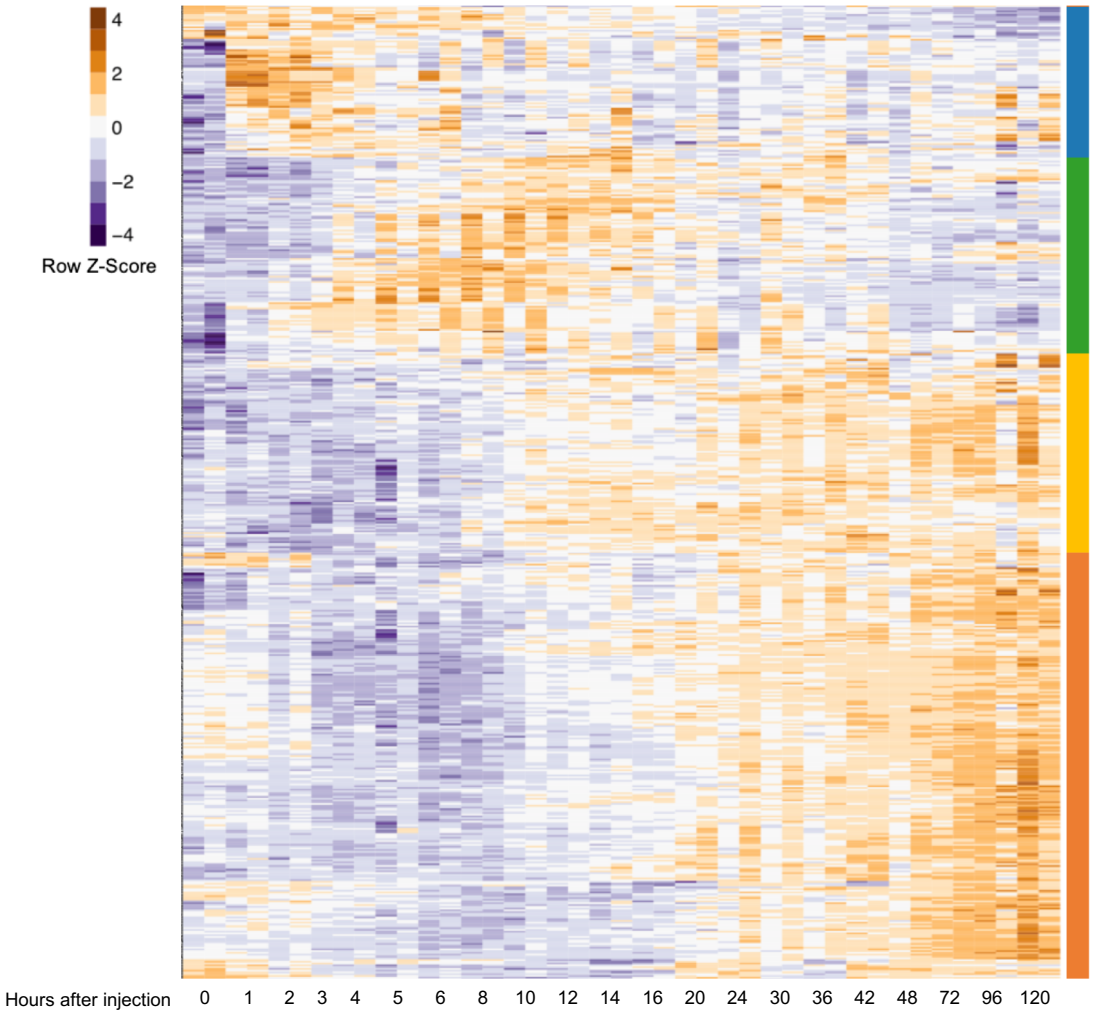
More on databases on April 27th

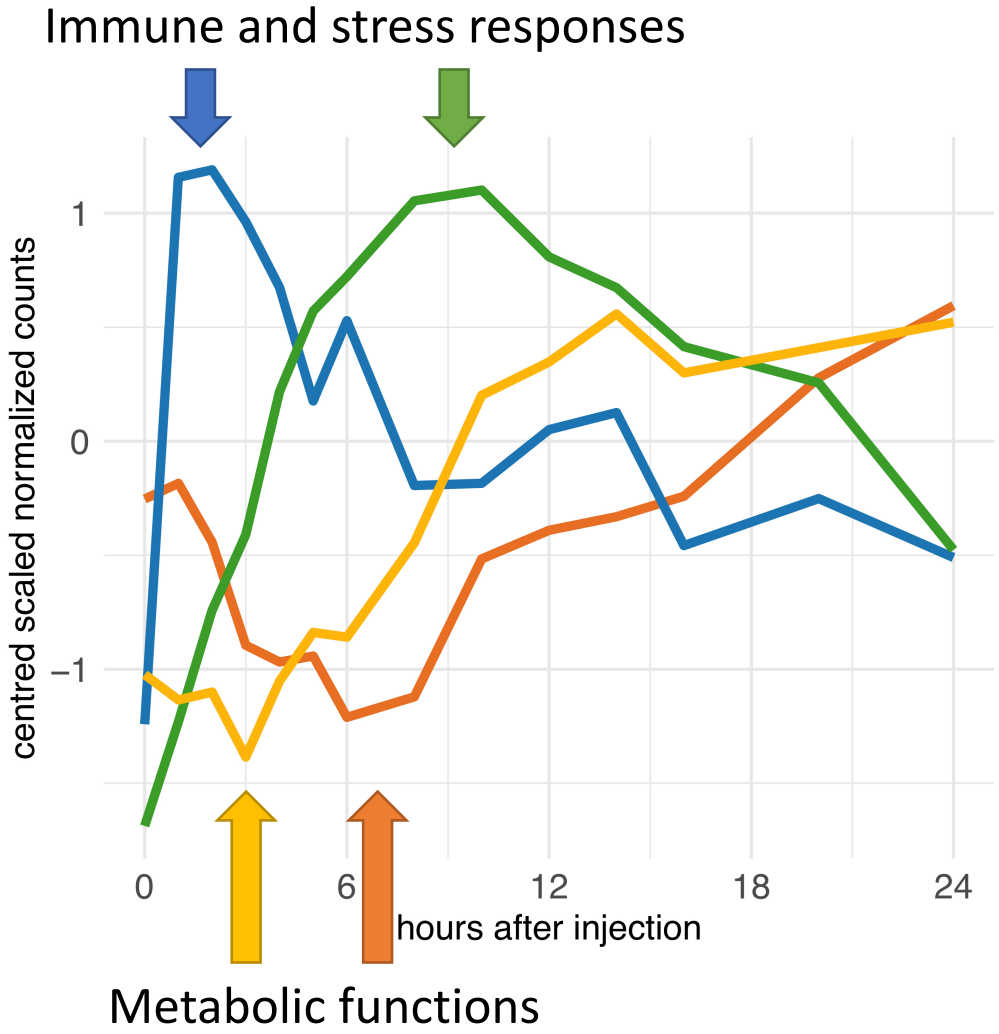
Example: Functional enrichment analysis on genes with highest variance

- Selected 20 genes with the highest variance across samples
- PANTHER Overrepresentation Test using GO Ontology database for *Drosophila Melanogaster*

GO biological process	#	expected	Fold Enrichment	P value
defense response to Gram-positive bacterium	7	0.06	> 100	3.03E-10
defense response to bacterium	8	0.25	31.5	1.94E-07
response to bacterium	12	0.3	39.48	3.66E-14
response to external biotic stimulus	13	0.45	28.91	4.98E-14
defense response to other organism	8	0.32	24.95	1.21E-06
defense response	9	0.51	17.57	1.56E-06
cellular response to heat	4	0.03	> 100	1.21E-04
antibacterial humoral response	4	0.04	96.07	2.50E-04
antimicrobial humoral response	5	0.11	46.79	1.76E-04
humoral immune response	6	0.13	44.57	9.10E-06

Example (cont'd): Summary of global dynamics using clusters – what pathways are overrepresented in each?

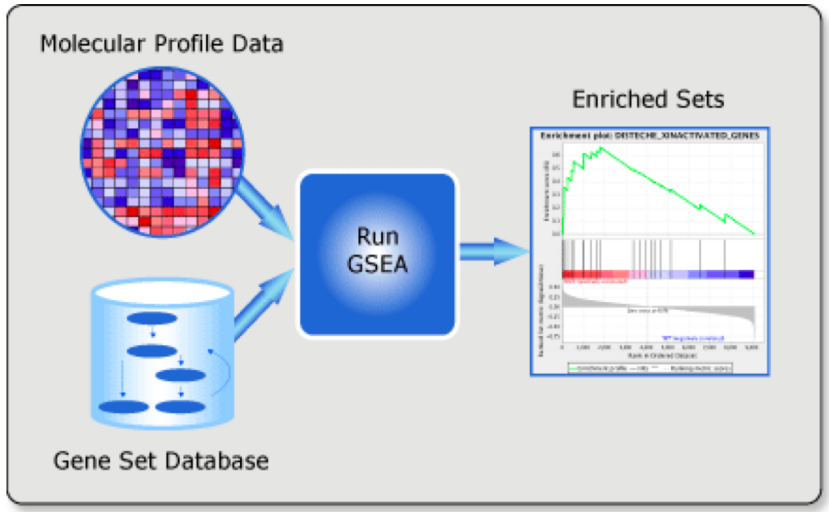




[Schlamp *et al.* 2019]

Gene Set Enrichment Analysis (GSEA)

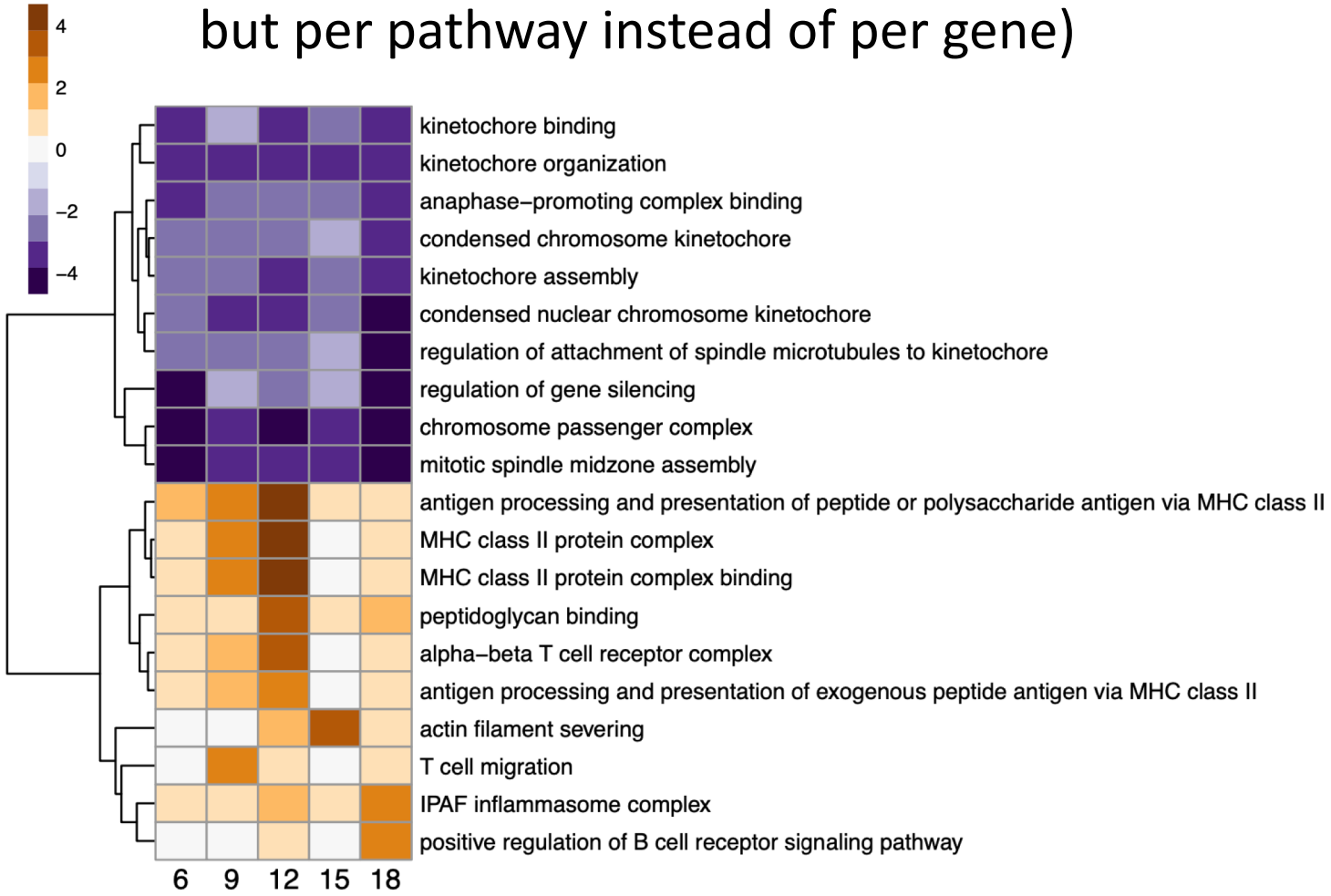
- Input:
  - genes + expression levels
  - database that assigns functions to genes
- Group of genes from X pathway might have significant expression changes as a group (when individually they do not)

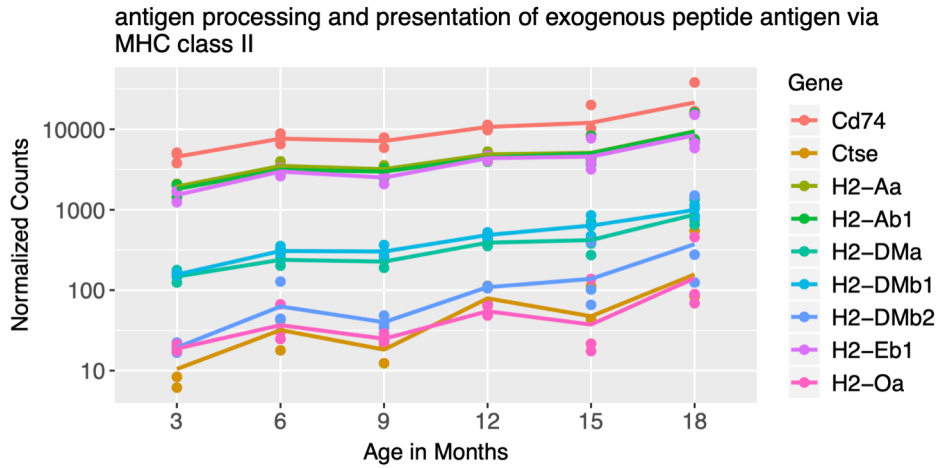


[Subramanian *et al.* 2005]

Gene Set Enrichment Analysis (GSEA)

- Output: one score and p-value per comparison (similar to differential expression, but per pathway instead of per gene)





[Zhang *et al.* 2020]

Note: consideration on inputs

- Gene names vs Gene IDs
- Check organism!
- Translate

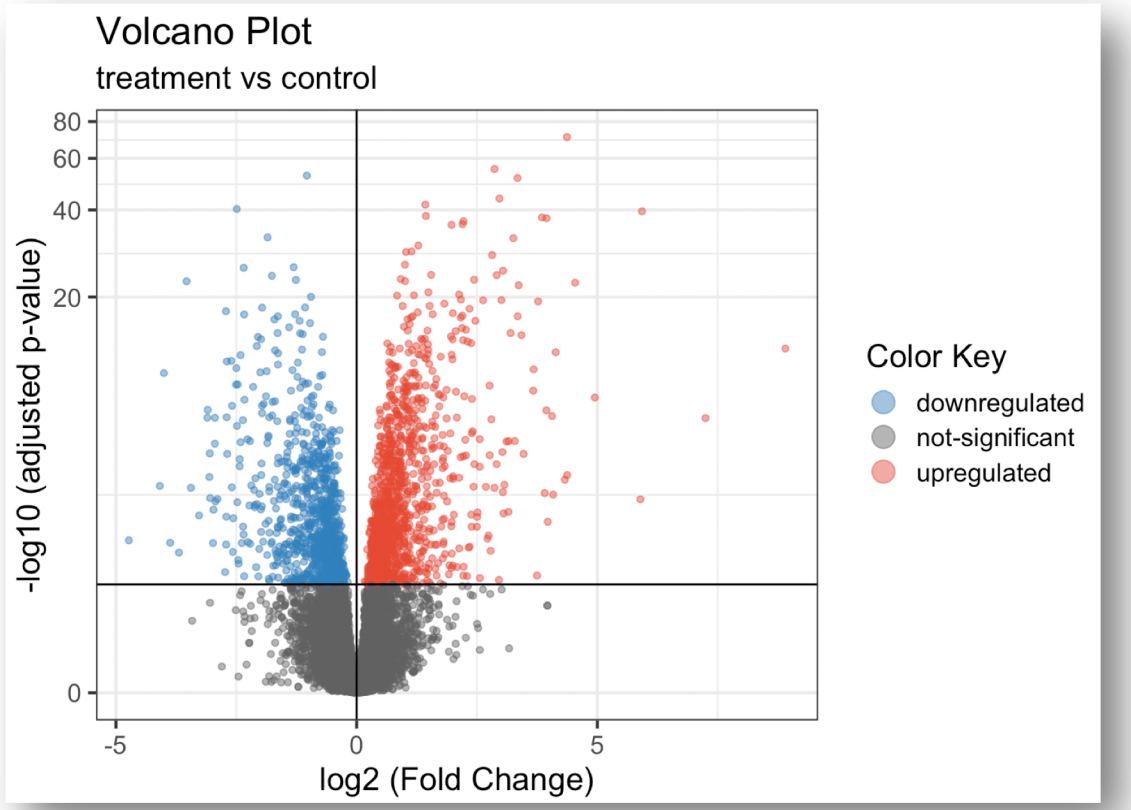
GeneID	GeneSymbol	entrezgene	HumanGene
ENSMUSG00000000001	Gnai3	14679	GNAI3
ENSMUSG00000000003	Pbsn	54192	PBSN
ENSMUSG00000000028	Cdc45	12544	CDC45
ENSMUSG00000000031	H19	NA	H19
ENSMUSG00000000037	Scml2	107815	SCML2
ENSMUSG00000000049	Apoh	11818	APOH
ENSMUSG00000000056	Narf	67608	NARF
ENSMUSG00000000058	Cav2	12390	CAV2
ENSMUSG00000000078	Klf6	23849	KLF6
ENSMUSG00000000085	Scmh1	29871	SCMH1
ENSMUSG00000000088	Cox5a	12858	COX5A
ENSMUSG00000000093	Tbx2	21385	TBX2

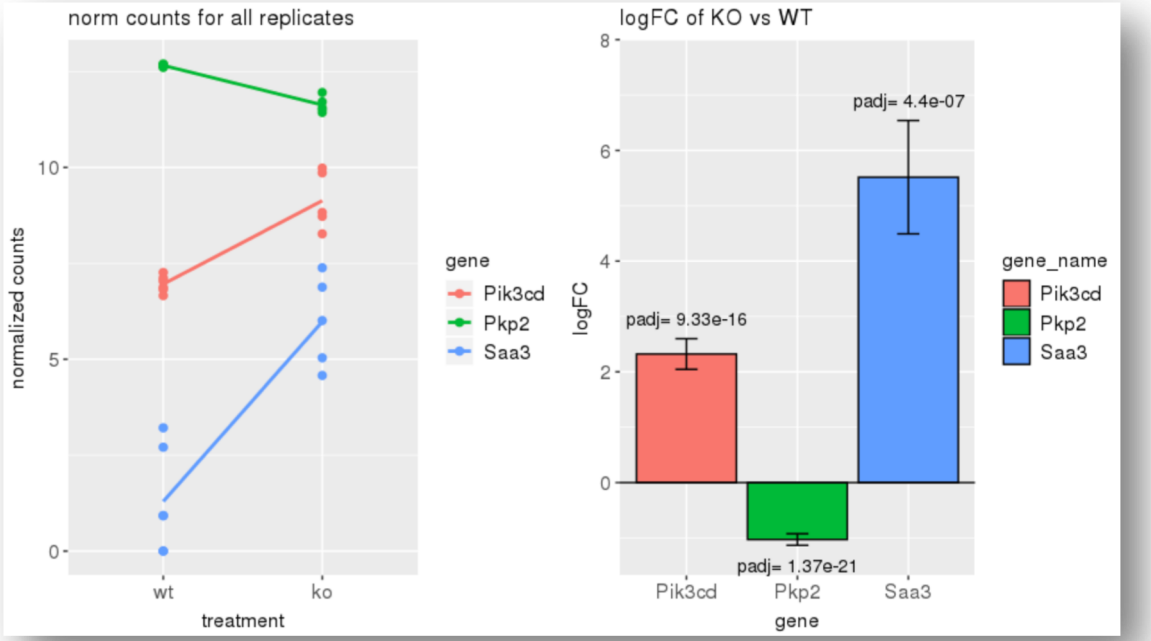
Summary

- Normalized counts
  - Pairwise scatterplots
  - Dimension reduction plots (PCA/MDS)
  - Distance heatmaps, dendrograms, clustering
  - Row Z-score heatmaps
  - Line plots, box plots
- Differential expression results
  - Bar plots
  - Volcano plots
  - Log fold change heatmaps
  - MA-plots
- Gene co-expression correlation heatmaps and networks
- Functional enrichment
- Genet Set Expression Analysis

NYU CVRC bioinformatics resources (coming soon)

- easy to use scripts to generate plots





Questions?

Florencia Schlamp, PhD | Email: Florencia.Schlamp@nyulangone.edu | Office: SB 604

Upcoming bioinformatics lectures

NYU Cardiovascular Research Center

- March 23rd - Inside the Black Box: The steps of RNA-seq data processing, data exploration, and data analysis
- April 27th - Intro to Machine Learning (guest lecturer from Cornell)
- TBD (May/June) - Analysis consideration when designing an RNA-seq experiment
- August 24th - Biodatabases (guest lecturer from Cornell)