CVRC Bioinformatics Workshop Series 2020

Day 2

Data Exploration, Part 1: Basic RNA-seq Data Manipulation

Florencia Schlamp, PhD

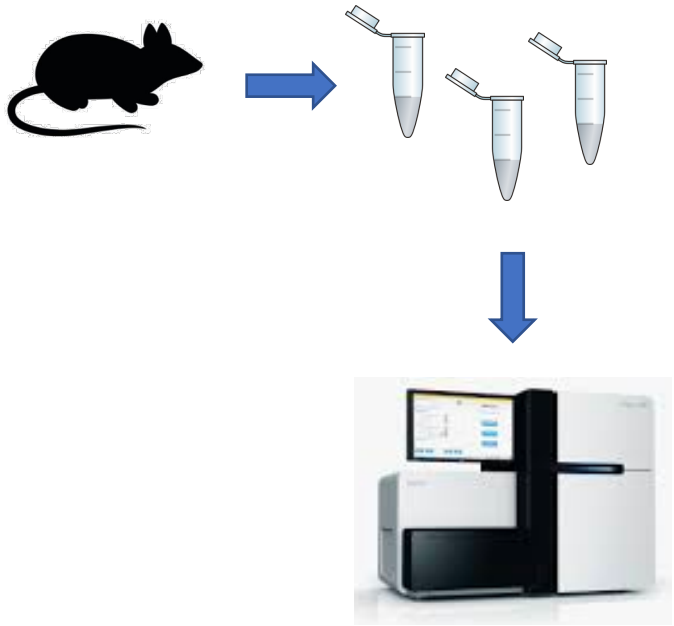
Friday, October 9th, 2020

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

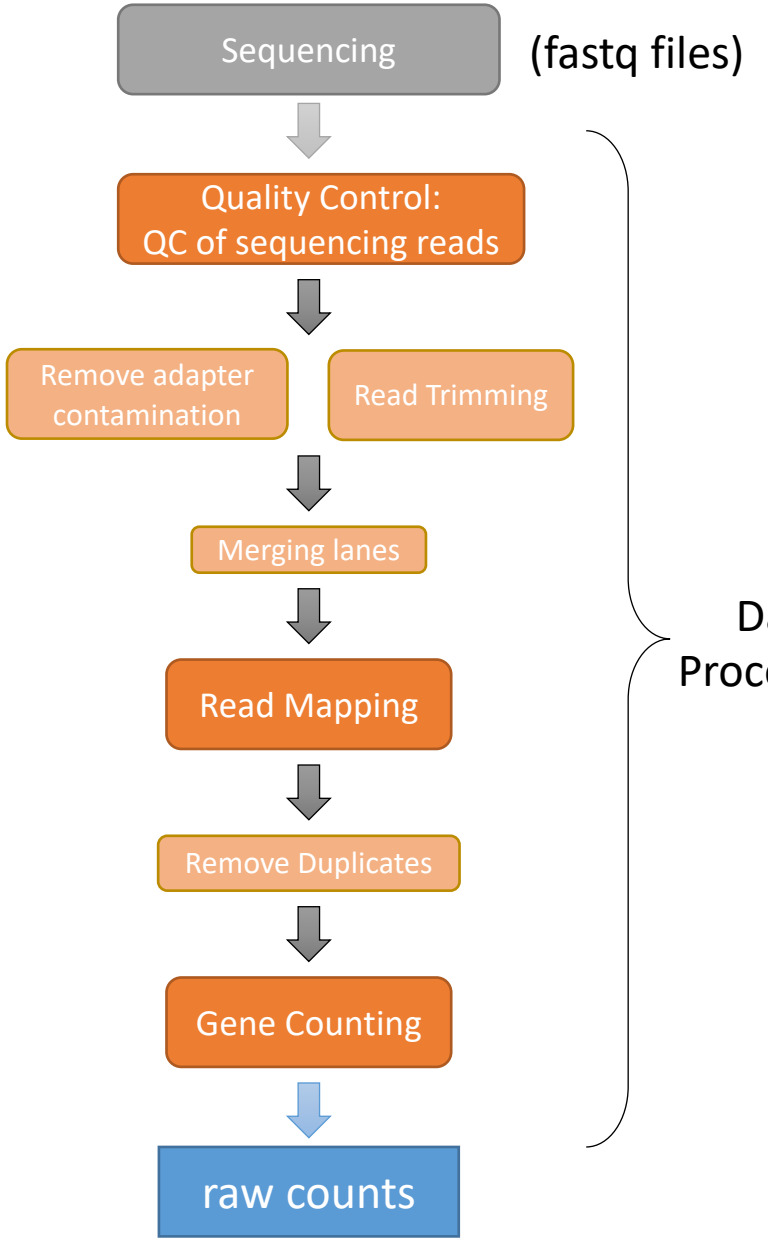
RNA sequencing



Data Generation

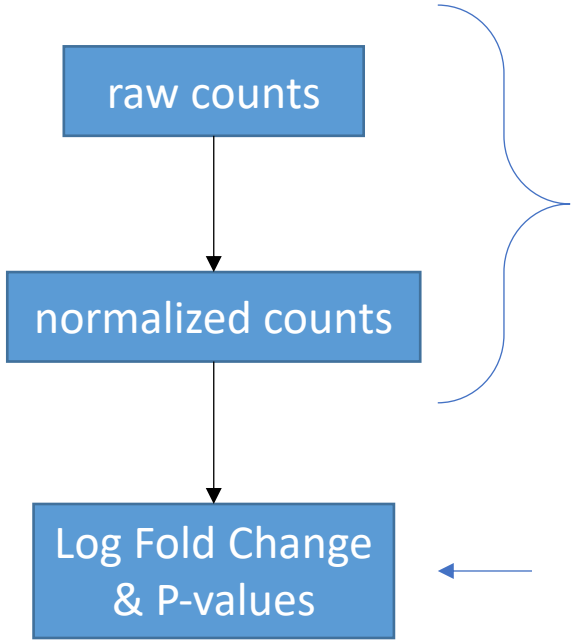
- Sequencing methods vary (read length, coverage, price, speed, accuracy, etc.)

More on design considerations TBD



More on data processing TBD

Data Types



this session (and next)

Days 4 and 5 (November)

Data
Exploration

Data
Analysis

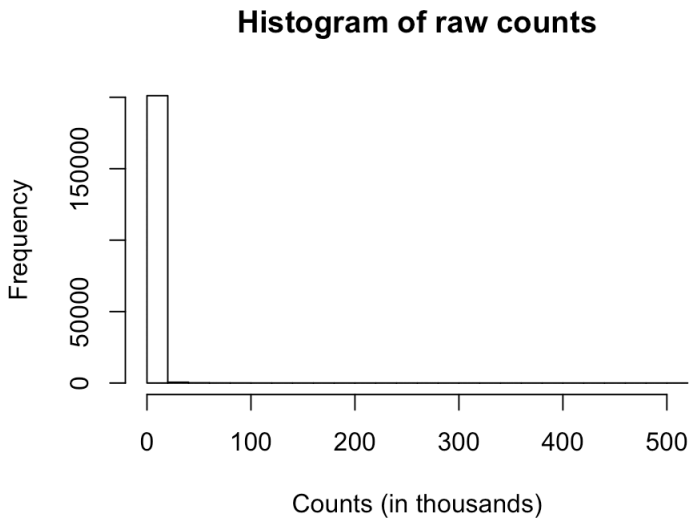
Data Types

raw counts

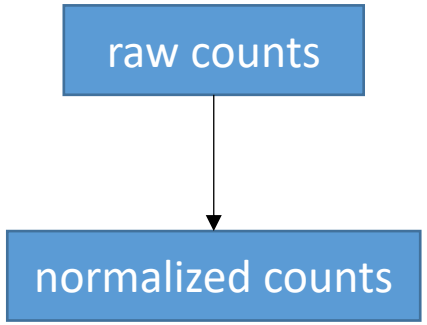
one value per gene per sample

	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

one value per gene per sample

Data Types

raw counts

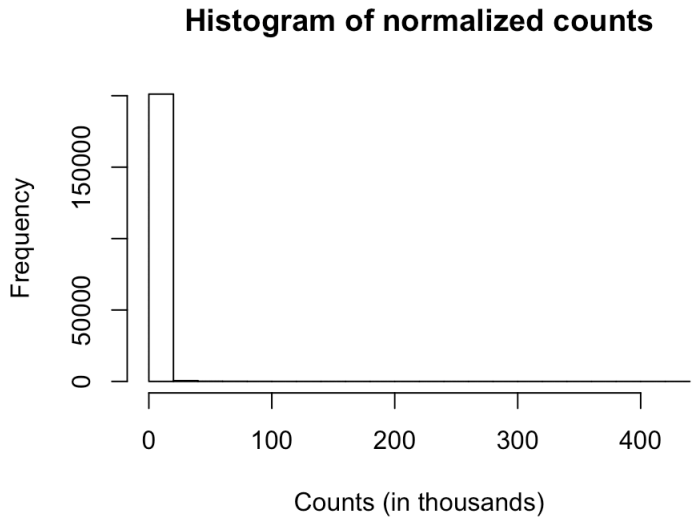
↓

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

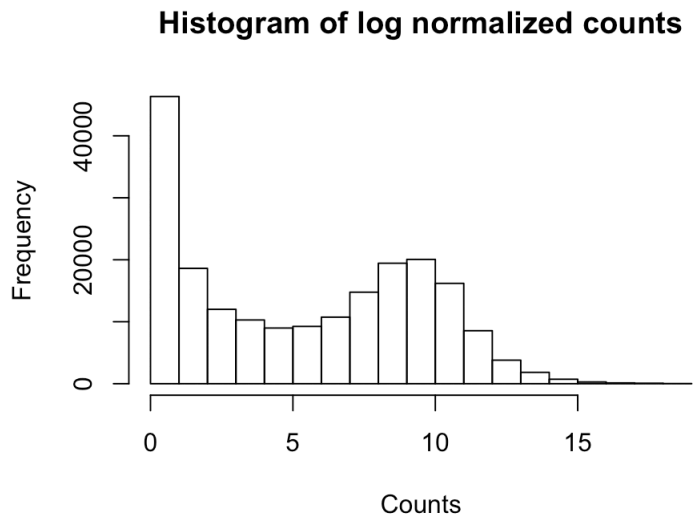
```
> range(norm_counts)
[1]      0.0 433308.8
```



2 normalized + log transformed

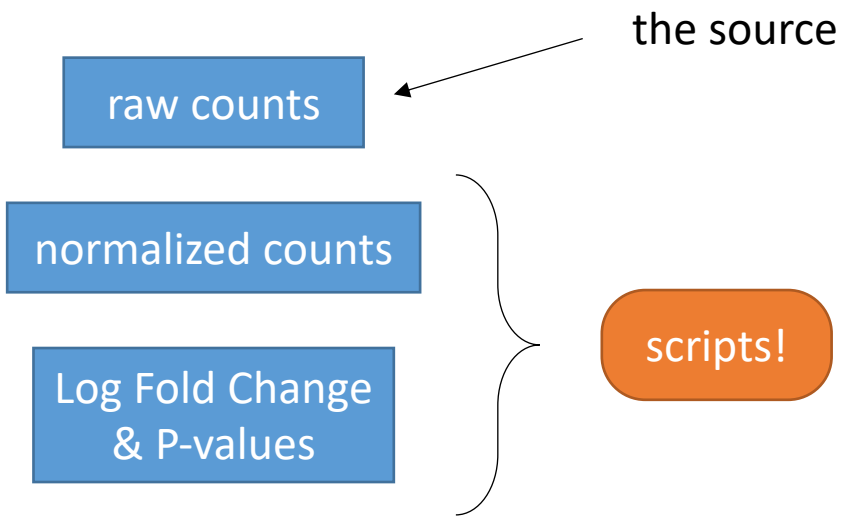
	C1	T1	C2	T2
ENSG00000000003	9.566103	9.118866	9.5866522	9.297471
ENSG00000000419	8.936573	9.224537	9.0341368	9.035542
ENSG00000000457	8.509104	8.207573	8.2770488	8.390563
ENSG00000000460	6.665591	6.546820	5.9775003	6.557310
ENSG00000000938	0.000000	0.000000	0.8870939	0.000000
ENSG00000000971	11.803575	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
```



Rigor and reproducibility note

- Files you should have / save:



Normalized counts (after data transformation)

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

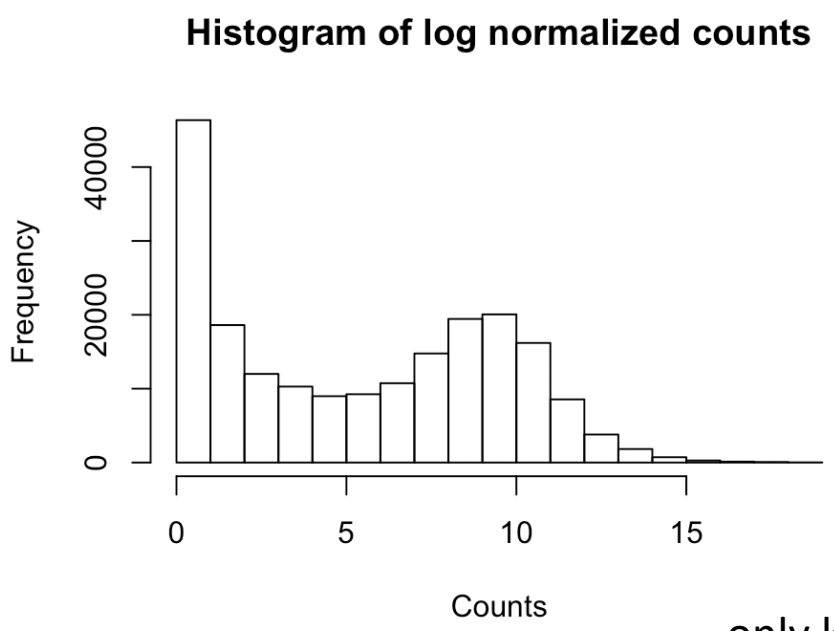
	C1	T1	C2	T2	C3	T3	C4	T4
ENSG00000000003	9.566103	9.118866	9.5866522	9.297471	10.0119186	9.5778244	9.825015	9.257247
ENSG00000000419	8.936573	9.224537	9.0341368	9.035542	9.0059121	9.0884190	8.875766	9.010804
ENSG00000000457	8.509104	8.207573	8.2770488	8.390563	8.1362475	8.2853502	8.538994	8.360729
ENSG00000000460	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)

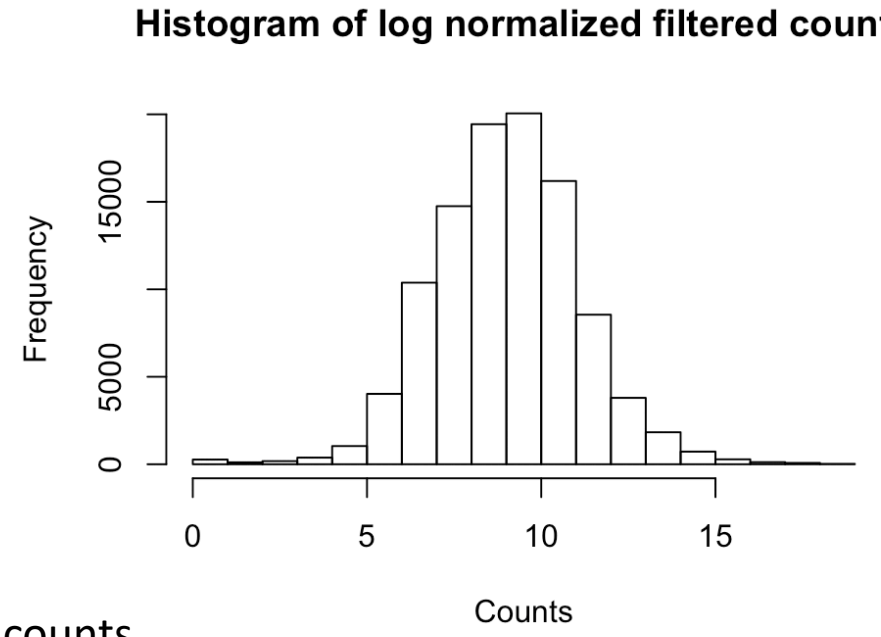
- Quality Control
  - filter out genes with low counts

HOW?

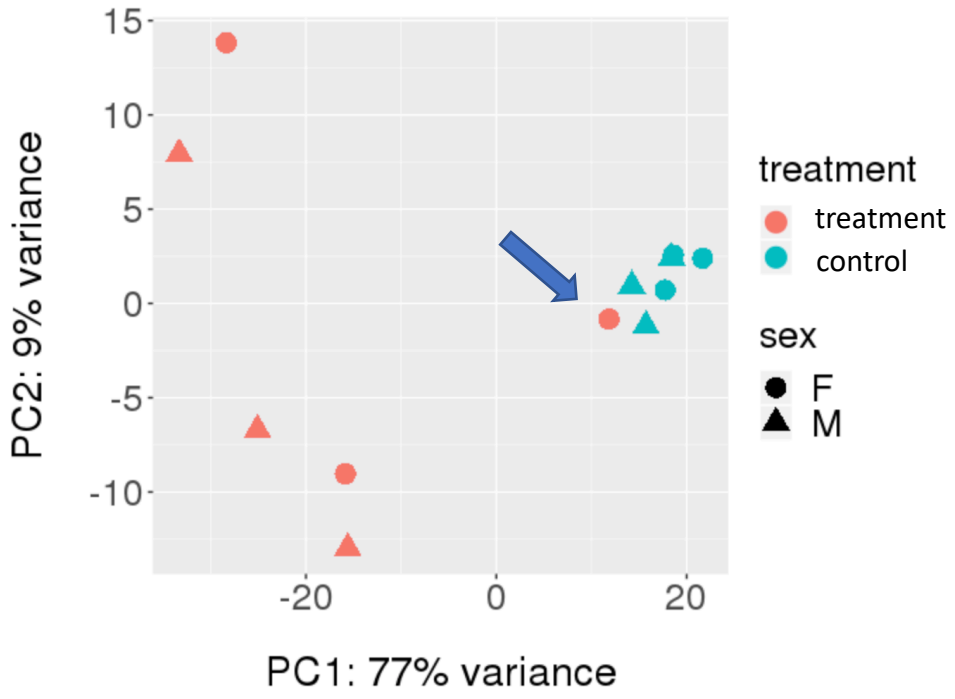
- Histograms (distribution of counts)



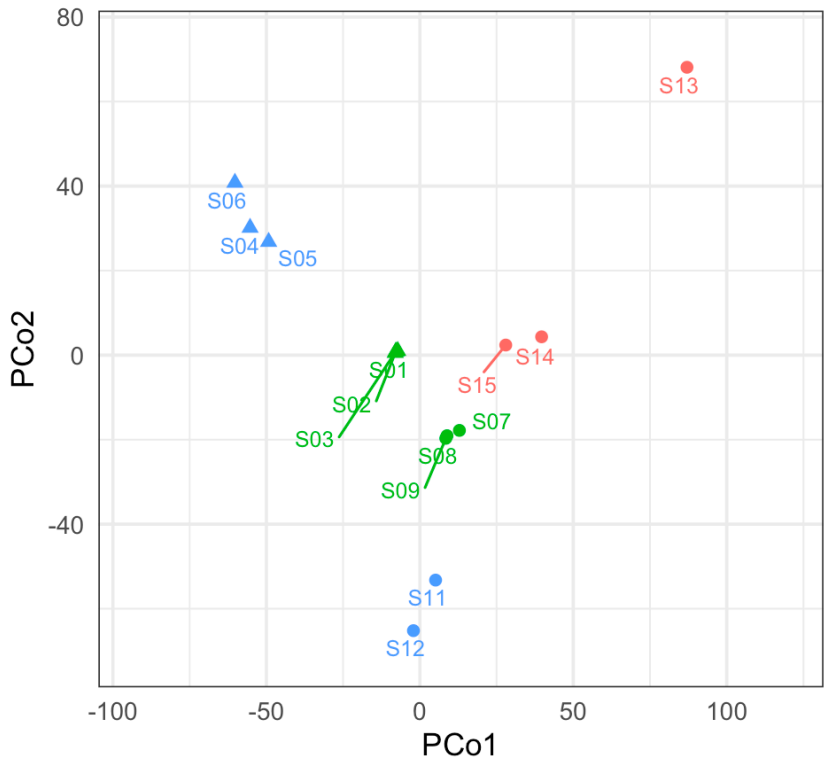
only keep genes with 6 counts
in at least 2 samples

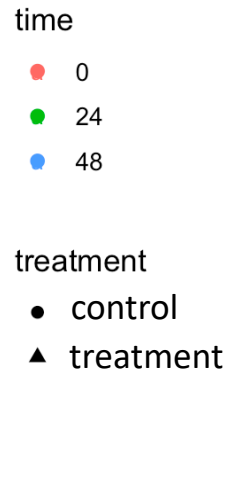


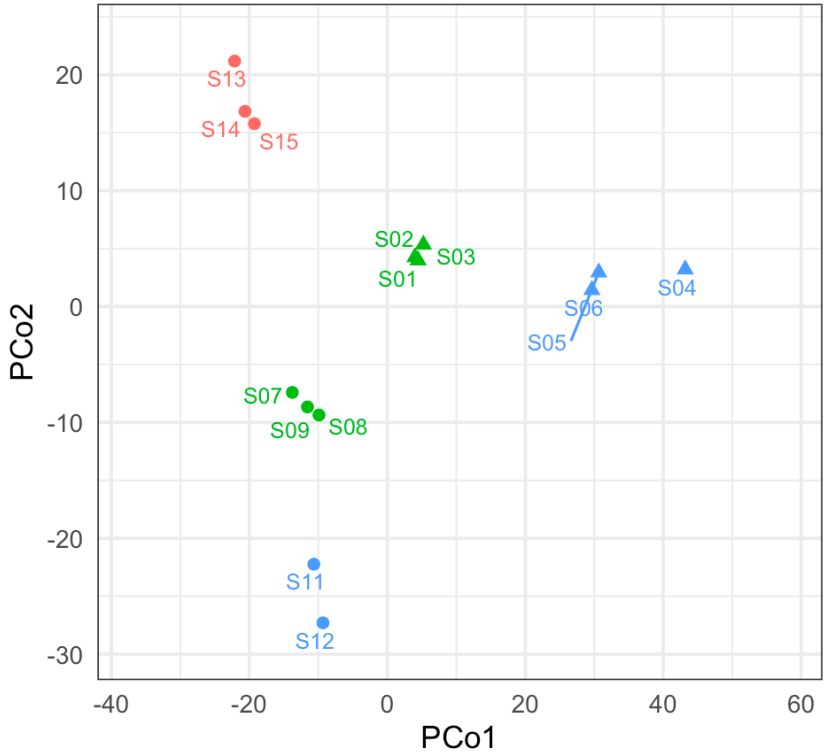
Normalized counts (after data transformation)

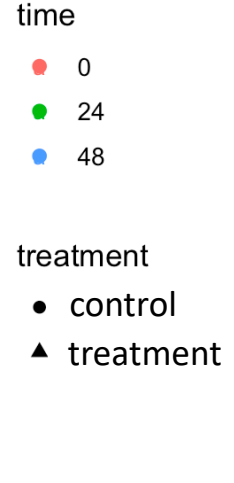


Example: improving clustering after removing genes with low counts









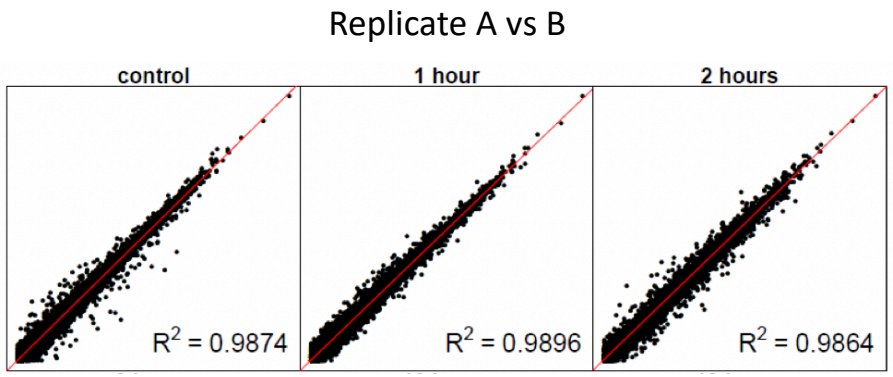
Only keep genes with 70 raw counts
in at least one sample

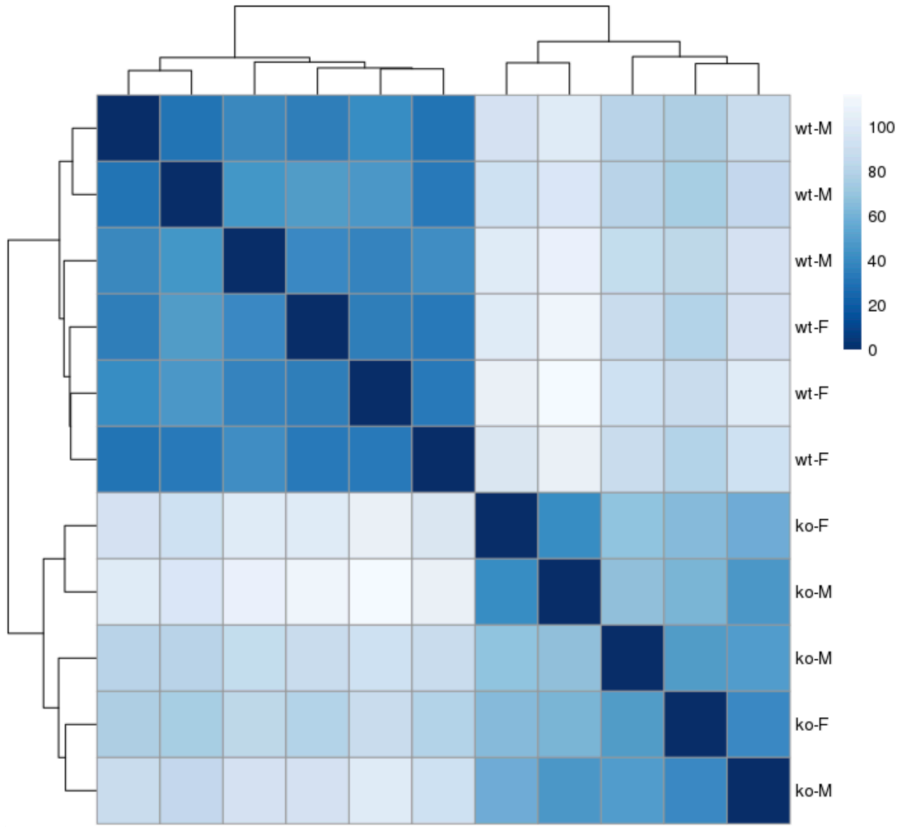
(29,656 to 11,343 genes)

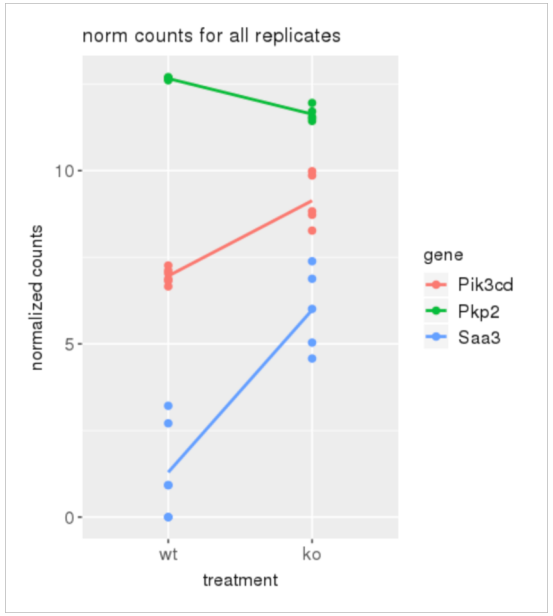
Next session

Day 3: “Data Exploration, Part 2: Basic RNA-seq Plots”

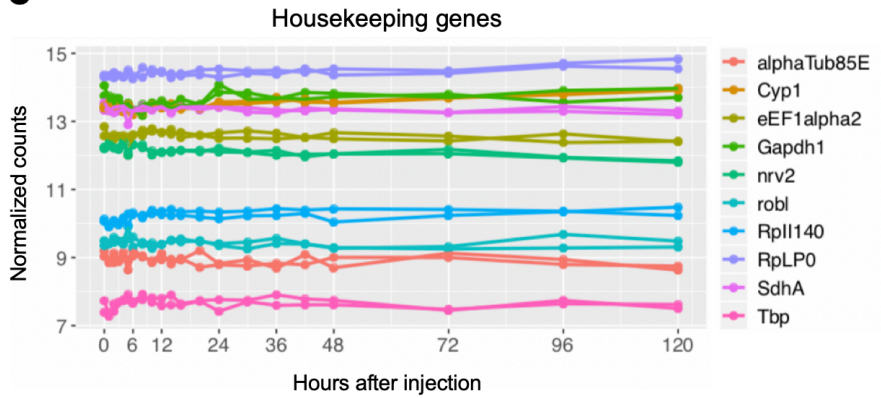
Friday October 23rd

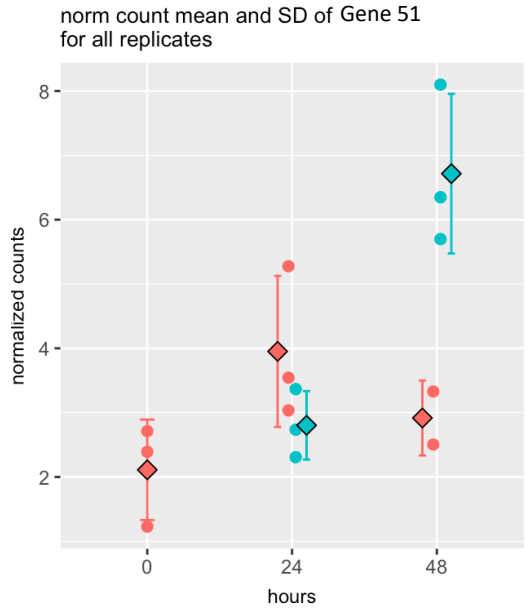

Replicate A vs B
control
1 hour
2 hours
R2 = 0.9874
R2 = 0.9896
R2 = 0.9864

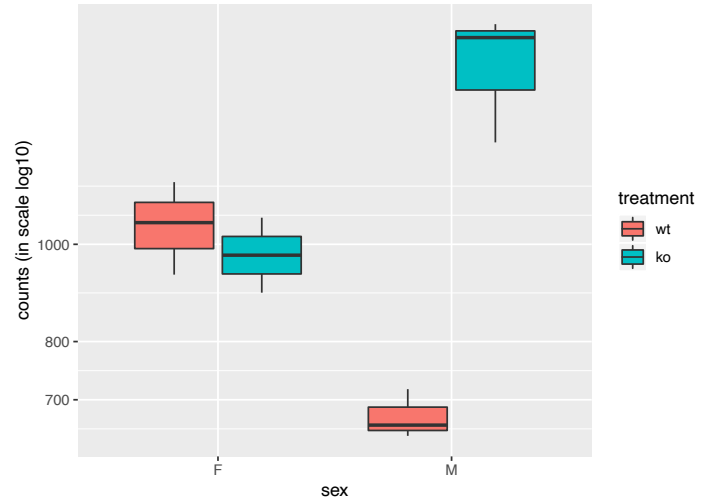

wt-M
wt-M
wt-M
wt-F
wt-F
wt-F
ko-F
ko-M
ko-M
ko-F
ko-M
100
80
60
40
20
0


norm counts for all replicates
normalized counts
10
5
0
wt
ko
treatment
gene
Pik3cd
Pkp2
Saa3

C


Housekeeping genes
Normalized counts
15
13
11
9
7
0
6
12
24
36
48
72
96
120
Hours after injection
alphaTub85E
Cyp1
eEF1alpha2
Gapdh1
nrv2
robl
RpII140
RpLP0
SdhA
Tbp


norm count mean and SD of Gene 51
for all replicates
normalized counts
8
6
4
2
0
24
48
hours
treatment
neg
chol


counts (in scale log10)
1000
800
700
F
M
sex
treatment
wt
ko

Questions?

Florencia Schlamp, PhD | Email: Florencia.Schlamp@nyulangone.edu