Access the slides and files here:

https://github.com/j-berg/bioinformatics_bootcamp

#5.2

DESeq2

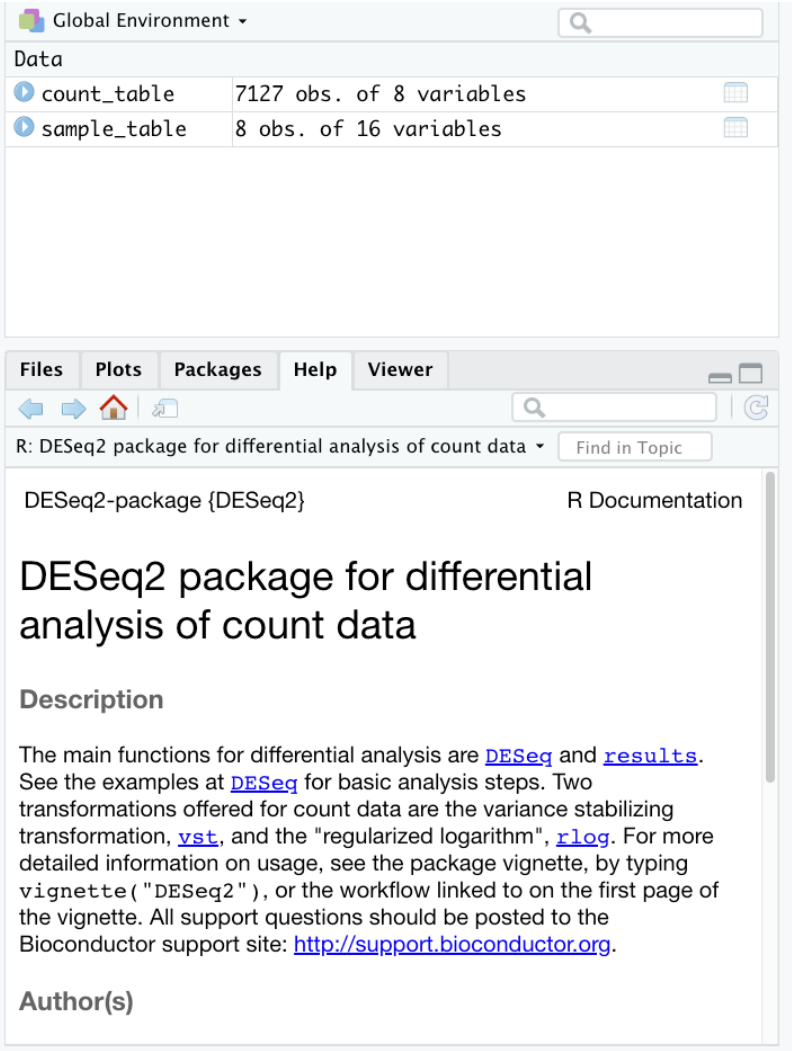
Multiple Hypothesis Testing

Data normalization

Scatter Plots

Analyzing count data using DESeq2

```
> library(DESeq2)
> count_table <- read.table("~/Desktop/bioinformatics-bootcamp/class_
5_2/sce_mct1_03hr_counts.txt", sep = '\t', header = TRUE, row.names =
 1, check.names=F)
> sample_table <- read.table(text = readLines("~/Desktop/bioinformati
cs-bootcamp/class_5_2/sce_mct1_03hr_metadata.txt", warn = FALSE), hea
der = TRUE, sep = '\t')
>
> head(count_table)
       14251X4 14251X6 14251X10 14251X12 14251X16 14251X18
ETS1-1       0       0        0        0        0        0
ETS1-2       0       0        0        0        0        0
ETS2-1       0       0        0        0        0        0
ETS2-2       0       0        0        0        0        0
HRA1         4       1        0        1        2        1
ICR1       191     128      178       86      118      100
       14251X22 14251X24
ETS1-1        0        0
ETS1-2        0        0
ETS2-1        0        0
ETS2-2        0        0
HRA1          4        3
ICR1        132       72
>
> head(sample_table)
         X Genotype     Media Time Conc...ng.uL. Volume..uL.
1  14251X4     a_WT Raffinose 3 hr         838.5          10
2  14251X6  mct1del Raffinose 3 hr         103.2          10
3 14251X10     a_WT Raffinose 3 hr        1124.2          10
4 14251X12  mct1del Raffinose 3 hr         293.7          10
5 14251X16     a_WT Raffinose 3 hr         325.2          10
6 14251X18  mct1del Raffinose 3 hr         401.3          10
            Sample.Type                Organism
1 Total RNA (eukaryote) Saccharomyces cerevisiae
2 Total RNA (eukaryote) Saccharomyces cerevisiae
3 Total RNA (eukaryote) Saccharomyces cerevisiae
```



Counts

	14251X4	14251X6	14251X10	14251X12	14251X16	14251X18	14251X22	14251X24
ETS1-1	0	0	0	0	0	0	0	0
ETS1-2	0	0	0	0	0	0	0	0
ETS2-1	0	0	0	0	0	0	0	0
ETS2-2	0	0	0	0	0	0	0	0
HRA1	4	1	0	1	2	1	4	3
ICR1	191	128	178	86	118	100	132	72
IRT1	410	615	429	718	391	760	510	806
ITS1-1	0	0	0	0	0	0	0	0
ITS1-2	0	0	0	0	0	0	0	0
ITS2-1	0	0	0	0	0	0	0	0
ITS2-2	0	0	0	0	0	0	0	0
LSR1	334	295	281	398	610	294	204	417

Metadata

	Genotype	Media	Time	Conc. (ng/uL	Volume (uL)	Sample Type	Organism	QC Conc. (ng	QC RIN	Seq Lib Protc	Prepped by C	Index Tag A	Index Tag Se	Sample Nam	Lib QC Conc.
14251X4	a_WT	Raffinose	3 hr	838.5	10	Total RNA (e	Saccharomyc	660	8.4	Illumina Tru	Y	Index 10 TAC	TAGCTT	SMN469	27
14251X6	mct1del	Raffinose	3 hr	103.2	10	Total RNA (e	Saccharomyc	69	8.9	Illumina Tru	Y	Index 16 CCG	CCGTCC	SMN471	19
14251X10	a_WT	Raffinose	3 hr	1124.2	10	Total RNA (e	Saccharomyc	817	7.9	Illumina Tru	Y	Index 5 ACAC	ACAGTG	SMN477	32
14251X12	mct1del	Raffinose	3 hr	293.7	10	Total RNA (e	Saccharomyc	234	8.8	Illumina Tru	Y	Index 11 GG(	GGCTAC	SMN479	29
14251X16	a_WT	Raffinose	3 hr	325.2	10	Total RNA (e	Saccharomyc	246	8.8	Illumina Tru	Y	Index 25 ACT	ACTGAT	SMN493	17
14251X18	mct1del	Raffinose	3 hr	401.3	10	Total RNA (e	Saccharomyc	308	9	Illumina Tru	Y	Index 6 GCCA	GCCAAT	SMN495	28
14251X22	a_WT	Raffinose	3 hr	292.9	10	Total RNA (e	Saccharomyc	216	9.3	Illumina Tru	Y	Index 19 GTC	GTGAAA	SMN501	28
14251X24	mct1del	Raffinose	3 hr	433.1	10	Total RNA (e	Saccharomyc	337	9.2	Illumina Tru	Y	Index 27 ATT	ATTCCT	SMN503	31

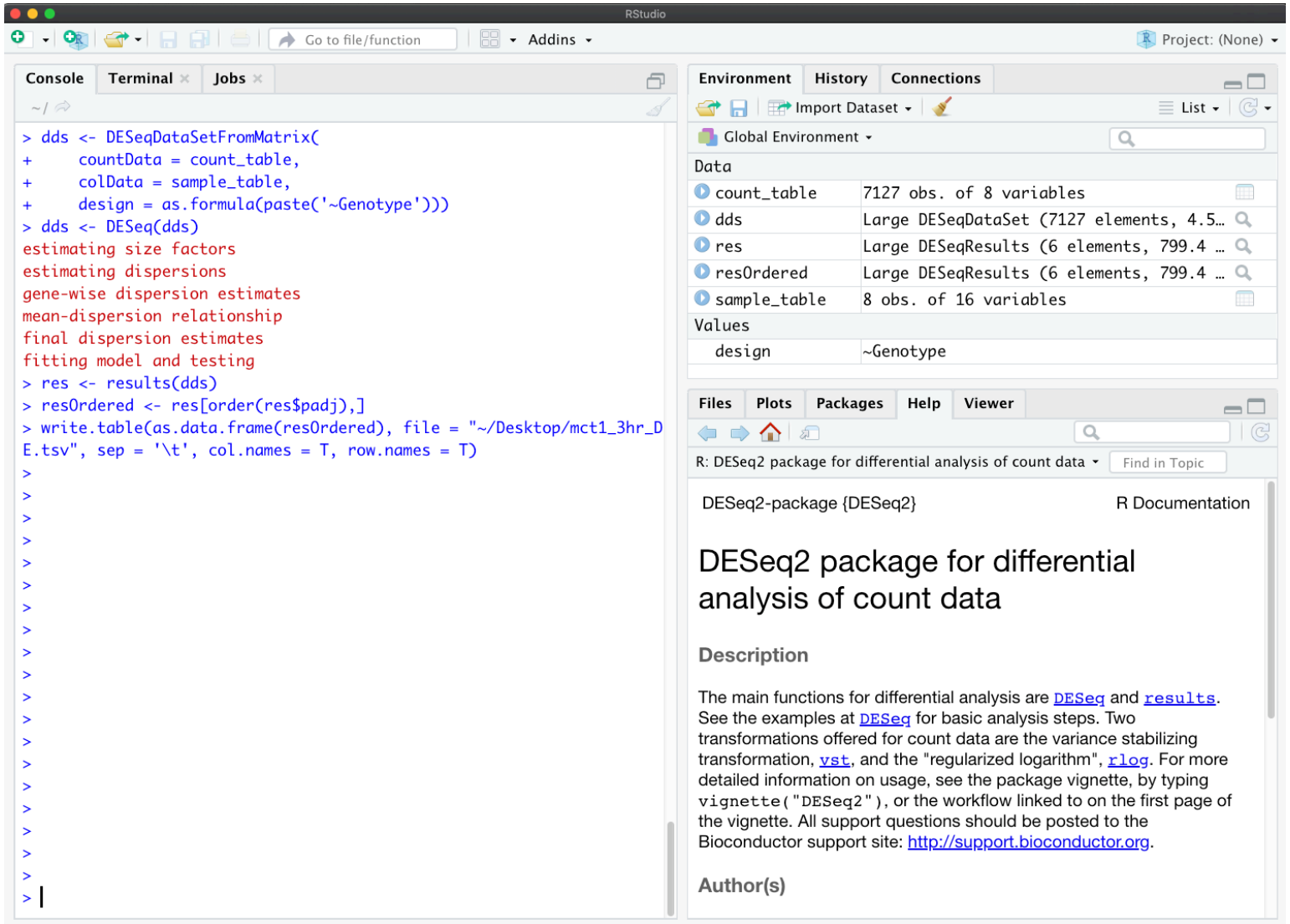
Commands

```
library(DESeq2)

count_table <- read.table("~/Desktop/bioinformatics-bootcamp/class_5_2/sce_mct1_03hr_counts.txt", sep = '\t', header = TRUE, row.names = 1, check.names=F)
head(count_table)

sample_table <- read.table(text = readLines("~/Desktop/bioinformatics-bootcamp/class_5_2/sce_mct1_03hr_metadata.txt", warn = FALSE), header = TRUE, sep = '\t')
head(sample_table)
```

Analyzing count data using DESeq2



Commands

```
dds <- DESeqDataSetFromMatrix(
    countData = count_table,
    colData = sample_table,
    design = ~Genotype)

dds <- DESeq(dds)
res <- results(dds)
resOrdered <- res[order(res$padj),]

write.table(as.data.frame(resOrdered), file = "~/Desktop/mct1_3hr_DE.tsv",
sep = '\t', col.names = T, row.names = T)
```

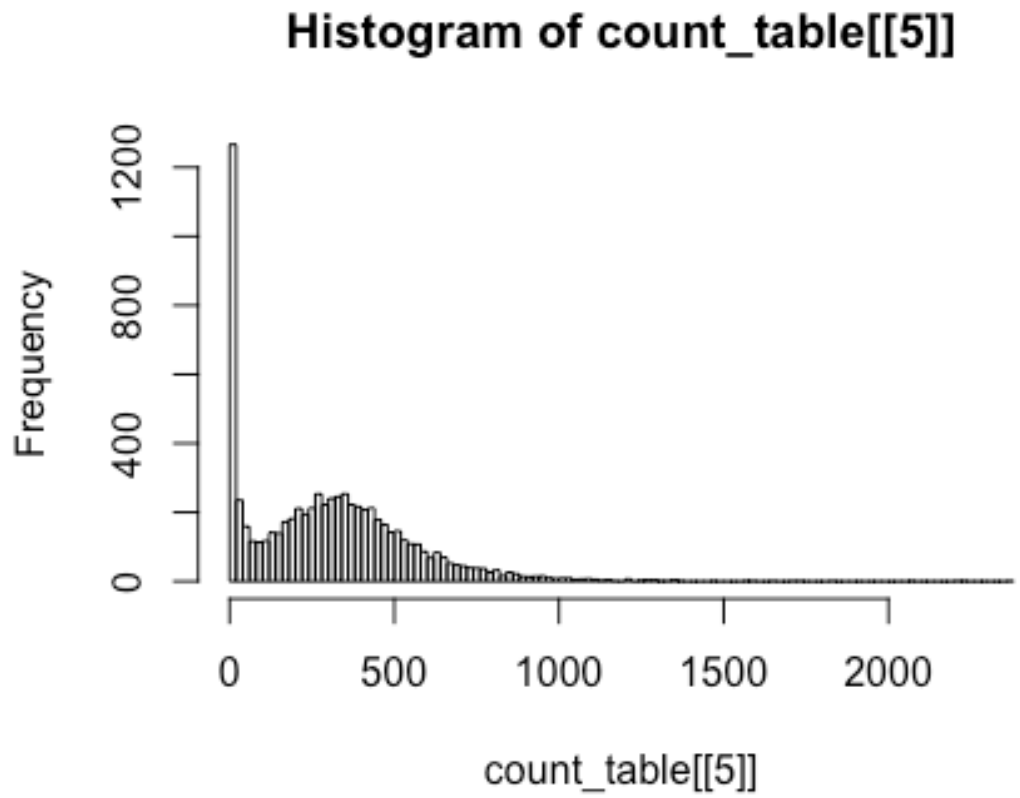
Output

	baseMean	log2FoldChai	lfcSE	stat	pvalue	padj
YOR221C	252.513057	-4.0339467	0.15978753	-25.245691	1.26E-140	8.32E-137
YHR043C	267.562037	-3.1089257	0.14671387	-21.190401	1.17E-99	3.86E-96
YML042W	805.486894	1.54856075	0.09492723	16.3131356	7.96E-60	1.75E-56
YLR303W	205.547163	7.30981239	0.47105544	15.5179449	2.62E-54	4.32E-51
YAL054C	766.114005	1.22135022	0.08142357	14.9999588	7.35E-51	9.68E-48
YPR001W	696.549289	1.4991631	0.10393206	14.4244534	3.63E-47	3.99E-44
YPR002W	838.089191	1.16662688	0.08766514	13.3077626	2.09E-40	1.96E-37
YGR234W	688.687757	-1.388017	0.10656542	-13.025023	8.82E-39	7.26E-36
YER024W	1225.24614	1.12154017	0.09916914	11.3093669	1.18E-29	8.64E-27
YEL071W	788.249848	1.0563048	0.09695971	10.8942656	1.23E-27	8.09E-25
YBR115C	303.861211	10.9226084	1.03445362	10.5588189	4.62E-26	2.77E-23

- Note: Your file paths for inputs and outputs may differ based on operating system

Biases in sequence libraries

- Gene counts for a given RNA-seq sample follow a negative binomial distribution
- Need to run appropriate statistical tests
- !!! Do not use a T-test, this assumes data is normally distributed which it is not!
- DESeq2 properly accounts for this



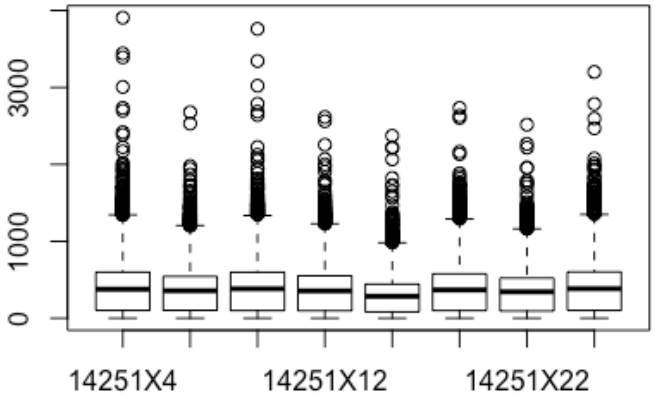
Library normalization

- The number of reads and the length of transcripts influence the number of measured reads
- A gene may have 500 counts between two samples, but if one sample had 100K reads and another had 100M total reads, 500 would account for different fractions of total reads
- A transcript that is 10kb will naturally fragment into 10x reads as a transcript that is only 1kb

- See https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

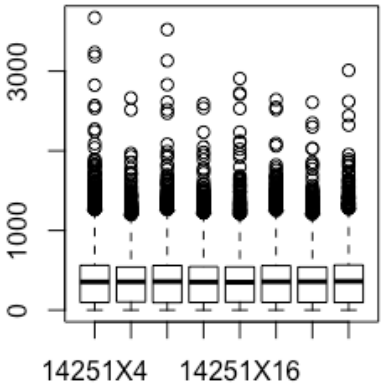
Performing normalization in R

- Since we will only be running a comparison of each gene to each other, we will demonstrate RPM to factor out library size variability



boxplot(as.matrix(count_table))

```
BiocManager::install("tweeDEseq")
library(tweeDEseq)
norm <- normalizeCounts(count_table, method="TMM")
```

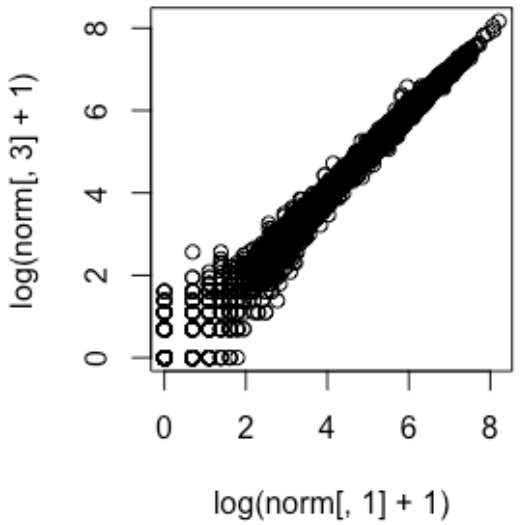


boxplot(norm)

Scatter plots to compare replicates in R

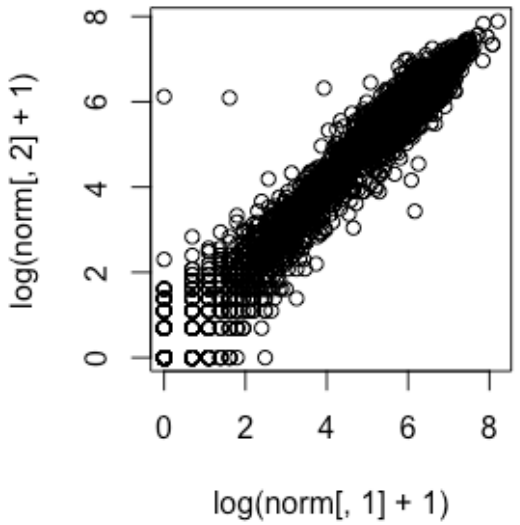
REPLICATES:

plot(log(norm[,1] + 1), log(norm[,3] + 1))



NOT REPLICATES:

plot(log(norm[,1] + 1), log(norm[,2] + 1))



Homework

- *Perform differential expression analysis of your read count table from before. Identify the 25 most up-regulated and down-regulated genes from your dataset. Do these make sense in the context of the model?*
- *Verify biological replicates look similar using scatter plots*