University of Massachusetts Medical School

Initial analysis

Summary I – Data types, file formats and utilities

- Annotation: Genomic regions
  - Genes
  - Peaks
  - *bedtools*
- Alignment: Map reads
  - BAM/SAM
  - *Samtools*
- Aggregation: Summary files
  - Wig (UCSC)
  - TDF (IGV)

Summary II – Data process

- Short read alignment (Bowtie, BWA)
  - Making the genome searchable: Hashing/BW
  - Seed an extend (hashing) vs suffix searches (BW)
  - New aligners are mix
- Spliced aligners (TopHat, STAR, GSNAP)
  - Map read fragments then strung them
  - Choosing the fragment size
  - Avoiding biases using information (junctions)
- Quantifying (RSEM/Cufflinks)
  - Read/Isoform assignment
  - Normalization procedures
- Differential expression (DESeq/EdgeR/Cufflinks)

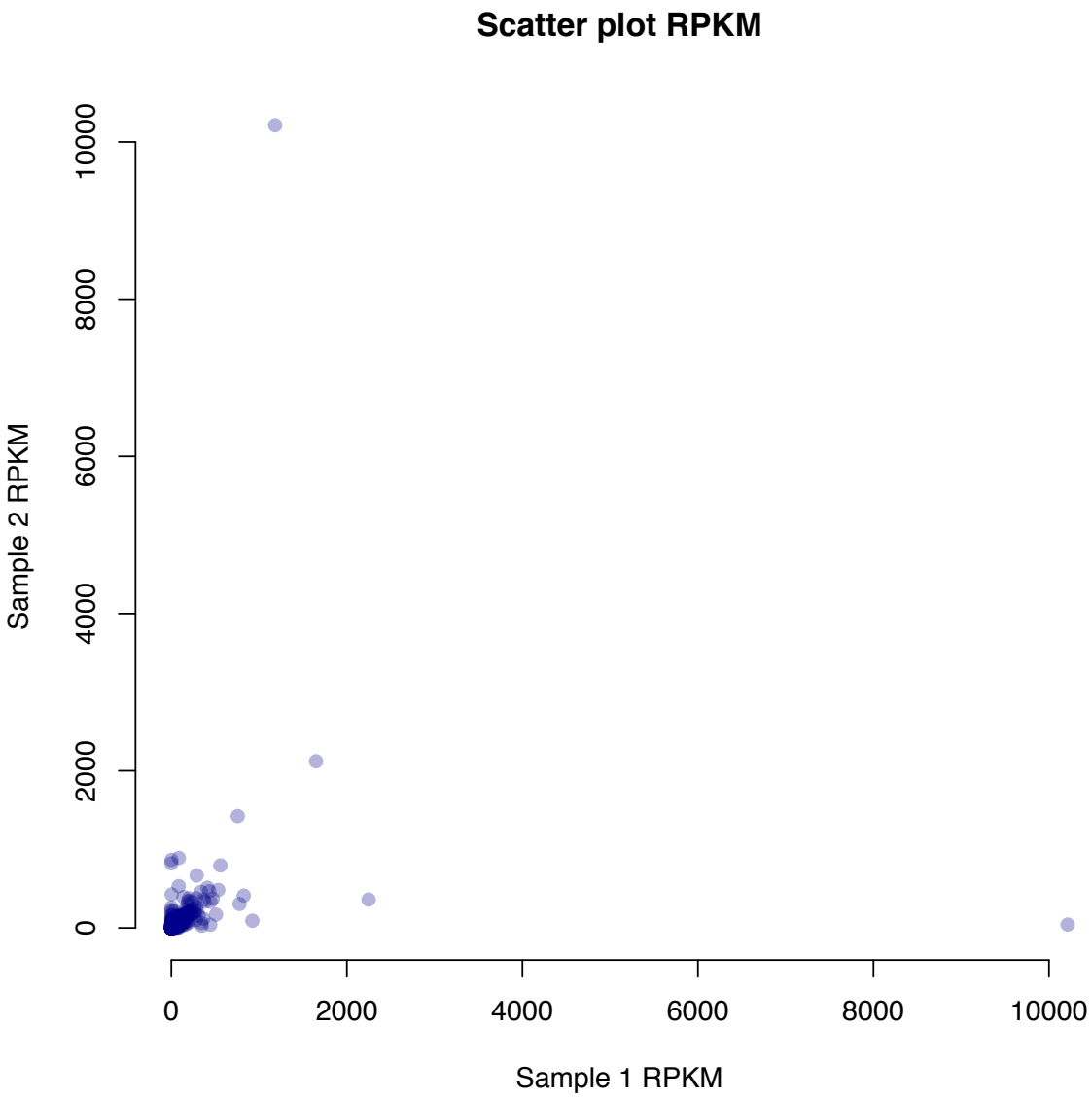
Summary III – Using a graphical user interface

- Galaxy – for knowledgable users who are not comfortable with UNIX
- All tools available
- Not great for many samples

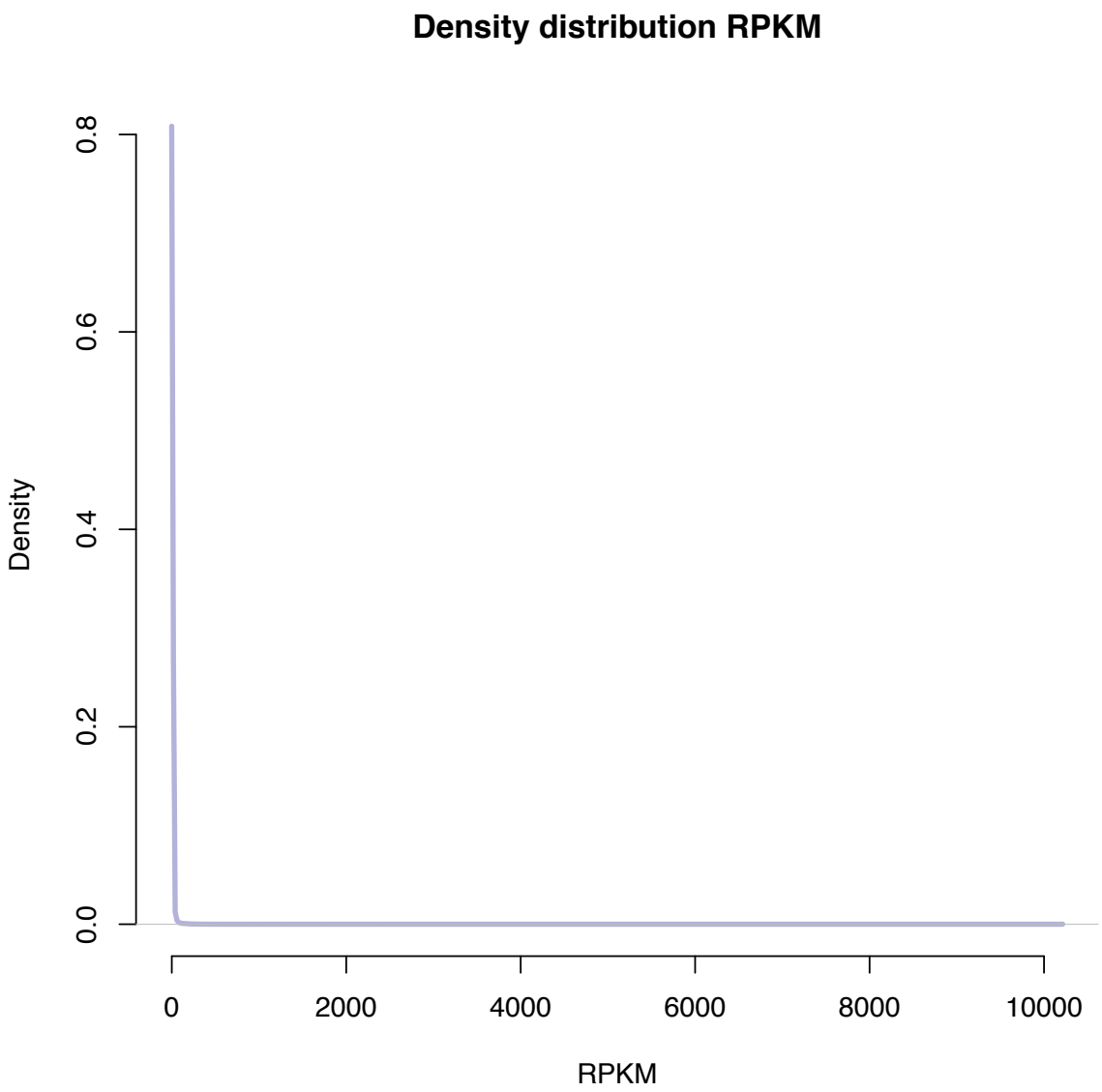
Todays topics

- Looking at ALL of your data

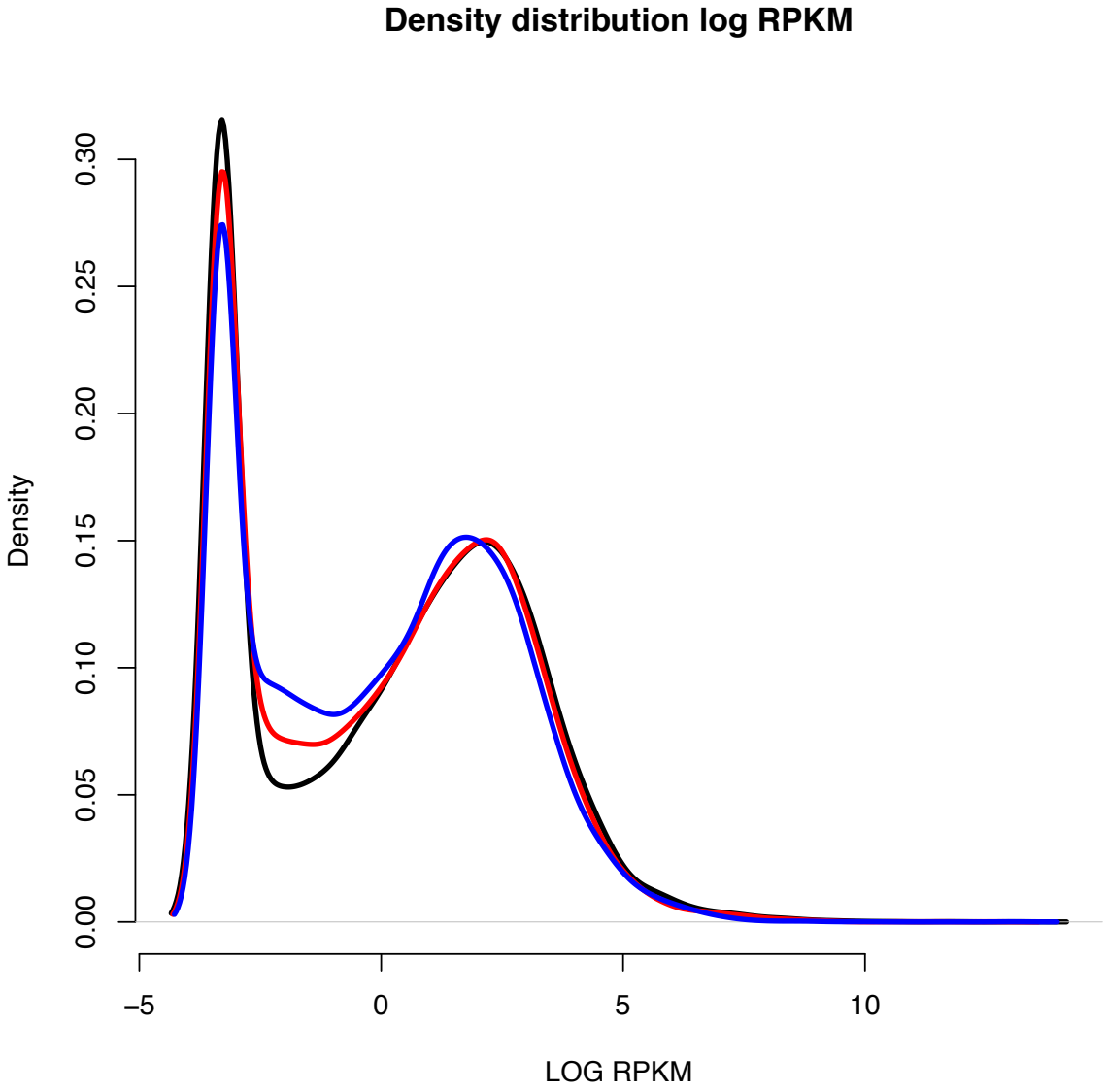
Comparing samples: Scatter plots



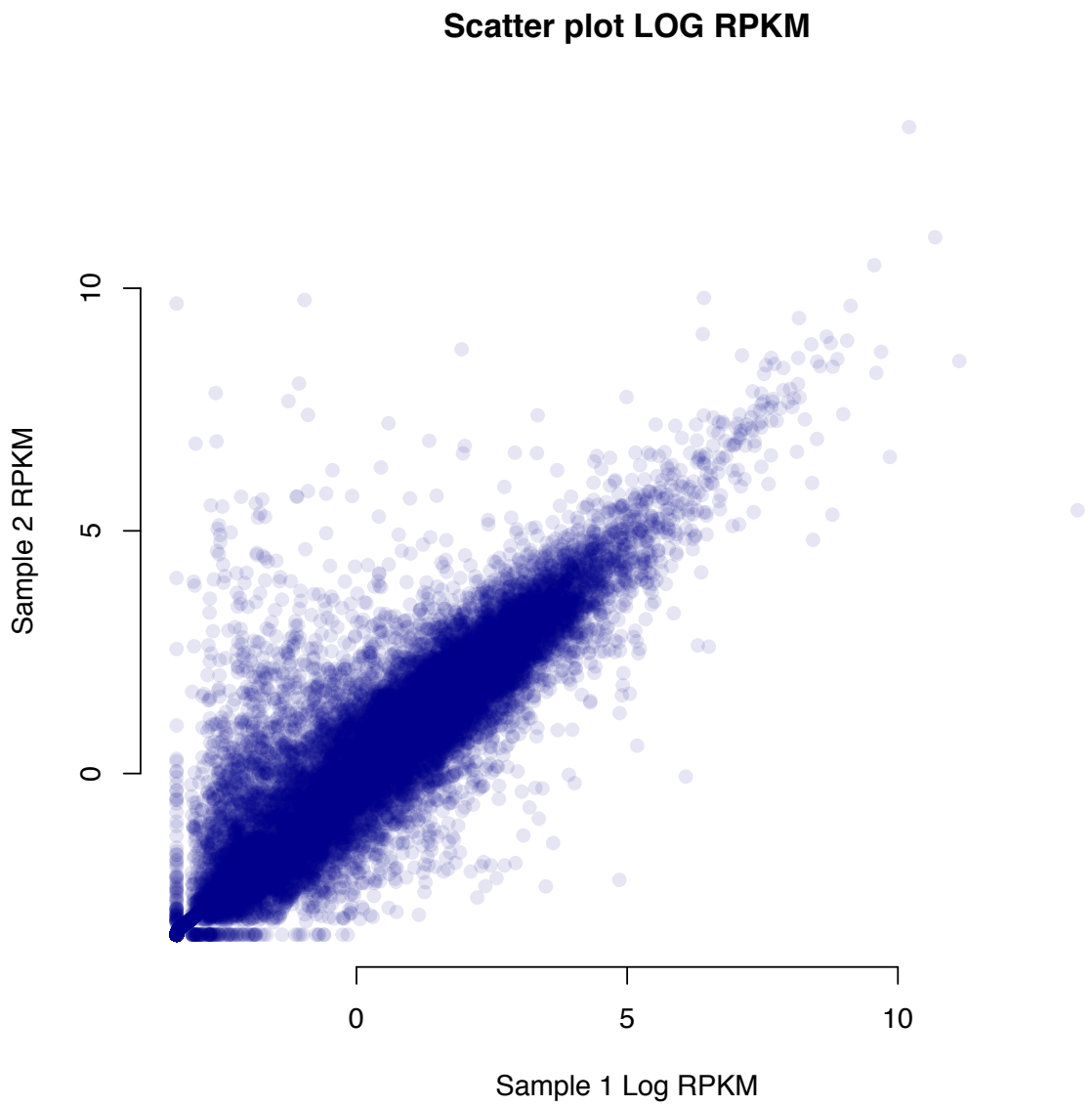
Raw counts/RPKMs are NOT Gaussian



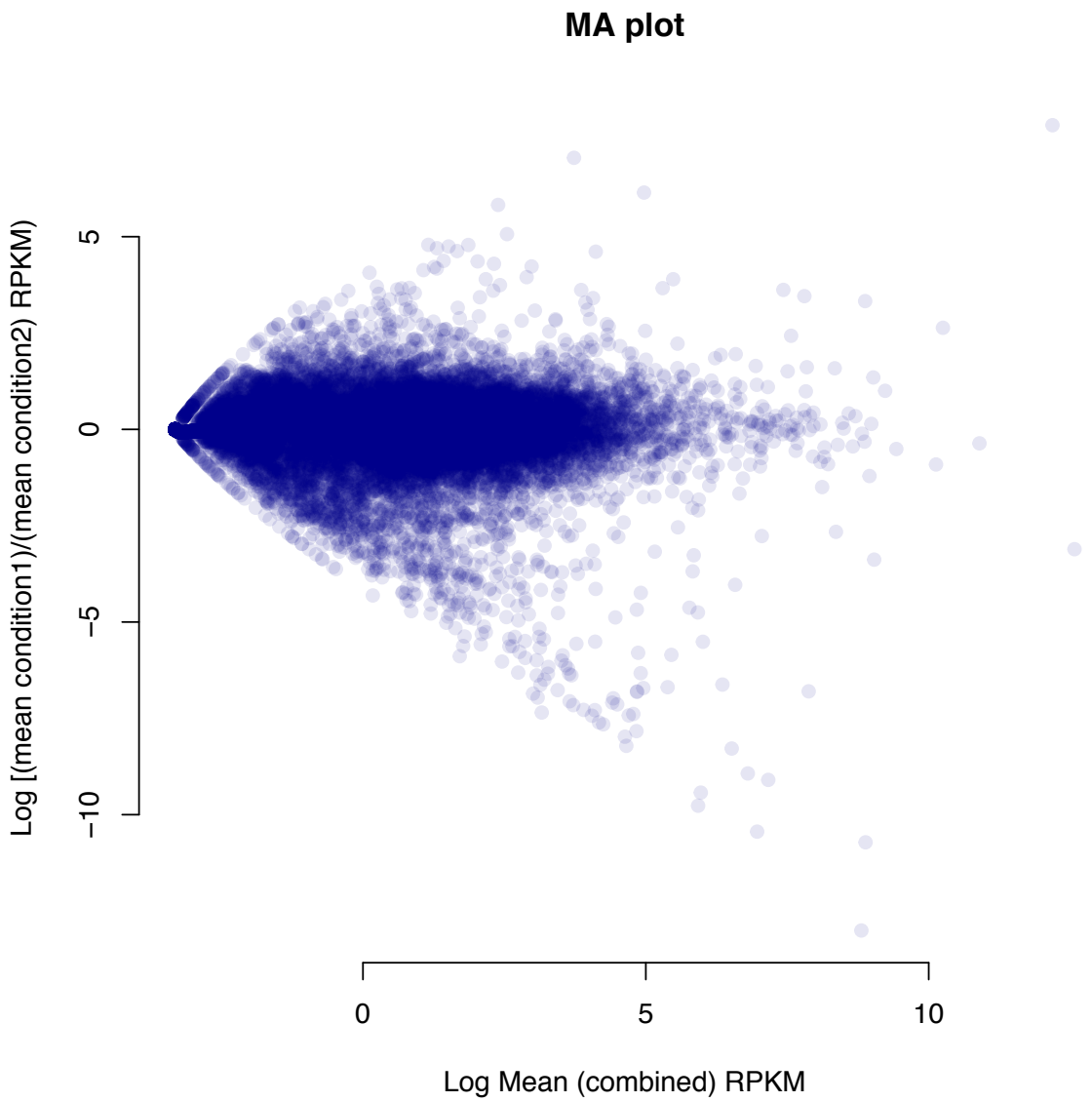
... they are more like Log-Gaussian



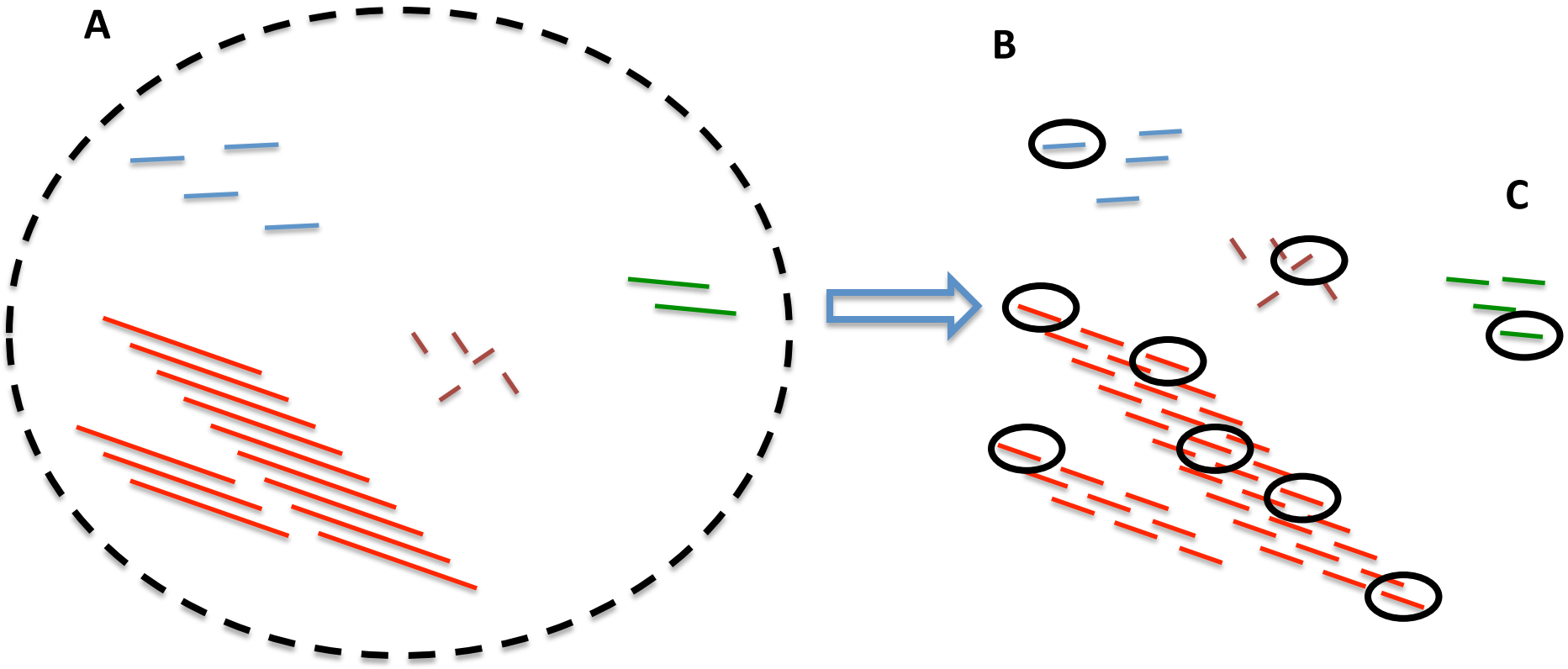
And log counts/RPKM can be scatter-plotted



Which can also be looked at as an “MA-Plot”



Differential analysis: Lets revisit quantification



RNA-Seq quantification: Infer # molecules in A from observed fragments in C

Quantification assumptions for differential expression

$$\mathcal{G} \propto f_g \times N$$

$\mathcal{G}$ = Read counts for gene g
f_g = the fraction of mRNA molecules for tene g
N = The total number of **aligned** reads

Note: We can at best estimate f_g

Modeling the RNA-Seq process

$$\mathcal{G} \propto f_g \times N$$

$$P(\mathcal{G}|N) = \binom{N}{\mathcal{G}} (f_g)^{\mathcal{G}} (1 - f_g)^{N-\mathcal{G}}$$

RNA-Seq counts should distribute "binomially"

Binomial? Why then we talk about Poisson

$$P(\mathcal{G}|N) = \binom{n}{\mathcal{G}} (f_g)^{\mathcal{G}} (1 - f_g)^{1-\mathcal{G}}$$

$$f_g \ll 1 \quad \text{and } \mathcal{G} \ll N \quad \text{and say } g = f_g \times M$$

M: # of mRNAs

g: # mRNAs for the gene

Binomial? Why then we talk about Poisson

$$P(\mathcal{G}|N) = \frac{N!}{\mathcal{G}!(N-\mathcal{G})!}(\frac{g}{M})^{\mathcal{G}}(1-\frac{g}{M})^{N-\mathcal{G}}$$

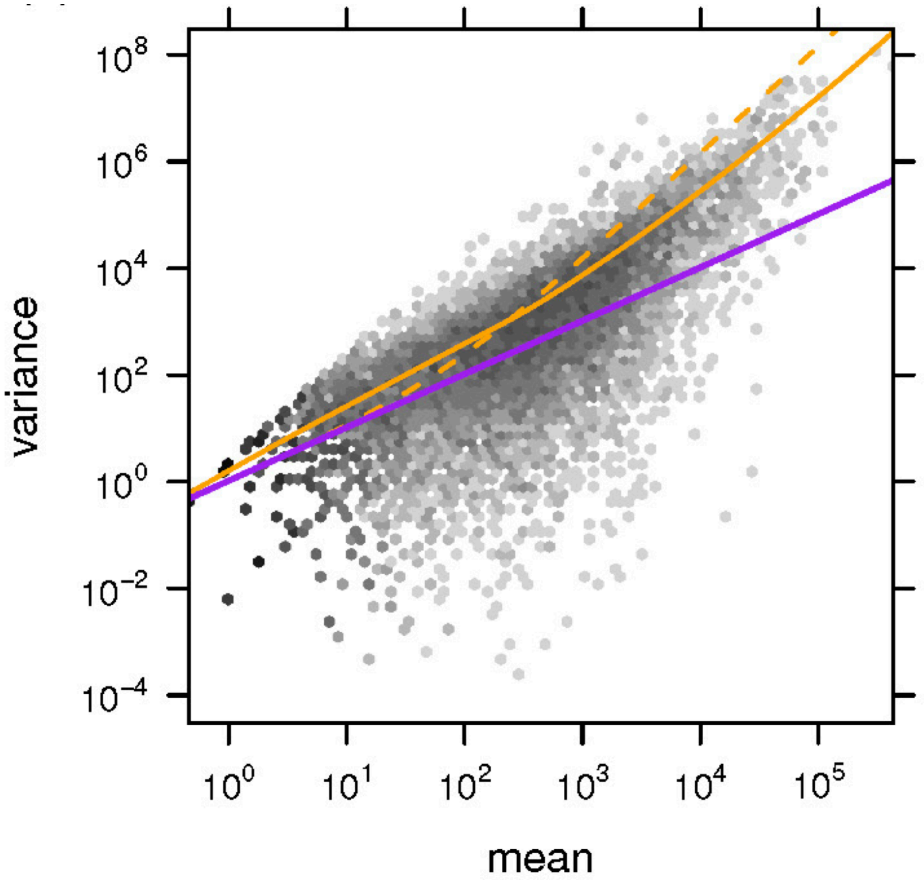
$$\approx \frac{N^{\mathcal{G}}}{\mathcal{G}!}\left(\frac{g}{M}\right)^{\mathcal{G}}(1+\frac{-g}{M})^{N}(1-\frac{g}{M})^{-\mathcal{G}}$$

$$= \frac{g^{\mathcal{G}}}{\mathcal{G}!}(1+\frac{-g}{M})^{N} \text{ and remember} (1+\frac{x}{n})^{n} \approx e^{x}$$

$$P(\mathcal{G}|N) \approx \frac{1}{\mathcal{G}!}g^{\mathcal{G}}e^{-g}$$

RNA-Seq counts can be approximated by a Poisson distribution

Poisson model does not work



Adapted from Anders, 2010

Solution: Use the negative binomial distribution

Hierarchical clustering – when are vector similar?

Gene	Cond1	Cond2	Cond3	Cond4
g_1	2.5	5	7.5	10
g_2	0.2	0.5	0.8	1.1
g_3	0.2	0.3	0.4	11
g_4	2.5	8	8	9

Clustering is about similarity:

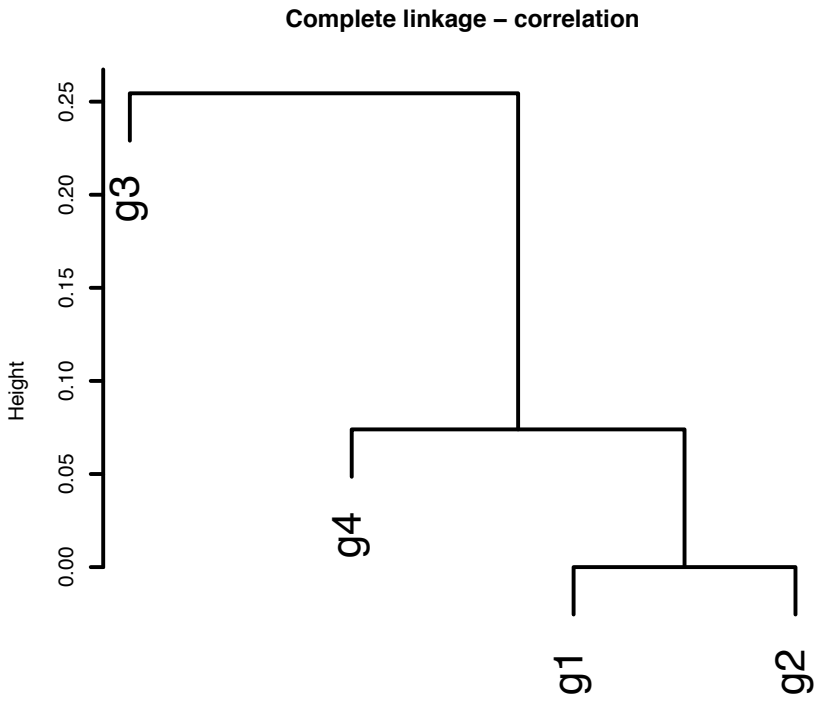
- Between two rows (specified by a distance function)
- Between two sets of rows (specified by the linkage method)

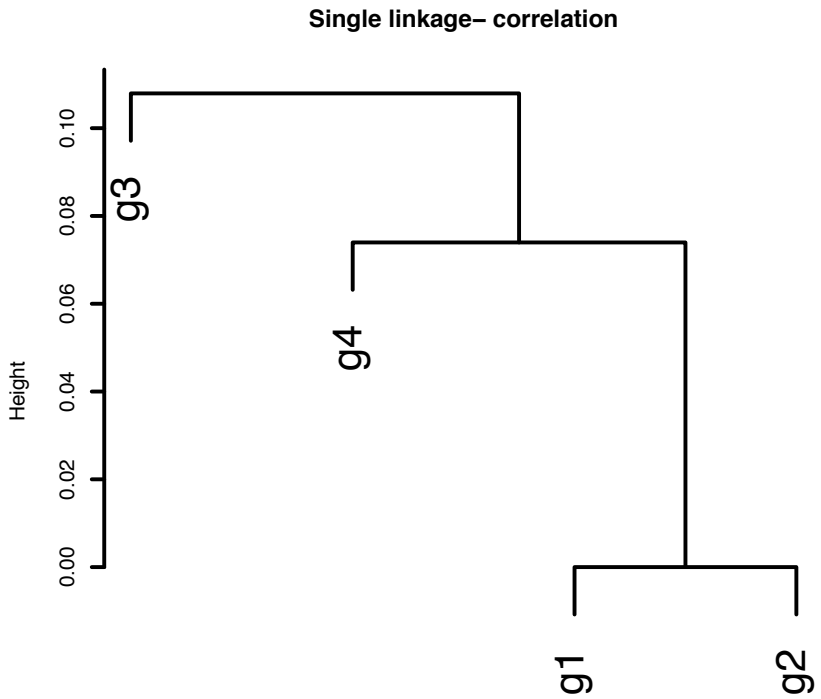
Common similarity approaches

- Distance between rows (or columns)
  - Correlation: `d(r,s) = (1- cor(r,s))/2`
  - Euclidean: $\texttt{d(r,s) = sqrt}(\Sigma_i(\texttt{r}_i - \texttt{s}_i)^2)$
- Linkage: Distance betwee two sets (`d(R,S)`)
  - Complete: $max\{d(r,s), s \in S, r \in R\}$
  - Average: $mean\{d(r,s), s \in S, r \in R\}$
  - Single: $min\{d(r,s), s \in S, r \in R\}$

Gene	Cond1	Cond2	Cond3	Cond4
g_1	2.5	5	7.5	10
g_2	0.2	0.5	0.8	1.1
g_3	0.2	0.3	0.4	11
g_4	2.5	8	8	9

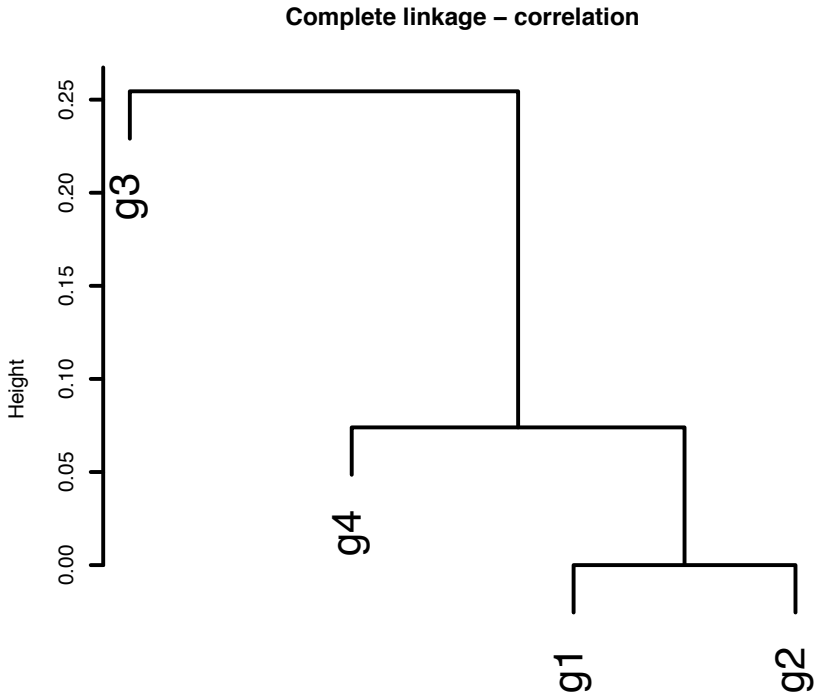
The effect of the linkage method

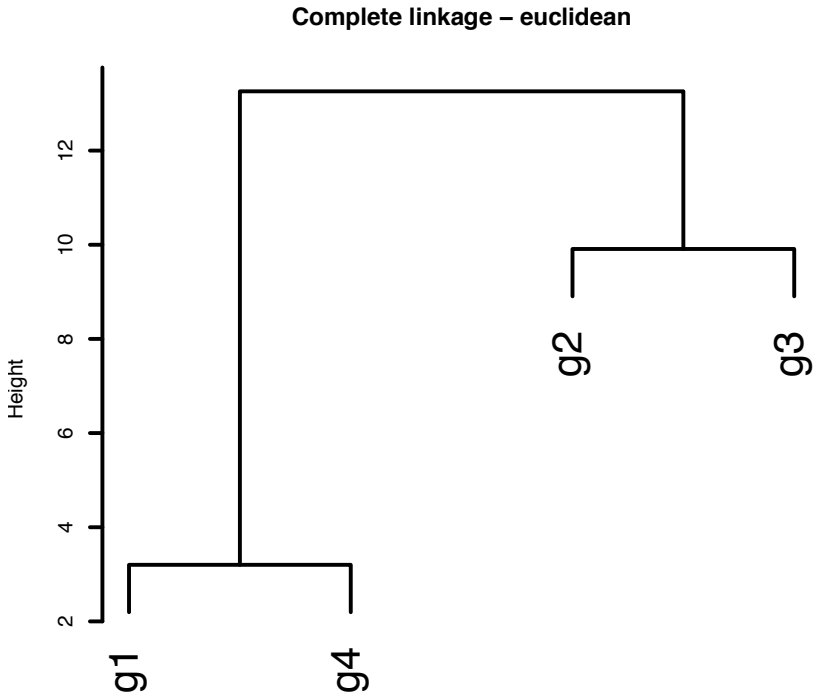




Gene	Cond1	Cond2	Cond3	Cond4
g_1	2.5	5	7.5	10
g_2	0.2	0.5	0.8	1.1
g_3	0.2	0.3	0.4	11
g_4	2.5	8	8	9

Effect of the distance!





Gene	Cond1	Cond2	Cond3	Cond4
g_1	2.5	5	7.5	10
g_2	0.2	0.5	0.8	1.1
g_3	0.2	0.3	0.4	11
g_4	2.5	8	8	9

Playing with clustering

```
#Define the toy matrix#
#######################
m =  rbind (c(2.5,5,7.5,10), c(0.2,0.5,0.8,1.1), c(0.2,0.3,0.4,11), c(2.5,8,8,9))

#Give column and row names#
###########################
rownames(m) = c("g1","g2","g3","g4");
colnames(m) = c("c1","c2","c3","c4");


#Compute the correlation distance matrix#
########################################
submat.dist = as.dist( (1 - cor(t(m)) )  /2 );


#Plot clustering with the three main methods#
#############################################
plot( hclust(submat.dist, method="complete",members=NULL), main="Complete linkeage -
correlation", sub="", xlab="", lwd=3);
plot( hclust(submat.dist, method="average",members=NULL),  main = "Average Linkeage -
correlation", sub="", xlab="", lwd=3);
plot( hclust(submat.dist, method="single",members=NULL),  main = "Single Linkeage- correlation",
sub="", xlab="", lwd=3);

#Plot clustering with the three main methods, using the euclidean distance#
############################################################################
plot( hclust(dist(m), method="complete",members=NULL), main="Complete linkeage - euclidean",
sub="", xlab="", lwd=3);
plot( hclust(dist(m), method="average",members=NULL),  main = "Average Linkeage -  euclidean",
sub="", xlab="", lwd=3);
plot( hclust(dist(m), method="single",members=NULL),  main = "Single Linkeage - euclidean",
sub="", xlab="", lwd=3);
```