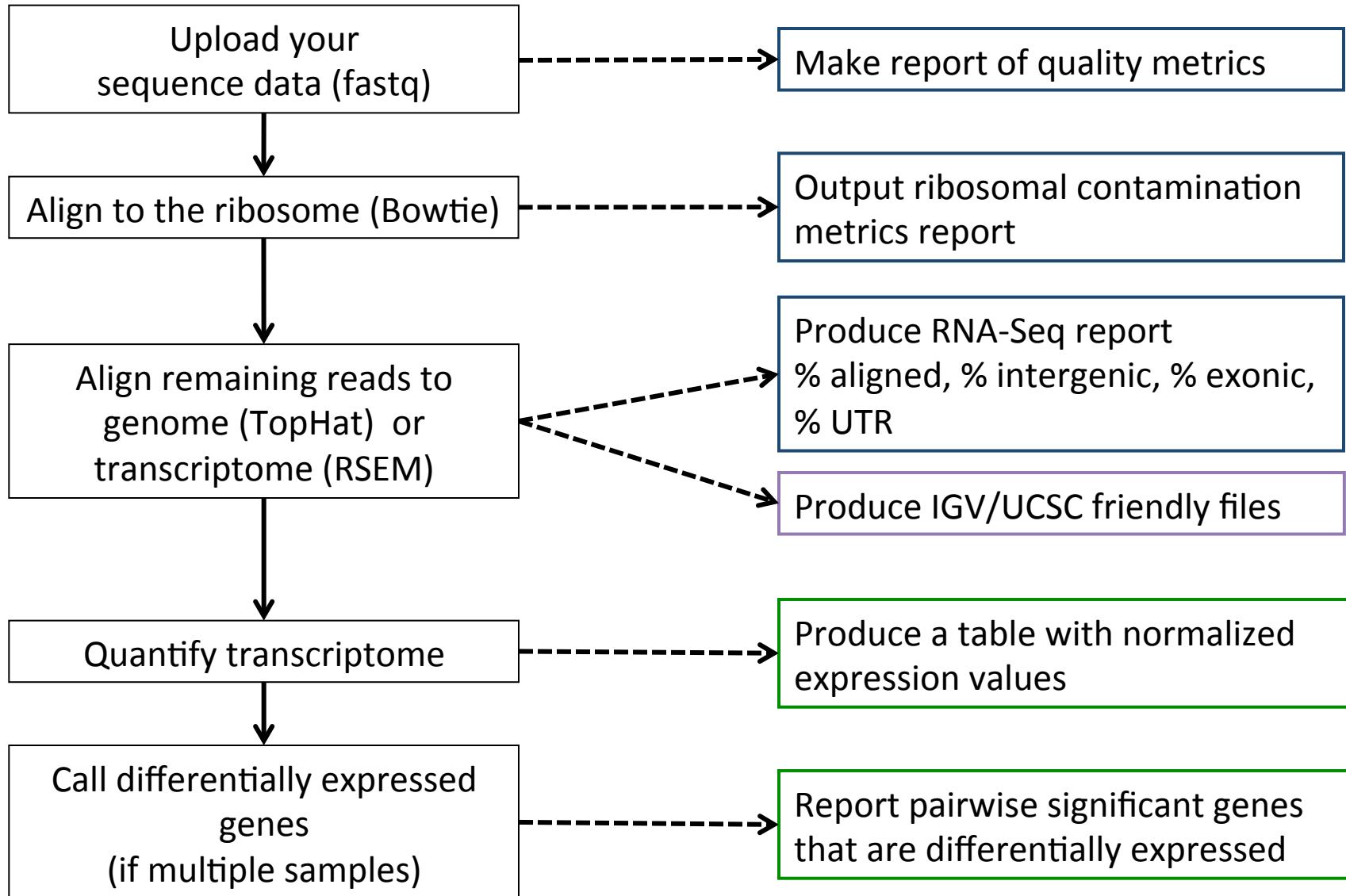




RNA-Seq primer

Our typical RNA quantification pipeline





Initial analysis

Summary I – Data types, file formats and utilities

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

Summary II – Data process

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

Visualization tricks & Tips

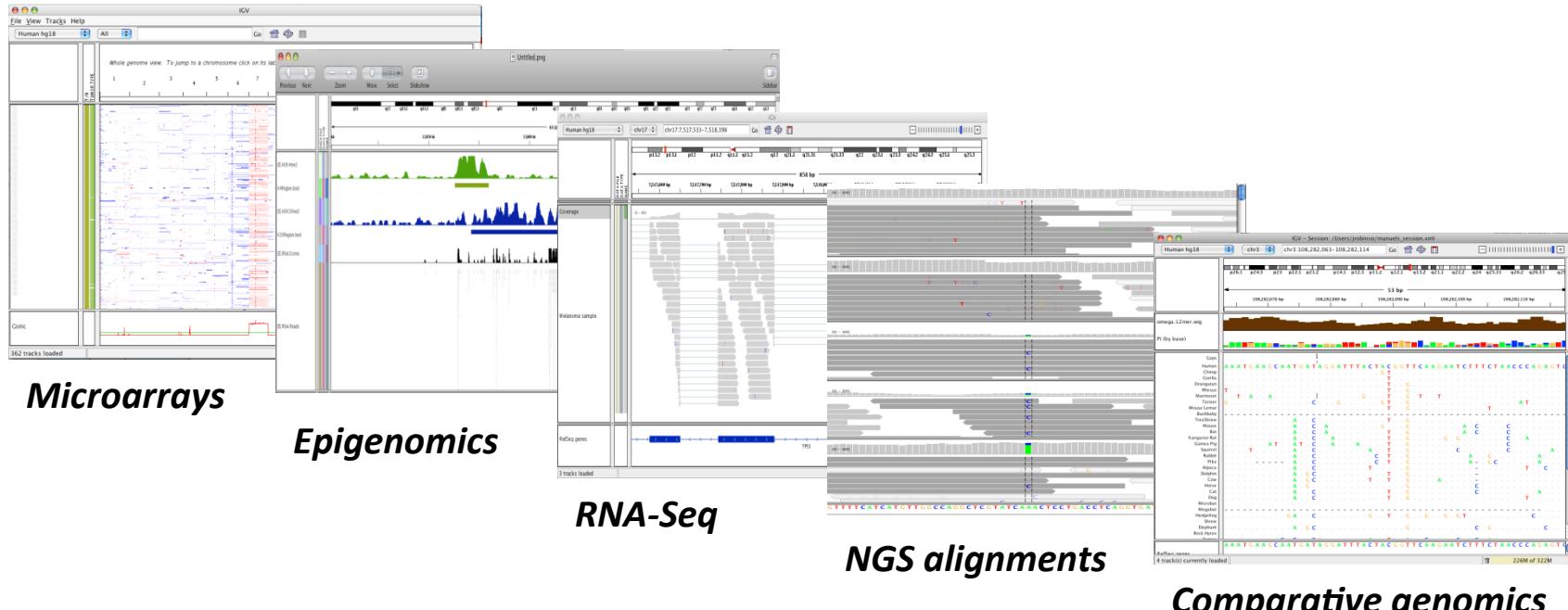
- Viewing normalized data
- Downsampling reads to avoid crashes
- Gene lists
- Sessions

IGV: Integrative Genomics Viewer

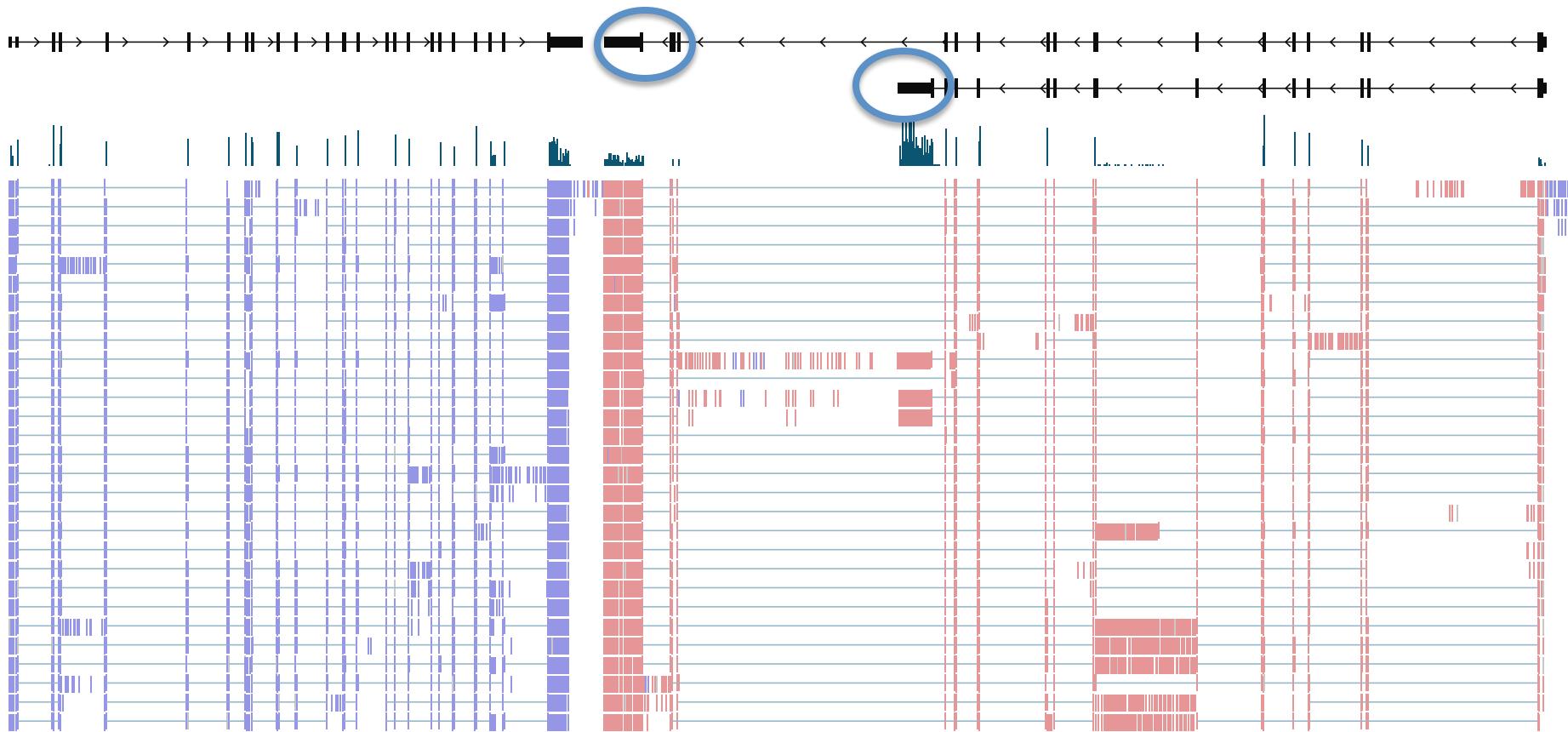


A desktop application

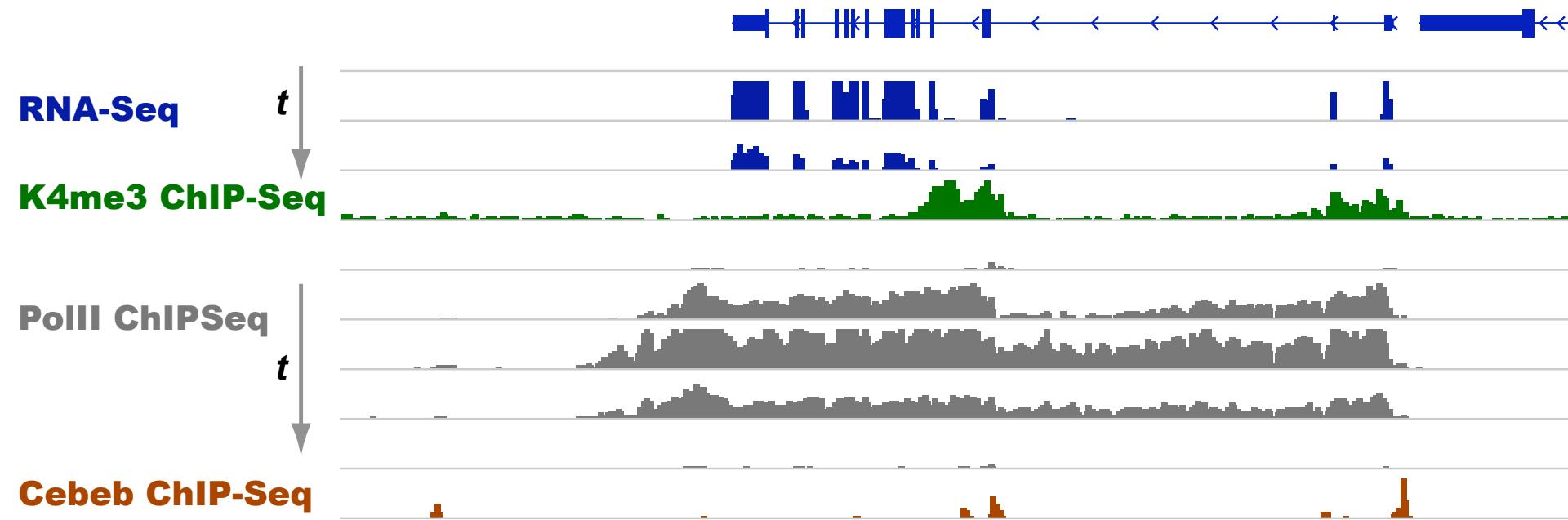
for the visualization and interactive exploration
of genomic data



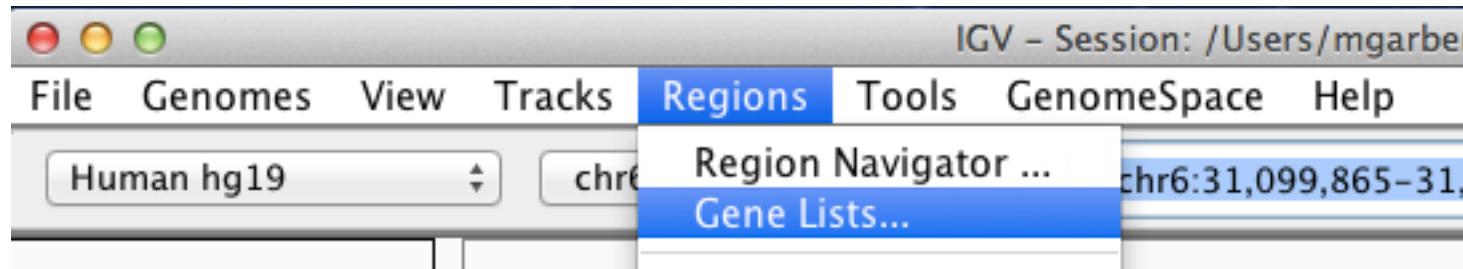
Visualizing read alignments with IGV — RNASeq



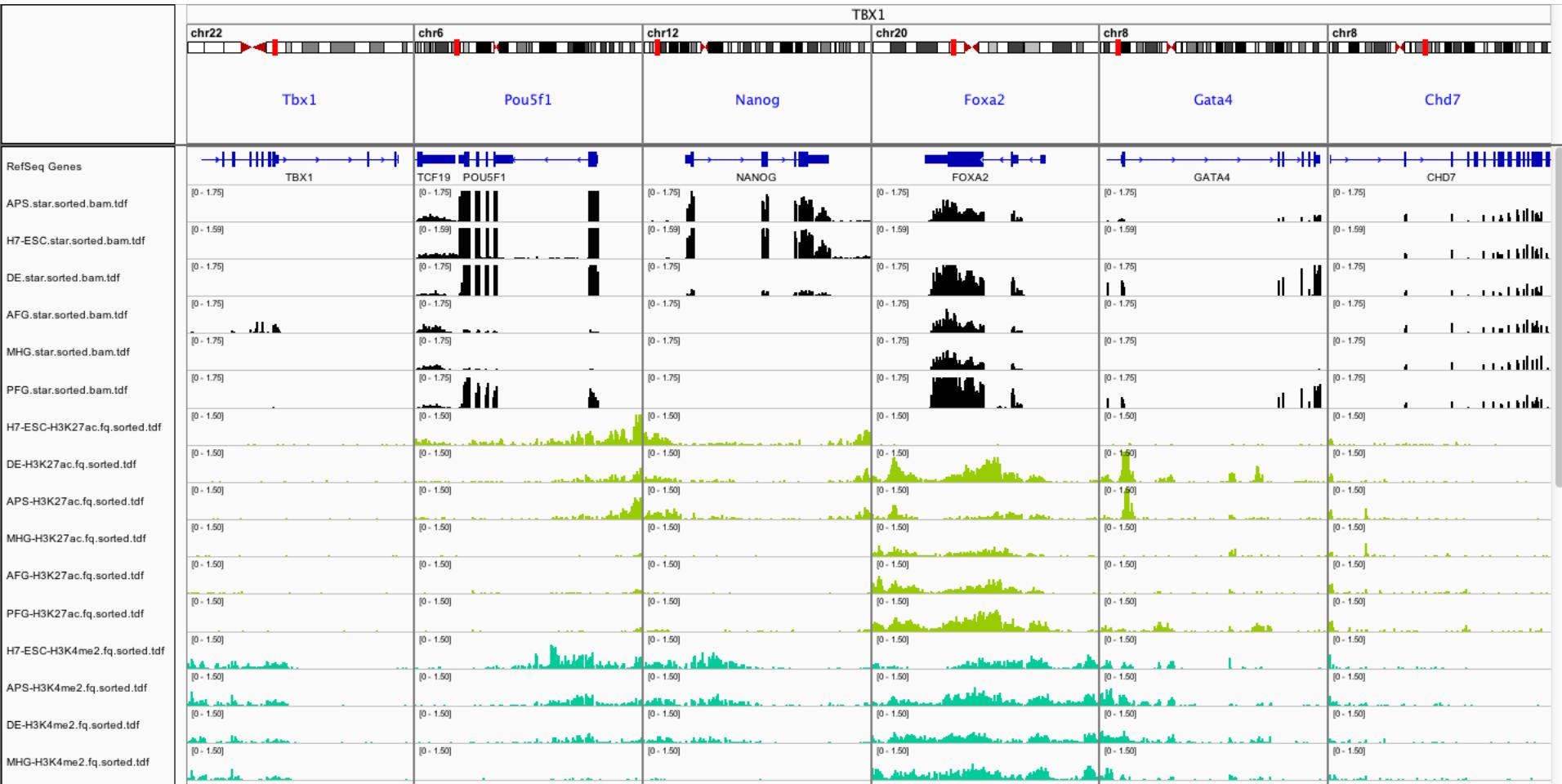
Visualizing read alignments with IGV — zooming out



Viewing several loci simultaneously: Gene lists



Viewing several loci simultaneously: Gene lists



Lets create a list using our small dataset

- Use the following genes
 - Fgf21
 - Bcat2
 - Rasip1
 - Naa60

Which are all within the dataset we selected.

Normalizing tracks

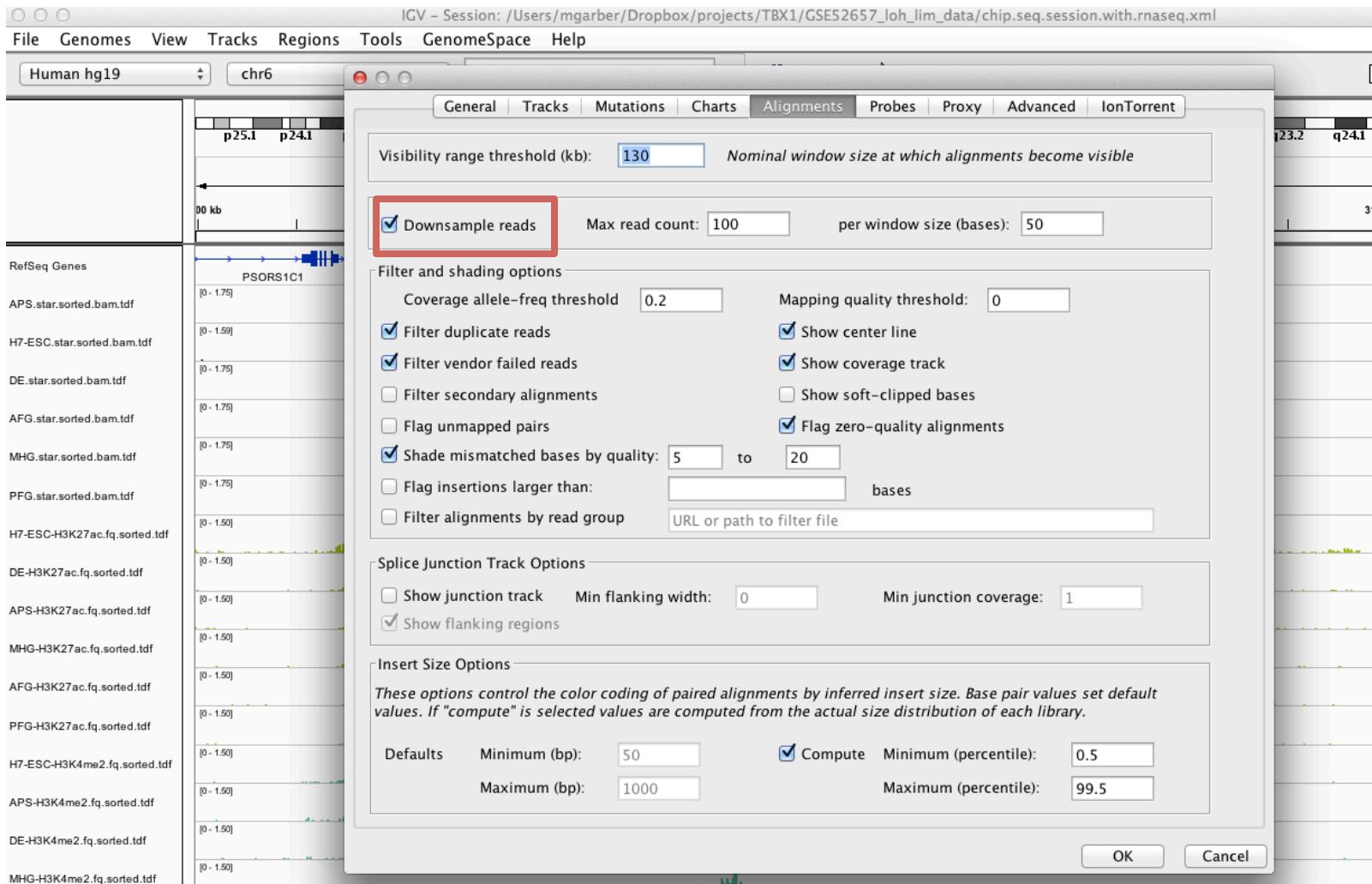
You can use simple read depth normalization for comparison of different tracks

The screenshot shows the IGV genome browser interface with a session titled "IGV - Session: /Users/mgarber/Dropbox/projects/TBX1/GSE52657_loh_lim_data/chip.seq.session.with.rnaseq.xml". The left panel lists genomic tracks: "Human hg19" (selected), "chr6", "RefSeq Genes", and numerous BAM and TDF files. The main panel displays tracks for genes like PSORS1C1 and H3K27ac across chromosomes 25 and 24. A context menu is open over one of the coverage tracks, specifically the "PSORS1C1" track. The menu has tabs for General, Tracks, Mutations, Charts, Alignments, Probes, Proxy, Advanced, and IonTorrent. The "Tracks" tab is selected. Inside the "Tracks" tab, there are two settings for track height: "Default Track Height, Charts (Pixels)" set to 40 and "Default Track Height, Other (Pixels)" set to 15. Below these is a field for "Track Name Attribute" with a placeholder and a note about labeling tracks from sample information files. At the bottom of the menu, three checkboxes are present: "Expand Feature Tracks" (unchecked), "Show Expand Icon" (unchecked), and "Normalize Coverage Data" (checked). A red box highlights the "Normalize Coverage Data" checkbox. A tooltip below it explains: "Applies to coverage tracks computed with igvtools (.tdf files). If selected coverage values are scaled by (1,000,000 / totalCount), where totalCount is the total number of features or alignments." The bottom right of the menu contains "OK" and "Cancel" buttons.

Configure your alignment display

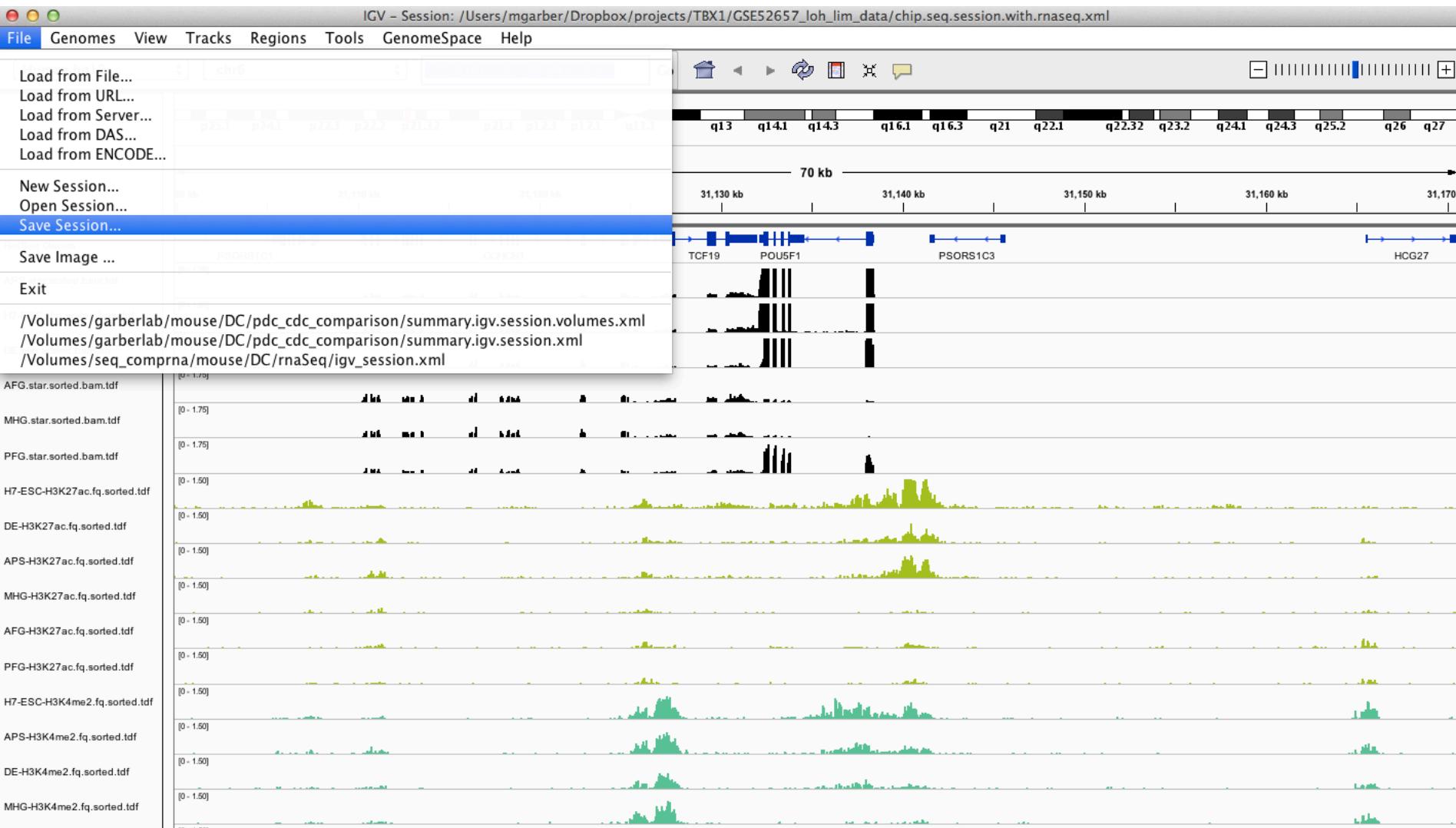
Downsampling reads is critical when loading the full alignments.

When you are loading reads, downsampling ensures that regions with high coverage result in IGV running out of memory.



Saving sessions

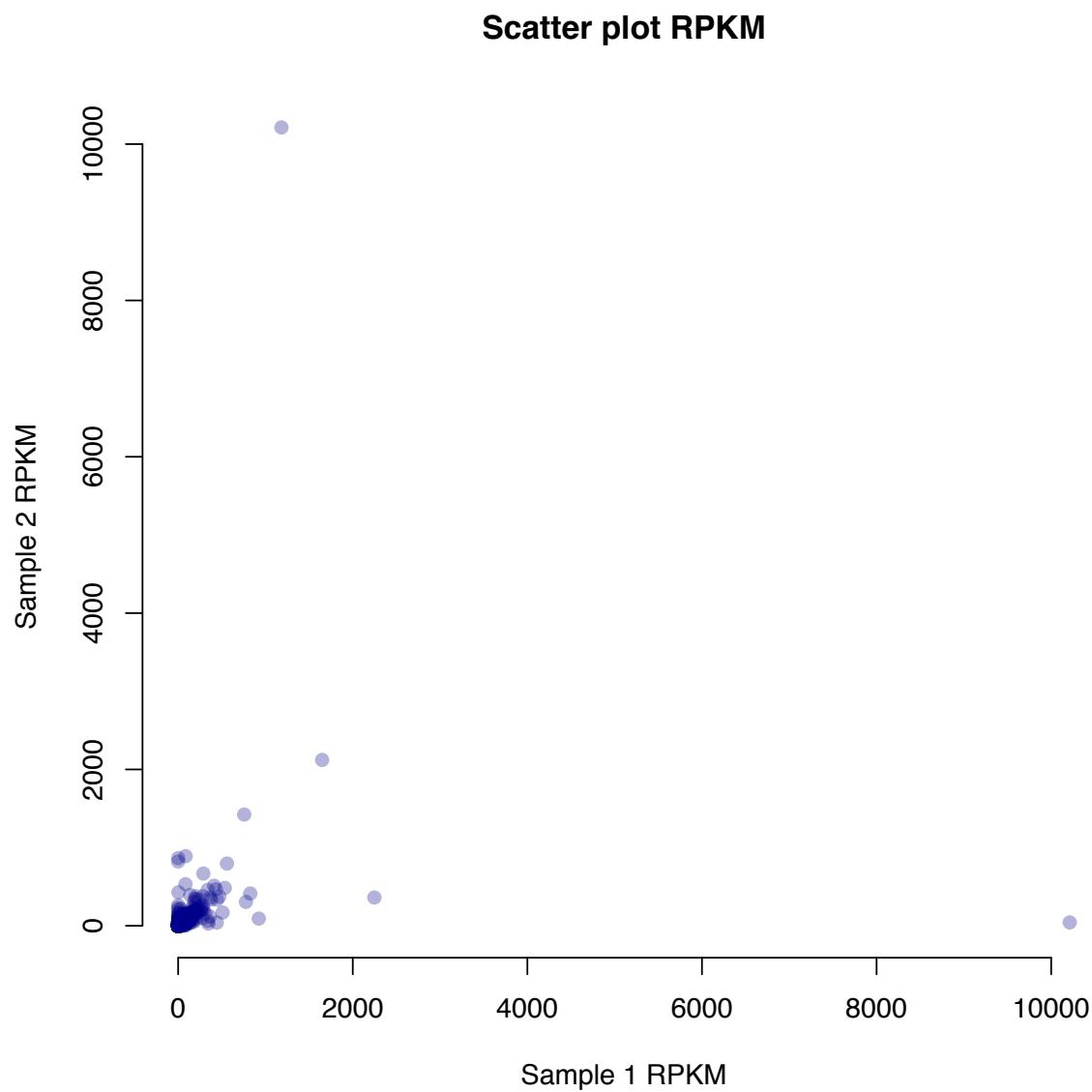
Sessions allows you to store a set of desired tracks along with any setting you want



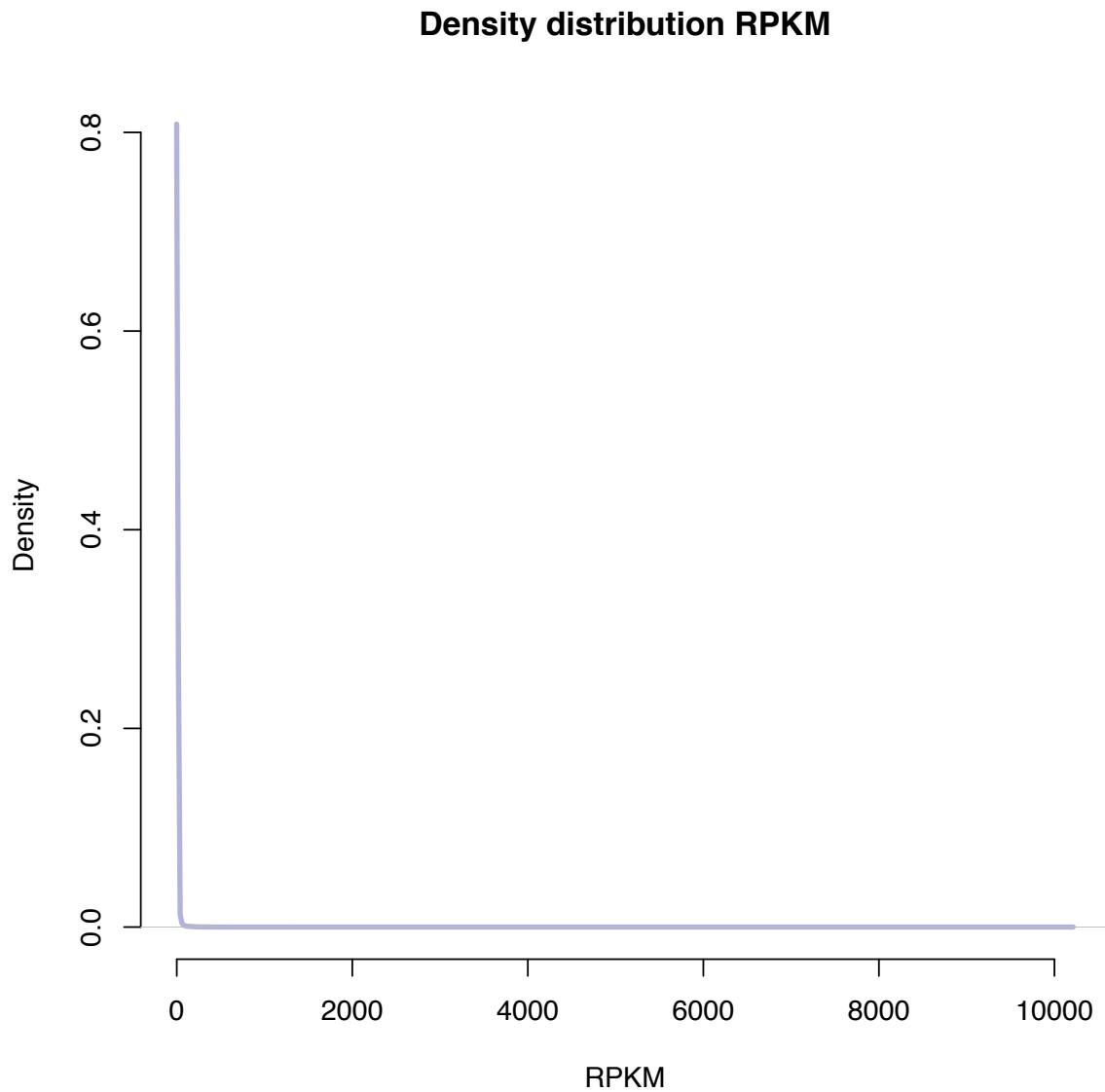
Todays topics

- Looking at ALL of your data

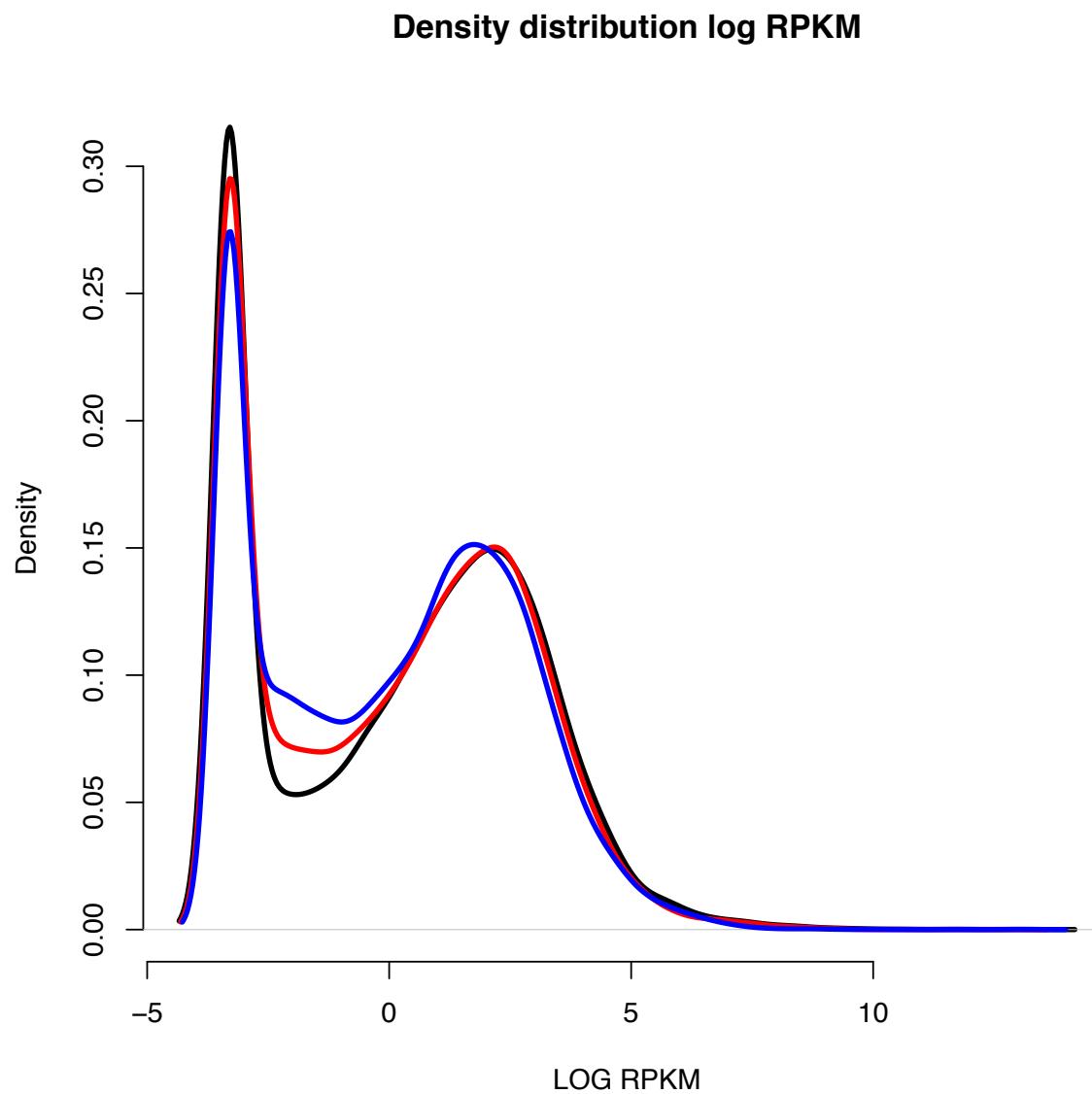
Comparing samples: Scatter plots



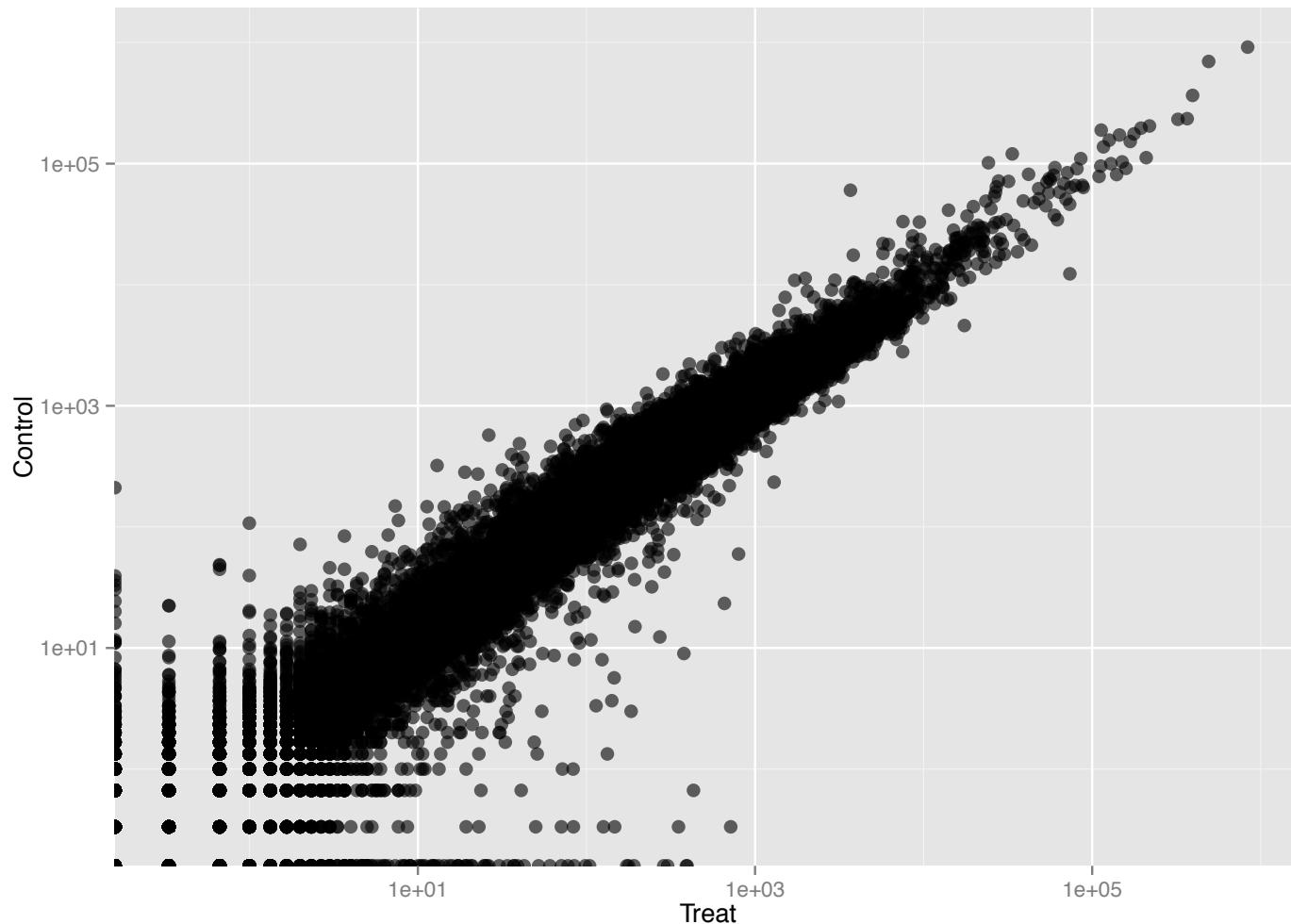
Raw counts/RPKMs are NOT Gaussian



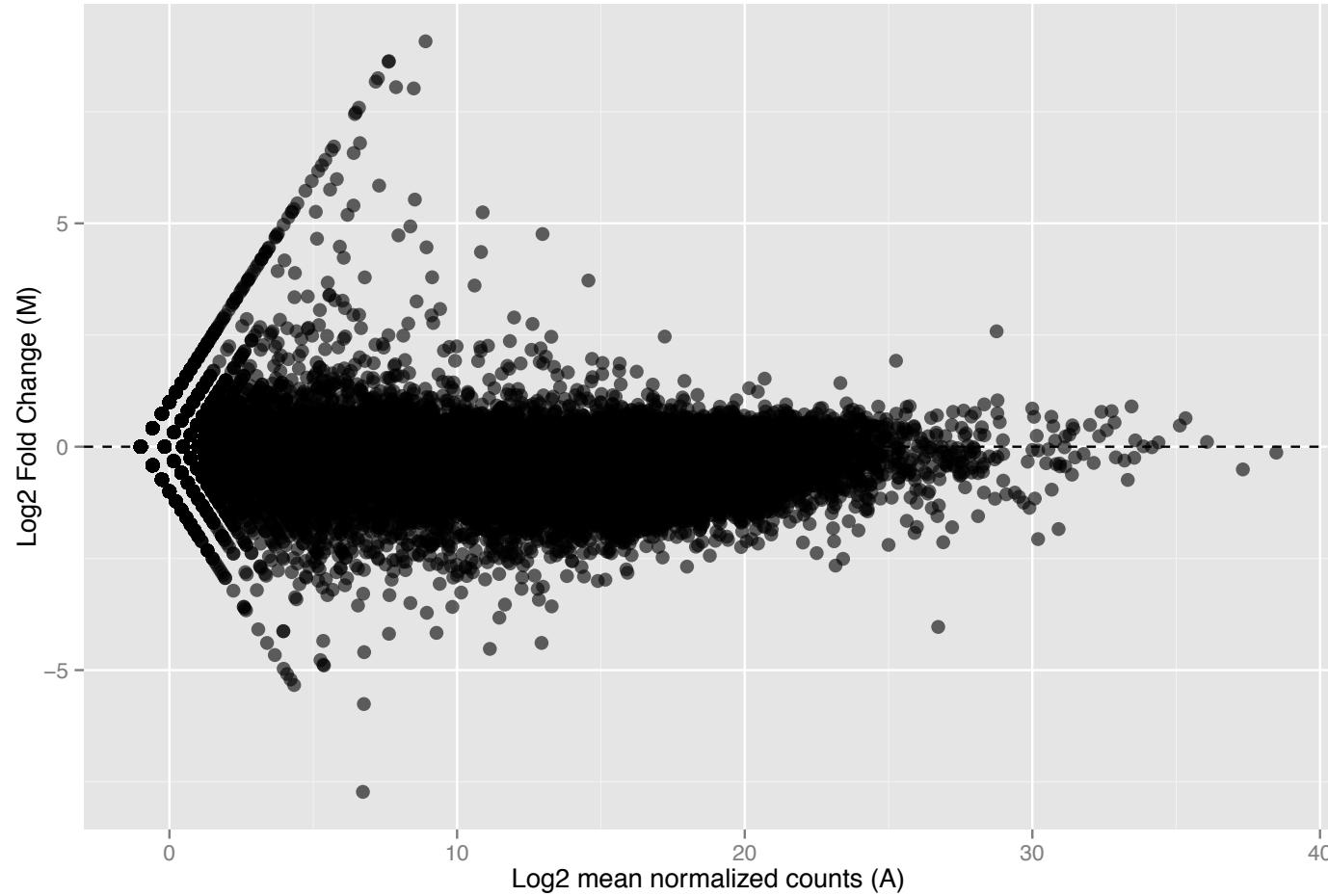
...they are more like Log-Gaussian



And log counts/RPKM can be scatter-plotted



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



Hierarchical clustering – vector similarity?

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)

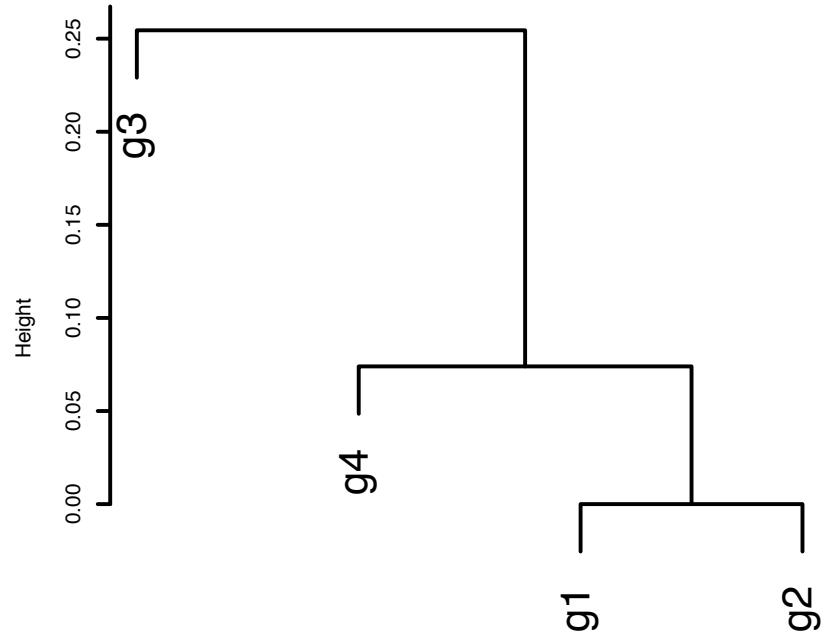
Common similarity approaches

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

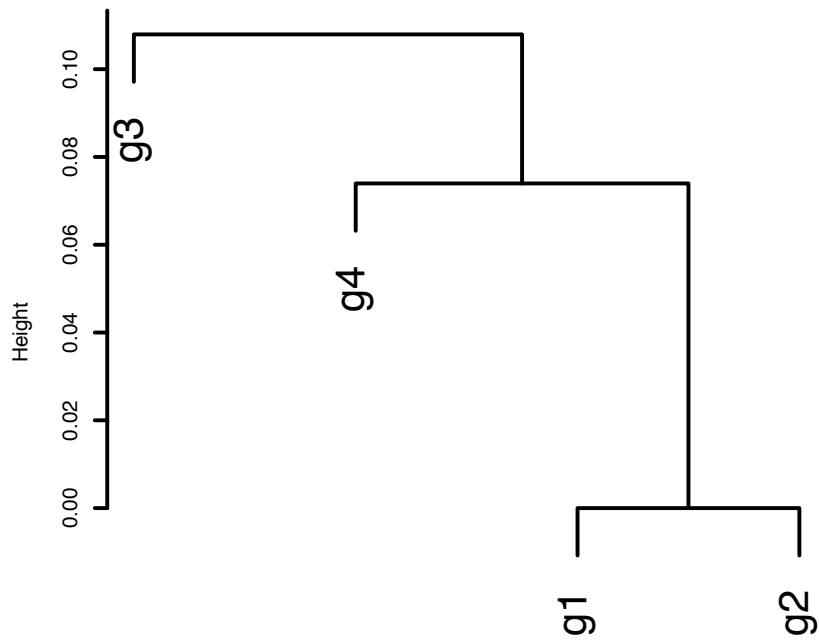
Gene	Cond1	Cond2	Cond3	Cond4
g_1	2.5	5	7.5	10
g_2	0.2	0.5	0.8	1.1
g_3	0.2	0.3	0.4	11
g_4	2.5	8	8	9

The effect of the linkage method

Complete linkage – correlation



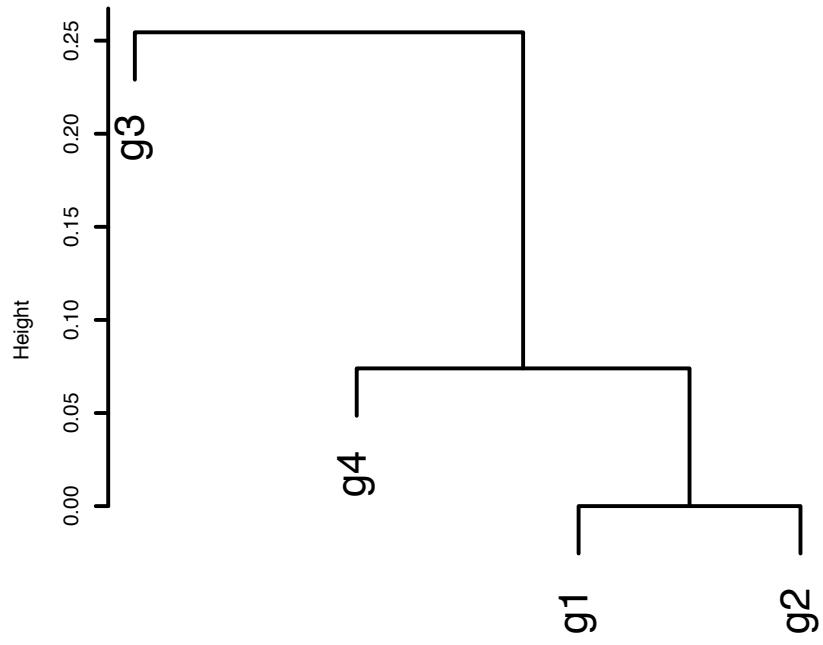
Single linkage- correlation



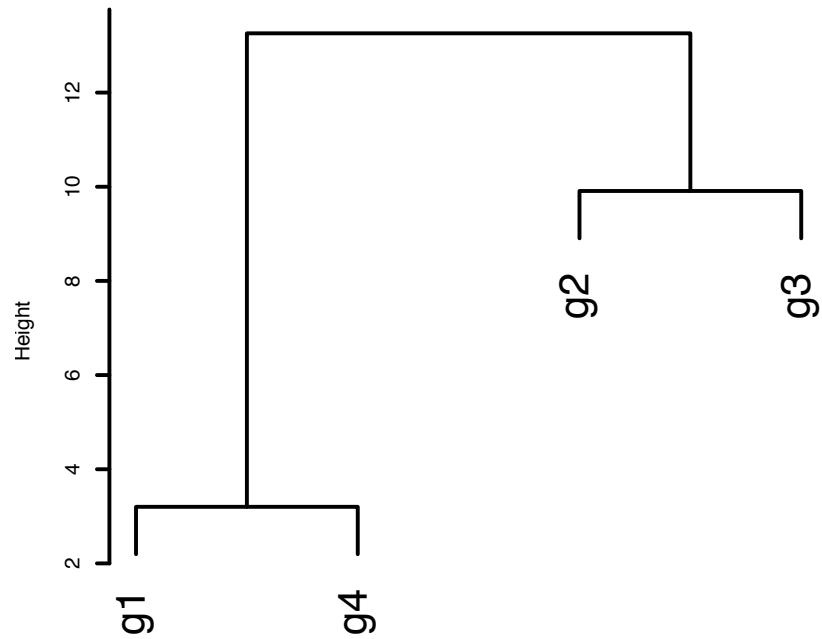
Gene	Cond1	Cond2	Cond3	Cond4
g ₁	2.5	5	7.5	10
g ₂	0.2	0.5	0.8	1.1
g ₃	0.2	0.3	0.4	11
g ₄	2.5	8	8	9

Effect of the distance!

Complete linkage – correlation



Complete linkage – euclidean



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);
```