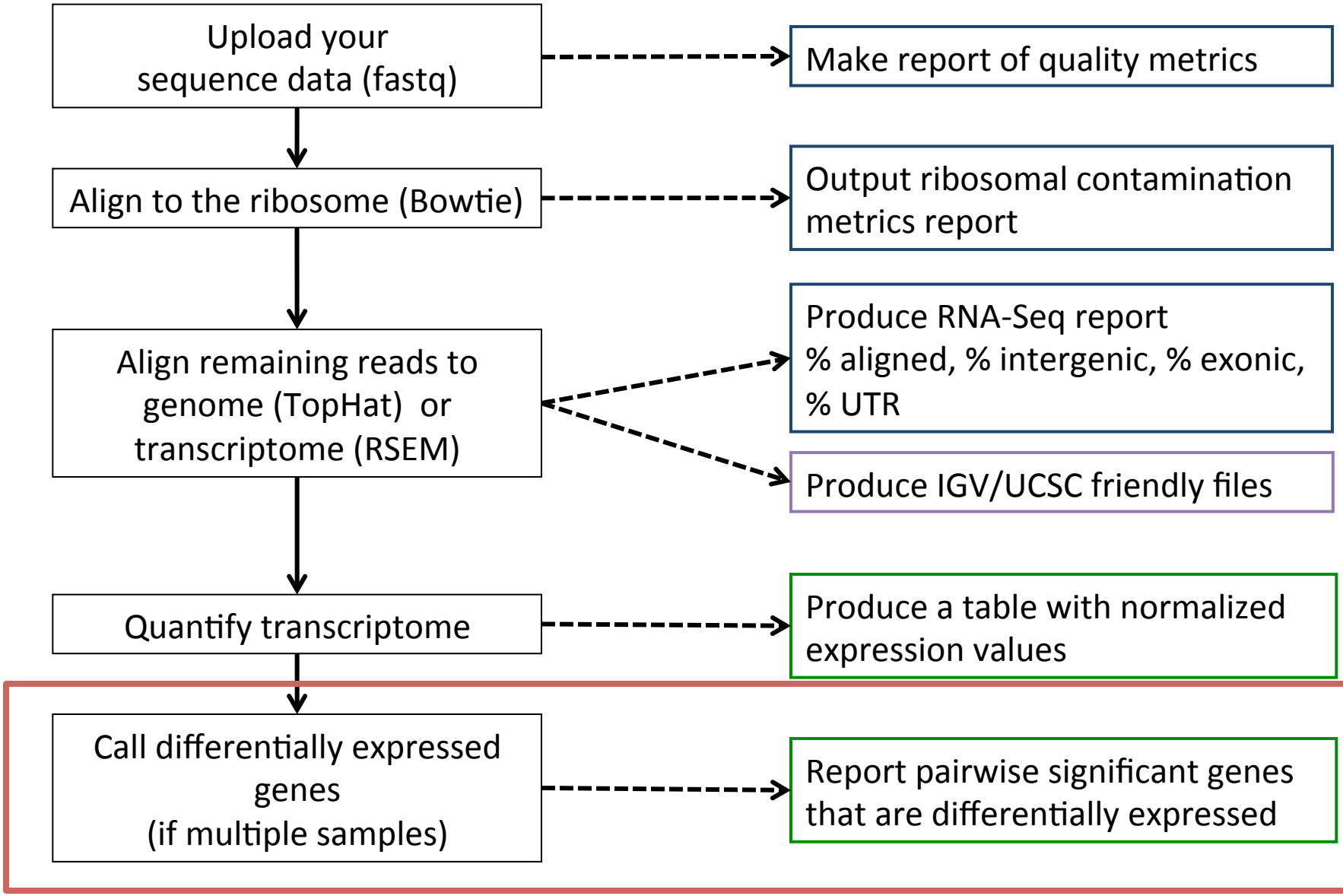
University of Massachusetts Medical School

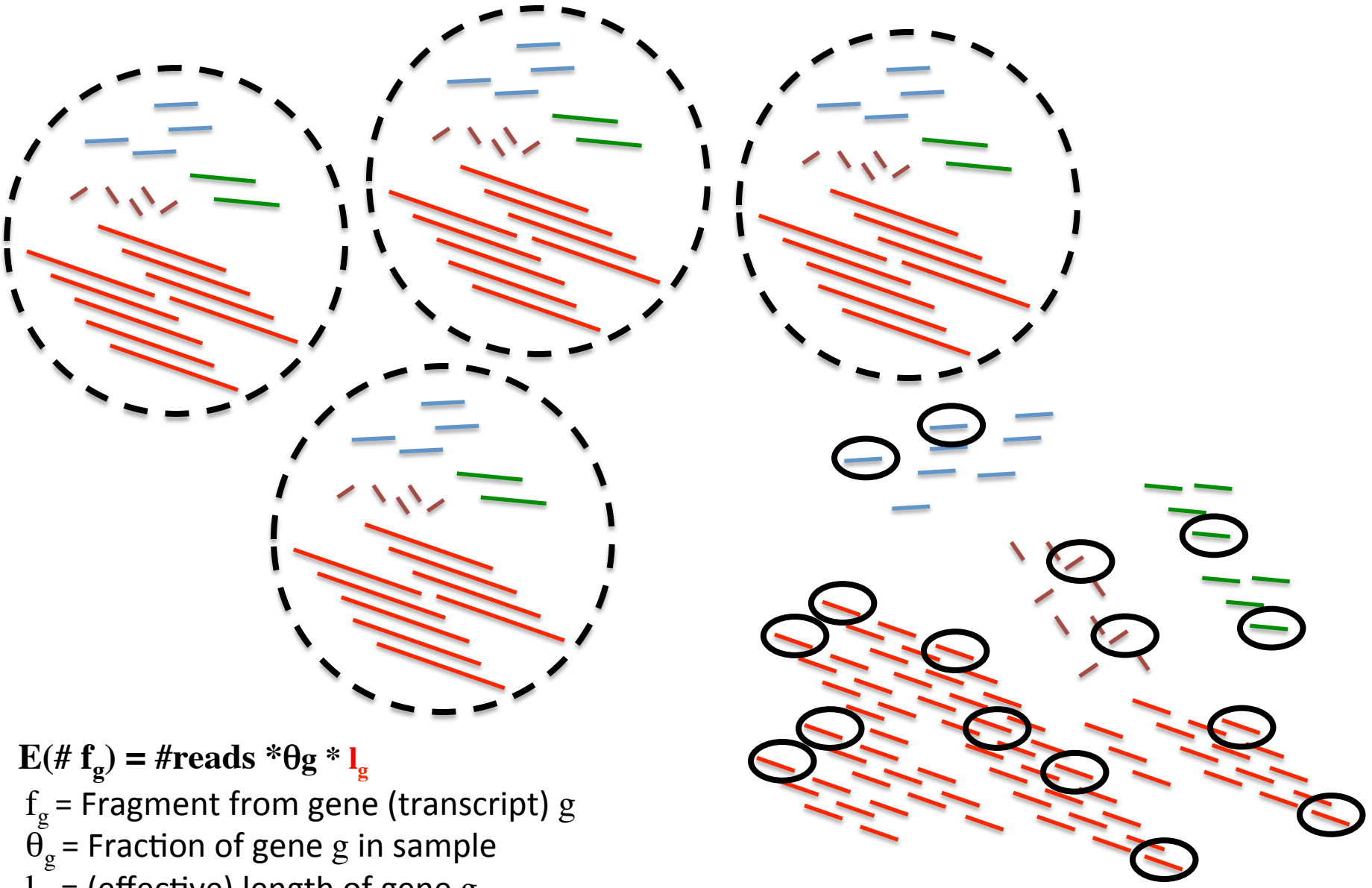
# Week 6
# Processing short reads

Data analysis

# Our typical RNA quantification pipeline



# A library satisfying assumptions 1 & 2



$\mathbf{E(\#\ f_g) = \#reads * \theta g * l_g}$

$f_g$ = Fragment from gene (transcript) g

$\theta_g$ = Fraction of gene g in sample

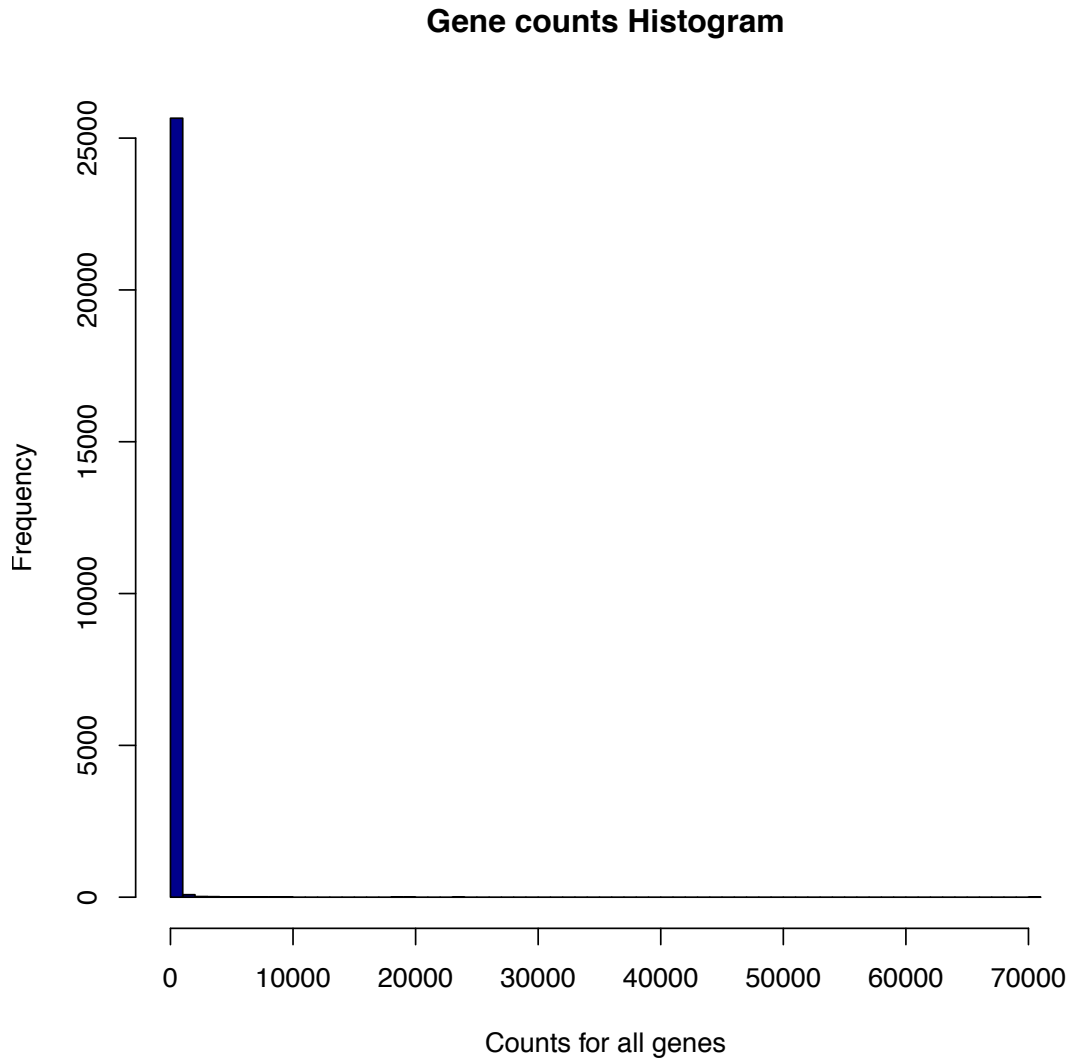
$l_g$ = (effective) length of gene g

# Summary

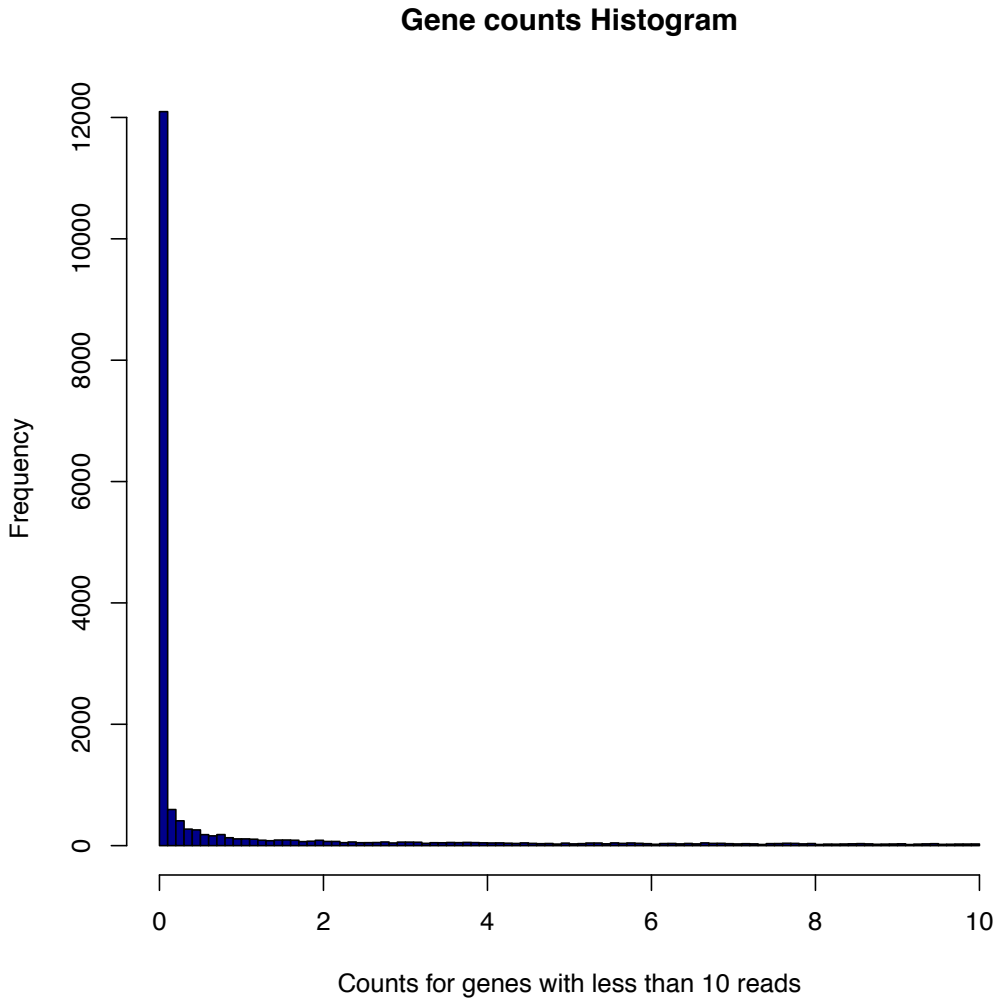
- RNA-Seq measures relative abundance
- Estimating relative abundance requires statistical modeling
- Protocols can simplify analysis (end-sequencing)
- Normalization strategy depends on goal:
  - FPKM, TPM great for comparing genes within samples
  - Normalized counts for differential expression
  - Extreme cases benefit from spike-in RNAs

# Comparing samples
## Scatter plots

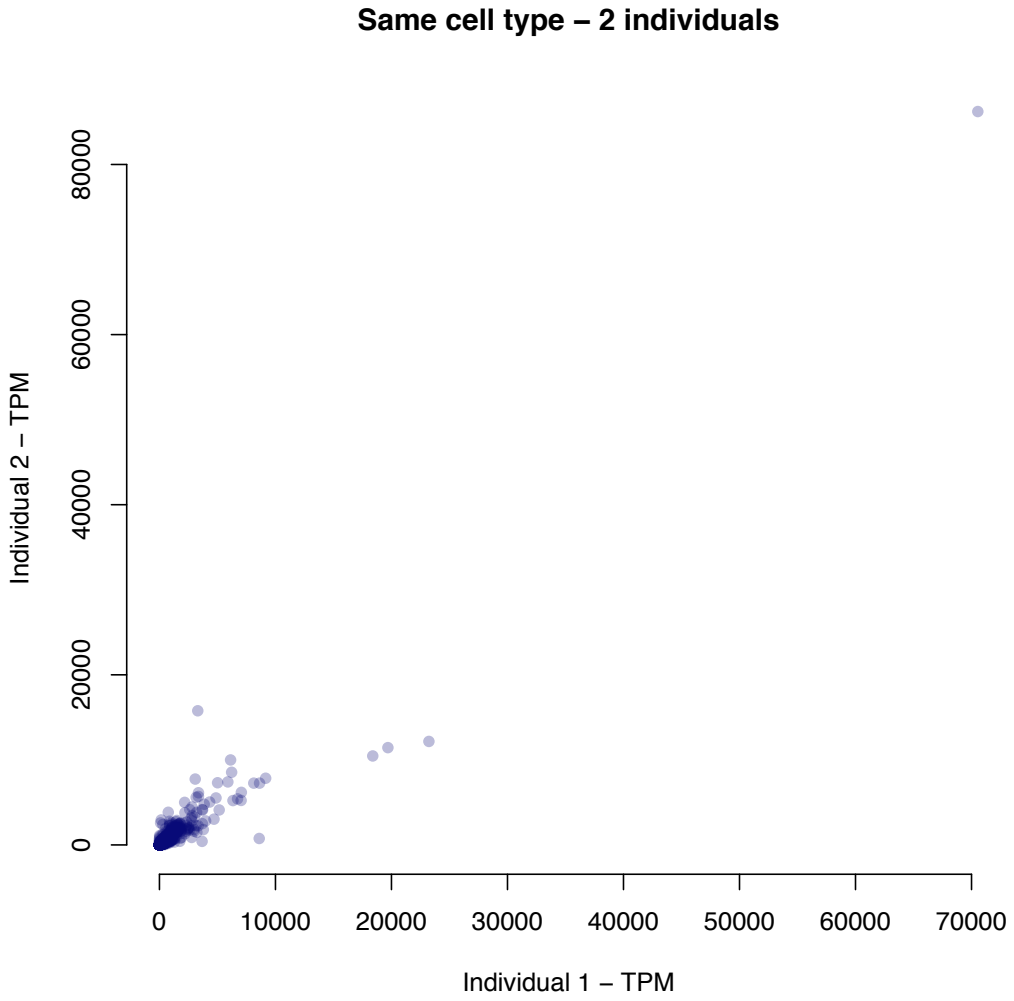
# Plotting data – the count distribution



# Plotting data – the count distribution



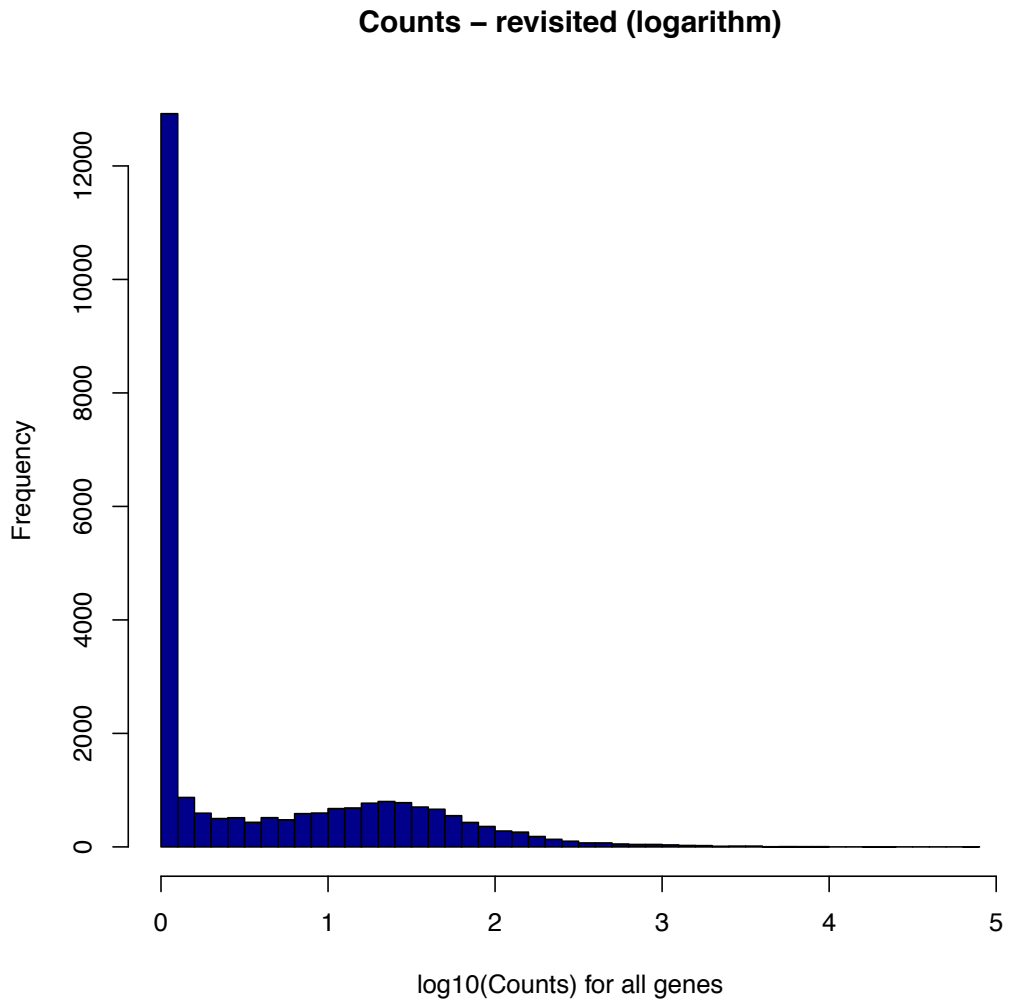
# Which makes an uninformative comparison



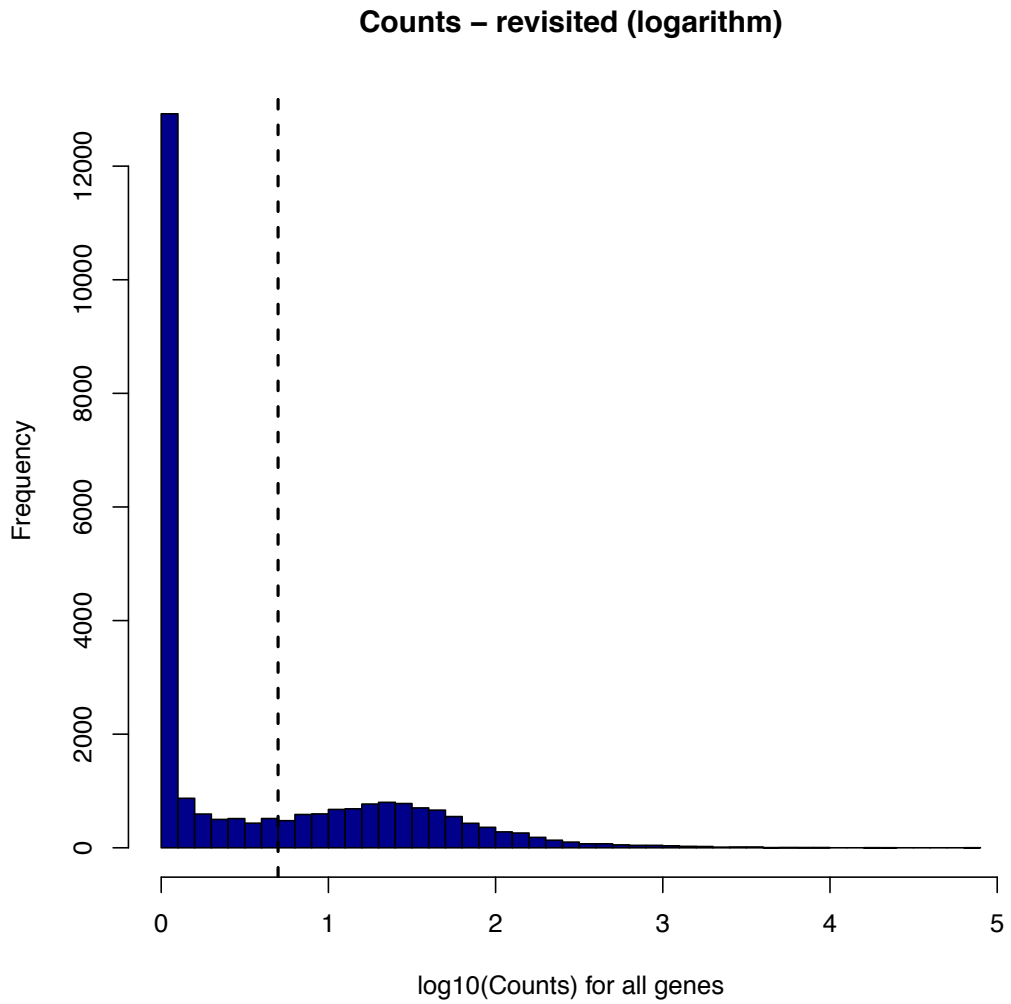

$R^2$=0.95

Sample 1 TPM

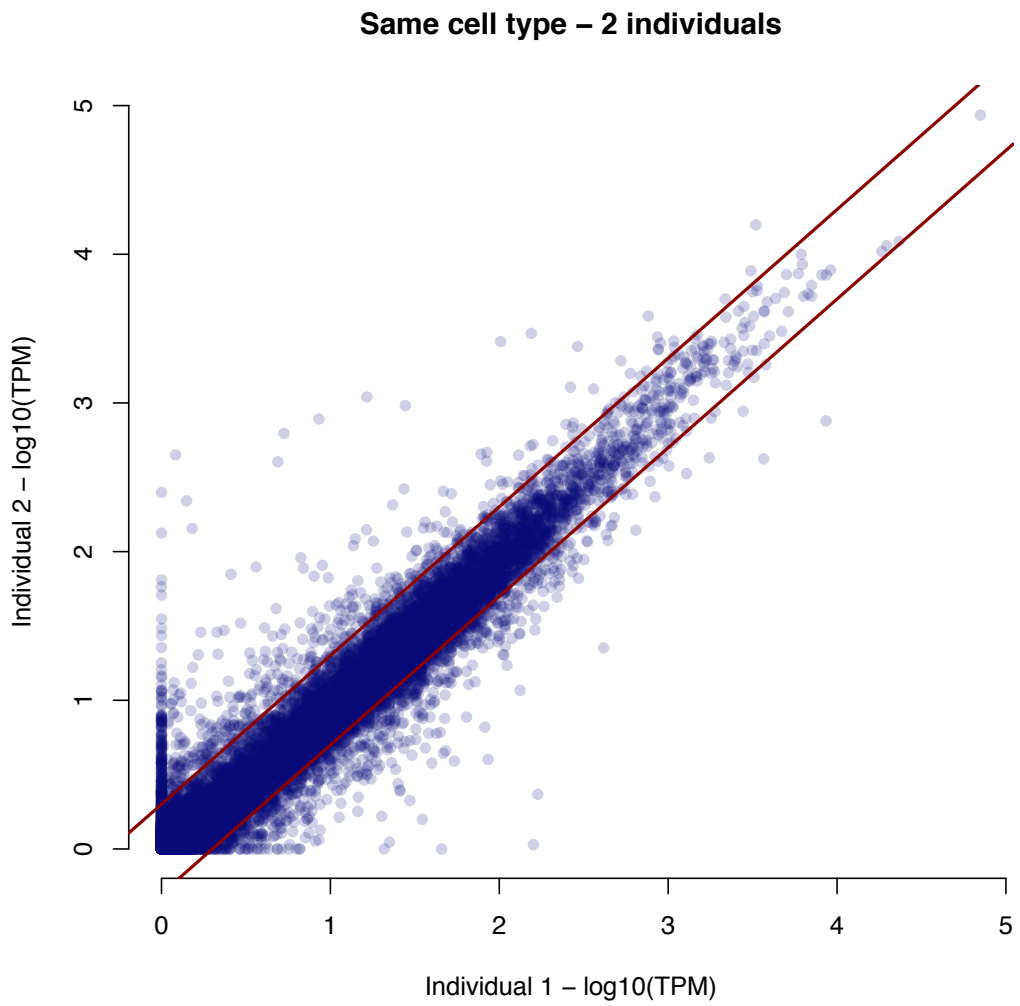
# Gene counts distribute as a multinomial



# Gene counts distribute as a multinomial

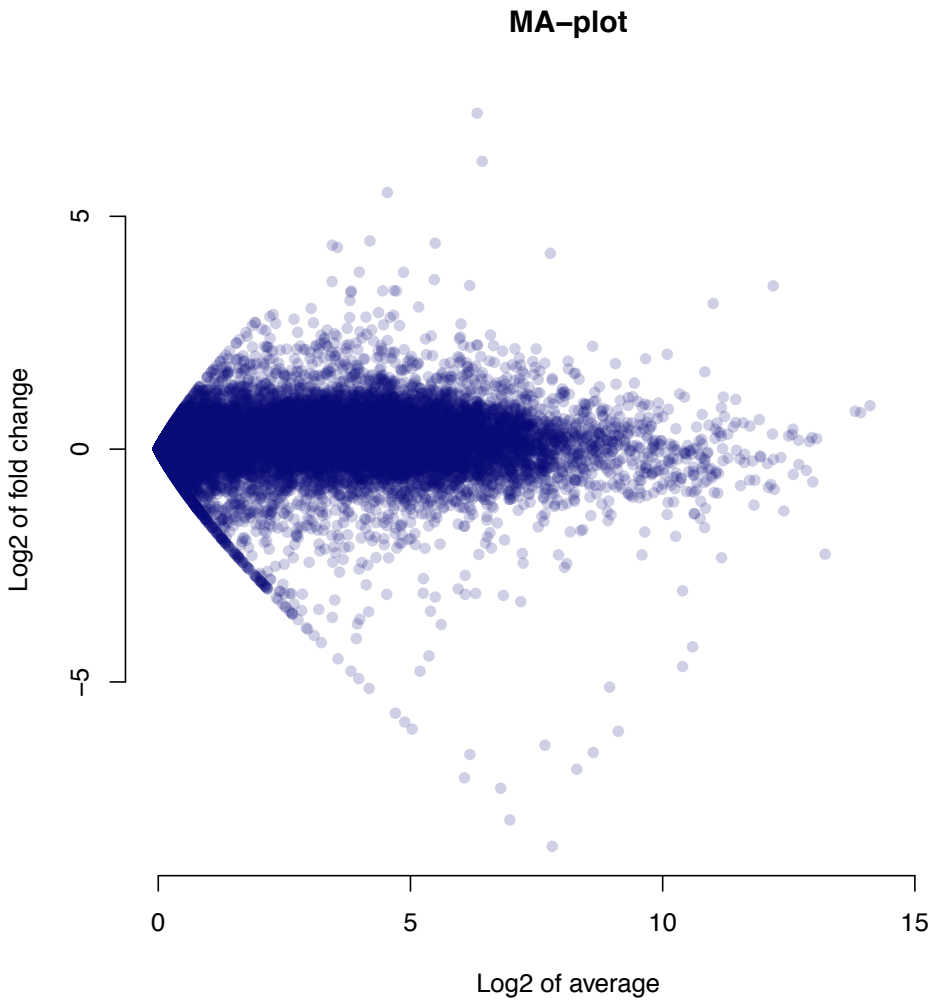


# And scatter-plots of log counts/RPKM are informative



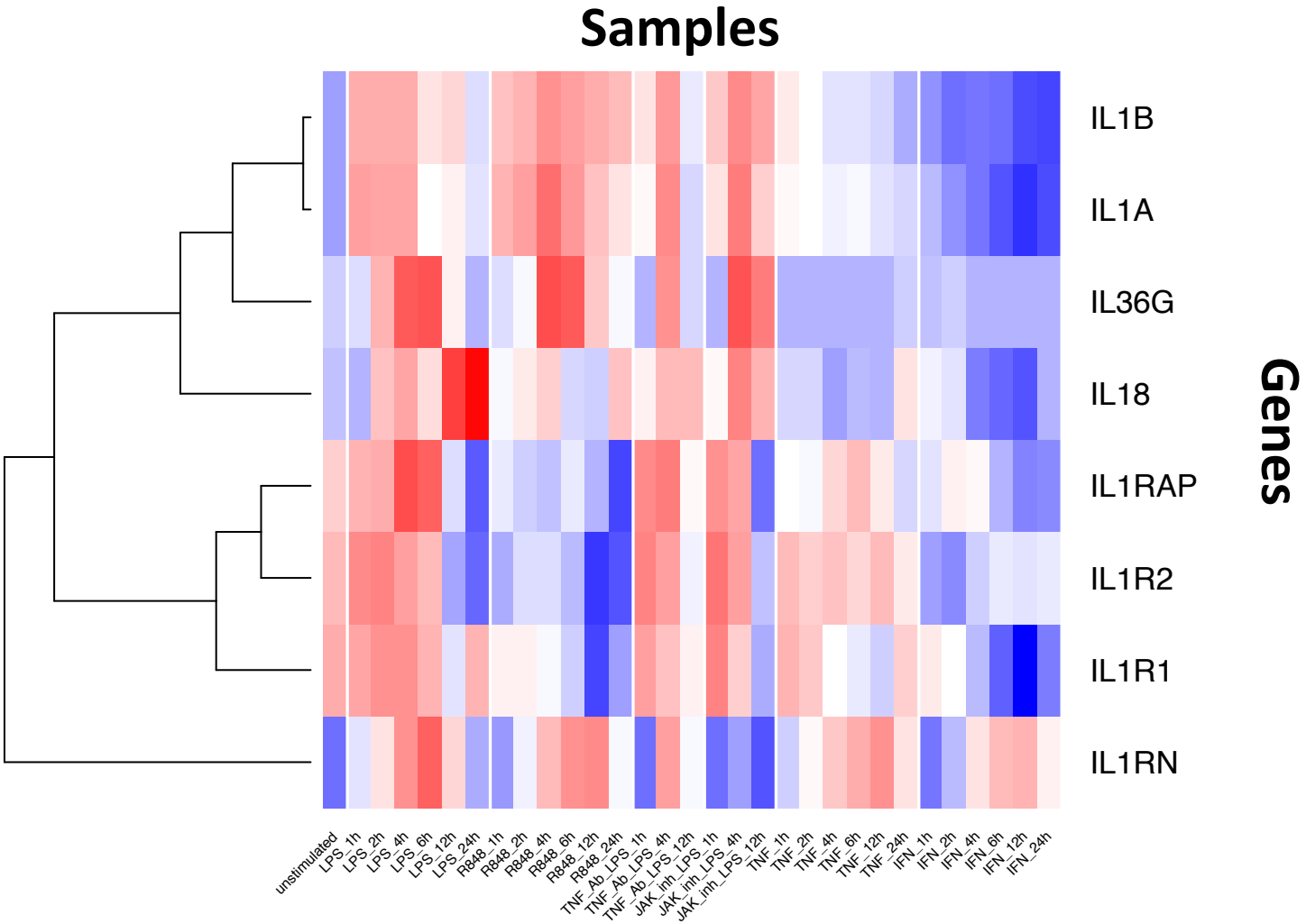

Cor=0.97

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



# Comparing samples
## Clustering

# Clustering – Similar patterns



# Hierarchical clustering – when are vector similar?

| Gene | Cond1 | Cond2 | Cond3 | Cond4 |
|---|---|---|---|---|
| $g_1$ | 2.5 | 5 | 7.5 | 10 |
| $g_2$ | 0.1 | 0.5 | 0.8 | 1.1 |
| $g_3$ | 0.2 | 0.3 | 0.4 | 11 |
| $g_4$ | 2.5 | 8 | 8 | 9 |

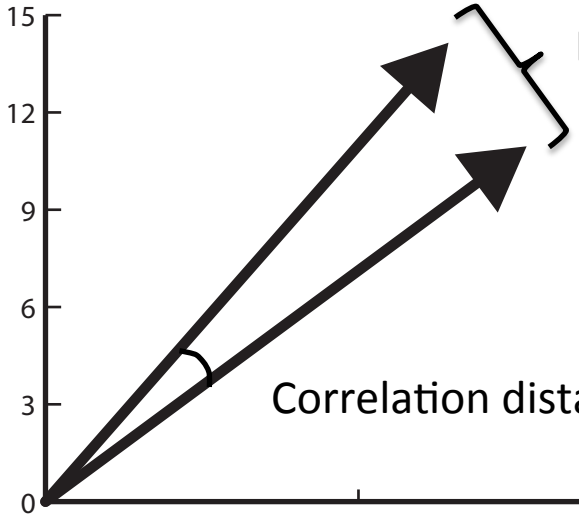
Clustering is about similarity:

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

# What is a distance?

The goal of clustering is to group together samples that are “similar”.

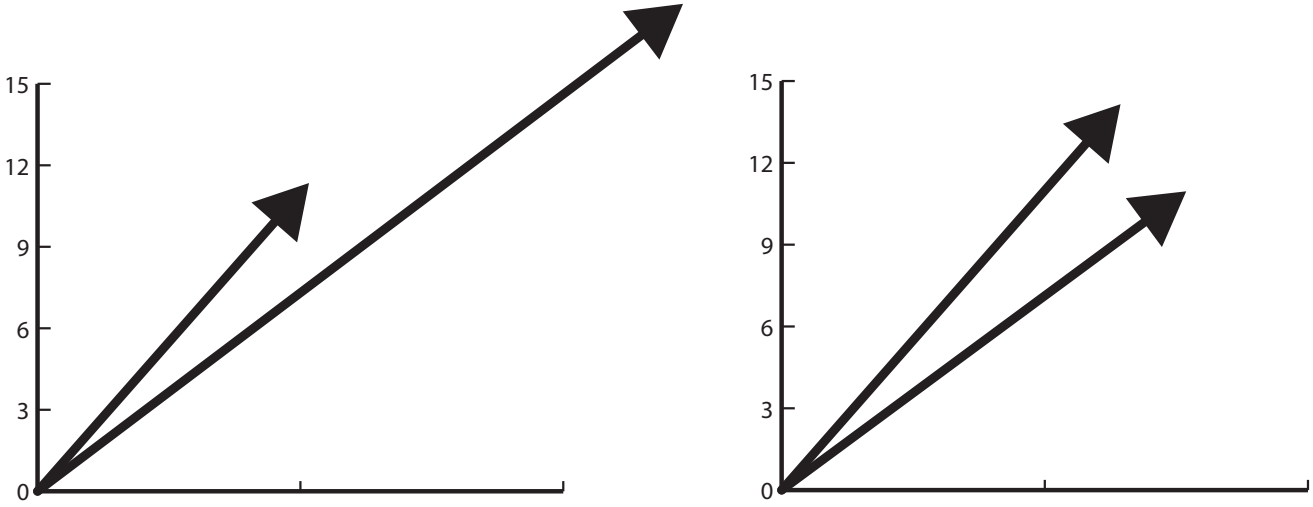
- When are two expression profiles “similar”?
- We consider each expression profile as a large “vector”. Each gene being a “dimension”



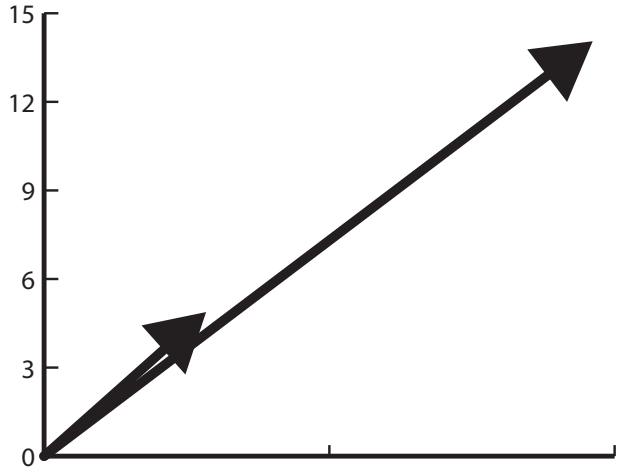

Euclidean distance $d(\vec{x},\vec{y}) = ||\vec{x}-\vec{y}|| = \sqrt{\sum_i (x_i - y_i)^2}$

Correlation distance $d(\vec{x},\vec{y}) = \frac{1 - cor(\vec{x},\vec{y})}{2}$

# What do difference distance care for?



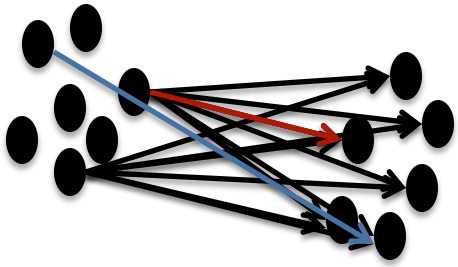

Similar correlation distance, very different euclidean distance




Correlation distance almost 0

# Similarity between groups of points

Linkage: Distance between two sets (`d(R,S)`)



Single Linkage $min\{d(r,s), s \in S,\ r \in R\}$

Complete Linkage $max\{d(r,s), s \in S,\ r \in R\}$

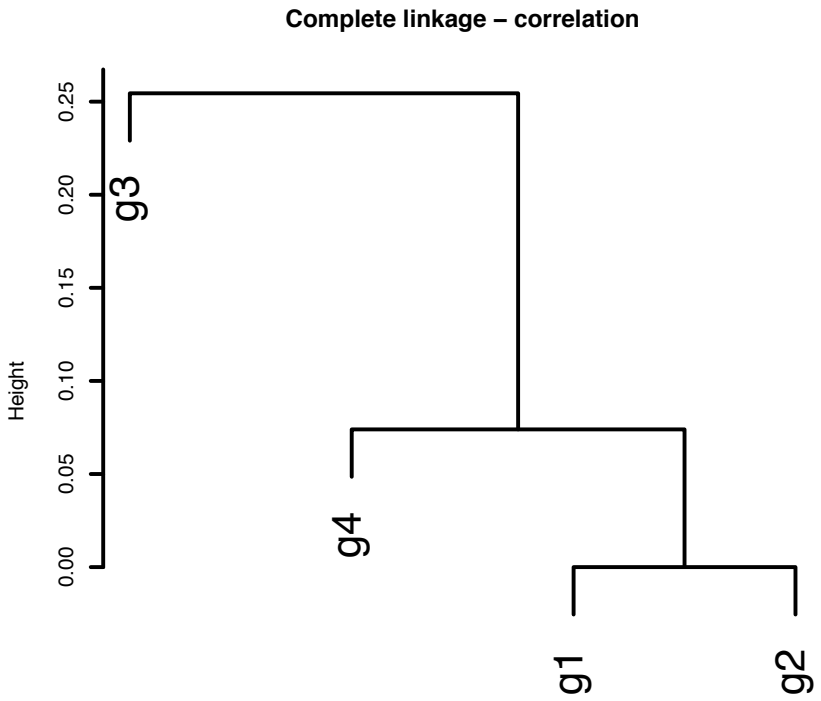
Average Linkage $mean\{d(r,s), s \in S,\ r \in R\}$

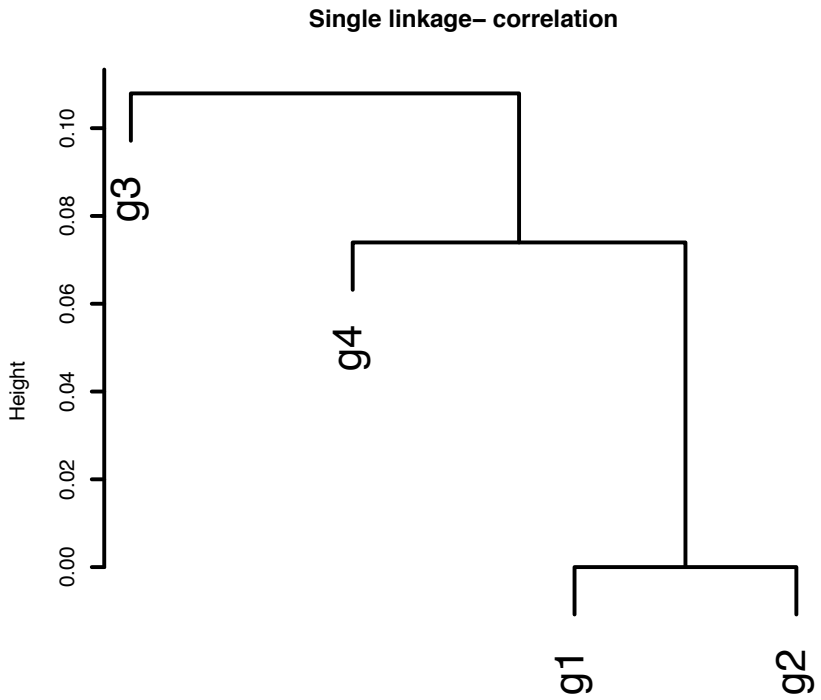
# Similarity between groups of points

- Linkage: Distance between 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 |

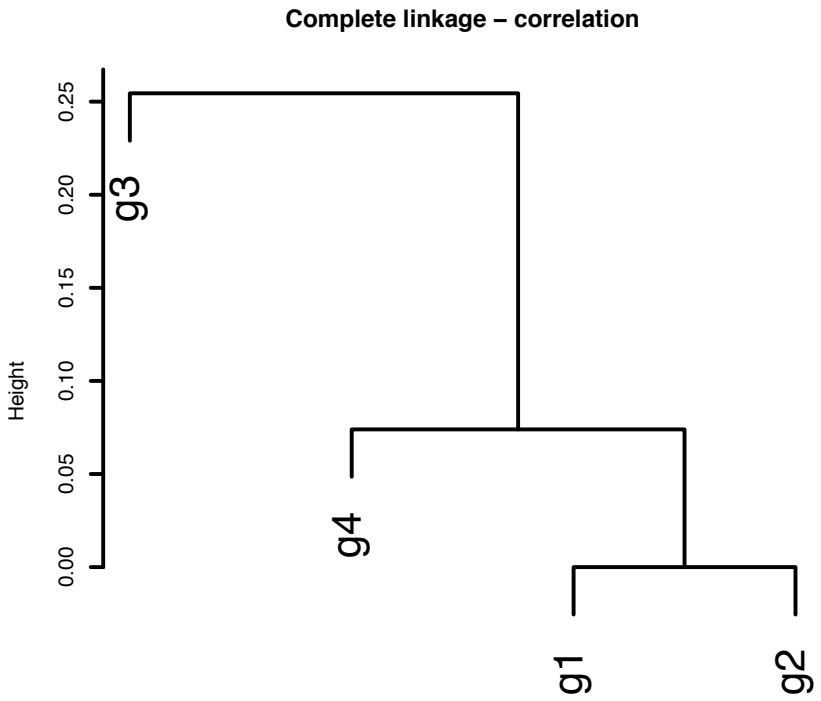
# The effect of the linkage method

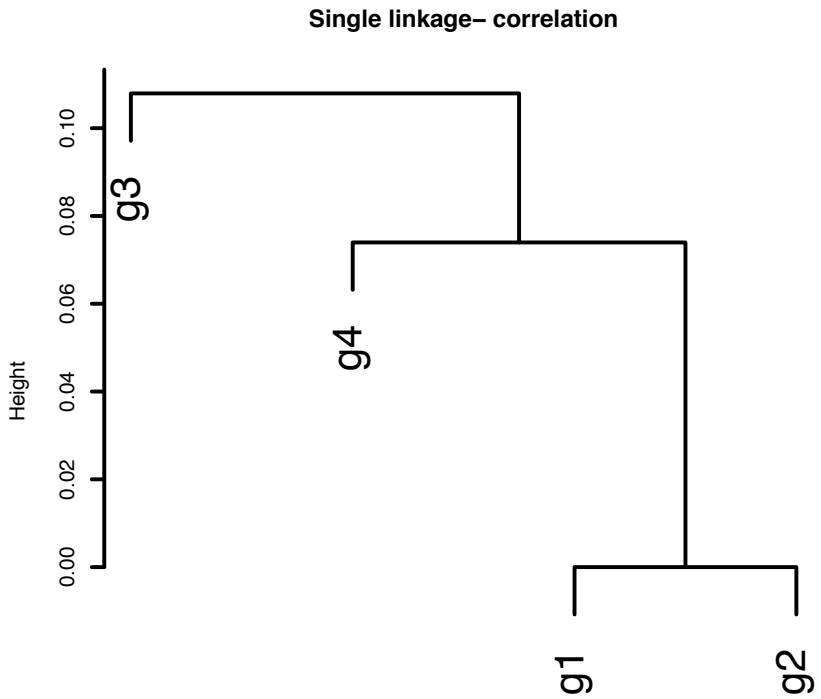



| Gene | Cond1 | Cond2 | Cond3 | Cond4 |
|---|---|---|---|---|
| $g_1$ | 2.5 | 5 | 7.5 | 10 |
| $g_2$ | 0.1 | 0.5 | 0.8 | 1.1 |
| $g_3$ | 0.2 | 0.3 | 0.4 | 11 |
| $g_4$ | 2.5 | 8 | 8 | 9 |

# The effect of the linkage method

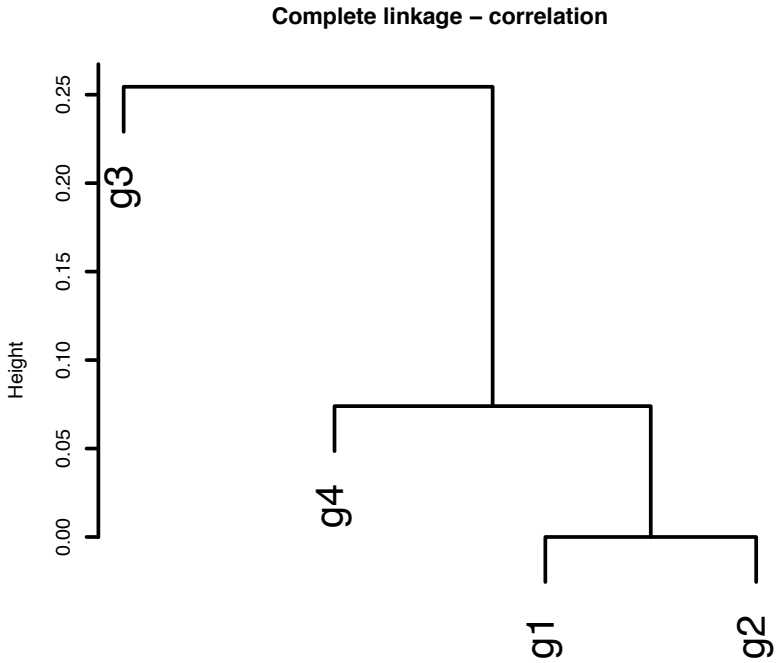


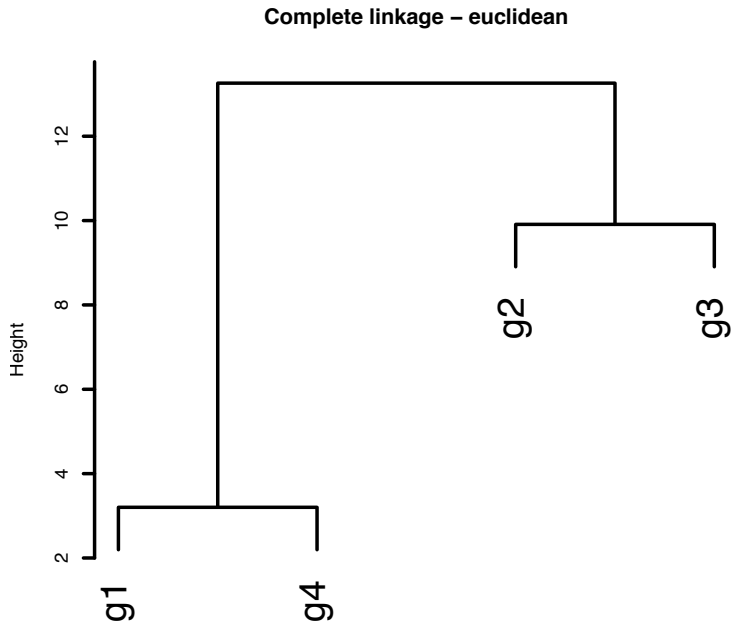




Correlation distance matrix

| Column1 | g1 | g2 | g3 | g4 |
|---|---|---|---|---|
| g1 | 0 | 0.00137174 | 0.10792118 | 0.07396763 |
| g2 | 0.00137174 | 0 | 0.12430885 | 0.05598953 |
| g3 | 0.10792118 | 0.12430885 | 0 | 0.25448339 |
| g4 | 0.07396763 | 0.05598953 | 0.25448339 | 0 |

# Effect of the distance!




| Column1 | g1 | g2 | g3 | g4 |
|---|---|---|---|---|
| g1 | 0 | 0.00137174 | 0.10792118 | 0.07396763 |
| g2 | | 0 | 0.12430885 | 0.05598953 |
| g3 | | | 0 | 0.25448339 |
| g4 | | | | 0 |

Correlation distance matrix

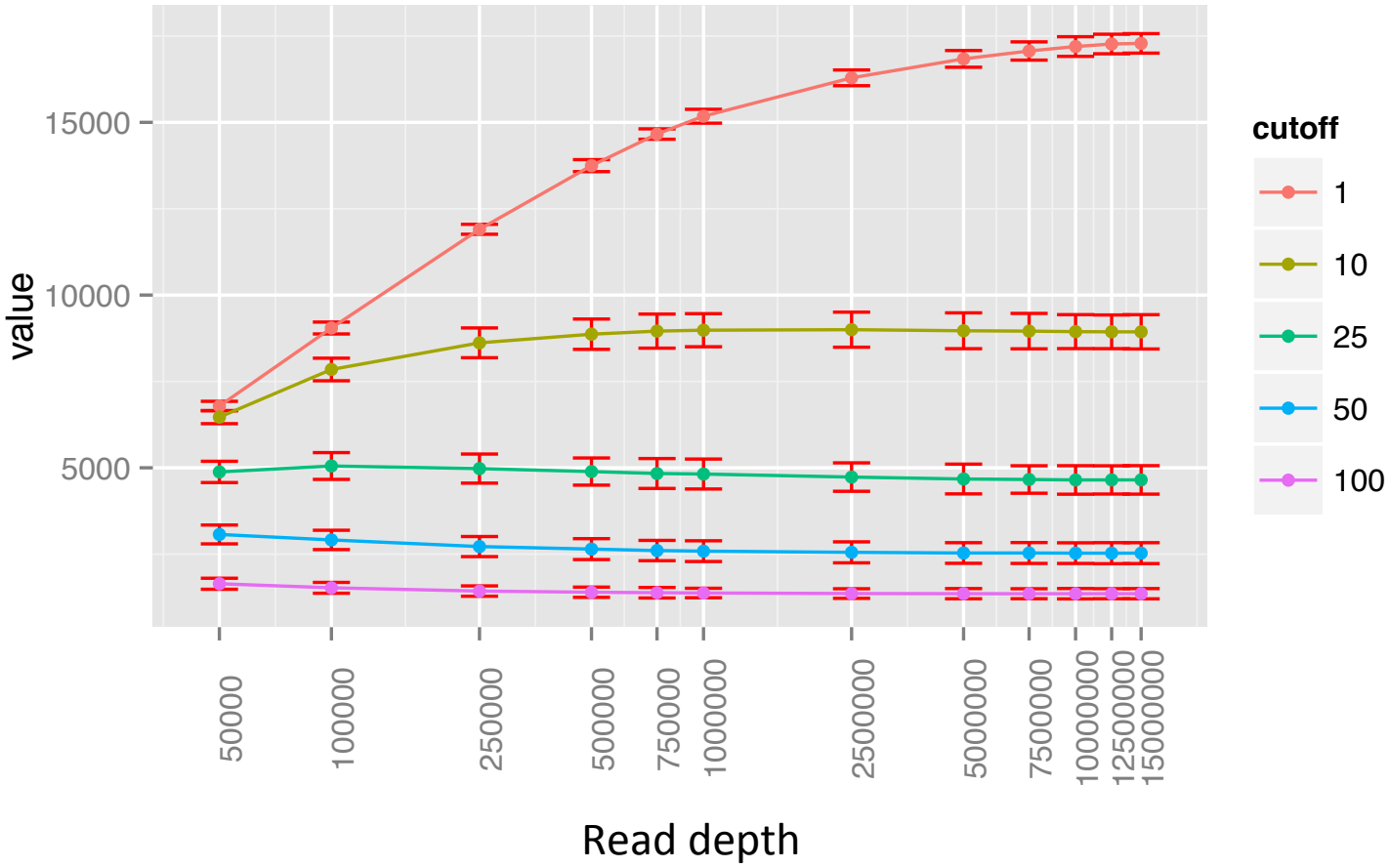
| Column1 | g1 | g2 | g3 | g4 |
|---|---|---|---|---|
| **g1** | 0.00 | 12.25 | 8.88 | 3.20 |
| **g2** | | 0.00 | 9.91 | 13.28 |
| **g3** | | | 0.00 | 11.24 |
| **g4** | | | | 0.00 |

Geometric (Euclidean) distance matrix

# Playing with clustering

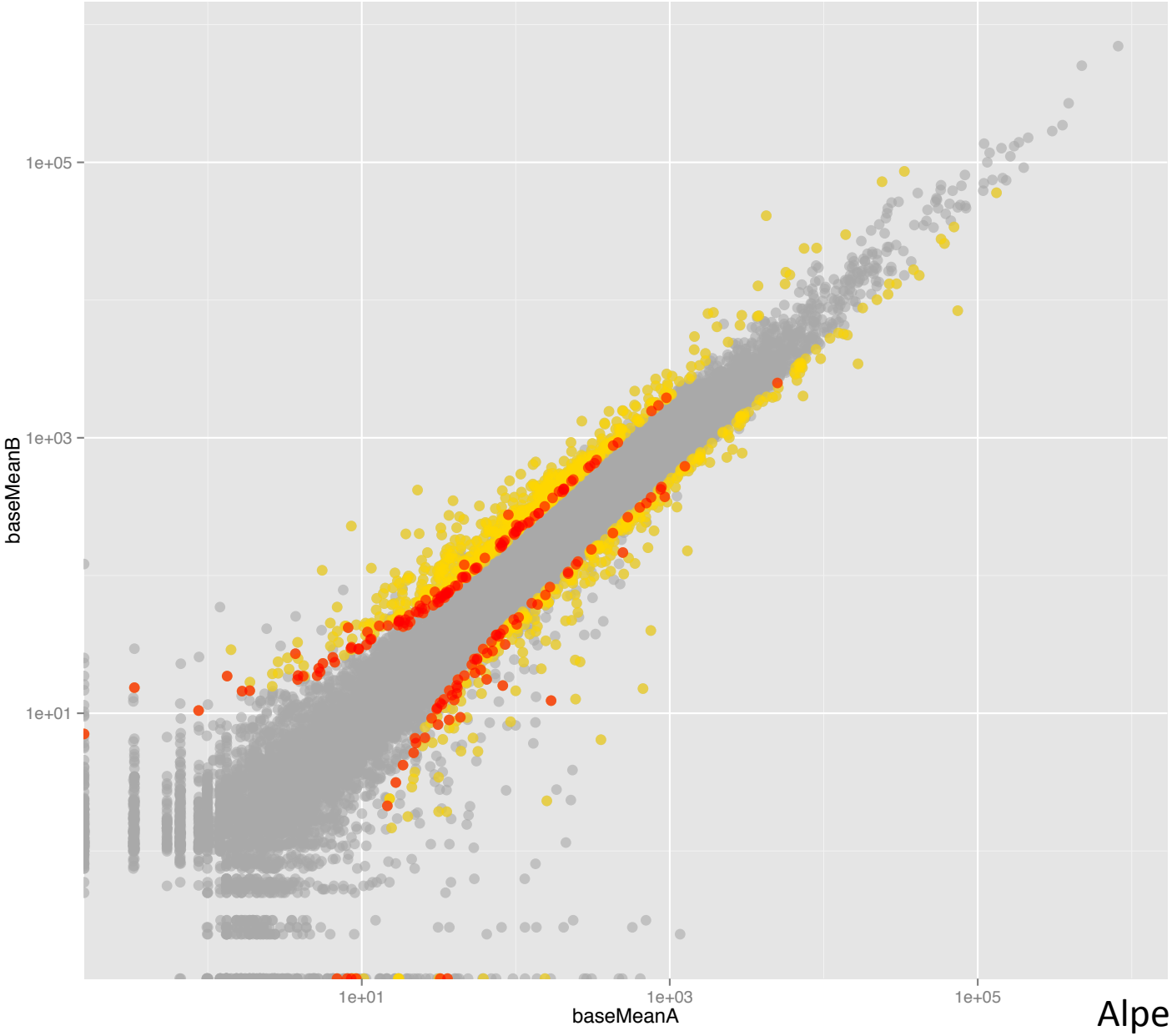
```
#Define the toy matrix#
#######################
m =  rbind (c(2.5,5,7.5,10), c(0.1,0.5,0.8,1.1), c(0.2,0.3,0.4,11), c(2.5,8,8,9))

#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);
```

# Robustness to low depth: Transcripts detected



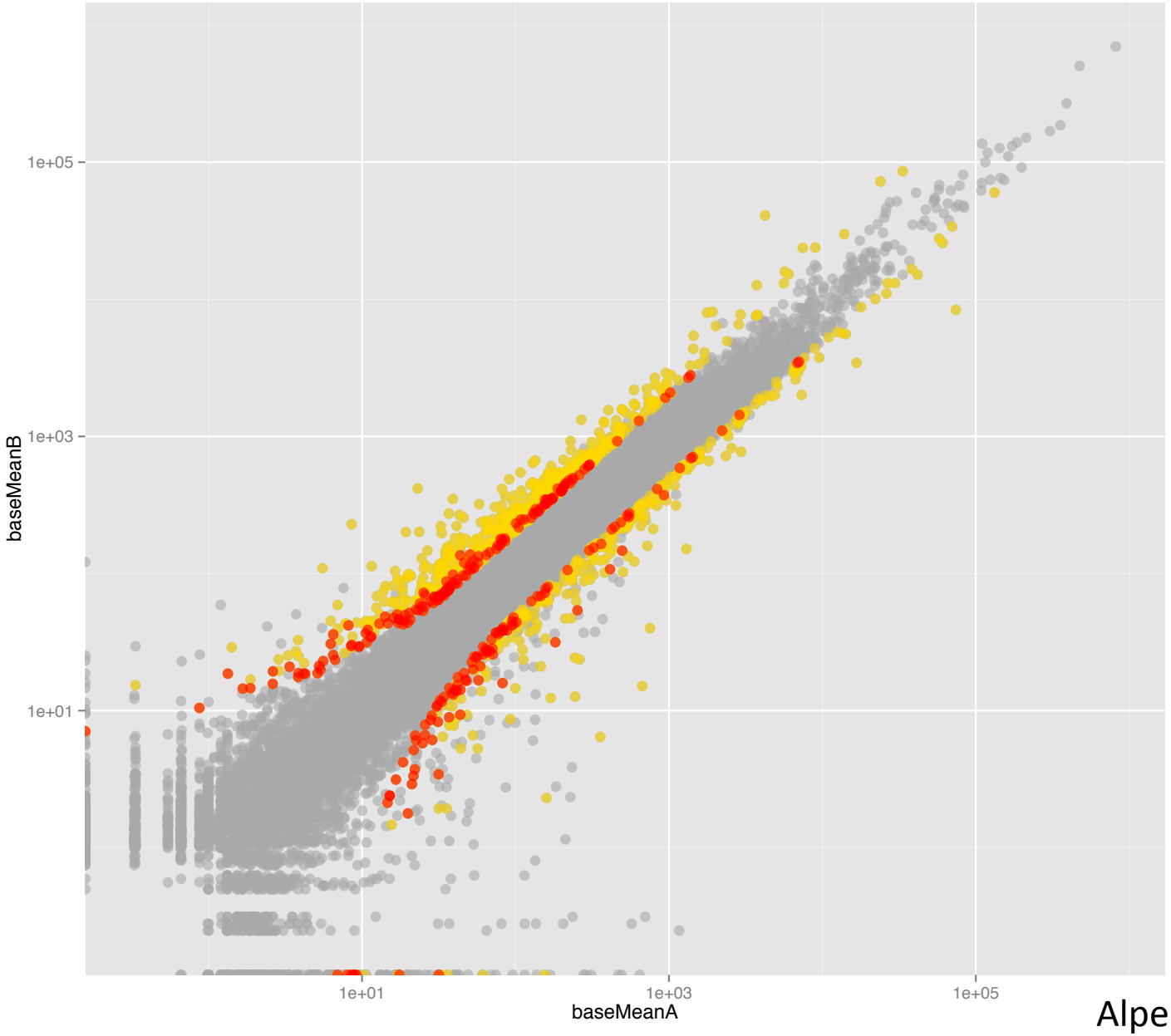

Alper Kucukural

# RSEM/DESeq: 15 mill reads in worm



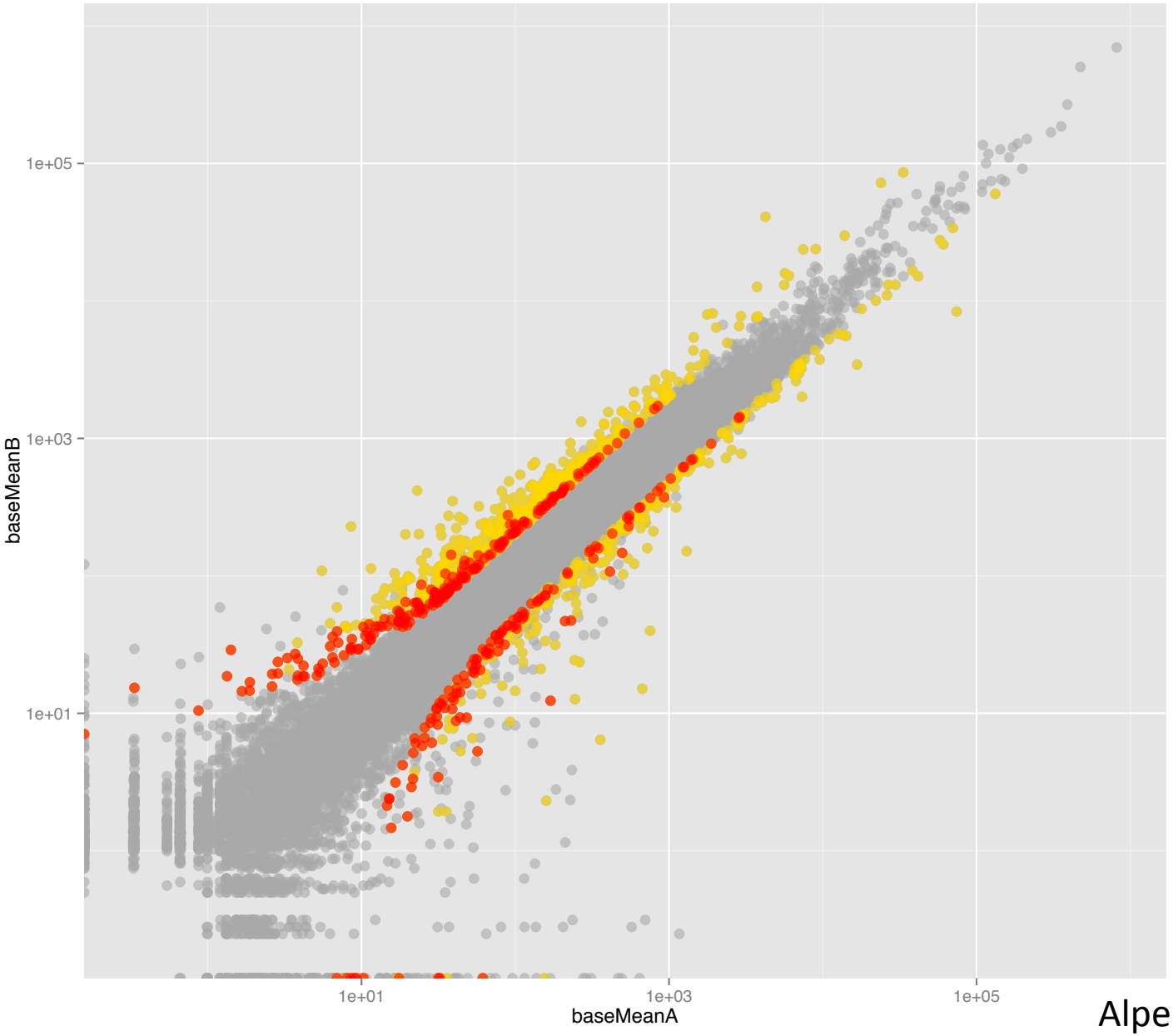

Alper Kucukural

# RSEM/DESeq: 10 mill reads in mouse



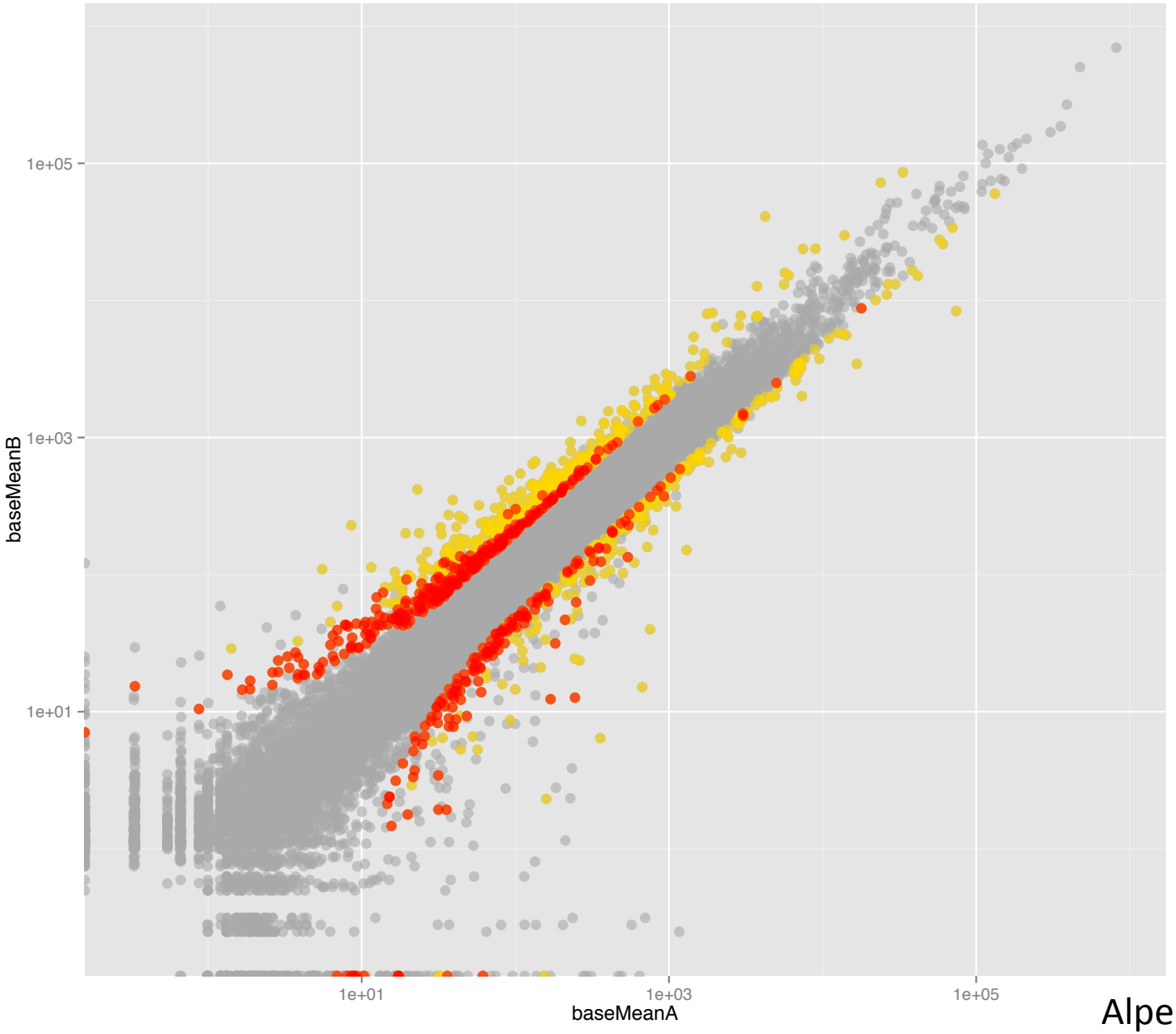

Alper Kucukural

# RSEM/DESeq: 7.5 mill reads in worm



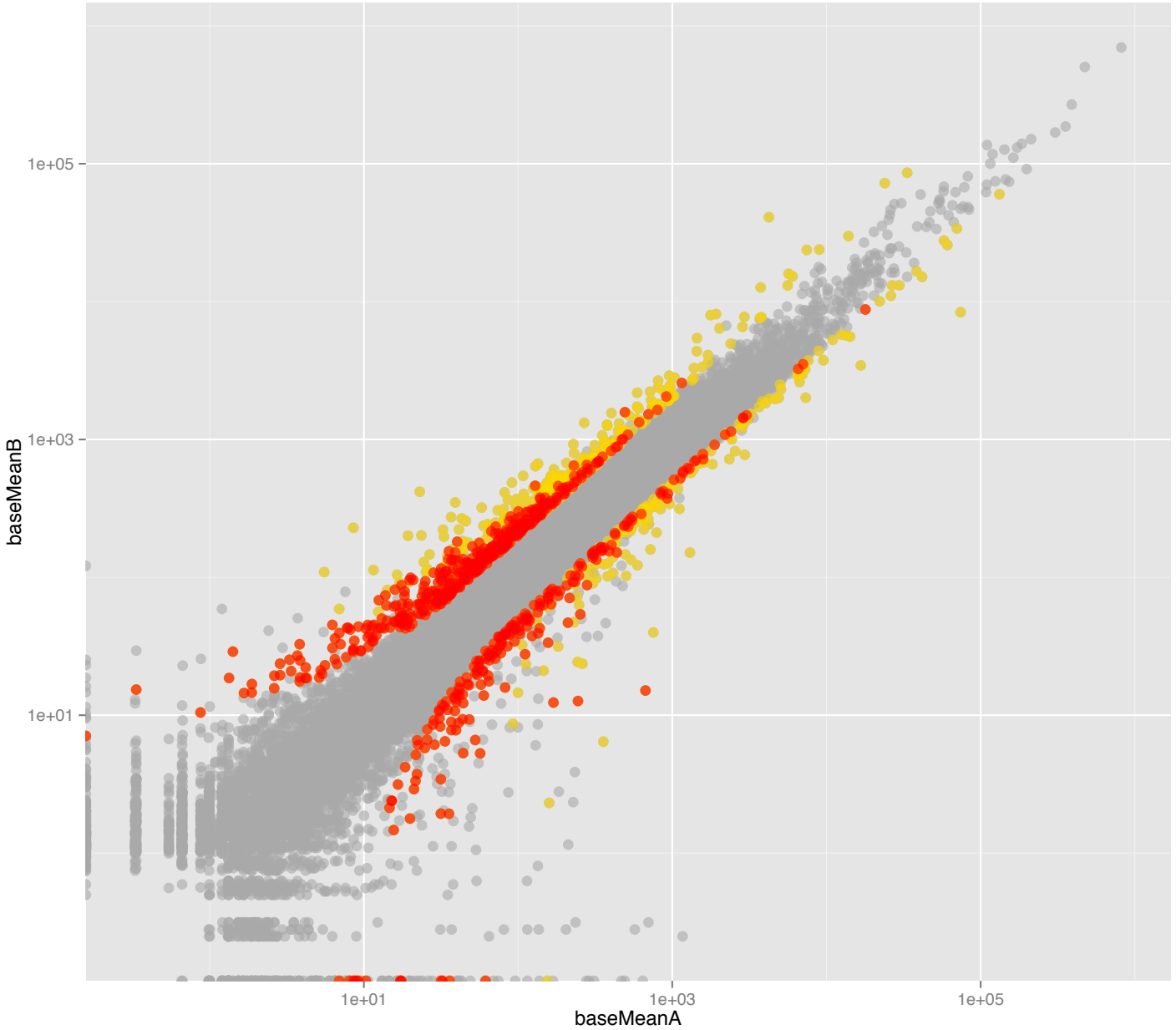

Alper Kucukural

# RSEM/DESeq: 5 mill reads in worm

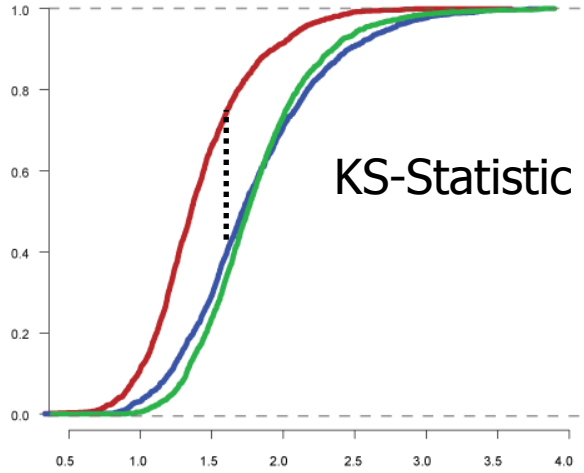


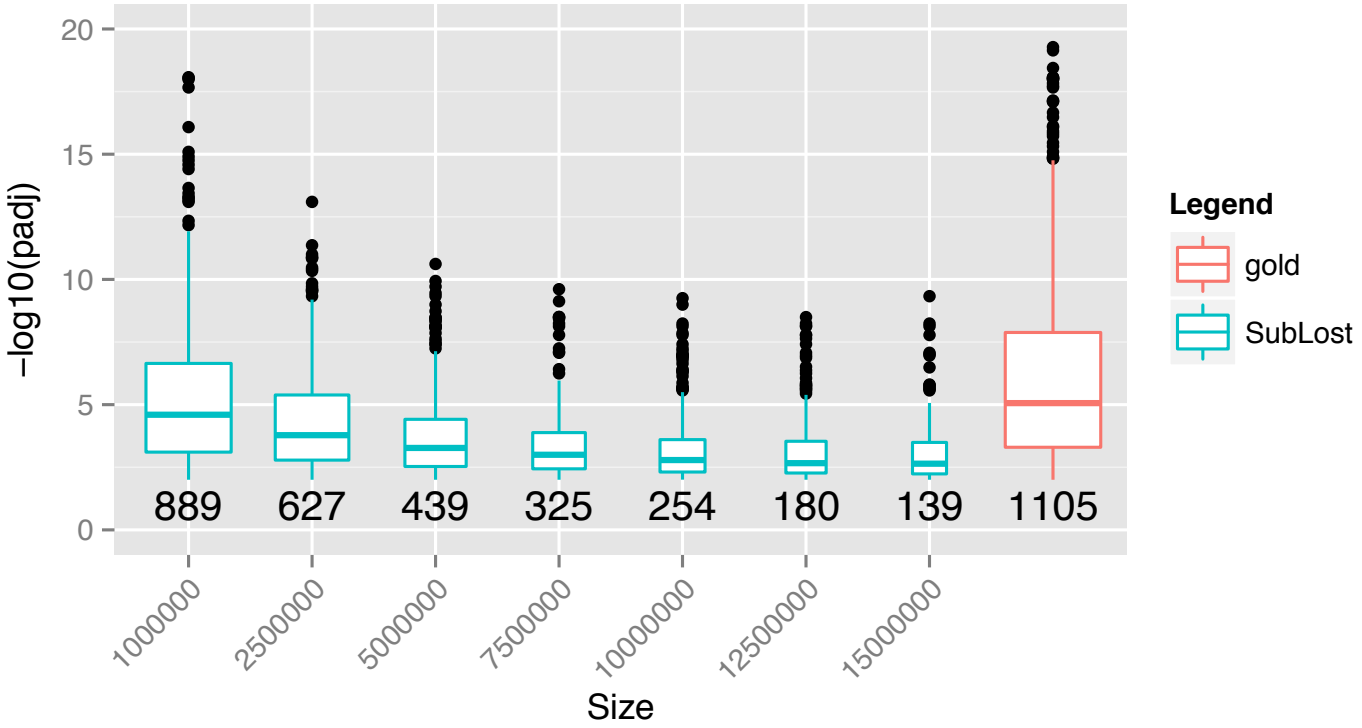

Alper Kucukural

# RSEM/DESeq: 2.5 mill reads in worm



# The loss is qualitatively small



# The loss is qualitatively small