STAT718/BIOL703: Genomic Data Science
Differential gene expression (DGE) analysis: Preprocessing

Dr. Yen-Yi Ho (hoyen@stat.sc.edu)

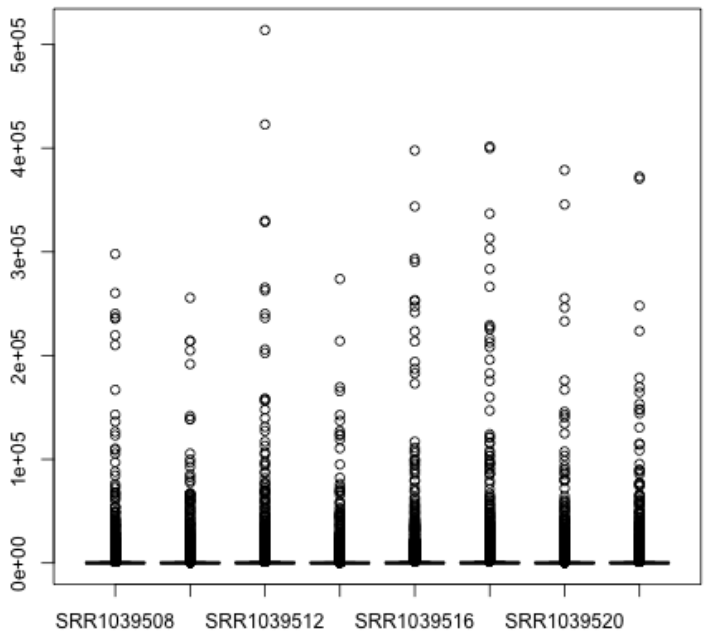
Learning Objectives

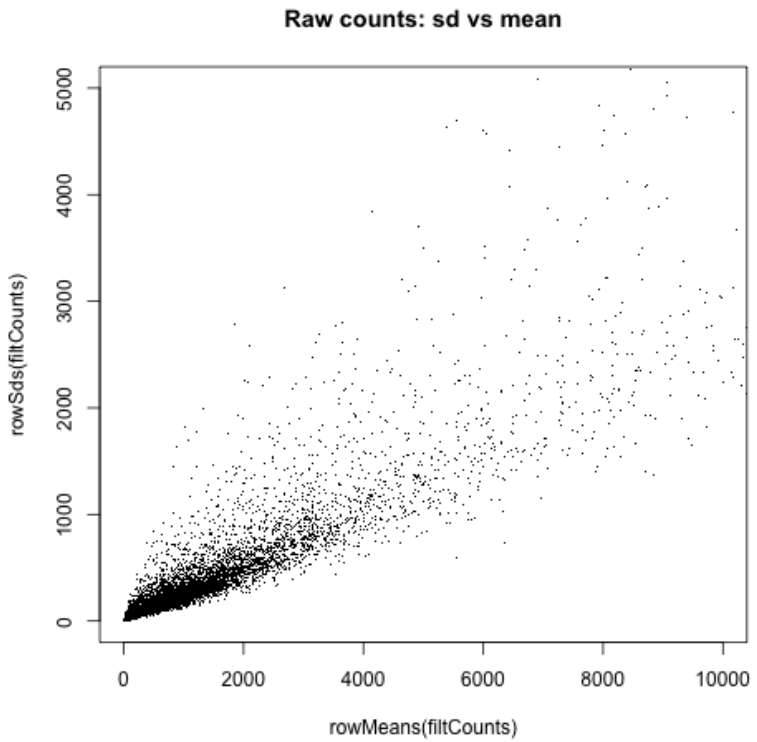
- RNAseq data preprocessing
- Explore different types of normalization methods
- Become familiar with the ‘DESeqDataSet‘ object
- Understand how to normalize counts using DESeq2

Filter the Gene

- Filtering out non-significant genes to decrease the impact of multiple testing
- Very lowly expressed reads

Raw Count Distribution

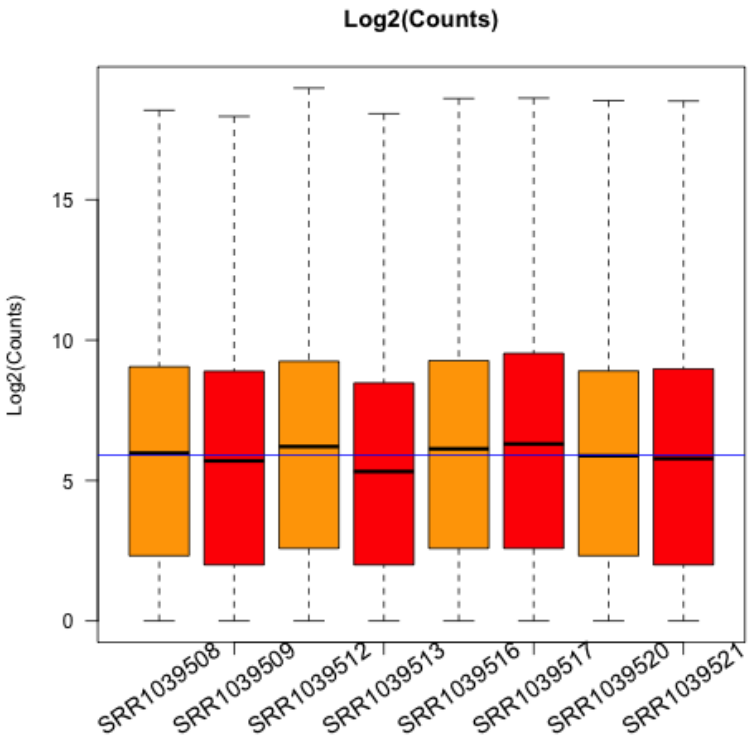


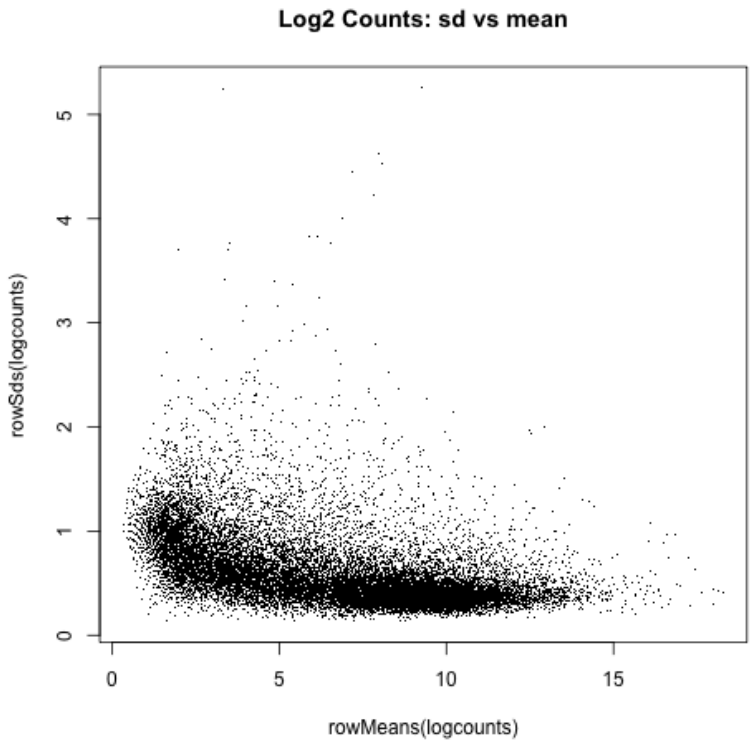


Data Transformation

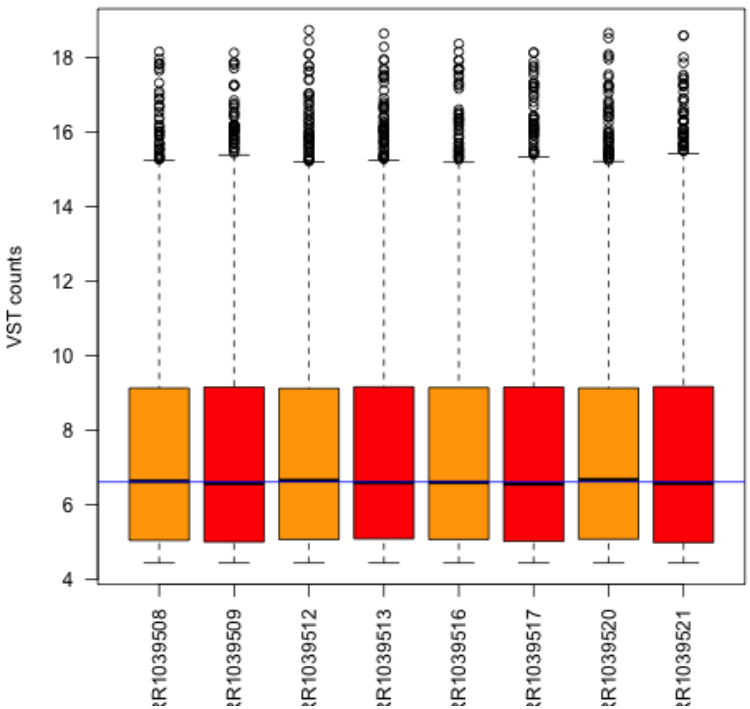
- $\log_2$ transformation
- Variance stabilizing transformation (VST)

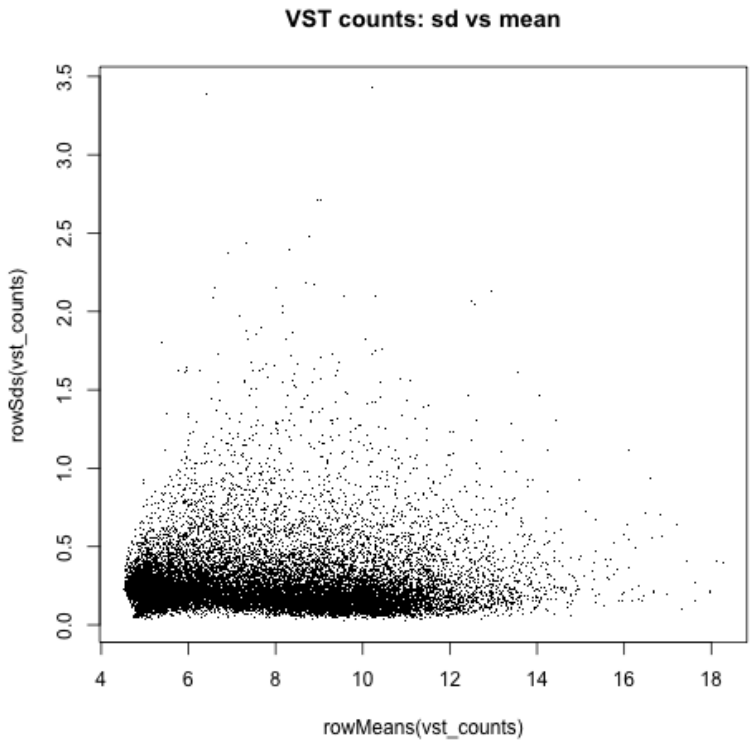
$\log_2$ Transformation



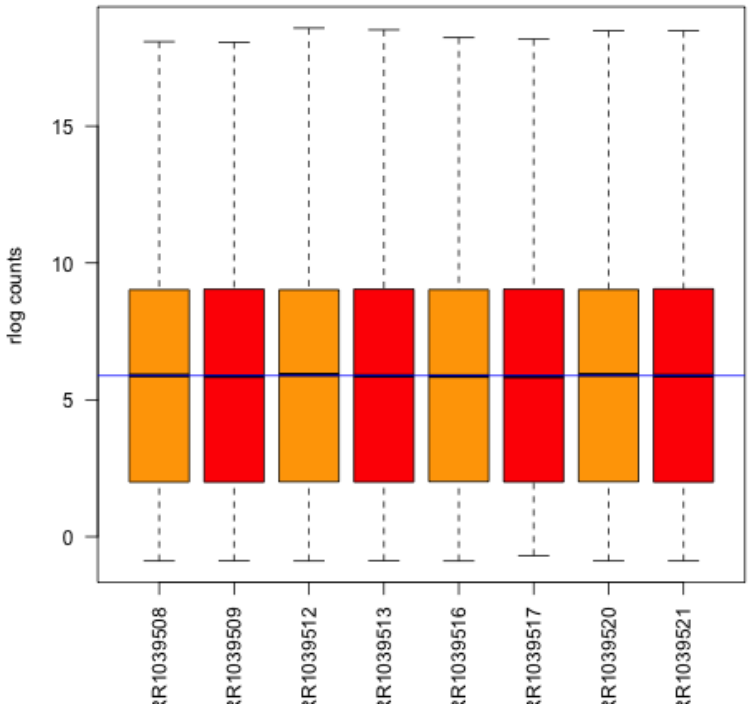


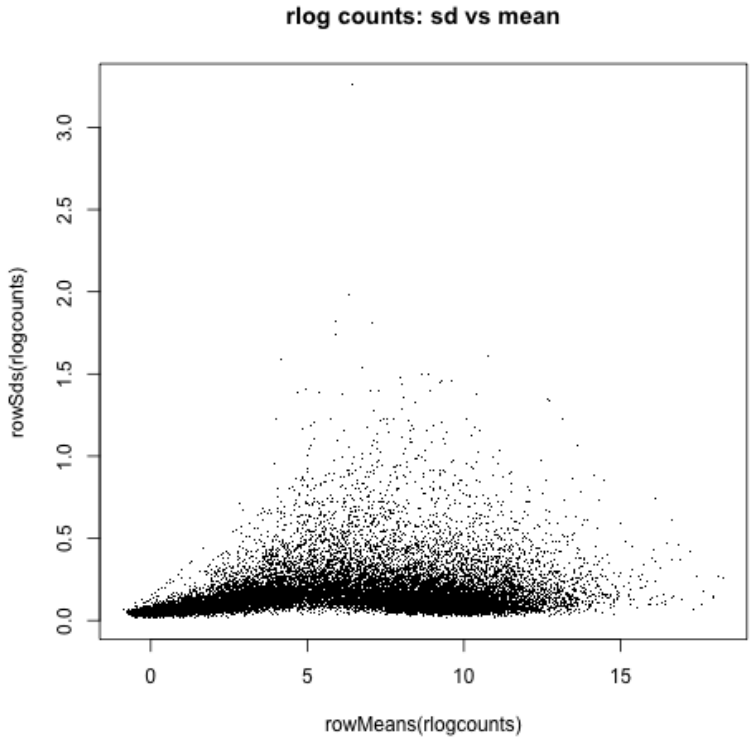
VST Transformation





rlog Transformation

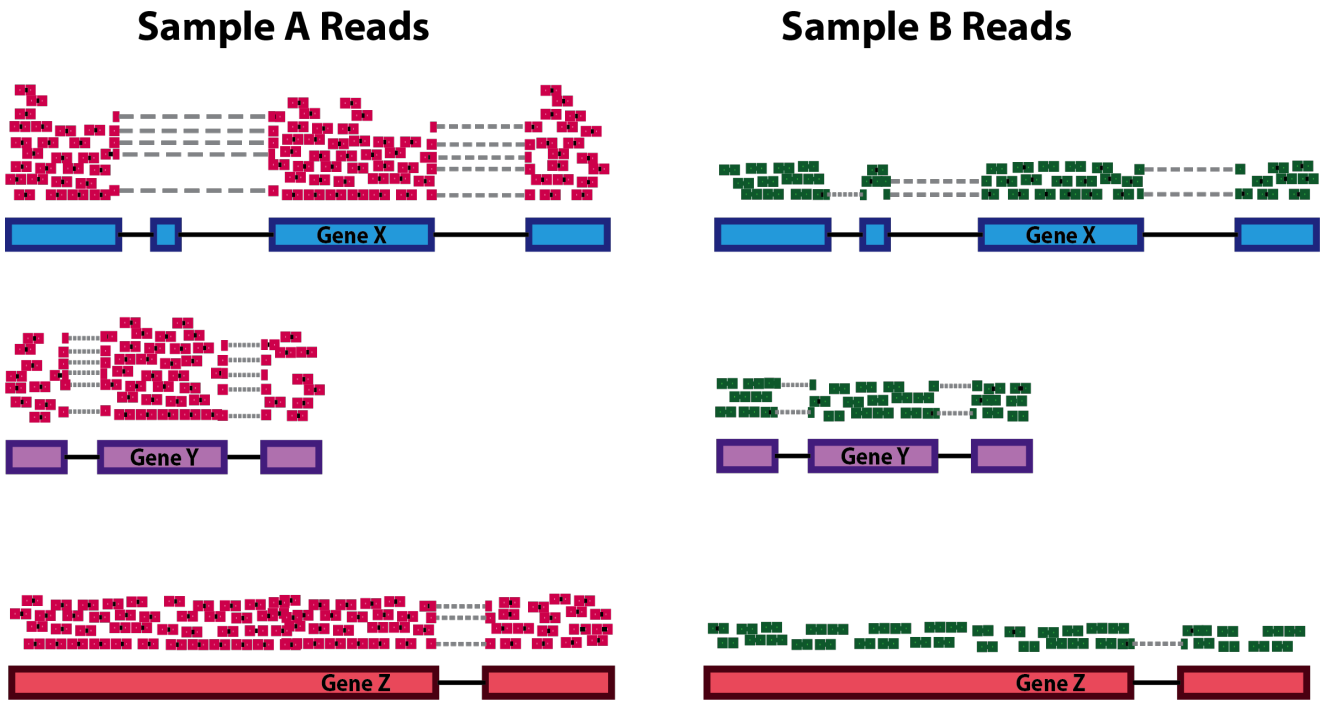




Normalization

Normalization is the process of scaling raw count values to account for the technical artifacts. In this way the expression levels are more comparable between and/or within samples. The main factors often considered during normalization are:

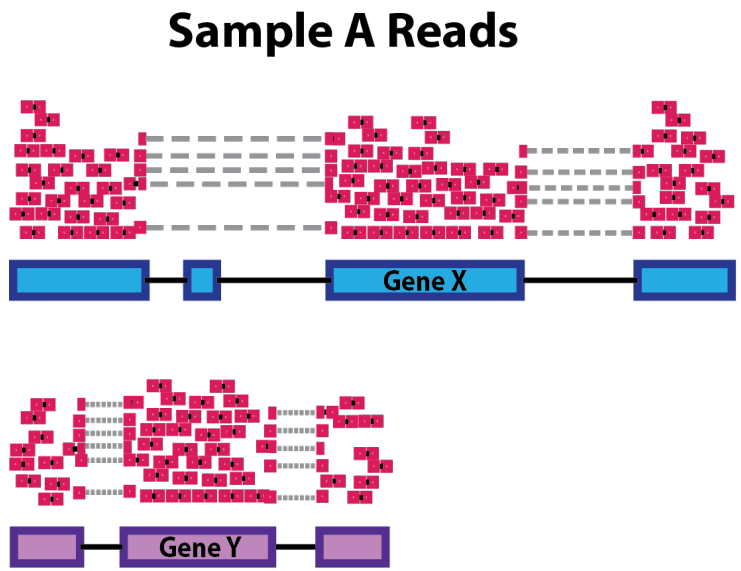
- Sequencing depth: comparison between samples



Normalization

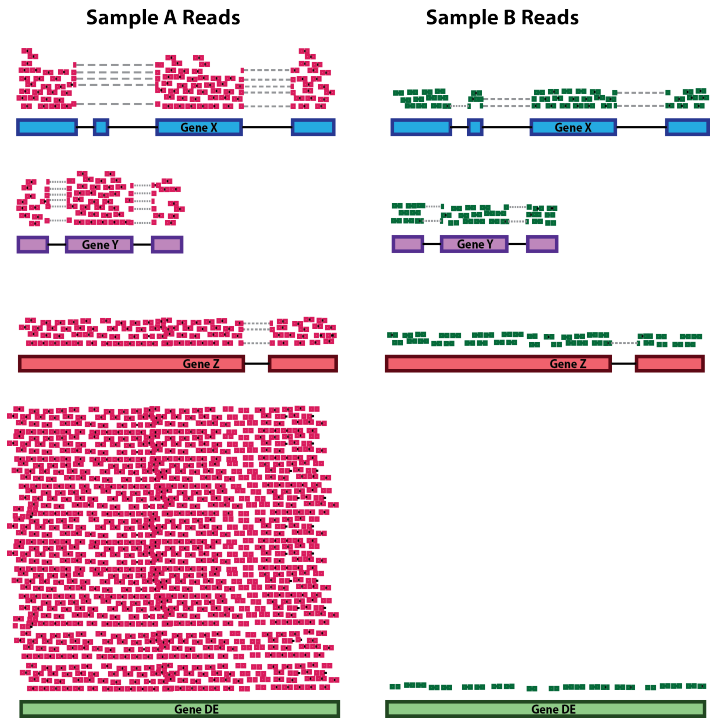
Normalization is the process of scaling raw count values to account for the technical artifacts. In this way the expression levels are more comparable between and/or within samples. The main factors often considered during normalization are:

- Sequencing depth : comparison between samples
- Gene length: between different genes within the same sample



Normalization

- Sequencing depth : comparison between samples
- Gene length: between different genes within the same sample
- RNA composition: a few highly differentially expressed genes between samples, differences in the number of genes expressed between samples, or presence of contamination can skew some types of normalization methods.



Common Normalization Methods

Normalization method	Description	Factors	Use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	NOT for within sample comparisons or DE analysis
TPM	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	Within & Between samples NOT for DE analysis
RPKM/FPKM	similar to TPM	sequencing depth and gene length	Between genes within a sample; NOT for between sample or DE analysis
DESeq2's median of ratios	divid by size factor	sequencing depth and composition	Between samples & DE analysis NOT for within sample
EdgeR's trimmed mean of M values (TMM)	weighted trimmed mean	sequencing depth and composition	Between samples & DE analysis NOT for within sample

RPKM/FPKM (not recommended)

- Using RPKM/FPKM normalization, the total number of RPKM/FPKM normalized counts for each sample will be different.

Gene	Sample A	Sample B
XCR1	5.5	5.5
WASHC1	73.4	21.8
...	...	
Total RPKM-normalized counts	10^6	1.5×10^6

Table: RPKM-normalized count table

DESeq2: Median of radios

- Step 1: creates a pseudo-reference sample (row-wise geometric mean)

For each gene, a pseudo-reference sample is created that is equal to the geometric mean across all samples.

Gene	sample A	sample B	reference
EF2A	1489	906	$\sqrt{1489 * 906} = 1161.5$
ABCD1	22	13	$\sqrt{22 * 13} = 16.9$
...			...

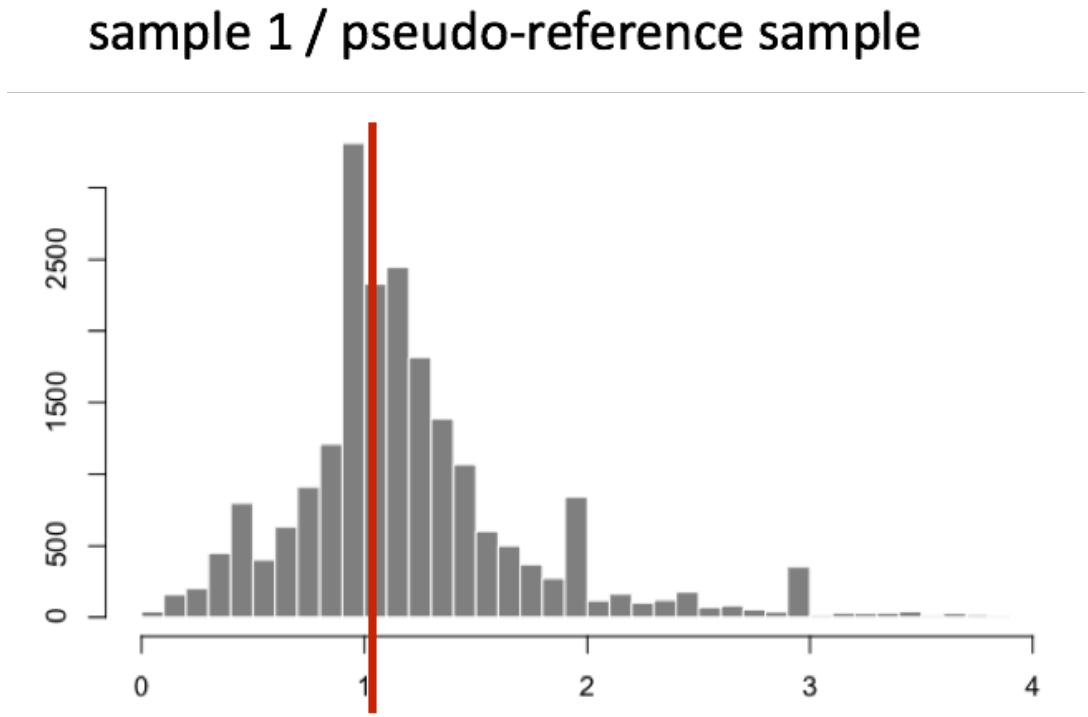
DESeq2: Median of radios

- Step 2: calculates ratio of each sample to the reference

Gene	sample A	sample B	reference	ratio of sampleA/ref	ratio of sampleB/ref
EF2A	1489	906	1161.5	1489/1161.5 = 1.28	906/1161.5=0.78
ABCD1	22	13	16.9	22/16.9 = 1.30	13/16.9 = 0.77
MEFV	793	410	570.2	793/570.2 = 1.39	410/570.2 = 0.72
BAG1	76	42	56.5	76/56.5 = 1.35	42/56.5 = 0.74
MOV10	521	1196	883.7	521/883.7 = 0.590	1196/883.7 = 1.35

DESeq2: Median of radios

- Step 3: calculate the normalization factor for each sample (size factor)



This method is robust to imbalance in up-/down-regulation and large numbers of differentially expressed genes.
Usually these size factors are around 1.

DESeq2: Median of radios

- Step 4: calculate the normalized count values using the normalization factor

Sample A median ratio=1.3
Sample B median ratio=0.7

Gene	sample A	sample B
EF2A	1489/1.3 =1145.39	906/0.77=1176.62
ABCD1	22/13.=16.92	13/0.77=16.88
...	...	...

Table: Normalized Counts

Count normalization using DESeq2 see R code posted on course website.