Lecture 7

Introduction to bioinformatics (MVE510)

Autumn, 2020

Additional reading: Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. Finotello F and Di Camillo F. *Briefings in Functional Genomics*, 14(2) 2014.

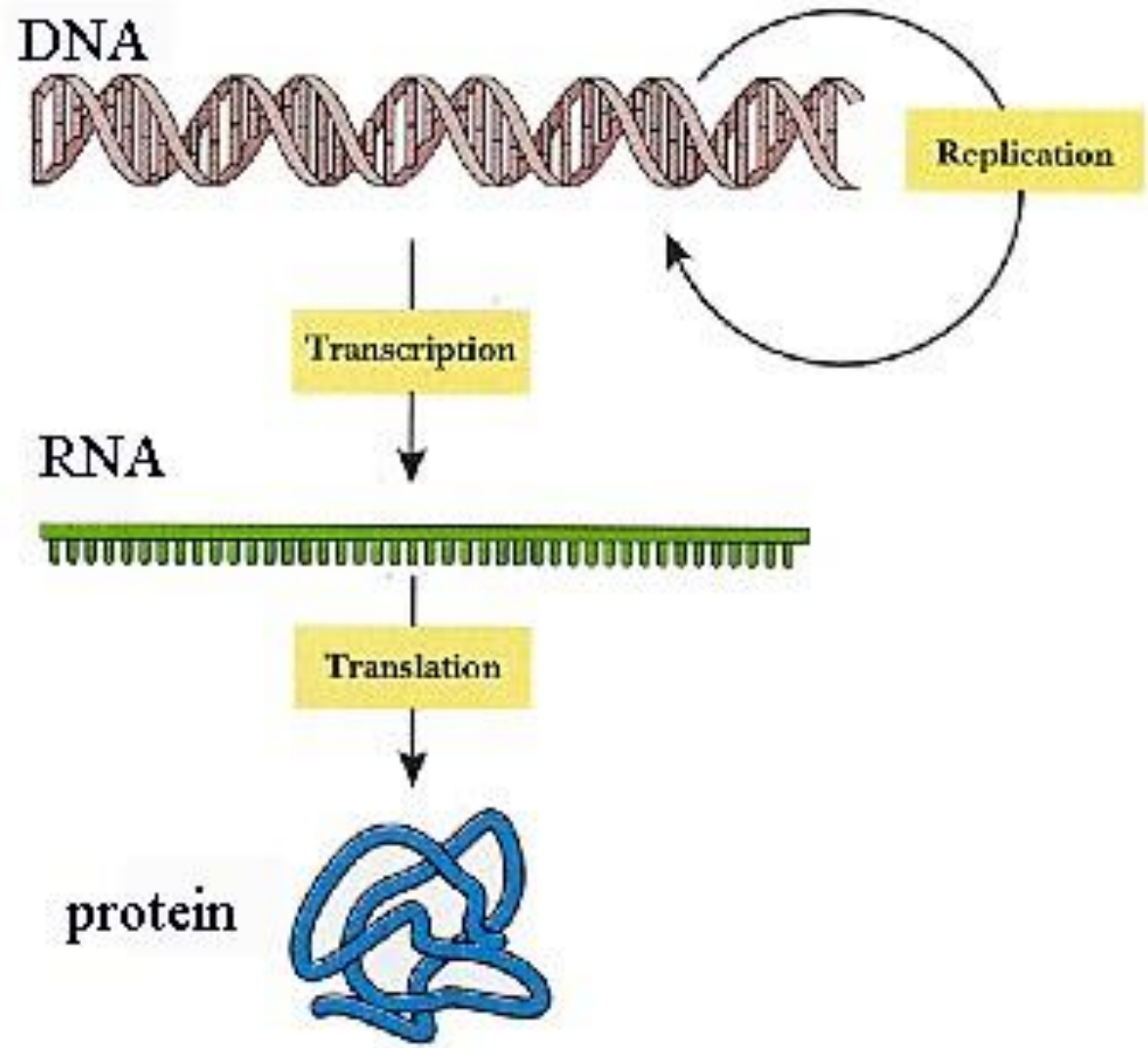
Repetition from last lecture

- Suffix trees are created from the suffixes of a DNA sequence. By traversing the tree we can find the position of any subsequence.
- The Burrows-Wheeler transform orders the genome in a specific way that makes identification of subsequences highly efficient.
- Read alignment using suffix arrays and the Burrows-Wheeler transform are as fast as a suffix tree but does not require the entire tree to be created and stored in memory.

Today agenda

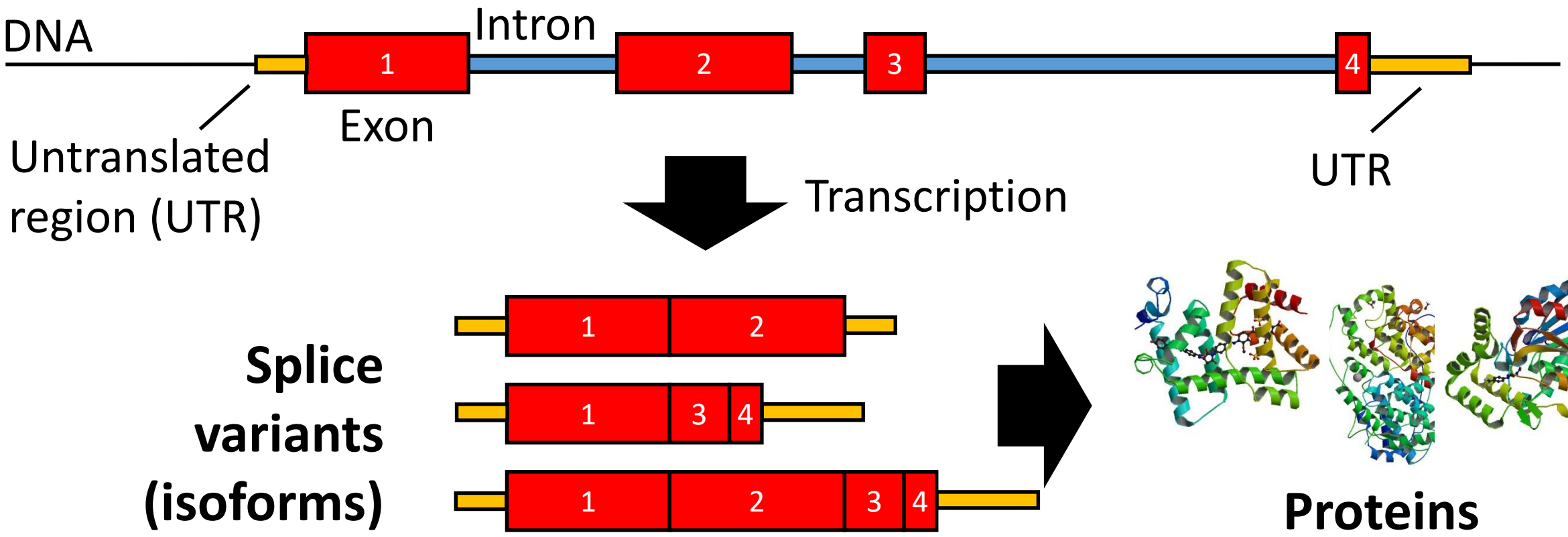
- RNA-seq: Gene expression analysis using next generation sequencing
- The principles of RNA-seq
- Data analysis of RNA-seq data
  - Gene quantification
  - Normalization
  - Identification of differentially expressed genes

The central dogma



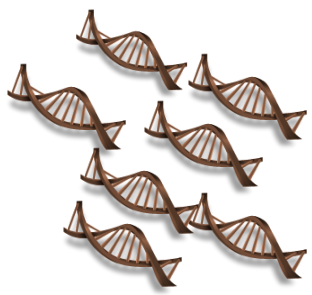
The structure of genes – some nomenclature

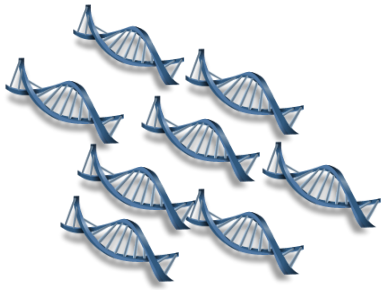
- The gene structure in many eukaryotic genomes is highly complex



The RNA-seq process

Reverse transcription





```
GATAGTACGGT
AATACGATAAA
ACGATAGGTAC
ACGTACGTGAC
CGGATATCCAT
TAAATTTAACG
AATACGATAAA
```

Extracted RNA (total or mRNA)

Complementary DNA (cDNA)

Sequencer

Reads

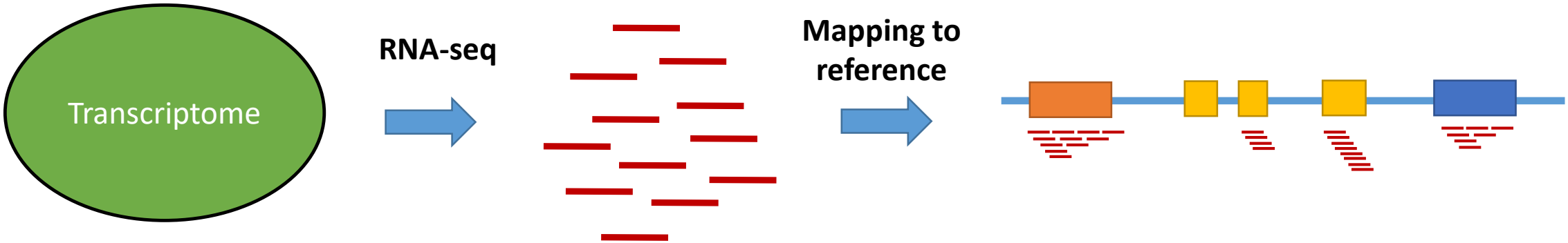
Analysis of RNA seq data

Three main steps

1. Quantification of the gene expression
   - From reads to a (semi)quantitative measurements of gene expression
2. Normalization
   - Correction of systematic errors within and between samples
3. Identification of differentially expressed genes (DEGs)
   - Find genes with a significant difference in gene expression

Quantification of gene abundances

- Read mapping is used to identify the origin of each fragment.



Challenges

- A large number of fragments (>10 million/sample)
- Short fragments (35-150 bases)
- Sequencing errors (substitutions, indels)

Garber et al, Computational methods for transcriptome annotation and quantification using RNA-seq, Nature Methods 8(6) 2011.

Quantification of gene abundances

Common references

- Genome
  - Requires high quality genome assembly
  - The mapper needs to be able to handle splicing (splice-aware)

- *De novo* assembled transcriptome
  - Construction of the transcripts from the sequence data
  - Hard to identify splicing. Representing mainly mature spliced transcripts.

Garber et al, Computational methods for transcriptome annotation and quantification using RNA-seq, Nature Methods 8(6) 2011.

Quantification of gene abundances

Gene quantification

- Count the total number of fragments for each gene

Exon quantification

- Counting the number of fragments for each exon.
- Splicing can be identified from fragments mapping junctions

Isoform quantification

- Estimates the abundance for each isoform.
- Requires mathematical models and complementary data such as a database of known splice variants.

Quantification of gene abundances

Multiple matches is common and caused by

- Genes with regions that are similar within the genome (e.g. recent paralogs)
- Repetitive regions in transcribed areas
- Errors in the reads caused by problems in the sequencing

Approaches to handle reads with multiple matches includes

- None – the reads are excluded from the analysis
- All – assign the reads to all its matching all regions
- Random – assign the read randomly to one of its machting regions
- Statistical predictive models

Counts

Samples

Genes

Gene	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8
ENSG00000000419	11	16	10	19	8	24	14	19
ENSG00000000457	28	22	17	21	18	15	21	13
ENSG00000001167	112	107	89	95	75	87	80	73
ENSG00000002016	1	1	1	1	0	0	0	0
ENSG00000002834	54	83	59	76	104	88	61	73
ENSG00000002919	1055	865	481	738	402	533	351	404
ENSG00000004700	79	162	125	124	45	118	99	146
ENSG00000004777	4	1	1	5	3	1	5	0
ENSG00000004779	256	327	227	234	278	364	214	276
ENSG00000004897	126	151	111	163	96	97	110	156
ENSG00000005175	49	59	57	37	26	47	52	53
ENSG00000005187	9	18	12	47	23	22	18	42
ENSG00000005339	895	404	422	985	393	188	613	635
ENSG00000005844	82	110	93	66	149	130	207	330
ENSG00000006007	180	173	147	196	132	186	148	208

Normalization

Data from different gene sand samples are not directly comparable due to systematic errors

Sequencing depth

The number of fragments sequenced in each sample differs

Gene lengths

Longer genes are represented by more RNA bases and will be represented by more fragments

Normalization

Sample quality

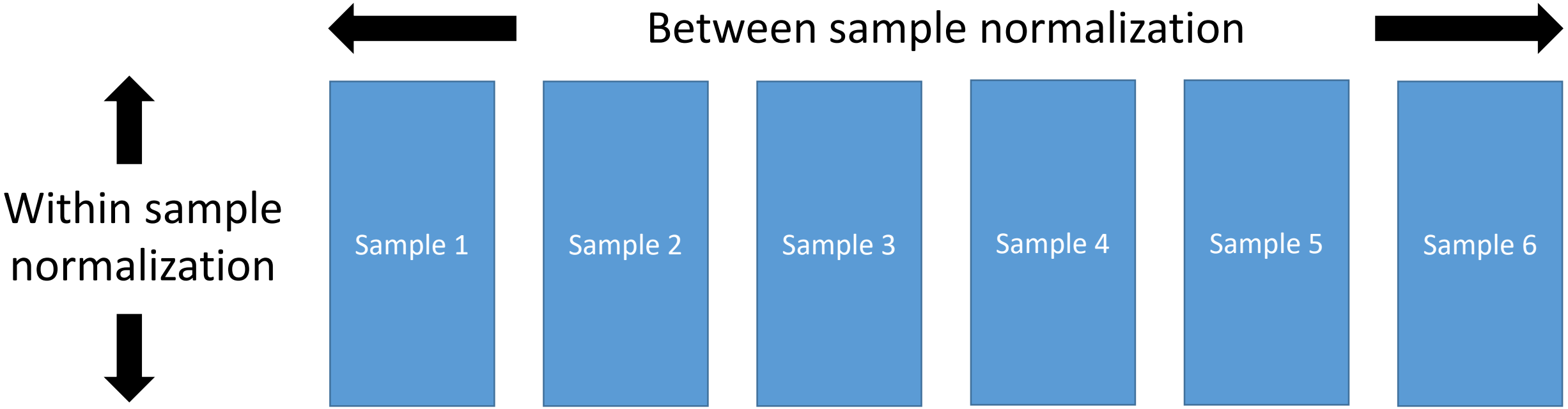
Errors affecting large part of a sample is introduced by quality issues in the sample preparation and sequencing.

Other systematic errors

- Sequencing errors
- Fragment distribution over the gene
- G/C-bias

Normalization aims to reduce the noise by identifying and removing systematic effects.

Normalization



Counts

Samples

Genes

Gene	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8
ENSG00000000419	11	16	10	19	8	24	14	19
ENSG00000000457	28	22	17	21	18	15	21	13
ENSG00000001167	112	107	89	95	75	87	80	73
ENSG00000002016	1	1	1	1	0	0	0	0
ENSG00000002834	54	83	59	76	104	88	61	73
ENSG00000002919	1055	865	481	738	402	533	351	404
ENSG00000004700	79	162	125	124	45	118	99	146
ENSG00000004777	4	1	1	5	3	1	5	0
ENSG00000004779	256	327	227	234	278	364	214	276
ENSG00000004897	126	151	111	163	96	97	110	156
ENSG00000005175	49	59	57	37	26	47	52	53
ENSG00000005187	9	18	12	47	23	22	18	42
ENSG00000005339	895	404	422	985	393	188	613	635
ENSG00000005844	82	110	93	66	149	130	207	330
ENSG00000006007	180	173	147	196	132	186	148	208

Counts

Samples

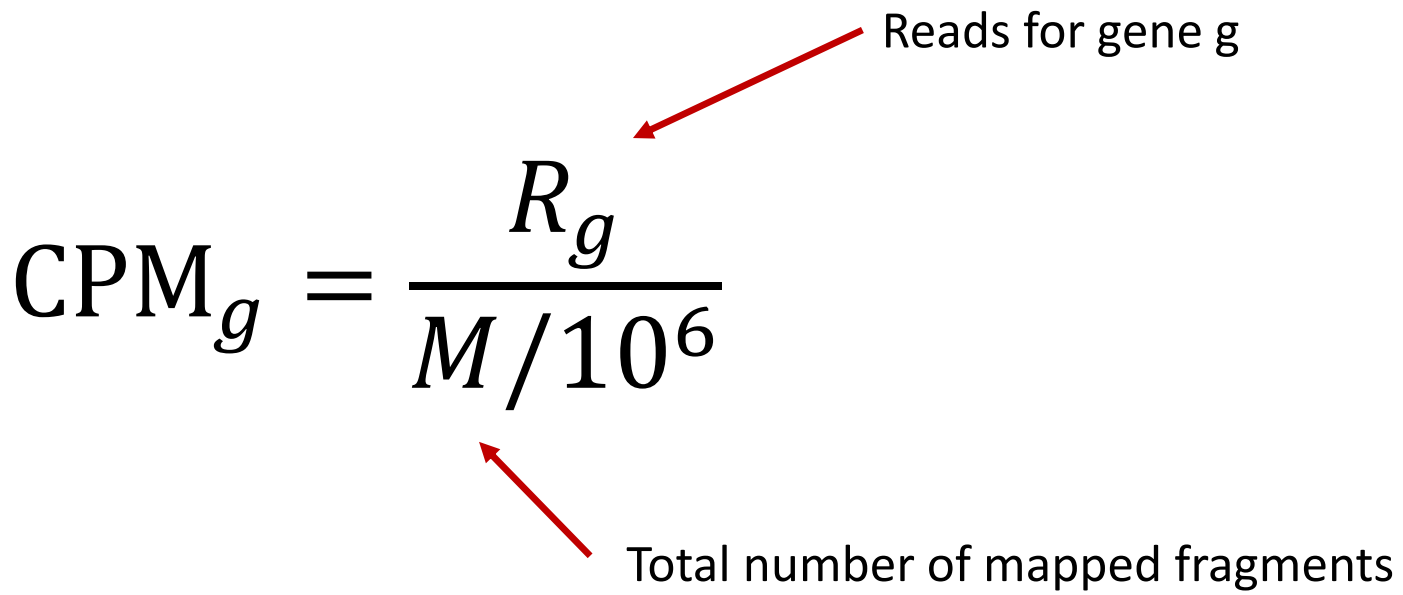
Genes

Gene	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8
ENSG00000000419	11	16	10	19	8	24	14	19
ENSG00000000457	28	22	17	21	18	15	21	13
ENSG00000001167	112	107	89	95	75	87	80	73
ENSG00000002016	1	1	1	1	0	0	0	0
ENSG00000002834	54	83	59	76	104	88	61	73
ENSG00000002919	1055	865	481	738	402	533	351	404
ENSG00000004700	79	162	125	124	45	118	99	146
ENSG00000004777	4	1	1	5	3	1	5	0
ENSG00000004779	256	327	227	234	278	364	214	276
ENSG00000004897	126	151	111	163	96	97	110	156
ENSG00000005175	49	59	57	37	26	47	52	53
ENSG00000005187	9	18	12	47	23	22	18	42
ENSG00000005339	895	404	422	985	393	188	613	635
ENSG00000005844	82	110	93	66	149	130	207	330
ENSG00000006007	180	173	147	196	132	186	148	208

Normalization

CPM – Counts per million mapped reads

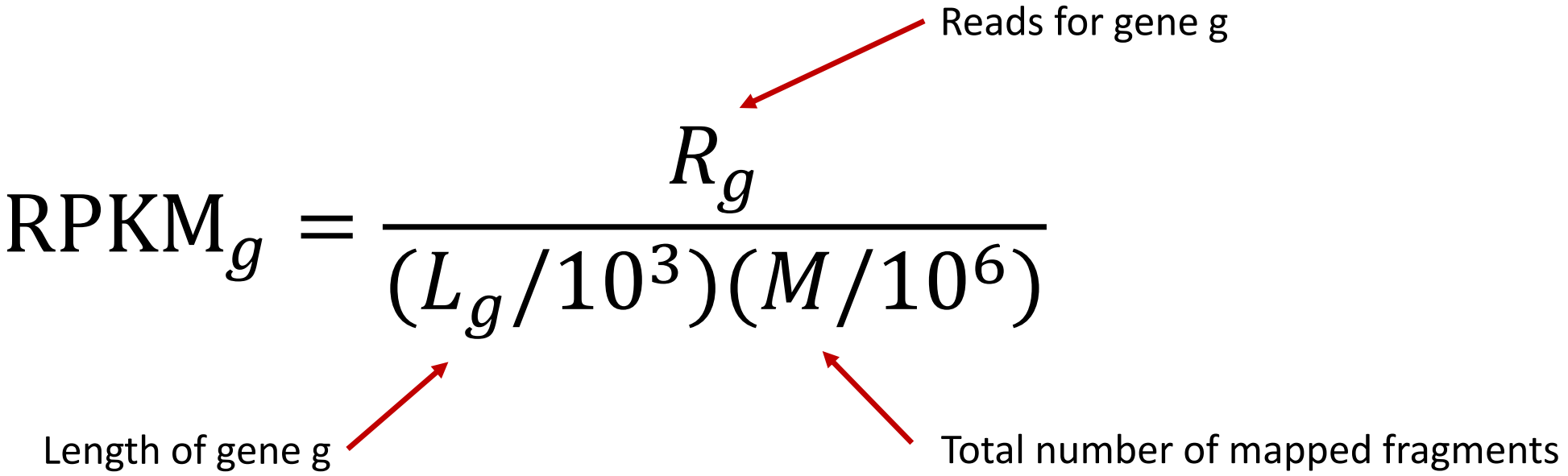
Estimates the relative abundance of a gene in relation to the total sequencing depth (total number of successfully mapped reads).



Normalization

RPKM – Reads per kilobase per million

Estimates the relative abundance for each gene in relation to its length and total sequencing depth (total number of successfully mapped reads).



Normalization

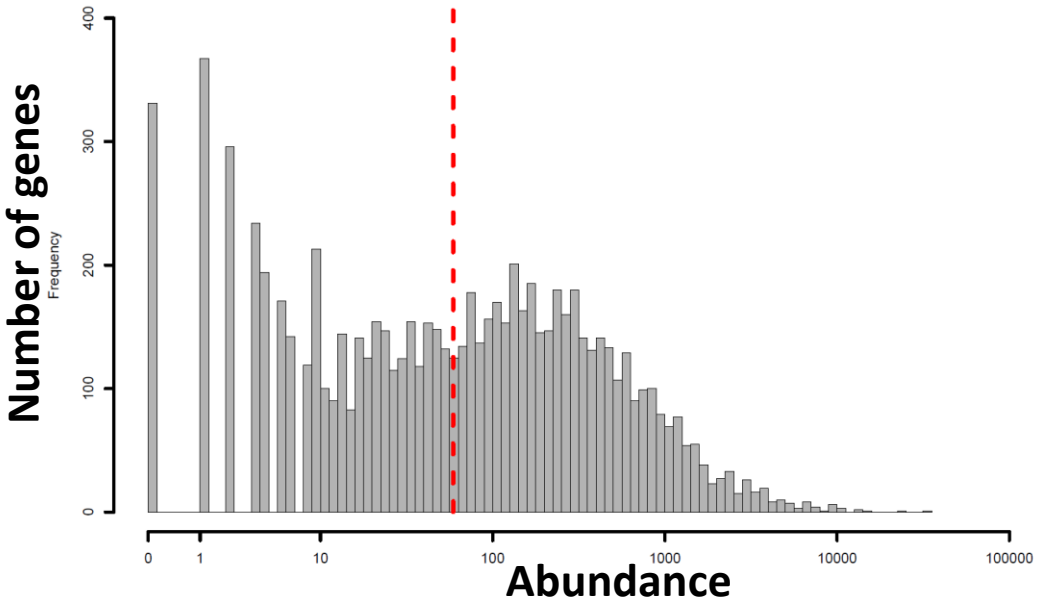
Normalization based on the total number of counts may introduce systematic errors

- The total RNA pool is often dominated by few highly expressed genes (e.g. actin). The crude rule of thumb is that “50% of the total number of RNA transcripts from 5% of the genes”.
- Variation in expression of the high abundant genes can therefore have big impact on the normalization.
- Even more important, systematic changes in their expression between the experimental conditions can lead to incorrect results.

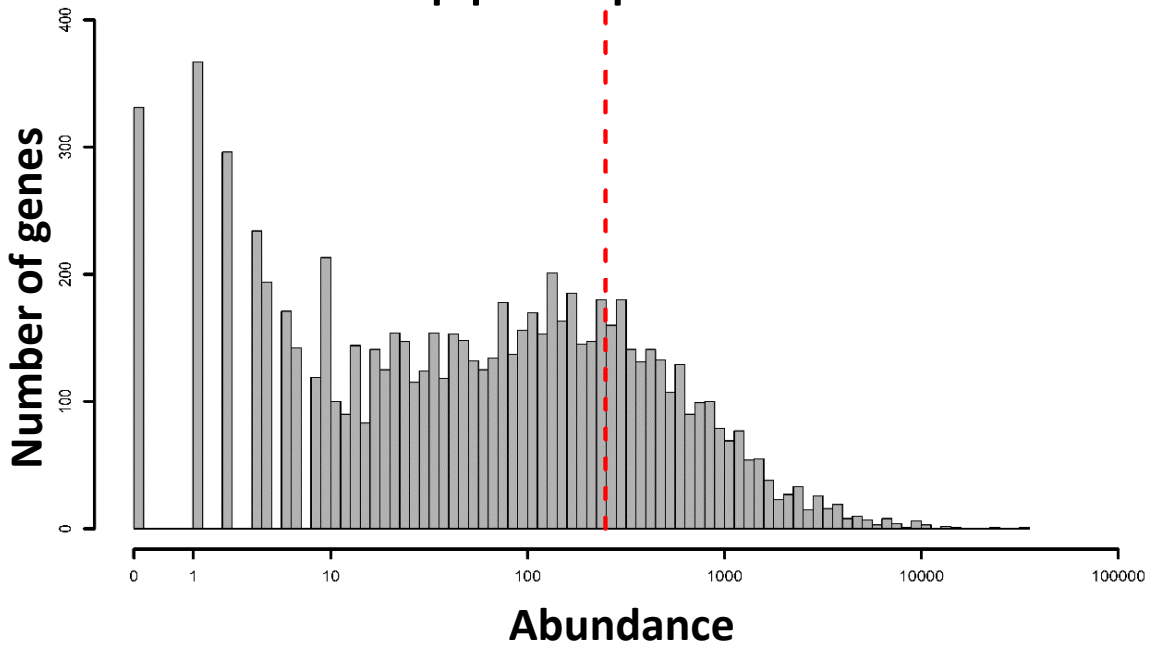
Normalization – robust alternatives

There are several robust alternative that can be used instead of the total number of mapped reads:

Median



Upper quartile



It is also possible to normalize in relation to a set of reference genes.

RNA-seq data is semiquantitative

Counts

	Sample 1	**Sample 2**
Gene 1	10	10
Gene 2	10	10
Gene 3	10	10
Gene 4	10	10
Gene 5	10	10
Total	**50**	**50**

Proportions

	Sample 1	**Sample 2**
Gene 1	0.2	0.2
Gene 2	0.2	0.2
Gene 3	0.2	0.2
Gene 4	0.2	0.2
Gene 5	0.2	0.2
Total	**50**	**50**

RNA-seq data is semiquantitative(!!)

Counts

	Sample 1	Sample 2	
Gene 1	10	30	Up-regulated gene
Gene 2	10	5	
Gene 3	10	5	
Gene 4	10	5	
Gene 5	10	5	
Total	**50**	**50**	

Proportions

	Sample 1	Sample 2	
Gene 1	0.2	0.6	
Gene 2	0.2	0.1	
Gene 3	0.2	0.1	
Gene 4	0.2	0.1	
Gene 5	0.2	0.1	
Total	**50**	**50**	

Normalization - TMM

Trimmed mean of M-values

Let Y_{ij} be the number of counts for gene i and sample j and N_j be the total number of reads for sample j. Select a sample (sample r) as the reference.

For a sample j define

$$M_i = \log_2 \frac{Y_{ij}}{N_j} \Big/ \frac{Y_{ir}}{N_r} = \log_2 \frac{Y_{ij}}{N_j} - \log_2 \frac{Y_{ir}}{N_r}$$

$$A_i = \log_2 \sqrt{\frac{Y_{ij}}{N_j} \times \frac{Y_{ir}}{N_r}} = \frac{1}{2}\left(\log_2 \frac{Y_{ij}}{N_j} + \log_2 \frac{Y_{ir}}{N_r}\right)$$

Normalization - TMM

TMM assumes that only a small proportion of the genes are differentially expressed. This means that M_i is close to zero the vast majority of the genes.

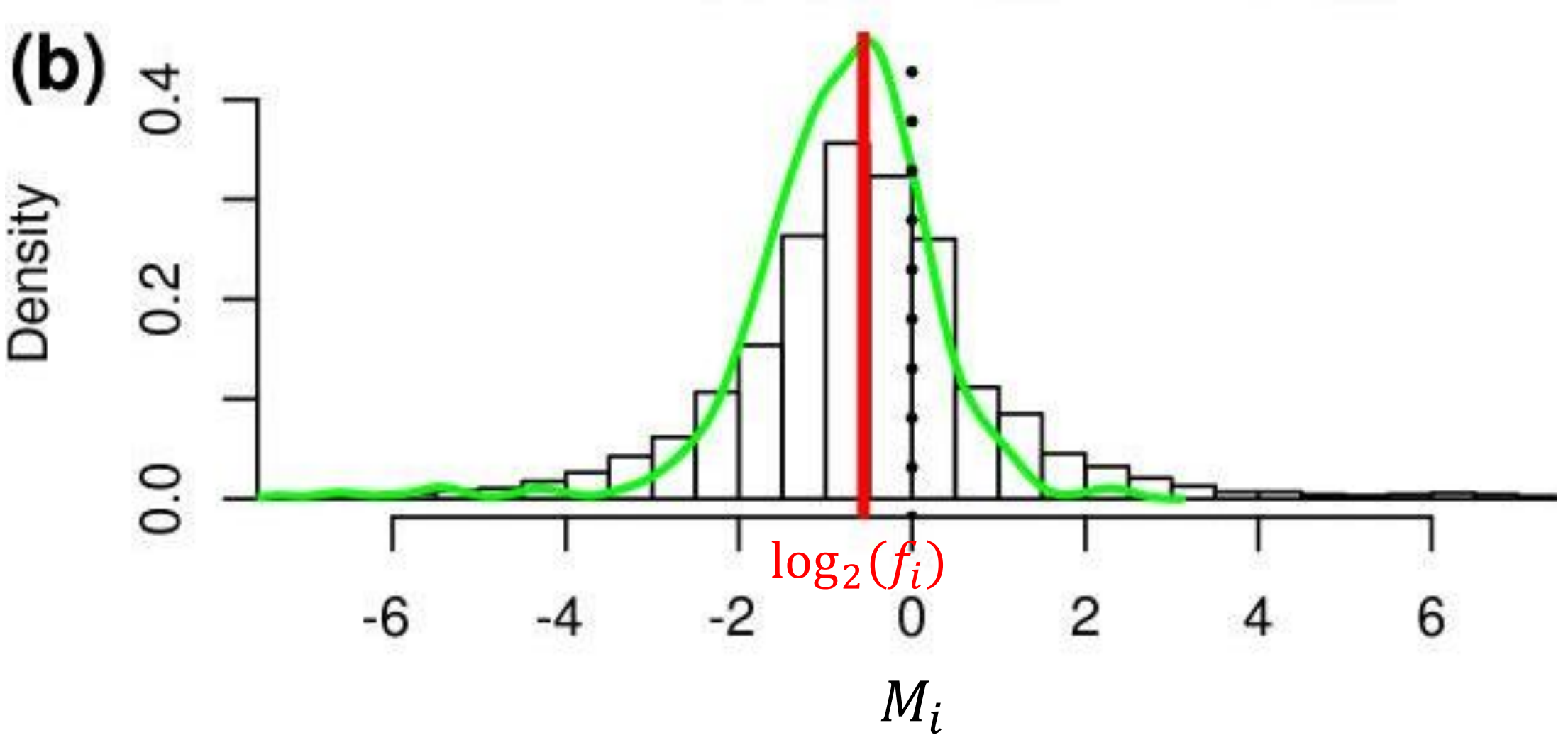
An adjustment factor for sample *j*, f_j is calculated as

$$\log_2(f_j) = \text{TrimmedMean}(M_i)$$

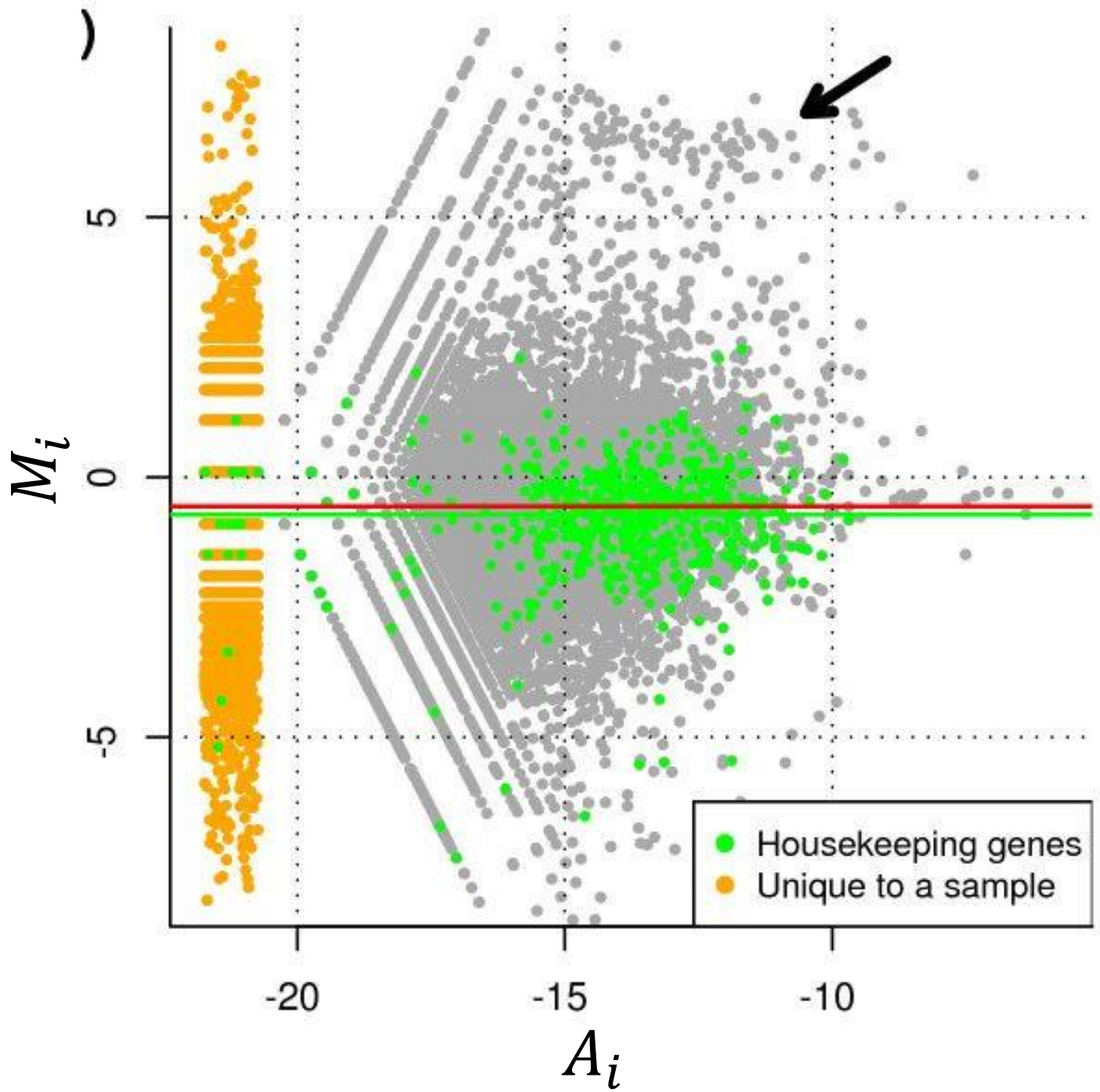
In the trimmed mean, the largest 30% and the lowest 30% of the M_i values are not included. This mean that only the middle 40% of the values are included in the mean.

A normalization factor is then calculated as $\widetilde{N}_j = f_j \times N_j$.

Total number of reads for sample j



Robinson & Oshlak, A scaling normalization method for differential expression analysis of RNA-seq data, Genome Biology 11 2010



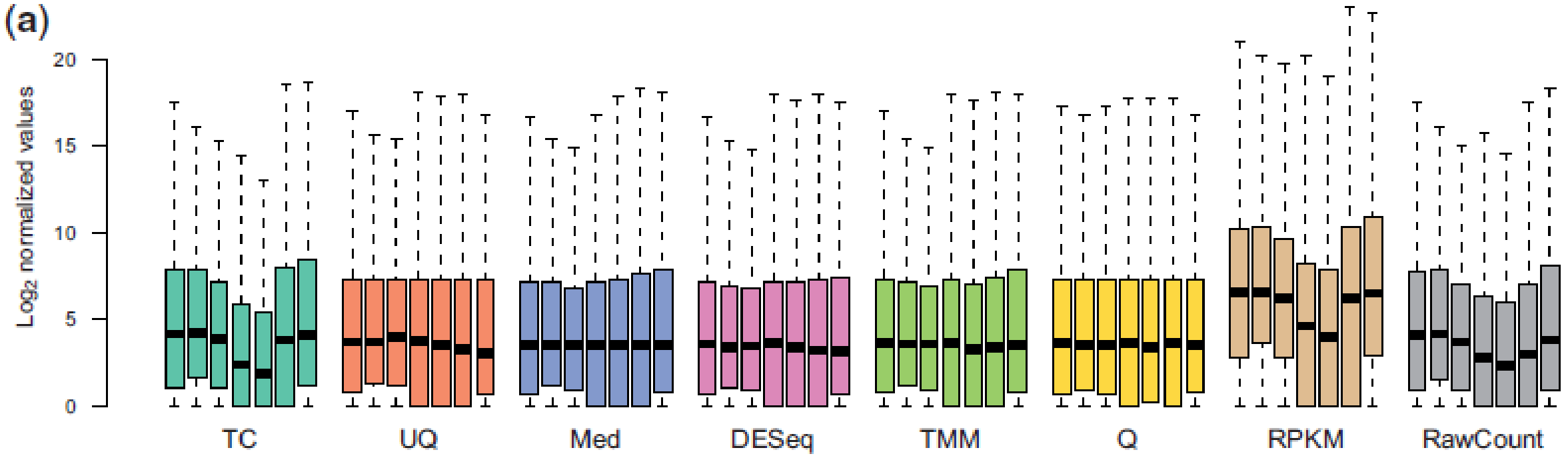
Robinson & Oshlak, A scaling normalization method for differential expression analysis of RNA-seq data, Genome Biology 11 2010

Normalization – two general remarks

Remark 1: Within sample normalization is often not be necessary! Systematic effects that are similar between samples may be canceled out when estimating the difference in expression.

Remark 2: Many methods for identification of differentially expressed genes works with counts. Some normalization methods modifies data in way that the counts are lost. It is therefore important to use a normalization method that is appropriate for the downstream statistical analysis.

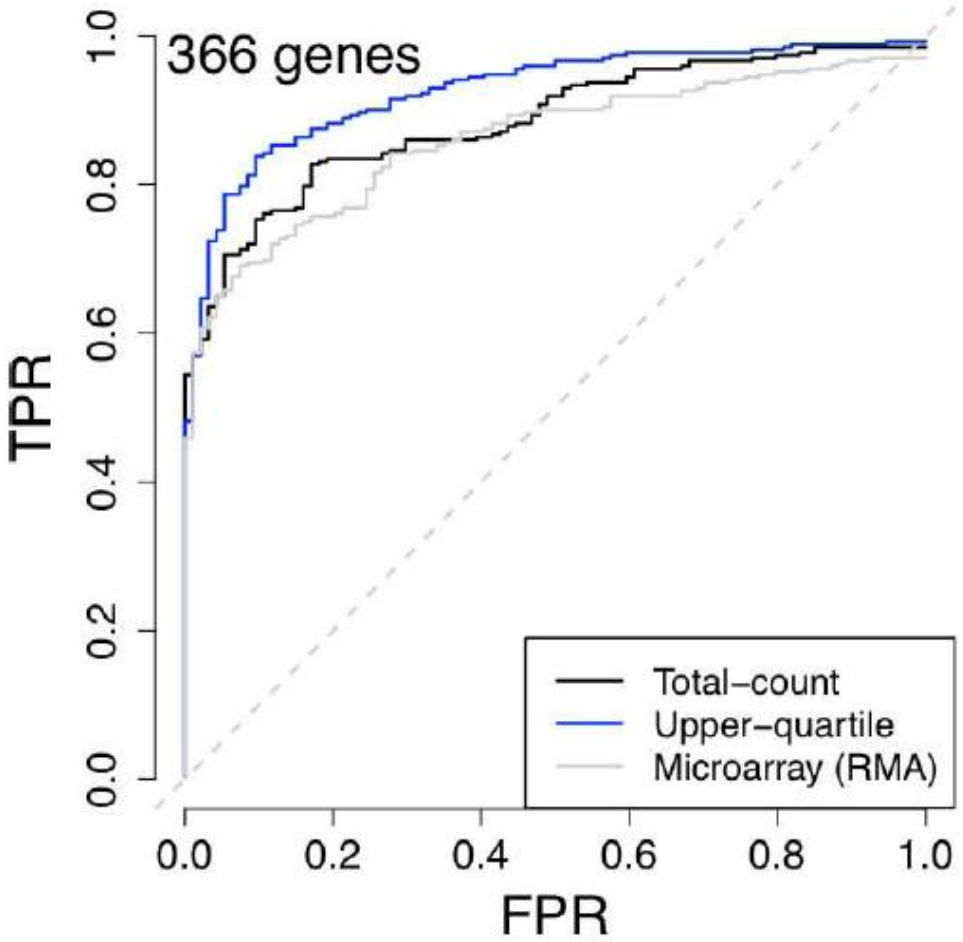
Normalization



TC=Total counts, UQ=Upper quartile, Med=Median, DESeq=Method implemented in DESeq, TMM=Trimmed Mean of M values, Q=Quantile, RPKM=Reads Per Kilobase per Million mapped reads.

Dillies et al, A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis, Briefings in Bioinformatics 14(6) 2013.

Improper normalization can reduce performance



- Receiver operating characteristics (ROC) curve on MAQC data.
- RNA sequencing of two reference pools of RNA.
- Comparison between RNA-seq and qPCR (golden standard).
- Upper-quartile normalization significantly improves identification of differentially expressed genes.

Bullard et al, Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments, BMC Bioinformatics 11(94) 2010.

Identification of differentially expressed genes

Aim: Identify genes with altered gene expression level between experimental conditions.

Strategy: Examine and analyze each gene separately.

Advantages

- Straight-forward and limits complexity of the analysis
- Provides gene-level information

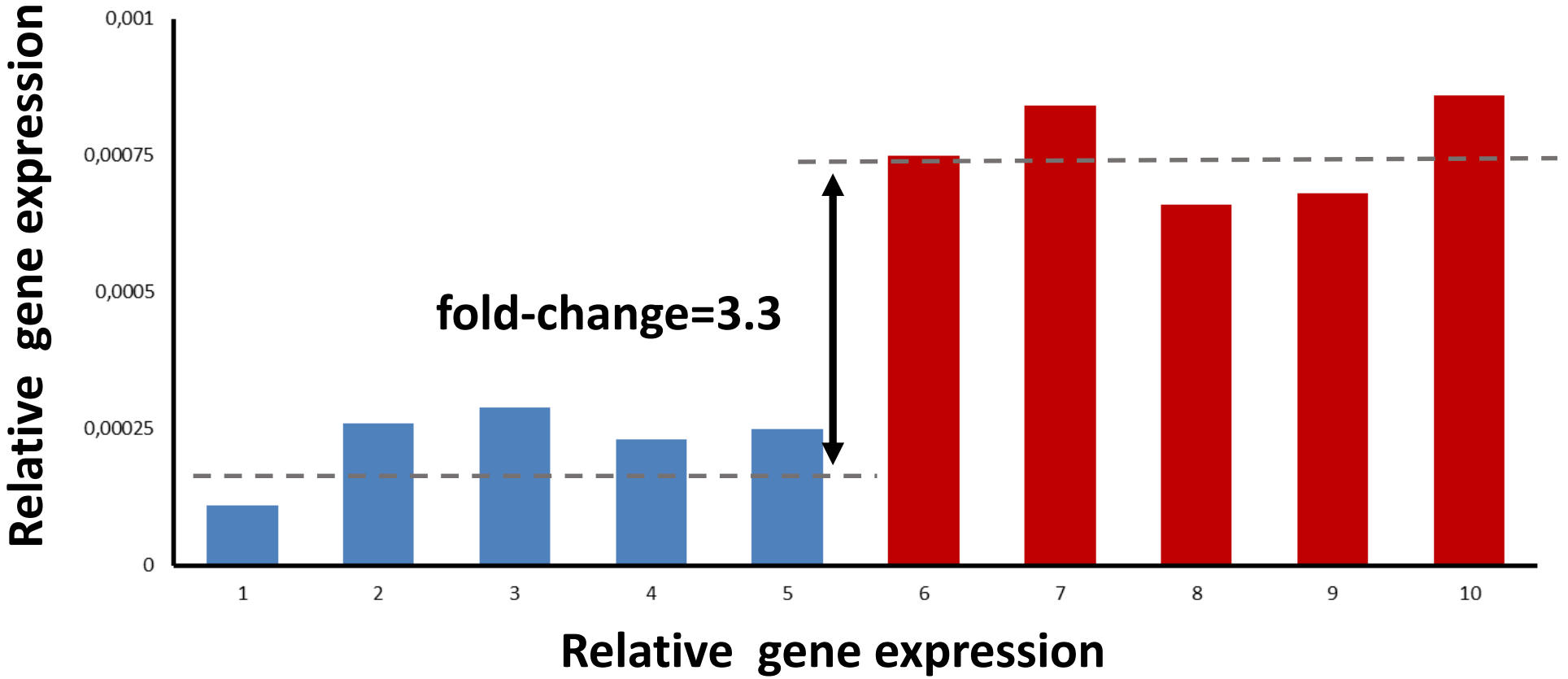
Disadvantages

- Many genes implies many comparisons
- Information for each gene is limited (e.g. few samples)

Identification of differentially expressed genes

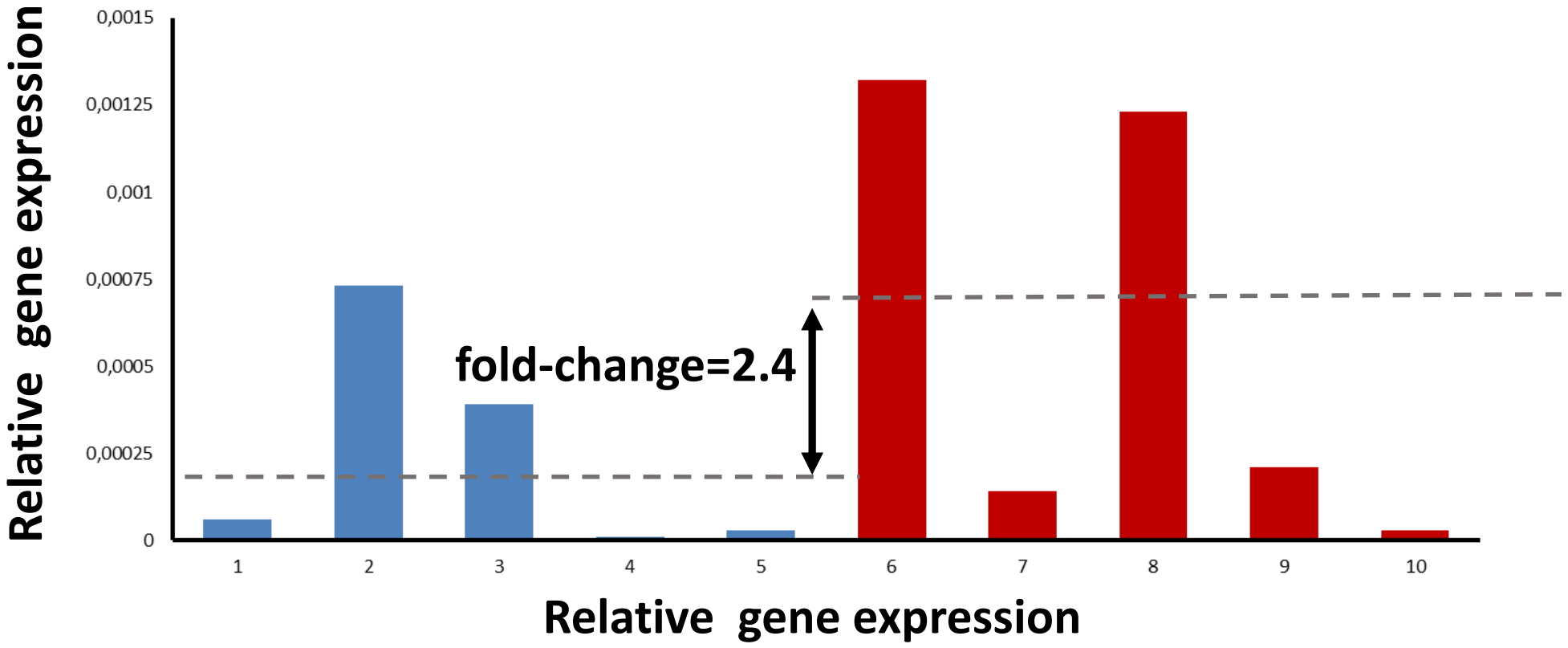
When can we considered the expression to be altered?

Example 1:



Identification of differentially expressed genes

Example 2:



The difference needs to be interpreted with respect to the variability!

Identification of differentially expressed genes

Technical variability

- Sample quality
- Sample preparation, e.g. RNA extraction and cDNA preparation
- Sequencing errors and quality, errors from the mapping

Biological variability

- Genotype and other effects on individual level
- Effects introduced by sampling and/or experimental setup
- Transcription is a stochastic process with spatial and temporal differences

Identification of differentially expressed genes

Challenges in the statistical analysis

1. Data consists of count (discrete)
2. Large number of genes
3. Typically few replicates (replicates << genes)
4. Large gene-specific variability

Statistical approach includes

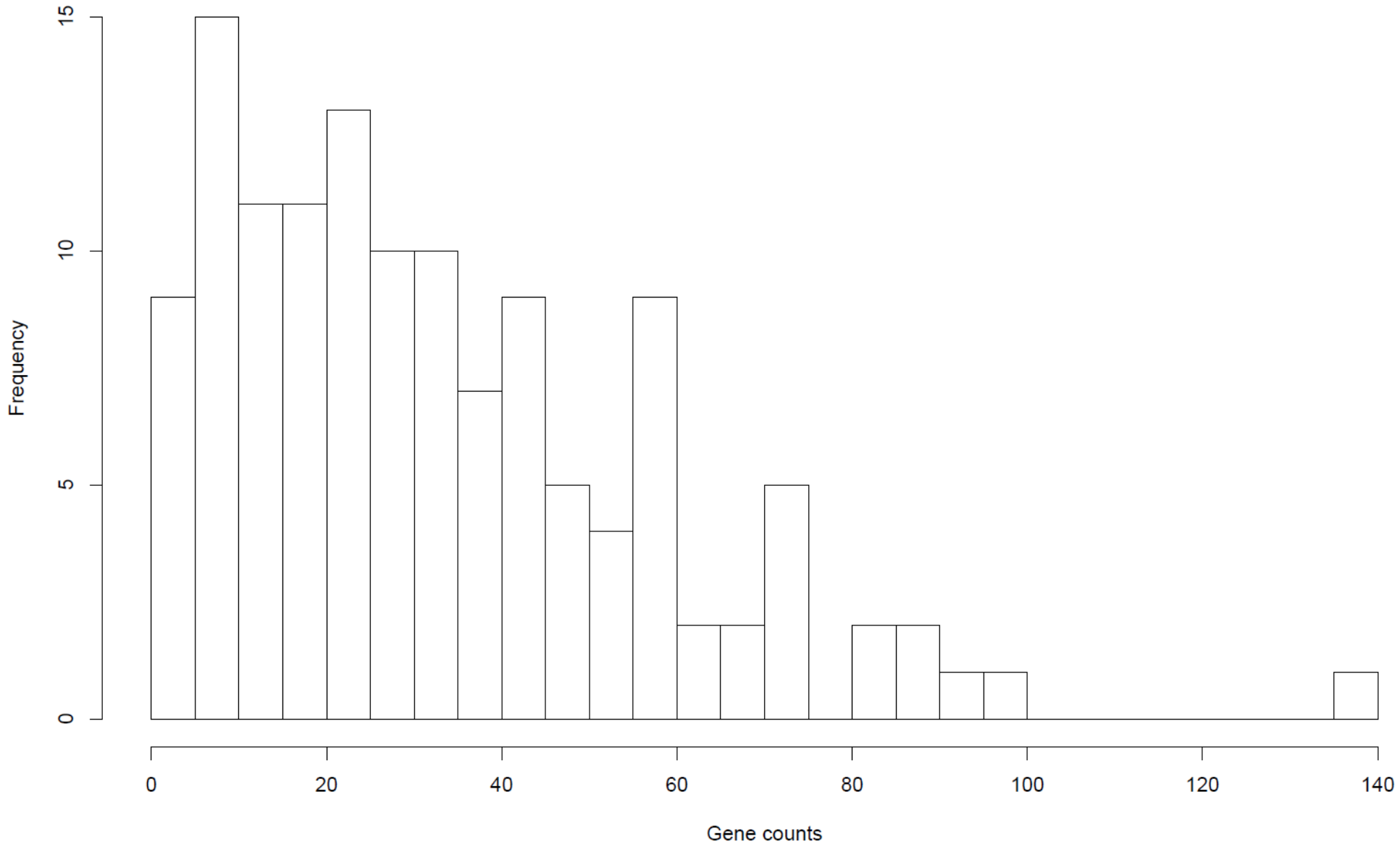
- A model describing the data and its variance structure
- A test for identification of differentially expressed genes

Three main approaches

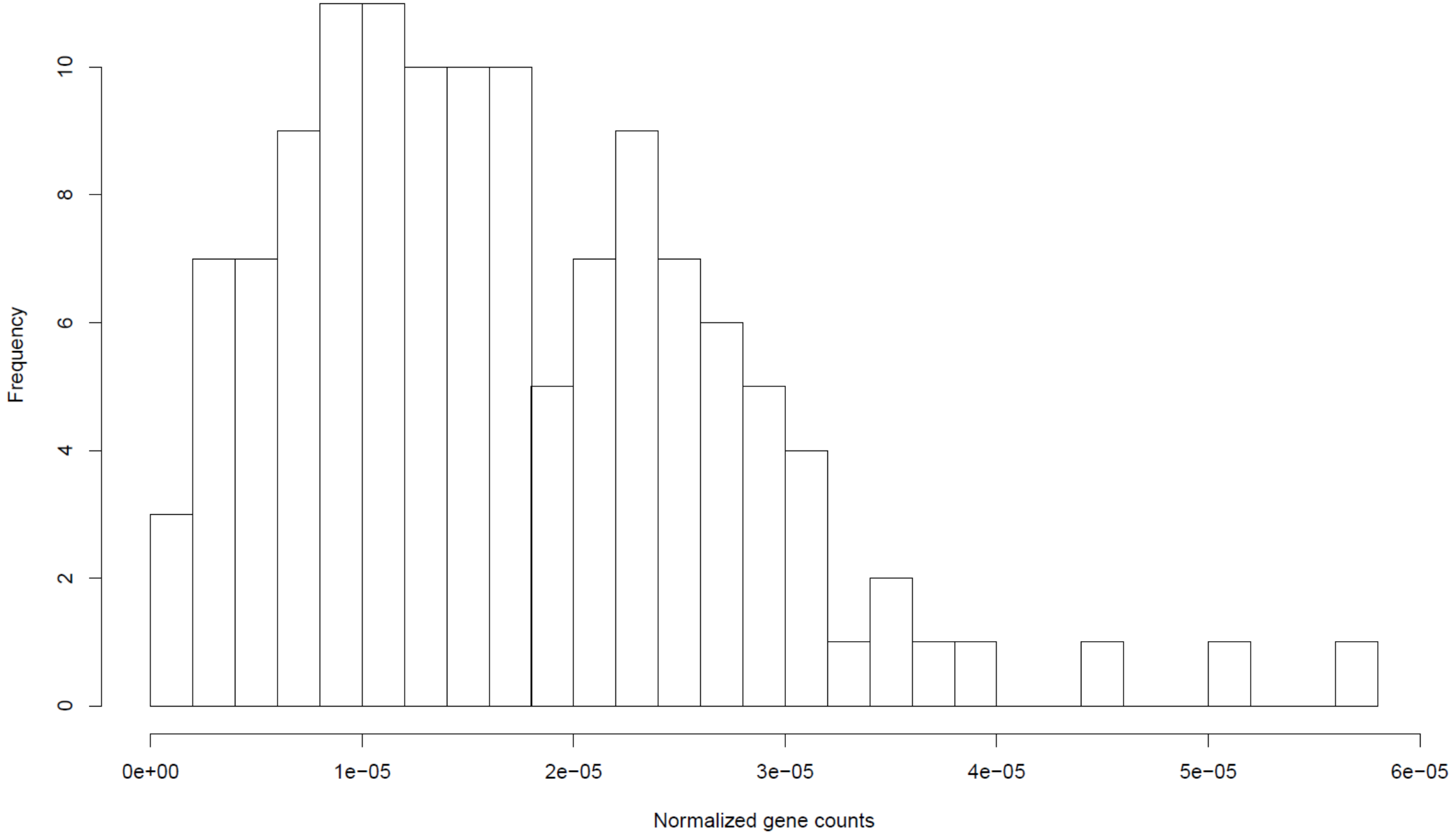
Methods based on normal assumptions

- Use a variance-stabilizing transformation to make the distribution of the data a close to a normal distribution as possible.
- Common transformation are the square-root and log transformations.
- Normally distributed data enabels the use of more flexible statistical tools such as linear models (the topic of the next lecture!)

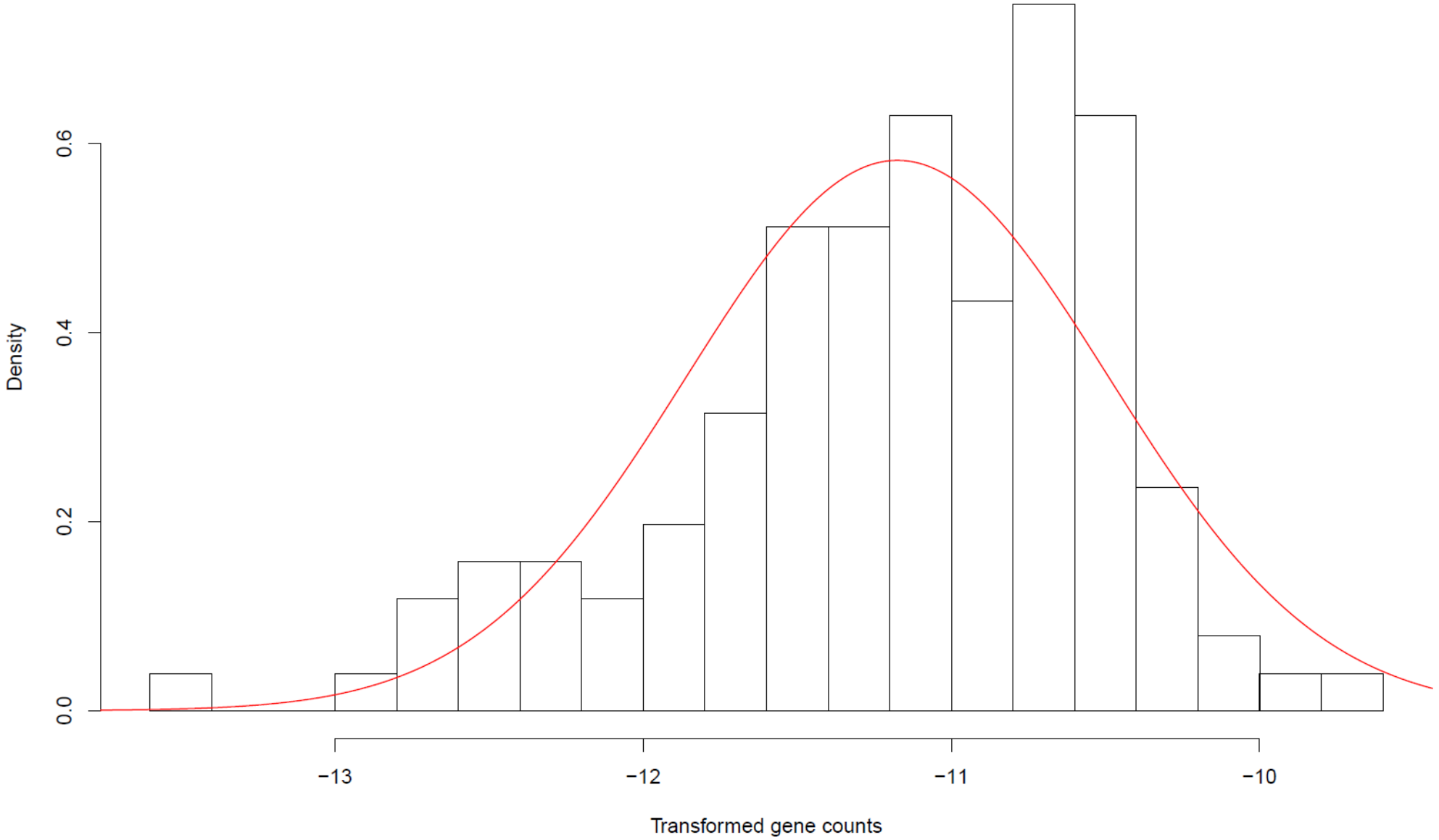
RNA-seq data from 129 samples for the gene FGR



After normalization



After log-transformation



Three main approaches

Methods based on non-parametric methods

- Makes no explicit distributional assumptions
- Can be sensitive to ties, i.e. values that are identical between samples. This can occur frequently for genes with low abundance and thus low counts.
- Common non-parametric methods include the Wilcoxon-Mann-Whitney test (Wilcoxon rank sum), the Kruskal-Wallis test and Fisher's exact test.

Three main approaches

Methods based on count distributions

- Assumes specific count distribution to the data. Common distributions include
  - Poisson distribution
  - Binomial distribution
  - Negative binomial distribution
- Often overdispersed count-models are required. These models add extra variability than present in the standard models.

Count models for RNA-seq data

Statistical fact 1

Count data has a mean-variance dependence.

Statistical fact 2

The variance of count data can be divided into two parts

$$\text{Total variance} = \text{Sampling variance} + \text{Overdispersion}$$

Random selection of DNA fragments

Dominated by biological variation

Count models for RNA-seq data

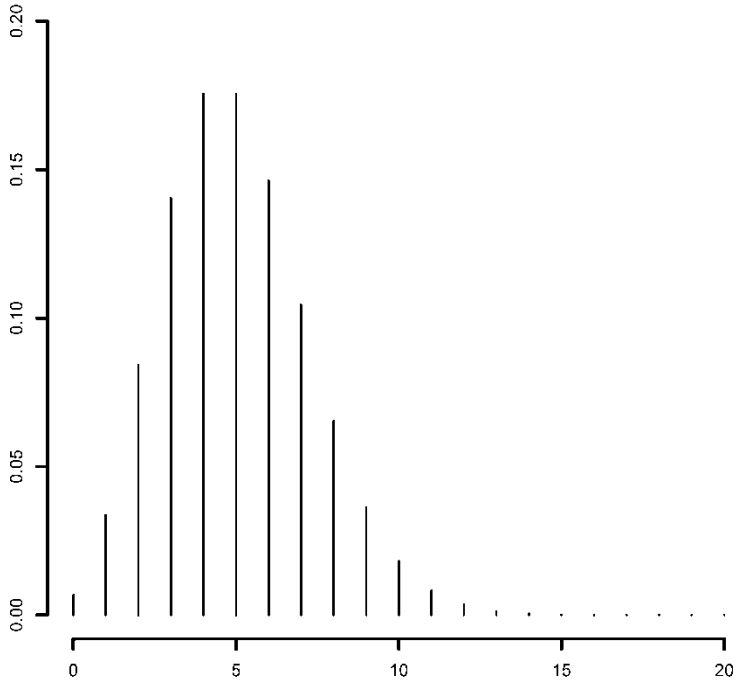
The Poisson model

- A basic model for describing counts
- The variability depends on the expression level

$$X_{gi} = \text{counts of gene g in sample i}$$

$$\text{Exp}[X_{gi}] = \mu_g$$

$$\text{Var}[X_{gi}] = \mu_g$$



The Poisson model assumes that two genes with the same expression level have the exact same variability!

Count models for RNA-seq data

The negative binomial model

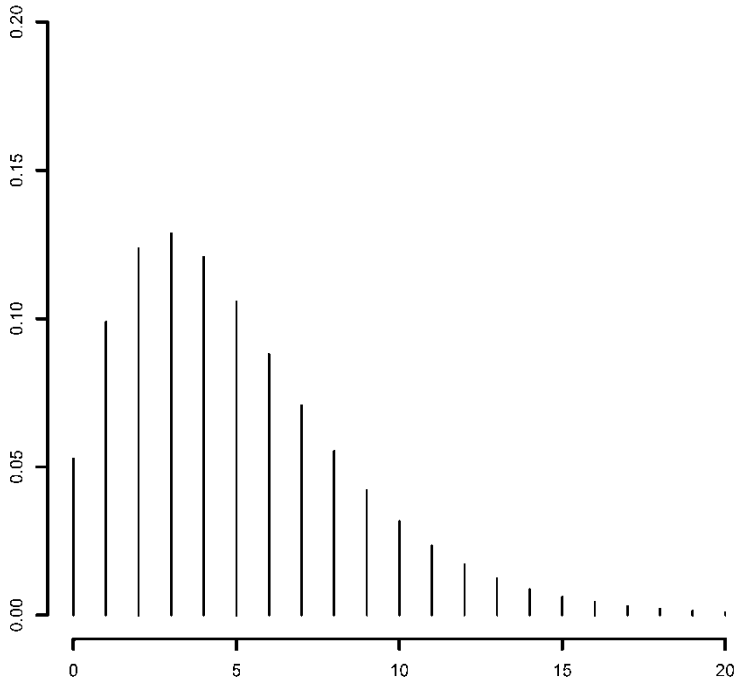
- Provides more variability than the Poisson model
- Variance divided into two parts
  1. Poisson noise from random sampling of DNA fragments
  2. Gene-specific technical and biological noise - *overdispersion*

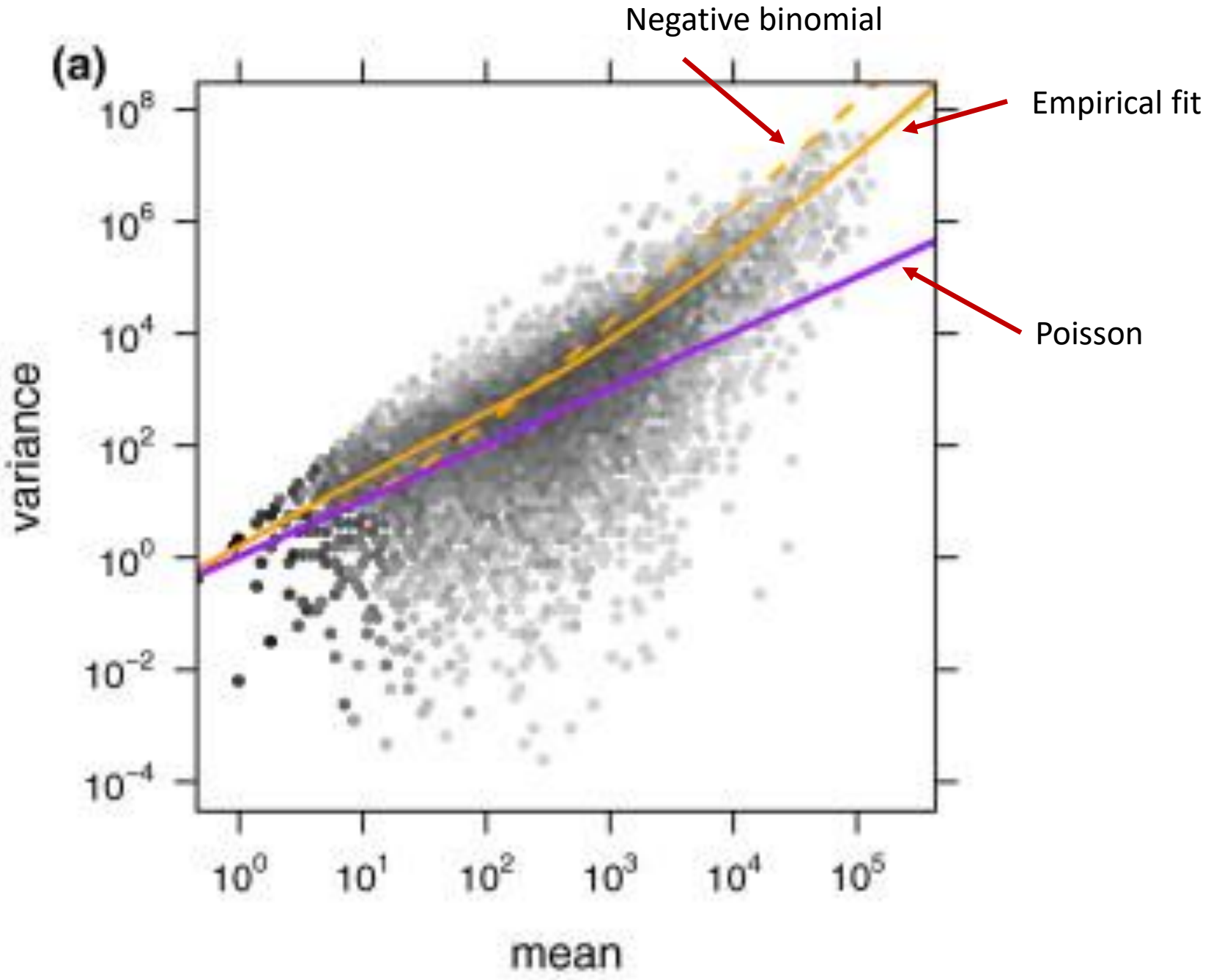
$$X_{gi} = \text{countes of gene g in sample i}$$

$$\text{Exp}[X_{gi}] = \mu_g$$

$$\text{Var}[X_{gi}] = \mu_g + \varphi_g \mu_g^2$$

Gene-specific overdispersion

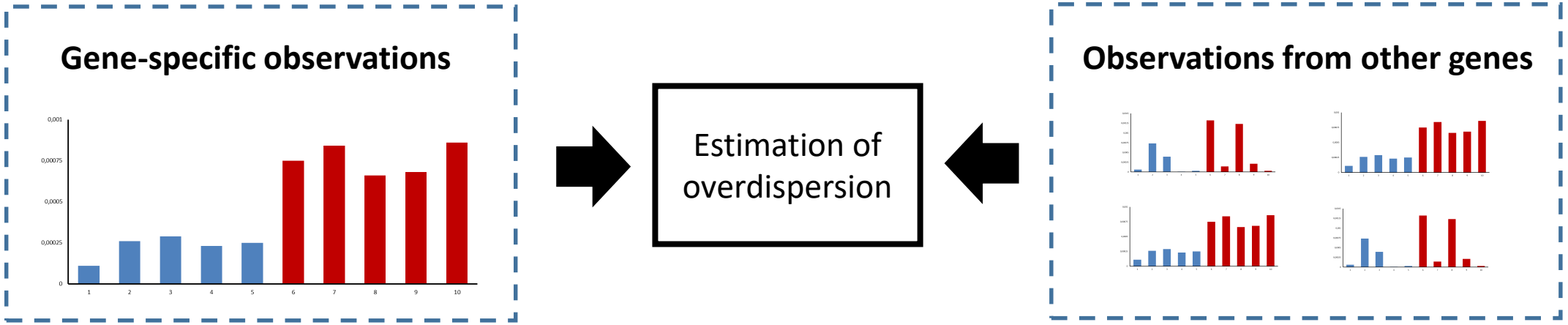




Anders & Huber, Differential expression analysis for sequence count data, Genome Biology 11:R106, 2010.

Count models for RNA-seq data

- Estimation of the gene-specific overdispersion is hard when few samples are available.
- Sharing of information between genes generates more robust estimates.



- This is called a shrinkage model. Often implemented by using Bayesian statistics.

Identification of differentially expressed genes

- The differentially expression is assessed fro each gene using a hypothesis test.

For each gene g,

H_0: Gene g is not differentially expressed

H_A: Gene g is differentially expressed

- The p-value provide information about rejecting H_0.
- Multiple testing: many tests – many p-values. More on this later!

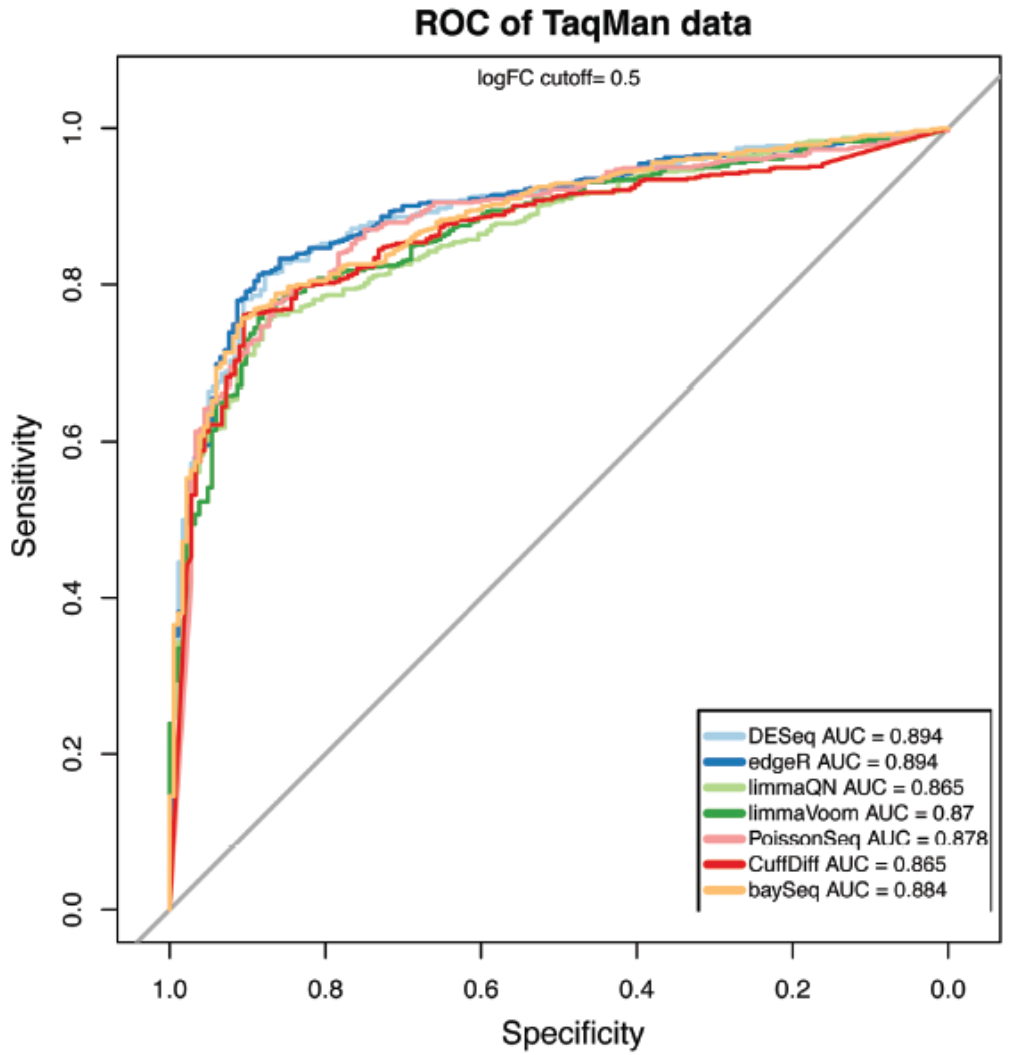
edgeR and DESeq

- Two R-packages for analysis of RNA-seq data
- Implements
  - Normalization (TMM and REL)
  - A statistical models based on negative binomial distributions
  - Robust estimation of gene-specific overdispersion
- Can handle many experimental designs, including comparisons of groups, regression and ANOVA-like analyses (GLM-based).
- Easy-to-use, well-tested and have reliable performance for most RNA-seq datasets.
- http://bioconductor.org/packages/release/bioc/html/edgeR.html
- https://bioconductor.org/packages/release/bioc/html/DESeq2.html

Methods for RNA-seq data

Method	Model	Sharing of overdisp between genes?	Multifactorial designs	Reference
Fisher's exact test	Hypergeometric	No	No	Bullard et al 2010
PoissonSeq	Overdisp. poisson	No	No	Li et al 2012
baySeq	Negative binomial	Yes, hierarchical Bayesian	No	Hardcastle et al 2010
edgeR	Negative binomial	Yes	Yes	Robinson et al 2010, McCarthy et al 2012
DESeq2	Negative binomial	Yes, empirical Bayes	Yes	Love et al 2014
voom	Approx. normal	Yes, empirical Bayes	Yes	Law et al 2014

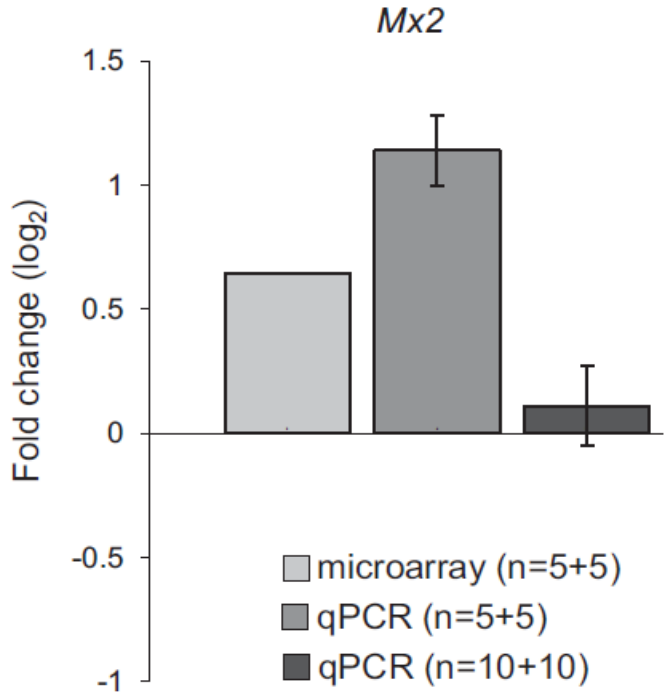
Which method is best?



Rapaport et al, Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data, Genome Biology 14:R95 2013.

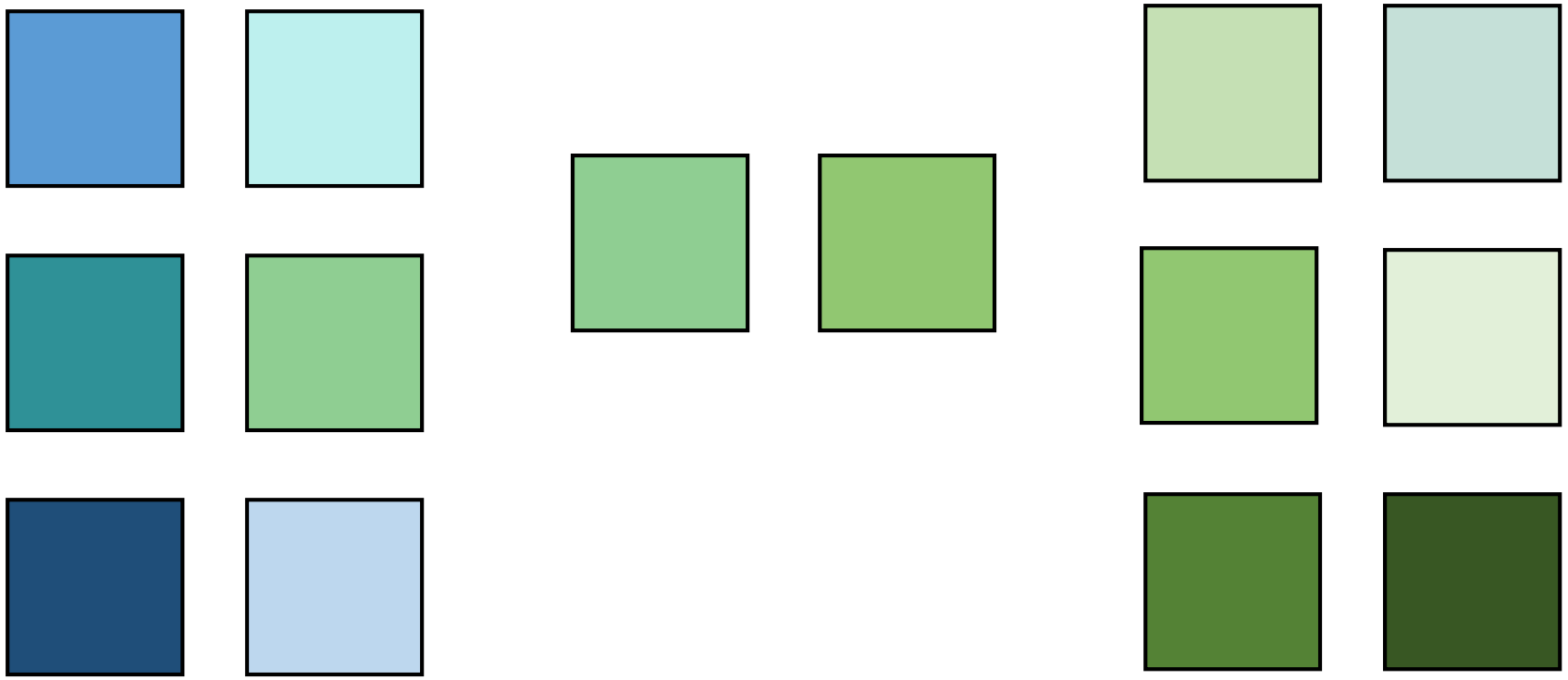
Interpretation of gene lists

- Gene lists should be interpreted with care
  - P-values are often biased due to incorrect model assumptions
- False positives are common!
- PCR verification of selected genes is highly recommended. Preferably on a independent and larger set of samples.



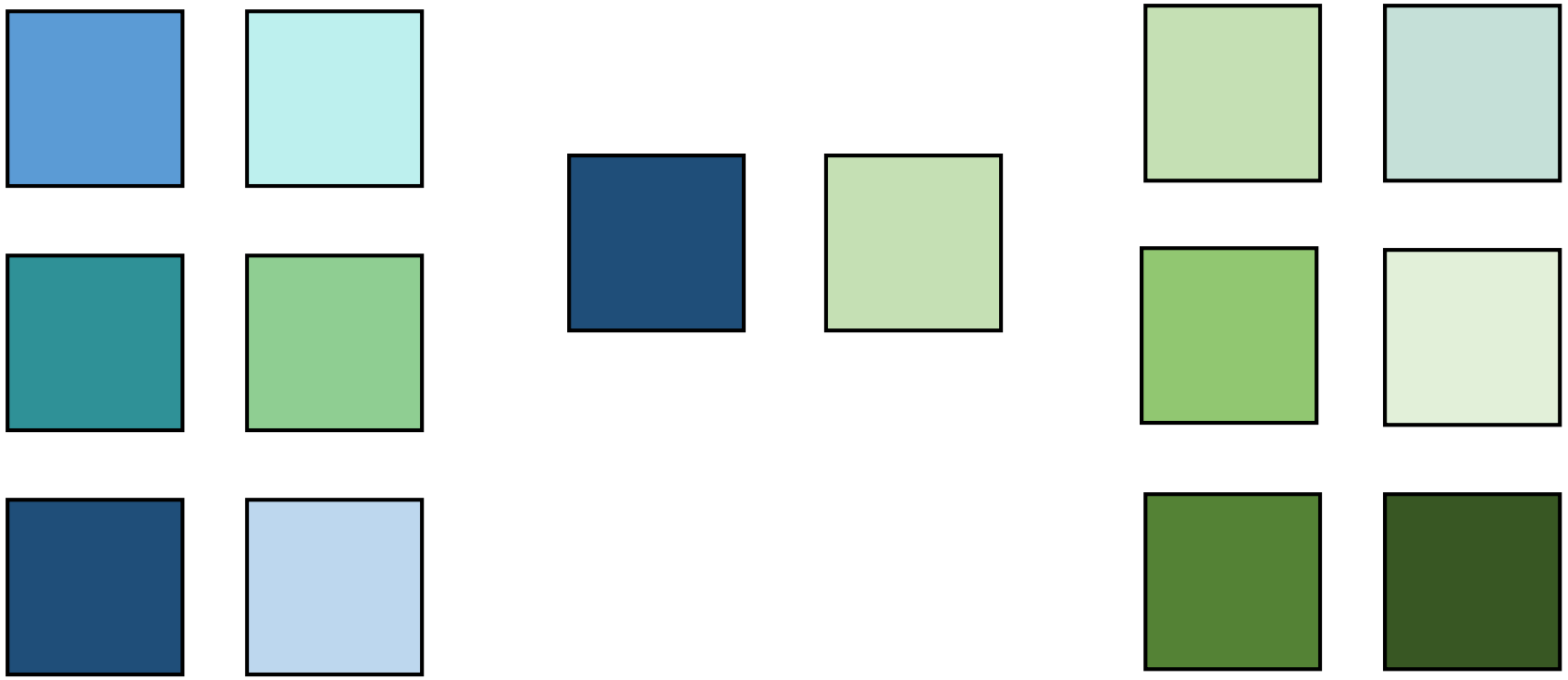
Experimental design of RNA-seq experiments

Biological replication is essential in high-throughput experiments!



Experimental design of RNA-seq experiments

Biological replication is essential in high-throughput experiments!



Experimental design

- Few biological replicates are often not representative for the entire population!
- How many replicates do you need in RNA-seq?
  - The more replicates you have, the higher is the power to identify differentially expressed genes!
  - User more replicates if you have
    - High biological and/or technical variability
    - Expect small changes in gene expression
    - Expect that samples may fail or have reduced quality.
- Completely unreplicated experiments is never (ever!) recommended!

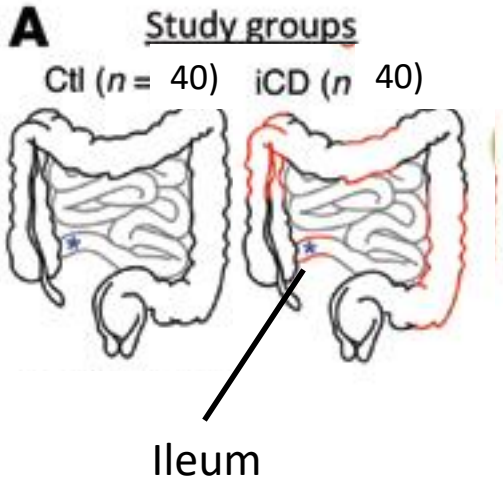
Experimental design

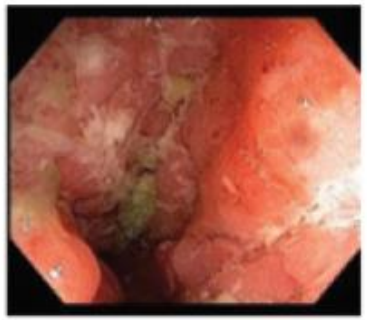
- Biological replicates should be prioritized over technical replicates
  - Technical replicates can not replace biological replicates
  - Biological variability is often substantially higher than the technical variability
  - The technical variability is often unnecessary to estimate separately. Only the total variability is of interest.

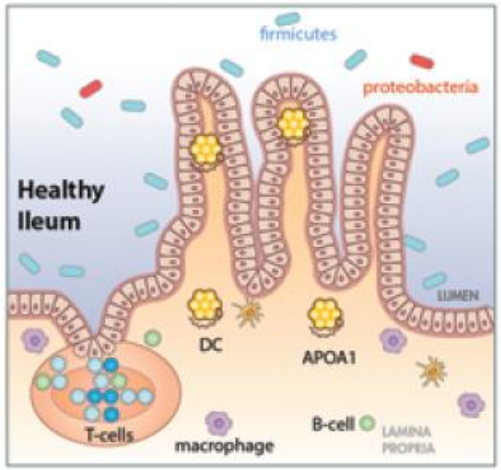
- Pooling of biological replicates should be avoided as a mean to reduce variability
  - Destroys information about the biological variability

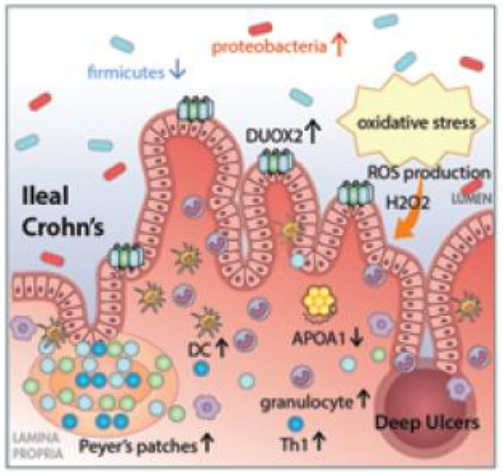
Introduction to computer exercise 3

Gene expression analysis of patients with and without irritated bowel disease (IBD). The aim is to understand the mechanisms behind Crohn's disease among children.




iCD-DU





Introduction to computer exercise 3

Samples

Genes

	SRR1782694	SRR1782695	SRR1782703	SRR1782712	SRR1782715	SRR1782717	SRR1782718	SRR1782719
ENSG00000000003	28384	21884	18484	26677	24318	17669	21352	16472
ENSG00000000005	50	0	50	150	50	100	0	49
ENSG00000000419	14900	23534	18651	17129	18633	16759	19626	16194
ENSG00000000457	14965	20808	19651	14428	18381	20800	15498	13258
ENSG00000000460	6488	9787	8693	6965	8047	9164	8604	6380
ENSG00000000938	2953	16721	15111	7177	14184	26891	9213	12585
ENSG00000000971	55143	80409	51342	50532	72553	102001	41232	34952
ENSG00000001036	60404	41600	35658	59473	49857	34174	38756	40068

patient.id	tissue	Sex	age.at.diagnosis	paris.age	diagnosis
SRR1782694	tissue: Ileal biopsy	Male	10.33	A1b	Not IBD
SRR1782695	tissue: Ileal biopsy	Male	10.83	A1b	Not IBD
SRR1782703	tissue: Ileal biopsy	Female	15	A1b	Not IBD
SRR1782712	tissue: Ileal biopsy	Female	16.92	A1b	Not IBD
SRR1782687	tissue: Ileal biopsy	Female	12.5	A1b	CD
SRR1782697	tissue: Ileal biopsy	Male	9.25	A1a	CD
SRR1782708	tissue: Ileal biopsy	Male	4.5	A1a	CD
SRR1782709	tissue: Ileal biopsy	Female	6.33	A1a	CD
SRR1782716	tissue: Ileal biopsy	Female	8.67	A1a	CD

Summary of todays lecture

- Analysis of RNA-seq data consist of three main steps
  1. Quantification of gene expression
  2. Normalization
  3. Identification of differentially expressed genes.
- Quantification of gene expression is based on mapping of the sequenced fragments against a reference. The reference typically consists of a genome or a *de novo* assembled transcriptome.
- Normalization within and between samples is necessary for comparability. Common methods includes RPKM, upper quartile and TMM.

Summary (cont)

- Identification of differentially expressed genes is done by hypothesis testing. Statistical models are used to describe the data and its variability.
- Estimation the gene-specific overdispersion is important. Sharing of variability between genes significantly improves the performance in datasets with few samples.
- Biological replication is essential. More biological replicates means higher power. Pooling should be avoided!