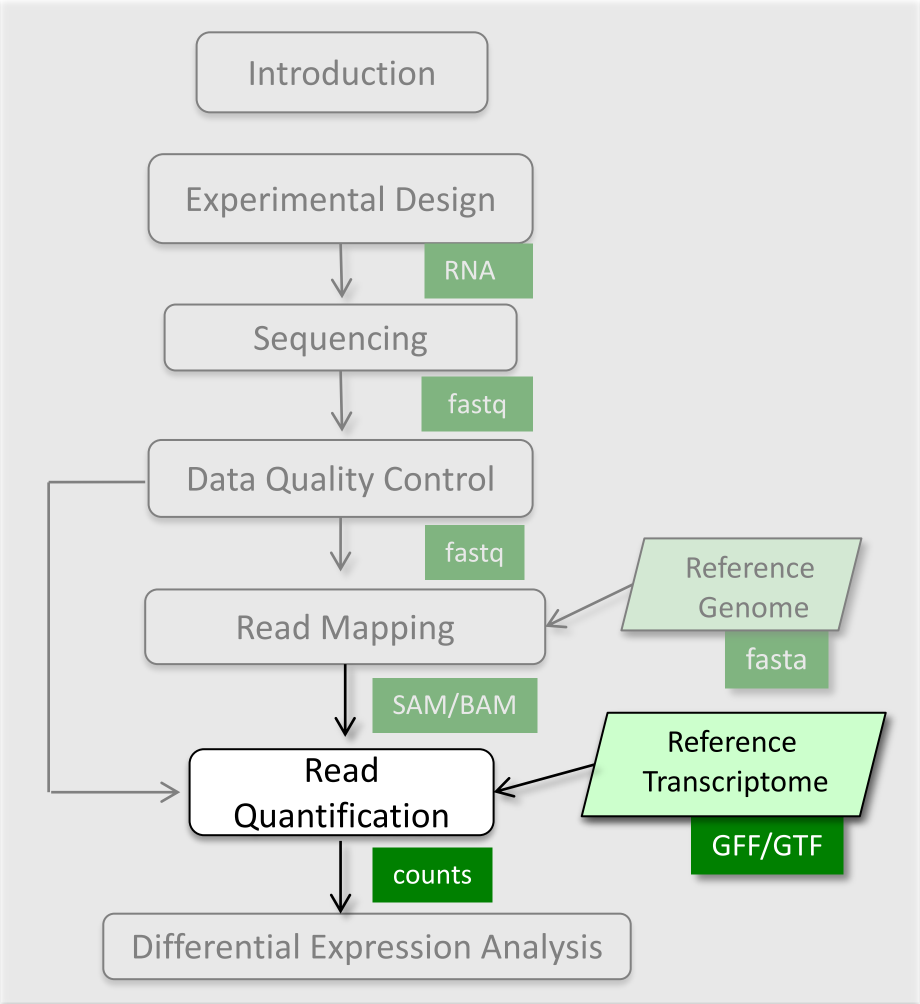
ETH zürich University of Zurich UZH functional genomics center zurich f. g. c. z.

RNA-seq Bioinformatics:

Read Quantification

Falko Noé

University of Zurich UZH | ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

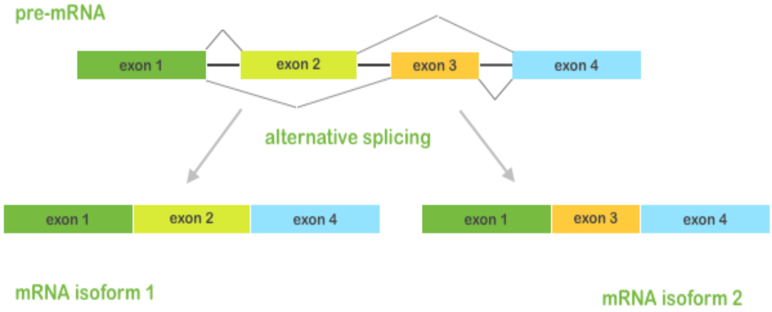


Expression quantification

- Read quantification (count models)
- Normalization
- Explorative analysis of the quantification

Expression quantification

- Expression quantification = finding the amount of sequenced reads assigned to a specific gene/transcript



- What to count
  - Gene level: Reads belong to a gene locus
  - Isoform level: Reads belong to an isoform

- How to count
  - Unique mapped reads

    Multi-reads are ignored – potentially biased quantification
  - All mapped reads

    Assignment of multi-reads need abundance estimation of overlapped genes/isoforms

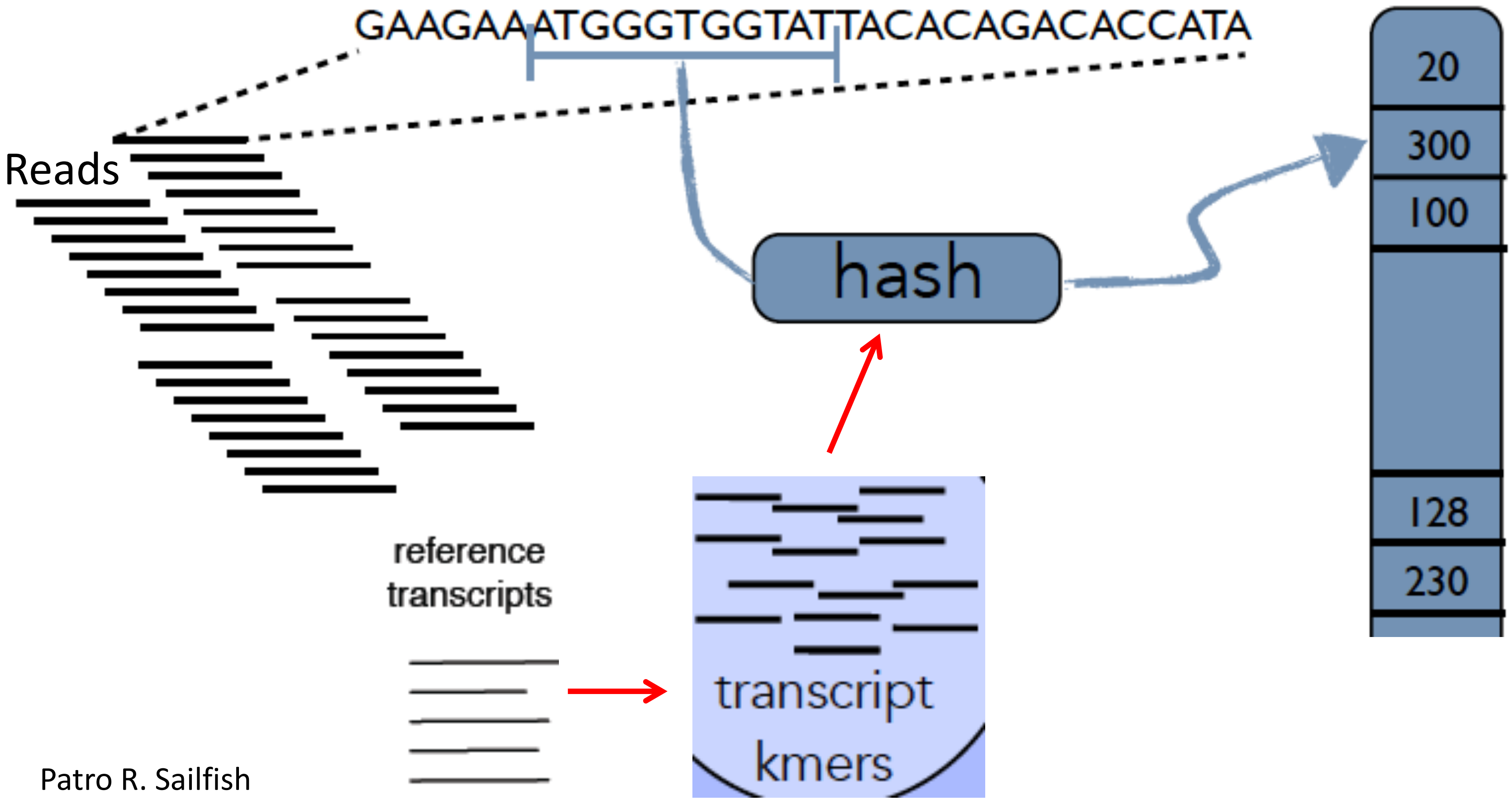
RNA-seq quantification: “Alignment independent/free”

- K-mer indexing and counting

Alignment dependent	Alignment free
Pre-process transcripts (e.g. build BWT)	Pre-process transcripts (e.g. build k-mer index)
Align reads to transcripts	Count k-mers in reads
Shuffle / allocate reads	Shuffle / allocate k-mers
Compute abundance	Compute abundance

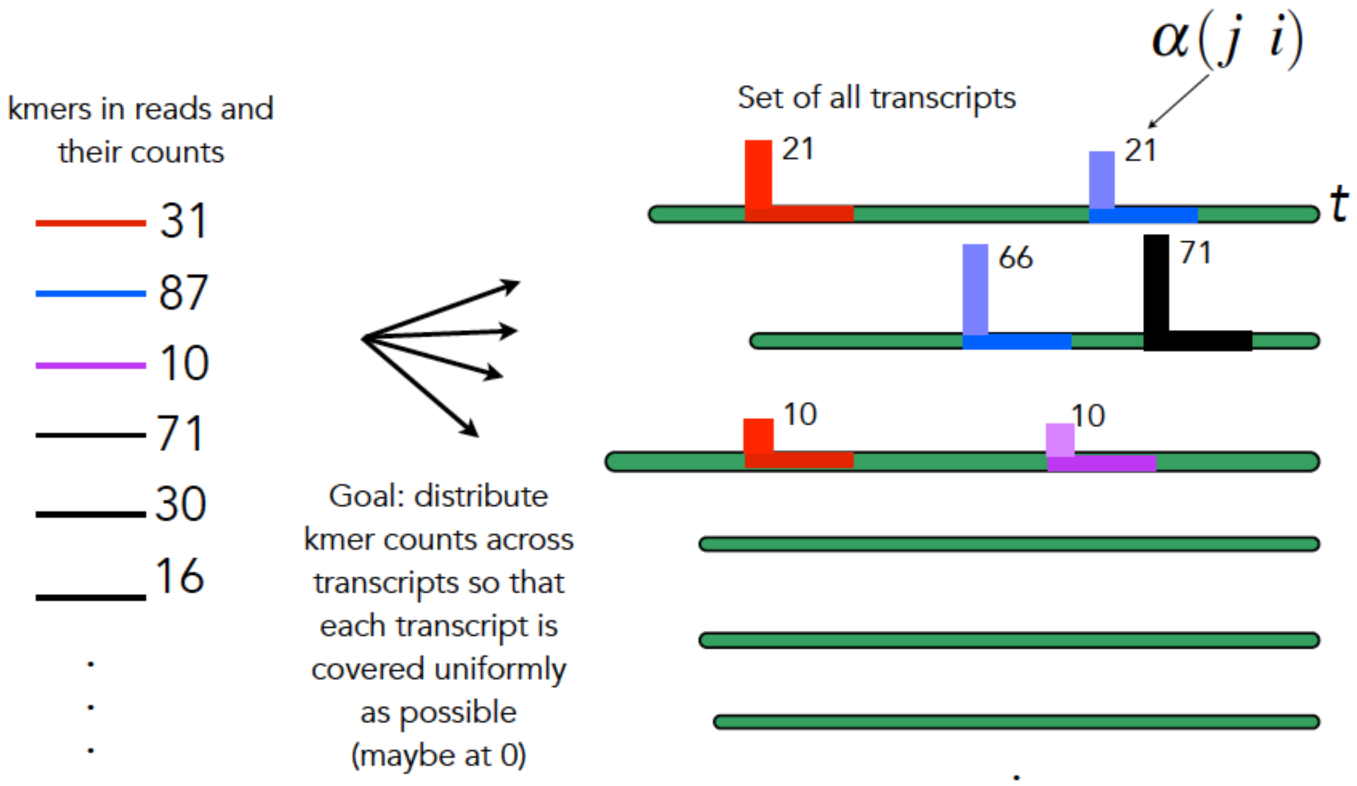
- Tools: **kallisto**, salmon and sailfish

RNA-seq quantification: K-mer indexing and counting



Patro R. Sailfish

Allocate k-mers to transcripts – EM estimation



kmers in reads and their counts

Set of all transcripts

Goal: distribute kmer counts across transcripts so that each transcript is covered uniformly as possible (maybe at 0)

Patro R. Sailfish

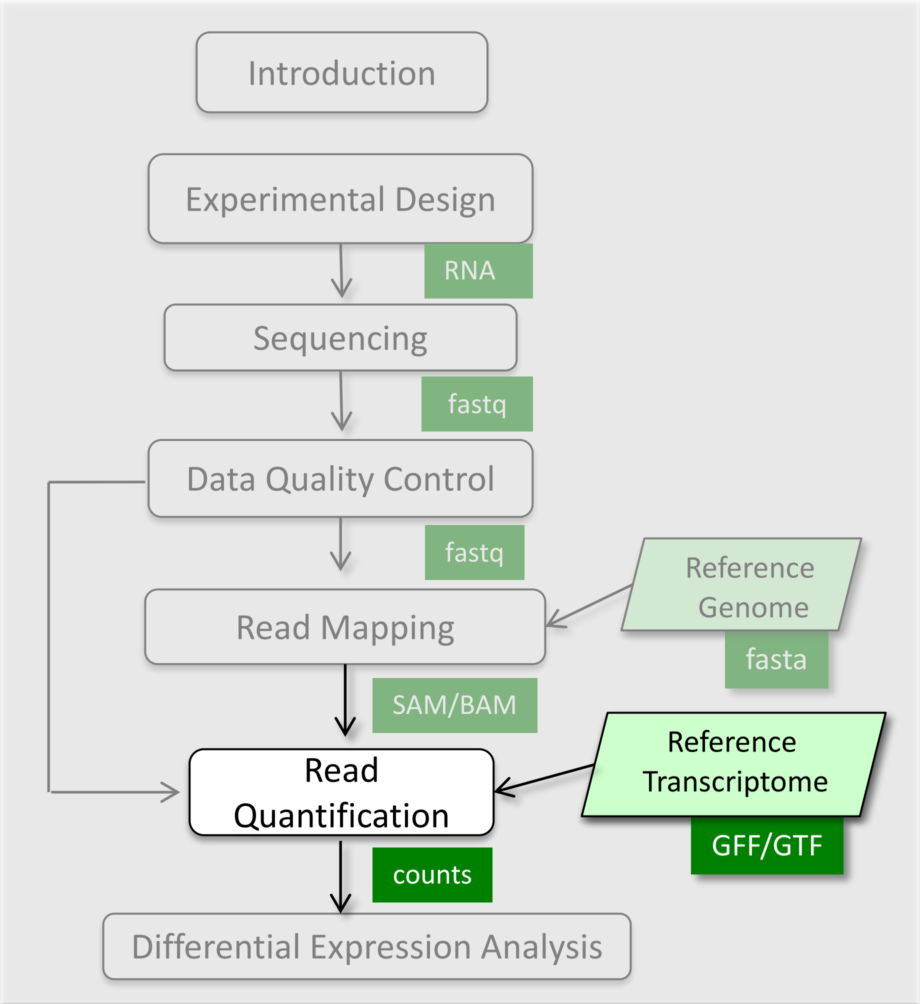
Count multi-reads via iterative estimation

1. Estimate abundances based on uniquely mapping reads only
2. For each multi-read, divide it between the transcripts to which it maps, proportionally to their abundances estimated in the first step
3. Re-compute abundances based on updated counts for each transcript
4. Continue with Step 2
5. Expectation-Maximization estimation

"RNA-Seq gene expression estimation with read mapping uncertainty"
Li, B., Ruotti, V., Stewart, R., Thomson, J., Dewey, C.
Bioinformatics, 2010

Alignment free counting

- Pros
  - Accurate and fast in quantifying **known transcripts**
  - Counts can be aggregated to get gene-level quantification
- Cons
  - Less well annotated genomes – less accurate results
  - RNA-seq is more than counting



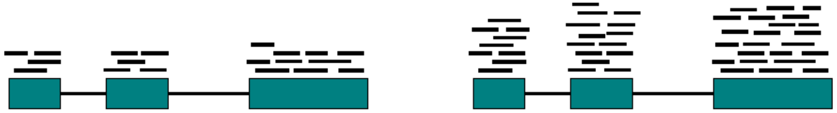
Expression quantification

- Read quantification (count models)
- Normalization
- Explorative analysis of the quantification

Number of reads ≠ Expression level

	Sample 1	Sample 2	Sample 3
Gene A	5	3	8
Gene B	17	23	42
Gene C	10	13	27
Gene D	752	615	1203
Gene E	1507	1225	2455

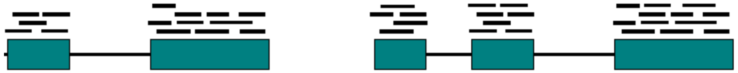
Gene D in the sample 3 have about twice as many reads aligned to it as in sample 2. What does it mean?



1) The gene is two times more expressed in sample 3 than in sample 2

2) Difference in sequencing depth between samples – sequencing depth

3) Longer isoform was expressed in sample 3 – transcript length



Normalization

	control				treated		
Gene 1	5	1	0	0	4	0	0
Gene 2	0	2	1	2	1	0	0
Gene 3	92	161	76	70	140	88	70
:	:				:		
:	:				:		
:	:				:		
Gene G	15	25	9	5	20	14	17

Correction multiplicative factor - scales the counts for each sample j .

C_j	1.1	1.6	0.6	0.7	1.4	0.7	0.8

Normalization

Raw count table

	control				treated		
Gene 1	5	1	0	0	4	0	0
Gene 2	0	2	1	2	1	0	0
Gene 3	92	161	76	70	140	88	70
:		:				:	
:		:				:	
:		:				:	
Gene G	15	25	9	5	20	14	17

Correction multiplicative factor - scales the counts for each sample j

Gene 3	92	161	76	70	140	88	70	
C_j	1.1	1.6	0.6	0.7	1.4	0.7	0.8	x
Gene 3	101.2	257.6	45.6	49	196	61.6	56	

Normalized counts

Normalization

Normalized count table

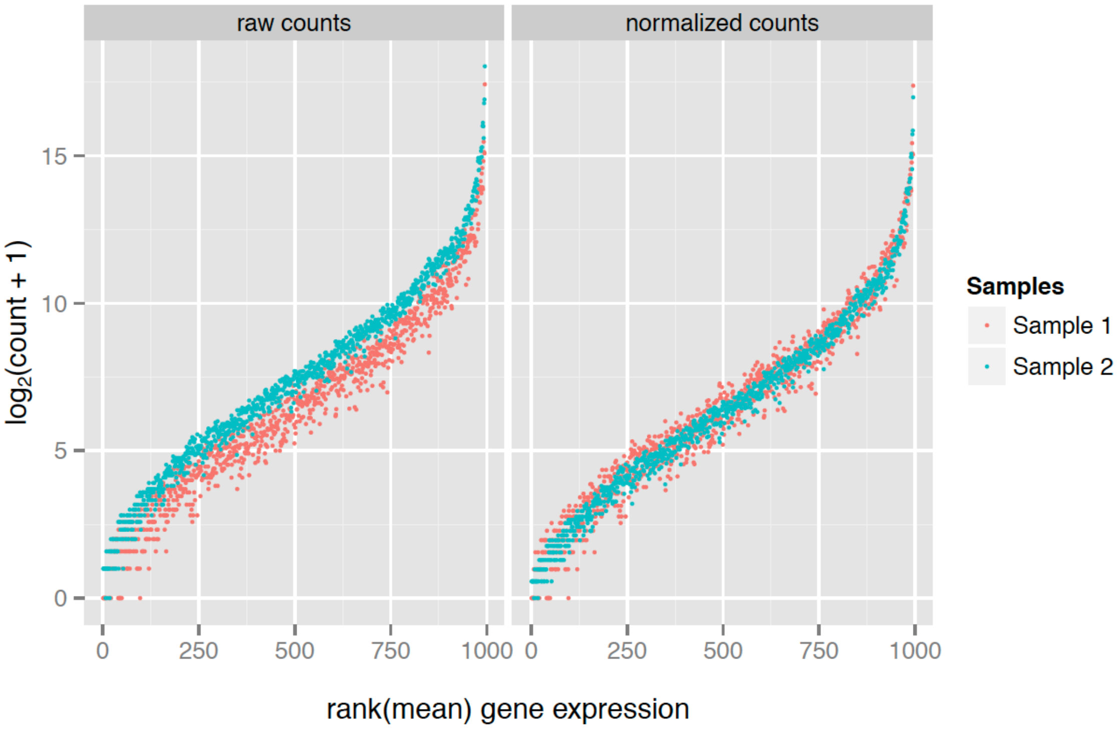
	control				treated			
Gene 1	5.5	1.6	0	0	**5.6**	**0**	**0**	
Gene 2	0	3.2	0.6	1.4	**1.4**	**0**	**0**	
Gene 3	101.2	257.6	45.6	49	**196**	**61.6**	**56**	
:		:				:		
:		:				:		
:		:				:		
Gene G	16.5	40	5.4	5.5	**28**	**9.8**	**13.6**	
								+
Lib. size	13.1	13.0	13.2	13.1	**13.2**	**13.0**	**13.1**	x 10^5

After normalization, the sequencing depth is almost equal

Total read count normalization

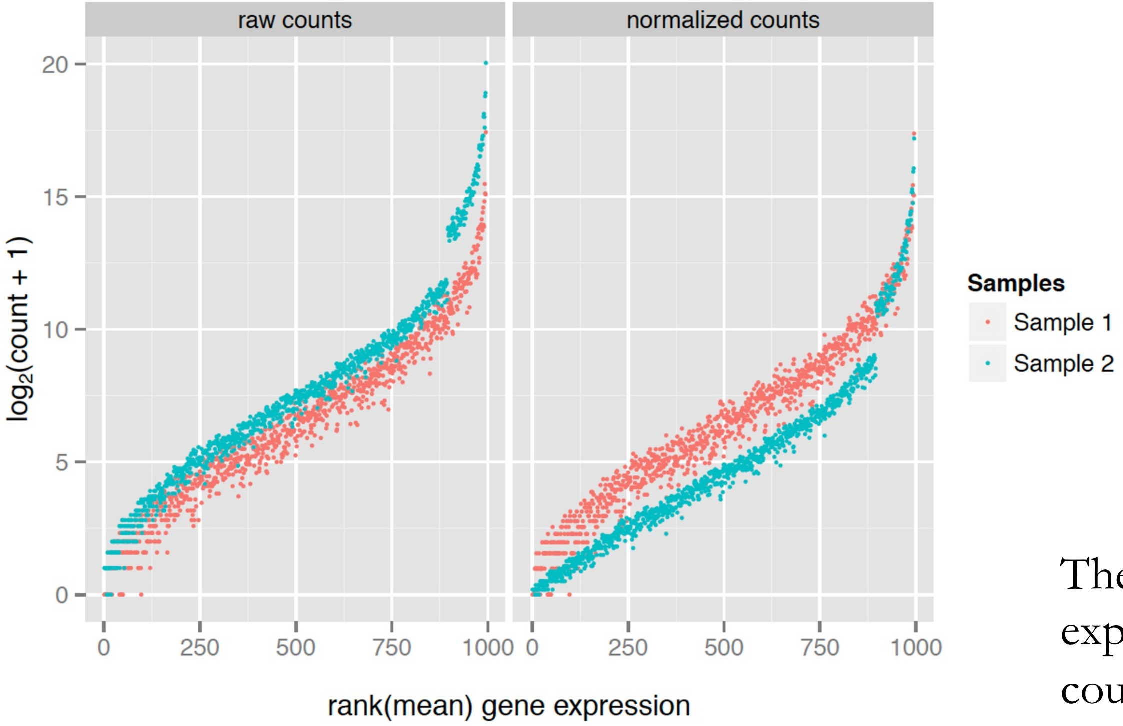
$$C_j = \frac{10^6}{D_j}$$

D_j: total number of reads in sample j



Total read count normalization: drawback

Example: 100 very highly expressed genes in one condition



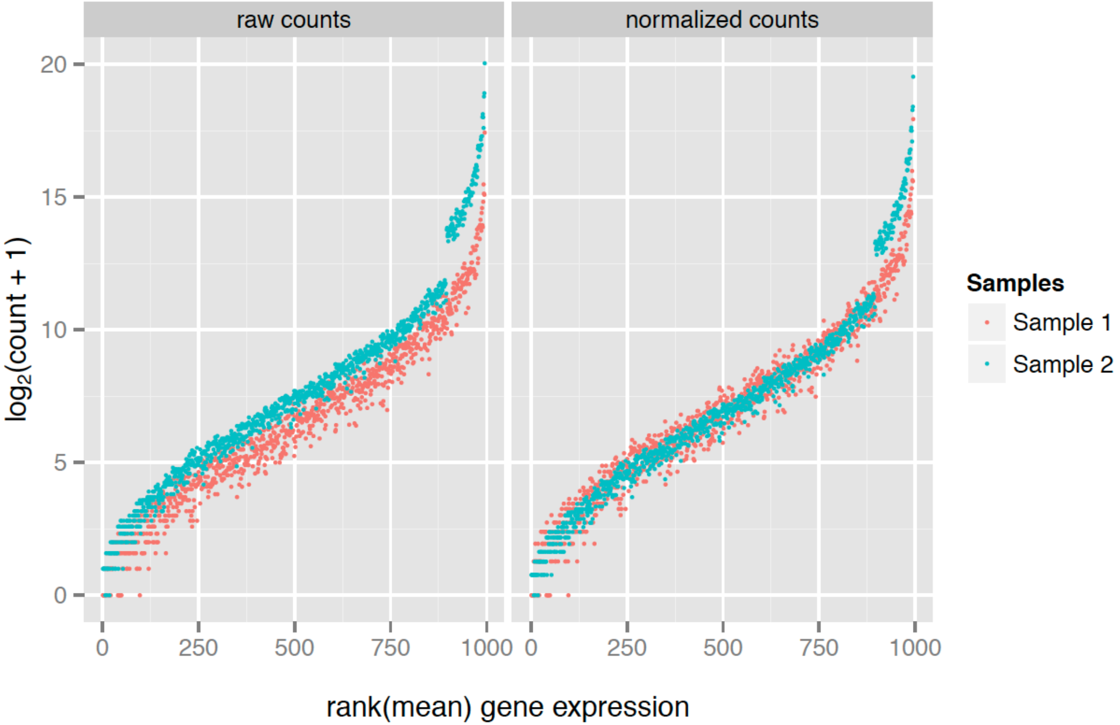
The small fraction of highly expressed genes will skew the counts of lowly expressed genes.

Upper quantile normalization

$$C_j = \frac{1}{D_j Q_j^{(p)}}$$

median normalized upper quantile (p-th percentile) of sample j

D_j: total number of reads in sample j



RPKM (FPKM) normalization

Reads (or Fragments, in the case of paired-end) per Kilobase per Million mapped reads

$$\mathrm{FPKM}_i = \frac{X_i}{\left(\frac{\widetilde{l}_i}{10^3}\right)\left(\frac{N}{10^6}\right)} = \frac{X_i}{\widetilde{l}_i N} \cdot 10^9$$

Counts per gene

Counts per Sample

length of transcript

Transcripts per million (TPM) normalization

$$\mathrm{TPM}_i = \frac{X_i}{\widetilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\widetilde{l}_j}} \right) \cdot 10^6$$

RPKM normalization

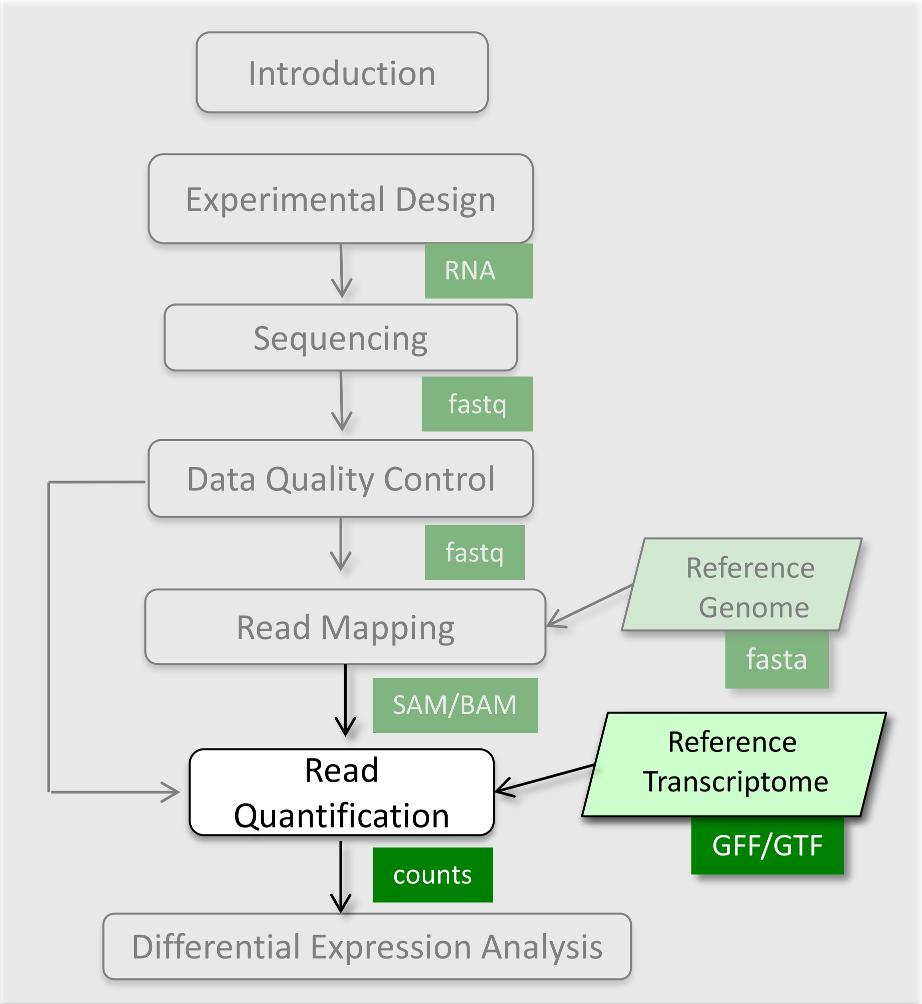
- Corrects for total library coverage
- Corrects for gene length
- Comparable between different genes within the same sample

FPKM normalization

- Only relevant for paired end libraries
- Effectively halves the raw counts

TPM normalization

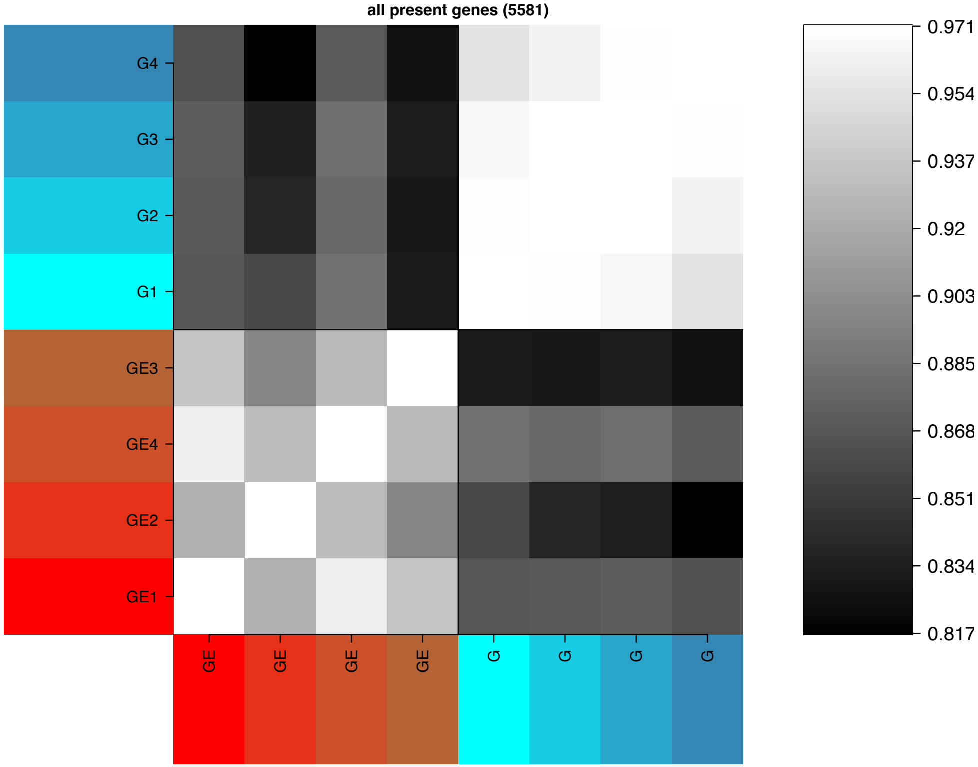
- Normalizes to transcript copies instead of reads
- Corrects for cases where the average transcript length differs



Expression quantification

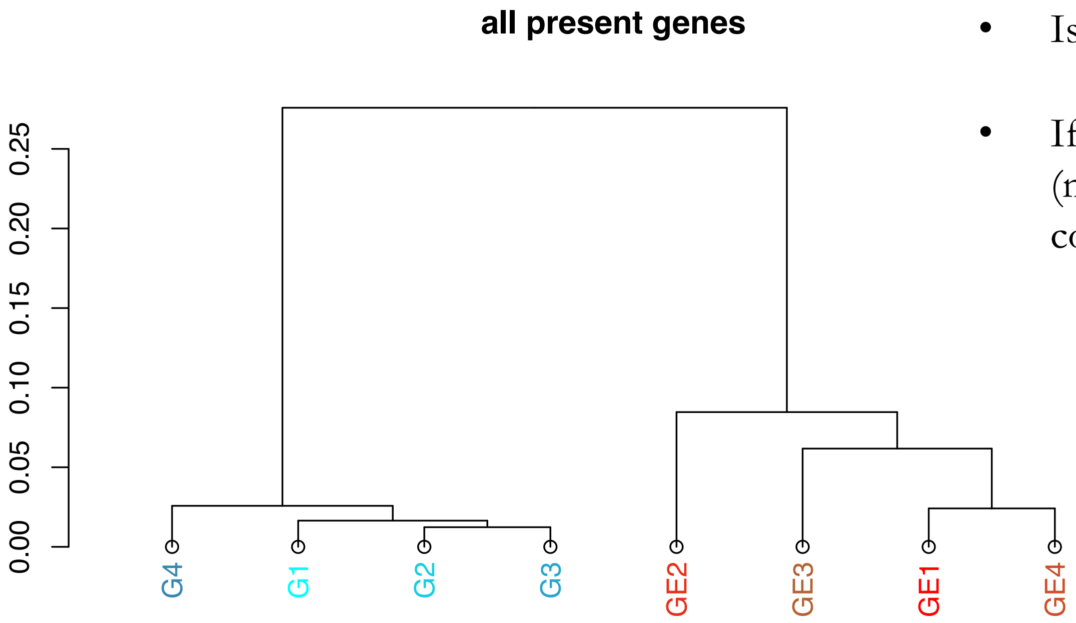
- Read quantification (count models)
- Normalization
- Explorative analysis of the quantification

How do samples' expression values correlate?



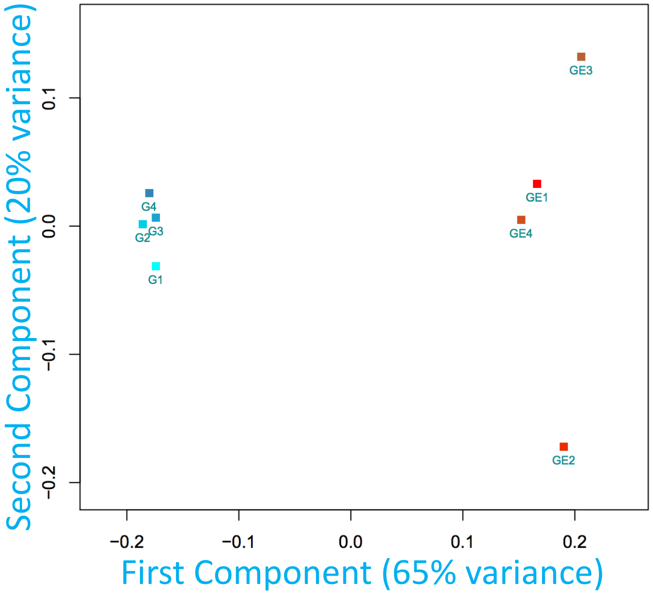
- Is the intra-group correlation much higher than the inter-group?
- How does this change if only the most varying genes are considered?

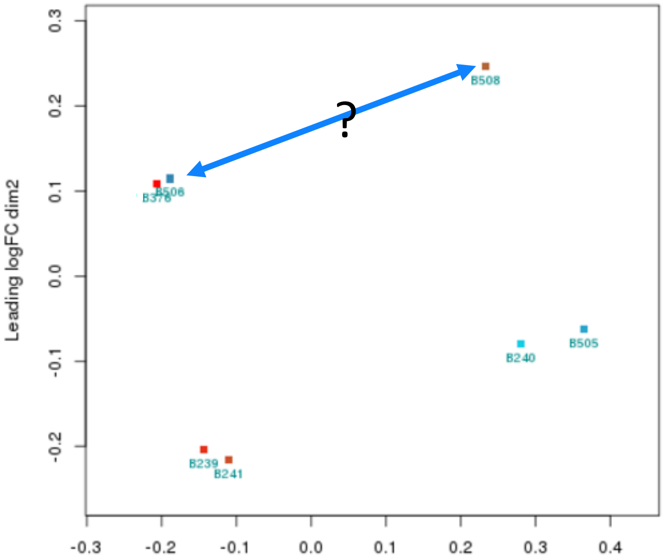
How do samples cluster?



- Do the groups cluster according to the biology?
- Is there a dominating separation?
- If not, can we identify possible cause (mis-labeling/sample swapping, hidden confounders) for sample outliers

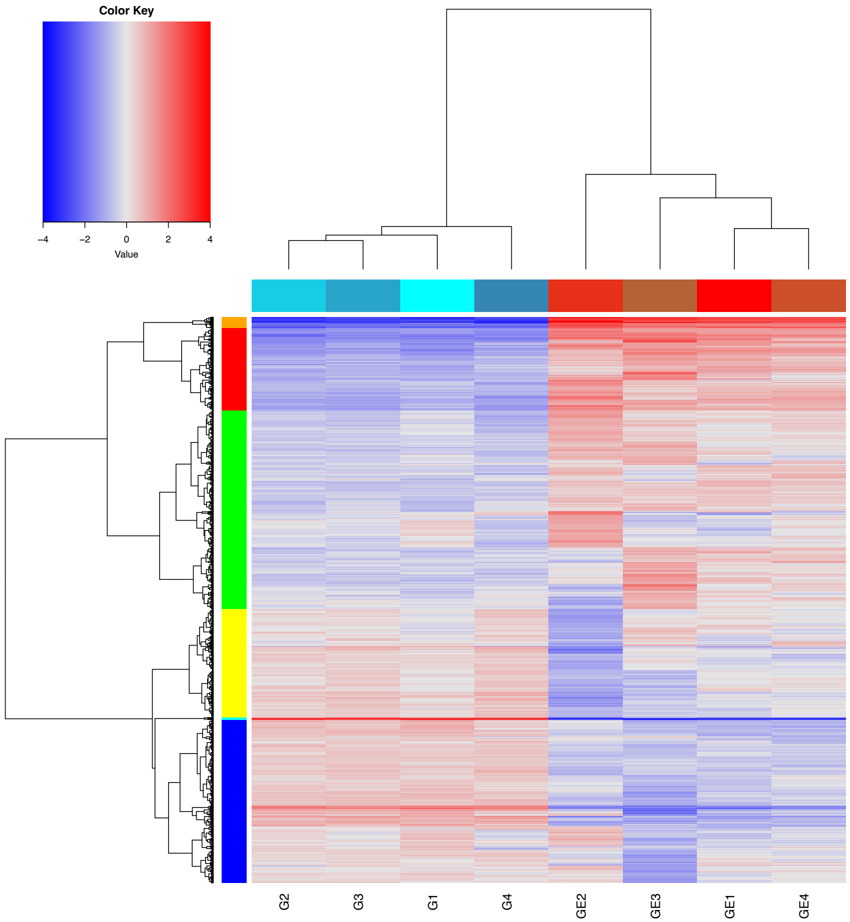
What does a principal components analysis (PCA) reveal?





- How do the samples group?
- Are there specific separation zones (e.g., half-planes, quadrants)?
- If not, can we identify possible cause (mis-labeling/sample swapping, hidden confounders) for sample outliers

How do the genes with high variance behave?



- Do they separate the groups?
- How do the genes cluster
- Are there dominant cluster?
- Are there very small clusters which might represent very strong effects?

Expression quantification: Summary

- Alignment dependent quantification = counting reads per gene
  - Different models: union, intersect with different level of stringency
- Alignment free quantification – kmer indexing and counting
  - Accuracy depends on annotation quality
- Shuffling of reads/kmers belong to multiple genes/transcripts
  - Expectation-Maximization Estimation
- Normalization
  - Needed to correct for varied sequencing depth and transcript length when necessary
  - RPKM is not effective, relative RPKM is TPM
- Explorative analysis of the counts
  - Identify outliers (samples, genes/transcripts)
  - Possible causes (mis-labeling, hidden confounders, GC, length)

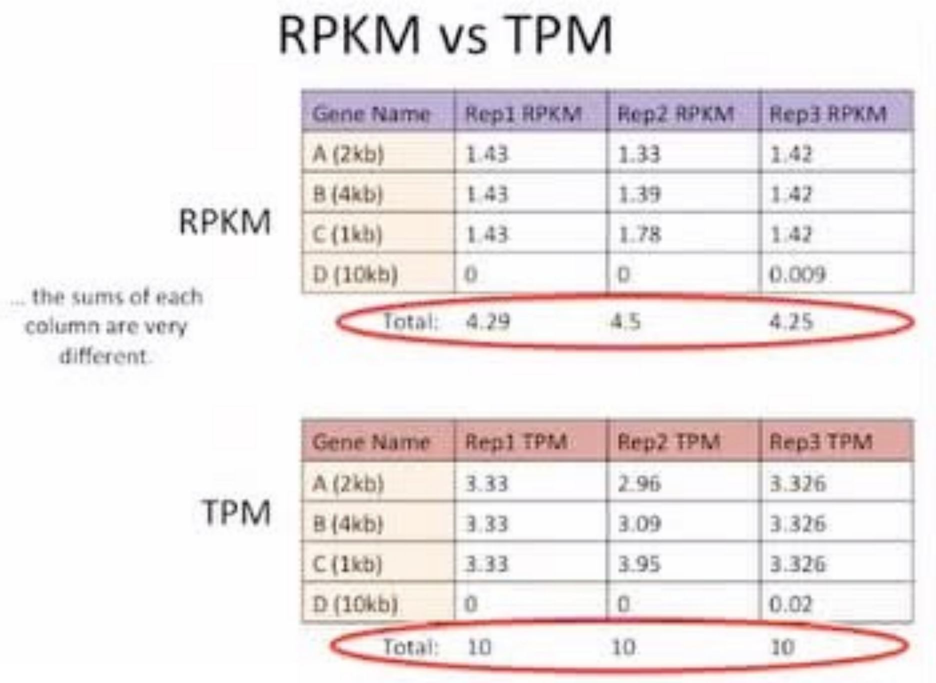
Supplementary slides

RPKM vs. TPM

- TPM values can be compared across samples
- RPKM values are not effectively normalized thus not comparable across samples
- Relative RPKM are TPM

$$\mathrm{TPM}_i = \left(\frac{\mathrm{FPKM}_i}{\sum_j \mathrm{FPKM}_j} \right) \cdot 10^6$$

  - 1.43/(1.43+1.43+1.43)*10=3.33



RPKM vs TPM

RPKM

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009
Total:	4.29	4.5	4.25

... the sums of each column are very different.

TPM

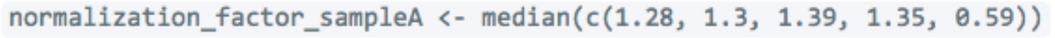
Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02
Total:	10	10	10

Normalization: EdgeR

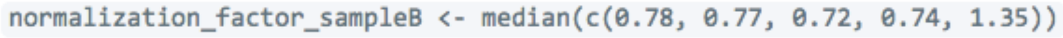
- Trimmed Mean of M-values (TMM)
- Reference samples: samples with average expressions closest to mean of all samples
- Test samples: all others
- For each test sample
  - Remove genes
    - with highest/lowest expression
    - with highest/lowest fold changes /log2 ratios
  - Normalization factor: Mean of log2 ratios between the test and reference, weighted by asymptotic variance estimated when the sample size approaches infinity

MD Robinson and A. Oshlack. Genome Biology. 2010

Normalization: DESeq2



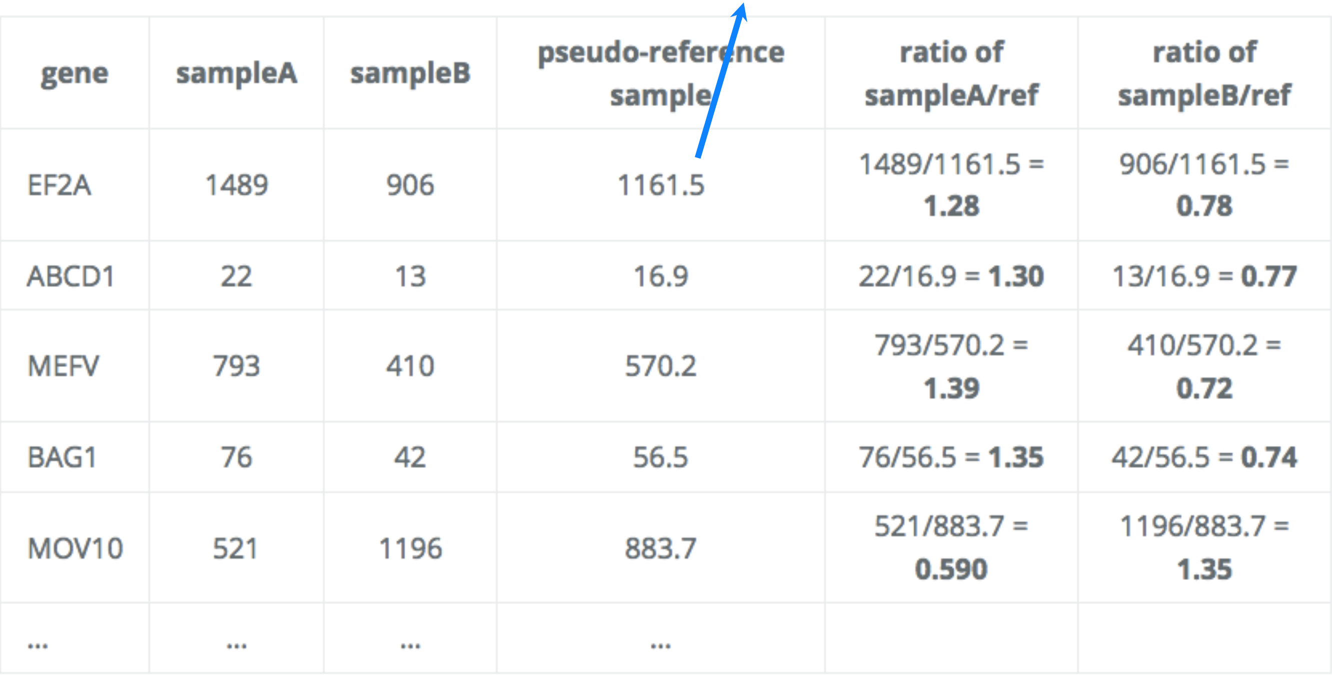
```
normalization_factor_sampleA <- median(c(1.28, 1.3, 1.39, 1.35, 0.59))
```



```
normalization_factor_sampleB <- median(c(0.78, 0.77, 0.72, 0.74, 1.35))
```

- Median of ratios

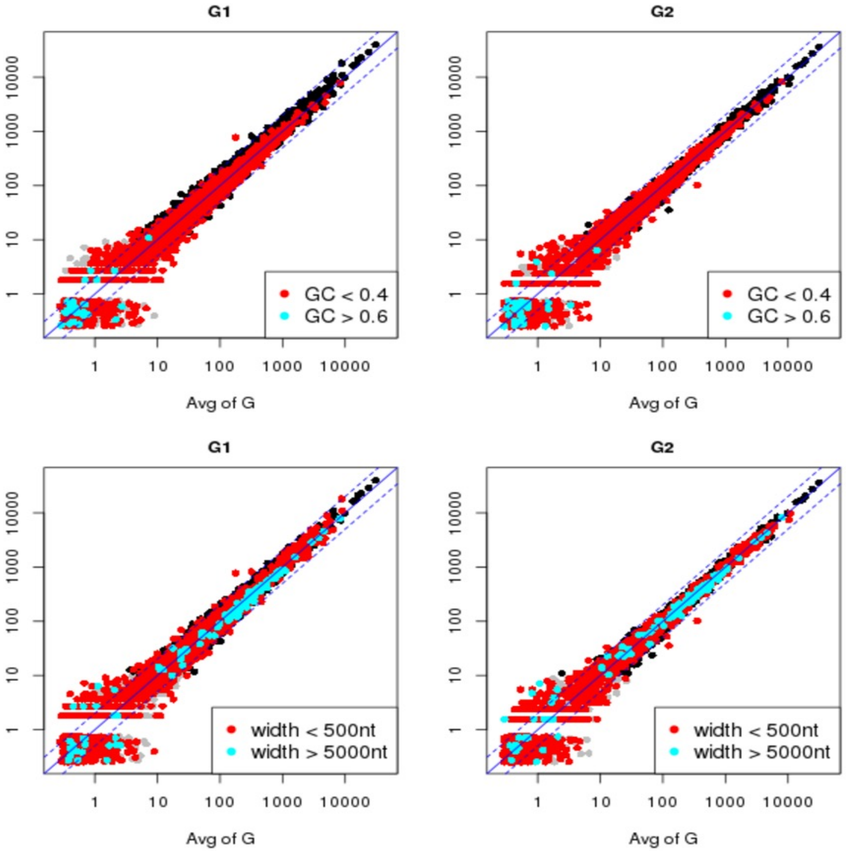
Geometric mean: sqrt(1489*906)=1161.5



gene	sampleA	sampleB	pseudo-reference sample	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**
...	...	...	...		

http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Compare individual samples to the mean of the group



- Try to identify possible causes (GC content, gene length, etc.) for expression outliers