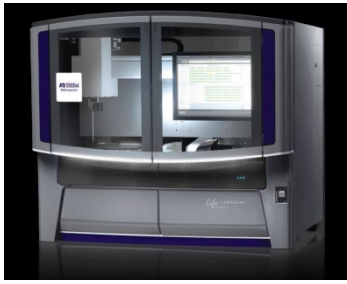
Analysis of RNA-Seq data

Counting, normalization, and statistical tests for differential expression

March 12, 2021

Dr. ir. Perry D. Moerland
Bioinformatics Laboratory
Amsterdam UMC
p.d.moerland@amsterdamumc.nl
www.bioinformaticslaboratory.nl

Next generation sequencing



Applications

- Full genomes (human, bacteria, viruses,....)
- Structural variation
- Variant detection
- Exome sequencing
- Metagenomics
- Chip-Seq
- DNA methylation
- .....
- **RNA-Seq**

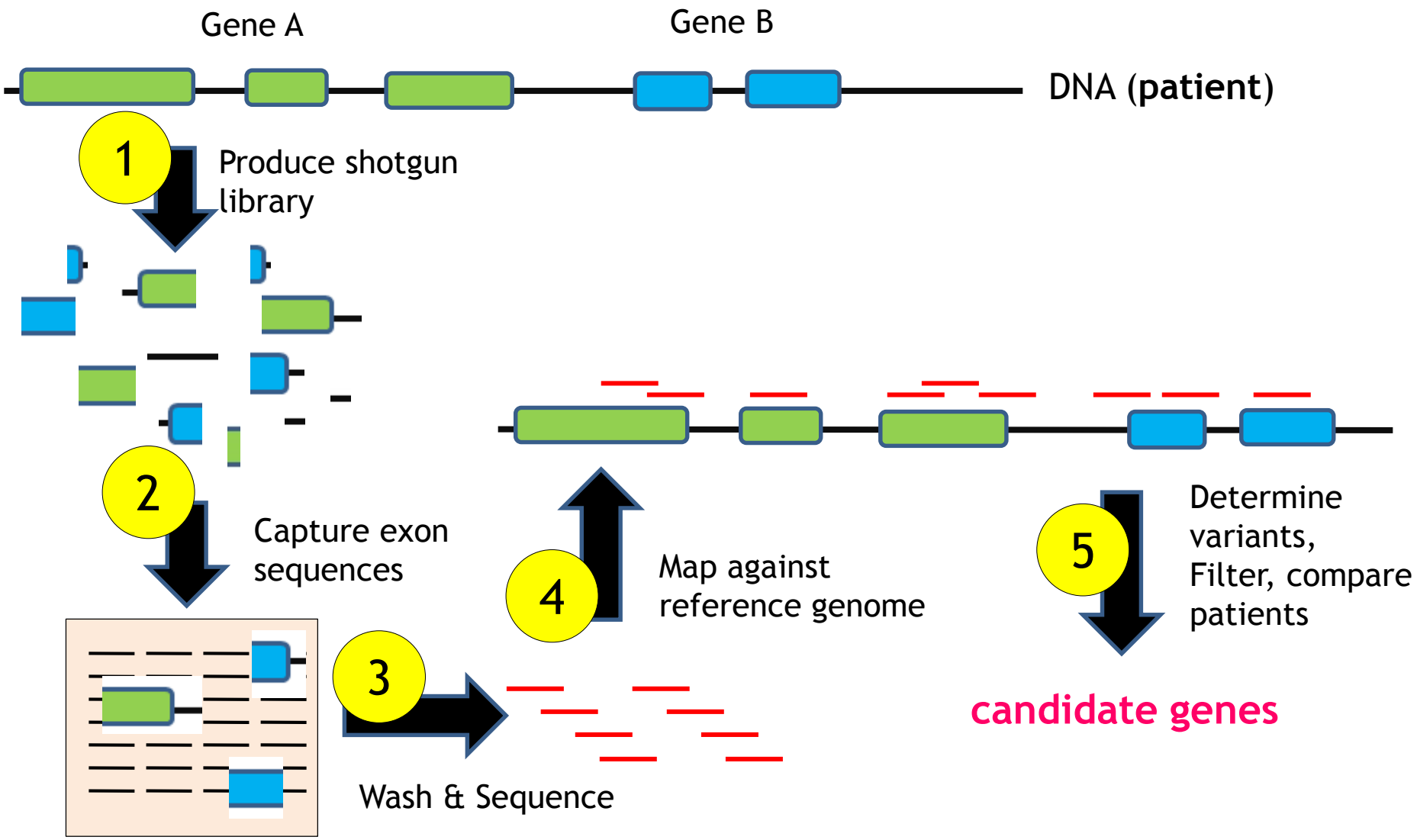
RNA-Seq: why?

- Functional questions:
  - Which RNA is expressed?
- Differential expression
  - Different patient groups
  - Different treatments (drugs on cell lines)
- Observe RNA specific features:
  - Alternative isoforms
  - Fusion transcripts
  - RNA editing

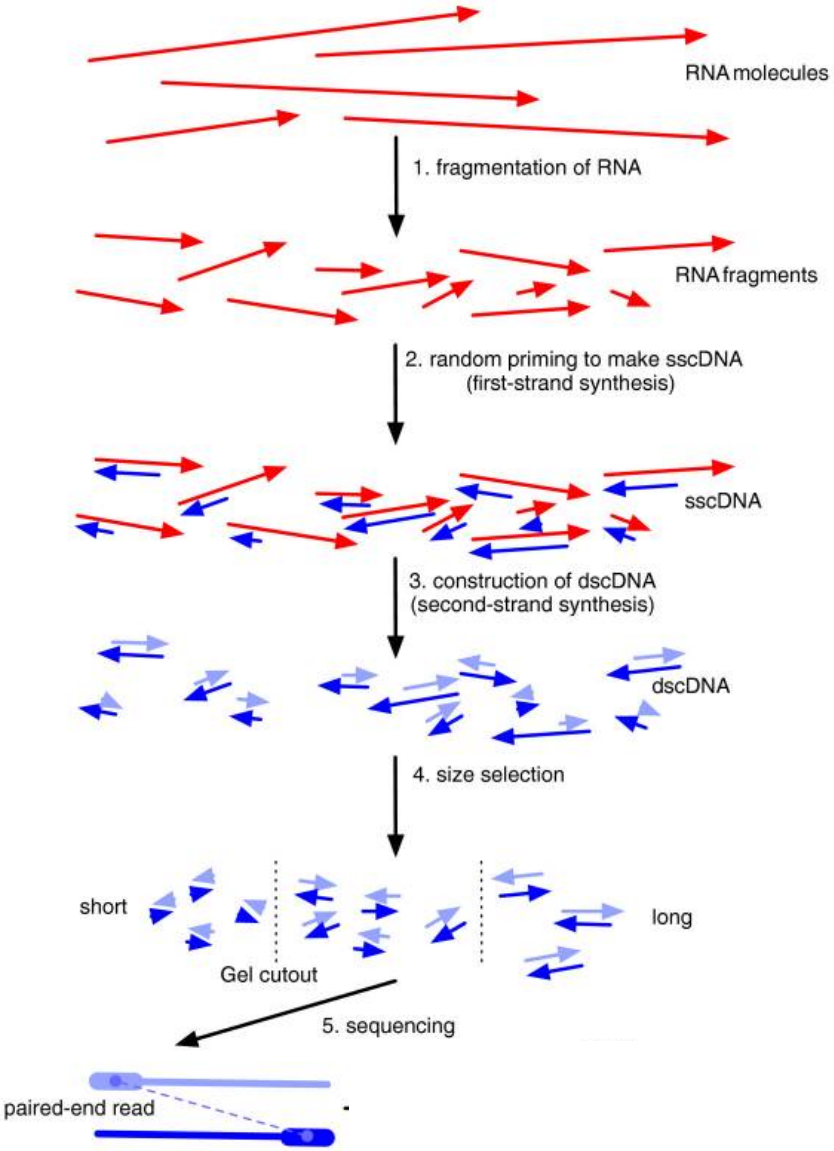
Challenges: RNA-Seq (versus DNA-Seq)

- RNAs consist of small exons that may be separated by large introns
  - Mapping reads to genome is more challenging
- The relative abundance of RNAs vary wildly
  - 10^5 – 10^7 orders of magnitude
  - Since RNA sequencing works by random sampling, a small fraction of highly expressed genes may consume the majority of reads
- RNAs come in a wide range of sizes
  - Small RNAs must be captured separately
- RNA is fragile compared to DNA (easily degraded)
- (Measurement biases and variability)

Exome-Seq: how does it work?

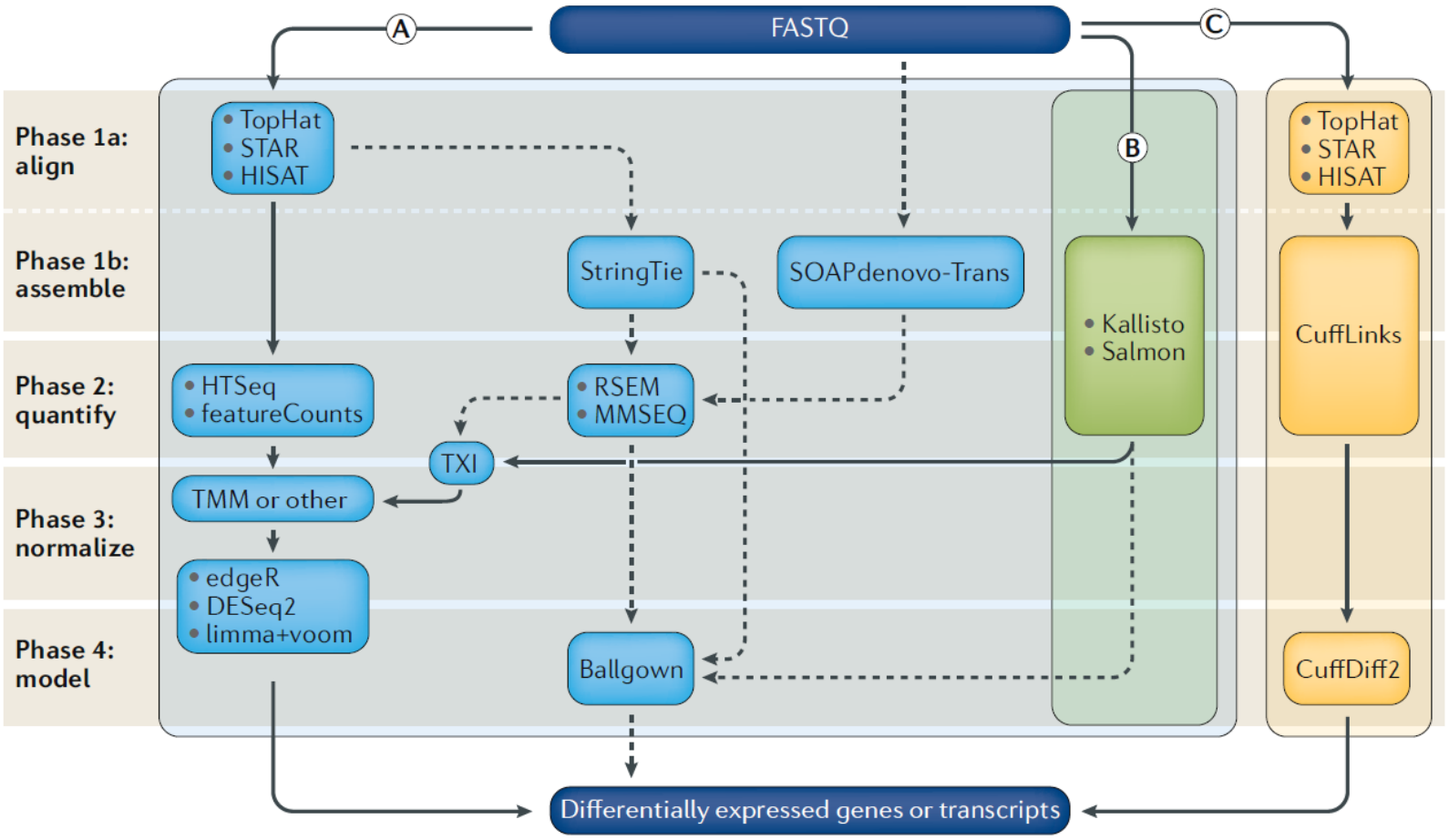


RNA-Seq: typical experimental protocol



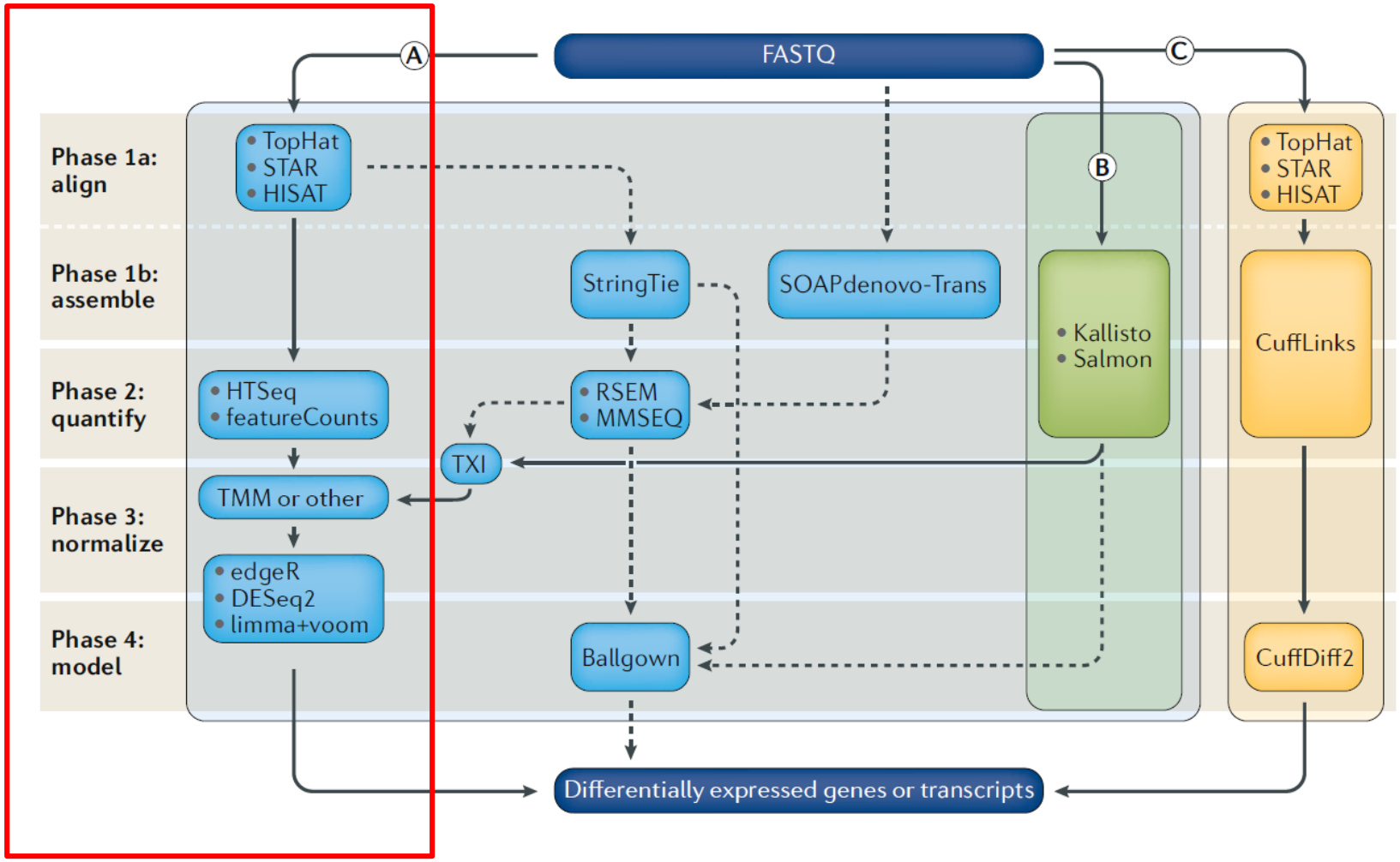
Roberts et al., Genome Biology, 12(3):R22 (2011)

RNA-Seq: data analysis workflow



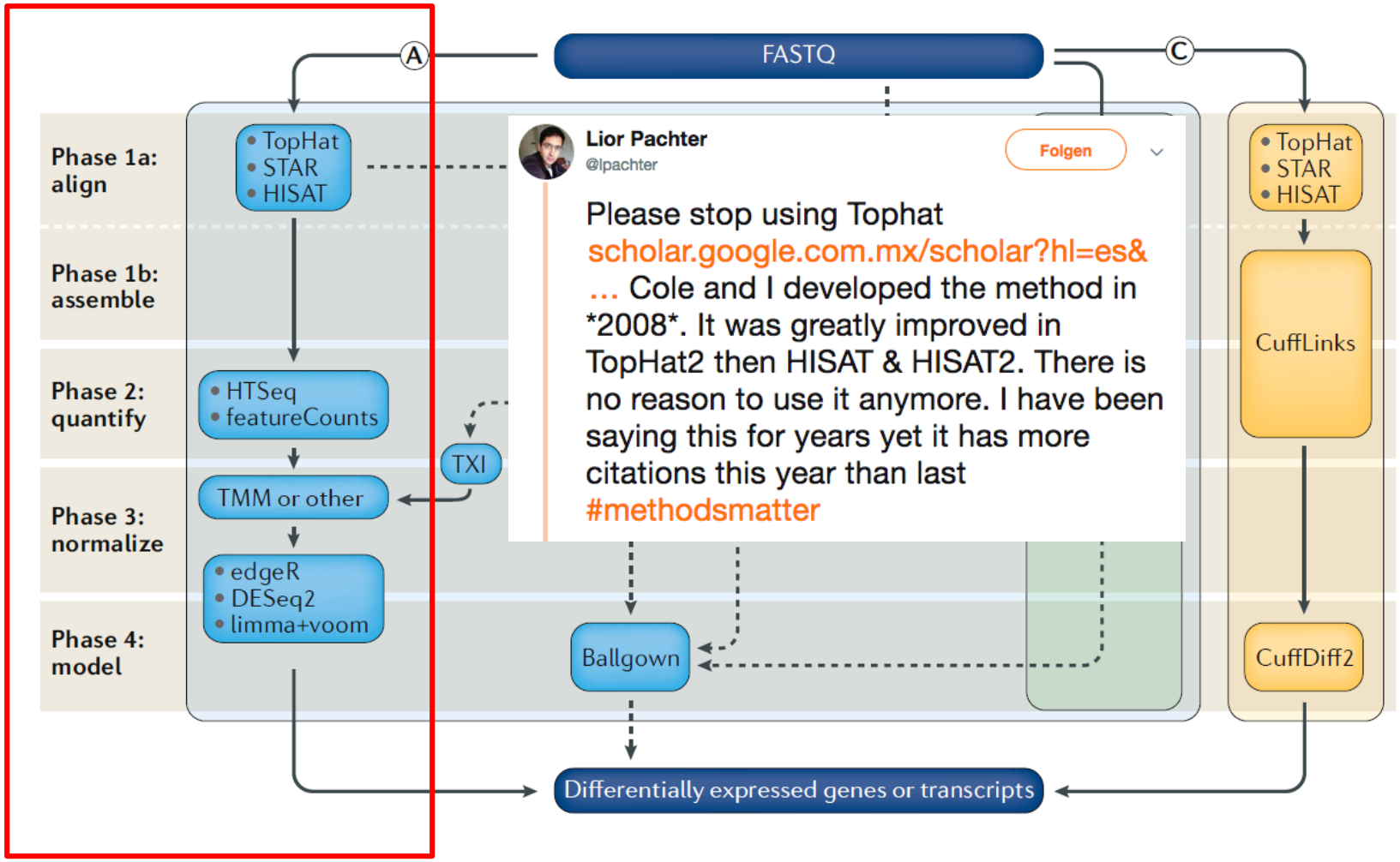
Stark et al., Nature Reviews Genetics, 20:631-56 (2019)

RNA-Seq: data analysis workflow



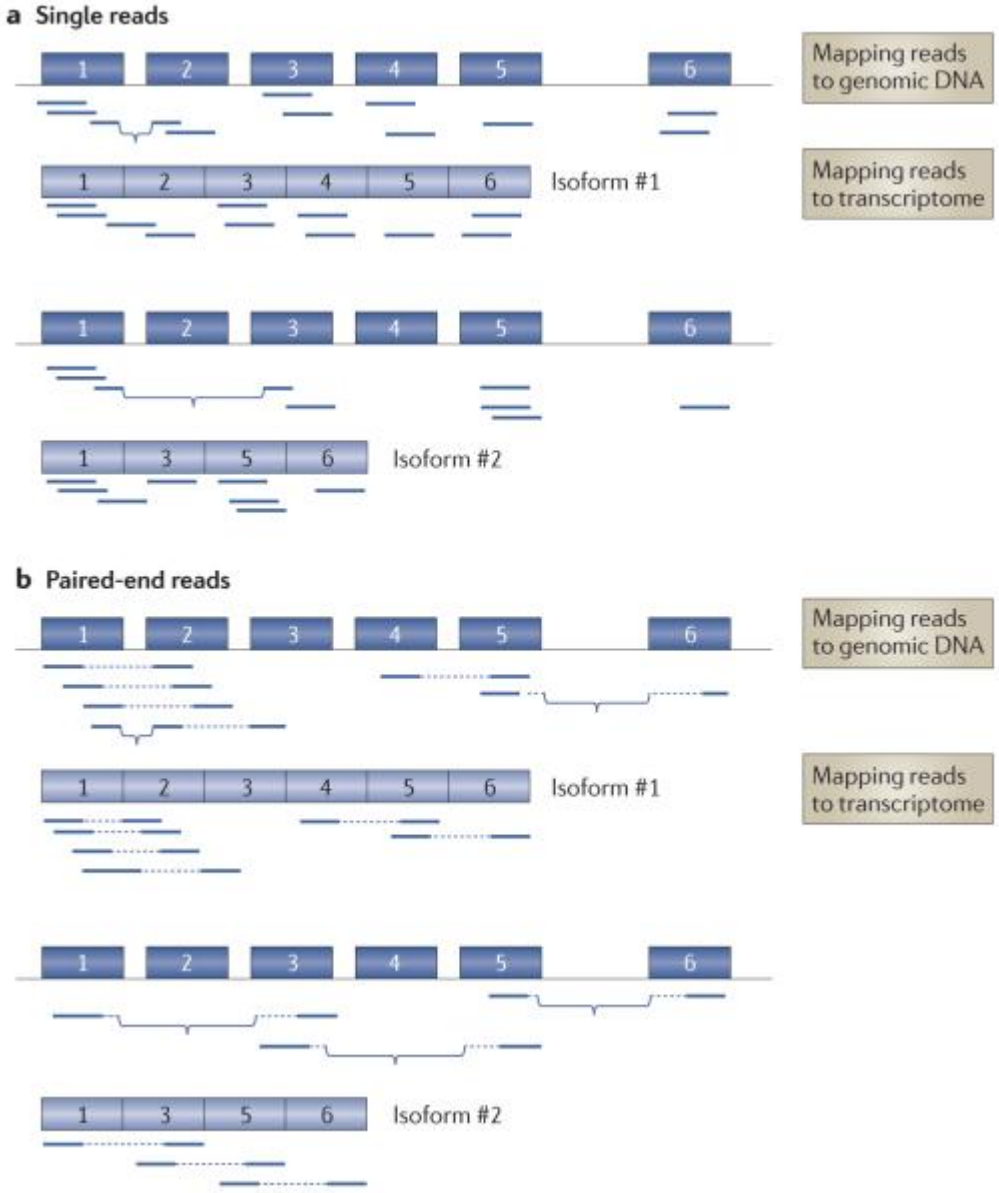
Stark et al., Nature Reviews Genetics, 20:631-56 (2019)

RNA-Seq: data analysis workflow



Stark et al., Nature Reviews Genetics, 20:631-56 (2019)

RNA-Seq: data after alignment



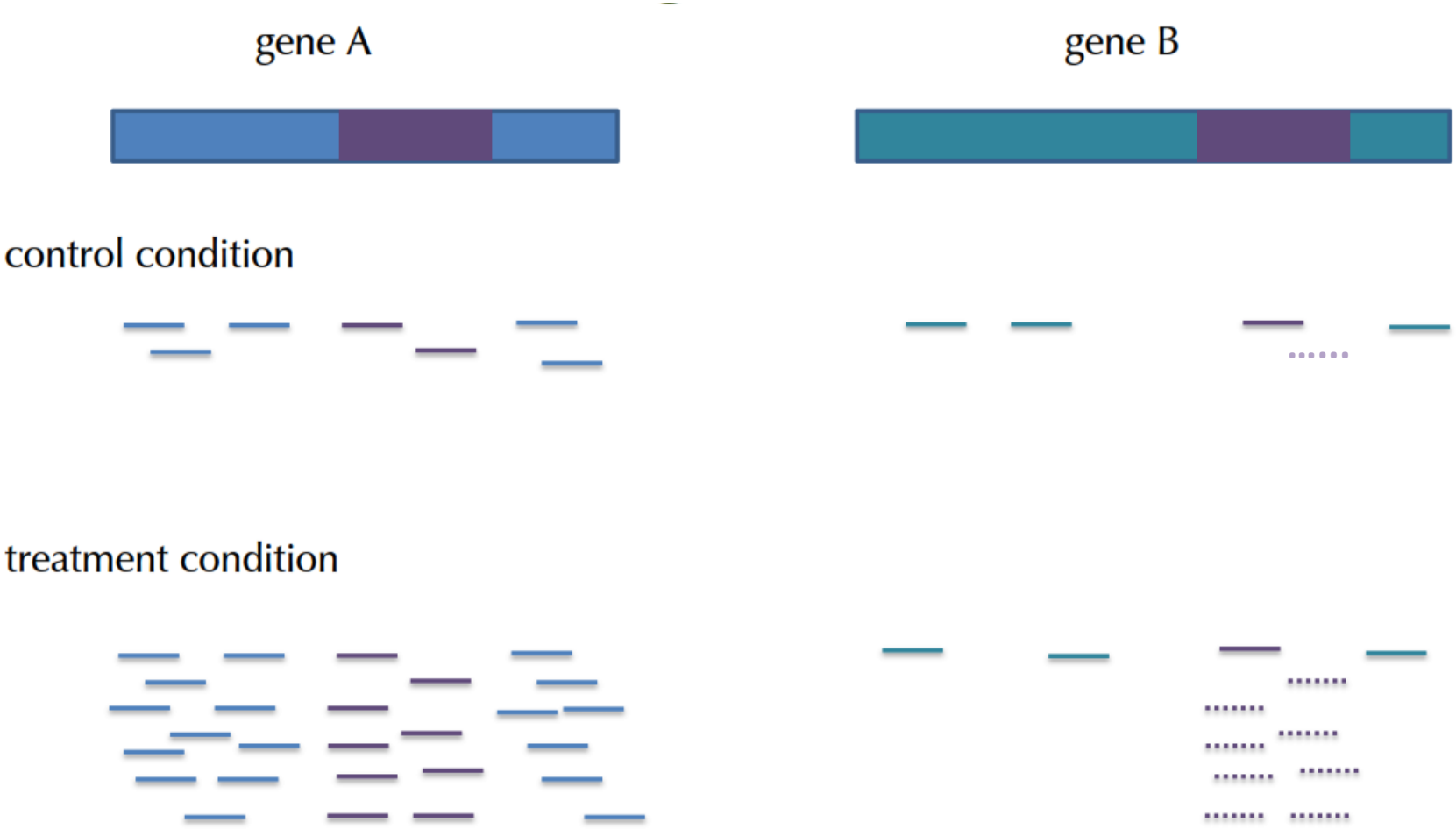
Ozsolak et al., Nat Rev Genet, 12(2):87–98 (2011)

Read counting: expression quantification

Basic rules

- Count reads, not base pairs
- In general, discard a read if
  - The alignment quality score is bad
  - (for paired-end reads) the mates do not map to the same gene
  - It cannot be uniquely mapped, for example if its alignment overlaps with several genes

Read counting: discard non-unique alignments



Context: differential expression

Read counting: union mode

- Define a *feature* (gene) as the union of all its exons (exon-union)
- For each position *i* in the read, set *S(i)* is the set of all features overlapping *i*
- In union mode set *S* is the union of sets *S(i)*

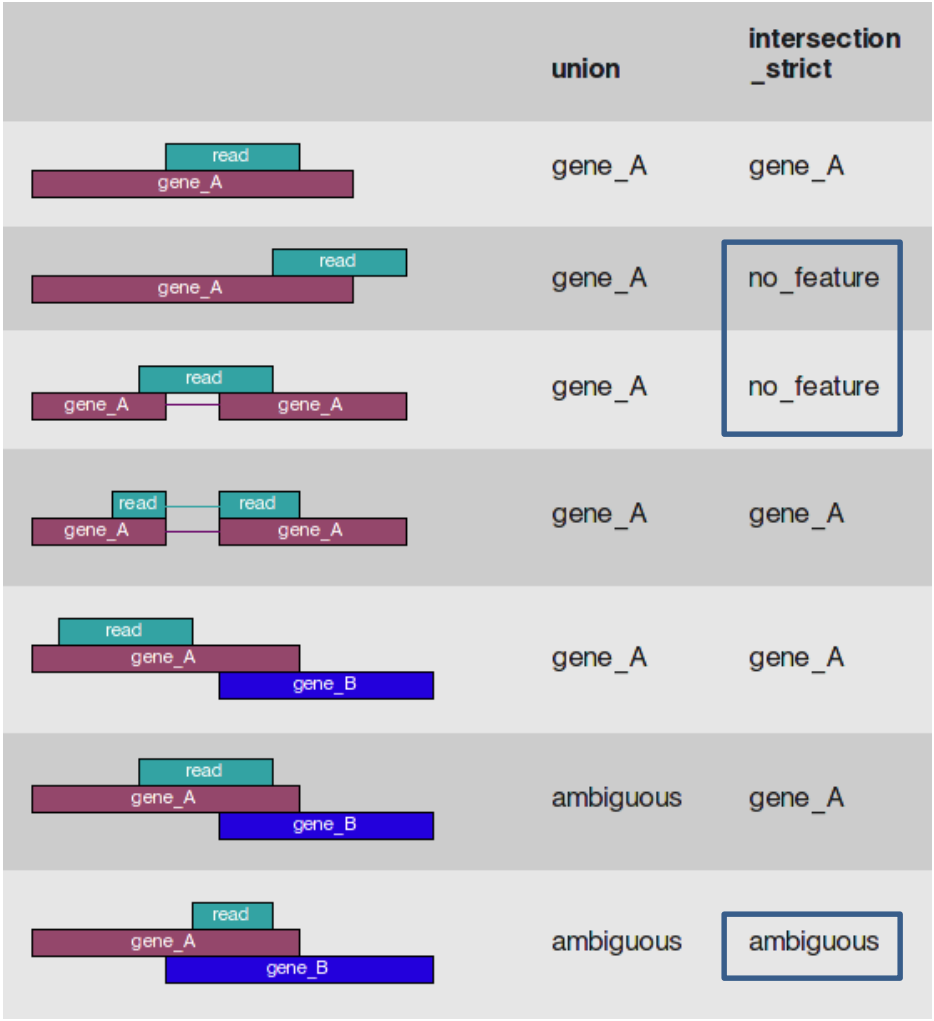
If size of S = 1: read counted

If size of S>1 : read not counted

If size of S=0 : read not counted

https://htseq.readthedocs.io/en/release_0.9.1/count.html

Read counting: strict intersection mode



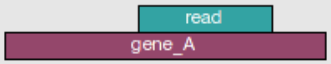
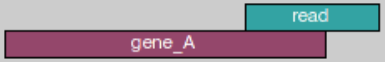
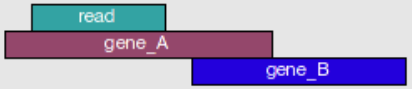
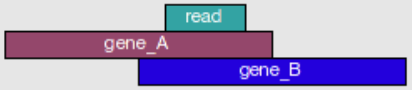
- In strict intersection mode set *S* is the intersection of sets *S(i)*

If size of S = 1: read counted

If size of S>1 : read not counted

If size of S=0 : read not counted

Read counting: non-empty intersection mode

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

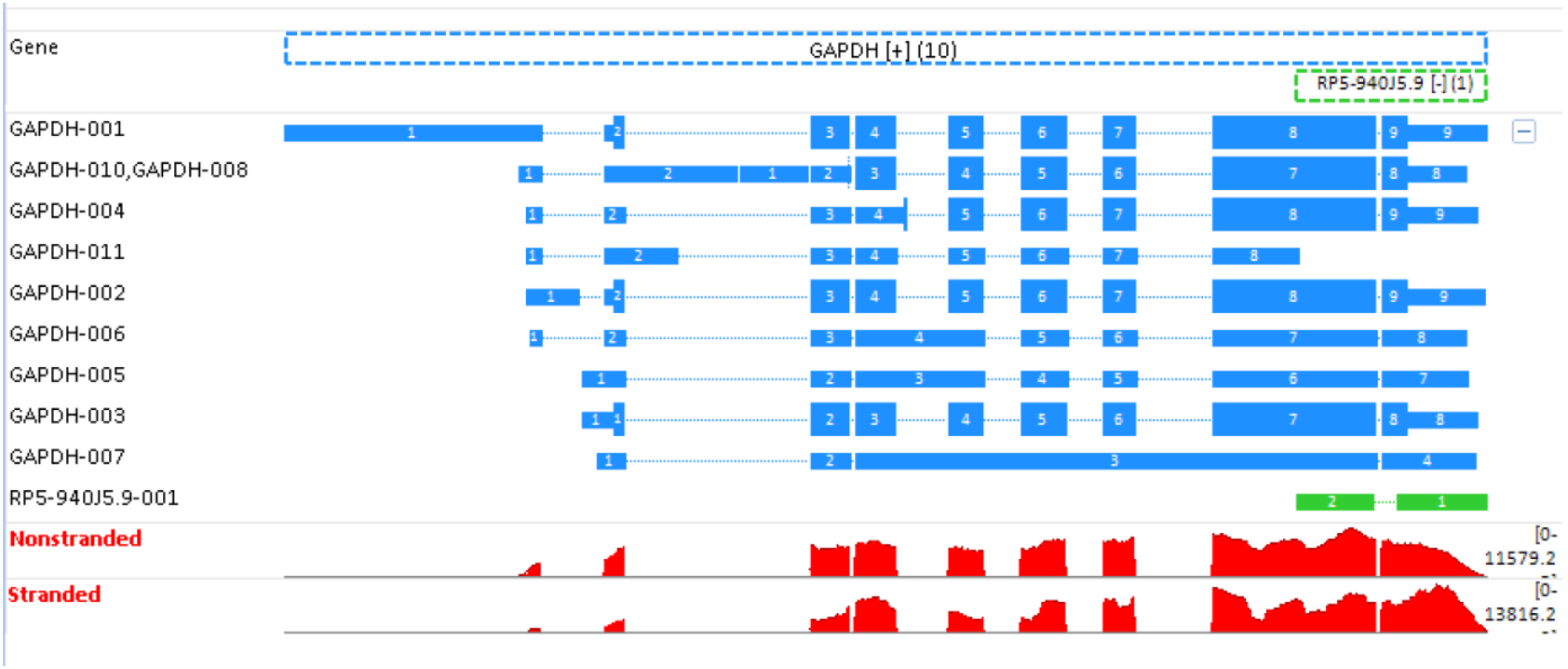
- In non-empty intersection mode set *S* is the intersection of all non-empty sets *S(i)*

If size of S = 1: read counted

If size of S>1 : read not counted

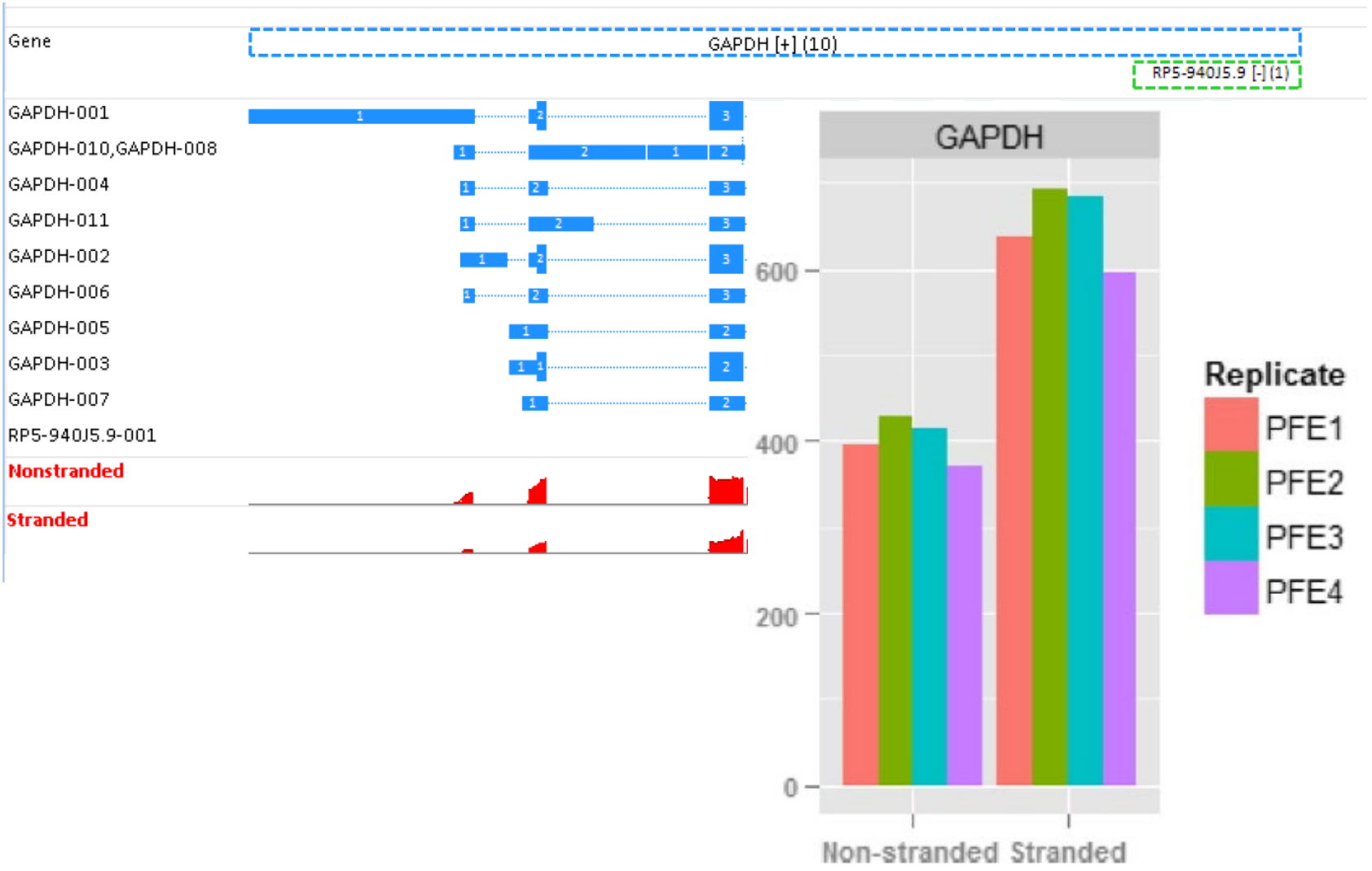
If size of S=0 : read not counted

Strand-specific RNA-Seq



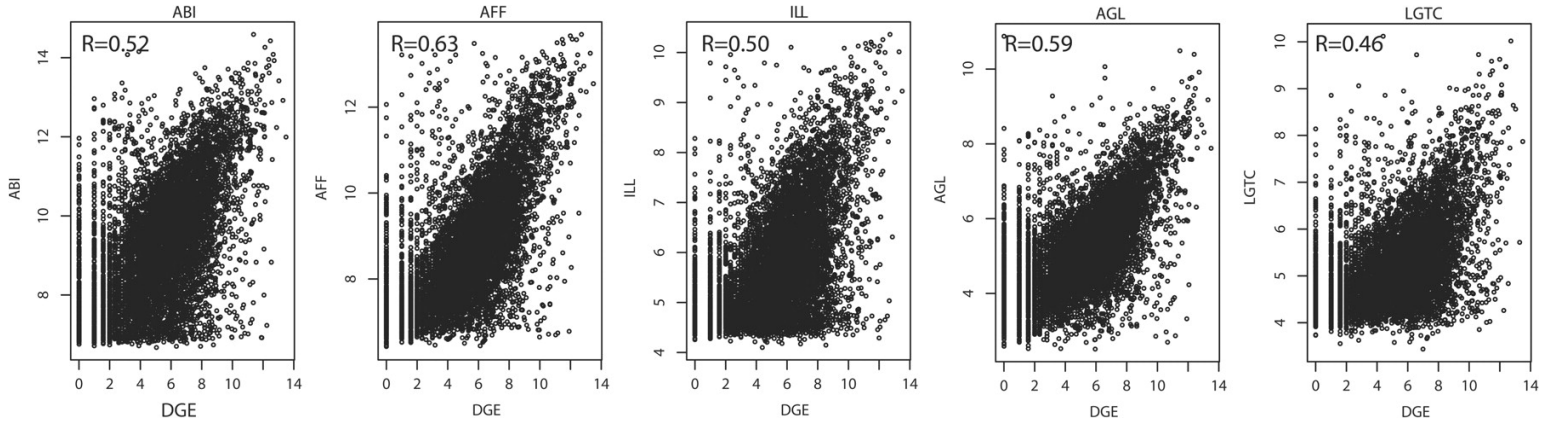
Zhao et al., BMC Genomics, 16:675 (2015)

Strand-specific RNA-Seq



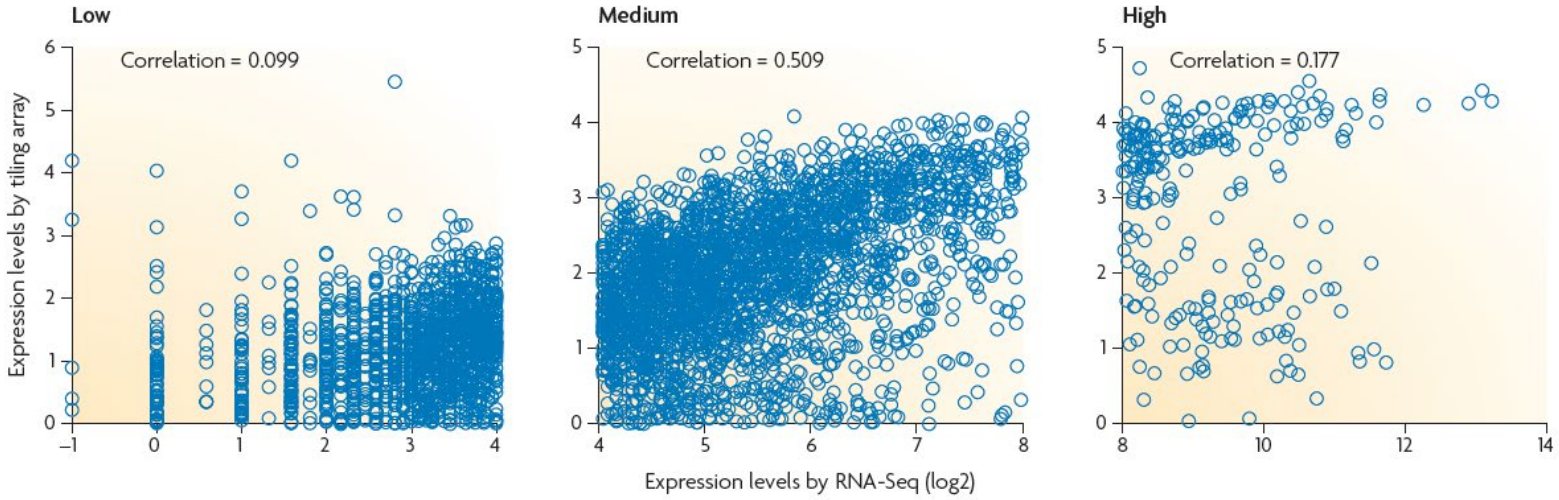
Zhao et al., BMC Genomics, 16:675 (2015)

RNA-Seq versus microarray (I)



't Hoen et al., Nucleic Acids Res (2008)

Wang et al., Nature Reviews Genetics (2009)



RNA-Seq versus microarray (II)

- Microarrays are not cheaper anymore
- RNA-seq has a wider measurement range
  - Lowly expressed transcripts:
    - Microarrays have high background signal -> poor measurement
    - RNA-seq can measure well if you sequence very deeply
  - Medium expressed transcripts:
    - Microarrays measure well
    - RNA-seq measures well if sequenced relatively deeply
  - Highly expressed transcripts:
    - Microarrays measure poorly because of saturation
    - RNA-seq measures well

RNA-Seq: a panacea?

“One particularly powerful advantage of RNA-Seq is that it can capture transcriptome dynamics across different tissues or conditions without sophisticated normalization of data sets’’

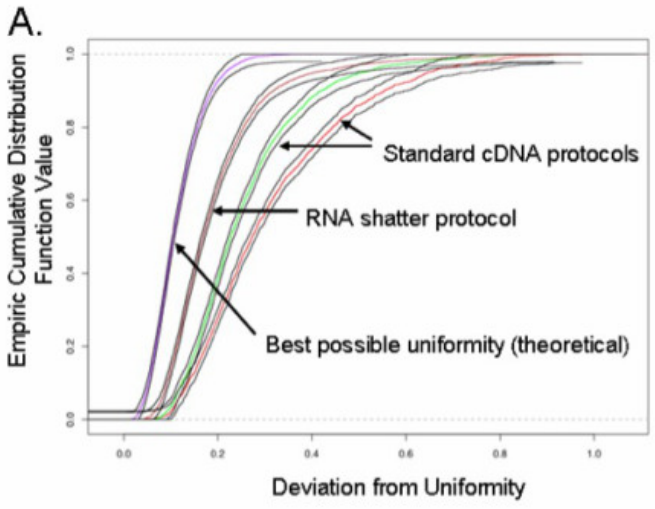
Wang et al., Nature Reviews Genetics (2009)

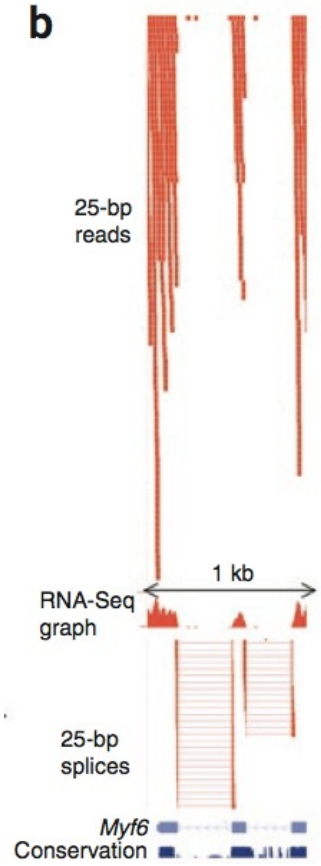
Normalization: remove systematic technical effects so that technical bias has minimal impact on the results

Really?

Mortazavi et al. (Nature Methods, 5(7):621, 2008) already identified various types of bias:

- Transcript length & library size
- Non-uniformity of coverage:
  - Can be reduced by adapting experimental protocols for library preparation

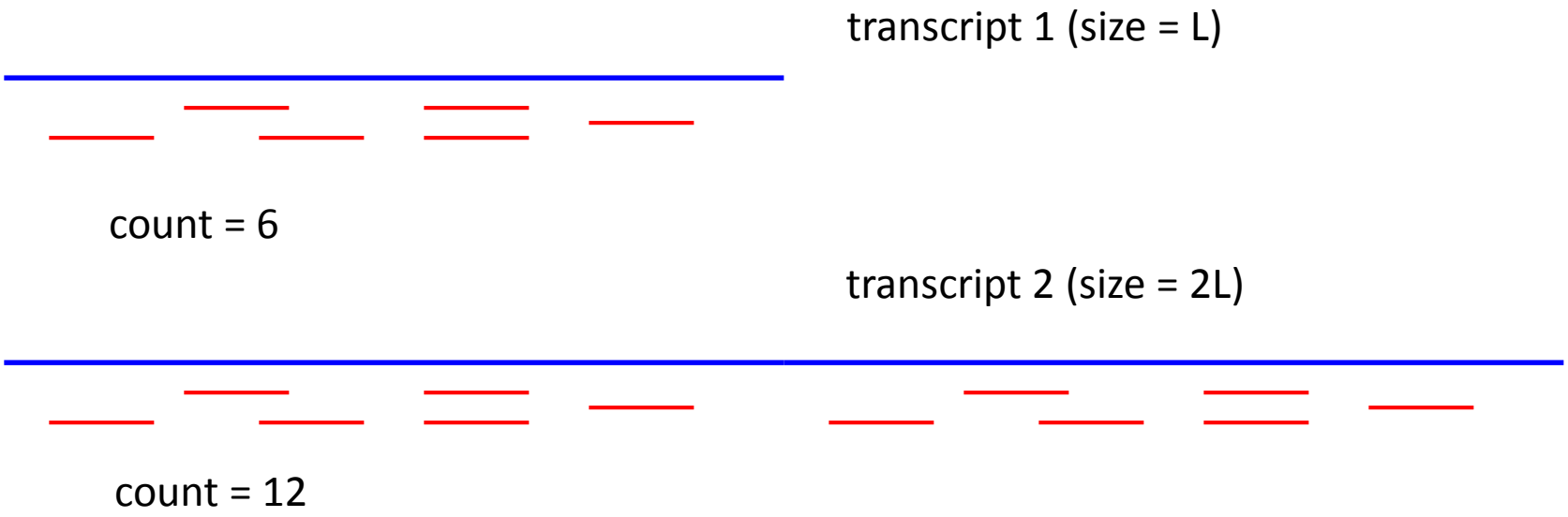




RNA-seq: types of bias

- Transcript length
- Library size
- Mappability of reads
  - Lower sequence complexity, repeats, ...
- Position
  - Fragments are preferentially located towards either the beginning or end of transcripts
- Sequence-specific
  - Likelihood of fragments being selected is biased by nucleotide composition
  - GC content: %GC

Transcript length: different transcripts, within sample



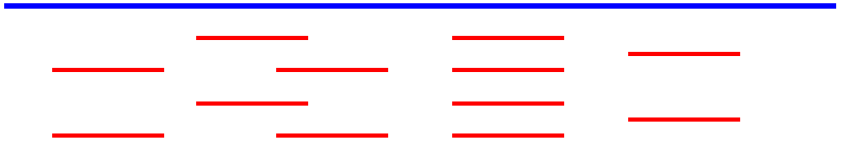
One cannot conclude that transcript 2 has a higher expression than transcript 1:
✓correct for transcript length

Library size: same transcript, between samples

transcript 1 (sample 1)

count = 6, library size = 600

transcript 1 (sample 2)



count = 12, library size = 1200

One cannot conclude that the transcript has a higher expression in sample 2 than in sample 1:

✓correct for library size

RPKM: Reads per kilobase per million mapped reads

- Unit of measurement

$$\text{RPKM} = \#\text{mapped reads} \times \frac{1000\text{bases} \times 10^6}{\text{length of transcript} \times \text{total \# of mapped reads}}$$

- RPKM reflects the molar concentration of a transcript in the starting sample by normalizing for
  - RNA length
  - Total number of reads in the measurement

- This facilitates comparison of transcript levels within and between samples

Mortazavi et al., Nature Methods, 5(7):621 (2008)

RPKM: Example

$$\text{RPKM} = \#\text{mapped reads} \times \frac{1000\text{bases} \times 10^6}{\text{length of transcript} \times \text{total \# of mapped reads}}$$

- Example 1:
  - 2500 base transcript with 900 alignments in a sample of 10 million reads (out of which 8 million reads can be mapped):

$$\text{RPKM} = 900 \times \frac{1000 \times 10^6}{2500 \times 8 \times 10^6} = 45$$

- Example 2:
  - Given a 40M read measurement, how many reads would we expect for a 1 RPKM measurement for a 2kb transcript?

$$\text{RPKM} = C \times \frac{1000 \times 10^6}{2000 \times 40 \times 10^6} = 1 \rightarrow C = 80$$

FPKM: Fragments per K per M

Difference between FPKM and RPKM?

- Paired-end RNA-Seq experiments produce two reads per fragment, but that doesn't necessarily mean that both reads will be mappable. For example, if the second read is of poor quality.
- If we were to count reads rather than fragments, we might double-count some fragments but not others, leading to a skewed expression value.
- Thus, FPKM is calculated by counting fragments, not reads.

Trapnell et al., Nature Biotechnology, 28(5):511 (2010)

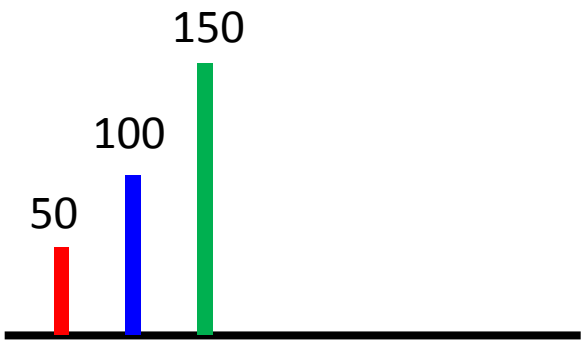
Normalization: between samples

- Differential gene expression – same gene *between* samples
  - Technical biases (gene length and nucleotide composition) are canceled out
  - Between-sample normalization is still essential for comparing counts from different libraries relative to each other.
- Simplest and commonly used normalization
  - Scale by the total number of reads in the library
- Problem if samples have a very different “composition”
  - Highly expressed genes present in only one condition lead to “undersampling” for the remaining genes

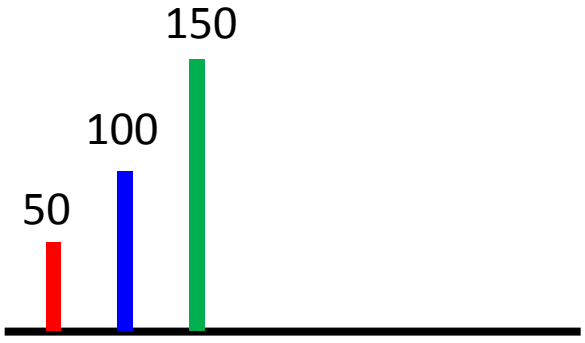
Thought experiment (I)

Suppose

- Two RNA populations (samples): A and B
- The same three genes expressed in both samples
- Numbers indicate number of transcripts / cell



Condition A



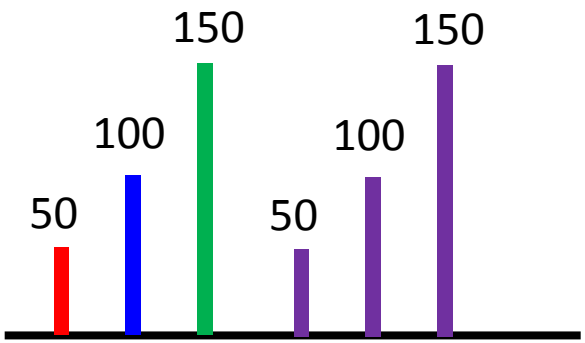
Condition B

No differential expression of these genes

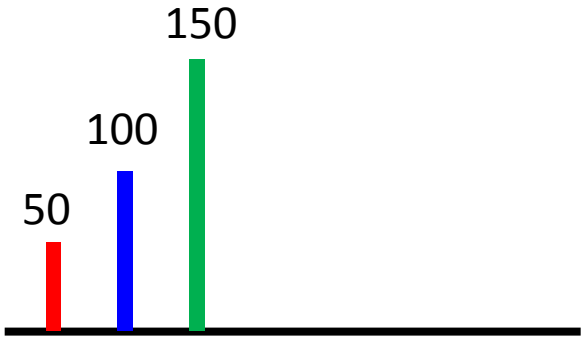
Thought experiment (II)

Suppose

- Two RNA populations (samples): A and B
- The same 3 genes expressed in both samples
- Numbers indicate number of transcripts / cell
- Now condition A has 3 additional genes not in B with equal number and expression



Condition A



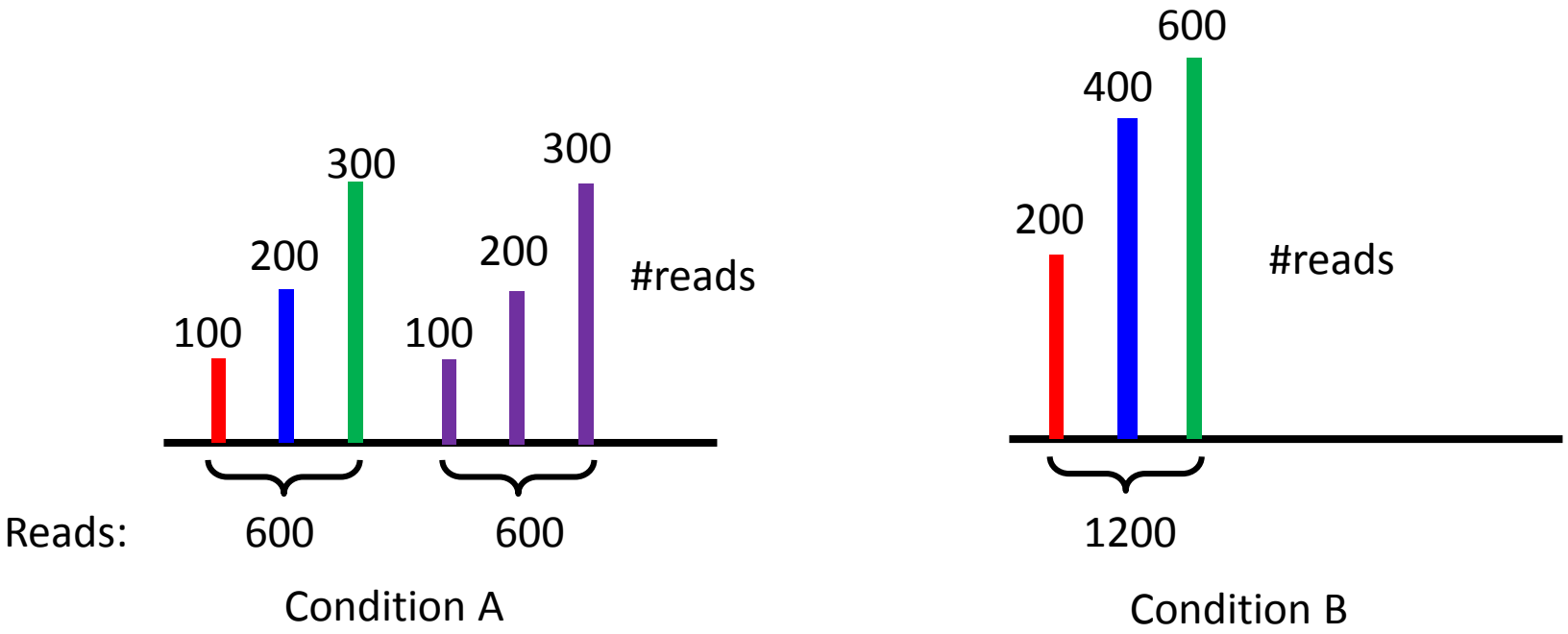
Condition B

Still no differential expression of first three genes. However, RNA production in A is twice the one in B.

Thought experiment (III)

Suppose we sequence both samples with the same depth (1200 reads)

These reads get proportionally 'distributed' over the expressed genes



(1) Correct normalization factor would adjust condition A by a factor of two

(2) Proportion of reads attributed to a gene in one library depends on expression properties of whole sample ➔ If a sample has larger RNA production , RNA-seq will undersample many genes

RPKM would fail in this example

$$\text{RPKM} = \frac{\#\text{ mapped reads } \times 10^6}{\text{total number of mapped reads}}$$

(assuming transcript lengths are the same)

In this example:

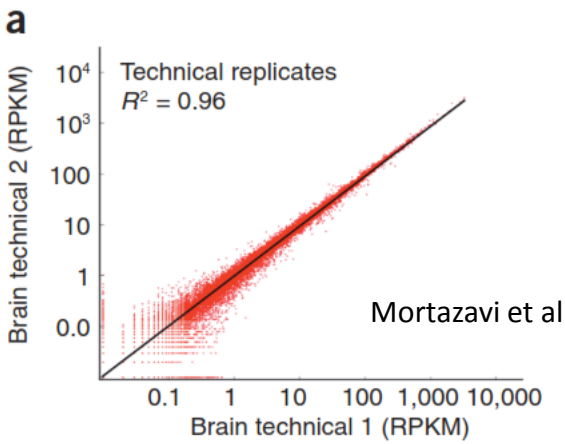
Condition A, first (red) gene: $$RPKM = \frac{100 \times 10^6}{1200} = 83333$$

Condition B, first (red) gene: $$RPKM = \frac{200 \times 10^6}{1200} = 166666$$

RPKM normalization would result in differential expression: we did not take total RNA production into account

When does RKPM fail?

- If samples have largely different RNA production
  - Many unique genes and/or highly expressed genes
  - If many genes in one sample have a very high expression compared to the other samples
- If RNA sample is contaminated
  - Reads that represent the contamination will take away reads from the true sample, thus dropping the number of reads of interest.
- If you can assume that your samples are 'comparable' then RPKM is OK
  - e.g., technical replicates



Mortazavi et al.

Taking total RNA production into account

- Total RNA production of sample *k* (S_k) cannot be estimated directly
- Relative RNA production of two samples:

$$f = \frac{S_1}{S_2}$$

Essentially a global fold change

can more easily be determined

- Assumption that the majority of the genes are not differentially expressed
- TMM: Trimmed Mean of M-values

Robinson and Oshlack, Genome Biology, 11(3):R25 (2010)

Y_{ig} = read counts for gene g in sample $i = 1, 2$

N_i = total read counts for sample $i = 1, 2$

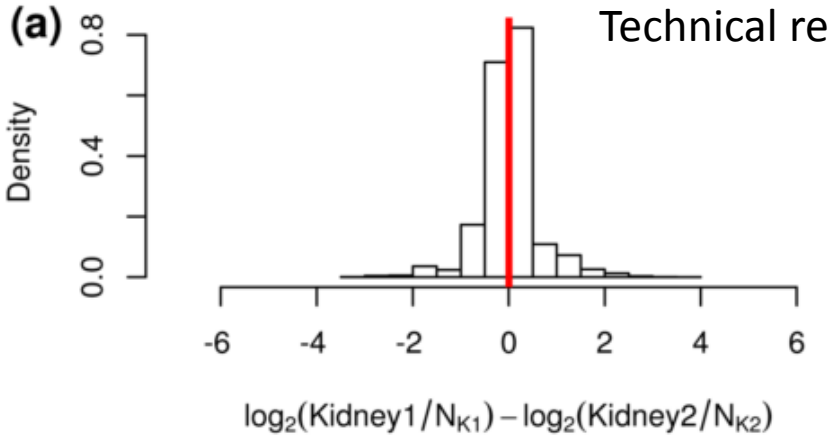
$$M = \log\left(\frac{Y_{1g}}{N_1}\right) - \log\left(\frac{Y_{2g}}{N_2}\right)$$

Gene-wise log-fold-change

$$A = \frac{1}{2}\left[\log\left(\frac{Y_{1g}}{N_1}\right) + \log\left(\frac{Y_{2g}}{N_2}\right)\right]$$

Gene-wise average expression

Example: TMM normalization (I)



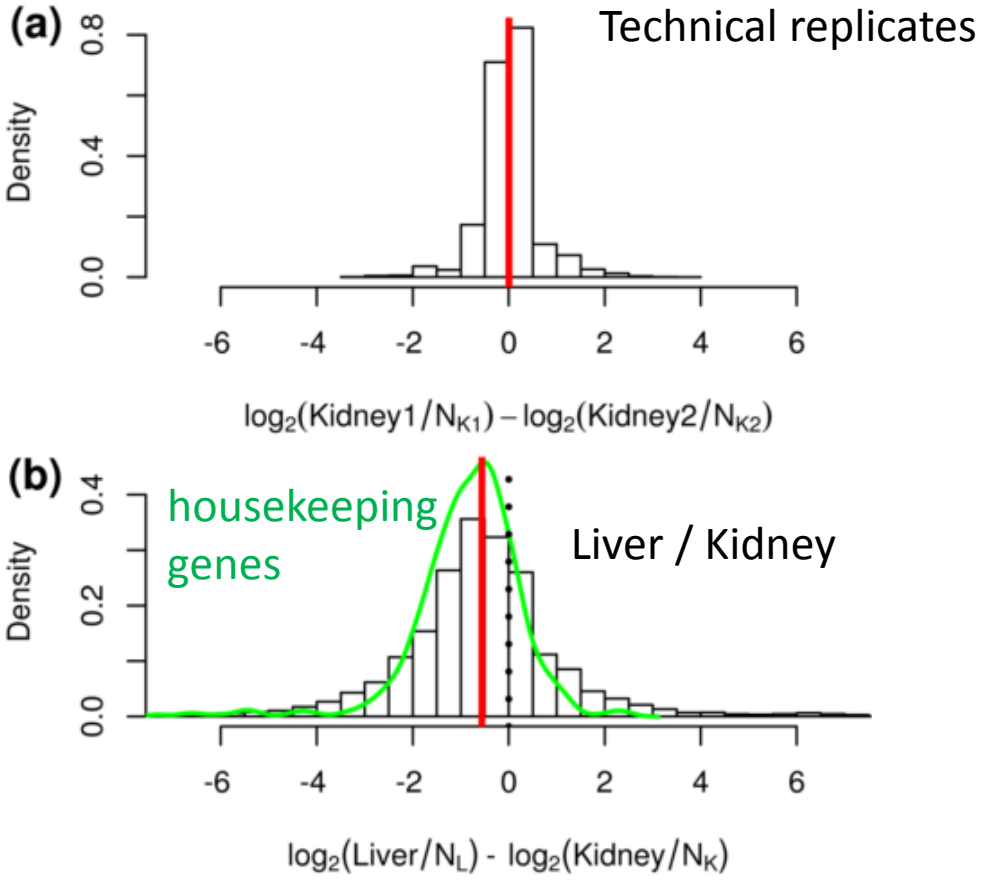
Technical replicates

$$M = \log\left(\frac{Y_{1g}}{N_1}\right) - \log\left(\frac{Y_{2g}}{N_2}\right)$$

mean log ratio ~0

Marioni et al., Genome Research, 18(9):1509 (2008)

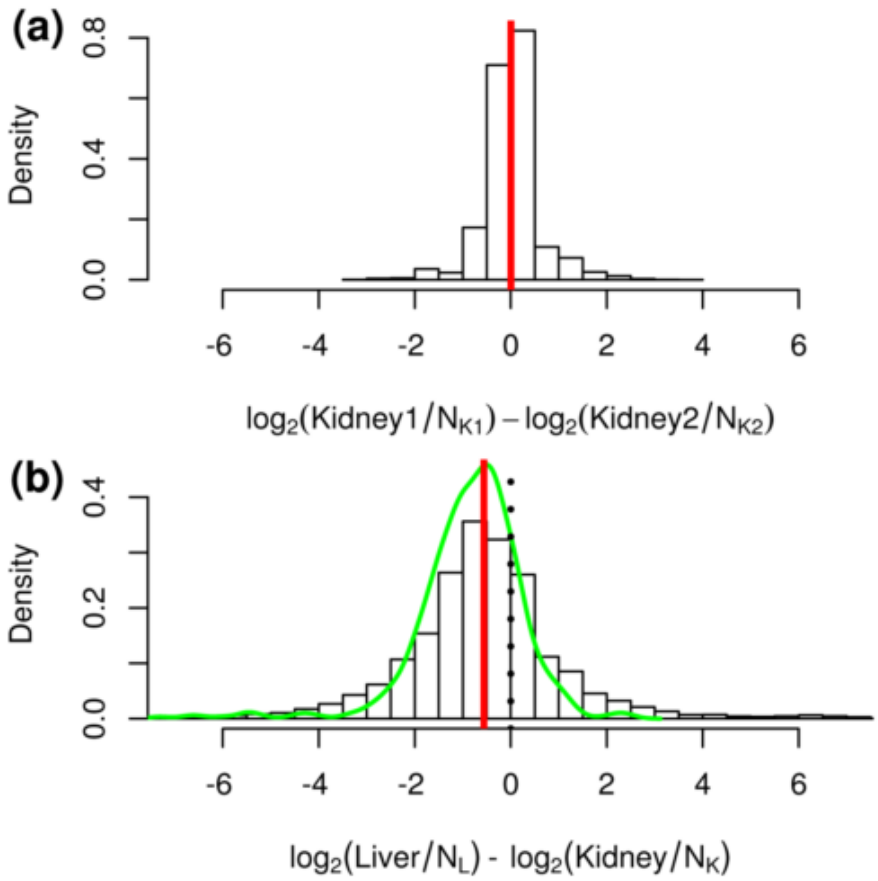
Example: TMM normalization (II)

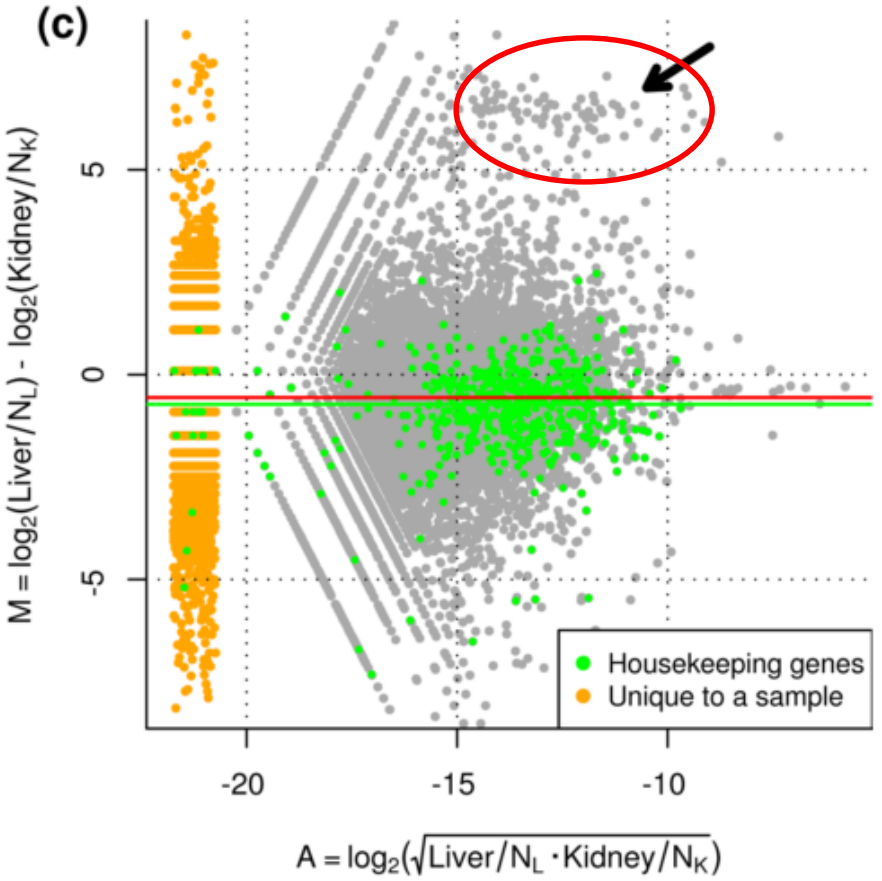


mean log ratio shifted to higher kidney expression

Example: TMM normalization (III)

A few strongly expressed, differentially expressed genes in liver
→ less sequence reads available for bulk of lower expressed liver genes
→ ratio = liver/kidney becomes smaller (i.e., shift of distribution towards kidney)





— TMM normalization factor

Then, from the trimmed subset of genes, calculate a relative scaling factor from a weighted average of M –values (for sample *k* and reference sample *r*, gene *g)*:

$$\lambda^{TMM} = \log_2(TMM_k^{(r)}) = \frac{\sum_{g \in G^*} w_{gk}^r M_{gk}^r}{\sum_{g \in G^*} w_{gk}^r}$$

$$f_k = \frac{S_1}{S_2}$$

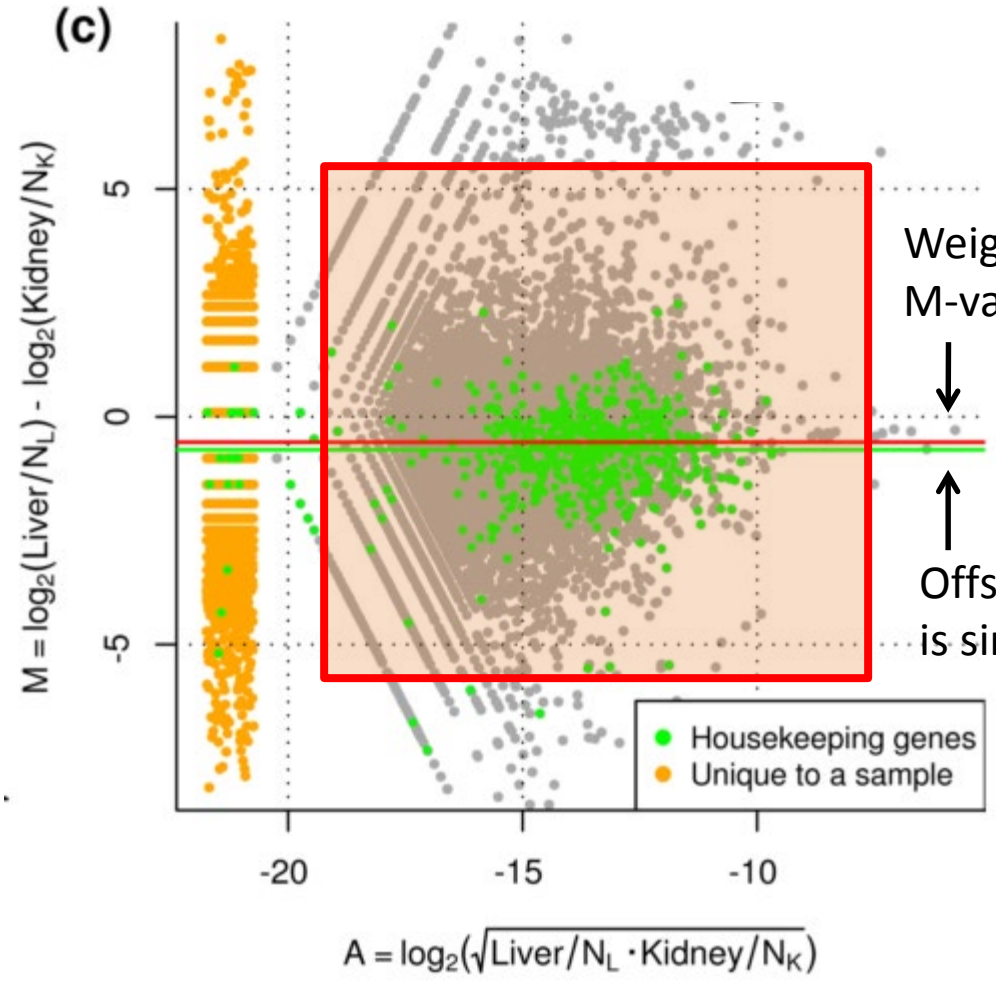
$$w_{gk}^r = \frac{N_k - Y_{gk}}{N_k Y_{gk}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}} \approx \frac{1}{\text{variance}}$$

Implemented in edgeR (R/Bioconductor). Similar method is used in DESeq2 (R/BioC).

A few strongly expressed, differentially expressed genes in liver
→ less sequence reads available for bulk of lower expressed liver genes
→ ratio=liver/kidney becomes smaller (i.e., shift of distribution towards kidney)

Trim the data

M: 30%, A: 5%



Weighted average of M-values: λ^{TMM}

Offset for HK genes is similar to λ

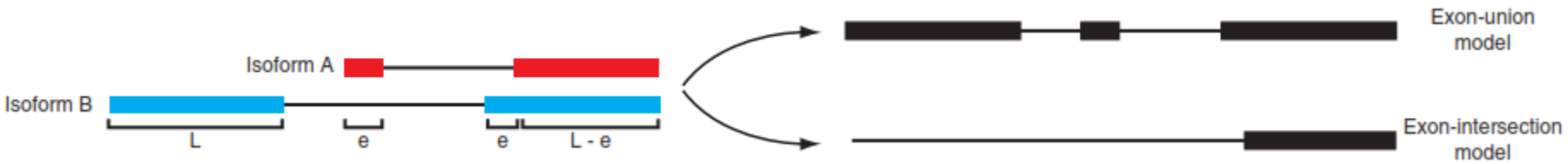
Other normalization methods

- Spike-ins
- Housekeeping genes (Bullard et al., 2010)
- Upper-quartile (Bullard et al., 2010). Counts are divided by (75^{th}) upper-quartile of counts for transcripts with at least one read
- Quantile normalization (Irizarry et al., 2003; developed for microarrays)

Comparison of normalization methods (Dillies et al., 2013)

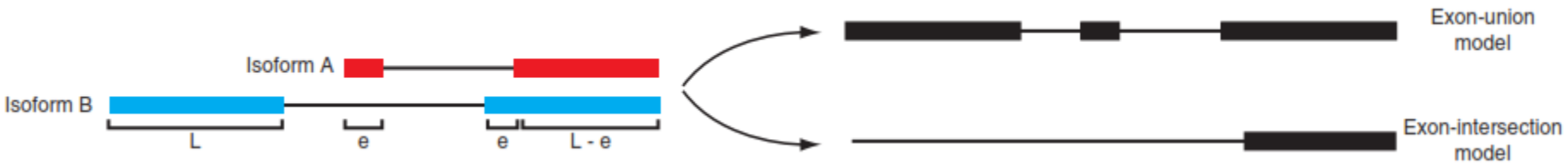
Bullard et al. (2010) BMC Bioinformatics, 11:94.
Irizarry et al. (2003) Biostatistics, 4(2): 249–64.
Dillies et al. (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing. Briefings in Bioinformatics

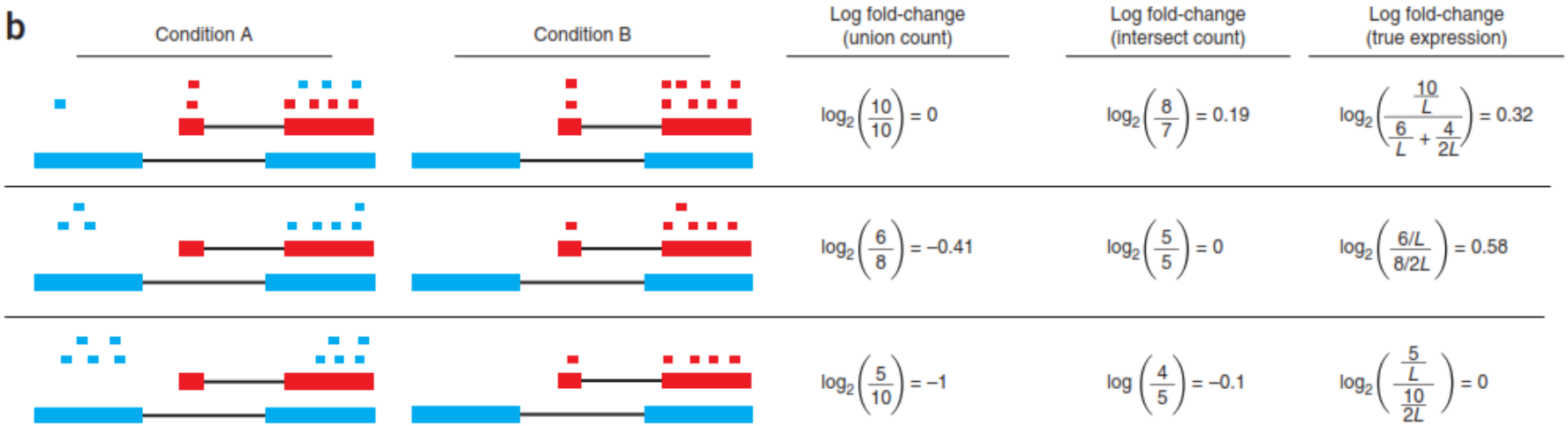
Read counting with isoforms: transcript length (I)



Trapnell et al., Nature Biotechnology (2013)

Read counting with isoforms: transcript length (II)



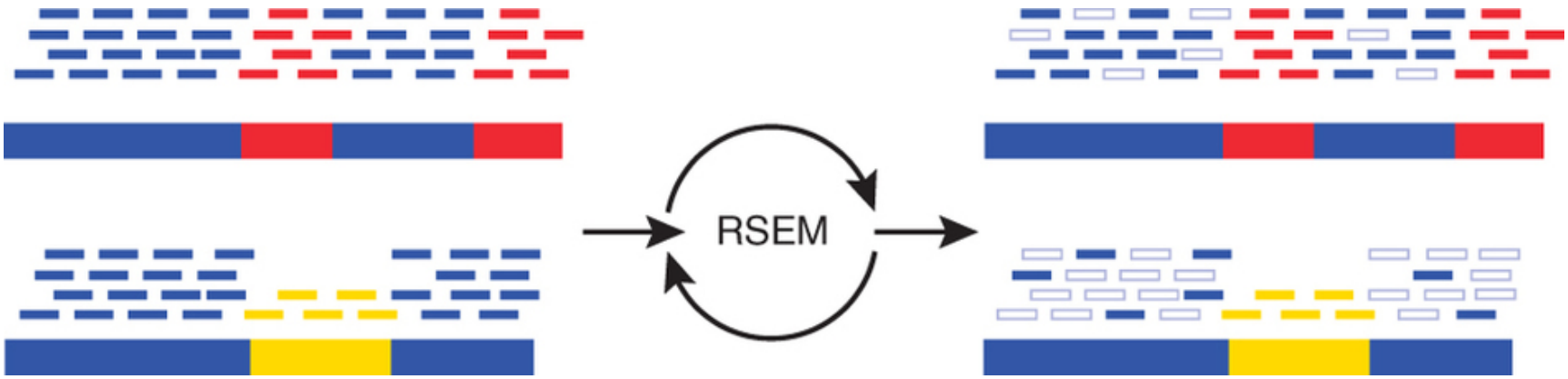


$$\log_2\left(\frac{B}{A}\right)$$

Correct for isoform length in case of different transcripts

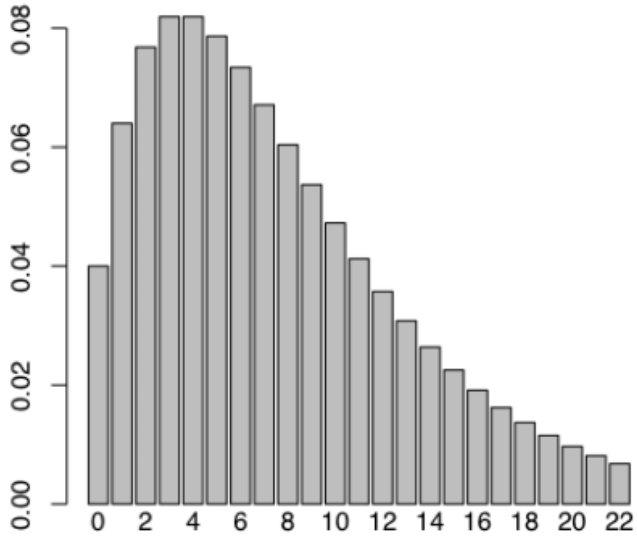
Trapnell et al., Nature Biotechnology (2013)

Transcript abundance: Expectation-Maximization



Differential expression analysis: challenges with RNA-Seq count data

- discrete, positive, skewed
  - no (log-)normal model
- small numbers of replicates
  - no rank based or permutation based methods
- sequencing depth (coverage) varies between samples
  - “normalisation”
- large dynamic range (0 ... 10^5)
  - heteroskedasticity matters between genes



Naive approach: non-parametric test (I)

```
# R code
x <- c(0,0,10,4)
y <- c(23,42,0,17)
wilcox.test(x, y)
```

```
        Wilcoxon rank sum test with continuity correction
data: x and y
W = 3, p-value = 0.1832
alternative hypothesis: true location shift is not equal to 0
```

Naive approach: non-parametric test (II)

```
# R code
x <- c(0,10,4)
y <- c(23,19,17)
wilcox.test(x, y)
```

```
        Wilcoxon rank sum test with continuity correction
data: x and y
W = 0, p-value = 0.1
alternative hypothesis: true location shift is not equal to 0
```

More powerful alternatives: edgeR, DESeq2 (and limma/voom)

Distribution

- Assumptions:
  - Several flow cell lanes are filled with aliquots of the *same* prepared library
  - The concentration of a certain transcript species is *exactly* the same in each lane
  - We get the same total number of reads from each lane
- For each lane, count how often you see a read from the transcript. Will the counts all be the same?

Distribution

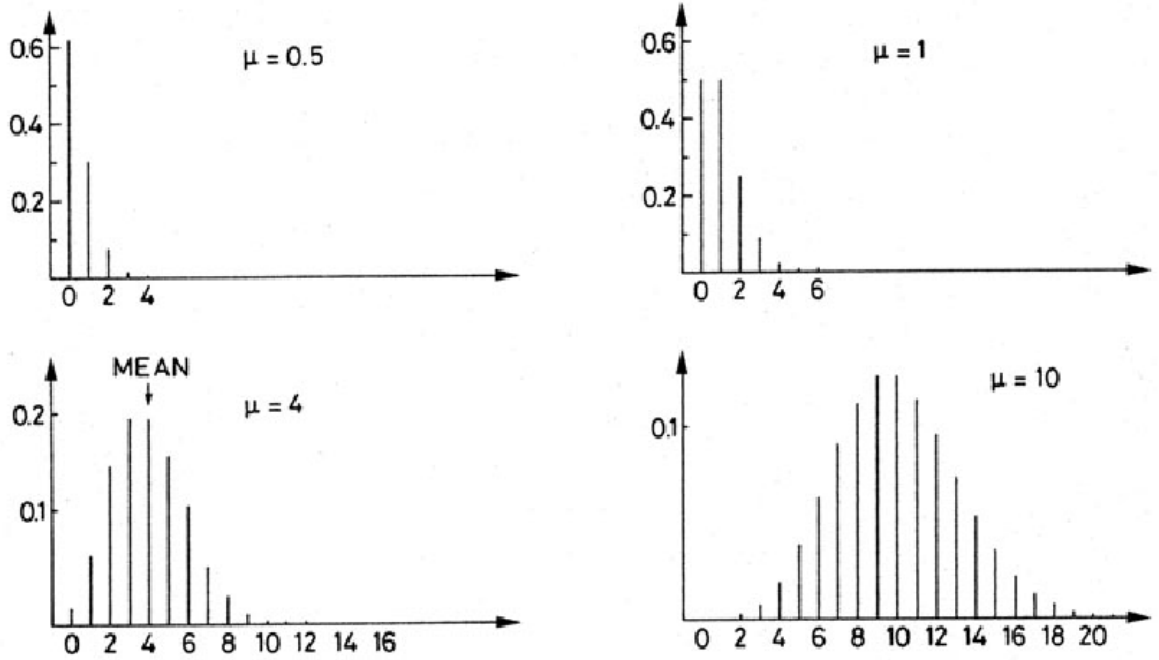
- Assumptions:
  - Several flow cell lanes are filled with aliquots of the *same* prepared library
  - The concentration of a certain transcript species is *exactly* the same in each lane
  - We get the same total number of reads from each lane
- For each lane, count how often you see a read from the transcript. Will the counts all be the same?

- Of course not: even for equal concentrations, the counts will vary. This *theoretically unavoidable* noise is called *shot noise*:

In this case counts are Poisson distributed

Poisson distribution

- The Poisson distribution turns up whenever things are counted
- Example: A short, light rain shower with μ drops/m^2
- What is the probability to find *k* drops on a paving stone of size 1 m^2?



Poisson distribution (II)

- For Poisson-distributed data, the variance is equal to the mean.
- Hence, no need to estimate the variance
  - according to several authors: Marioni et al. (2008), Wang et al. (2010), Bloom et al. (2009), Kasowski et al. (2010), Bullard et al. (2010)
- Really?
- Is HTS count data Poisson-distributed?
- To sort this out, we have to distinguish *two* sources of noise
  1. Shot noise
  2. …

Sample noise

- Now consider
  - Several lanes contain samples from biological replicates
  - The concentration of a given transcript varies around a mean value with a certain standard deviation.

This standard deviation cannot be calculated, it has to be *estimated* from the data.

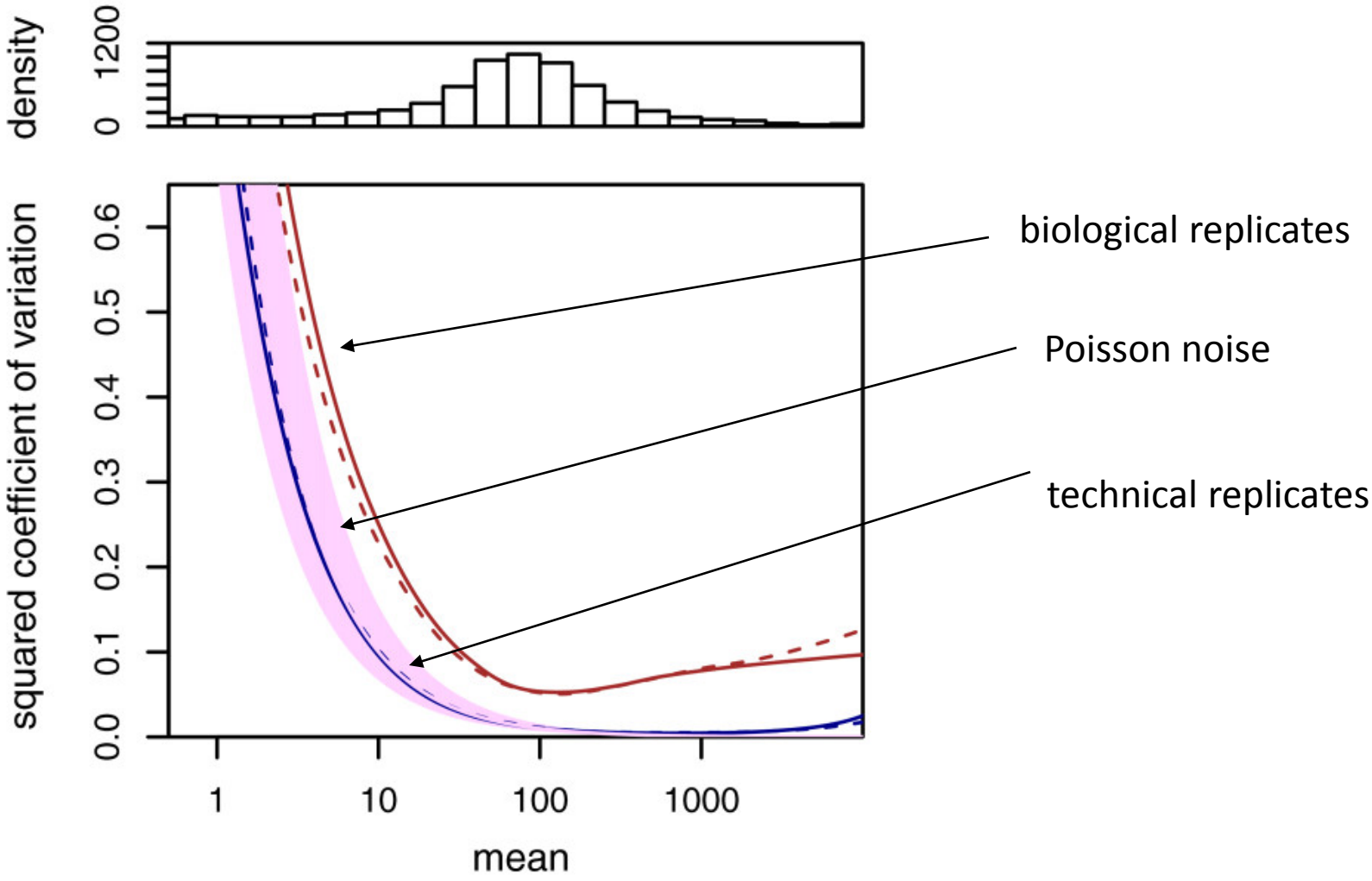
Technical and biological replicates

- Nagalakshmi *et al.* (2008) have found that counts for the same gene from different *technical* replicates have a variance equal to the mean (Poisson)

- Counts for the same gene from different *biological* replicates have a variance exceeding the mean (overdispersion)

- Marioni *et al.* (2008) have looked and confirmed the first fact (and confused everybody by ignoring the second fact).

Technical and biological replicates (II)



Summary: noise

We distinguish:

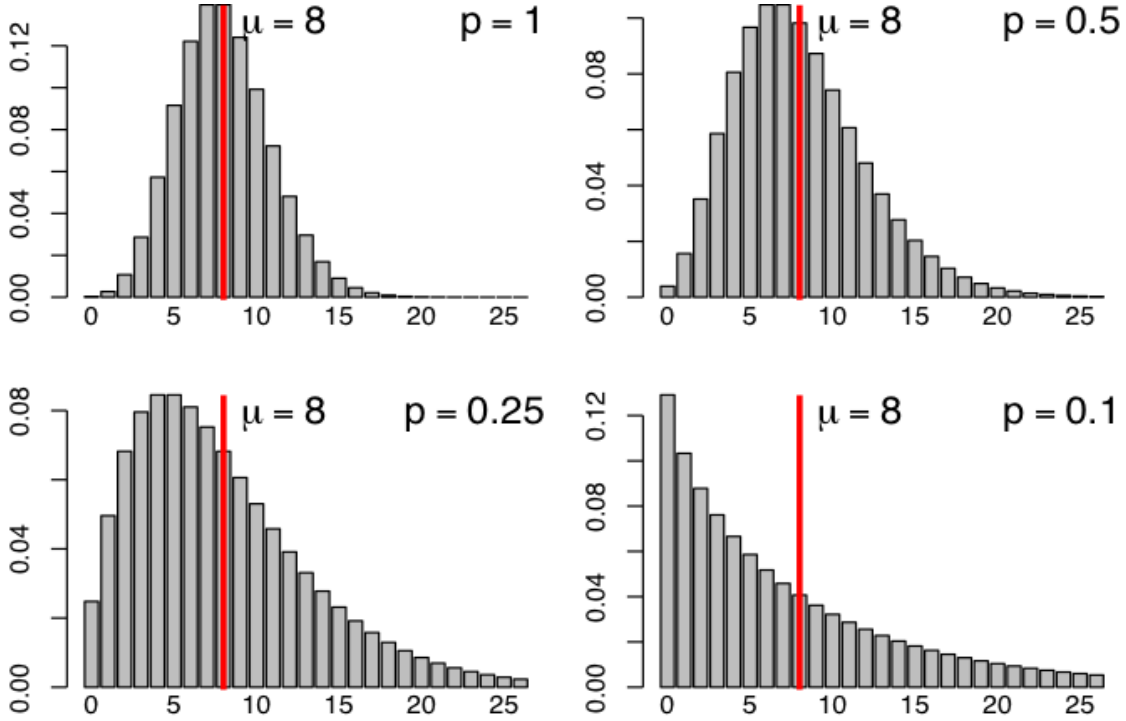
- Shot noise
  - unavoidable, appears even with perfect replication
  - dominant noise for weakly expressed genes
- Technical noise
  - from sample preparation and sequencing
  - negligible (if all goes well)
- Biological noise
  - Unaccounted for differences between samples
  - Dominant noise for strongly expressed genes

Negative-binomial distribution

Variance of negative-binomial distribution: mean + α.mean^2

Shot noise

Technical + biological noise

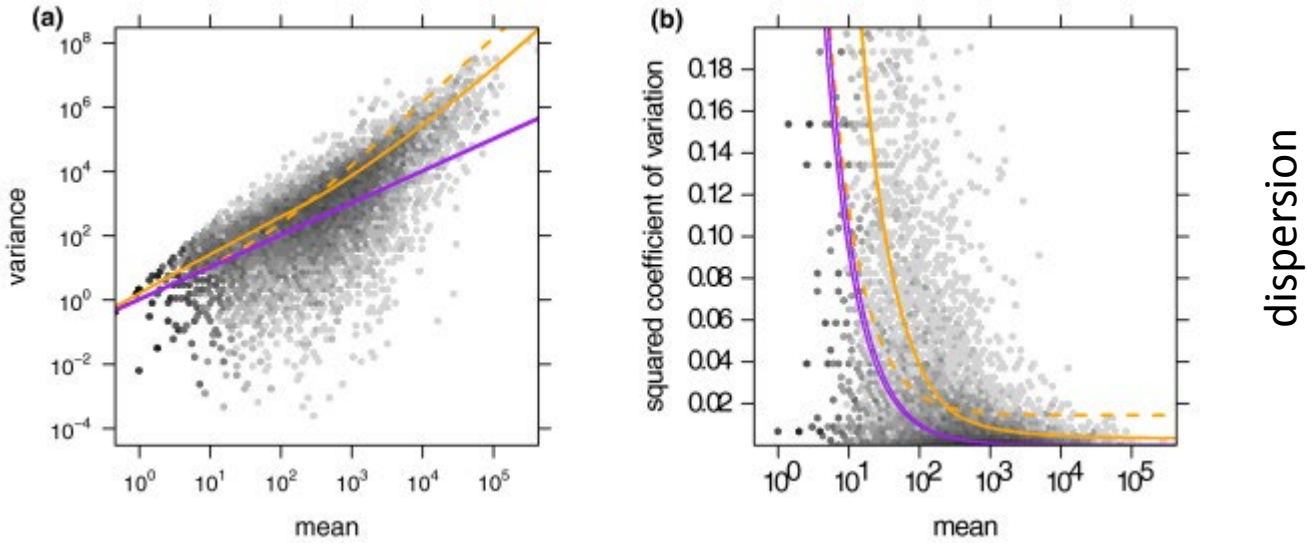


$p=1/\alpha$

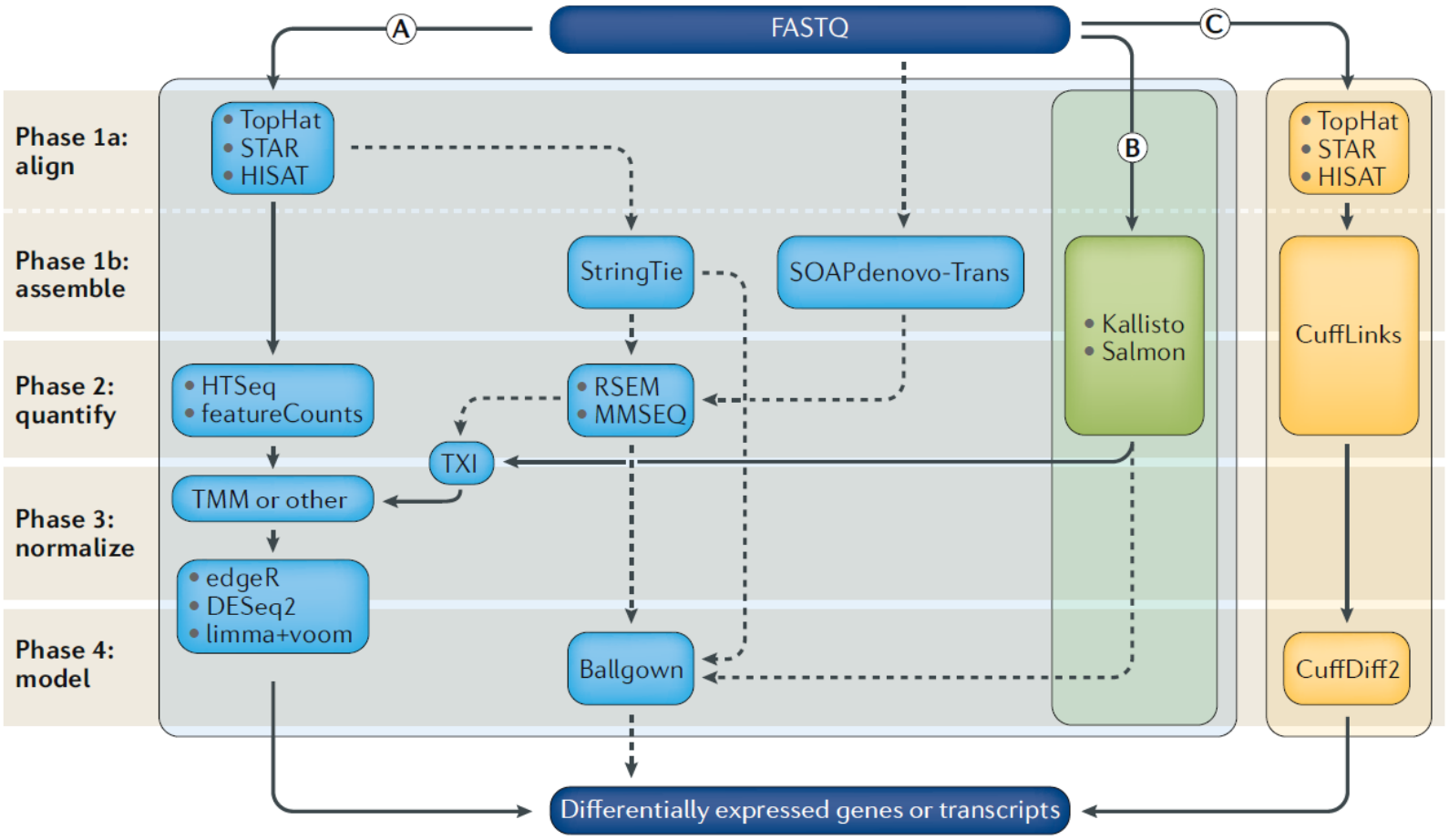
Negative-binomial distribution

- RNA-Seq
  - Often few biological replicates
- Tricks needed to reduce the number of parameters:
  - DESeq(2) & edgeR : mean expression is a good predictor of the variance, i.e., genes with a similar expression level also have similar variance.

    across replicates – fit a smooth curve and then shrink



RNA-Seq: data analysis workflow



Stark et al., Nature Reviews Genetics, 20:631-56 (2019)