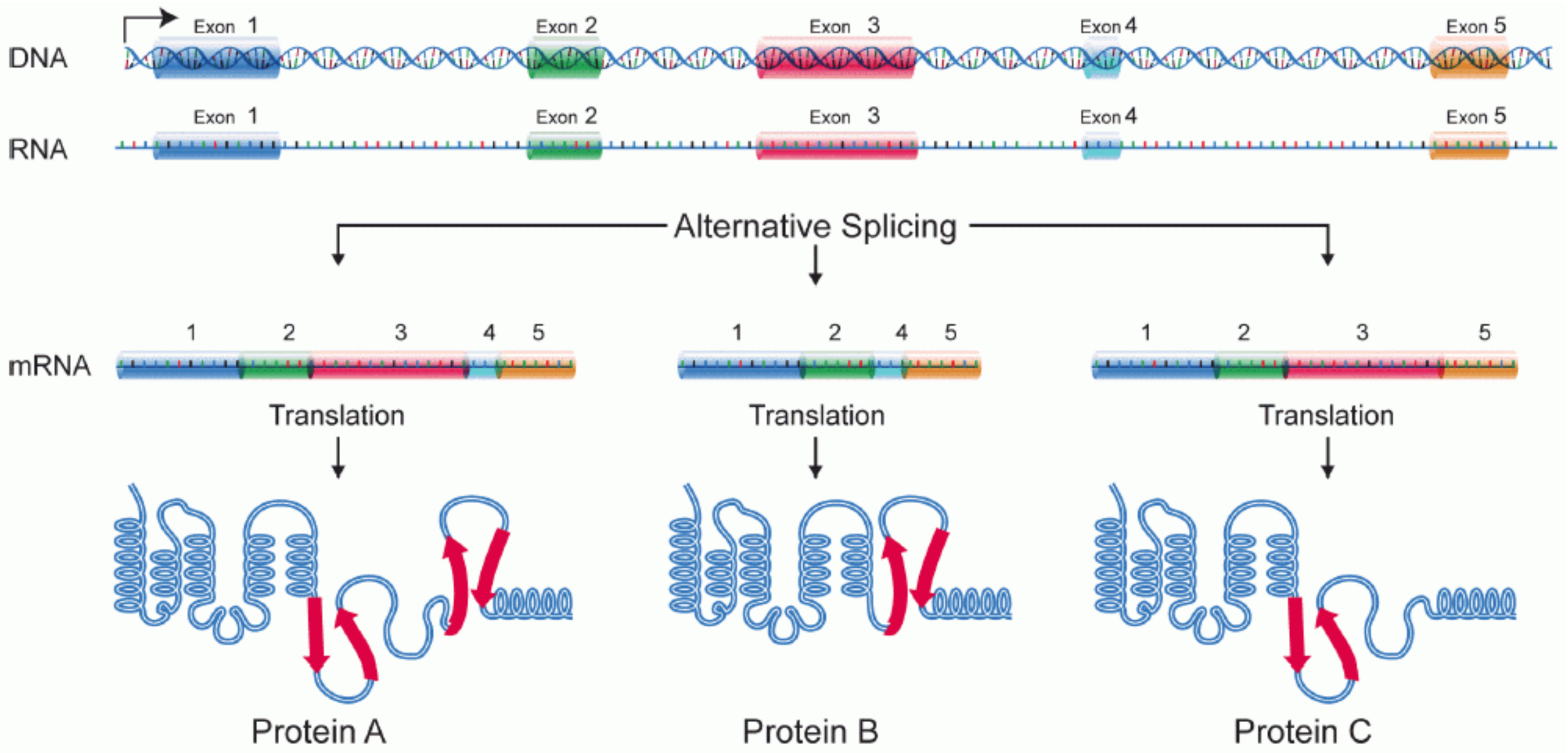
Differential expression analysis

Charlotte Soneson

Friedrich Miescher Institute for Biomedical Research &
SIB Swiss Institute of Bioinformatics

Hinxton, April 9, 2019


DNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
RNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Alternative Splicing
mRNA
1 2 3 4 5
1 2 4 5
1 2 3 5
Translation
Translation
Translation
Protein A
Protein B
Protein C

Wikipedia

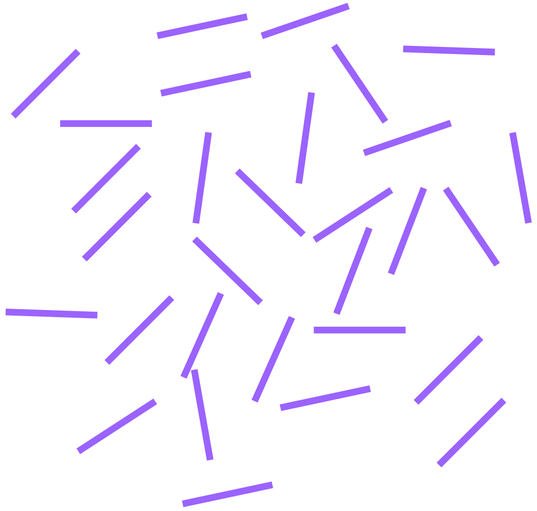
Differential analysis types for RNA-seq

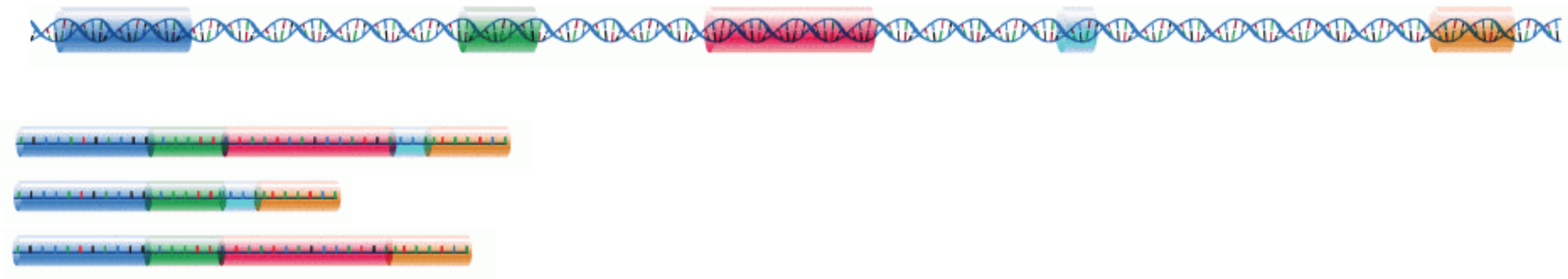
- Does the total output of a gene change between conditions? **DGE**
- Does the expression of individual transcripts change? **DTE**
- Does *any* isoform of a given gene change? **DTE+G**
- Does the isoform composition for a given gene change? **DTU/DIU/DEU**
- (Does *anything* change? GDE*)

- need **different** abundance quantification of transcriptomic features (genes, transcripts, exons)

*https://liorpachter.wordpress.com/2018/02/15/gde%C2%B2-dge%C2%B2-dtu%C2%B2-dte%E2%82%81%C2%B2-dte%E2%82%82%C2%B2/

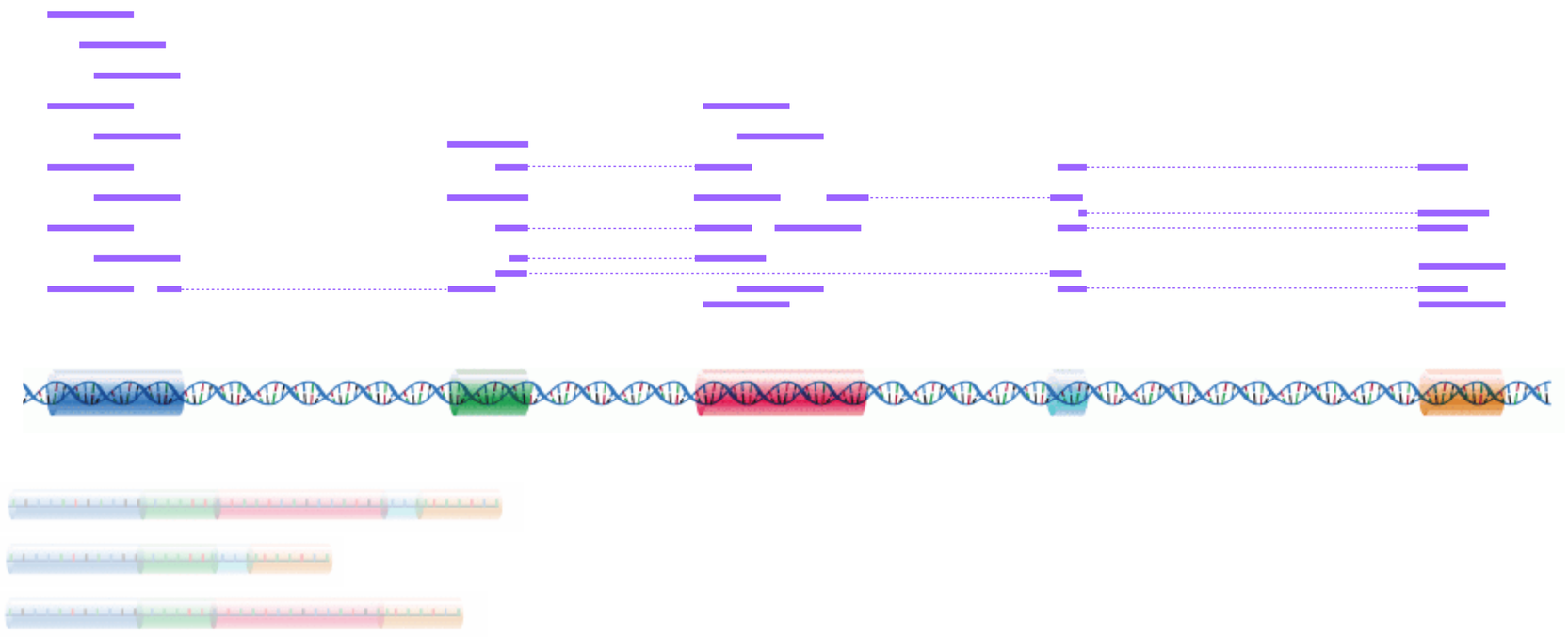
Step I: Abundance quantification





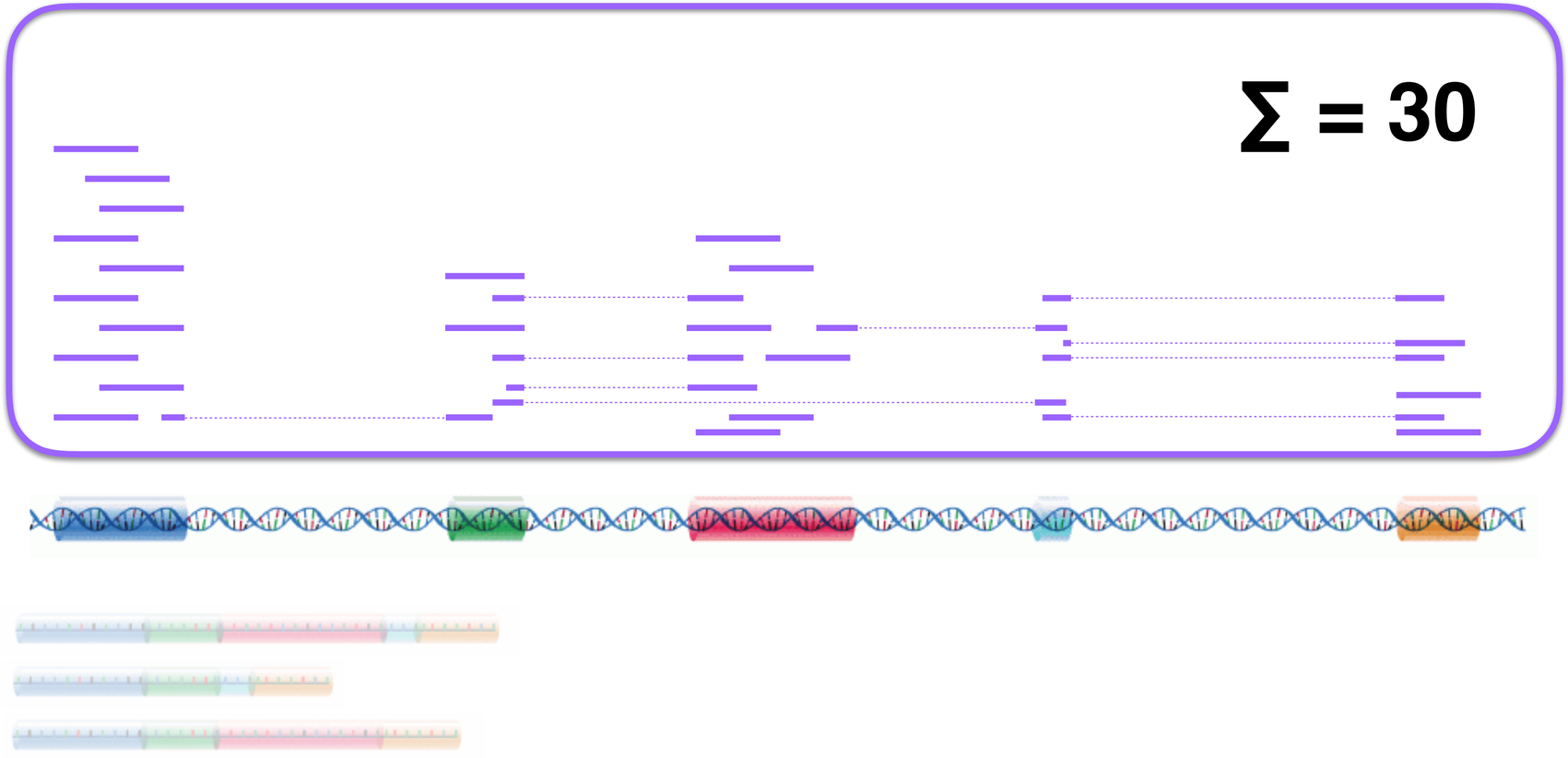
Step I: Abundance quantification

Gene-level counts, often obtained by genome alignment + overlap counting



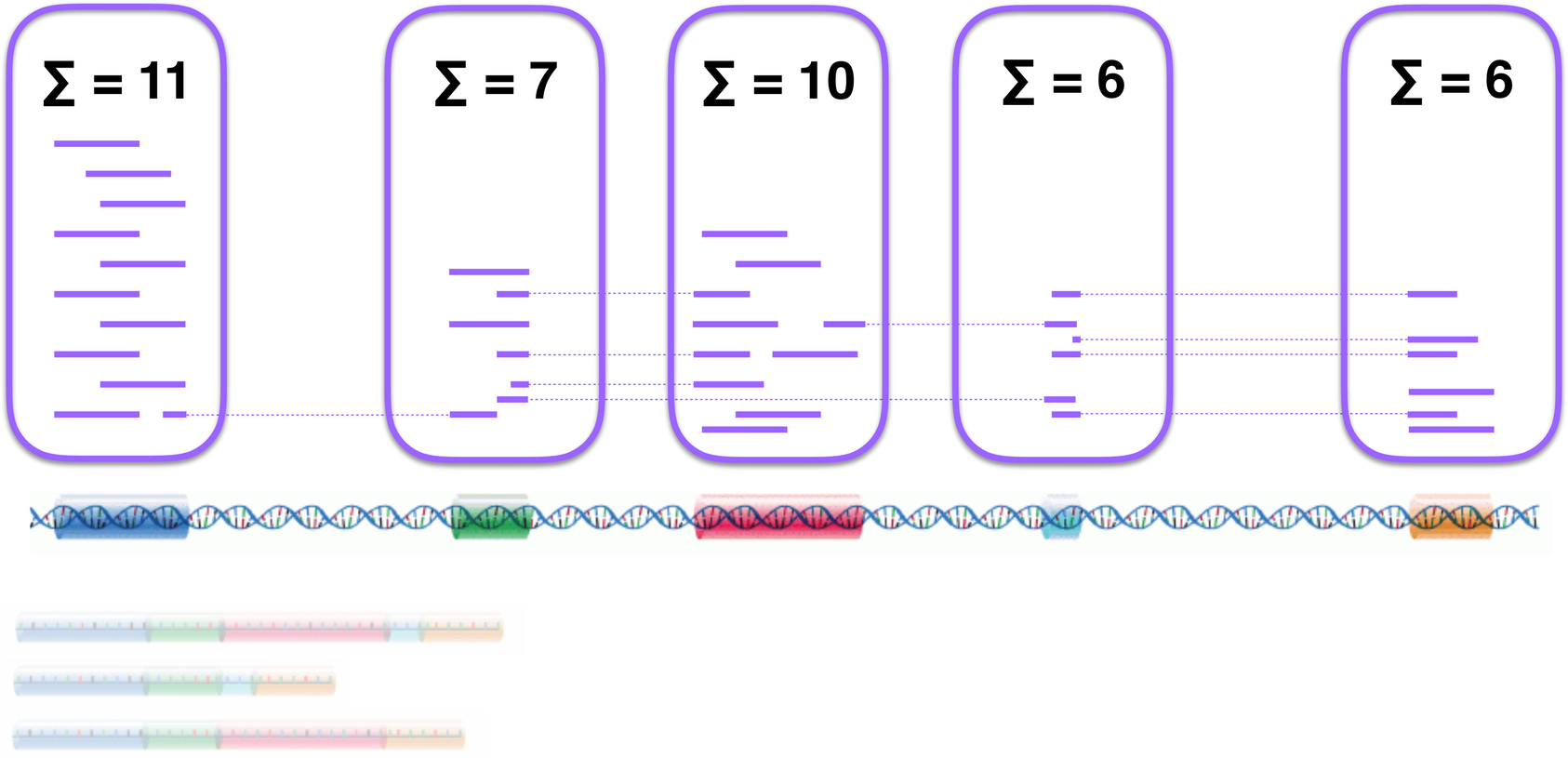
Step I: Abundance quantification

Gene-level counts, often obtained by genome alignment + overlap counting



Step I: Abundance quantification

Exon-level counts, often obtained by genome alignment + overlap counting



Step I: Abundance quantification

Equivalence-class counts, e.g. obtained by "alignment-free" estimation methods

- **Salmon** (Patro et al, *Nat Methods* 2017)
- **kallisto** (Bray et al, *Nat Biotechnol* 2016)

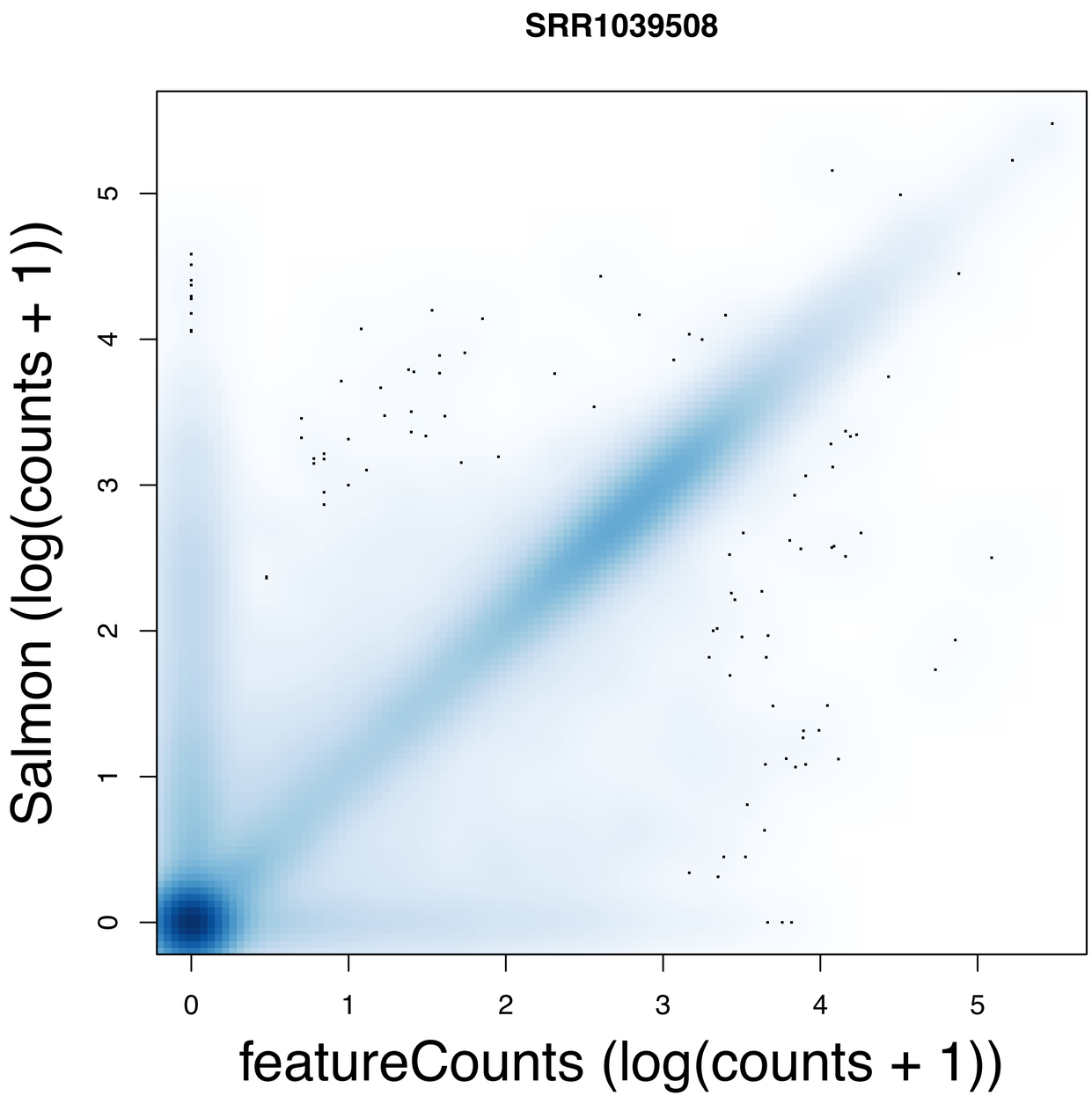
Step I: Abundance quantification

Transcript-level counts, e.g. obtained by "alignment-free" estimation methods

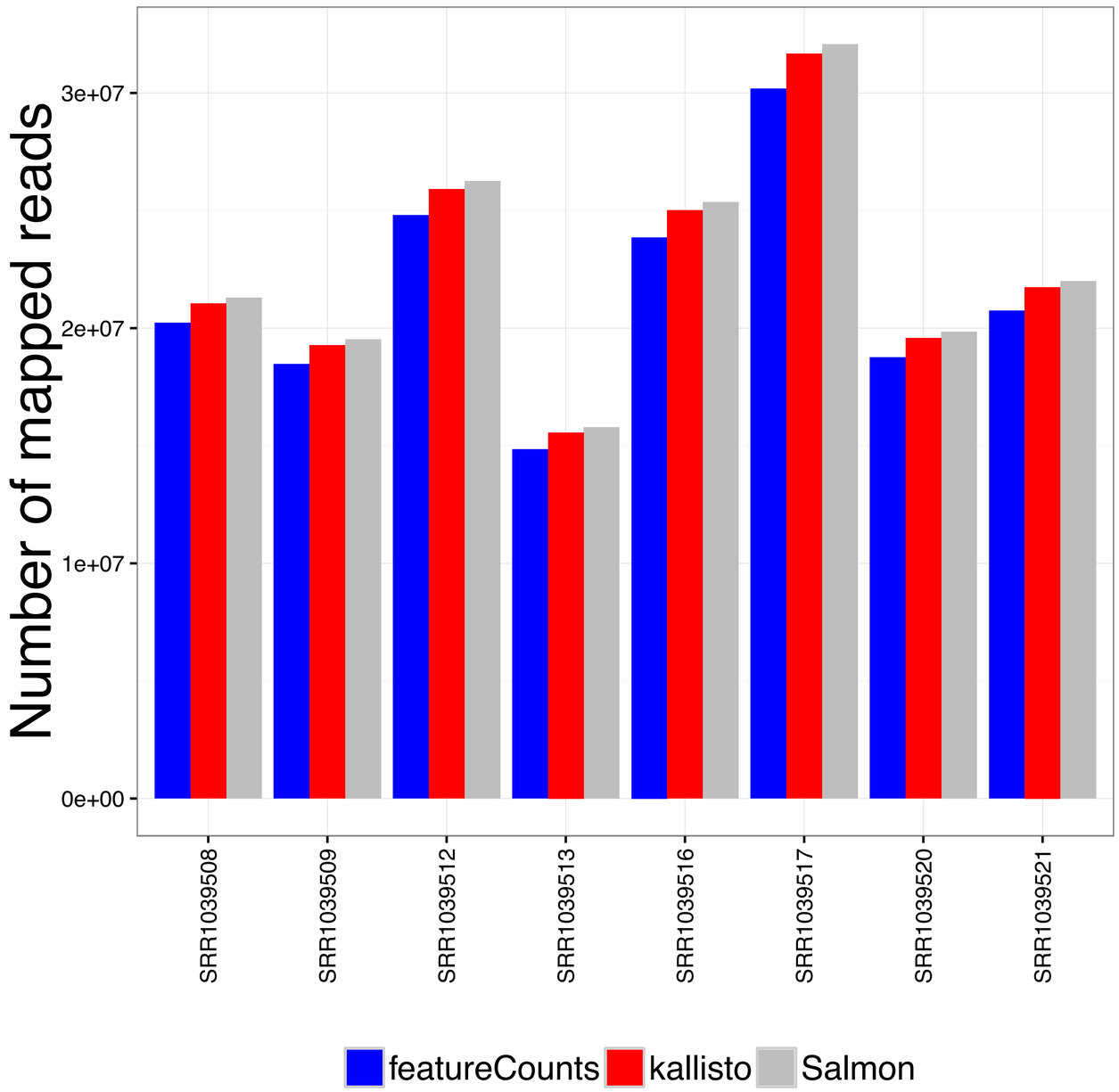
Step I: Abundance quantification

Gene-level counts, obtained by summation of transcript counts

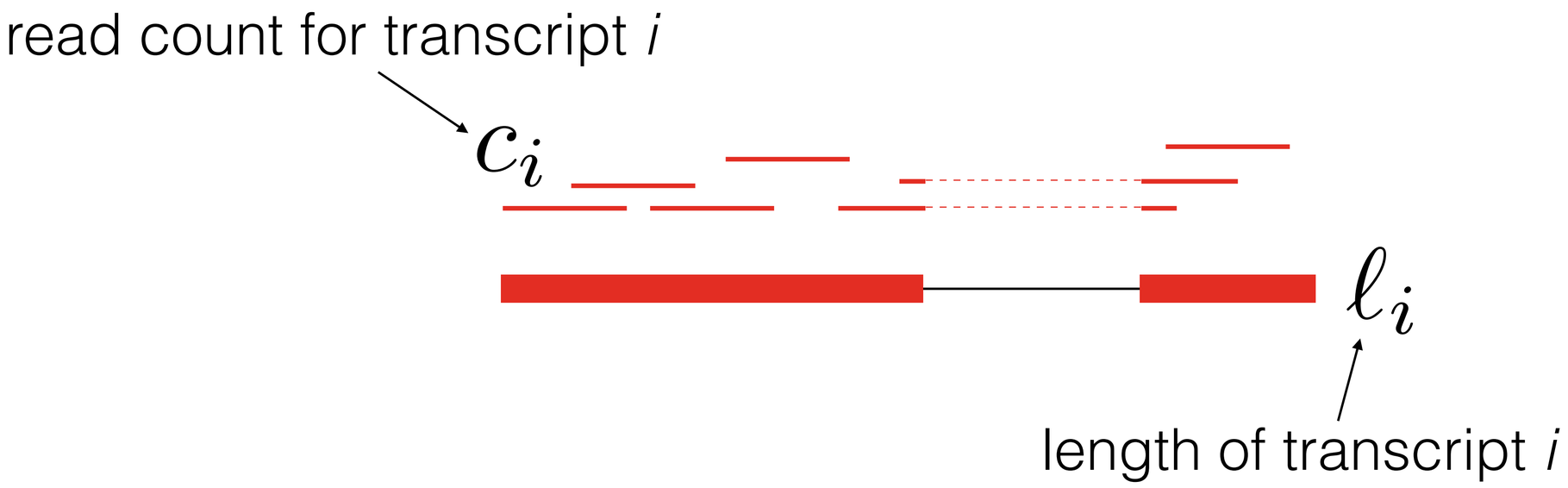
Gene-level counts mostly similar between approaches



kallisto and Salmon can use slightly more reads



Abundance units



Abundance units

read count for transcript *i*

c_i

ℓ_i

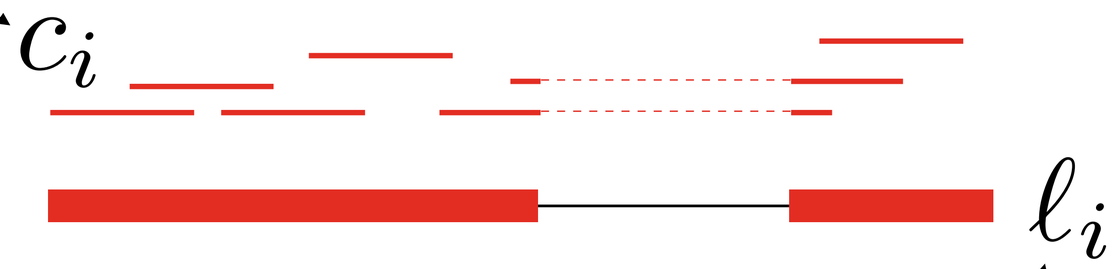
fragment length

length of transcript *i*

$$t_i = \frac{c_i r}{\ell_i}$$

Abundance units

read count for transcript *i*



fragment length

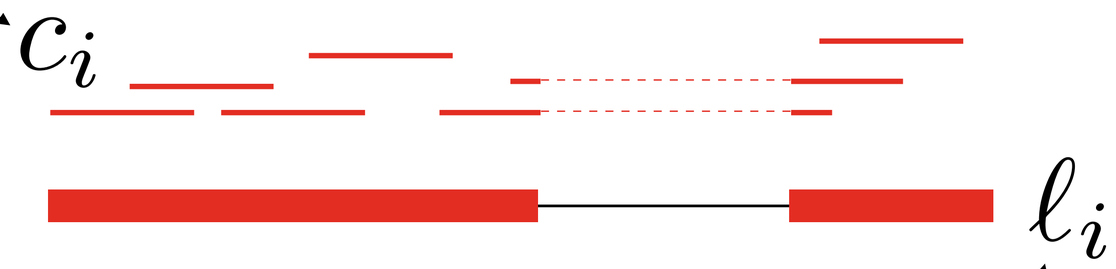
length of transcript *i*

$$t_i = \frac{c_i r}{\ell_i}$$

$$TPM_i = 10^6 \cdot \frac{t_i}{\sum_k t_k}$$

Abundance units

read count for transcript *i*



fragment length

length of transcript *i*

$$t_i = \frac{c_i r}{\ell_i}$$

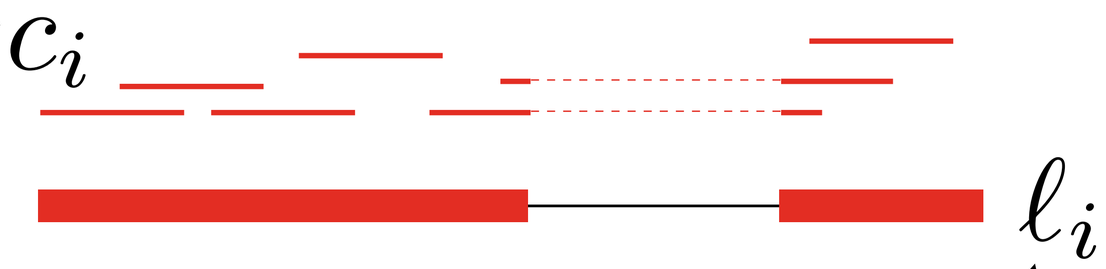
$$TPM_i = 10^6 \cdot \frac{t_i}{\sum_k t_k}$$

library size

$$RPKM_i = 10^9 \cdot \frac{c_i}{\ell_i \boxed{\sum_k c_k}} = 10^9 \cdot \frac{t_i}{\sum_k (t_k \ell_k)}$$

Abundance units

read count for transcript *i*



fragment length

length of transcript *i*

$$t_i = \frac{c_i r}{\ell_i}$$

$$TPM_i \propto RPKM_i$$

$$\sum_i TPM_i = 10^6$$

$$TPM_i = 10^6 \cdot \frac{t_i}{\sum_k t_k}$$

library size

$$RPKM_i = 10^9 \cdot \frac{c_i}{\ell_i \boxed{\sum_k c_k}} = 10^9 \cdot \frac{t_i}{\sum_k (t_k \ell_k)}$$

An example why FPKM cannot compare across samples

True expression

	true exp. 1	**sample 1**	true exp. 2	**sample 2**
gene 1	100	**10**	100	**20**
gene 2	100	**10**	100	**20**
gene 3	100	**10**	100	**20**
gene 4	100	**10**	0	**0**
gene 5	100	**10**	0	**0**
gene 6	100	**10**	0	**0**
total reads	600	**60**	300	**60**

Measured expression

	true exp. 1	**RPKM 1**	true exp. 2	**RPKM 2**
gene 1	100	**0.166..**	100	**0.333..**
gene 2	100	**0.166..**	100	**0.333..**
gene 3	100	**0.166..**	100	**0.333..**
gene 4	100	**0.166..**	0	**0**
gene 5	100	**0.166..**	0	**0**
gene 6	100	**0.166..**	0	**0**
total reads	600	**1**	300	**1**

FPKM

Differential expression analysis

- Input: expression/abundance matrix
  (features x samples) + grouping/sample annotation

	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516	SRR1039517	SRR1039520	SRR1039521
ENSG00000000003	693	451	887	416	1148	1069	774	581
ENSG00000000005	0	0	0	0	0	0	0	0
ENSG00000000419	466	515	623	364	590	794	419	510
ENSG00000000457	326	274	372	223	356	450	308	297
ENSG00000000460	91	75	61	48	110	95	100	82
ENSG00000000938	0	0	2	0	1	0	0	0

- Output: result table (one line per feature)

	logFC	logCPM	LR	PValue	FDR
ENSG00000109906	-5.882117	4.120149	924.1622	5.486794e-203	3.493826e-198
ENSG00000165995	-3.236681	4.603028	576.1025	2.641667e-127	8.410672e-123
ENSG00000189221	-3.316900	6.718559	562.9594	1.909251e-124	4.052512e-120
ENSG00000120129	-2.952536	7.255438	506.3838	3.881506e-112	6.179067e-108
ENSG00000196136	-3.225084	6.911908	463.2175	9.587512e-103	1.221008e-98
ENSG00000101347	-3.759902	9.290645	449.9697	7.323427e-100	7.772231e-96
ENSG00000211445	-3.755609	9.102440	433.4656	2.861624e-96	2.603138e-92
ENSG00000162692	3.616656	4.551120	402.0266	1.994189e-89	1.587300e-85
ENSG00000171819	-5.705289	3.474697	389.3431	1.150502e-86	8.140055e-83
ENSG00000152583	-4.364255	5.491013	376.1995	8.363745e-84	5.325782e-80

Differential expression analysis - input

	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516	SRR1039517	SRR1039520	SRR1039521
ENSG00000000003	693	451	887	416	1148	1069	774	581
ENSG00000000005	0	0	0	0	0	0	0	0
ENSG00000000419	466	515	623	364	590	794	419	510
ENSG00000000457	326	274	372	223	356	450	308	297
ENSG00000000460	91	75	61	48	110	95	100	82
ENSG00000000938	0	0	2	0	1	0	0	0

- **Most** RNA-seq methods (e.g., edgeR, DESeq2, voom) need **raw counts** (or equivalent) as input
- **Don't** provide these methods with (e.g.) RPKMs, FPKMs, TPMs, CPMs, log-transformed counts, normalized counts, …
- Read documentation carefully!

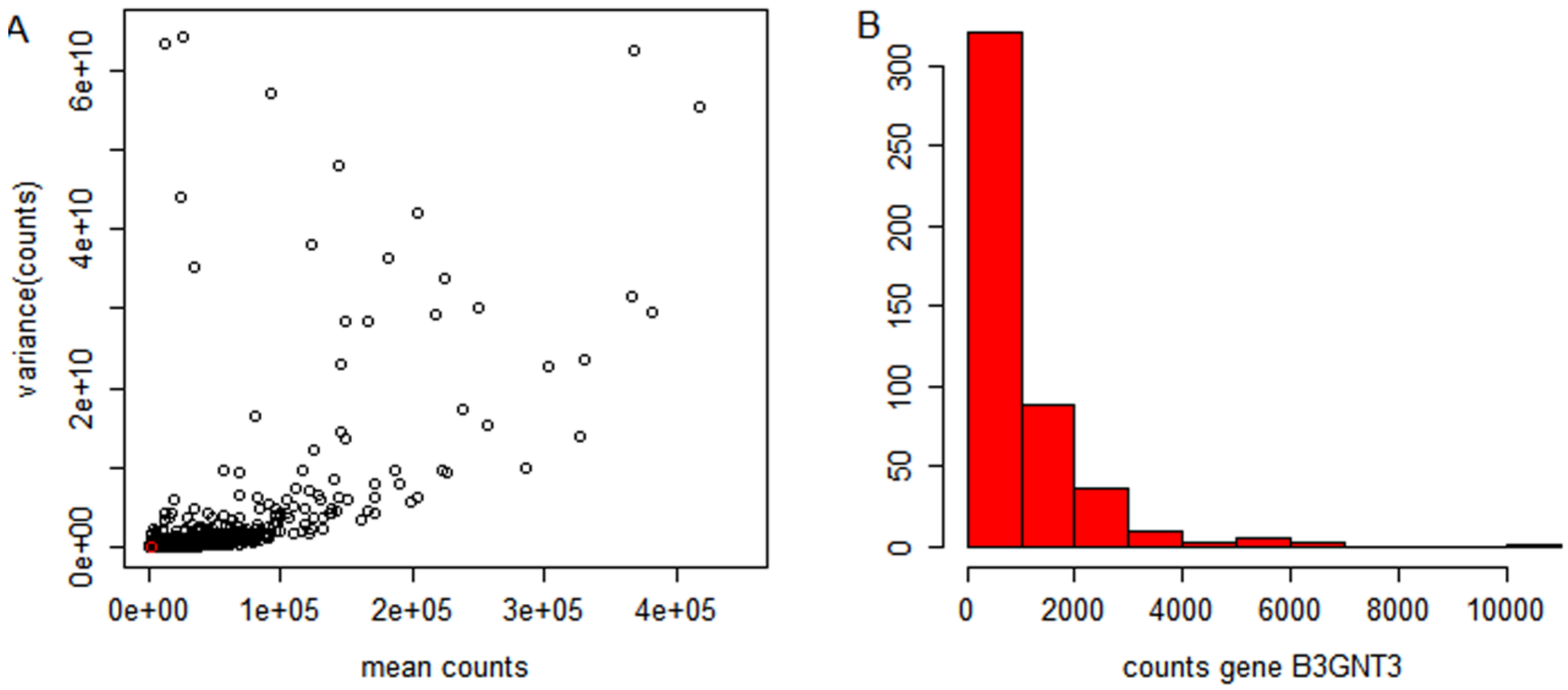
Challenges for RNA-seq data

- Choice of statistical distribution
- Normalization between samples
- Few samples -> difficult to estimate parameters (e.g., variance)
- High dimensionality (many genes) -> many tests

Challenges for RNA-seq data

- **Choice of statistical distribution**
- Normalization between samples
- Few samples -> difficult to estimate parameters (e.g., variance)
- High dimensionality (many genes) -> many tests

Characteristics of RNA-seq data

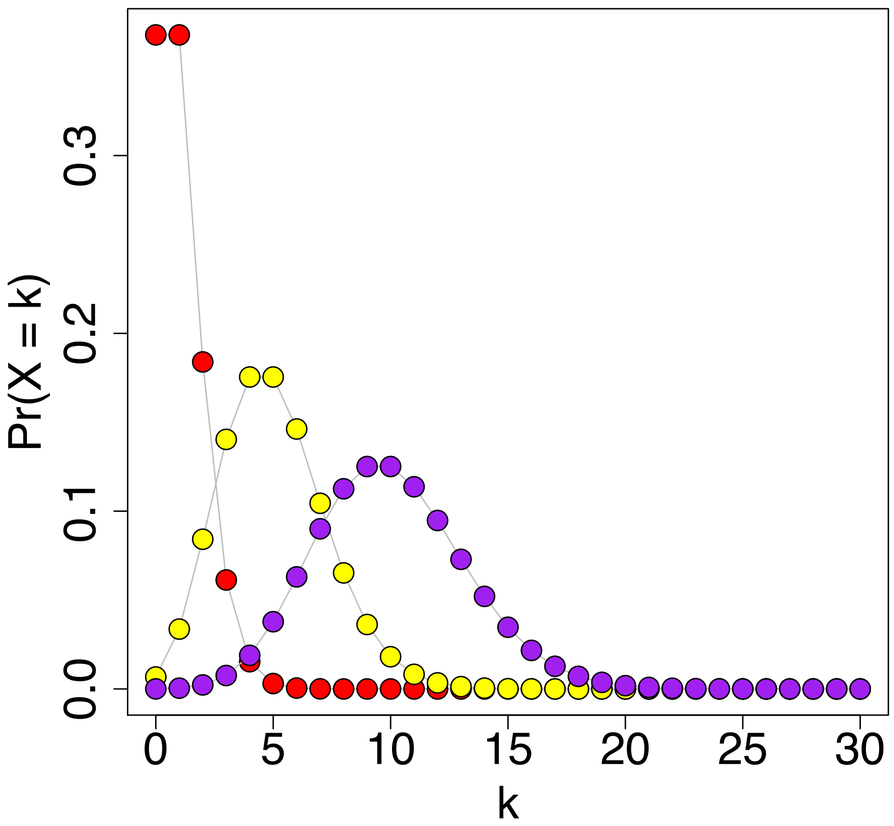


- Variance depends on the mean count
- Counts are non-negative and often highly skewed

Zwiener et al., PLoS One 2014

Modeling counts - the Poisson distribution

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$



Modeling counts - the Poisson distribution

- A famous use of the Poisson distribution was given by von Bortkiewicz (1898) in *Das Gesetz der kleiner Zahlen*

- He studied the number of soldiers in the Prussian army who got kicked by horses, over a number of years and corps

# horsekicks (k)	# obs	fraction	
0	109	0,545	
1	65	0,325	
2	22	0,11	
3	3	0,015	
4	1	0,005	

Modeling counts - the Poisson distribution

- A famous use of the Poisson distribution was given by von Bortkiewicz (1898) in *Das Gesetz der kleiner Zahlen*

- He studied the number of soldiers in the Prussian army who got kicked by horses, over a number of years and corps

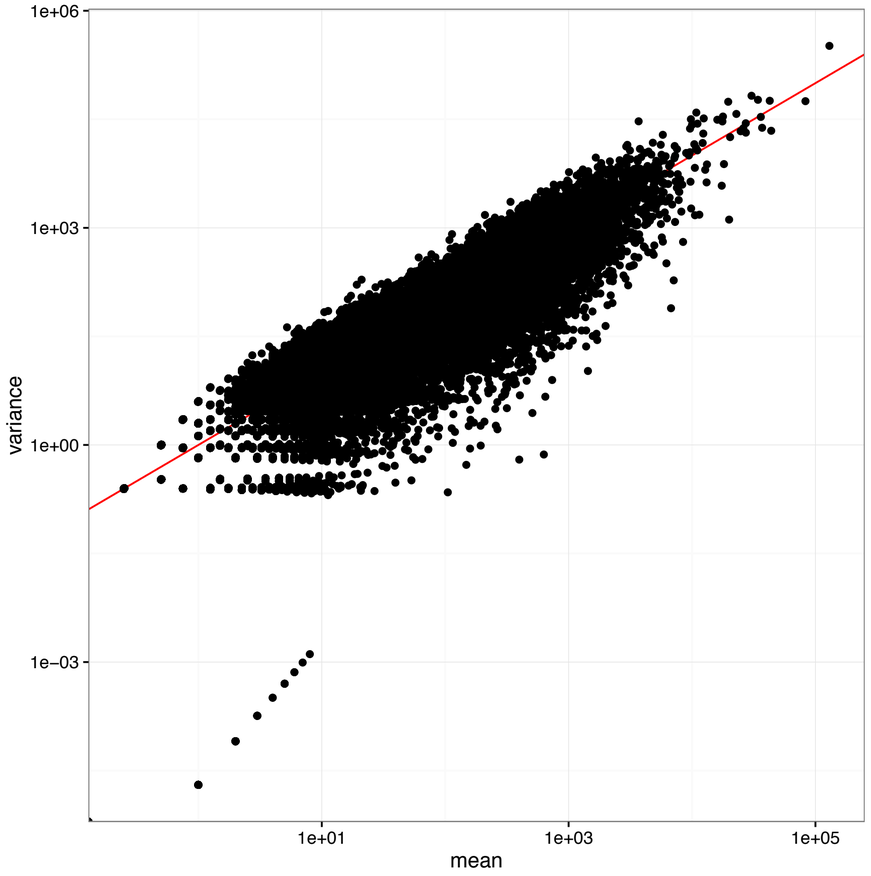
# horsekicks (k)	# obs	fraction	$\frac{0.61^k}{k!}e^{-0.61}$
0	109	0,545	0,543
1	65	0,325	0,331
2	22	0,11	0,101
3	3	0,015	0,0206
4	1	0,005	0,00313

The Poisson distribution for RNA-seq counts

- For RNA-seq data:
  - reads ~ soldiers
  - mapping to gene A ~ being kicked by a horse
- Assumes that the probability of a read mapping to gene A is the same for all samples within a class

Modeling counts

- **Poisson distribution**
  - Quantifies sampling variability
  - $var(X) = \mu$
  - Represents technical replicates well (mRNA proportions are identical across samples)

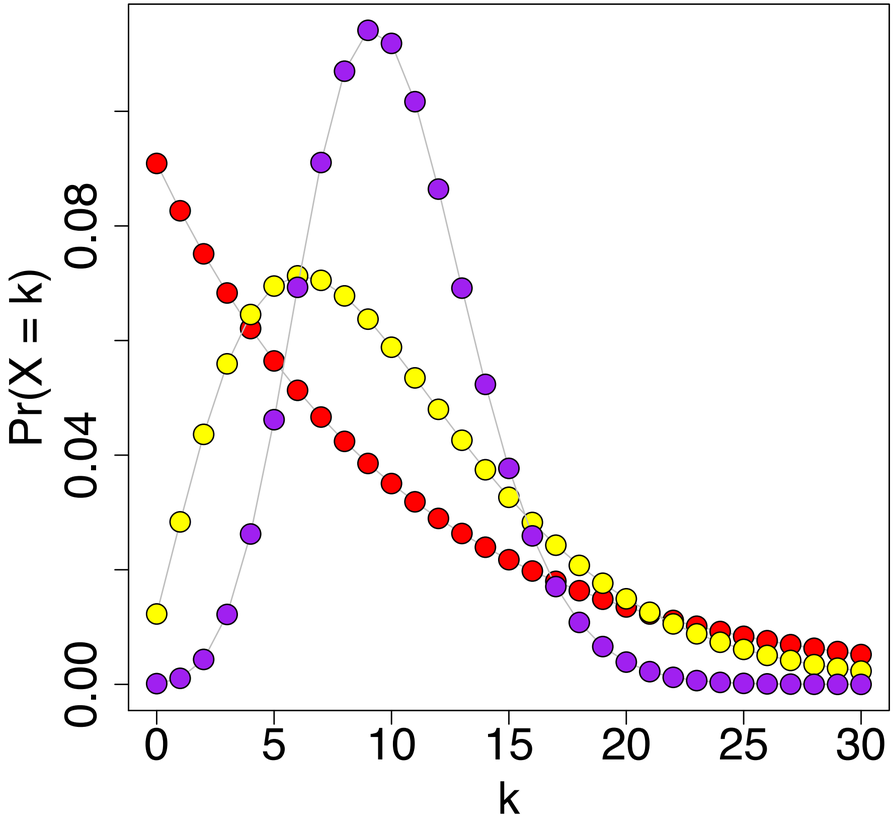


Example from SEQC data, same sample sequenced across multiple lanes

Modeling counts - the Negative Binomial distribution

$$P(X = k) = \binom{k + r - 1}{k} \cdot (1 - p)^r p^k$$

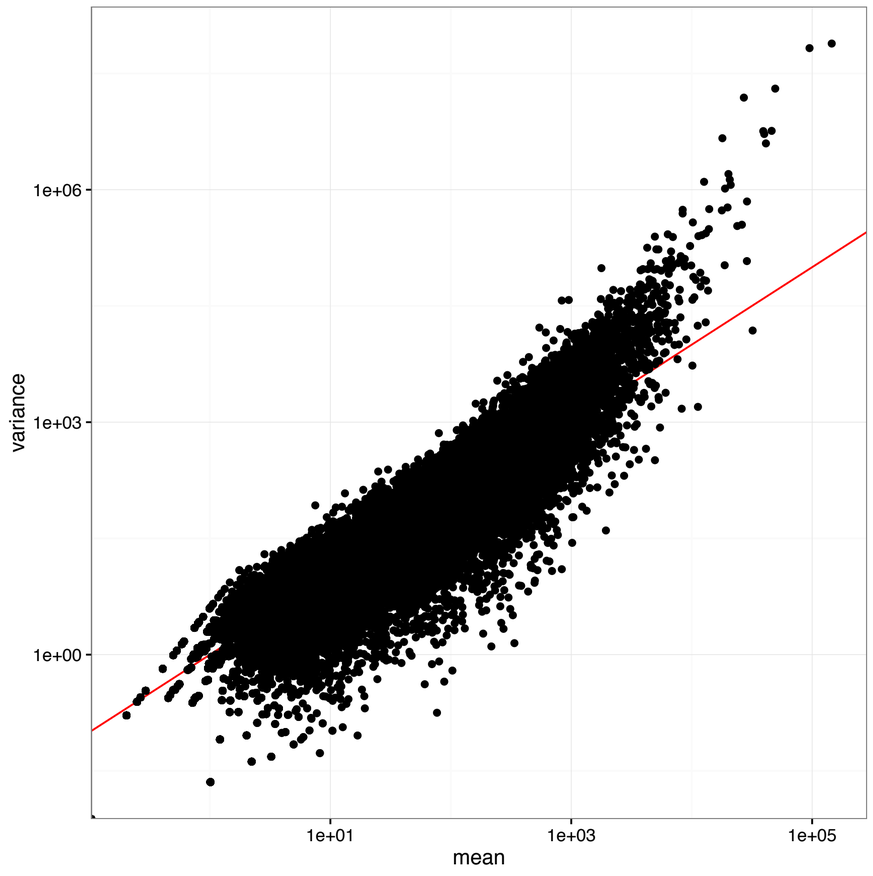
Generalizes the
Poisson distribution



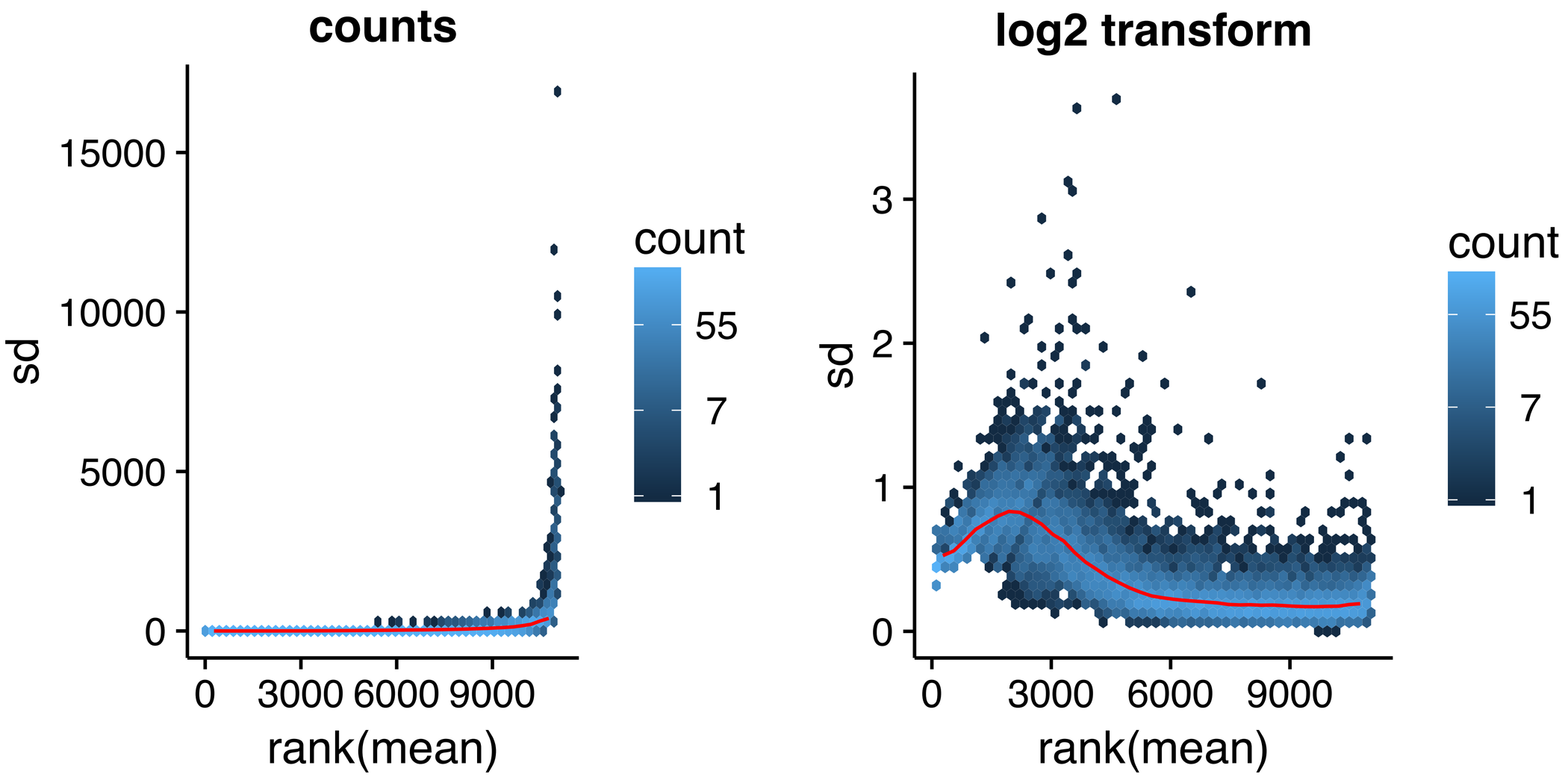
Modeling counts

- **Negative binomial distribution**
  - $var(X) = \mu + \theta\mu^2$
  - θ = dispersion
  - $\sqrt{\theta}$ = "biological coefficent of variation”
  - Allows mRNA proportions to vary across samples
  - Captures variability across biological replicates better

Example from SEQC data, replicates of the same RNA mix

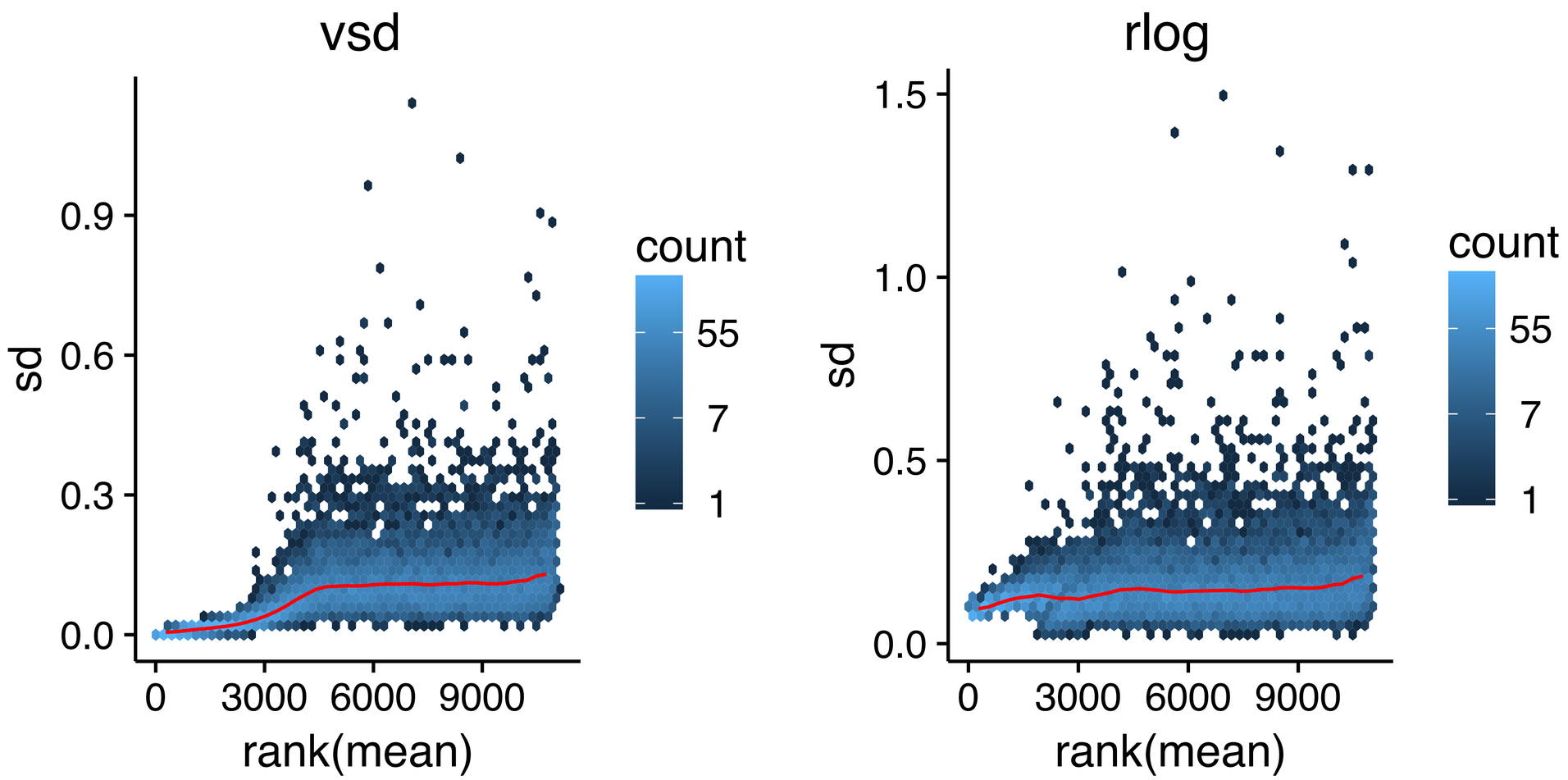


Data transformations - log



Data transformations - DESeq2

- Two approaches: rlog, variance stabilizing transformation
- Aim: remove dependence of variance on mean after transformation



Challenges for RNA-seq data

- Choice of statistical distribution
- **Normalization between samples**
- Few samples -> difficult to estimate parameters (e.g., variance)
- High dimensionality (many genes) -> many tests

Normalization

- Observed counts depend on:
  - abundance
  - gene length
  - sequencing depth
  - sequencing biases
  - …
- “As-is”, not directly comparable across samples

Normalization

raw count for gene *i* in sample *j*

normalization factor

relative abundance

$$C_{ij} \sim NB(\mu_{ij} = s_{ij} q_{ij}, \theta_i)$$

dispersion

- s_{ij} is a normalization factor (or *offset*) in the model
- counts are not explicitly scaled
  - important exception: voom/limma (followed by explicit modeling of mean-variance association)

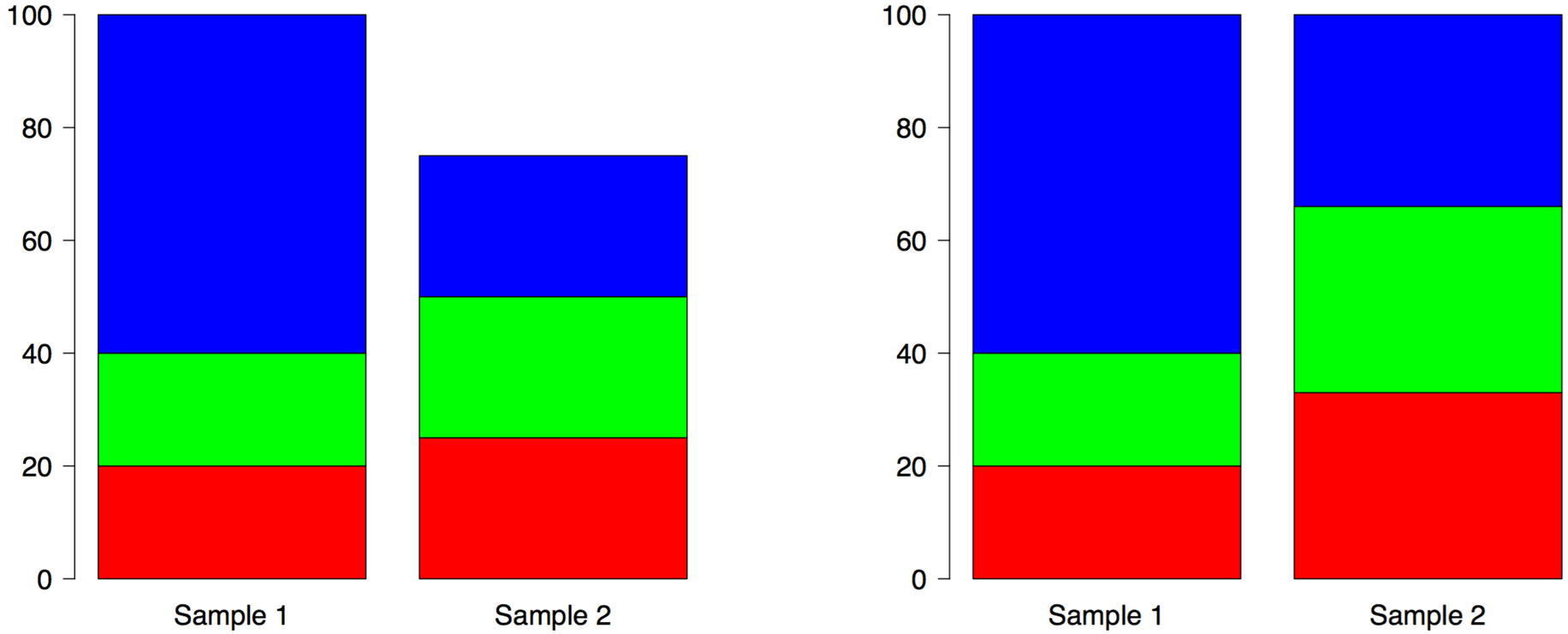
How to calculate normalization factors?

- Attempt 1: **total count** (library size)
  - Define a reference sample (one of the observed samples or a “pseudo-sample”) - gives a “target library size”
  - Normalization factor for sample *j* is defined by

$$\frac{\text{total count in sample } j}{\text{total count in reference sample}}$$

The influence of RNA composition

- Observed counts are relative
- High counts for some genes are “compensated” by low counts for other genes

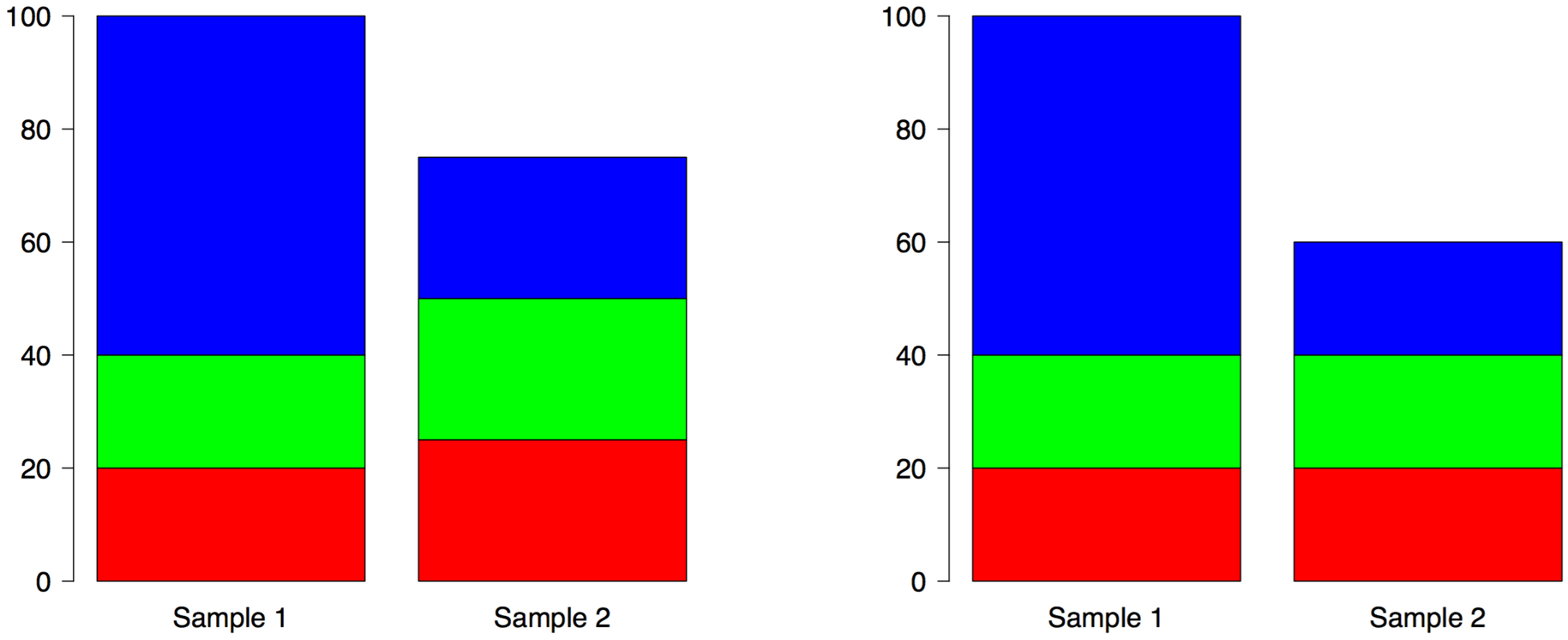


How to calculate normalization factors?

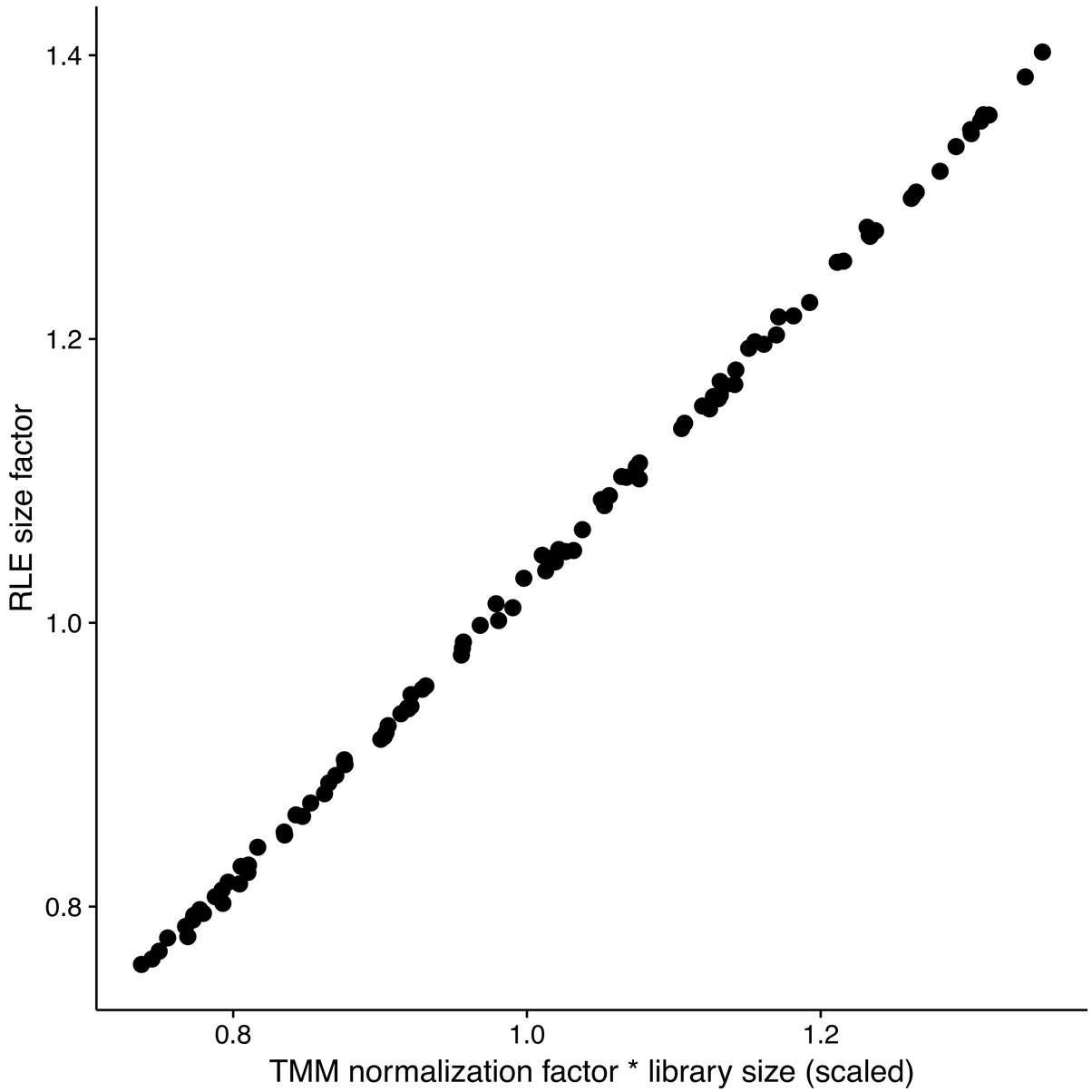
- Attempt 2: total count (library size) * compensation for differences in composition
- Idea: use only non-differentially expressed genes to compute the normalization factor
- Implemented by both edgeR (TMM) and DESeq2 (median count ratio)
- Both these methods assume that most genes are not differentially expressed

How to calculate normalization factors?

- Attempt 2: total count (library size) * compensation for differences in composition



“Effective library sizes” (edgeR) vs “size factors” (DESeq2)



References

- van den Berge, Hembach, Soneson, Tiberi et al: RNA sequencing data: hitchhiker's guide to expression analysis. PeerJ Preprints 6:e27283v2 (2018) - **review of RNA-seq**
- Orjuela, Huang, Hembach et al: ARMOR: an Automated Reproducible MOdular workflow for preprocessing and differential analysis of RNA-seq data. bioRxiv doi:10.1101/575951 (2019) - **RNA-seq workflow**
- Rue-Albrecht, Marini, Soneson & Lun: iSEE: Interactive SummarizedExperiment Explorer. F1000Research 7:741 (2018) - **iSEE**
- Love, Soneson & Patro: Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. F1000Research 7:952 (2018) - **DTU workflow**
- Robinson et al.: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139-140 (2010) - **edgeR**
- Love et al.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15:550 (2014) - **DESeq2**
- Law et al.: voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology 15:R29 (2014) - **voom**
- Patro et al.: Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods 14:4170419 (2017) - **Salmon**
- Bray et al.: Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology 34(5):525-527 (2016) - **kallisto**
- Patro et al.: Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nature Biotechnology 32:462-464 (2014) - **Sailfish**
- Pimentel et al.: Differential analysis of RNA-Seq incorporating quantification uncertainty. bioRxiv http://dx.doi.org/10.1101/058164 (2016) - **sleuth**
- Wagner et al.: Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences 131:281-285 (2012) - **TPM vs FPKM**
- Soneson et al.: Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research 4:1521 (2016) - **ATL offsets (tximport package)**
- Li et al.: RNA-seq gene expression estimation with read mapping uncertainty. Bioinformatics 26(4):493-500 (2010) - **TPM, RSEM**
- Soneson, Matthes et al.: Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. Genome Biology 17:12 (2016)
- Schurch et al.: How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 22:839-851 (2016)
- Dillies et al.: A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics 14(6):671-683 (2013)
- Soneson & Delorenzi: A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14:91 (2013)
- Anders et al.: Detecting differential usage of exons from RNA-seq data. Genome Research 22(10):2008-2017 (2012) - **DEXSeq**
- Love et al.: RNA-Seq workflow: gene-level exploratory analysis and differential expression. F1000 Research 4:1070 (2016) - **RNA-seq workflow**
- Law et al.: RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000 Research 5:1408 (2016) - **RNA-seq workflow**
- Chen et al: From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000 Research 5:1438 (2016) - **RNA-seq workflow**