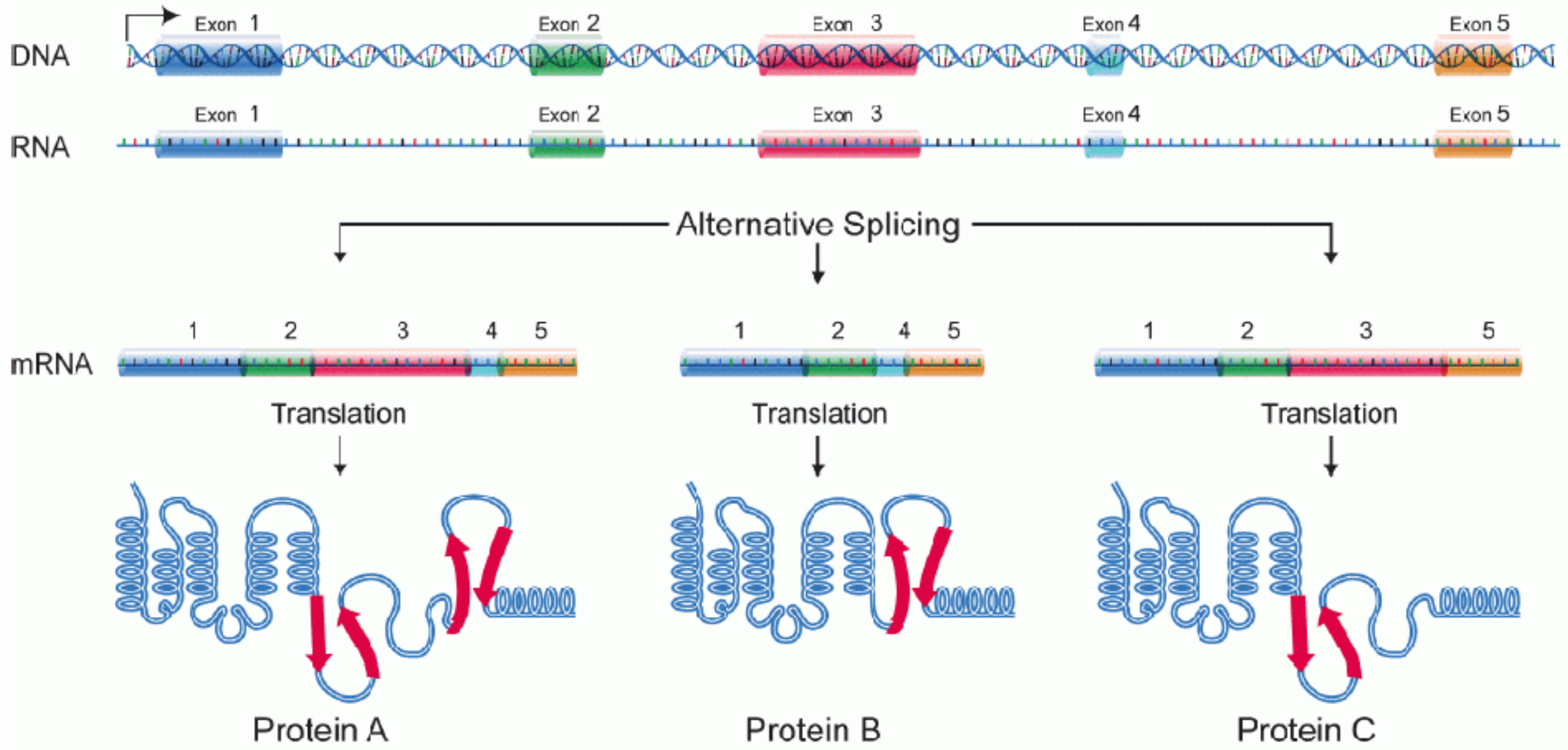
Differential gene expression analysis

Charlotte Soneson

CSAMA, Brixen, July 10 2018


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**

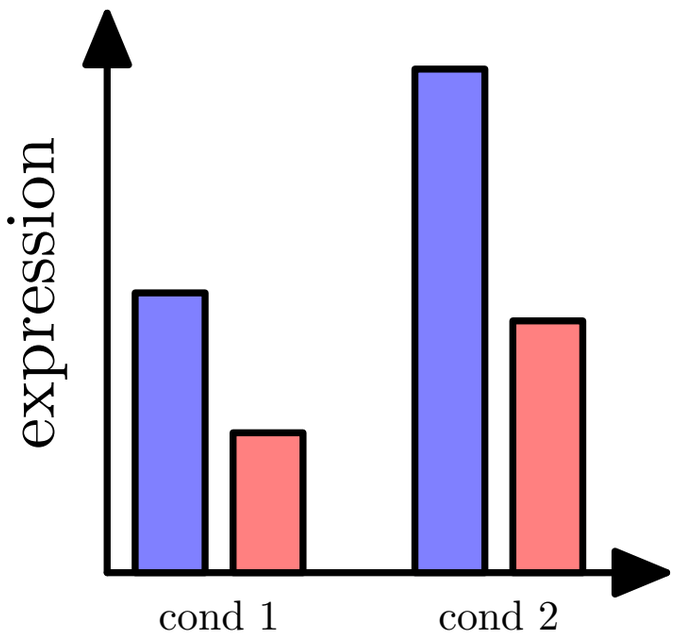
- Does *any* isoform of a given gene change? **DTE+G**

- Does the isoform composition for a given gene change? **DTU/DIU/DEU**

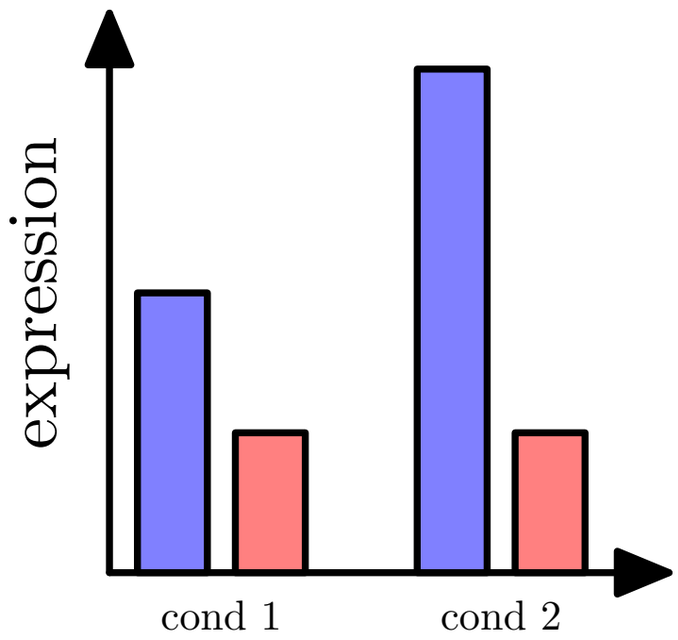
- need **different** abundance quantification of transcriptomic features (genes, transcripts, exons)

Differential behaviour types are not mutually exclusive

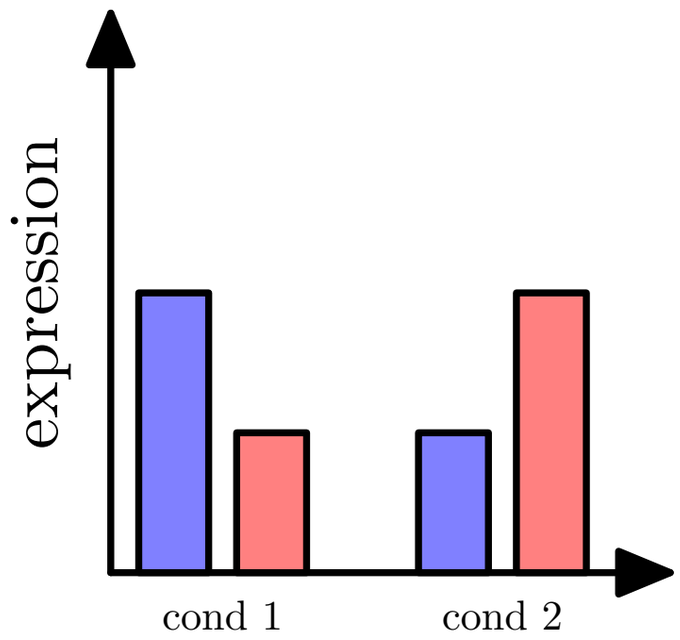
DGE
(also **DTE**)



DTE
(also **DGE, DTU**)



DTU
(also **DTE**)



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!

Differential expression analysis with DESeq2/edgeR

```
> library(DESeq2)
> dds <- DESeqDataSetFromMatrix(cnts, DataFrame(cond), ~ cond)
converting counts to integer mode
> dds <- DESeq(dds)
estimating size factors
estimating dispersions
gene-wise dispersion estimates
mean-dispersion relationship
final dispersion estimates
fitting model and testing
> res <- results(dds)
```

```
> library(edgeR)
> dge <- DGEList(cnts, samples = data.frame(cond))
> dge <- calcNormFactors(dge)
> design <- model.matrix(~cond, data = dge$samples)
> dge <- estimateDisp(dge, design = design)
> fit <- glmQLFit(dge, design = design)
> lrt <- glmQLFTest(fit)
> res <- topTags(lrt)
```

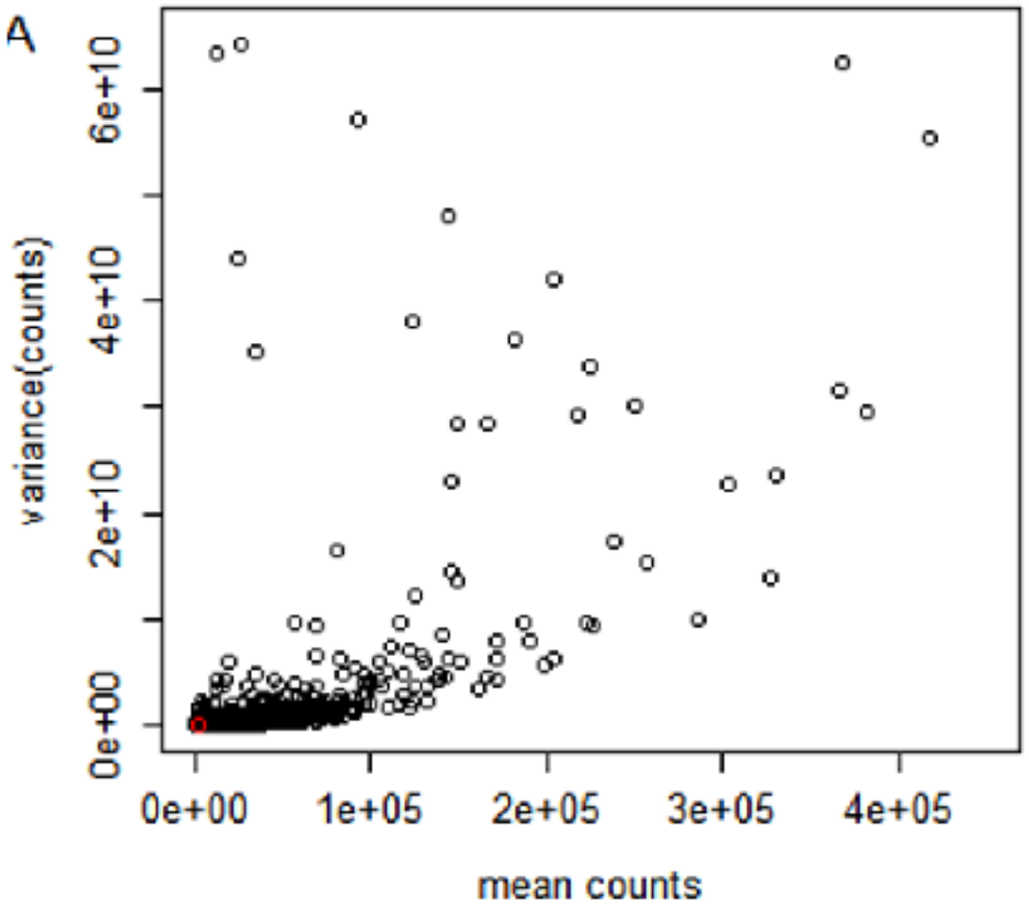
Challenges for RNA-seq data

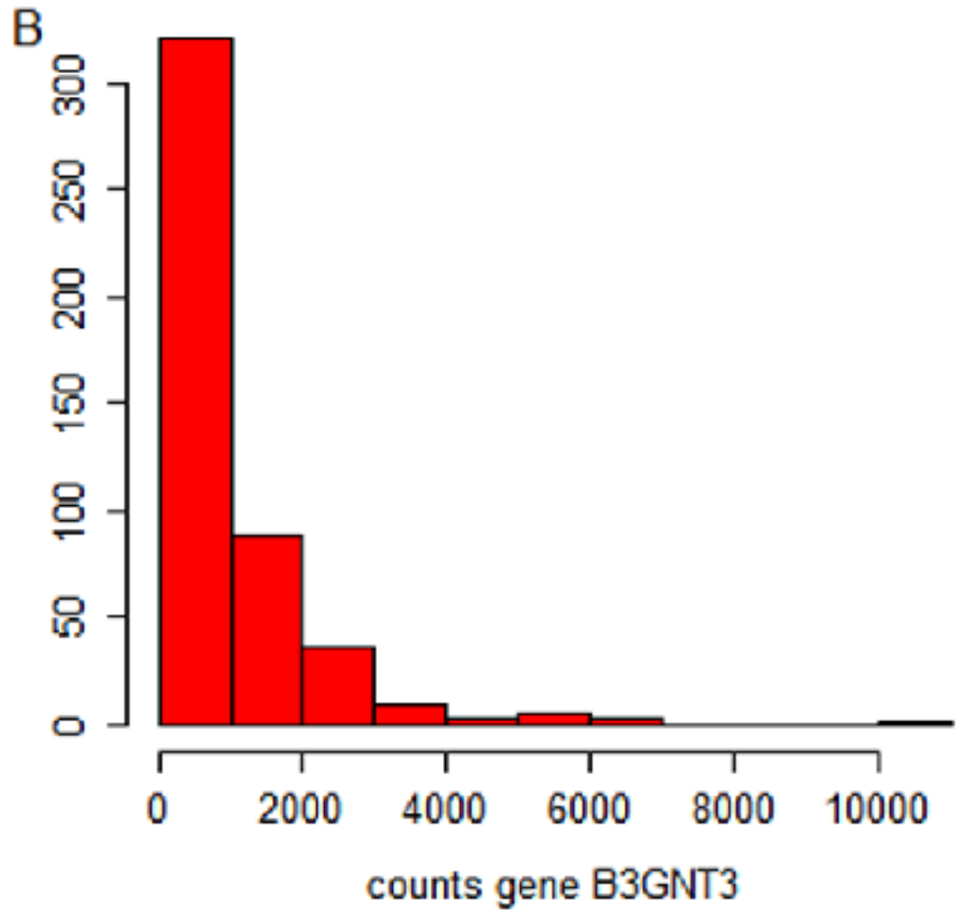
- Choice of statistical distribution
- Normalization between samples
- Few samples -> difficult to estimate parameters (e.g., variance)
- High dimensionality (many genes) -> many tests

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



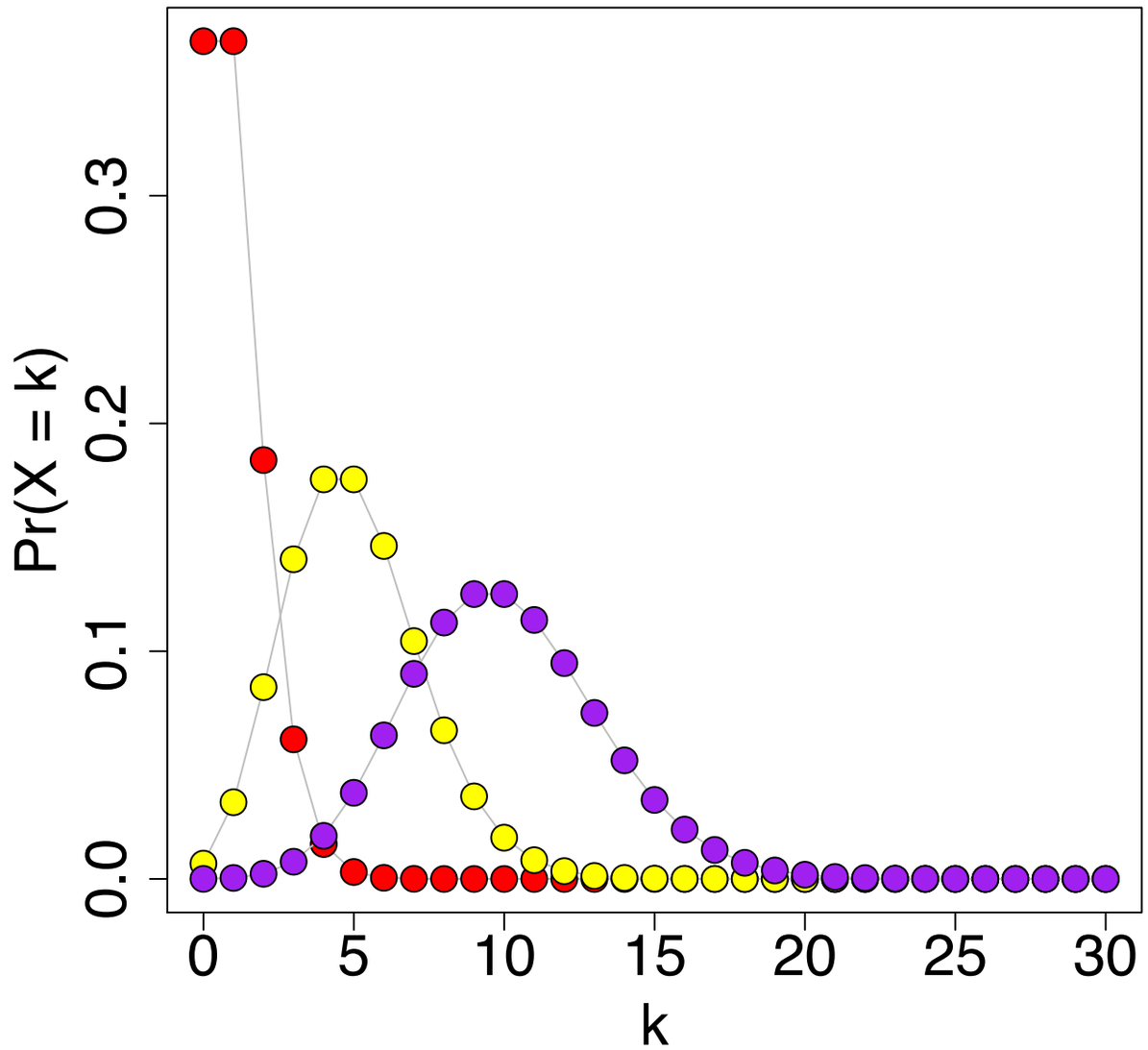


- Variance depends on the mean count
- Counts are non-negative and often highly skewed

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

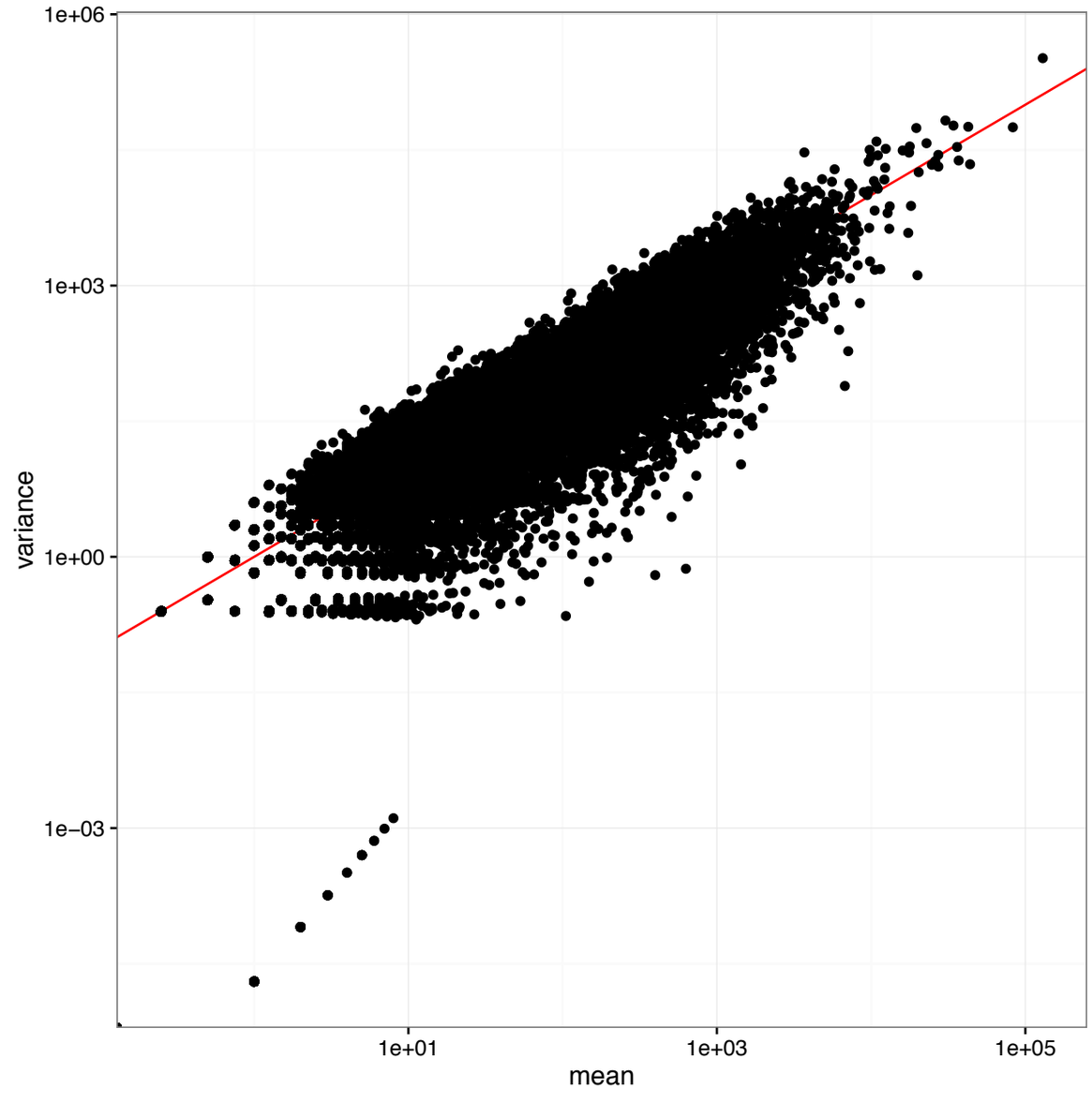
# horsekicks (k)	# obs	fraction	$\frac{0.61^k}{k!}e^{-0.61}$
0	109	0,545	0,543
1	65	0,325	0,331
2	22	0,11	0,101
3	3	0,015	0,0206
4	1	0,005	0,00313

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)

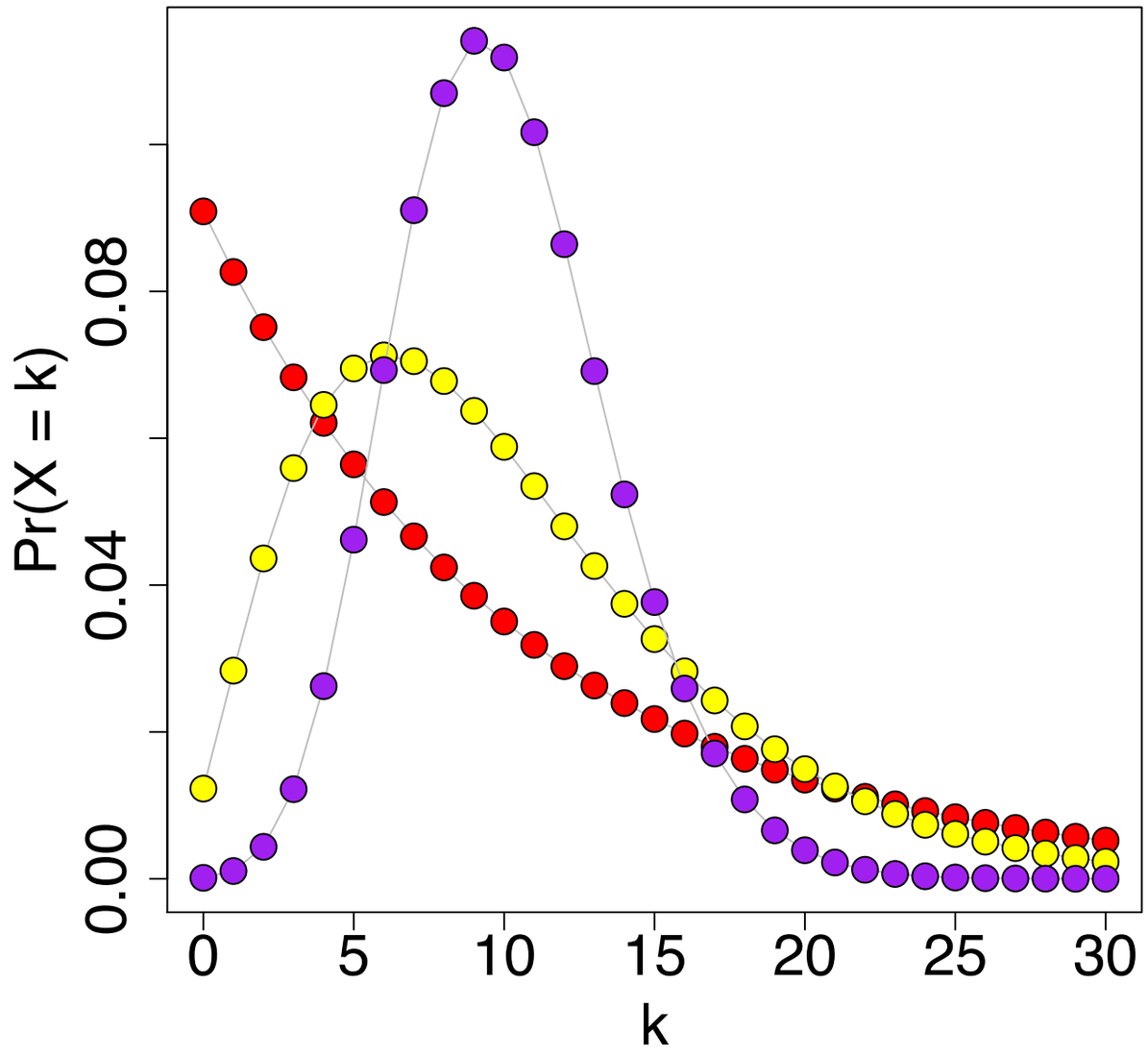


Example from SEQC data, same sample sequenced across multiple lanes

Modeling counts - the Negative Binomial distribution

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

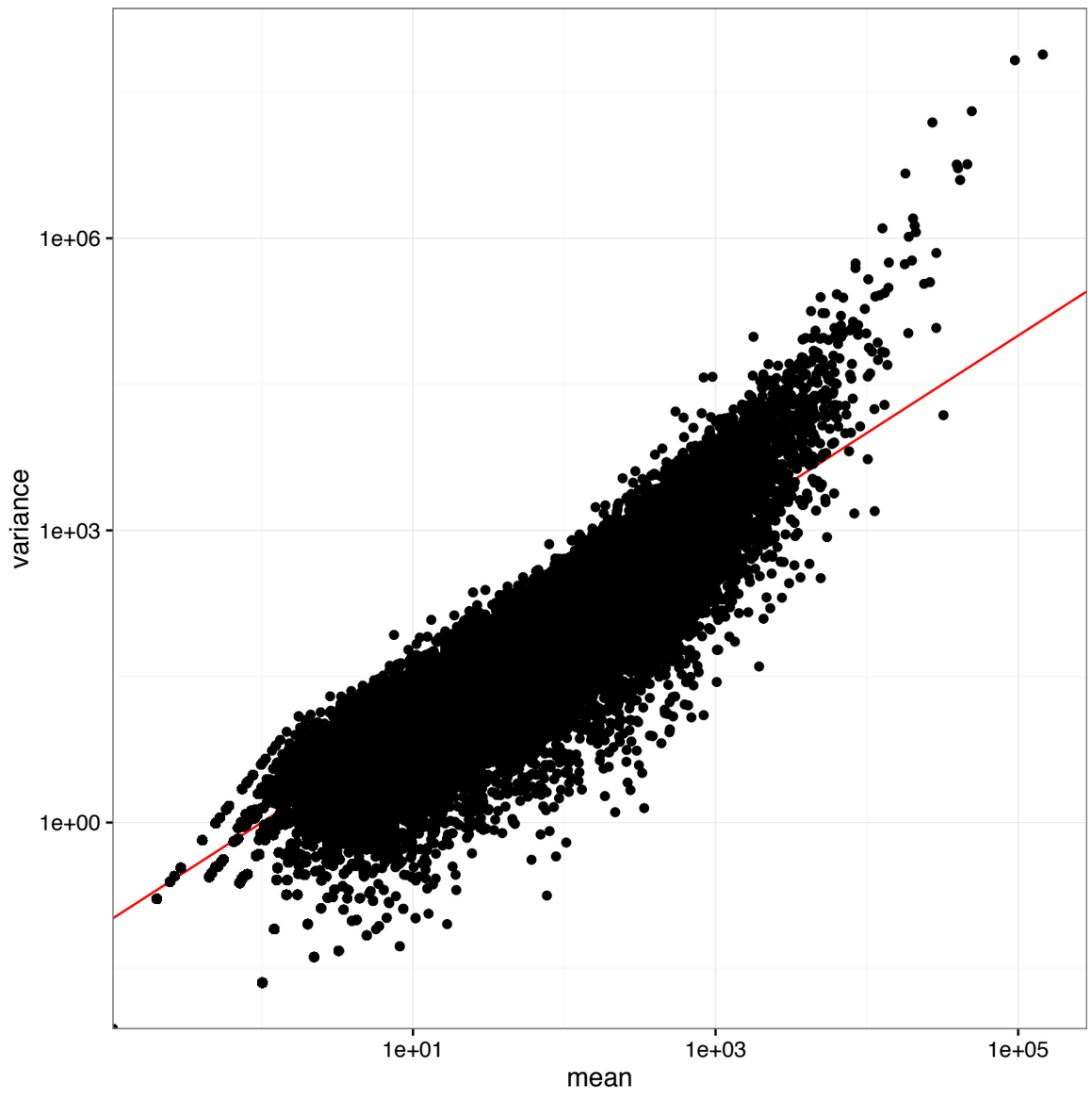
Generalizes the
Poisson distribution



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
  - Used by **DESeq2**, **edgeR** and other packages for RNA-seq analysis

Example from SEQC data, replicates of the same RNA mix



With count data…

- *linear* modeling (and thus t-tests, ANOVA, etc) is no longer suitable for inference

- *Generalized linear models* to the rescue!

Extension to single-cell data

- Single-cell RNA-seq data has lots of zeros, which can not always be explained by a Negative Binomial distribution (dropouts, “excess zeros”)

- The resulting distribution is a *zero-inflated* Negative Binomial

- One solution is to downweight the influence of zeros that are likely to come from the zero inflation when fitting the model

- Weights from the **zinbwave** package can be incorporated into DESeq2/edgeR models

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*

relative abundance

normalization factor

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

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)

Adapted from Mike Love

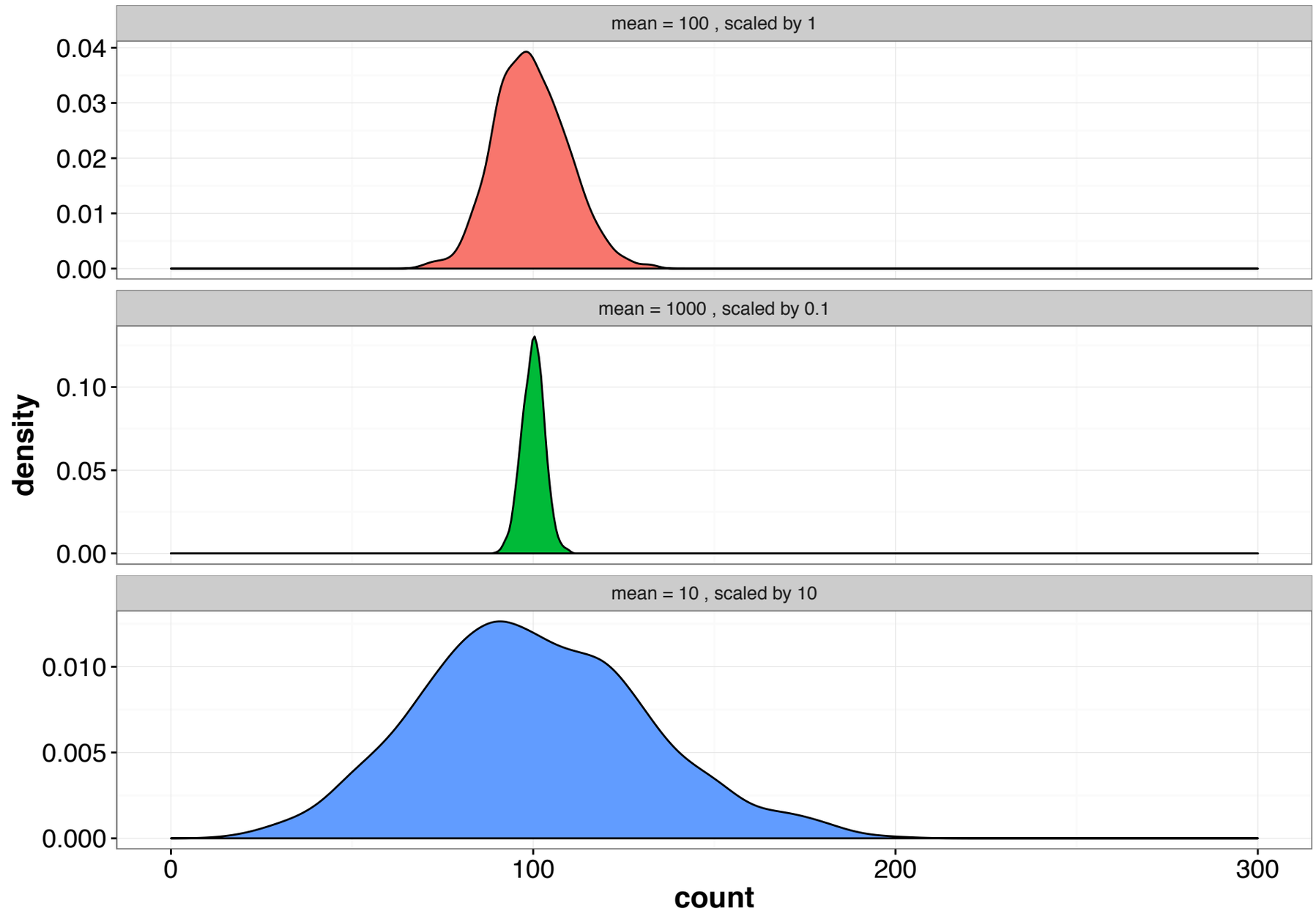
Why offset rather than scaling?

- Poisson distributed variables with different means, scaled to have mean = 100

Raw count
mean = 100

Raw count
mean = 1000,
scaled by 0.1

Raw count
mean = 10,
scaled by 10



Adapted from Mike Love

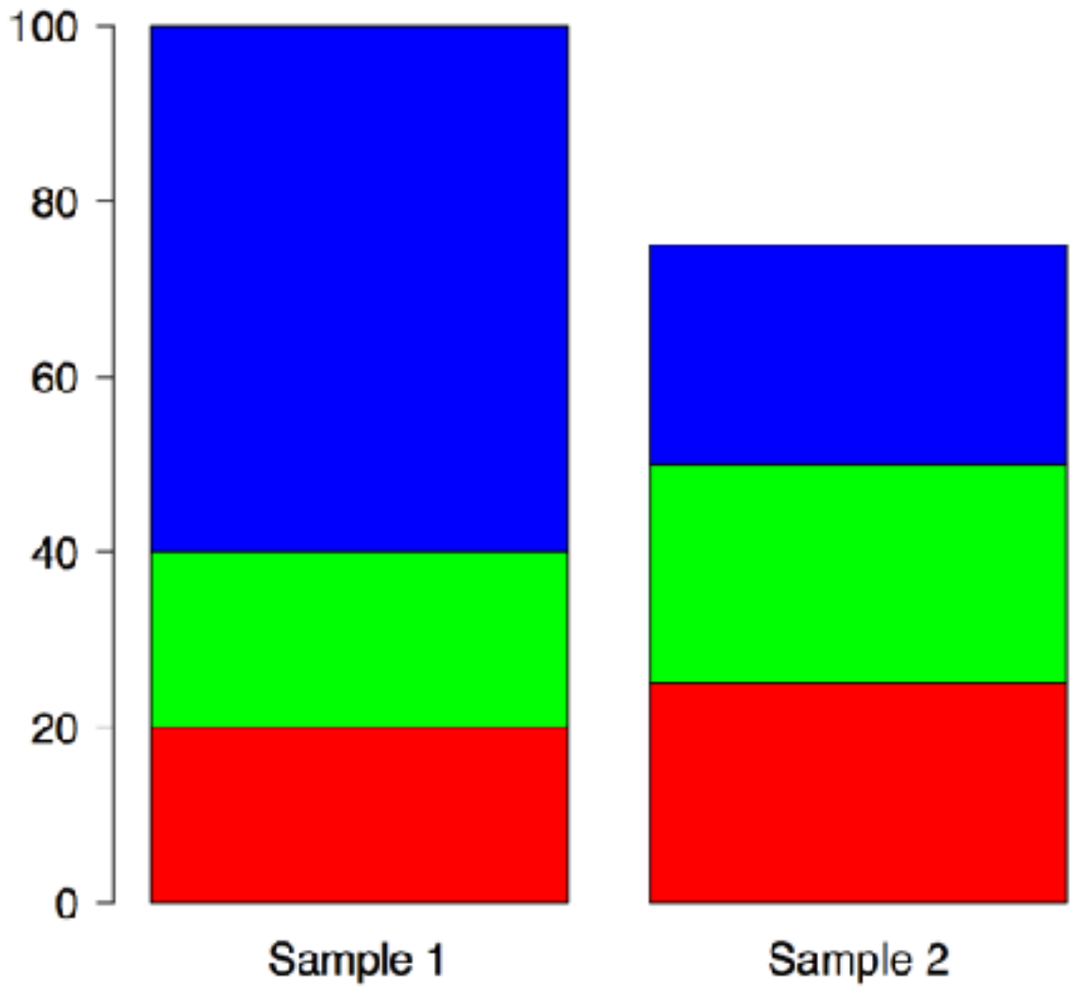
How to calculate normalization factors?

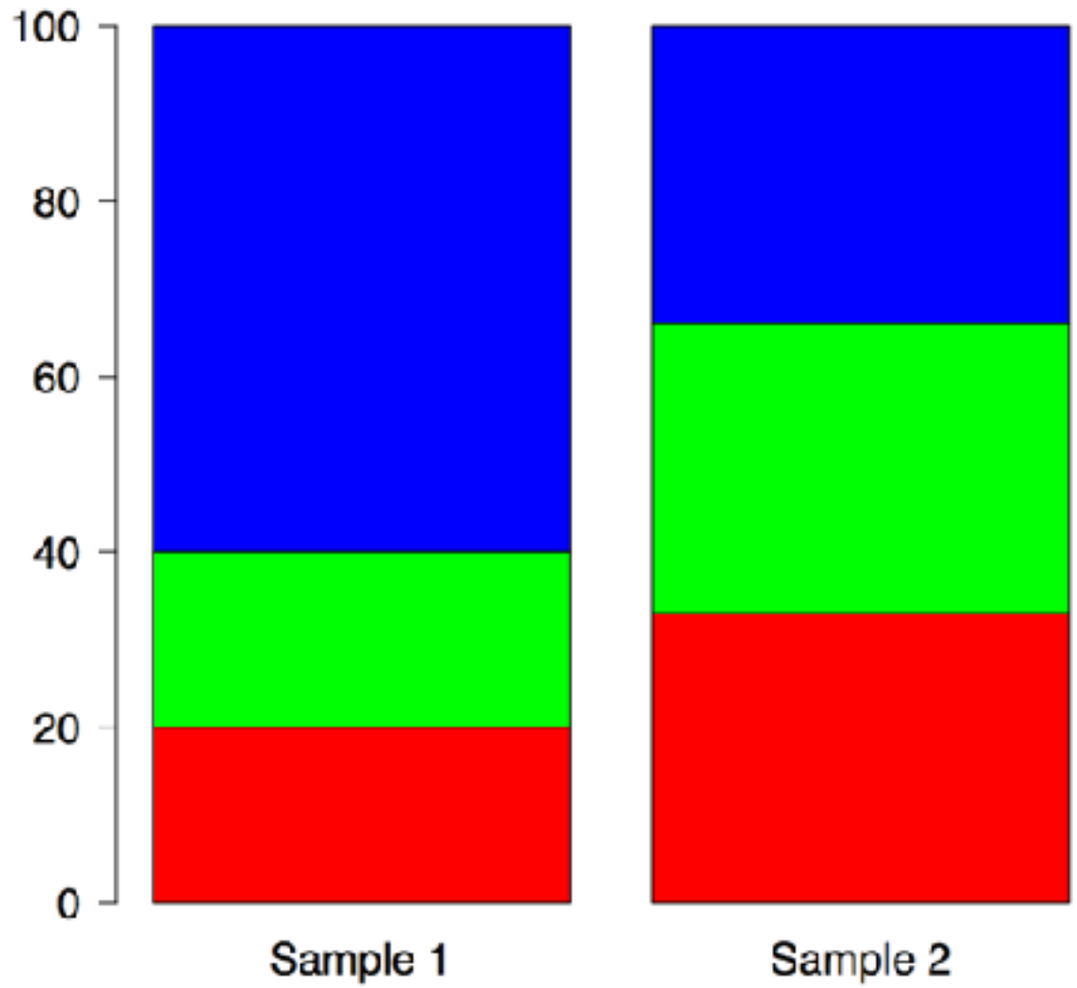
- Attempt 1: **total count** (library size)
  - Define a reference sample (one of the observed samples or a “pseudo-sample”) - gives a “target library size”
- Normalization factor for sample *j* is defined by

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

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

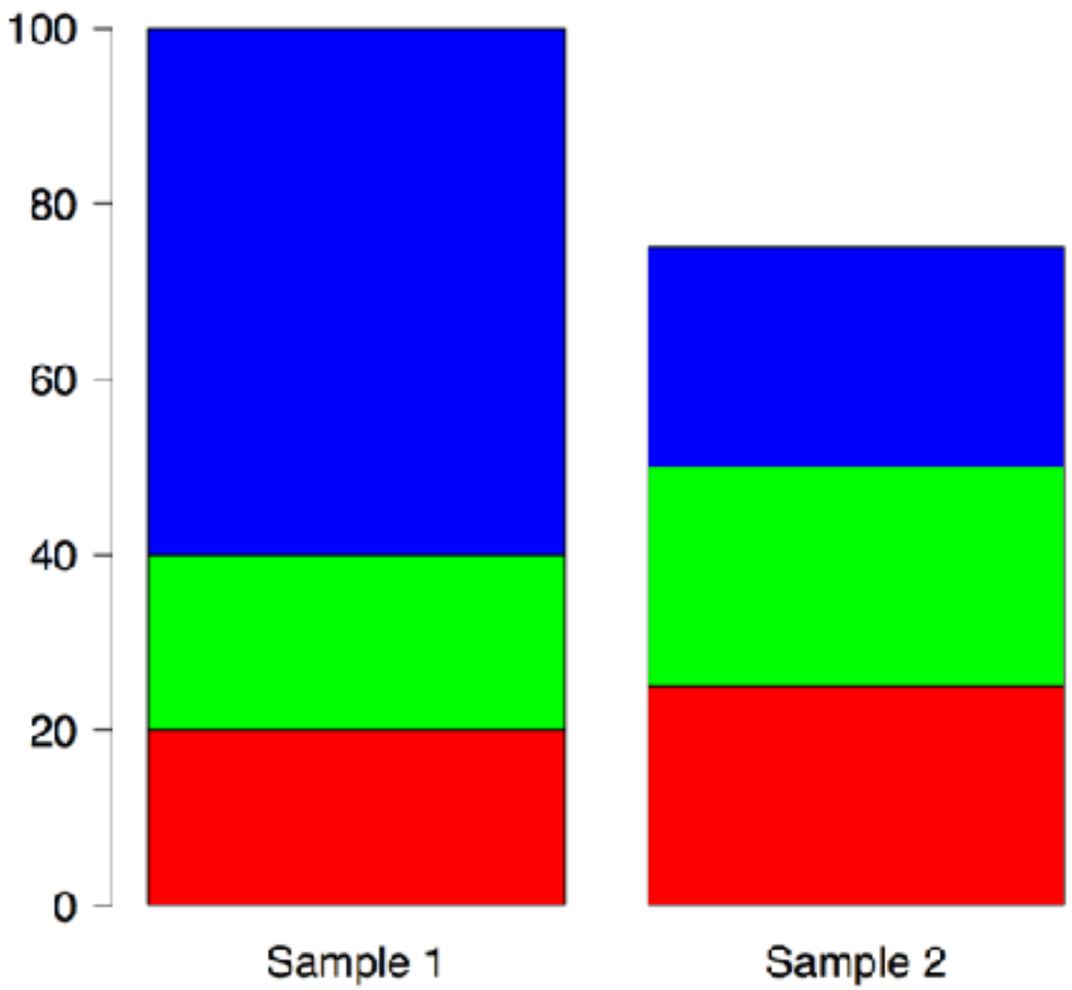
- Idea: use only non-differentially expressed genes to compute the normalization factor

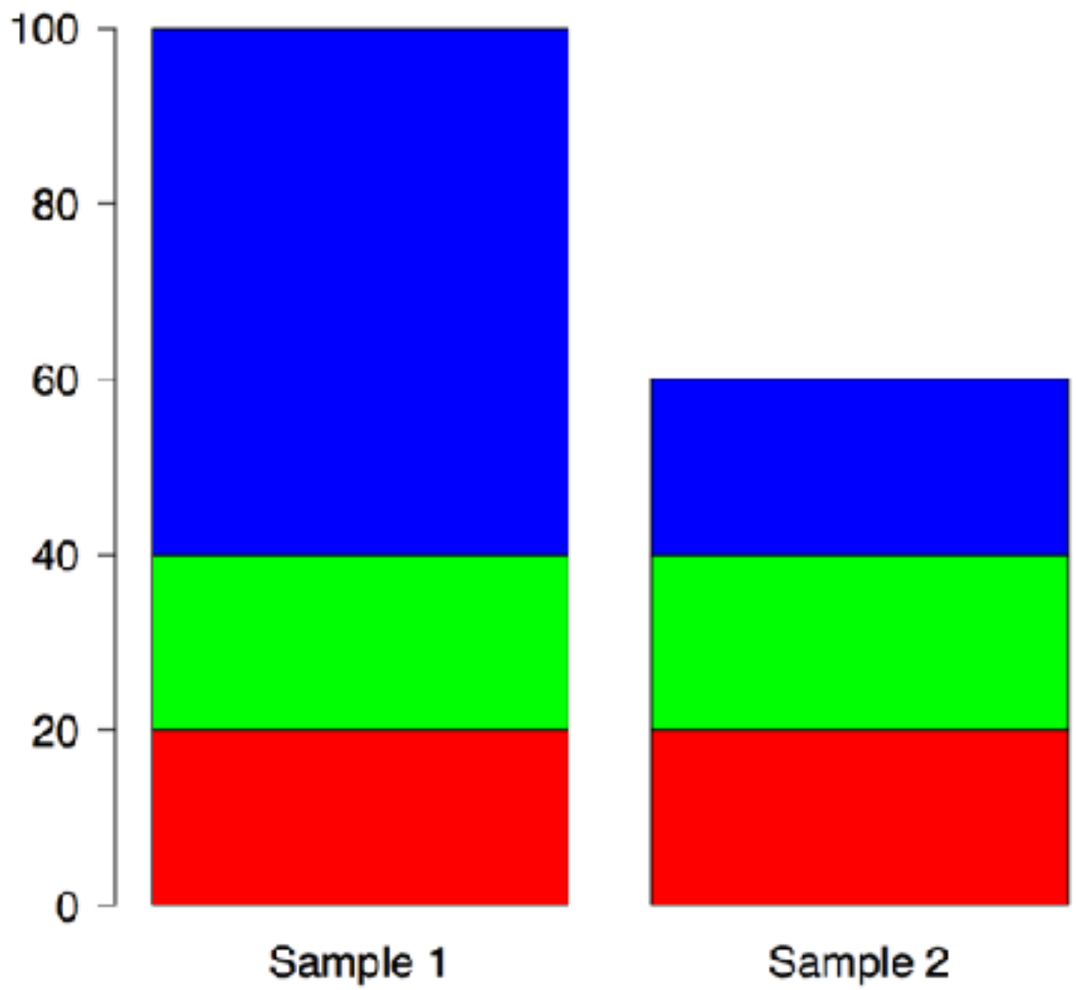
- Implemented by both edgeR (TMM) and DESeq2 (median count ratio)

- Both these methods assume that most genes are not differentially expressed

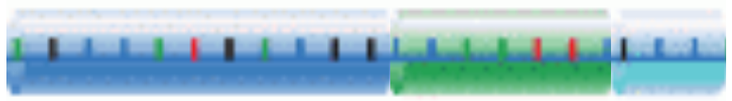
How to calculate normalization factors?

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





Impact of differential isoform usage on gene-level counts

T1 

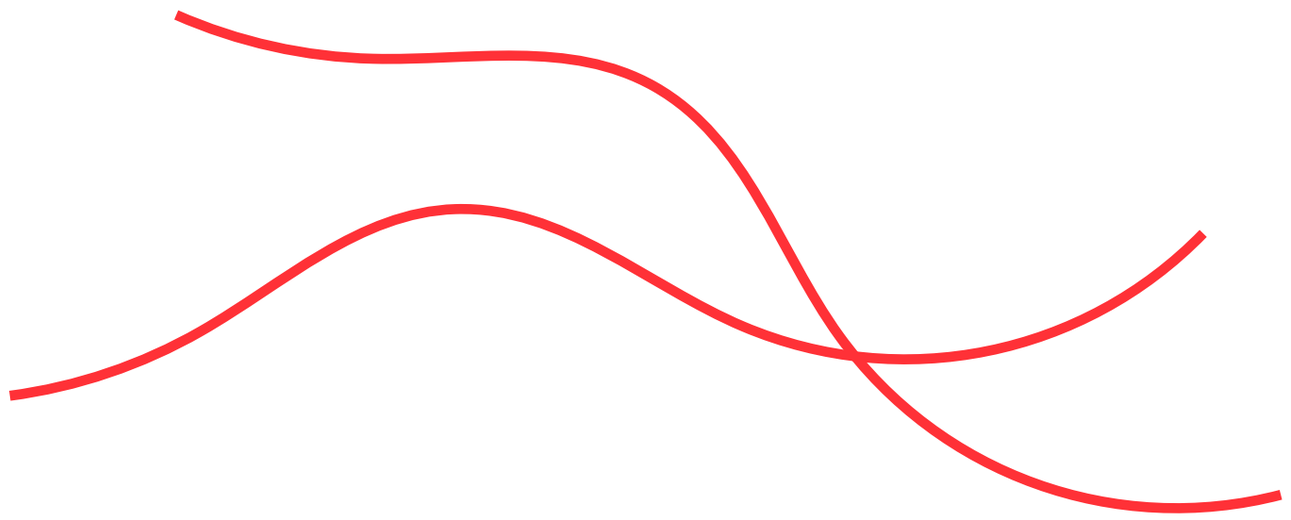
length = **L**

T2 

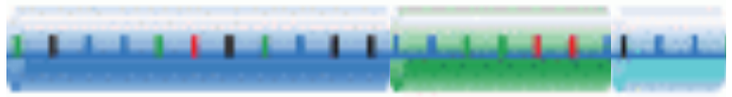
length = **2L**

sample 1

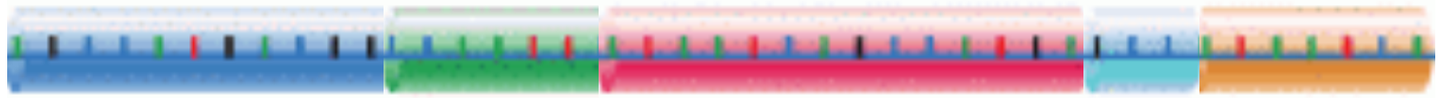
sample 2



Average transcript lengths

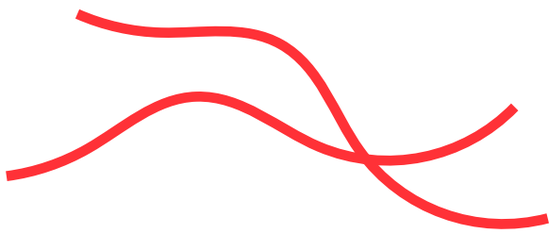
T1 

length = **L**

T2 

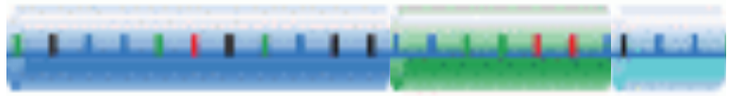
length = **2L**

$$ATL_{g1} = 1 \cdot L + 0 \cdot 2L = L$$

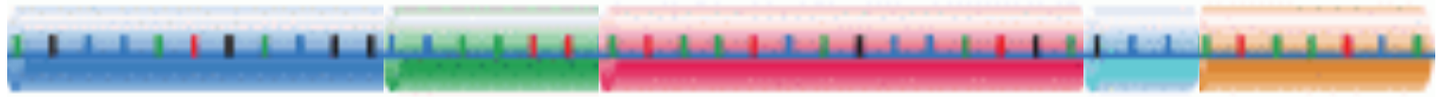


$$ATL_{g2} = 0 \cdot L + 1 \cdot 2L = 2L$$

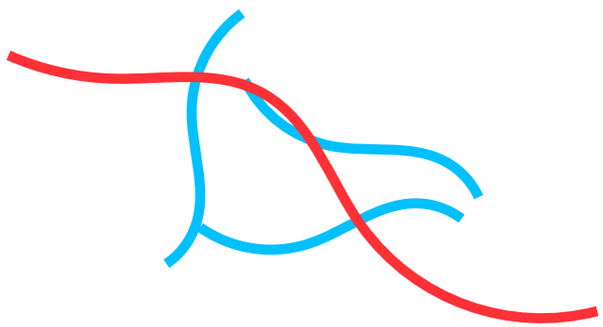
Average transcript lengths

T1 

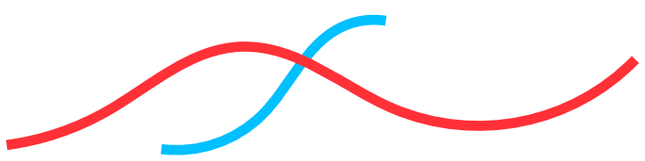
length = **L**

T2 

length = **2L**

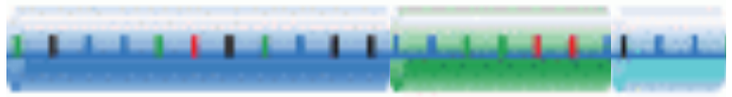


$$ATL_{g1} = 0.75 \cdot L + 0.25 \cdot 2L = 1.25L$$



$$ATL_{g2} = 0.5 \cdot L + 0.5 \cdot 2L = 1.5L$$

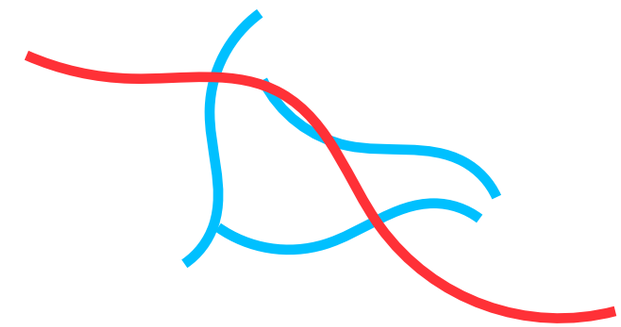
Average transcript lengths

T1 

length = **L**

T2 

length = **2L**



$$ATL_{g1} = 0.75 \cdot L + 0.25 \cdot 2L = 1.25L$$

$$ATL_{g2} = 0.5 \cdot L + 0.5 \cdot 2L = 1.5L$$

weights obtained from transcript TPM estimates

Offsets (“scaling factors”)

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

relative abundance

scaling factor

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

dispersion

- Extend scaling factor for given sample and gene to include the average length of the transcripts from the gene that are present in the sample

Adapted from Mike Love

Offsets ("average transcript lengths")

- Similar to correction factors for library size, but sample-**and** gene-specific
- Transcript abundance levels (TPMs) can be obtained from (e.g.) Salmon or kallisto
- Average transcript length for gene *g* in sample *s*:

$$ATL_{gs} = \sum_{i \in g} \theta_{is} \bar{\ell}_{is}, \qquad \sum_{i \in g} \theta_{is} = 1$$

$\bar{\ell}_{is}$ = effective length of isoform i (in sample s)
θ_{is} = relative abundance of isoform i in sample s

Getting ATL offsets with tximport

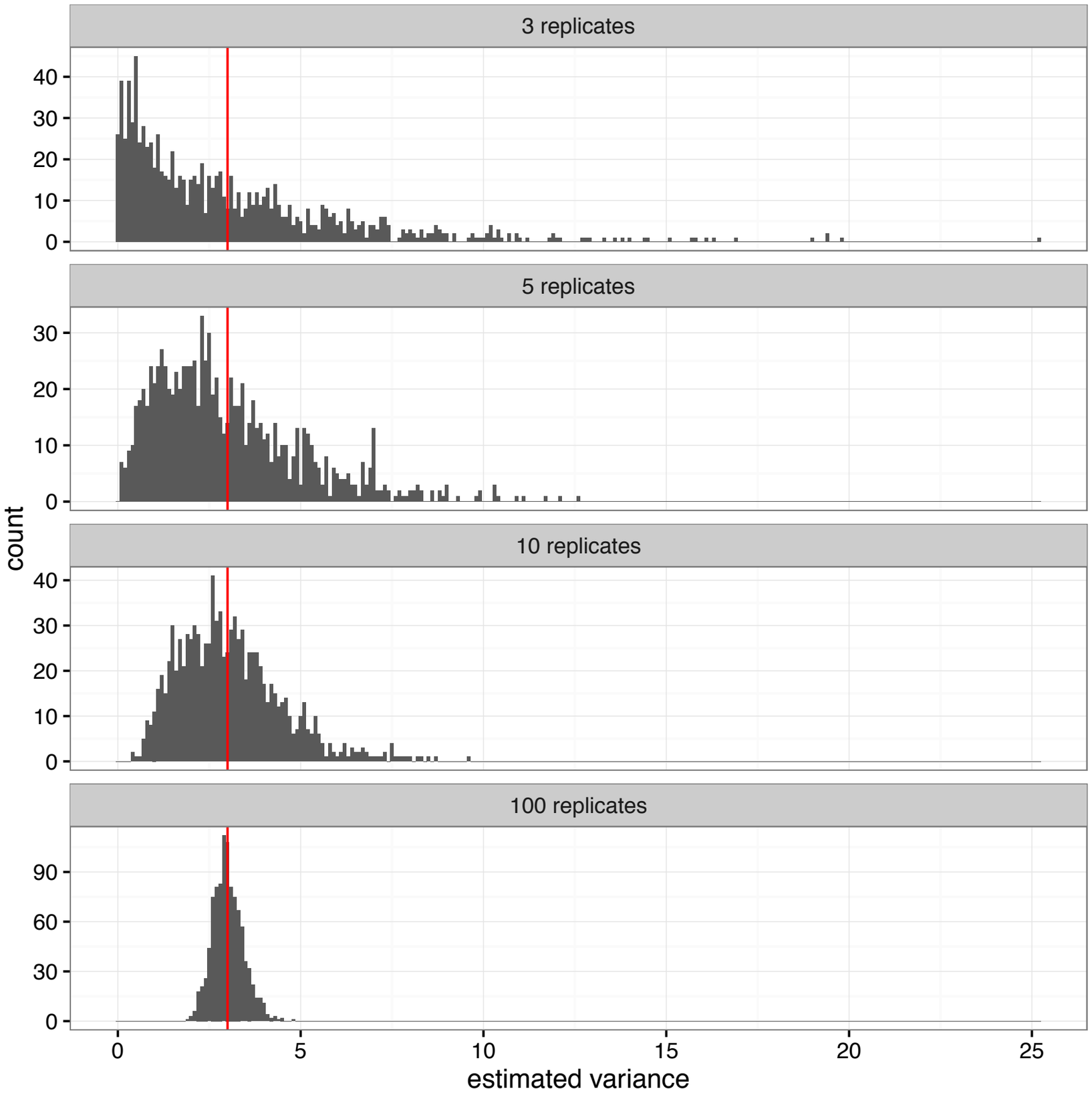
```
> txi <- tximport(files = salmon_files, type = "salmon", txOut = FALSE, tx2gene = tx
2gene)
reading in files with read_tsv
1 2 3 4 5 6 7 8
summarizing abundance
summarizing counts
summarizing length
> txi$length[16363:16367, ]
                 SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516
ENSG00000186469   2707.434   2727.511   2598.809  2966.5202   2287.617
ENSG00000186470   2152.745   2057.866   2048.808  2166.0751   2081.689
ENSG00000186471    384.702    384.702    384.702   384.7020    384.702
ENSG00000186472    592.864  16985.700   1509.290   583.0114   2726.088
ENSG00000186474    842.865    842.865    842.865   842.8650    842.865
                 SRR1039517 SRR1039520 SRR1039521
ENSG00000186469   2629.130   2473.666   2436.929
ENSG00000186470   2070.421   2233.212   2184.065
ENSG00000186471    384.702    384.702    384.702
ENSG00000186472   2726.088  16990.700   2726.088
ENSG00000186474    842.865    842.865    842.865
```

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

Example:
estimate variance of normally distributed variable

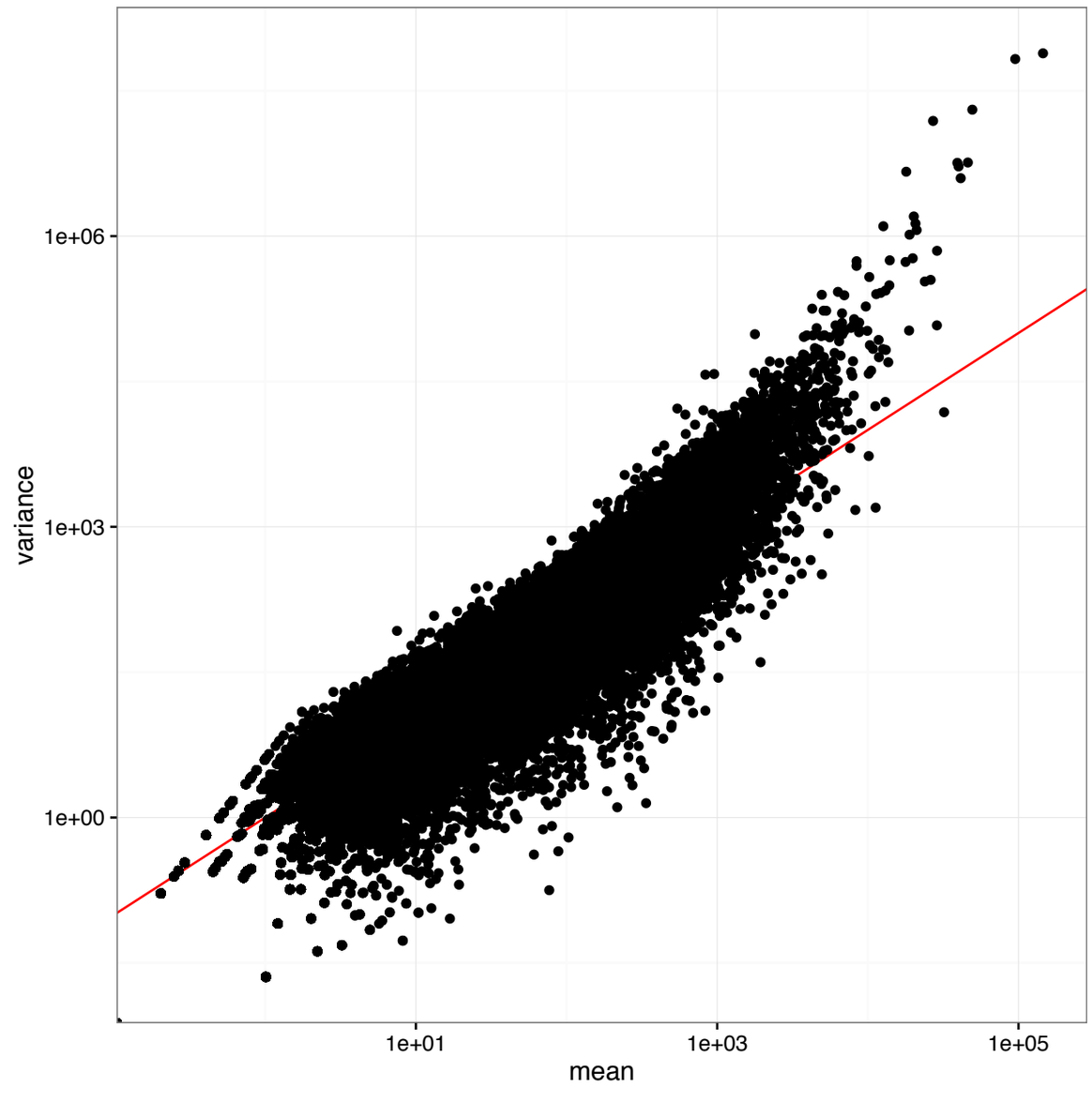
True value = 3



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

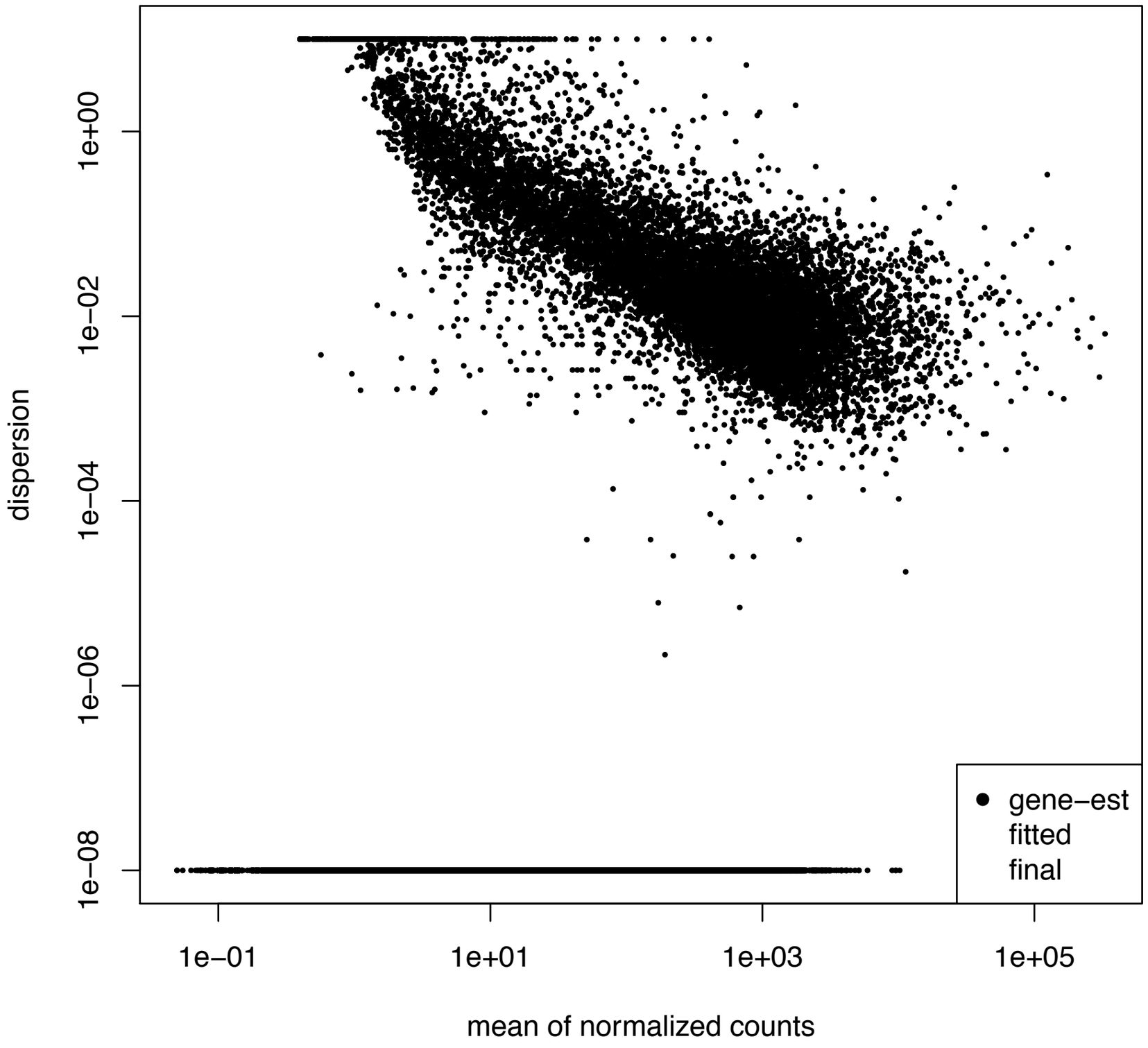
Example from SEQC data, replicates of the same RNA mix



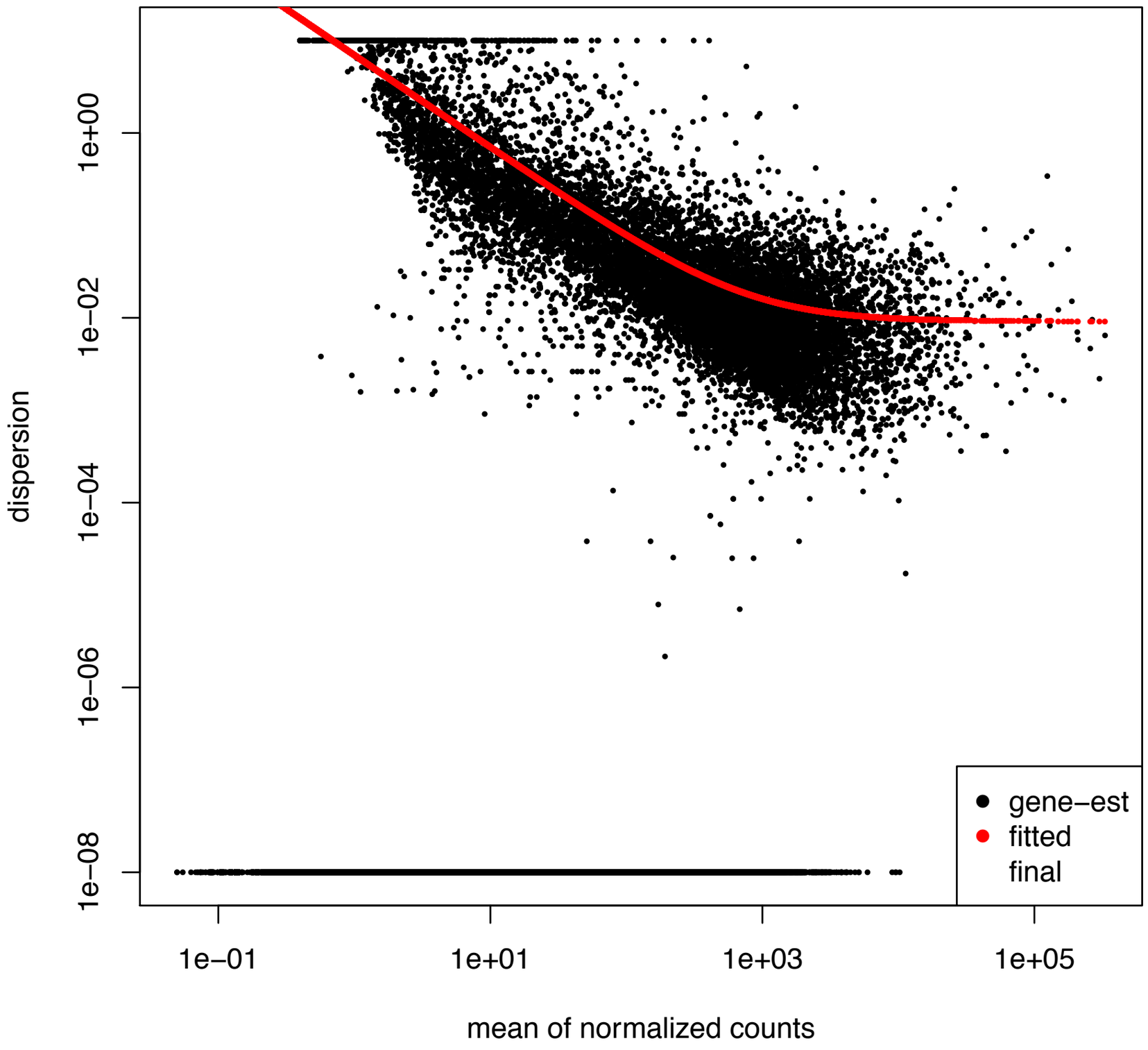
Shrinkage dispersion estimation

- Take advantage of the large number of genes
- Shrink the gene-wise estimates towards a center value defined by the observed distribution of dispersions across
  - all genes (“common” dispersion estimate)
  - genes with similar expression (“trended” dispersion estimate)

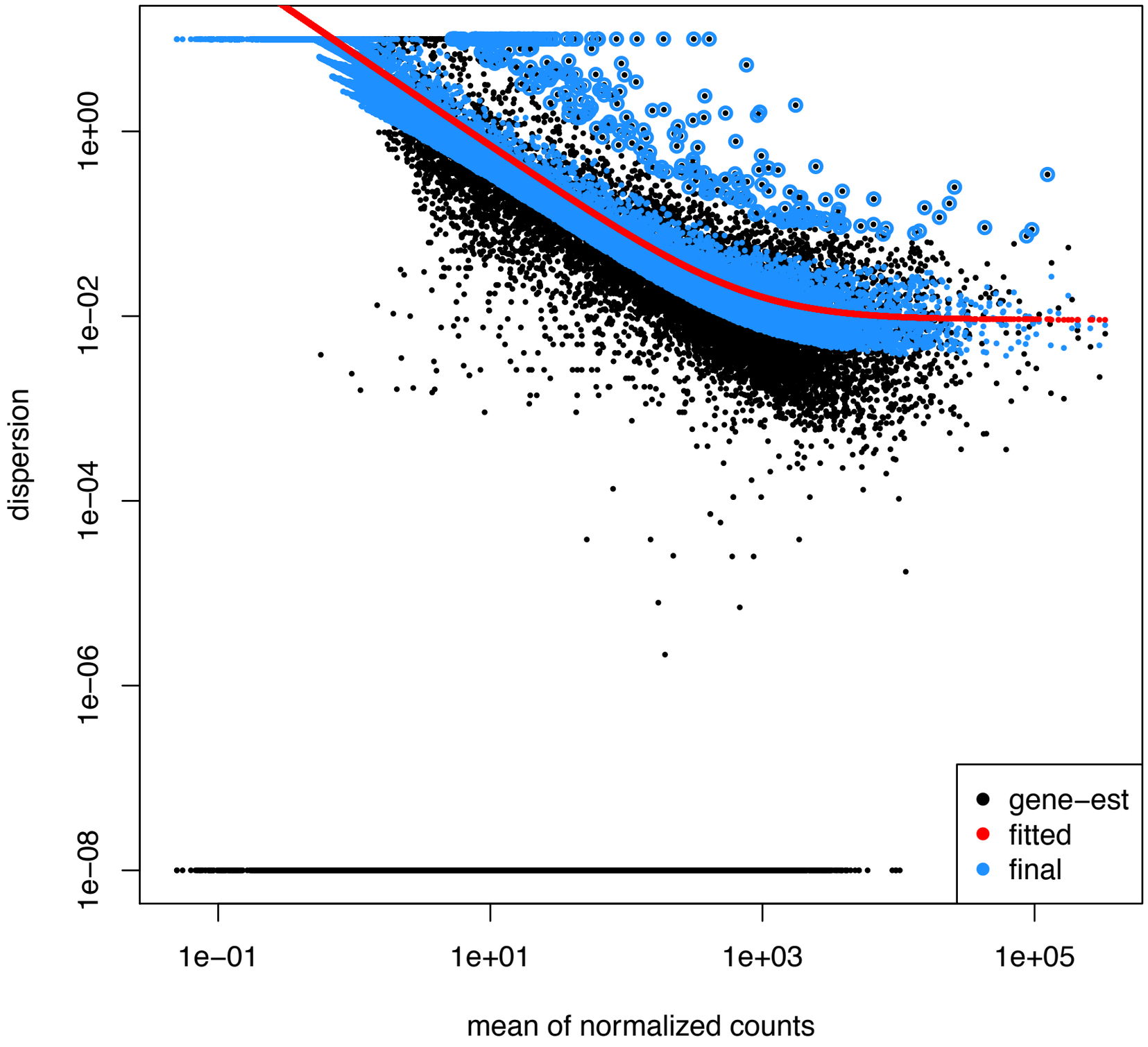
Shrinkage dispersion estimation



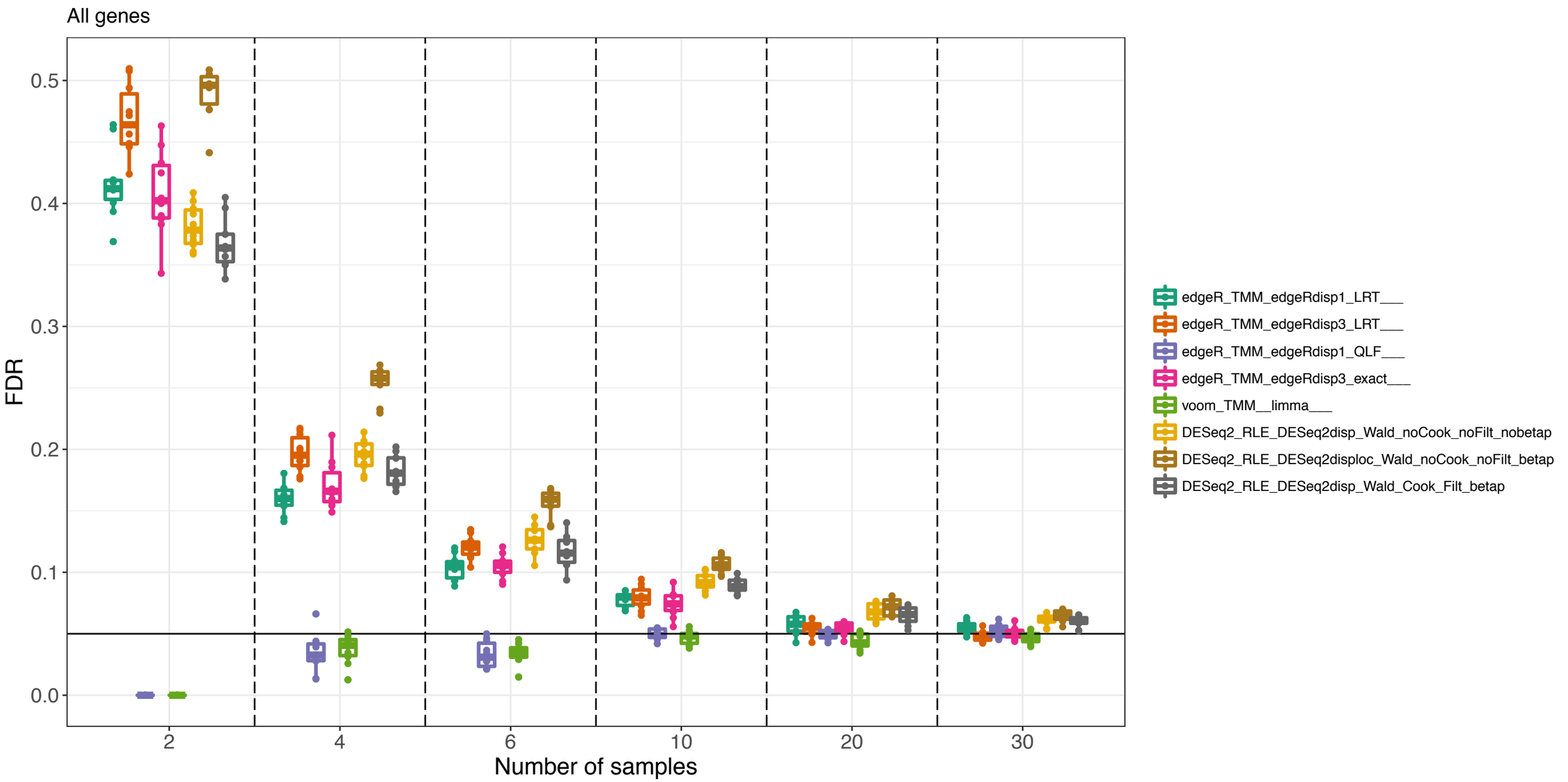
Shrinkage dispersion estimation



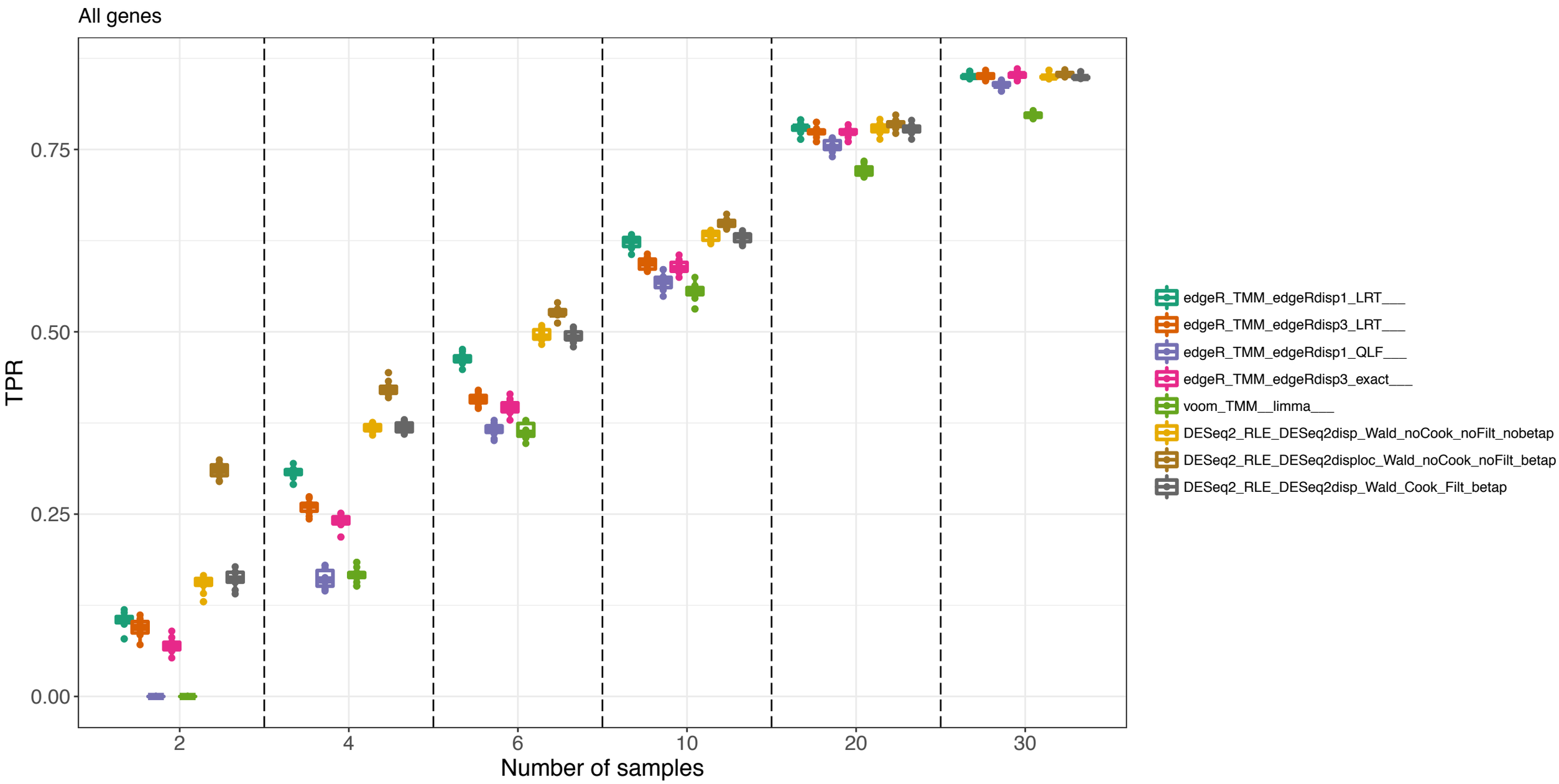
Shrinkage dispersion estimation



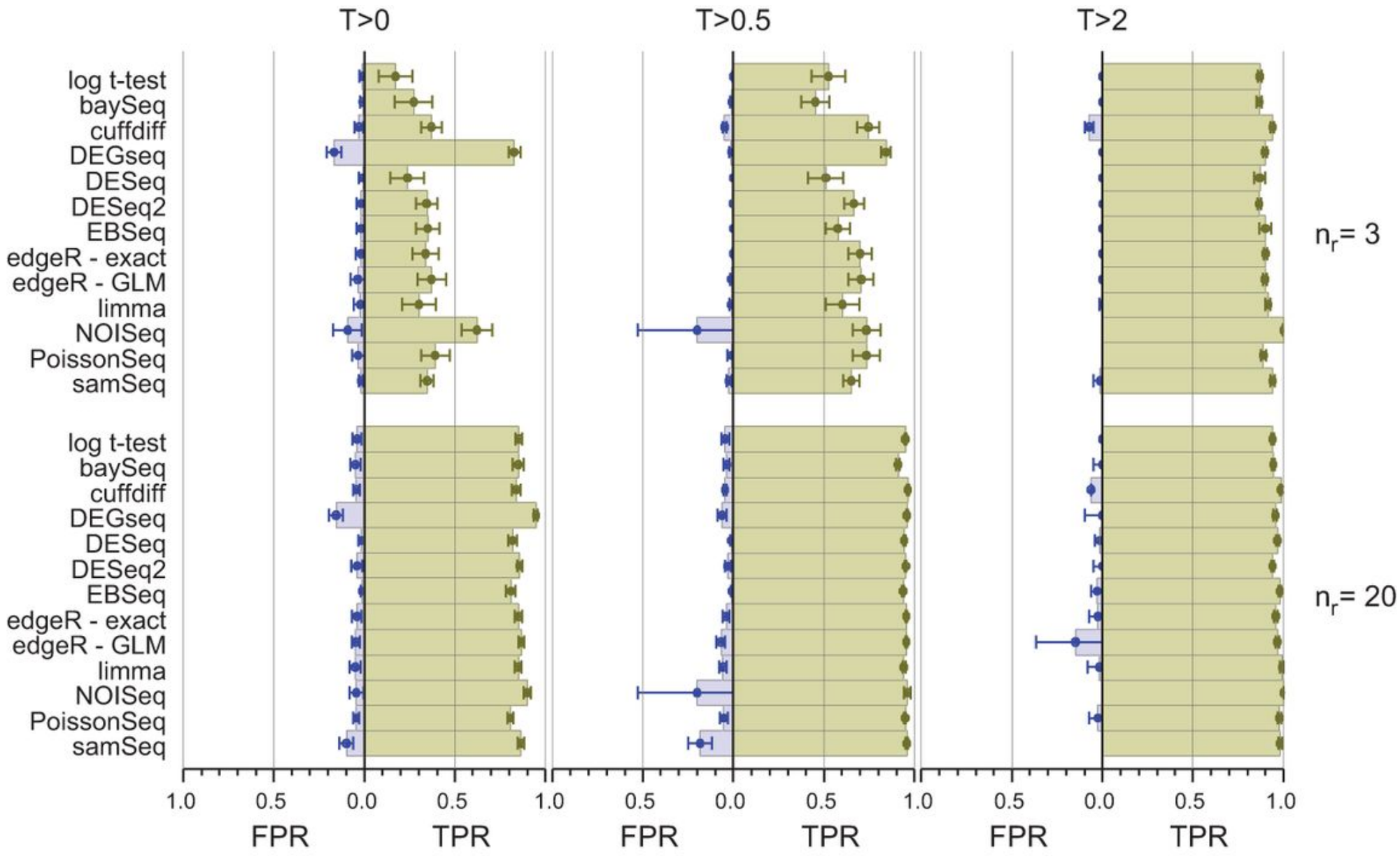
As sample size increases, methods perform better



As sample size increases, methods perform better



Strong signals can be detected with few samples



Schurch et al., RNA (2016)

Method comparison