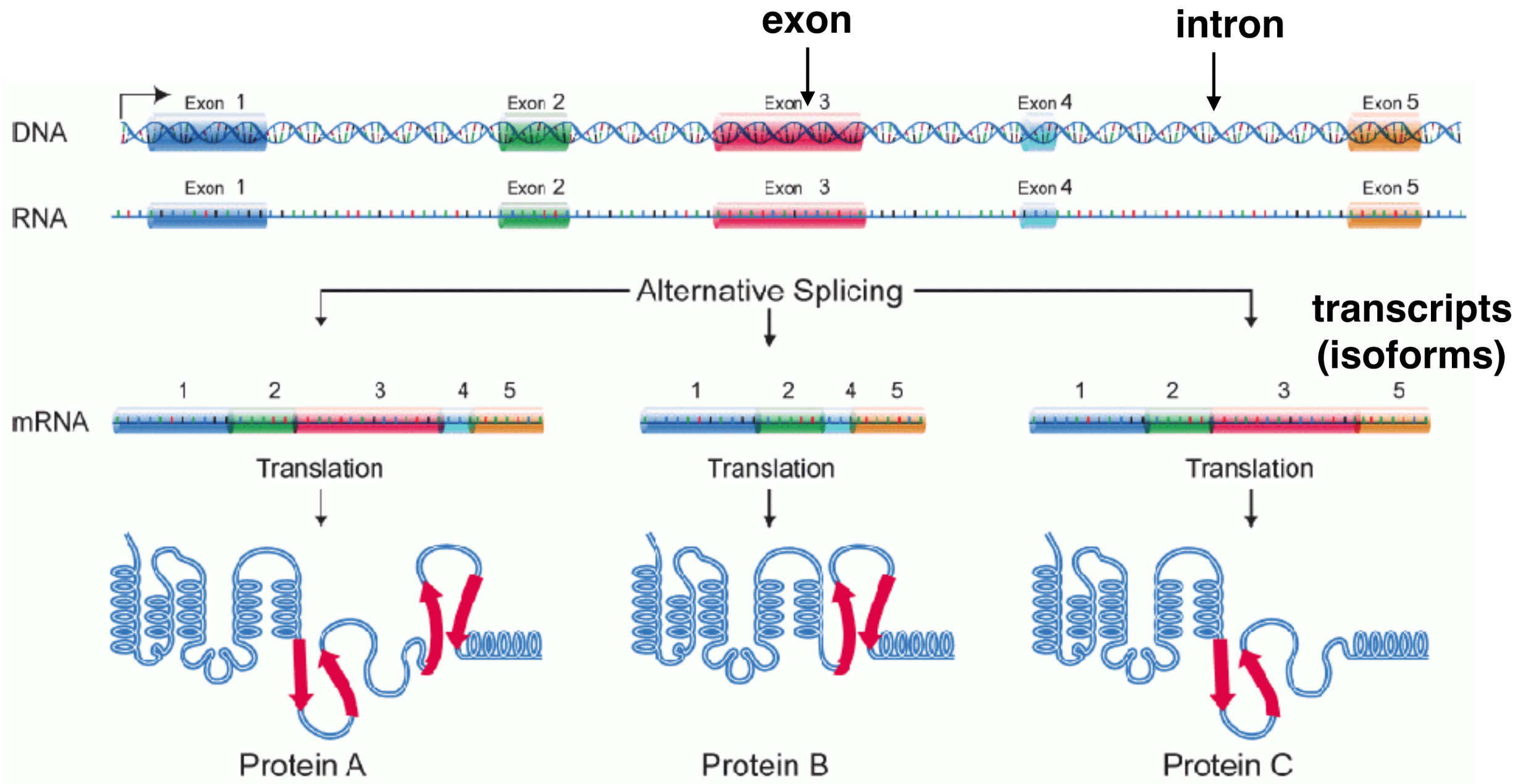
An introduction to RNA-seq

Charlotte Soneson

BIO634, September 18, 2018


exon
intron
DNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
RNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Alternative Splicing
transcripts (isoforms)
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? **D**ifferential **G**ene **E**xpression

- Does the expression of individual transcripts change? **D**ifferential **T**ranscript **E**xpression

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

- Does the isoform composition for a given gene change? **D**ifferential **T**ranscript **U**sage**/D**ifferential **E**xon **U**sage

- need **different** computational approaches (quantifications + tests)

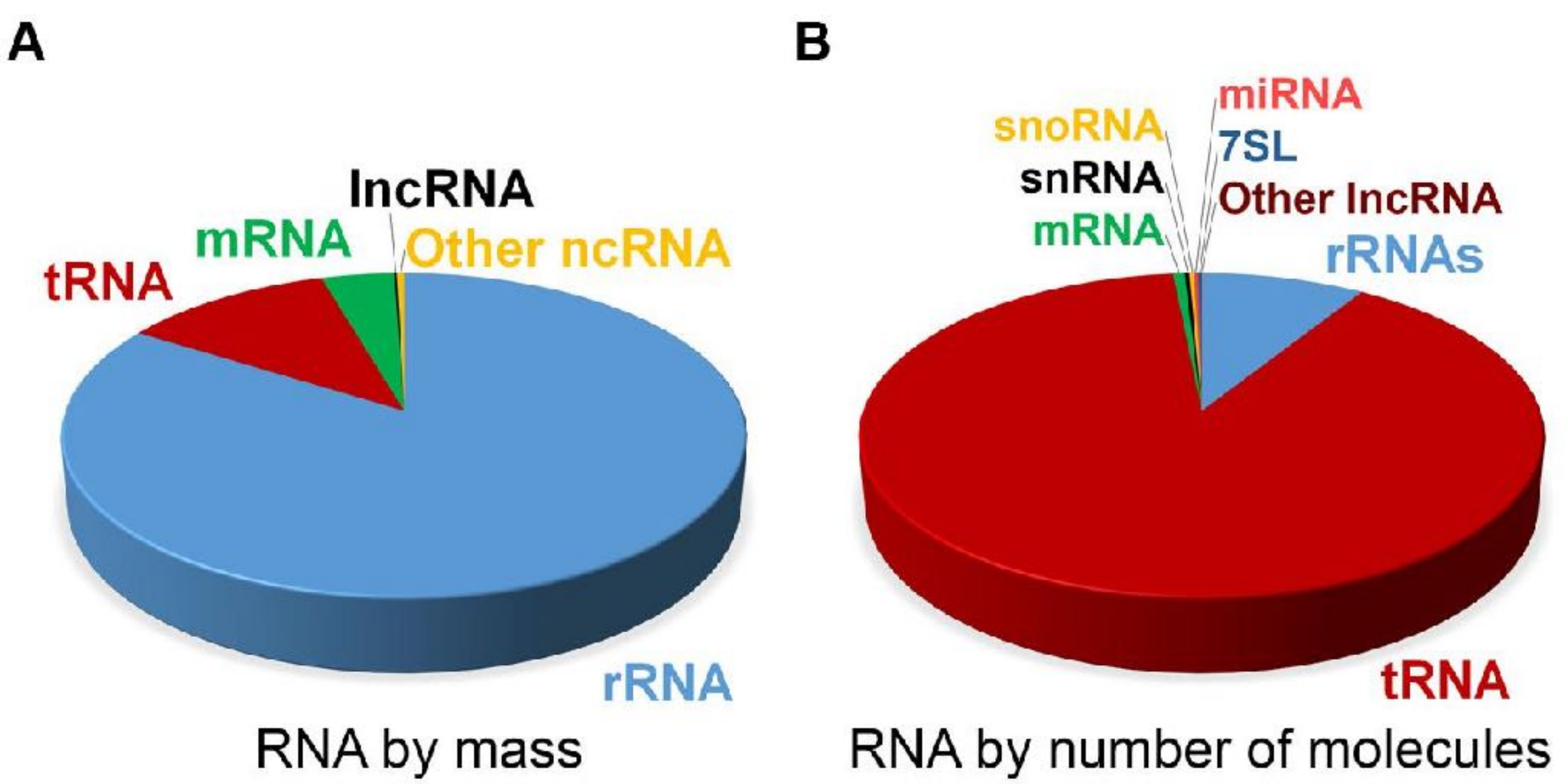
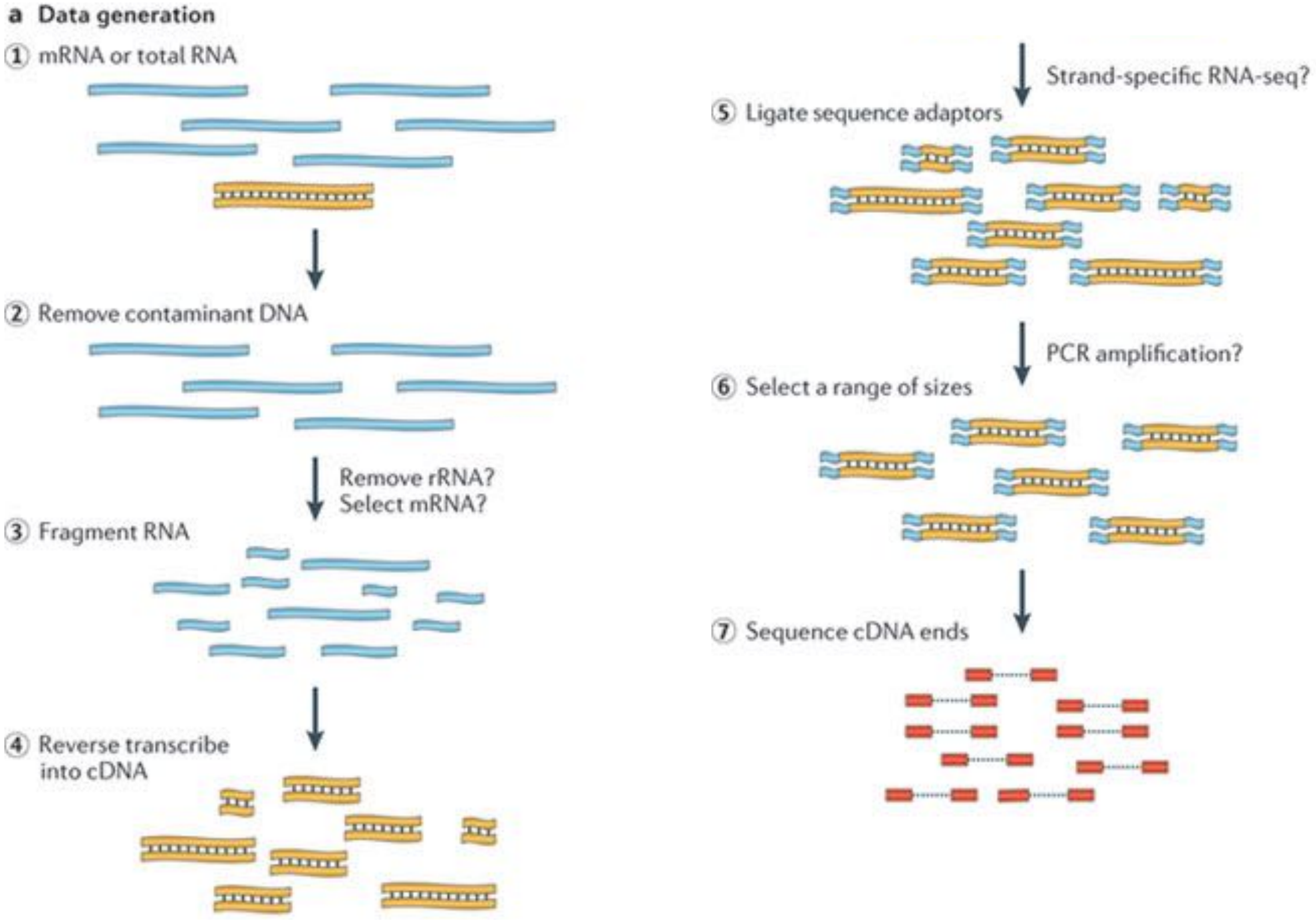


FIGURE 1. Estimate of RNA levels in a typical mammalian cell. Proportion of the various classes of RNA in mammalian somatic cells by total mass (A) and by absolute number of molecules (B). Total number of RNA molecules is estimated at roughly 10^7 per cell. Other ncRNAs in (A) include snRNA, snoRNA, and miRNA. Note that due to their relatively large sizes, rRNA, mRNA, and lncRNAs make up a larger proportion of the mass as compared to the overall number of molecules.

Palazzo & Lee, Front. Genet. 2015

Sequencing



Martin & Wang, Nat. Rev. Genetics (2011)

Single- vs paired-end sequencing

- Each fragment can be sequenced from one end only, or from both ends

- Single-end cheaper and faster

- Paired-end provide improved ability to localize the fragment in the genome and resolve mapping close to repeat regions - less multimapping reads

Strand-specificity

- In “standard” protocols, we don’t know from which strand a read stems

- Various “strand-specific” protocols allow us to keep this information

- Strand-specificity leads to lower number of ambiguous reads (overlapping multiple genes)

RESEARCH ARTICLE OPEN ACCESS

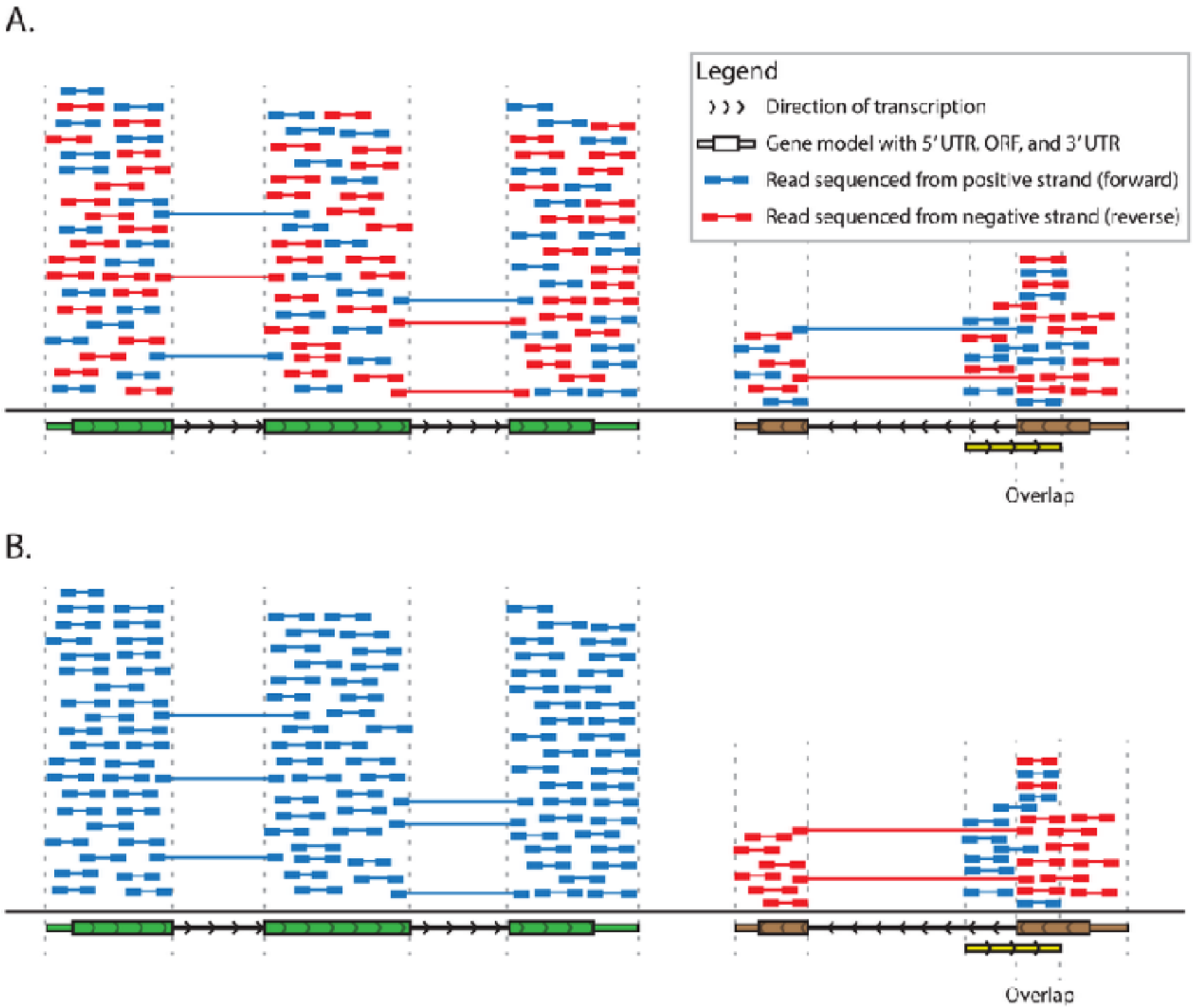
Differentially expressed genes from RNA-Seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols

Susan M. Corley ✉, Karen L. MacKenzie, Annemiek Beverdam, Louise F. Roddam and Marc R. Wilkins

BMC Genomics 2017 18:399 DOI: 10.1186/s12864-017-3797-0 © The Author(s). 2017 ReadCube

Received: 16 December 2016 Accepted: 16 May 2017 Published: 23 May 2017

Strand-specificity



https://doi.org/10.1371/journal.pcbi.1004393.g006

Library preparation is not bias-free

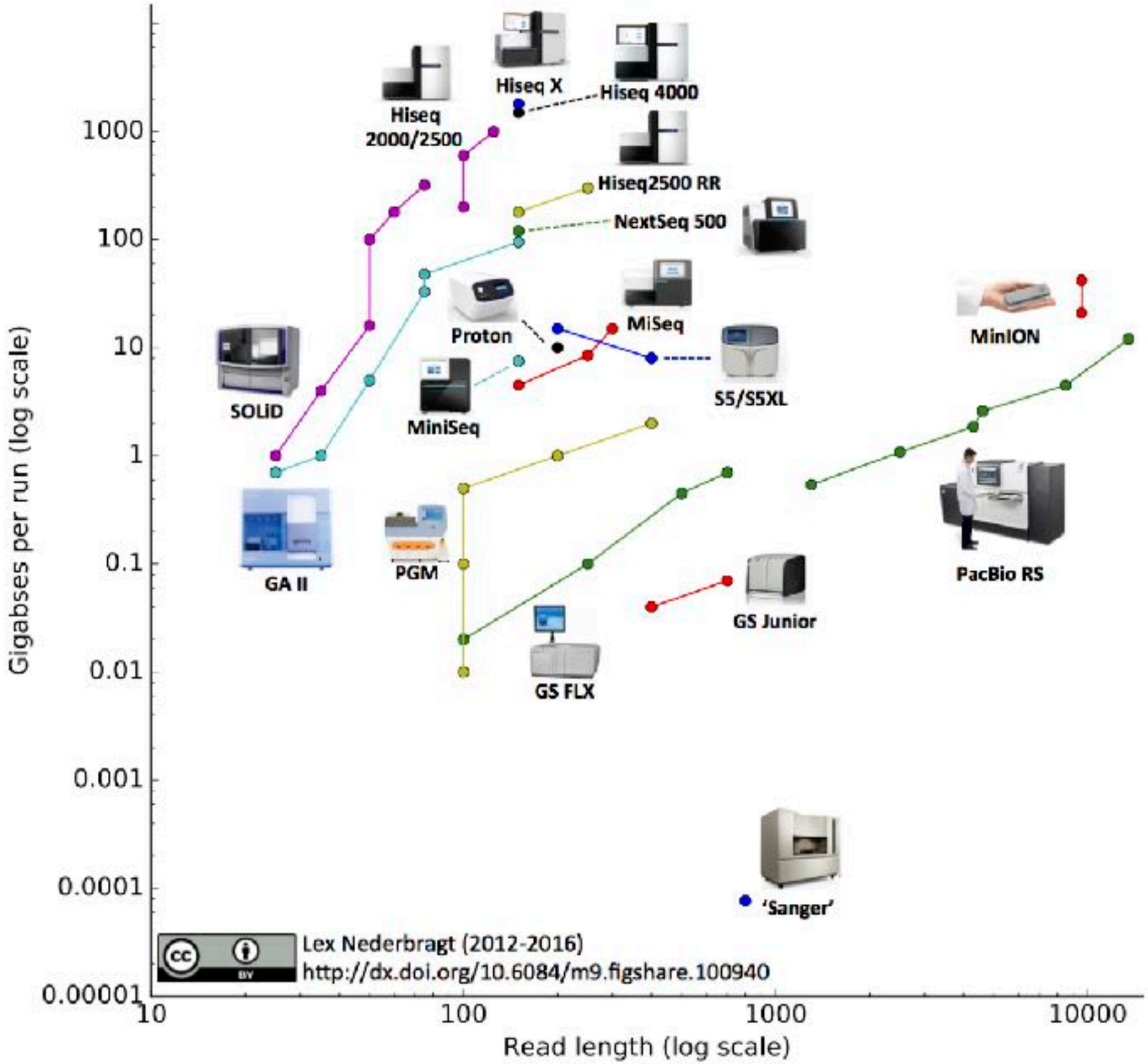
- Selective loss of GC-rich/poor regions
- RNA fragmentation not competely random
- Random hexamer priming for reverse transcription favors certain sequences
- Biased adapter ligation

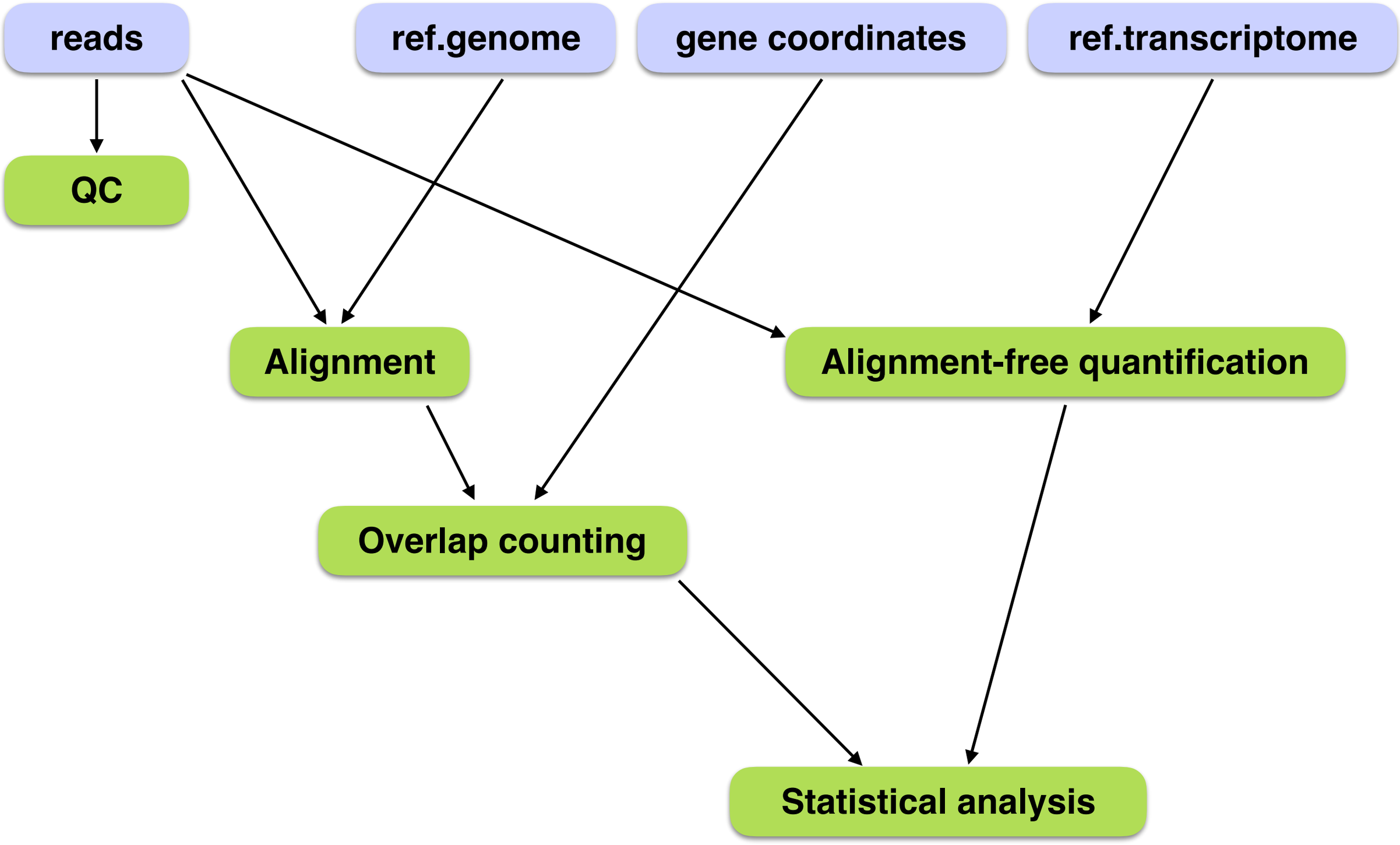
Review Article

Library preparation methods for next-generation sequencing: Tone down the bias

Erwin L. van Dijk[a,], Yan Jaszczyszyn[b], Claude Thermes[a]*

Which sequencing technology?




reads
ref.genome
gene coordinates
ref.transcriptome
QC
Alignment
Alignment-free quantification
Overlap counting
Statistical analysis

Raw reads - FASTQ format

- Combines sequence and base quality information

- Four lines per sequence (read)
    - ID line (starting with @)
    - sequence
    - another ID line (starting with +)
    - base qualities

- For paired-end sequencing: one file for “first” reads and one for “second” reads

FASTQ format - sequence ID line

```
@D7MHBFN1:202:D1BUDACXX:4:1101:1340:1967 1:N:0:CATGCA
NATCTTCGGATCACTTTGGTCAAATTGAAACGATACAGAGAAGATTGTAAGTAACAATATTTACCAAGGTTCGAGTCATACTAACTCGTTGTCCTATAGT
+
#1=DDFFFHHHHHJJJJJJJHIJIJJJIJIJGIIIJJJJJJIIJIJJJHIIFGIIIIJJJJJIIEHJIIHHGFFF@?ADFEDDEDCDDBDDBDCDDDDEC
```

- D7MHBFNI - unique instrument name
- 202 - run ID
- D1BUDACXX - flowcell ID
- 4 - flowcell lane
- 1101 - tile number within lane
- 1340 - x-coordinate of cluster within tile
- 1967 - y-coordinate of cluster within tile
- 1 - member of pair (1 or 2). Older versions: /1 and /2
- Y/N - whether the read failed quality control (Y = bad)
- 0 - none of the control bits are on
- CATGCA - index sequence (barcode)

FASTQ format - base qualities

- For each letter, estimate the probability of being erroneous (p)
- Phred score $Q = -10 \cdot \log_{10}(p)$

Phred score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

Quality format encoding

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................
 .........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 .....................................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 .....................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                               0........9.............................40
                                    3.....9.............................40
  0.......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

“Capital letters = good quality” (with Illumina 1.8+)

http://www.wikipedia.com/FASTQ_format

The (human) reference genome

- A “representative example” of the human genome sequence

- New versions are released periodically (the latest, GRCh38, in December 2013)

- Coordinates are not comparable across versions

The reference genome

- Typically provided as a **fasta** file - general sequence representation

- Two lines per sequence (e.g., chromosome)
  - Header line (starting with >)
  - Sequence

```
>chr1
......
GTTCTTGTTCTGTGTTCTTATAACCATACCAGAATTTTCTTCATCACAGA
CAGAGACTAAACTCTTTCTTCTCTTACCTTTCCTTTGATAATATTTTTGA
TCCAGGAATGGGGATAATTTTGCAGTTAAAATTTTCTTTTTATGATGGAA
GGTGAGGAGGAGAGAGAGGTTTACATTAGAAGTGACCCAACTCCATTTTC
TTCCAATGGTTTTTTTCAGTTTTATTTTTTTAAAGCGTGAACAGAGAATA
GTCACCTGATCAATTTAAATATGTCAAAAAGTGAAAGAAAAATCTCTCTT
TTAAAGGAAATGAGGGCAGTAACACAACCAAGGAATCAAAATTCAGGTTG
AGGCTGACCTTTGACCTGCAACTATGCTACTCCATGAACAGCAAGTAGGA
AATGGCTGATTTCATGAAGGTGGACTGGCATCAGAGGAGGCGAGGGATCC
AGGGTTCCTGATGAGTGGCAACATTCCTTGGTCTTTTGAGTTTGTTTGAT
TGGTGAATCAAATTTAGGTGACAGCCAGCTAAAGAGAGTGAGGGTGGCTG
TCTTGTGAATGGGAAGTGACCAAGCTTGAAAGCACAGACTgtggtggctc
......
```

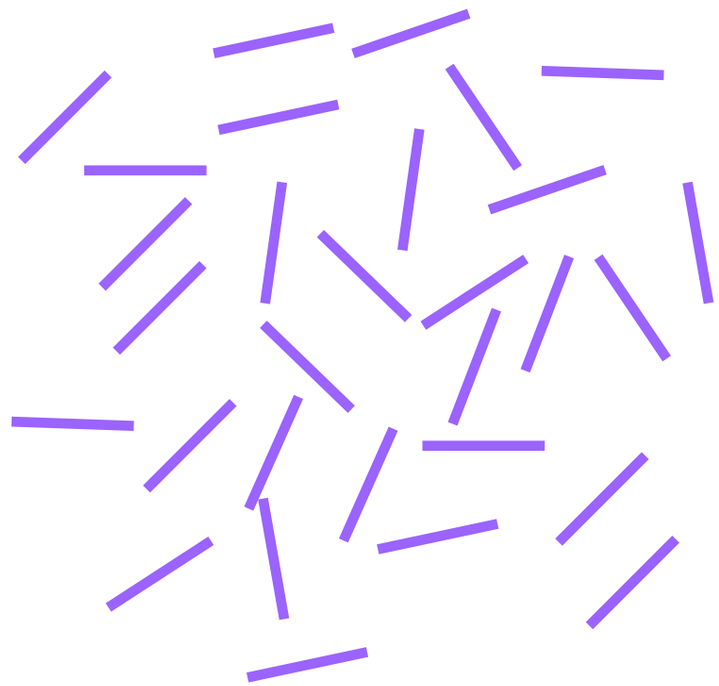
Locations of genes on reference genome

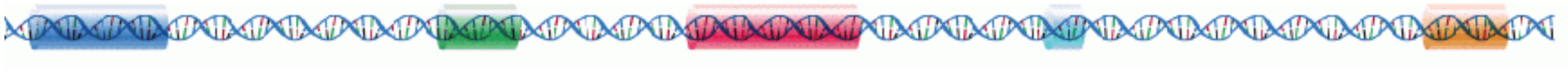
- Typically provided in a **gtf** (gene transfer format) file
- Similar to **gff**, but more restrictive

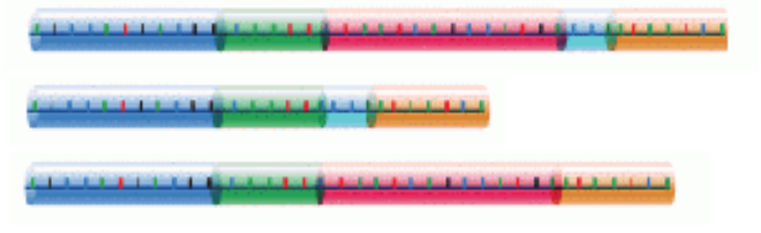
```
seqname  source  feature  start  end  score  strand  frame  attribute

2R   protein_coding   exon  5139815   5141712   .   -   .   gene_id "FBgn0020621"; transcript_id
"FBtr0112897"; exon_number "10"; gene_name "Pkn"; gene_biotype "protein_coding"; transcript_name "Pkn-RG";
exon_id "FBgn0020621:1";
2R   protein_coding   CDS   5141572   5141712   .   -   0   gene_id "FBgn0020621"; transcript_id
"FBtr0112897"; exon_number "10"; gene_name "Pkn"; gene_biotype "protein_coding"; transcript_name "Pkn-RG";
protein_id "FBpp0111810";
2R   protein_coding   stop_codon 5141569   5141571   .   -   0   gene_id "FBgn0020621"; transcript_id
"FBtr0112897"; exon_number "10"; gene_name "Pkn"; gene_biotype "protein_coding"; transcript_name "Pkn-RG";
```

http://www.ensembl.org/info/website/upload/gff.html

Abundance quantification

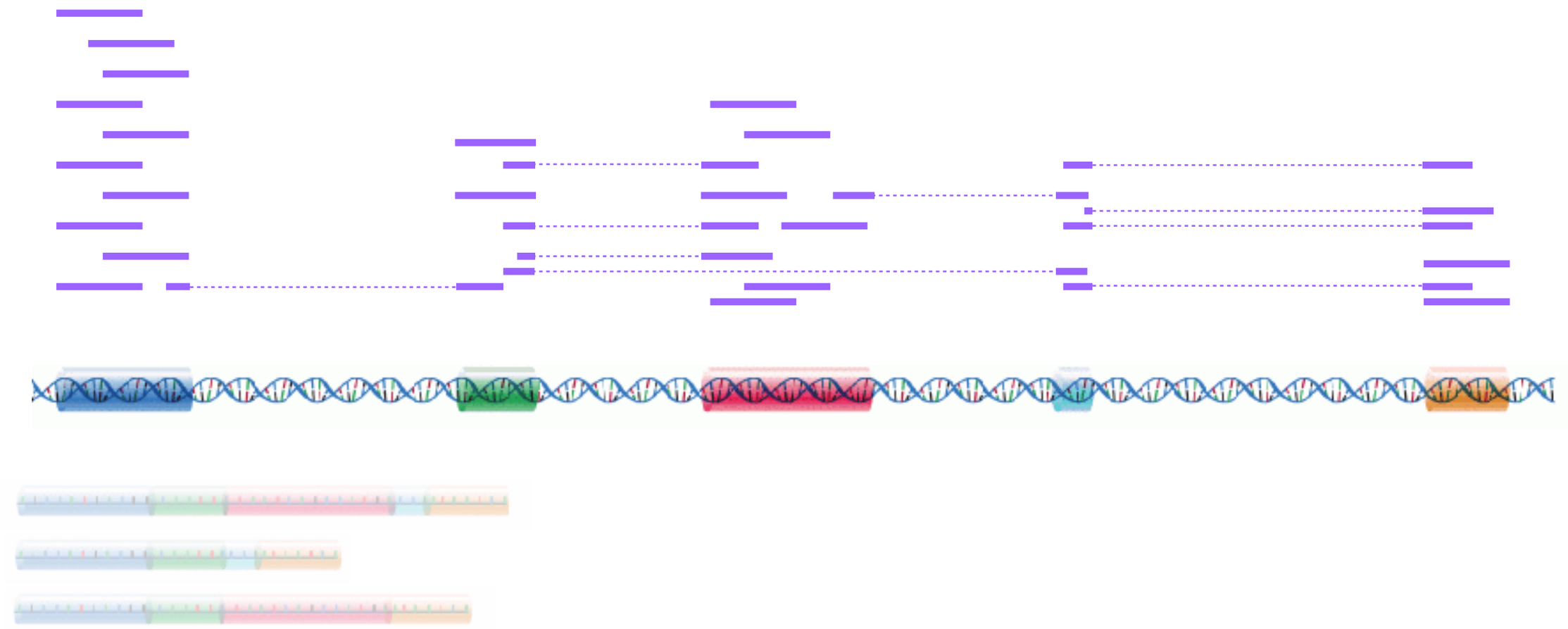






Abundance quantification

Genome alignment of RNA-seq requires a splice-aware aligner (STAR, HISAT2)



Representing alignments - SAM format

- Header

```
@SQ  SN:chr1   LN:249250621
@SQ  SN:chr2   LN:243199373
@SQ  SN:chr3   LN:198022430
@SQ  SN:chr4   LN:191154276
```

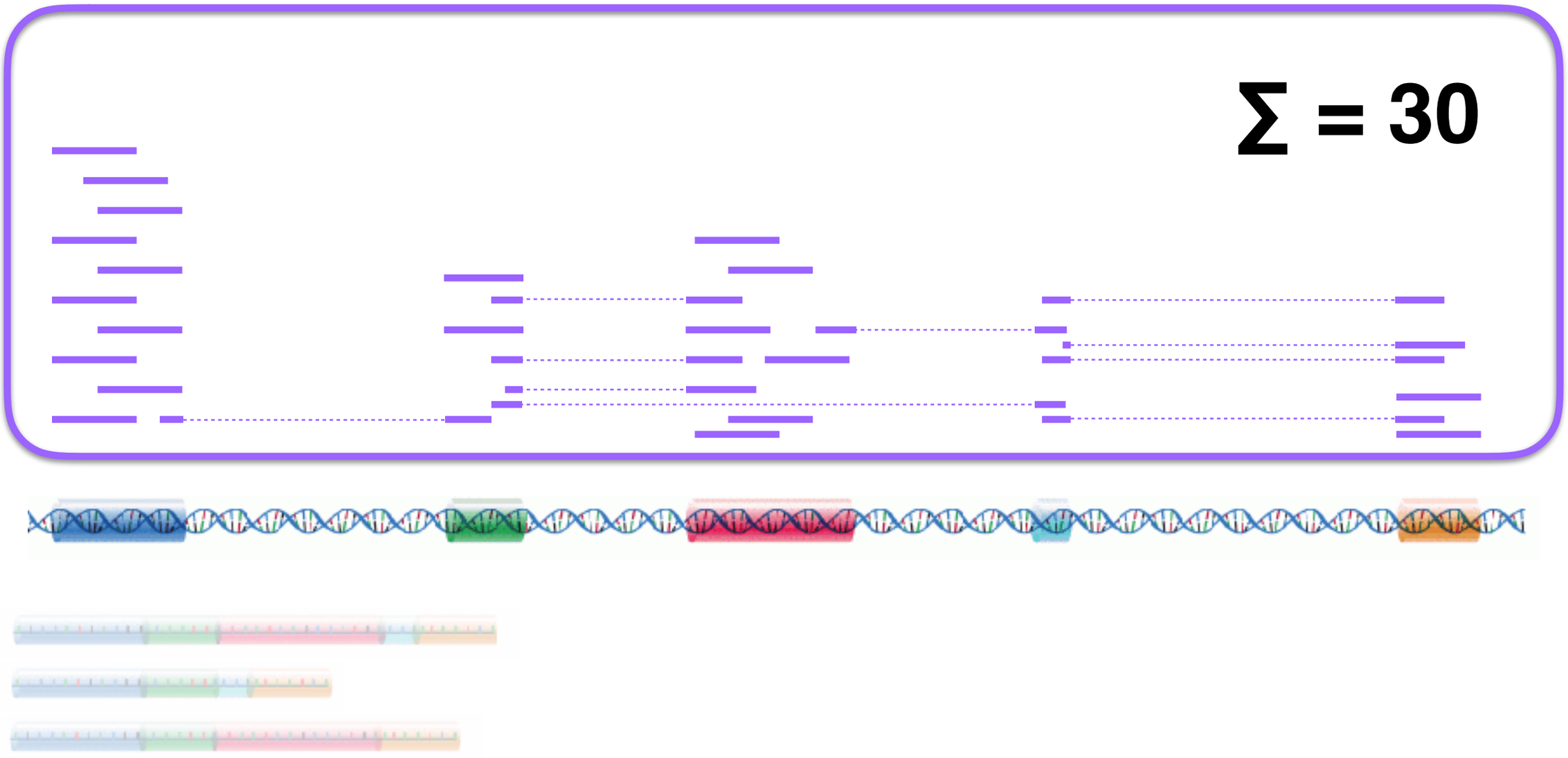
- Body

```
seq.13906018  0    chr10    101948233 255  101M *    0    0
GTCCACAGTCCTTTCTCTGAAACCCTTGGGNNAAGTTGTTTCAGAATTANGNAA   CBCFFFFFHHHHHJJJJJJJJJJJJJJJJ##11?
DHIIIIJJHIJJJJ#0#07    0L:A:F    IH:i:1    HI:i:1
```

  - Typically, one line per alignment
  - BAM = binary SAM

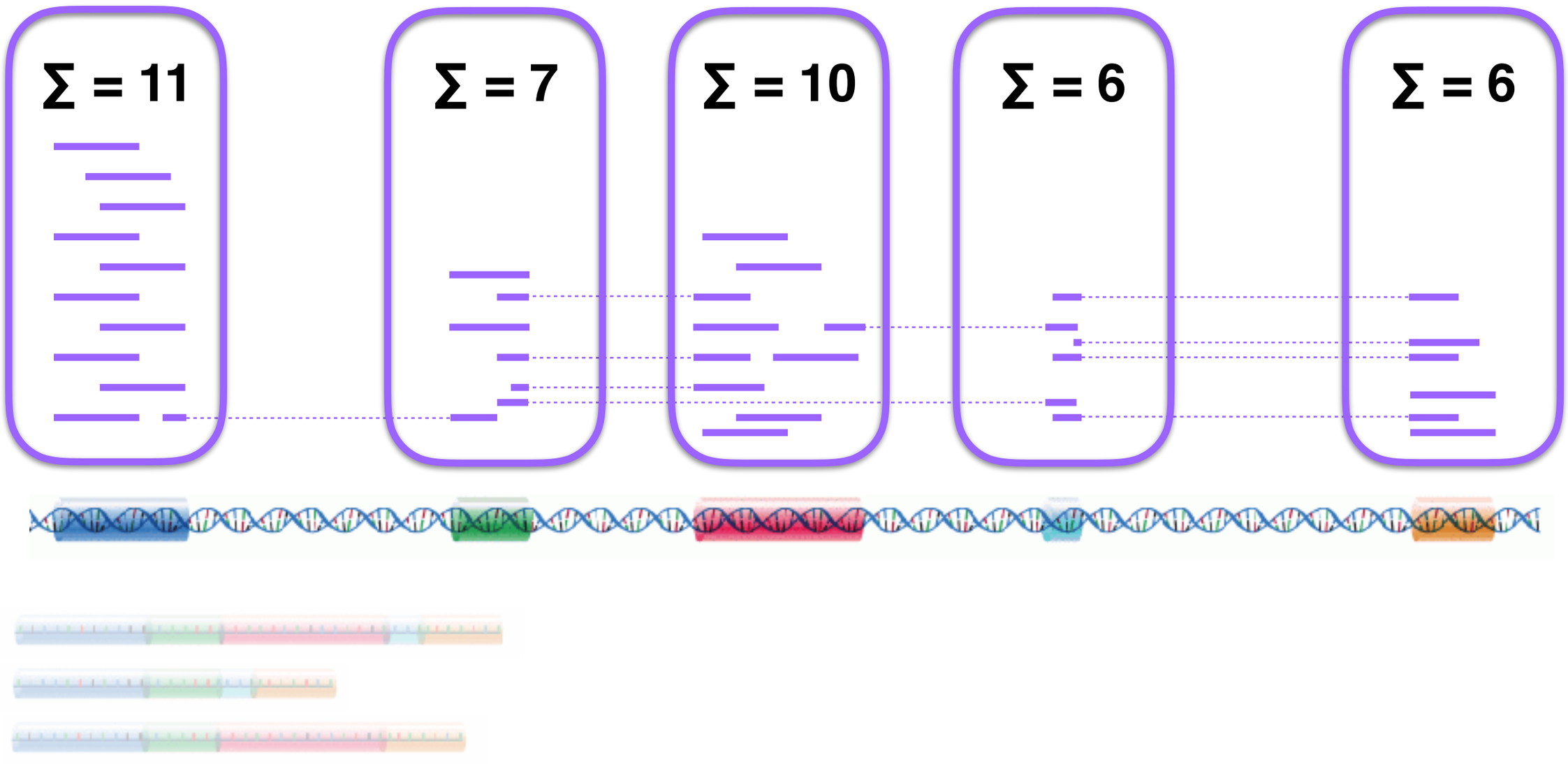
Abundance quantification

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



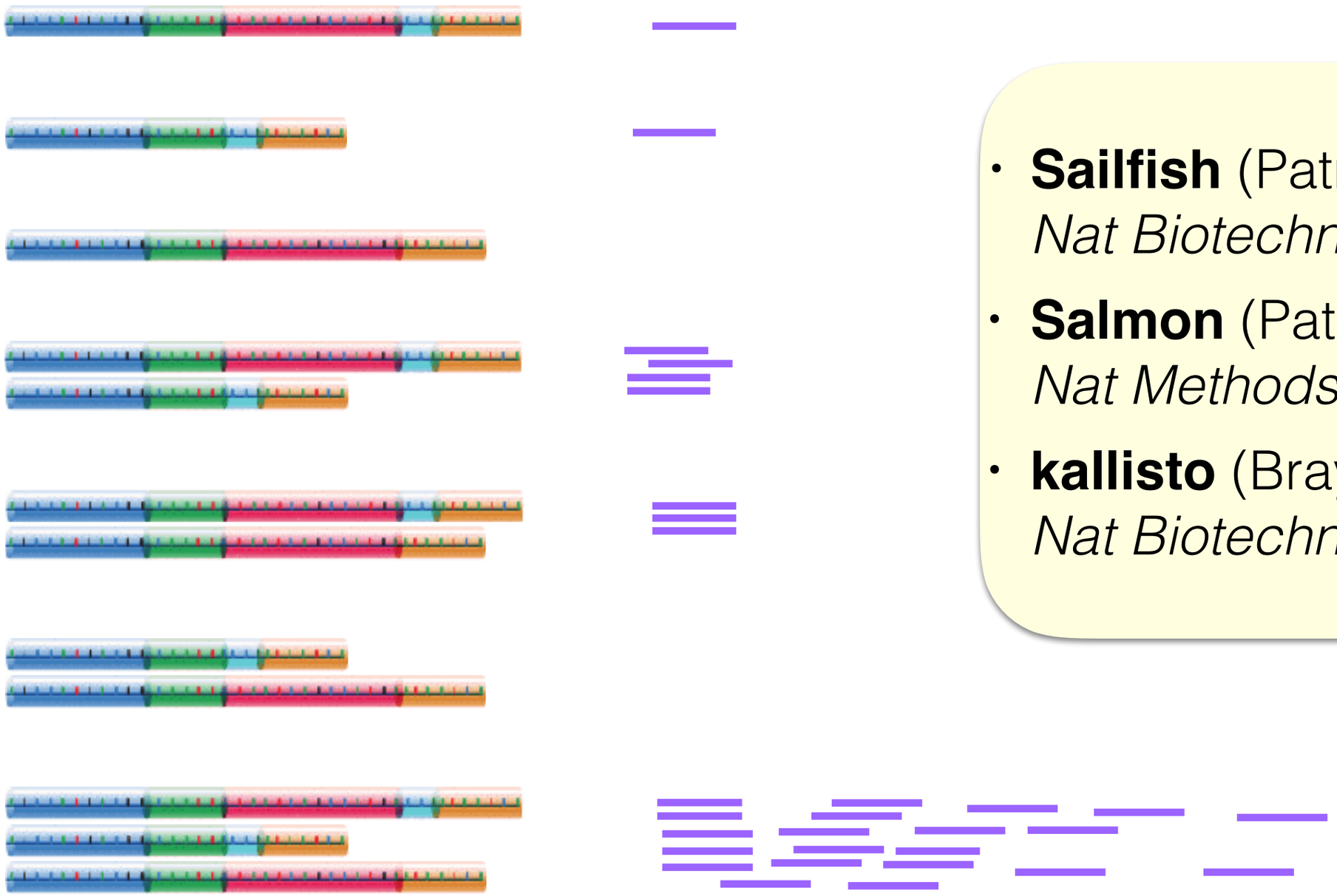
Abundance quantification

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



Abundance quantification

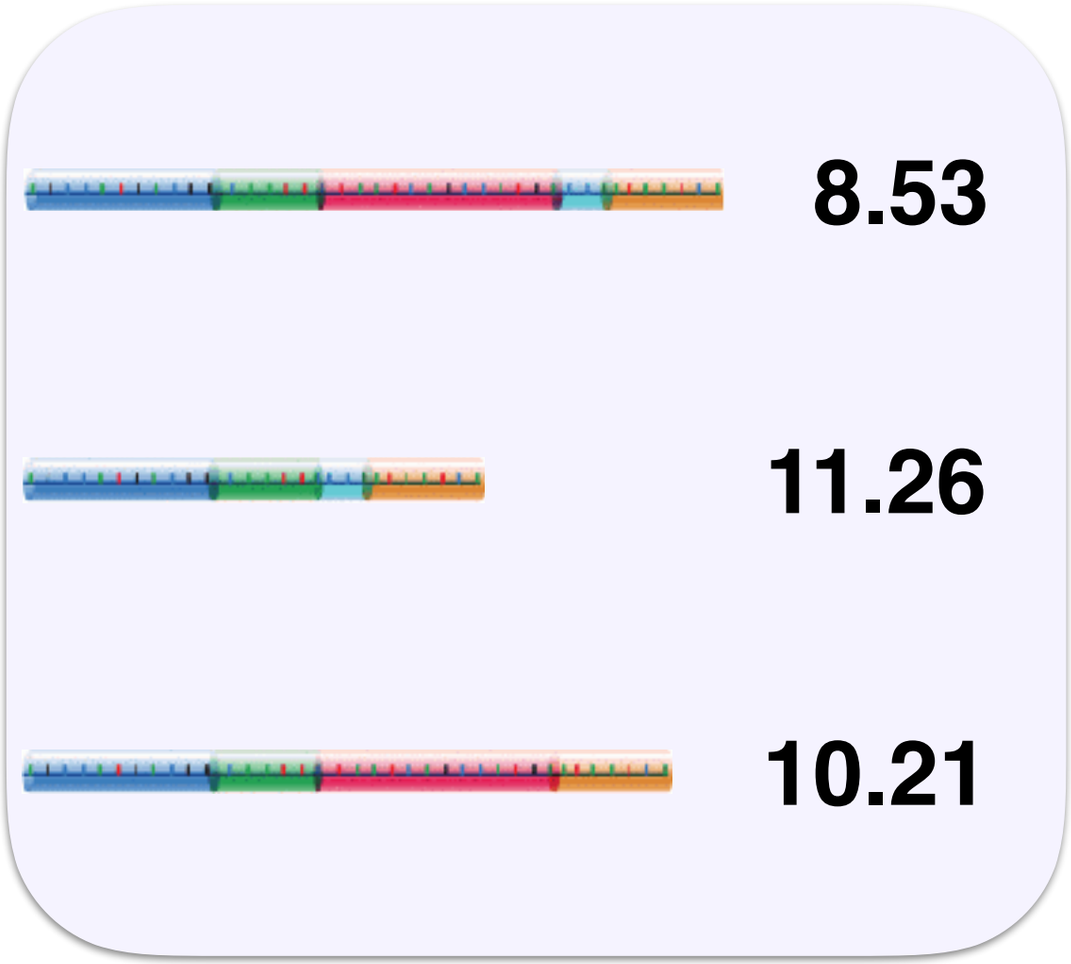
Transcript-level counts, e.g. obtained by "alignment-free" estimation methods



- **Sailfish** (Patro et al, *Nat Biotechnol* 2014)
- **Salmon** (Patro et al, *Nat Methods* 2017)
- **kallisto** (Bray et al, *Nat Biotechnol* 2016)

Abundance quantification

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



Abundance quantification

Gene-level counts, obtained by summation of transcript counts

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)
```

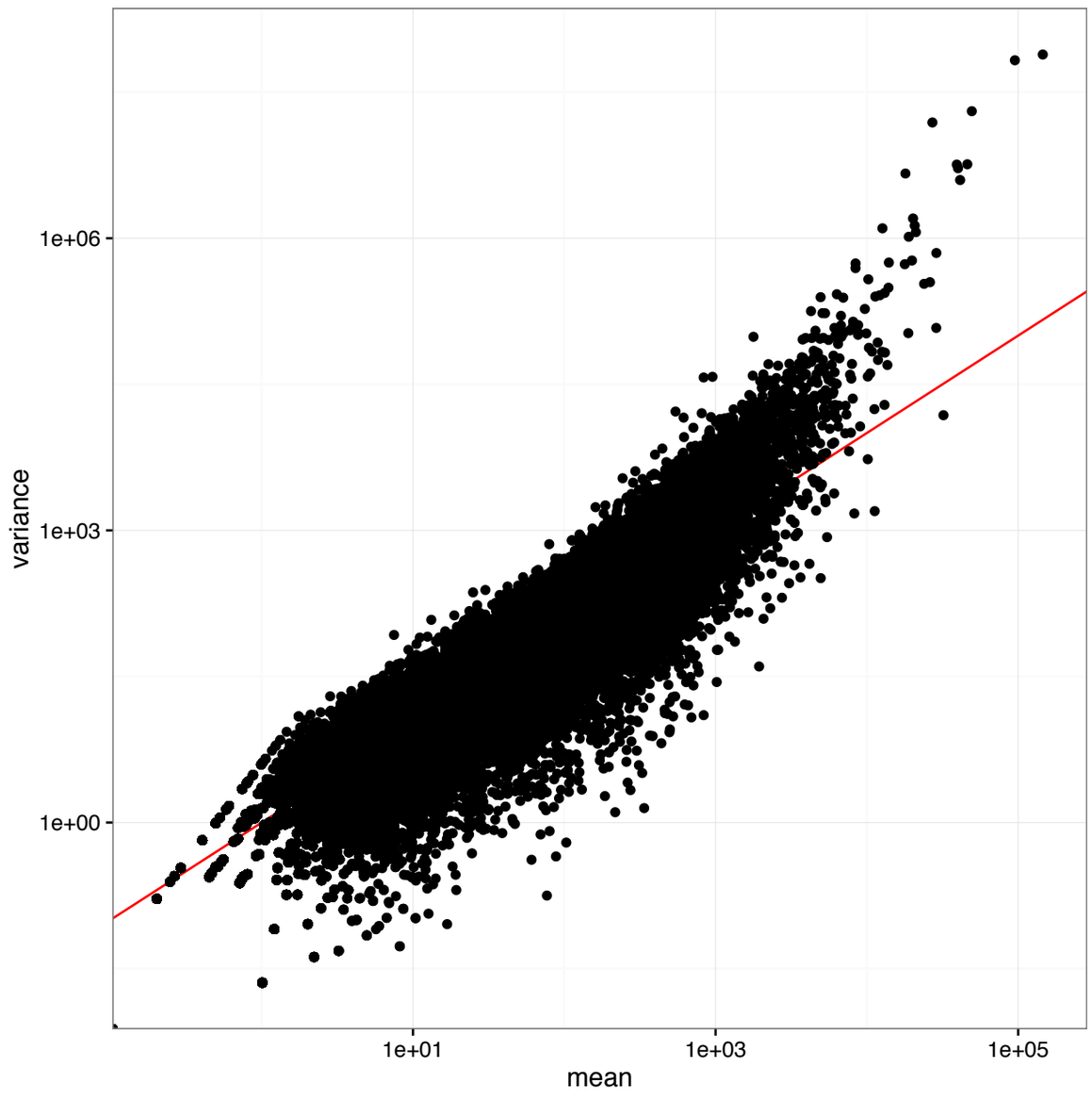
Challenges for RNA-seq data

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

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



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

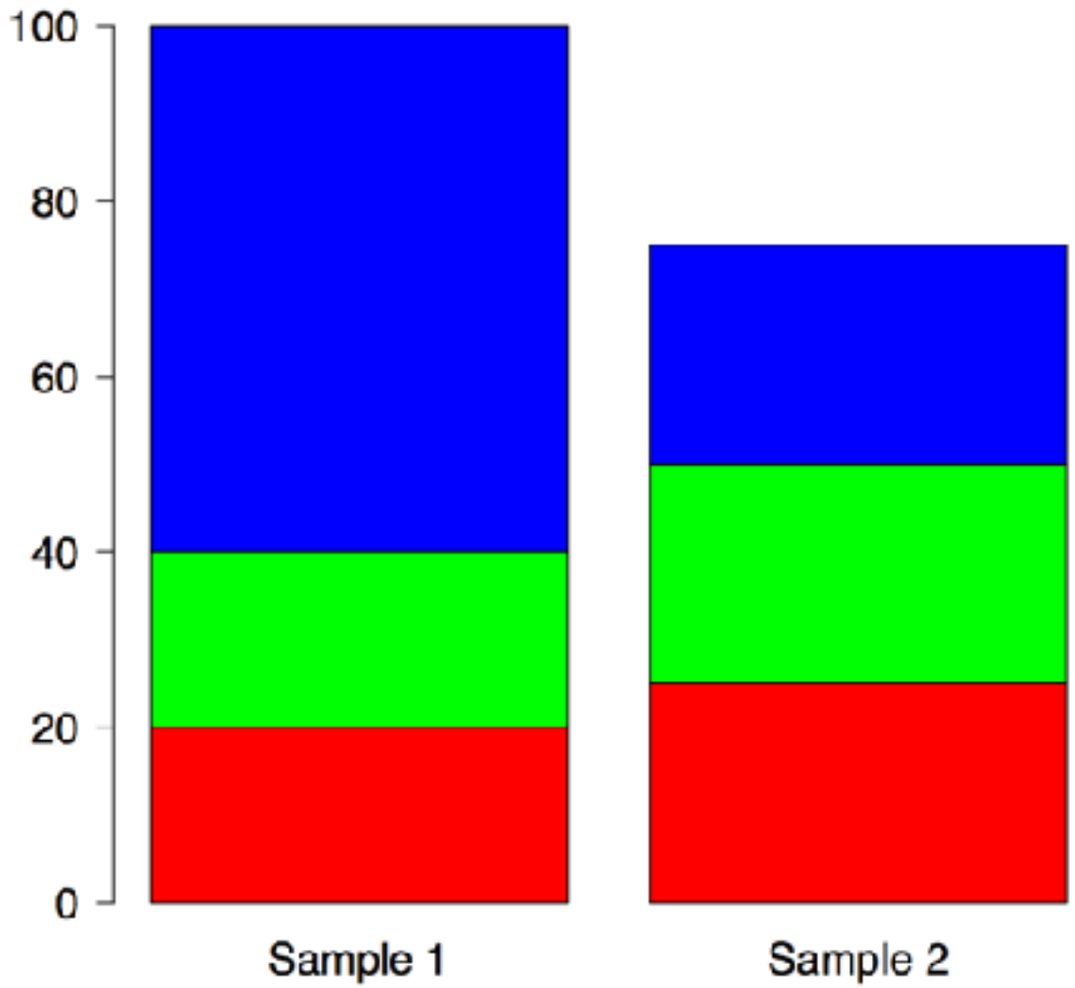
- counts are not explicitly scaled

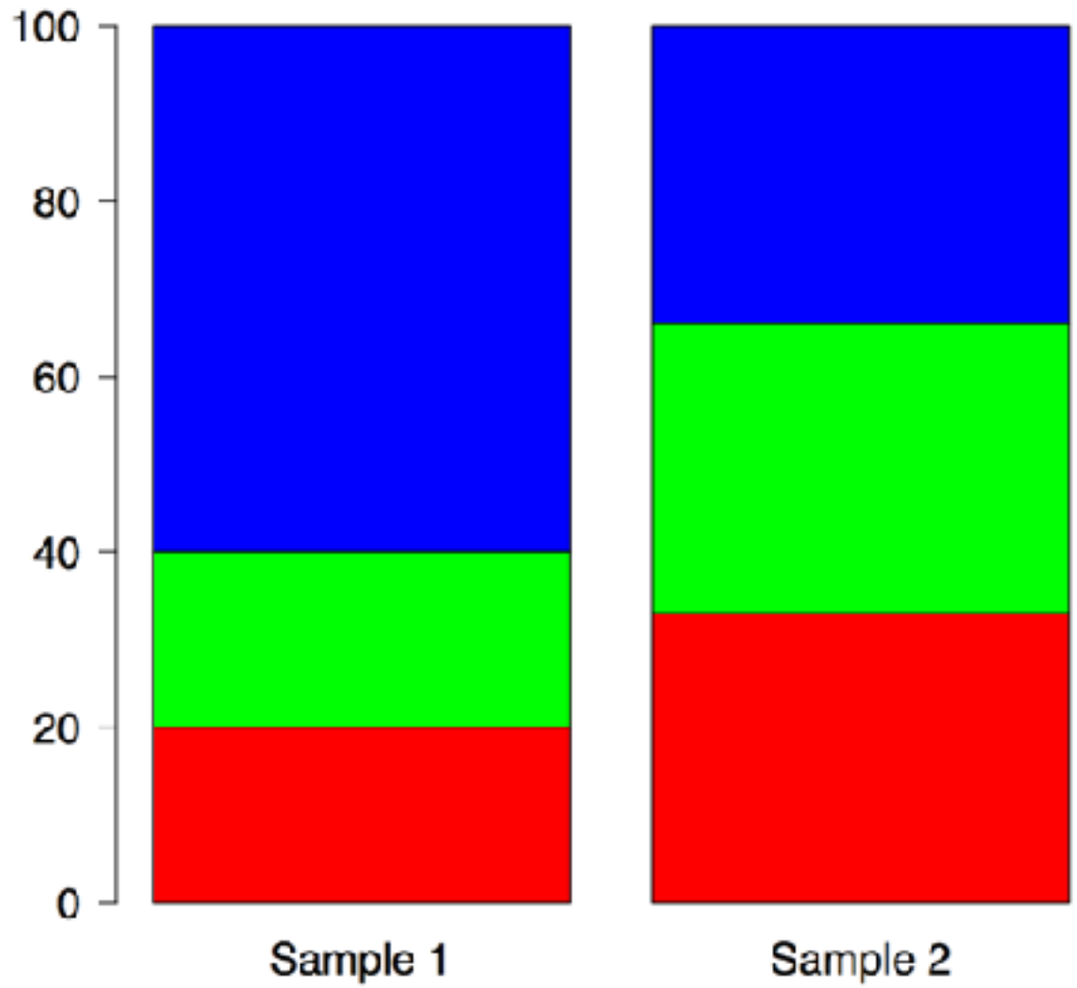
  - important exception: voom/limma (followed by explicit modeling of mean-variance association)

Adapted from Mike Love

The influence of RNA composition

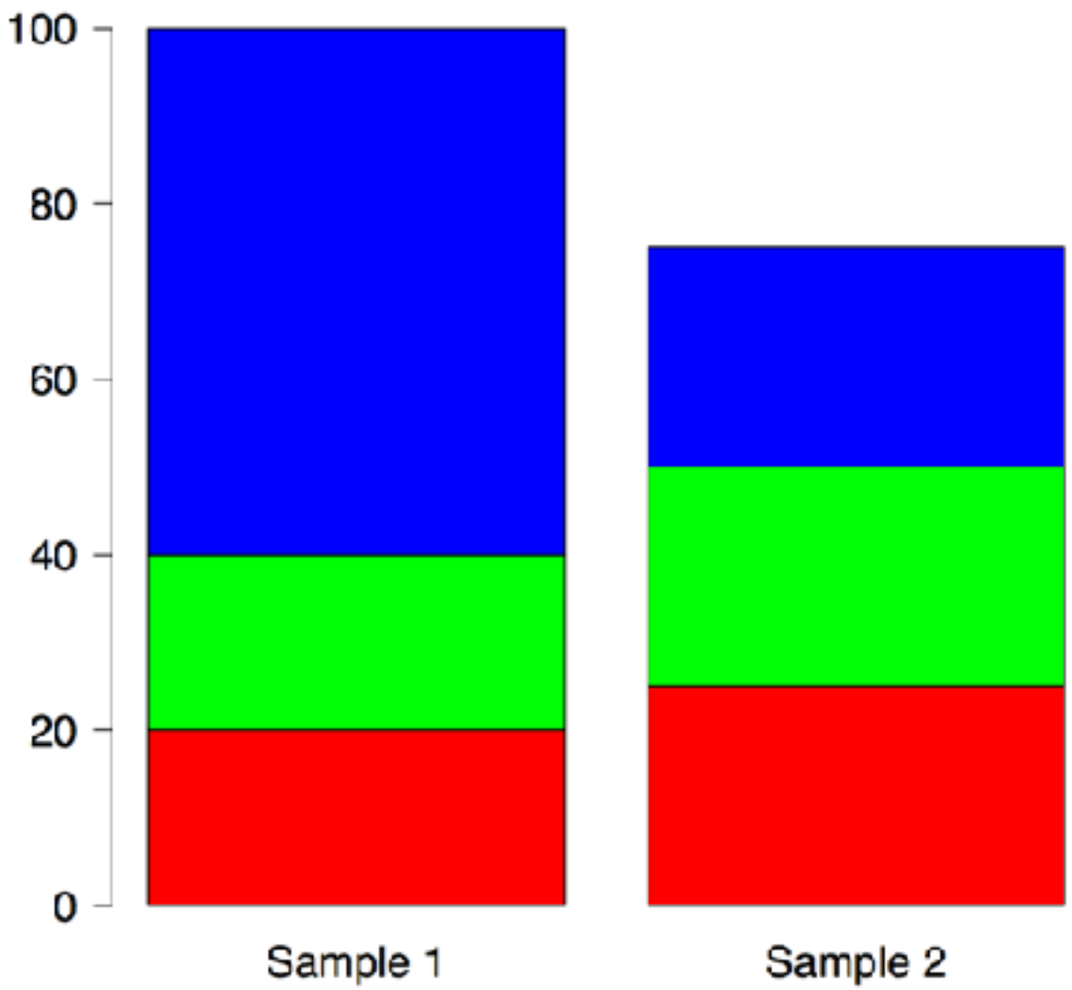
- Observed counts are relative
- High counts for some genes are “compensated” by low counts for other genes

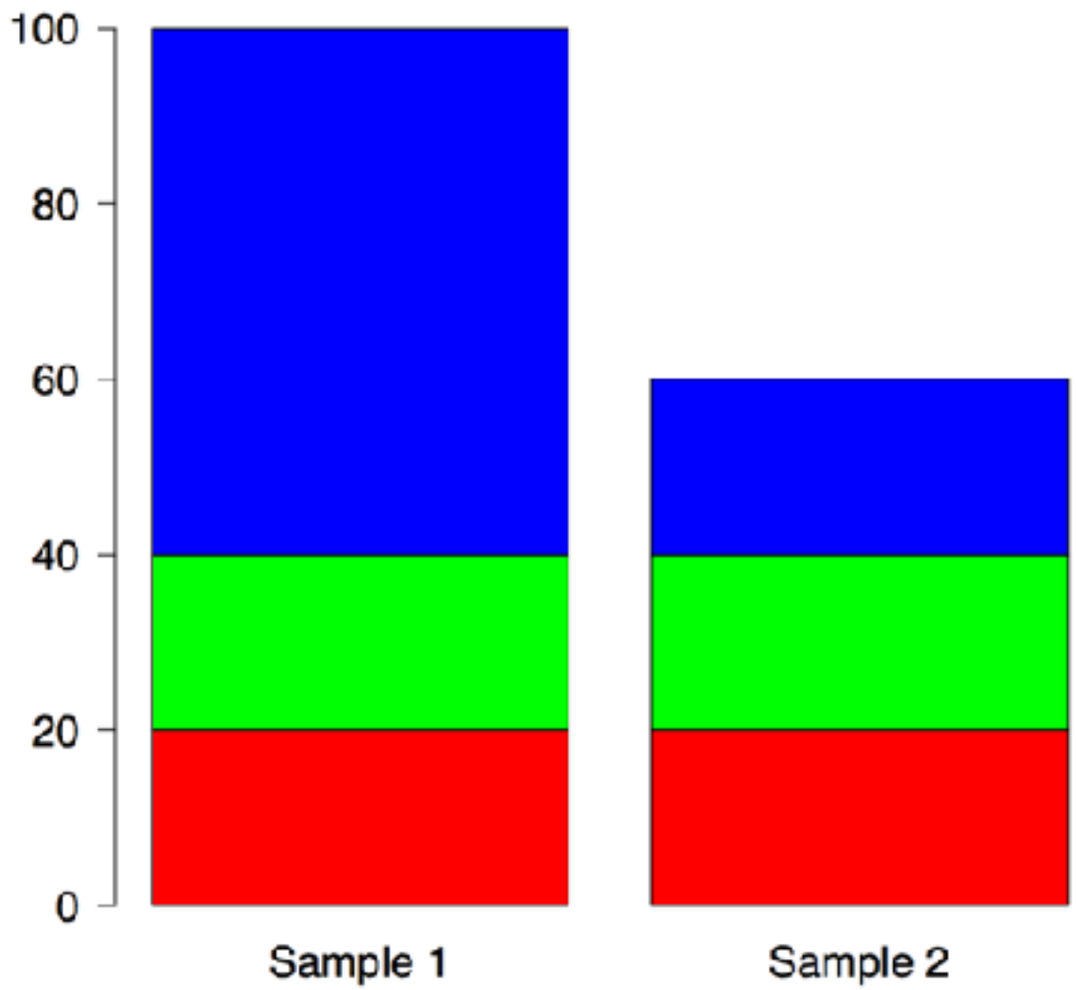




How to calculate normalization factors?

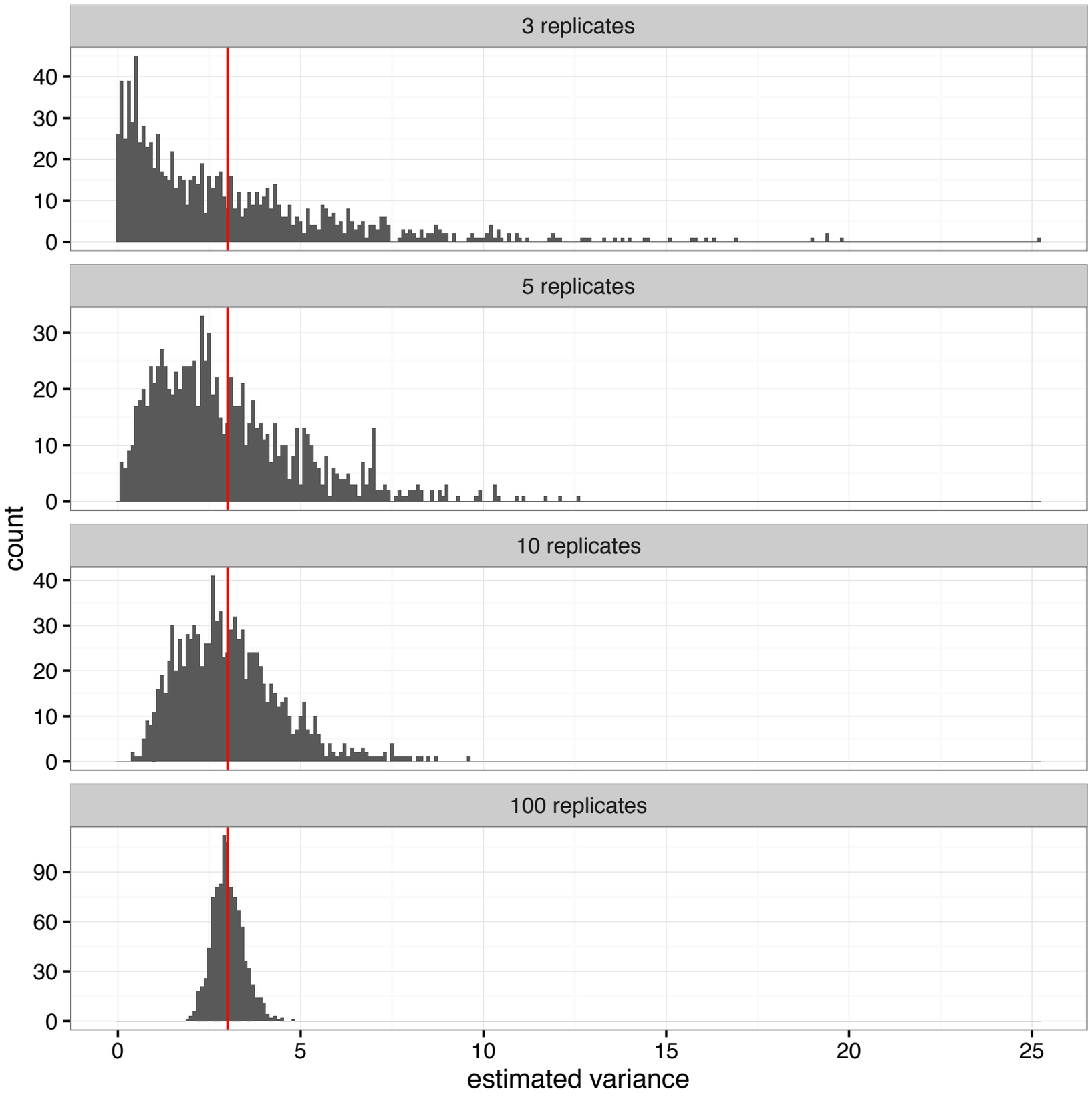
- Attempt 2: total count (library size) * compensation for differences in composition





Example:
estimate variance of normally distributed variable

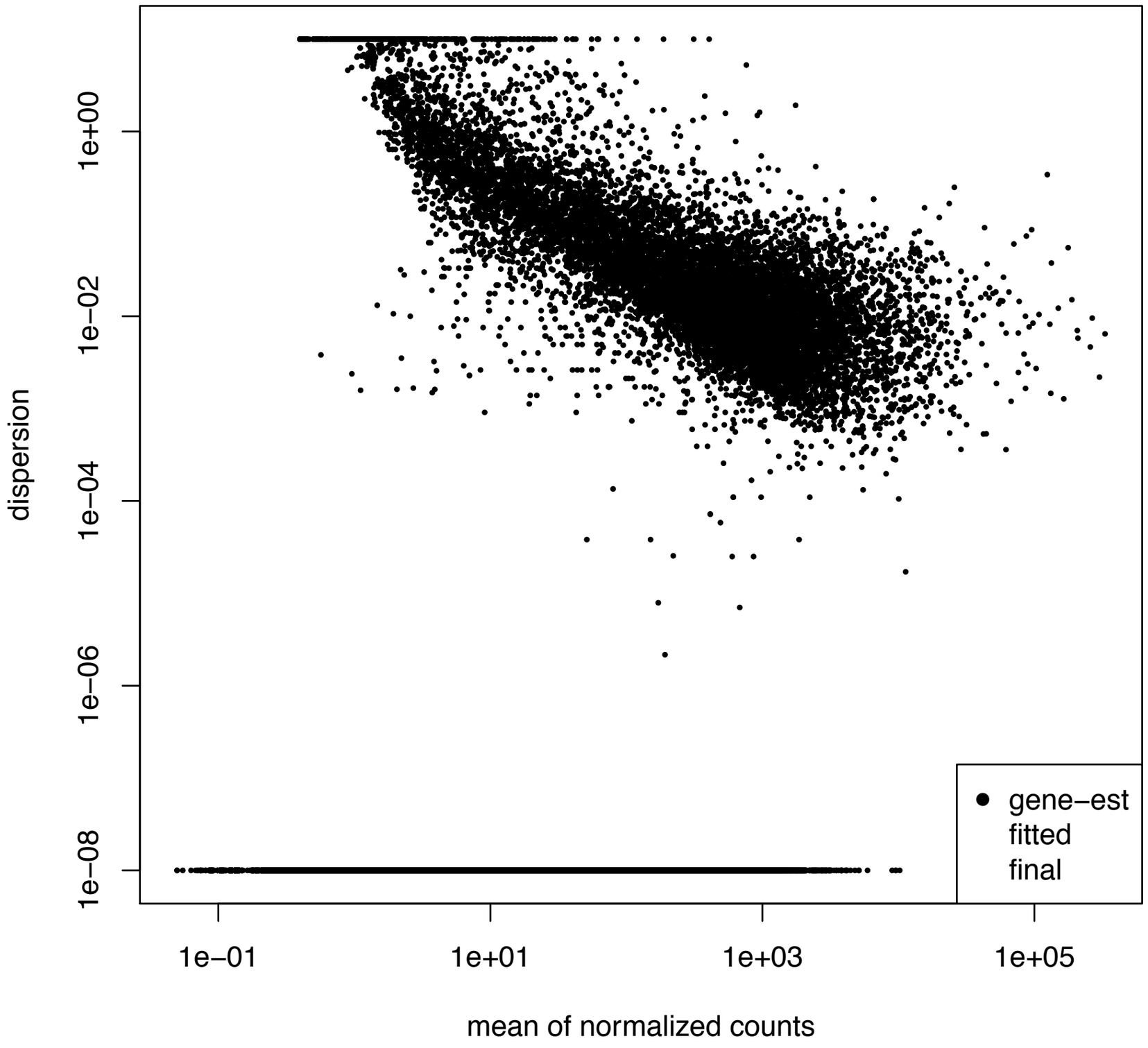
True value = 3



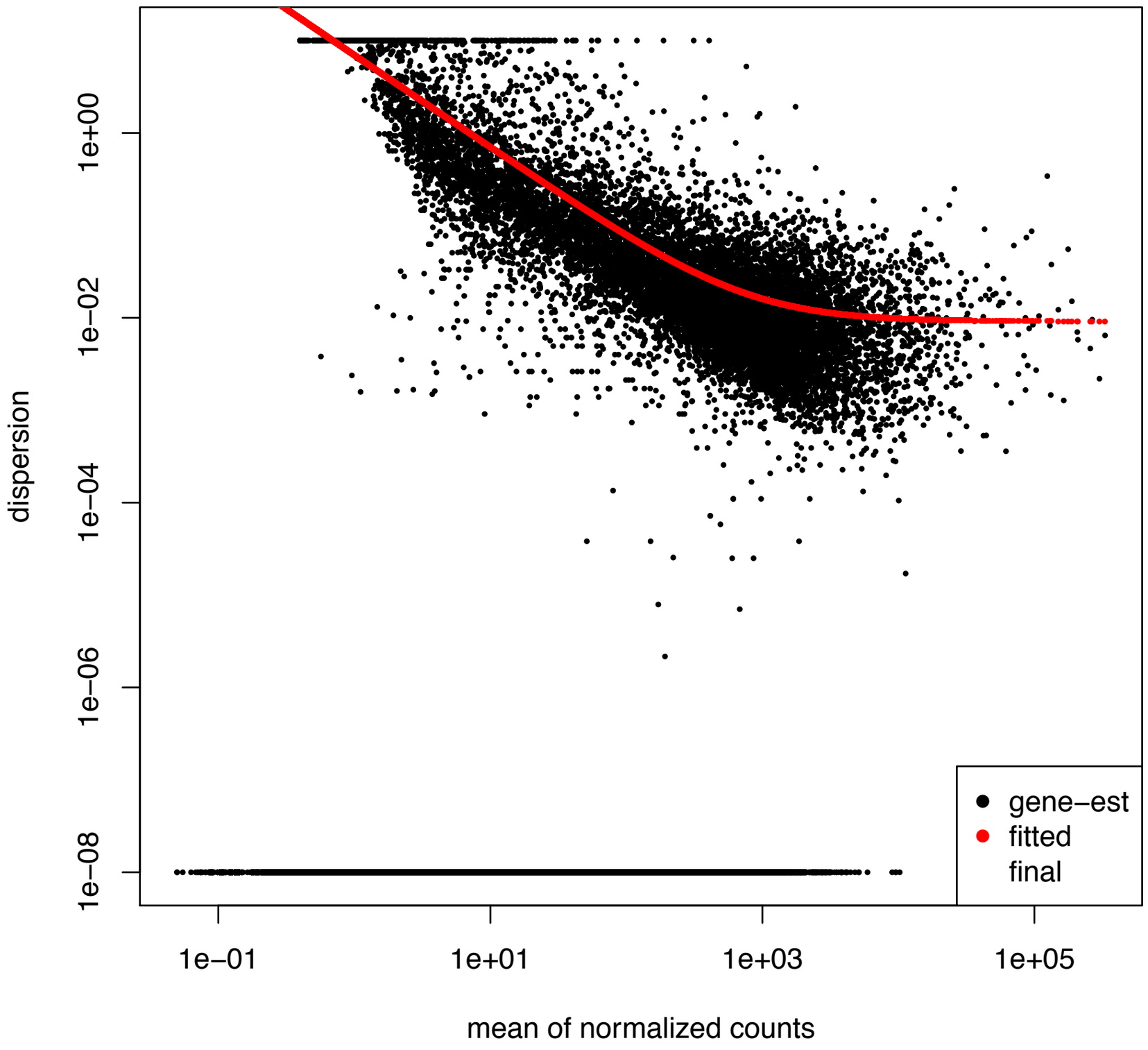
Shrinkage dispersion estimation

- 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)

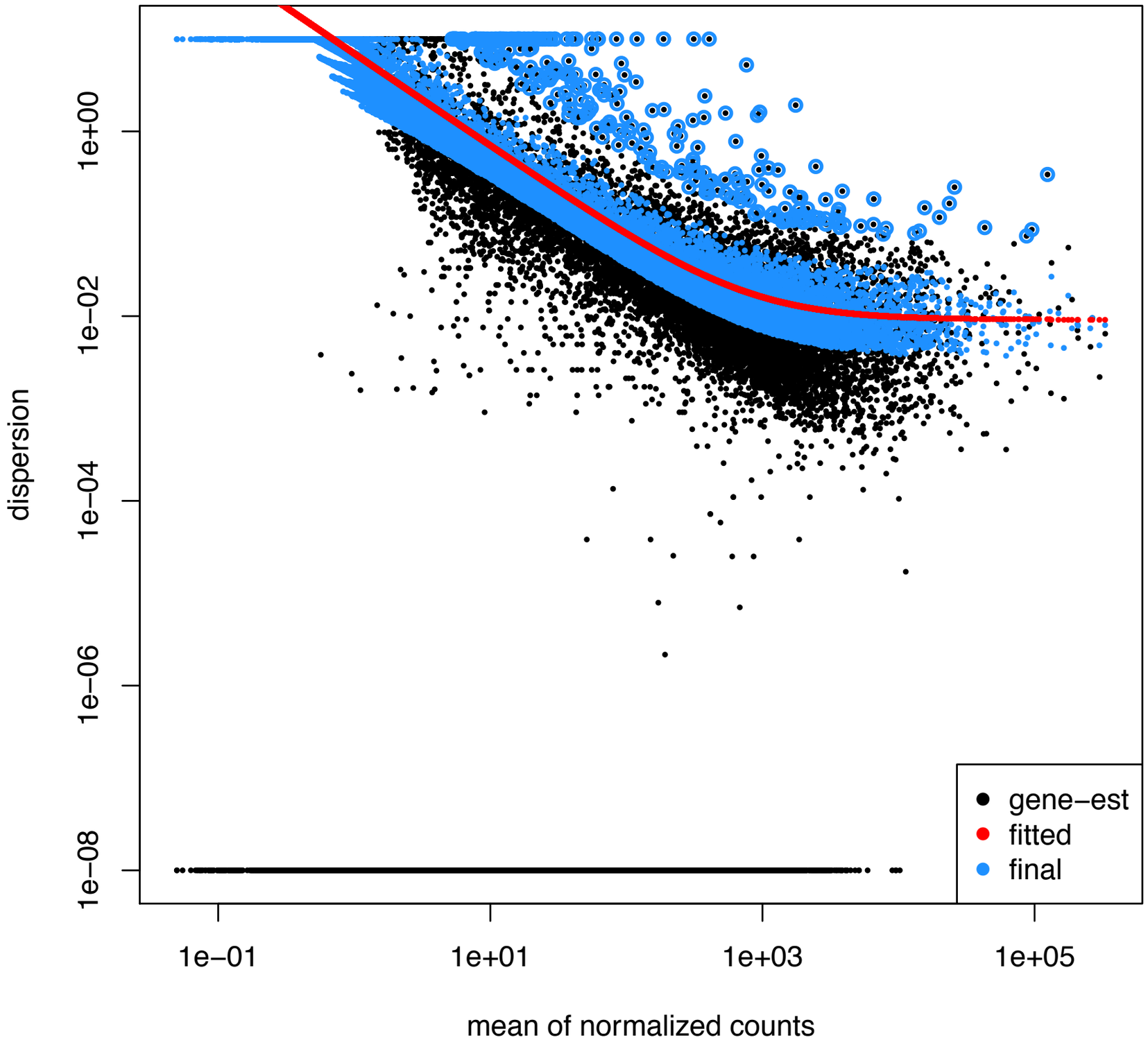
Shrinkage dispersion estimation



Shrinkage dispersion estimation



Shrinkage dispersion estimation



What is a p-value?

- The p-value is the probability of obtaining a test statistic *at least as extreme* as the one observed, *if the null hypothesis is true* (i.e., if there is no true signal in the data)

- Hence, if we get a p-value of **0.05**, it means that there is a **5%** chance of getting that extreme results even in the absence of real signal!

What does this mean for high-throughput studies?

- Assume that we perform 10,000 tests (one for each gene)…
- … and that there is no true signal at all in the data
- Then we would expect to get around 500 p-values below 0.05
- Relying solely on p-values would be misleading!

We need to change perspective

- Instead of limiting the false positive probability for *each individual test*, try to limit
  - the probability of obtaining *any* false positives (FWER)
  - the fraction of false positives among the significant genes (FDR)

Benjamini-Hochberg correction - controlling the FDR

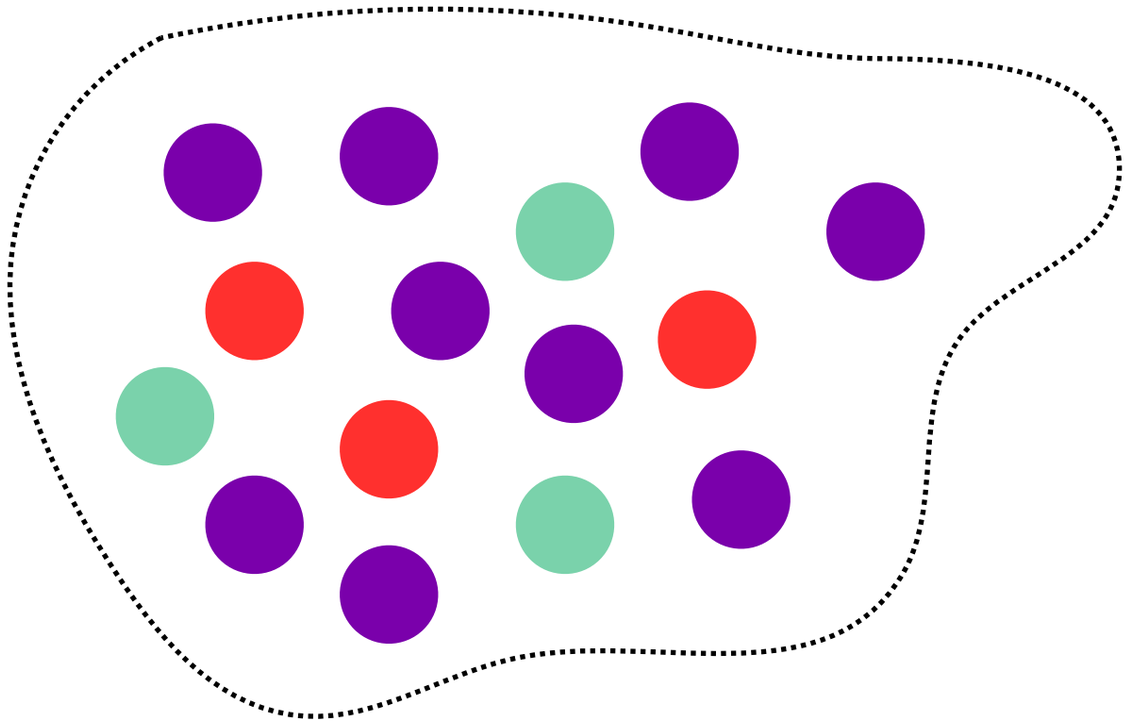
- Assume we are performing N tests
- Intuition:
  - for each threshold α, we can estimate the expected number of false discoveries by αN
  - Compare this to the actual number of discoveries at that threshold (N_α)
  - Choose a so that $\alpha N / N_\alpha \leq 0.05$ (or another desired threshold)

Interpreting the FDR

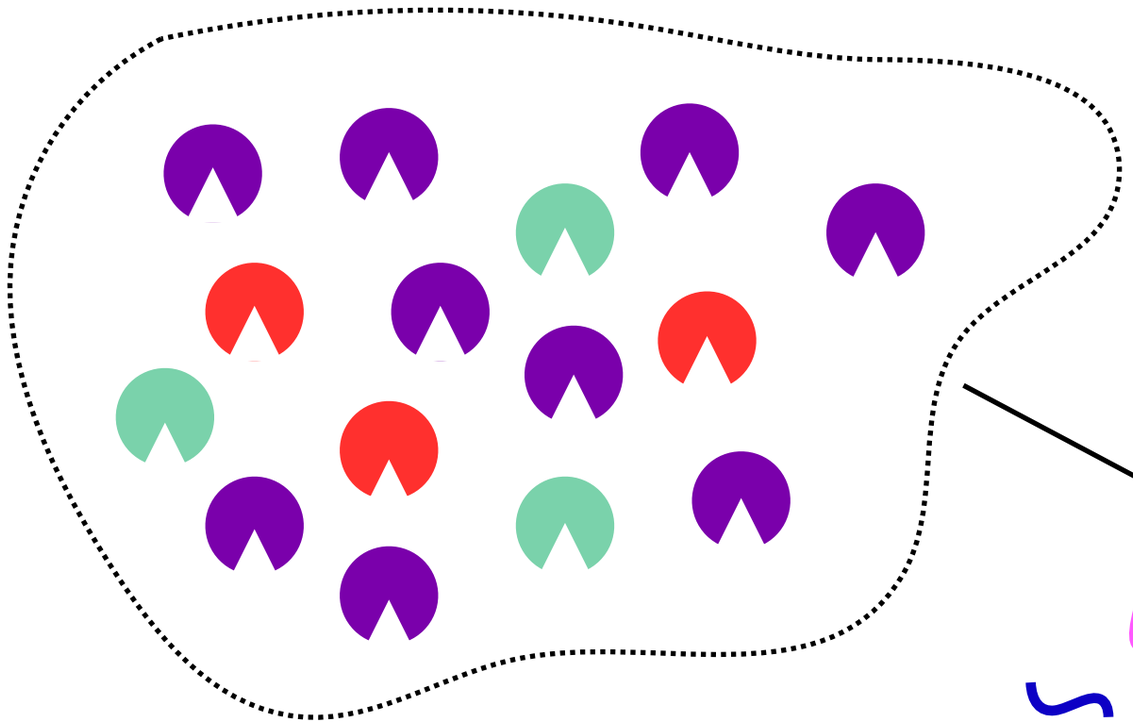
- The FDR is a measure for a *set* of genes
- In a set of genes with FDR = 0.05, approximately 5% can be expected to be false discoveries
- However, we don't know *which ones*! It could be the most significant!
- *q-values* are gene-wise significance measures ("adjusted p-values") - the smallest FDR we have to accept in order to call the gene significant

Single-cell RNA-seq

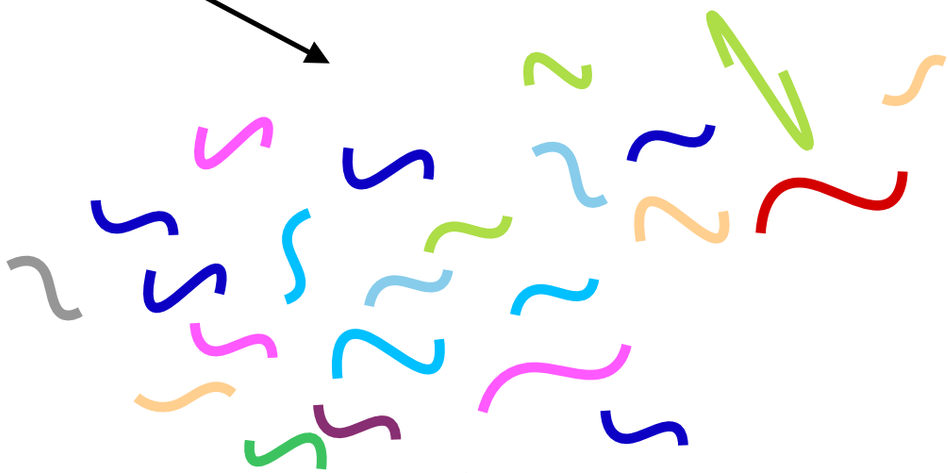
Bulk RNA-seq



Bulk RNA-seq



RNA extraction (polyA selection, ribosomal RNA depletion)



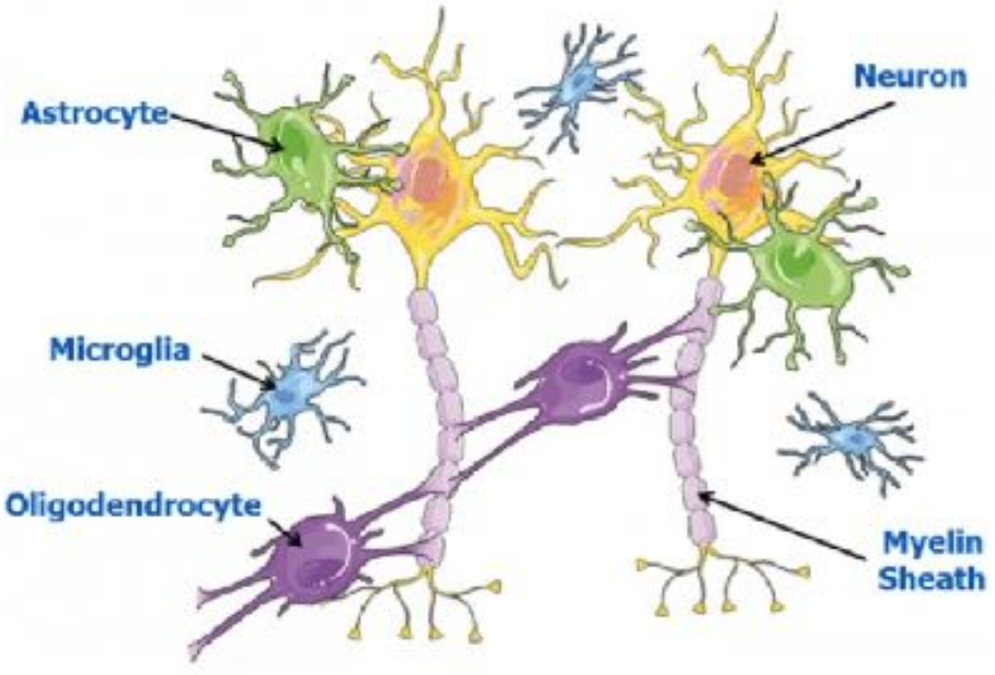
Fragmentation

cDNA synthesis Amplification

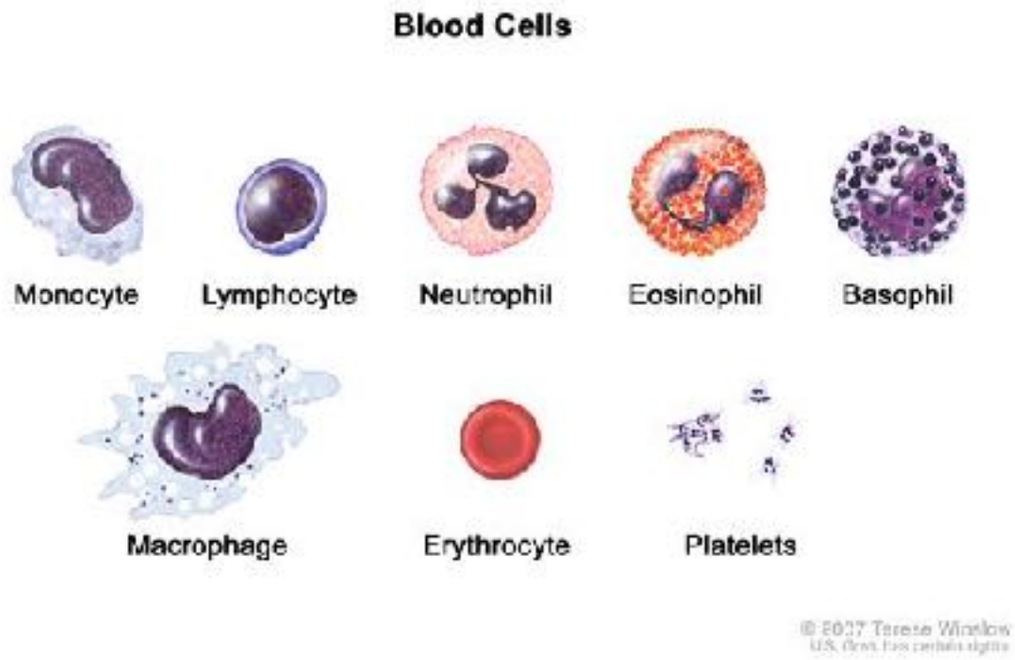
Sequencing Abundance quantification

gene	count
gene1	32
gene2	4
gene3	0
gene4	1928
...	

There are lots of different sorts of cells



- Wikipedia lists >200 cell types present in the human body
- Cell *type* often refers to more “permanent” aspects of a cell’s identity



- Cell *state* instead reflects aspects arising in more transient processes (differentiation, cell cycle, circadian rhythm)
- Samples are often heterogeneous in terms of cell types/states

Wagner et al., Nature Biotechnology 2016

https://www.cancer.gov/publications/dictionaries/cancer-terms/def/red-blood-cell, http://facts.net/brain/

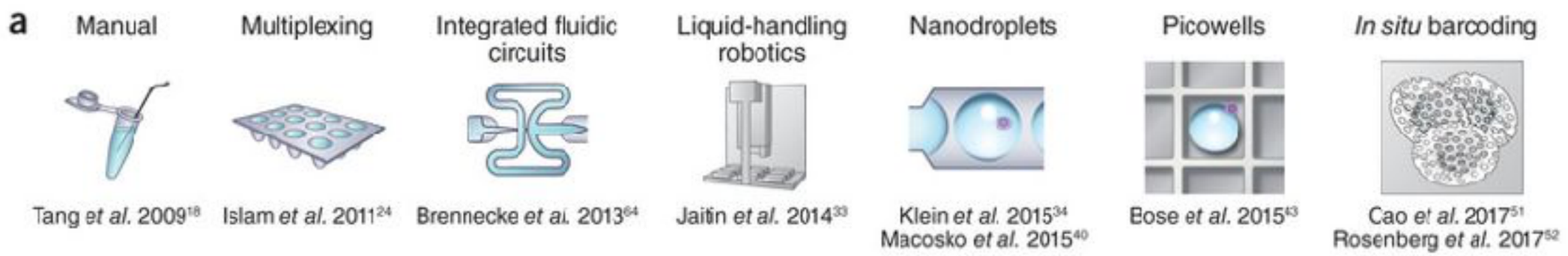
Opportunities with single-cell (RNA-seq) assays

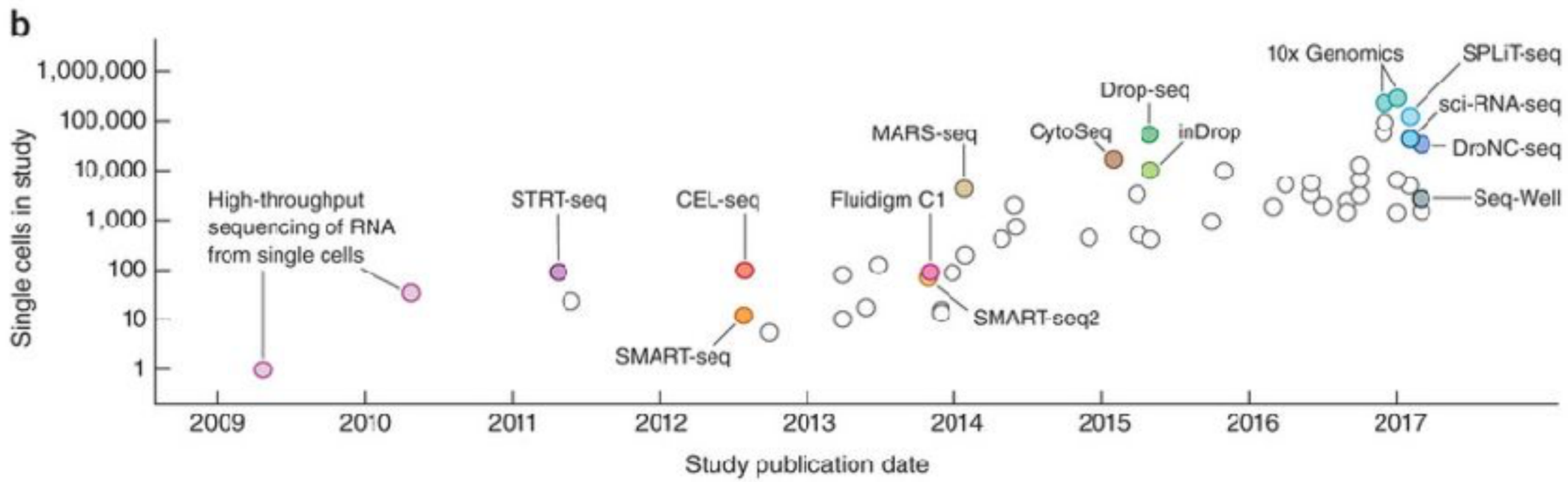
- Quantify abundance of individual cell types, and detect new ones
- Discriminate between differences in gene abundance due to changes in cell type composition, and differences due to changes in the expression levels within a given cell type
- Study the variability (more generally, the distribution) of a gene's expression among cells
- ...

scRNA-seq is not a single assay

- SmartSeq(2)
- CEL-Seq
- 10x Genomics (GemCode, Chromium)
- DropSeq
- STRT-Seq

- MARSseq
- QuartzSeq
- inDrop
- ...

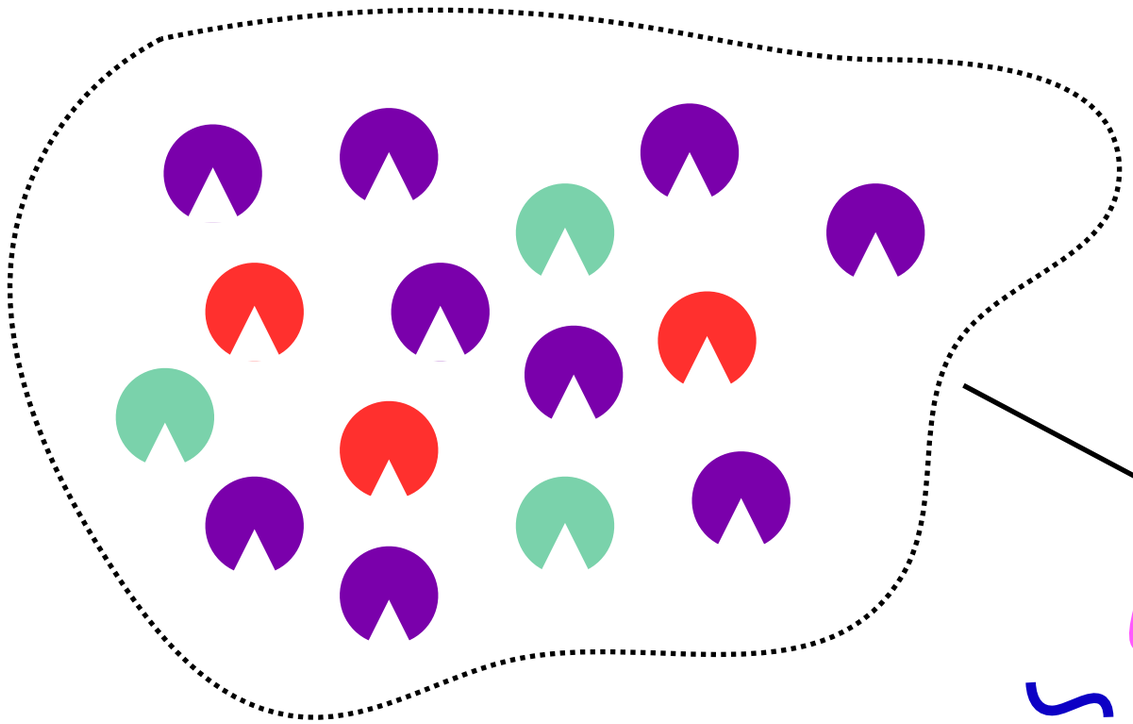




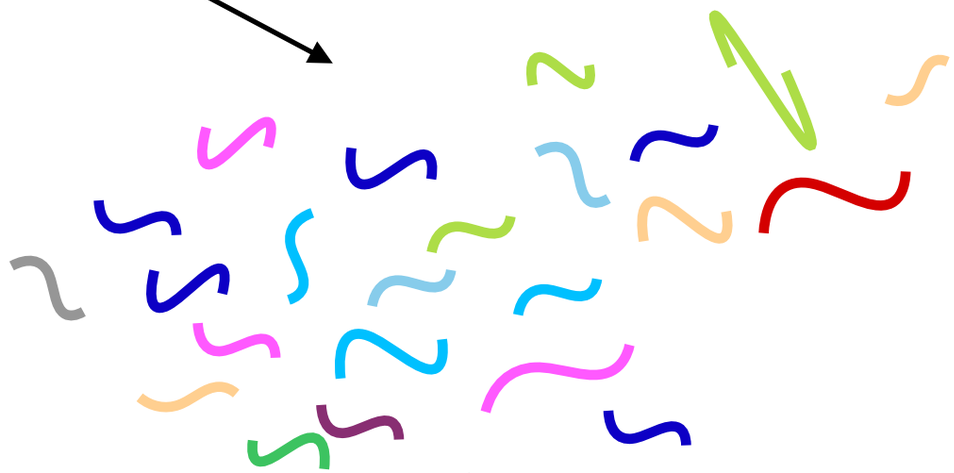
Svensson et al., Nature Protocols 2018

Long-read RNA-seq

“Bulk” RNA-seq



RNA extraction (polyA selection, ribosomal RNA depletion)



Fragmentation

cDNA synthesis Amplification

Sequencing Abundance quantification

gene	count
gene1	32
gene2	4
gene3	0
gene4	1928
...	

Why long reads?

- With Illumina sequencing, we need to fragment and amplify the (c)DNA before sequencing

- Application areas for long-read technologies:
  - genome assembly
  - metagenomics
  - transcriptomics
  - DNA/RNA modification detection

- Advantages for RNA-seq: identification of splice isoforms, transcriptome annotation

http://bioconductor.org/

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home | Install | Help | Developers | About

About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 1211 software packages, and an active user community. Bioconductor is also available as an AMI (Amazon Machine Image) and a series of Docker images.

News

- Bioconductor 3.3 is available.
- Bioconductor F1000 Research Channel launched.
- Orchestrating high-throughput genomic analysis with *Bioconductor* (abstract) and other recent literature.
- Read our latest newsletter and course material.
- Use the support site to get help installing, learning and using Bioconductor.

Install »

Get started with *Bioconductor*

- Install *Bioconductor*
- Explore packages
- Get support
- Latest newsletter
- Follow us on twitter
- Install R

Learn »

Master *Bioconductor* tools

- Courses
- Support site
- Package vignettes
- Literature citations
- Common work flows
- FAQ
- Community resources
- Videos

Use »

Create bioinformatic solutions with *Bioconductor*

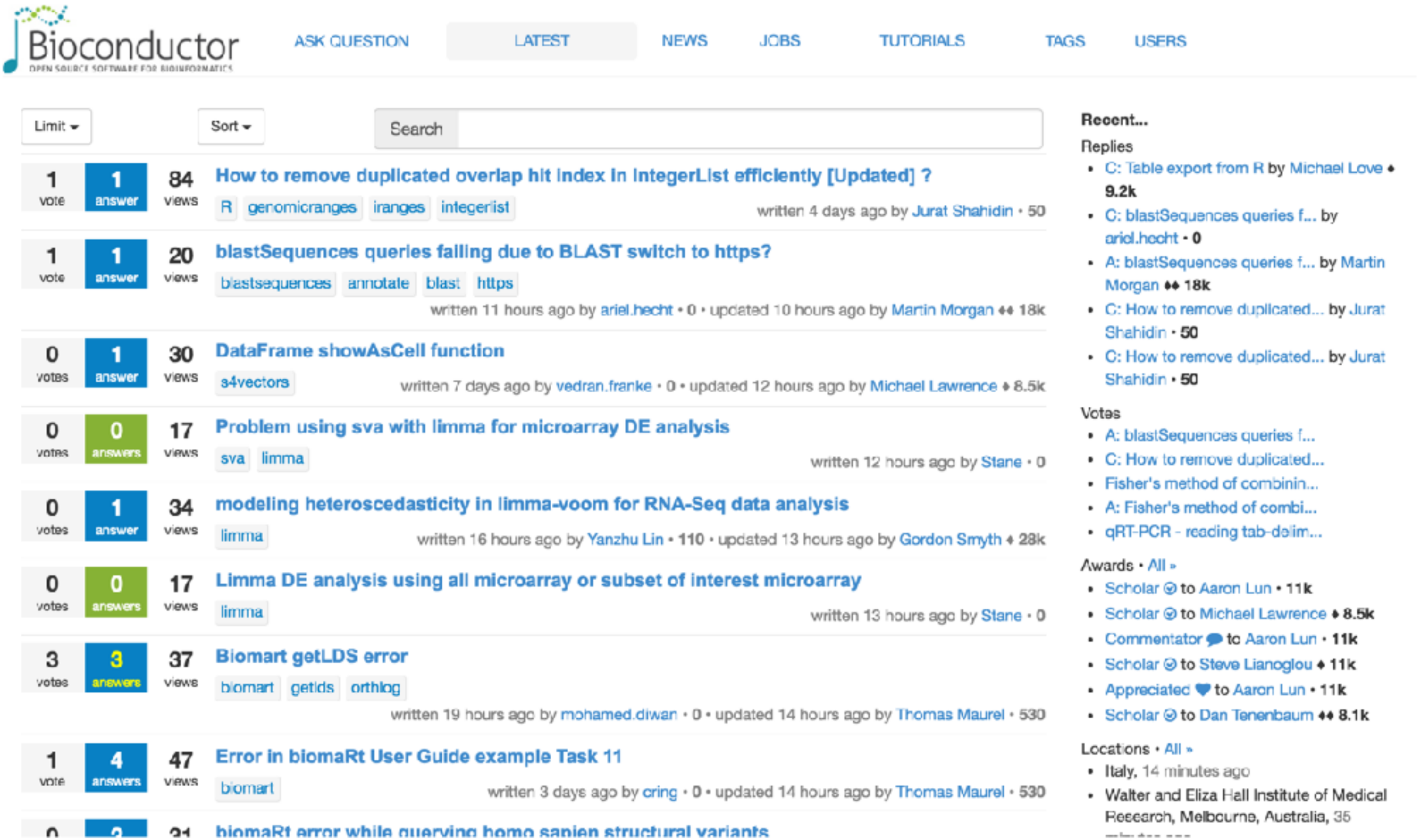
- Software, Annotation, and Experiment packages
- Amazon Machine Image
- Latest release annoucement
- Support site

Develop »

Contribute to *Bioconductor*

- Developer resources
- Use Bioc 'devel'
- 'Devel' Software, Annotation and Experiment packages
- Package guidelines
- New package submission
- Build reports

https://support.bioconductor.org/



https://bioconductor.org/packages/release/bioc/html/Rsubread.html

Rsubread

Subread sequence alignment for R

Bioconductor version: Release (3.3)

Provides powerful and easy-to-use tools for analyzing next-gen sequencing read data. Includes quality assessment of sequence reads, read alignment, read summarization, exon-exon junction detection, fusion detection, detection of short and long indels, absolute expression calling and SNP calling. Can be used with reads generated from any of the major sequencing platforms including Illumina GA/HiSeq/MiSeq, Roche GS-FLX, ABI SOLiD and LifeTech Ion PGM/Proton sequencers.

Author: Wei Shi and Yang Liao with contributions from Jenny Zhiyin Dai and Timothy Triche, Jr.

Maintainer: Wei Shi <shi at wehi.edu.au>

Citation (from within R, enter `citation("Rsubread")`):

Liao Y, Smyth GK and Shi W (2013). "The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote." *Nucleic Acids Research*, **41**, pp. e108.

Installation

To install this package, start R and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("Rsubread")
```

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("Rsubread")
```

PDF	R Script	Rsubread Vignette
PDF		Reference Manual
Text		NEWS

Some suggestions for further reading

- 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.: Accurate, fast, and model-aware transcript expression quantification with Salmon. bioRxiv http://dx.doi.org/10.1101/021592 (2015) - **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**
- Goeman & Bühlmann: Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 23(8): 980-987 (2007) - **competitive vs self-contained gene set tests**