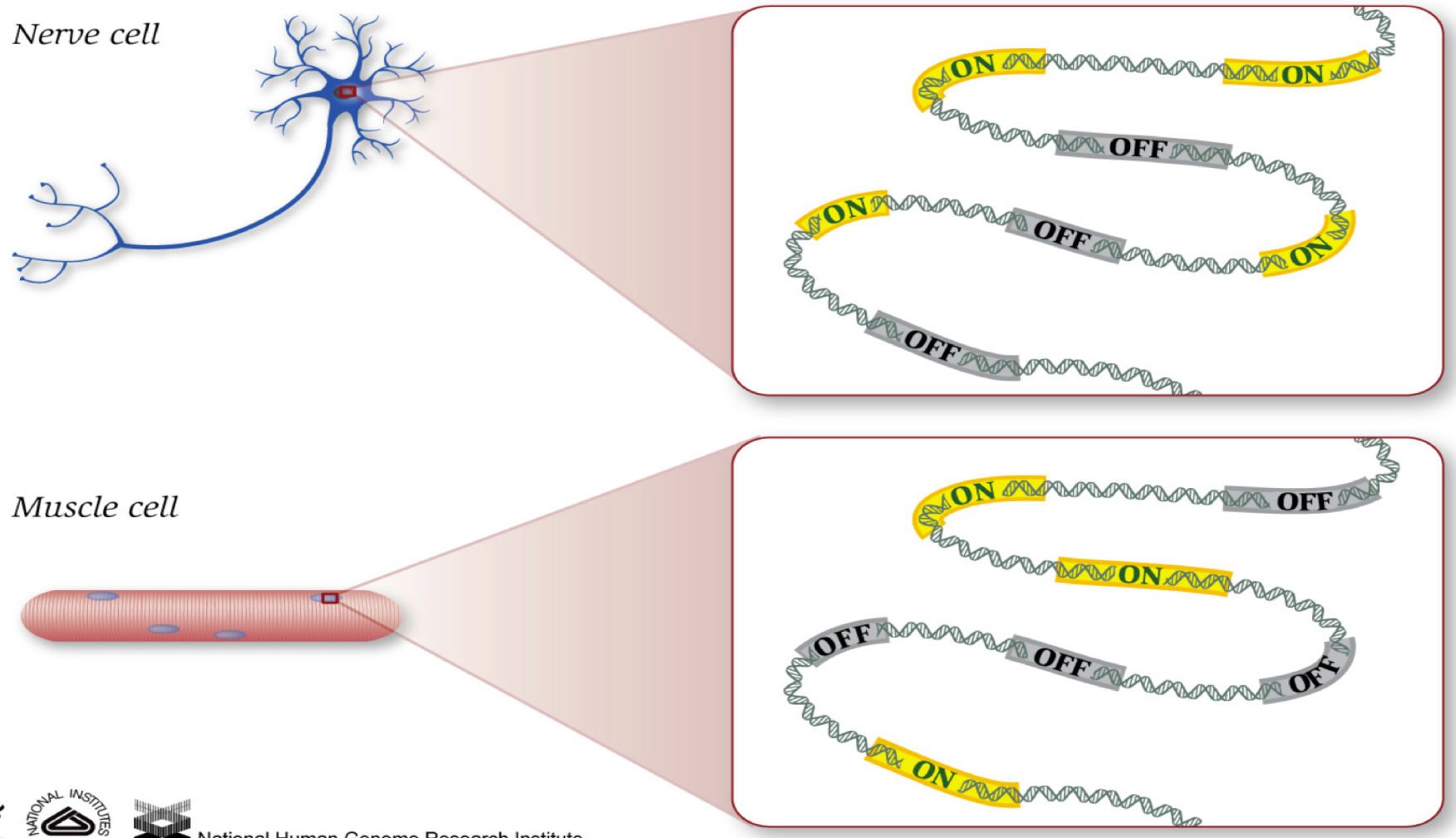
Introduction to RNA-seq Analysis

Mudassar Iqbal

WT ICD Informatics Bootcamp

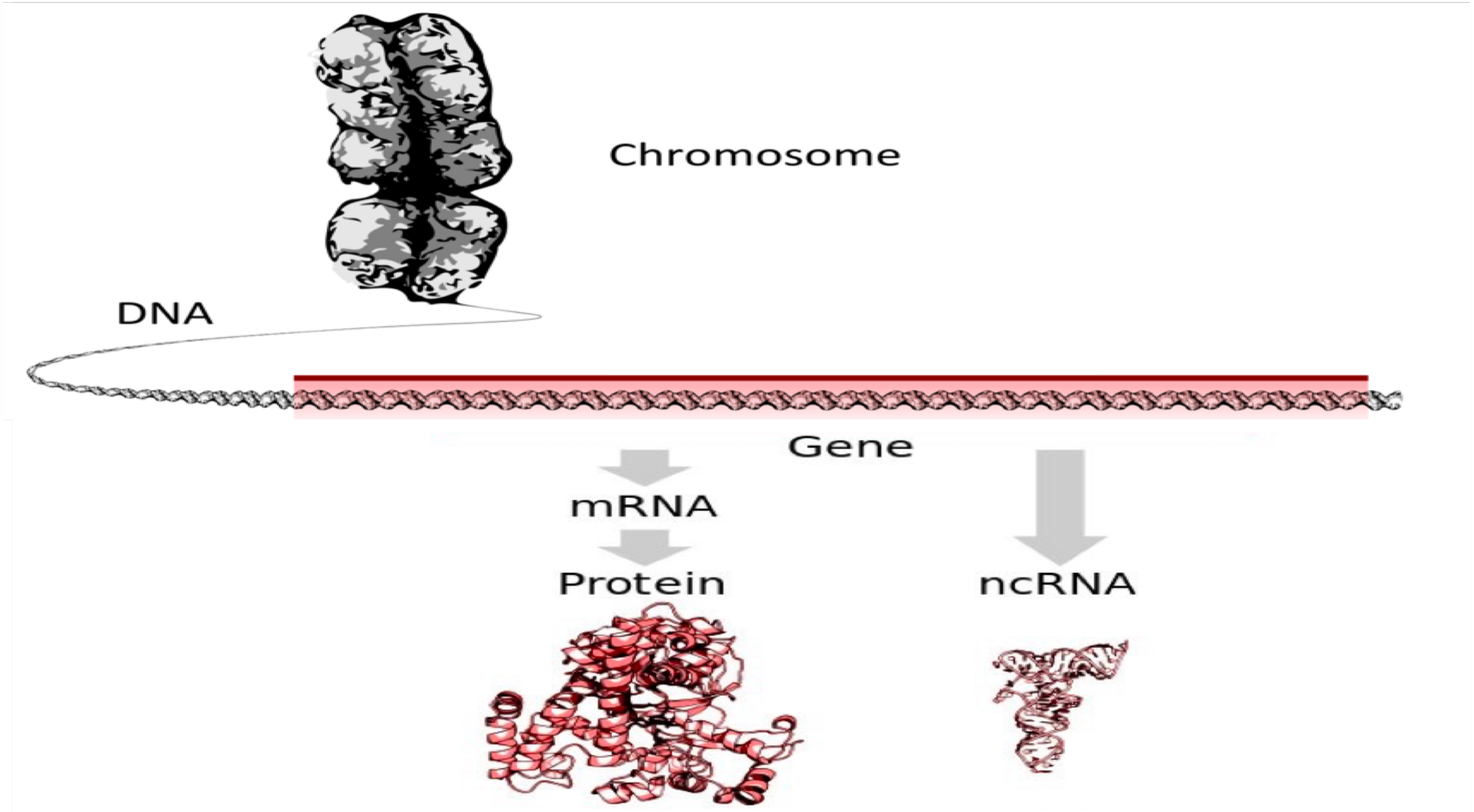
9^{th} – 13^{th} November 2020

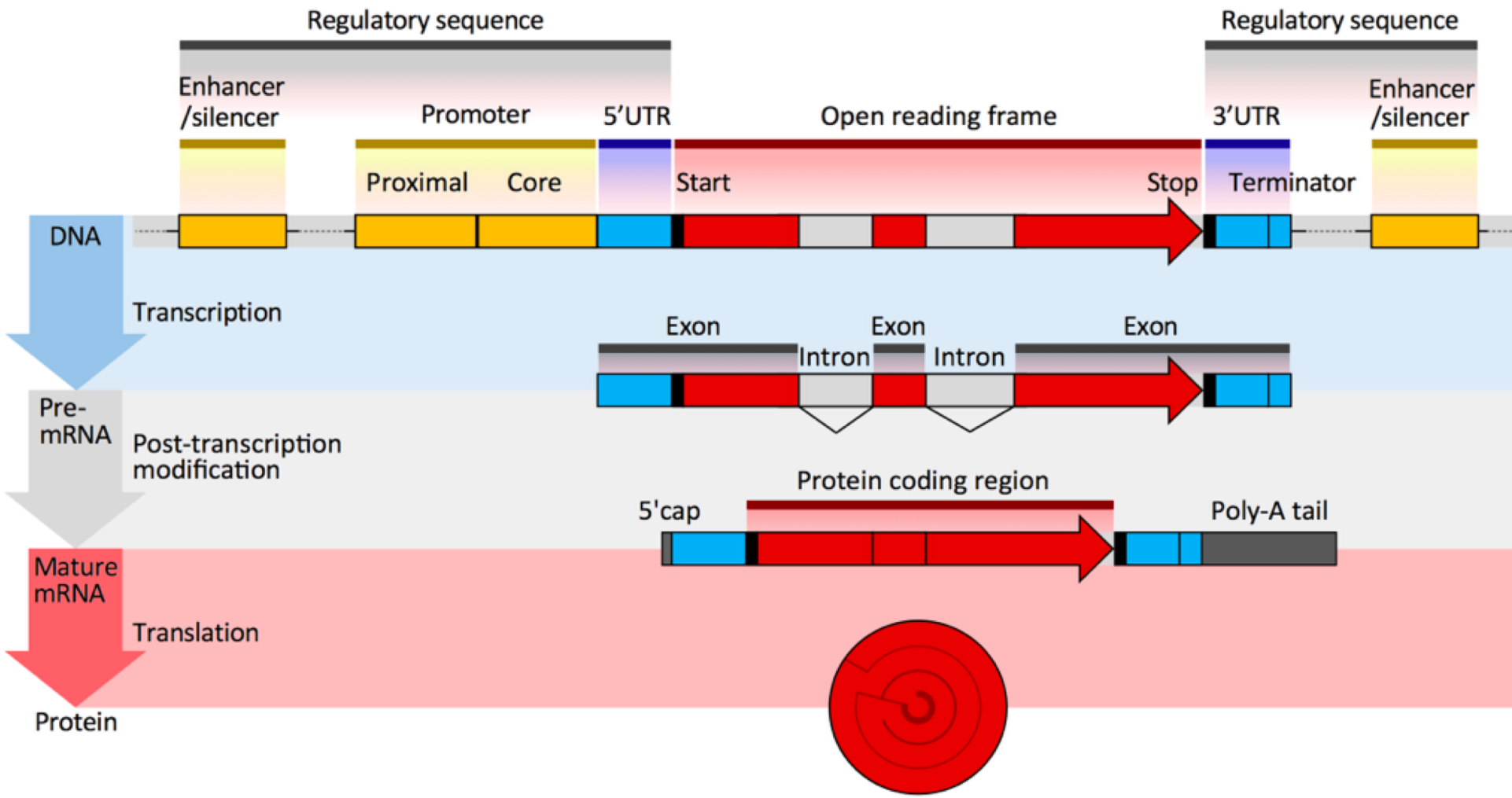
Gene expression: same DNA, different function



National Human Genome Research Institute

Transcription & Translation, etc.

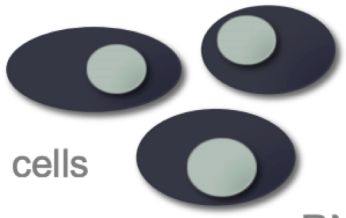




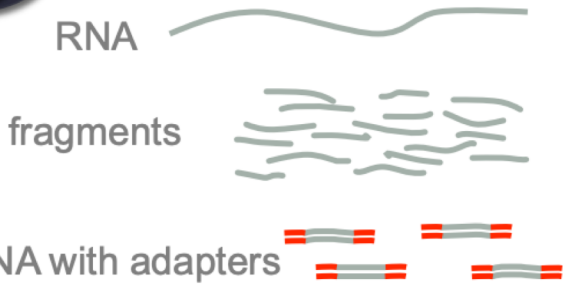
Trancriptomics

- The transcriptome is defined as a collection of all the transcript readouts present in a cell.
- RNA-seq data can be used to explore and/or quantify the transcriptome of an organism, which can be utilized for the following types of experiments:
  - **Differential Gene Expression**: *quantitative* evaluation and comparison of transcript levels
  - **Transcriptome assembly**: building the profile of transcribed regions of the genome, a *qualitative* evaluation.
  - Building **gene regulatory networks**

RNA-seq workflow overview



Total RNA extraction



- mRNA enrichment
- fragmentation
- cDNA library

Sequencing

- cluster generation
- sequencing by synthesis
- image acquisition

Bioinformatics

https://chagall.med.cornell.edu/RNASEQcourse/Slides_Nov2019_Day1.pdf

RNA-seq platforms

Illumina almost has a *de facto* monopoly on high-throughput sequencing

Platform (Manufacturer)	Commercial release	Typical read length	Maximum throughput per run	Single read accuracy	RNA-Seq runs deposited in the NCBI SRA (Oct 2016) [74]
454 (Roche, Basel, Switzerland)	2005	700 bp	0.7 Gbp	99.9%	3548
Illumina (Illumina, San Diego, CA, USA)	2006	50–300 bp	900 Gbp	99.9%	362903
SOLiD (Thermo Fisher Scientific, Waltham, MA, USA)	2008	50 bp	320 Gbp	99.9%	7032
Ion Torrent (Thermo Fisher Scientific, Waltham, MA, USA)	2010	400 bp	30 Gbp	98%	1953
PacBio (Pacbio, Menlo Park, CA, USA)	2011	10,000 bp	2 Gbp	87%	160

NCBI, National Center for Biotechnology Information; SRA, Sequence Read Archive; RNA-Seq, RNA sequencing.

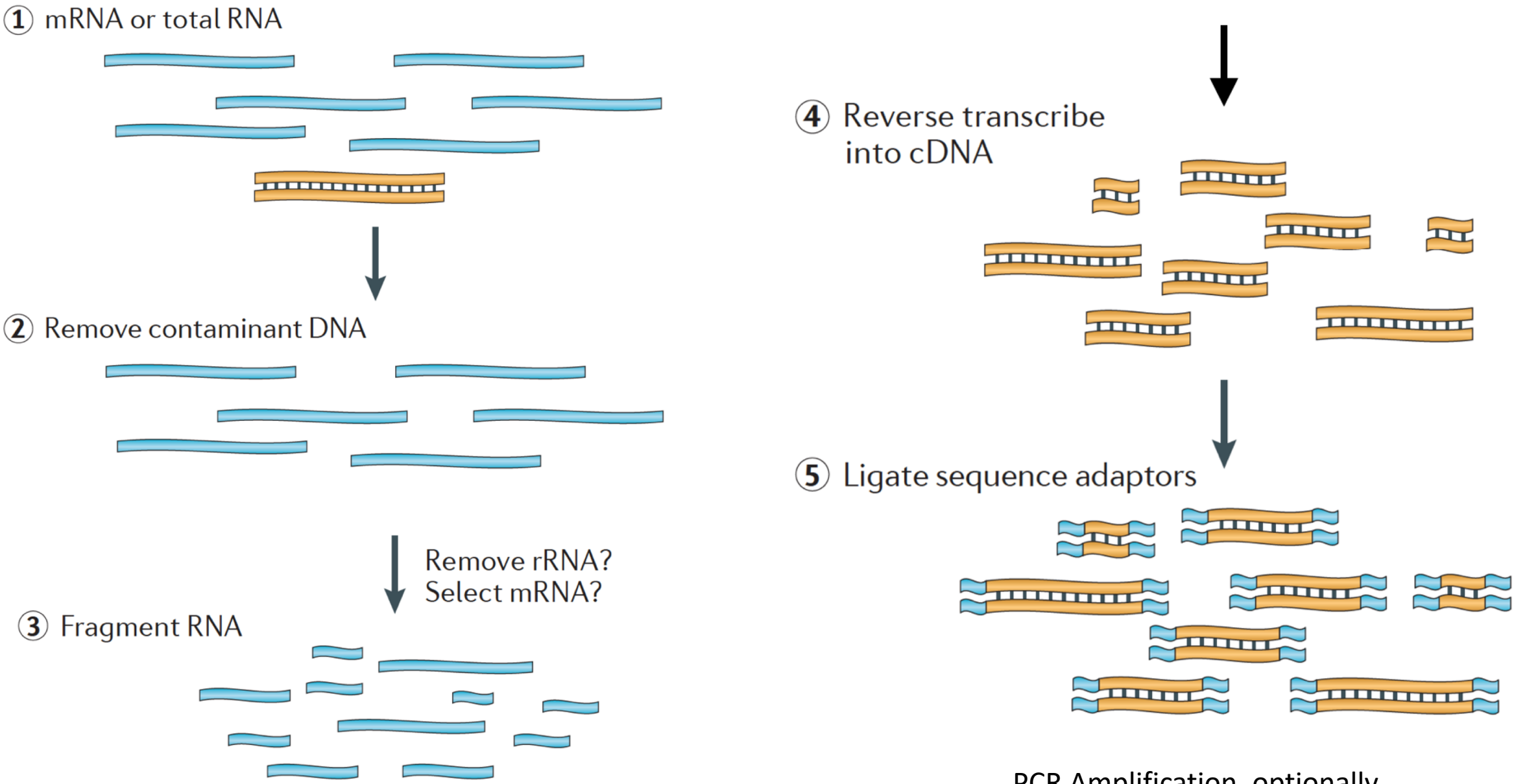
https://doi.org/10.1371/journal.pcbi.1005457.t002

currently, all mainstream RNA-seq solutions rely on copying RNA into cDNA prior to sequencing

direct sequencing of RNA (Nanopore) is in its infancy, but will be useful to detect modified bases and avoid biases from the amplification steps

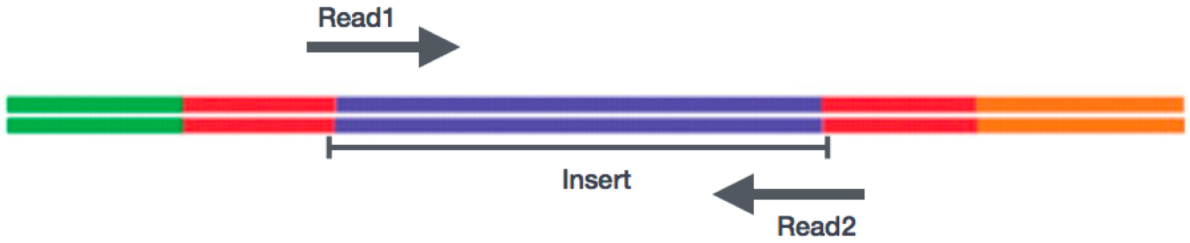
Lowe et al. (2017) PLoS Comp Bio. doi: 10.1371/journal.pcbi.1005457

Illumina library preparation



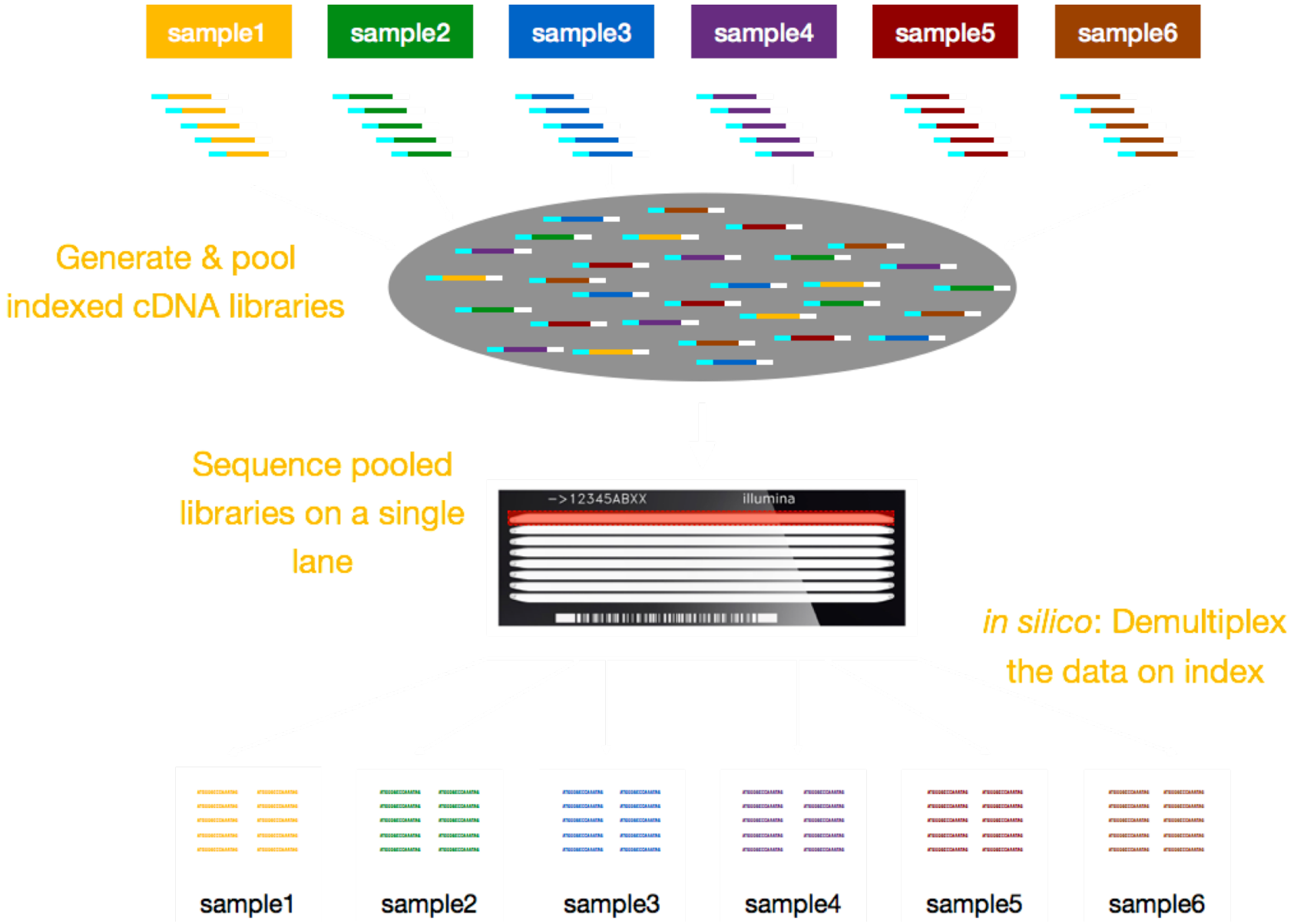
PCR Amplification, optionally

Illumina Sequencing



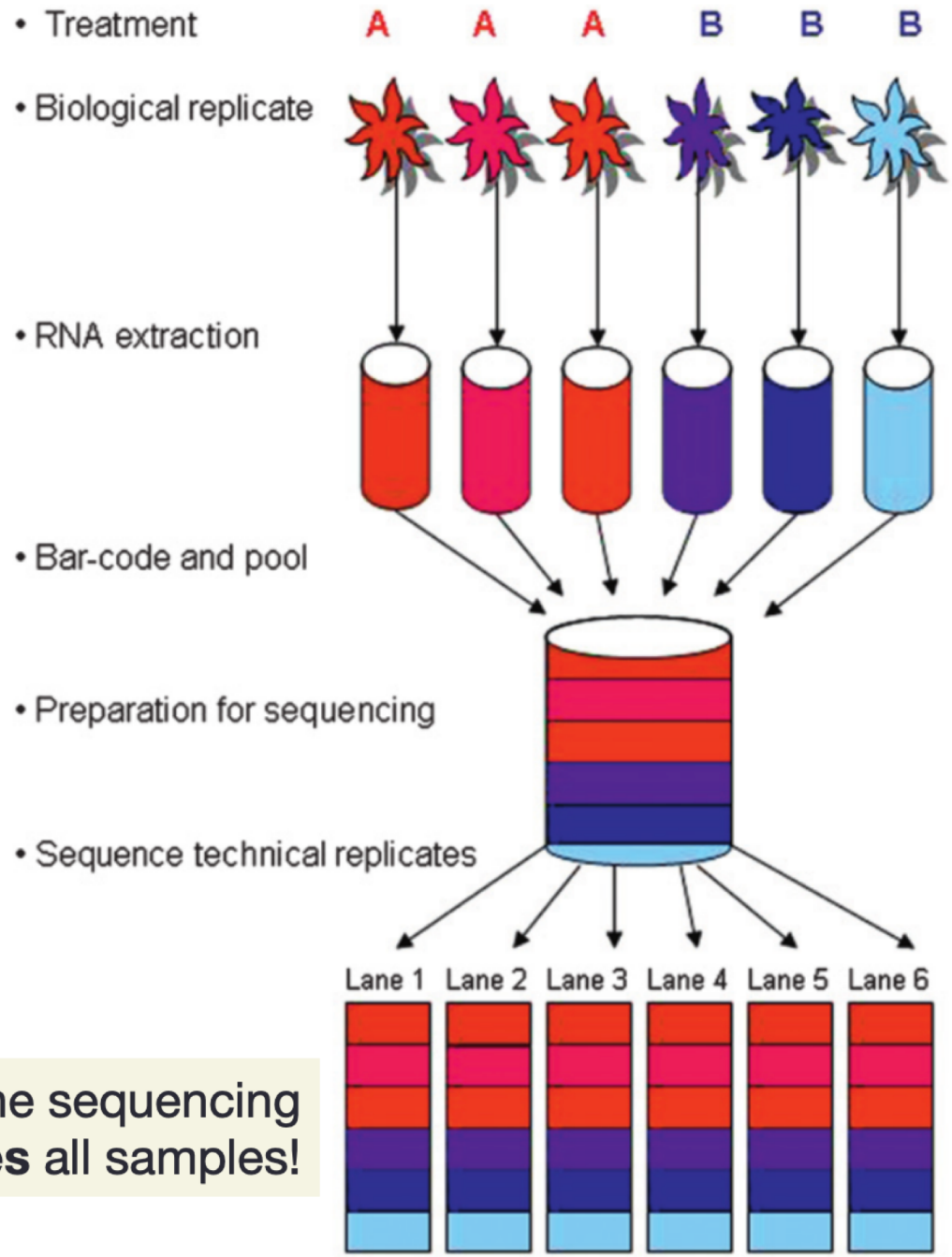
- SE - Single end dataset => Only Read1
- PE - Paired-end dataset => Read1 + Read2
  - can be 2 separate FastQ files or just one with interleaved pairs

Multiplexing



Typical RNA-seq set-up

- keep the **technical nuisance** factors (harvest date, RNA extraction kit, sequencing date…) to a **minimum**
- cover only as much of the **biological variation as needed** (but keep possible limitations for the final conclusions in mind)



Make sure the sequencing core **multiplexes** all samples!

Experimental planning considerations - Replicates

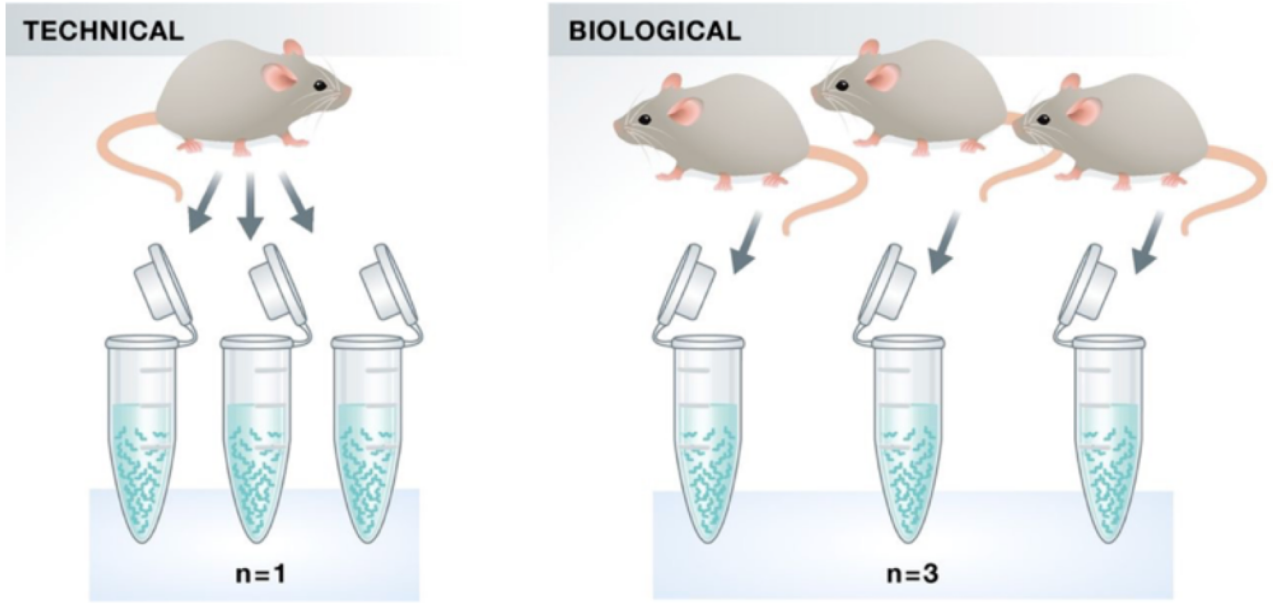


*Image credit: Klaus B., EMBO J (2015) **34**: 2727-2730*

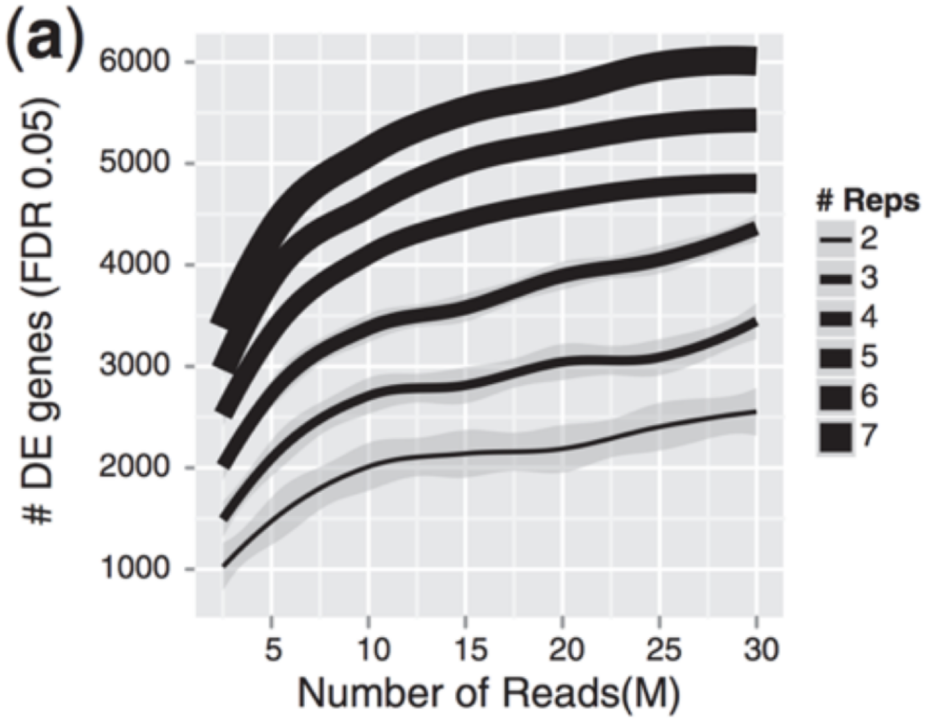
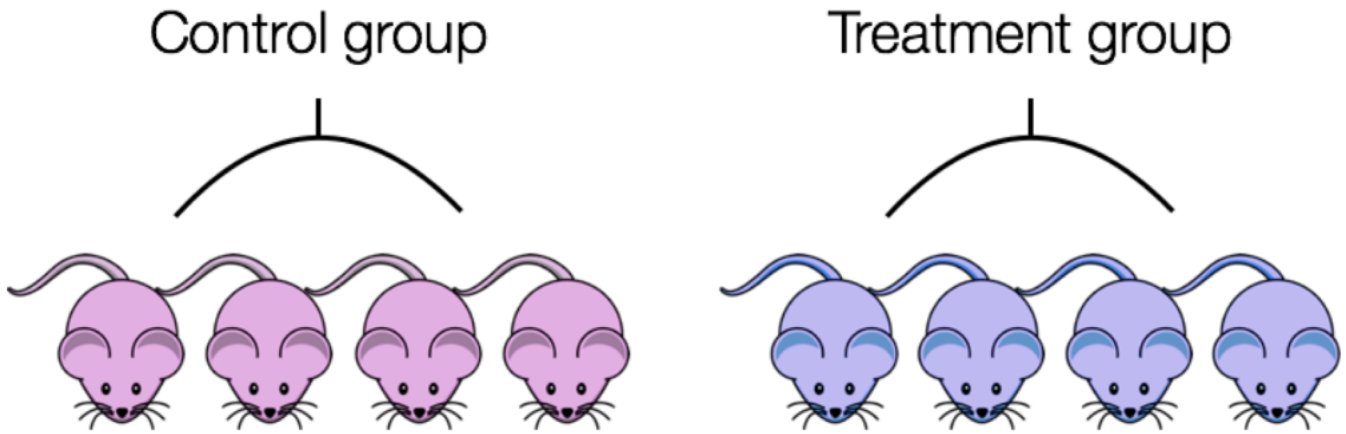
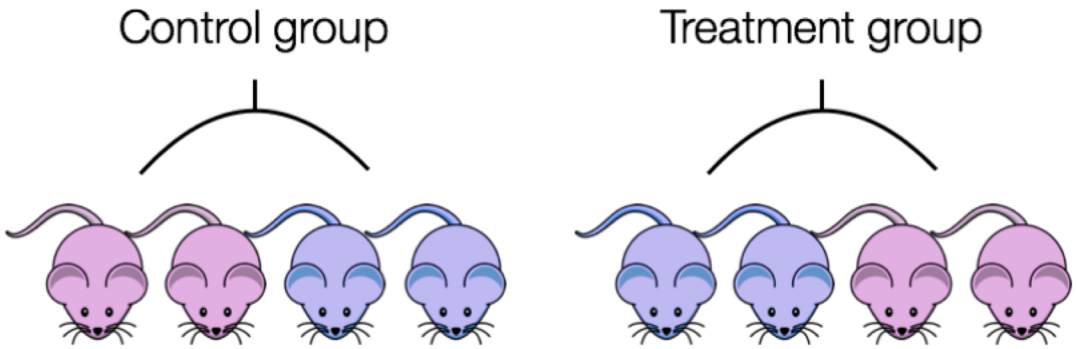


*Image credit: Liu, Y., et al., Bioinformatics (2014) **30**(3): 301–304*

Experimental planning considerations- Confounding



Avoid



Recommended

https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/experimental_planning_considerations.html

Bioinformatics workflow of RNA-seq analysis

Images
.tif

Base calling & demultiplexing
Bustard/RTA/OLB, CASAVA

FASTQC ► Raw reads
.fastq

Mapping
STAR

Aligned reads
.sam/.bam

Counting
featureCounts

Read count table
.txt

Normalizing
DESeq2, edgeR

Normalized read count table
.Robj

DE test & multiple testing correction
DESeq2, edgeR, limma

List of fold changes & statistical values
.Robj, .txt

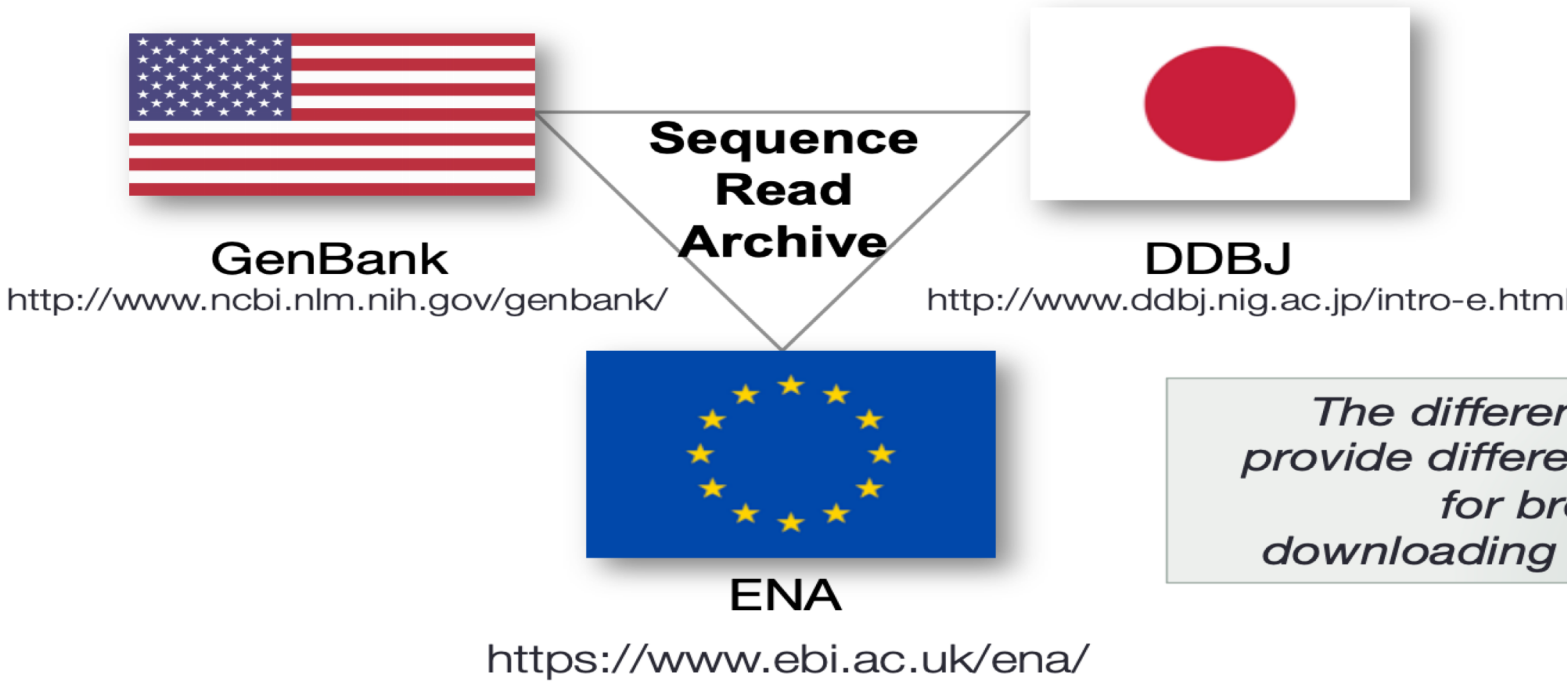
Filtering
Customized scripts

Downstream analyses on DE genes

https://chagall.med.cornell.edu/RNASEQcourse/Slides_Nov2019_Day1.pdf

Where are all the reads?

The sequence read archive (**SRA**) is the main repository for publicly available DNA and RNA sequencing data of which 3 instances are maintained world-wide.



The different mirrors provide different routes for browsing & downloading the data.

Detailed information about the SRA: O'Sullivan et al. (2018) Managing Sequence Data. In *Bioinformatics: Data, Sequence Analysis, and Evolution*. doi: 10.1007/978-1-4939-6622-6_4

Public RNA-seq resources

Collection	Number of samples/ libraries	Reference
TCGA	12,000	Cancer Genome Atlas Research Network. 2013. PMID: 24071849
GTEX	11,000	Carither et al., 2015. PMID: 26484571
Human Protein Atlas	8,00	Uhlen et al. 2015. PMID: 25613900
ENCODE	2,300	David et al. 2018. PMID: 29126249
GEUVADIS	1,100	Lappalainen et al. 2013. PMID: 24037378
Cancer Cell Line Encyclopedia	650	Barretina et al. 2012. PMID: 22460905
Leucegene (AML focus)	550	Lavalée et al. 2018. PMID: 29550835

The **recount** resource offers **processed read counts** of >2,000 different studies!
`https://jhubiostatistics.shinyapps.io/recount/`

Unmapped Reads: FASTQ file format

= FASTA + **q**uality scores

1 read ⇔ 4 lines!

```
1 @ERR459145.1 DHKW5DQ1:219:D0PT7ACXX:2:1101:1590:2149/1
2 GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC
3 +
4 @7<DBADDDBH?DHHI@DH>HHHEGHIIIGGIFFGIBFAAGAFHA’5?B@D
```

1. @Read ID and sequencing run information
2. sequence
3. + (additional description possible)
4. quality scores

Quality control of raw reads: FastQC

http://www.bioinformatics.babraham.ac.uk/projects/fastqc

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

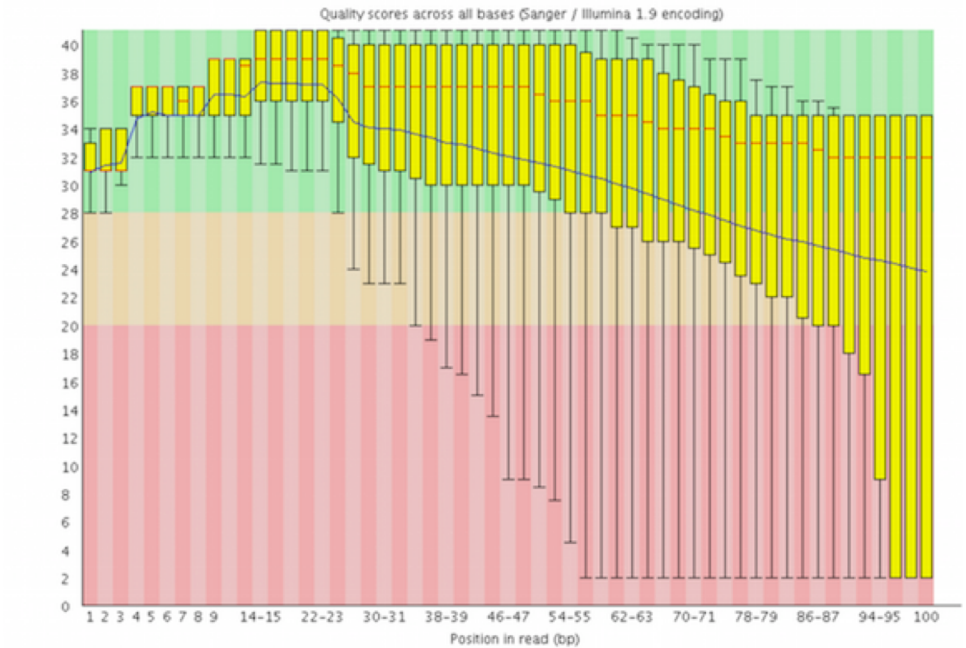
not specific for RNA-seq data!

Two basic questions of QC

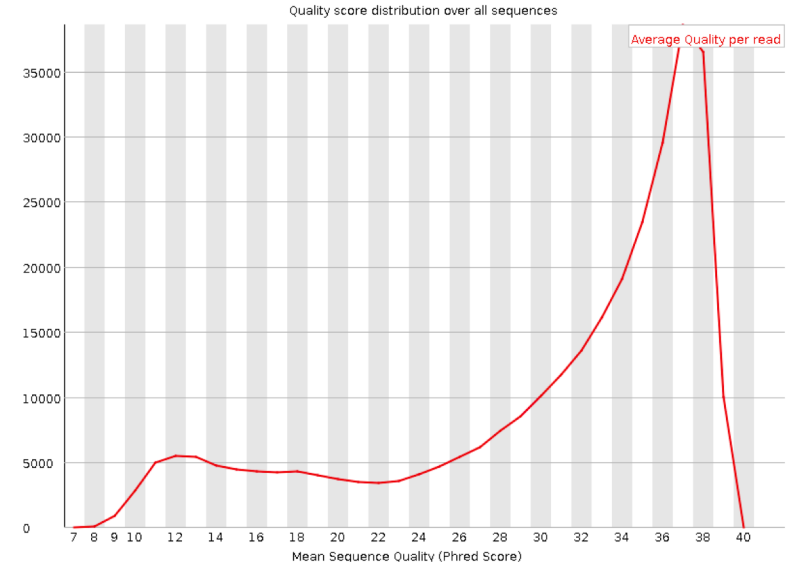
- How successful was the actual **sequencing**?
  - consistently high base call confidence

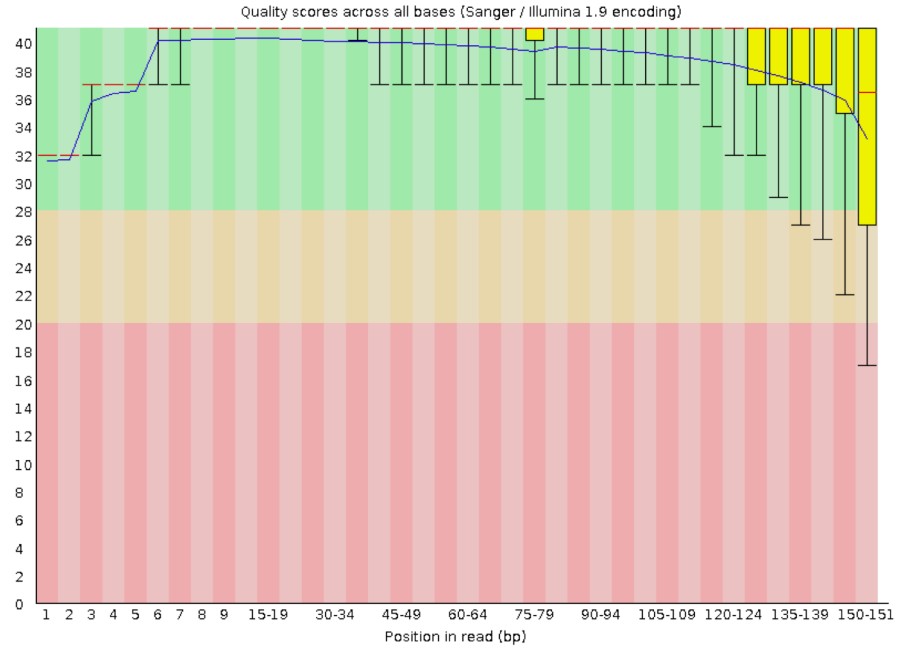
- Did our **library pep** generate a **faithful representation** of the DNA/RNA molecules in our samples?
  - ideally, the entire universe of transcripts has been sufficiently sampled (diverse library)
  - no contaminations (rRNA, foreign DNA, adapters, primers, …)
  - no bias towards fragments of certain GC contents/sizes
  - no degradation [cannot be assessed without alignment]

Per base sequence quality



Per sequence quality scores





More QC details

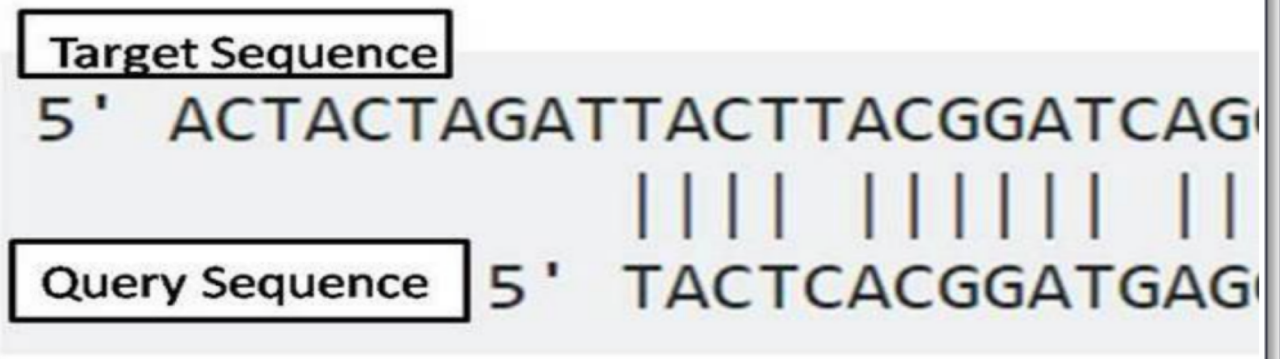
- Zhou, X., & Rokas, A. (2014). **Prevention, diagnosis and treatment of high-throughput sequencing data pathologies.** *Molecular Ecology*, *23*(7), 1679–1700. https://doi.org/10.1111/mec.12680
- https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq
- https://sequencing.qcfail.com/
- https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

Read Mapping

Different philosophies of transcript quantification

alignment followed by **counting** of reads overlapping with genes/exons

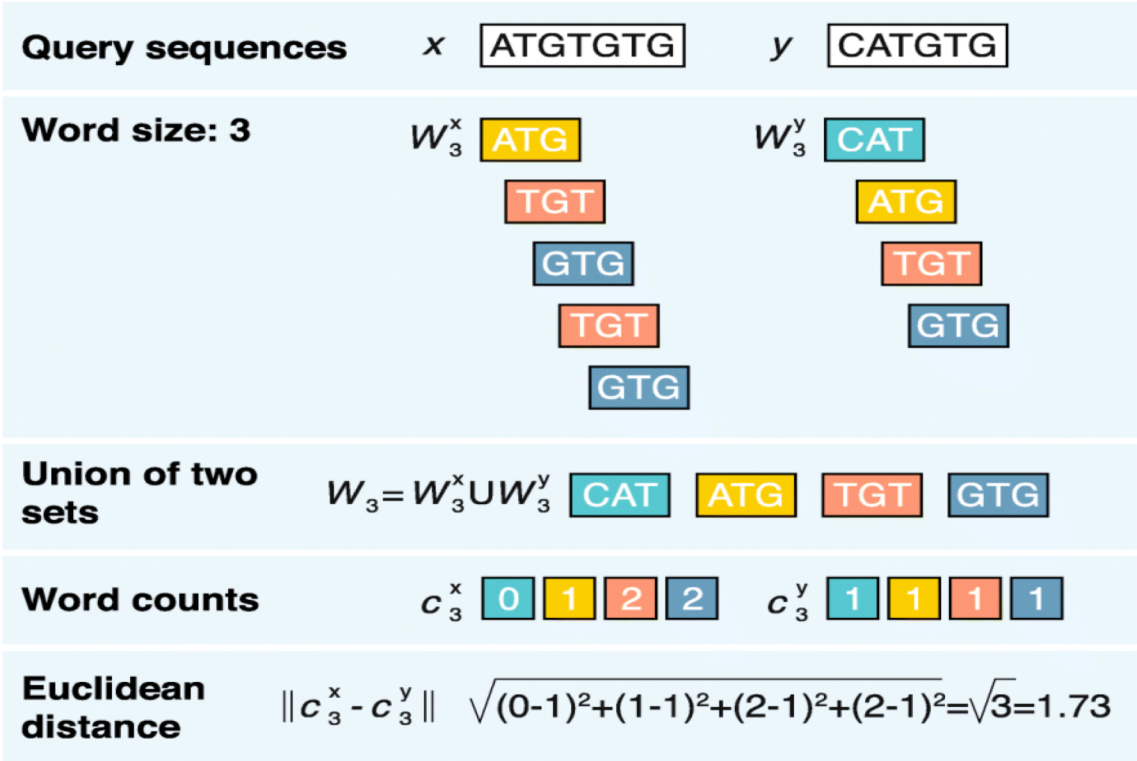
e.g. STAR + featureCounts



Both approaches absolutely rely on excellent reference sequences.

estimating expression levels of individual isoforms/genes based on **alignment-free k-mer matching**

salmon, kallisto



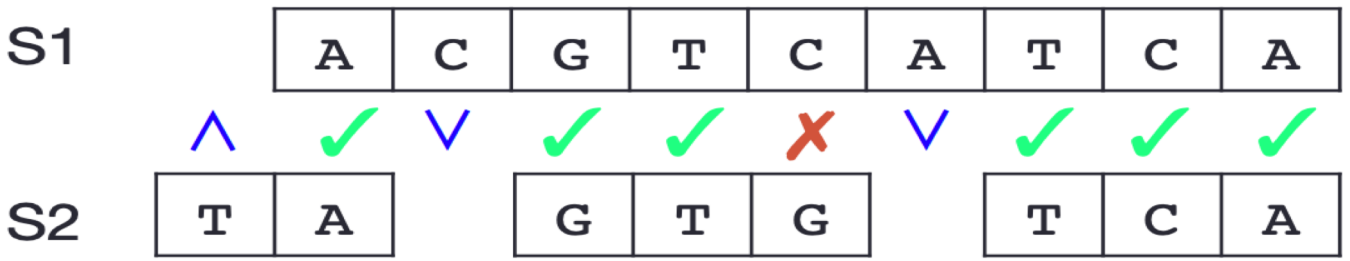
Zielezinski et al. (2017). Genome Biology. doi: 10.1186/s13059-017-1319-7

Read alignment basics

S1: A C G T C A T C A

S2: T A G T G T C A

Alignment = lining up the letters of two (or more) strings so that each letter in S1 either matches a gap or another letter in S2.



edit distance
= number of changes that are needed to match S1 and S2

To find the best alignment, we need:

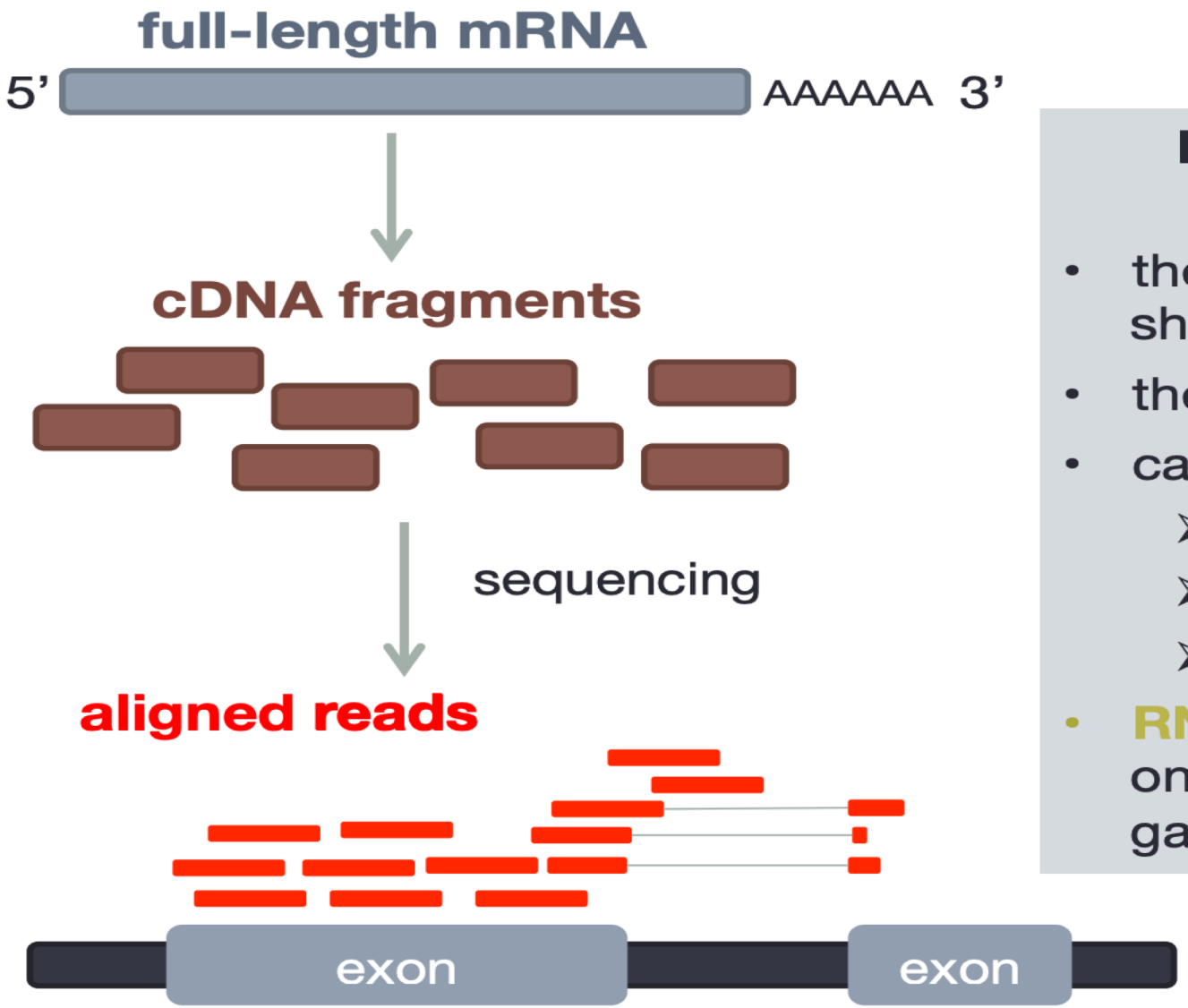
choices made by the programmer of a given tools

- **scoring function** for the edit distance
- efficient alignment-solving **algorithm**

Needleman-Wunsch | Smith-Waterman | BLAST

https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-096-algorithms-for-computational-biology-spring-2005/lecture-notes/lecture5_newest.pdf

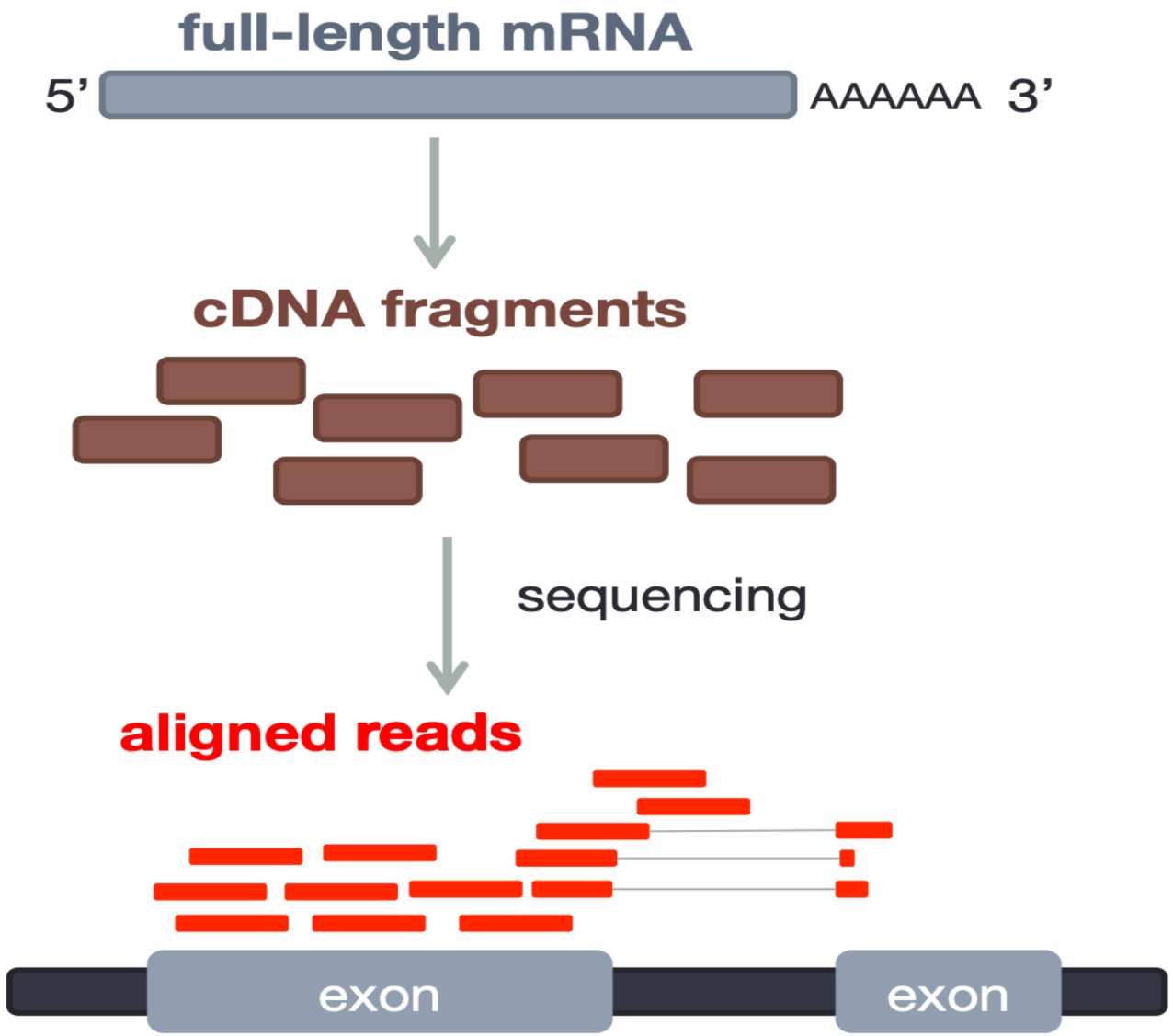
Aligning short RNA-seq reads



Particular challenges of Illumina sequencing:

- the query sequences (= reads) are very short
- there are millions of them!
- cannot expect 100% exact matches
  - seq. errors
  - biological variation
  - reference errors
- RNA-seq: some cDNA fragments can only be aligned if one allows for gigantic gaps (= **introns**)

Aligning short RNA-seq reads



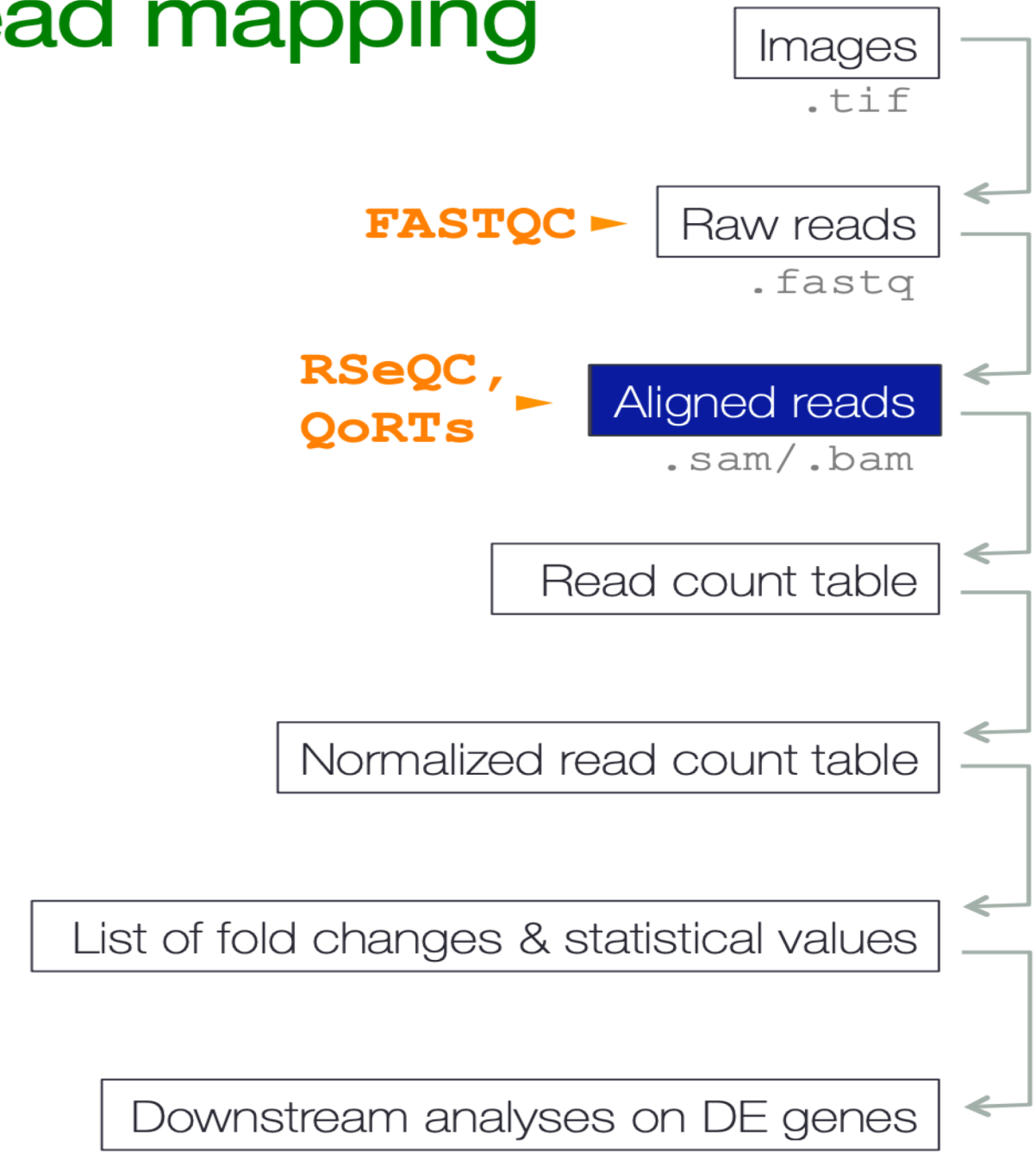
Spliced alignment tools usually need:

1) **reference genome** for the **alignment**
2) **annotation** to inform decisions about where to allow **gaps** in the alignment

greatest downside of alignment approach: it's resource-intensive!

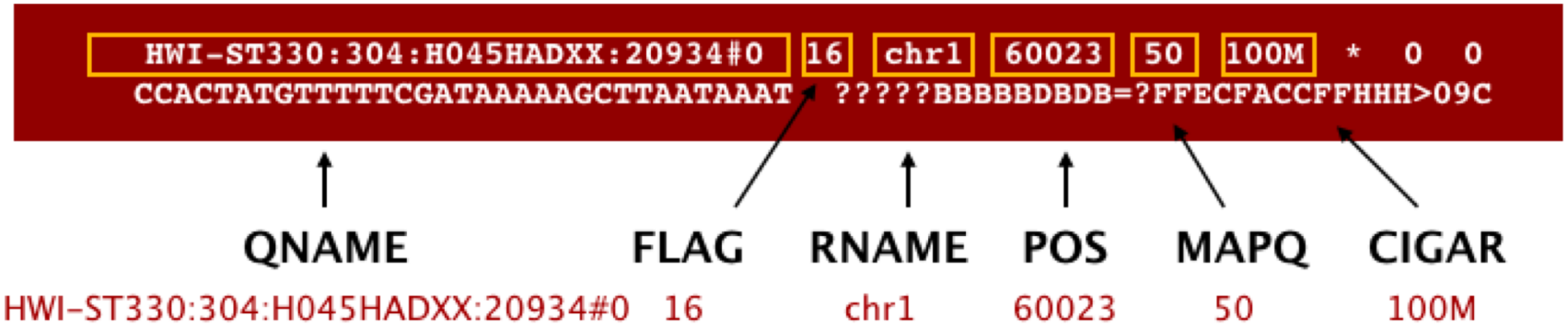
… and the result is **not inherently quantitative** (it's just read coordinates, really)!

Read mapping



1. Theoretical background
2. Reference & annotation files
3. Files with aligned reads: SAM/BAM
4. Looking at aligned reads

Alignment output: SAM/BAM

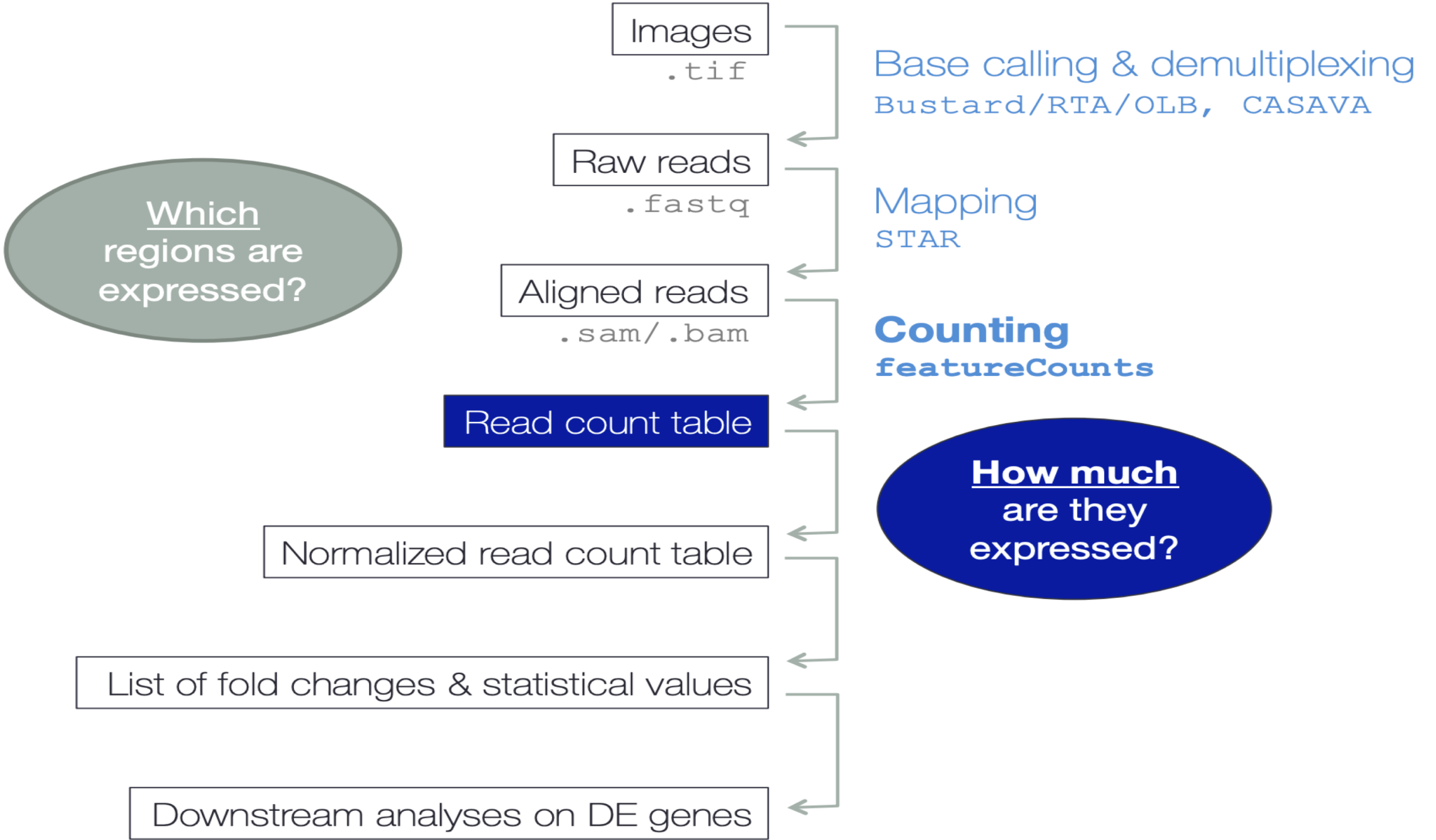


Alignment Stats.

Number of mapped reads:	38,147,300
Total number of alignments:	45,404,813
Number of secondary alignments:	7,257,513
Number of non-unique alignments:	0
Aligned to genes:	29,395,576
Ambiguous alignments:	2,493,354
No feature assigned:	12,435,808
Missing chromosome in annotation:	1,080,075
Not aligned:	1,824,541

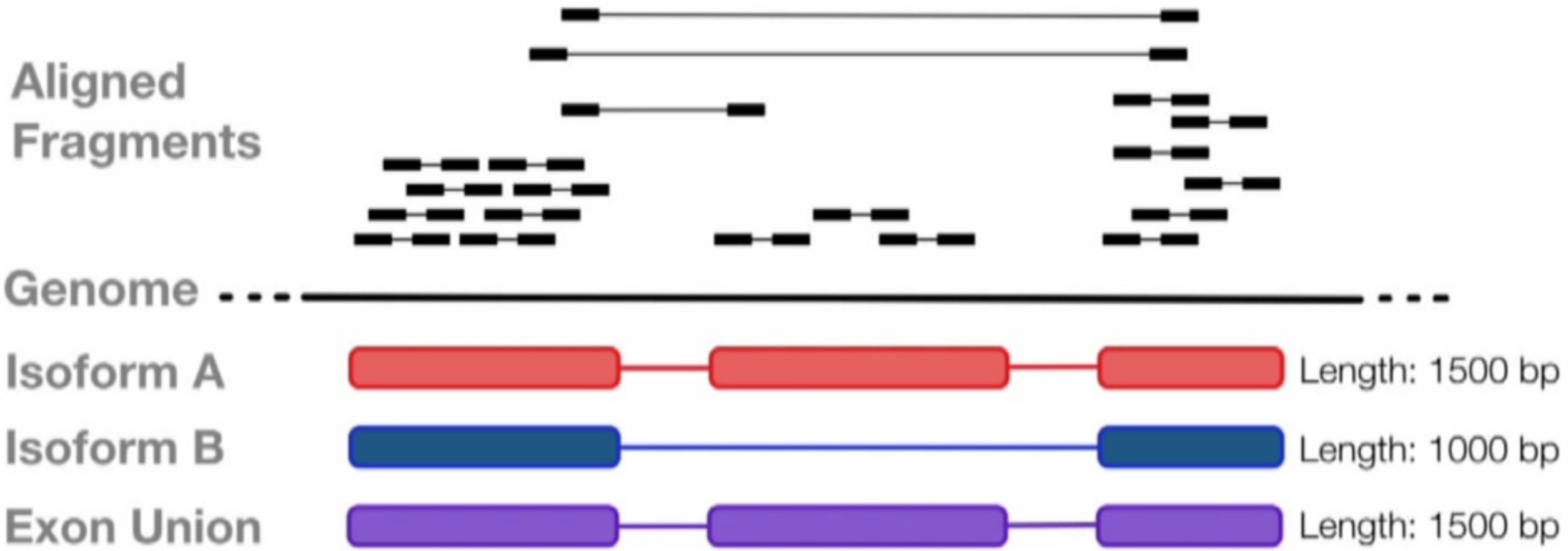
Counting Reads

Bioinformatics workflow of RNA-seq analysis

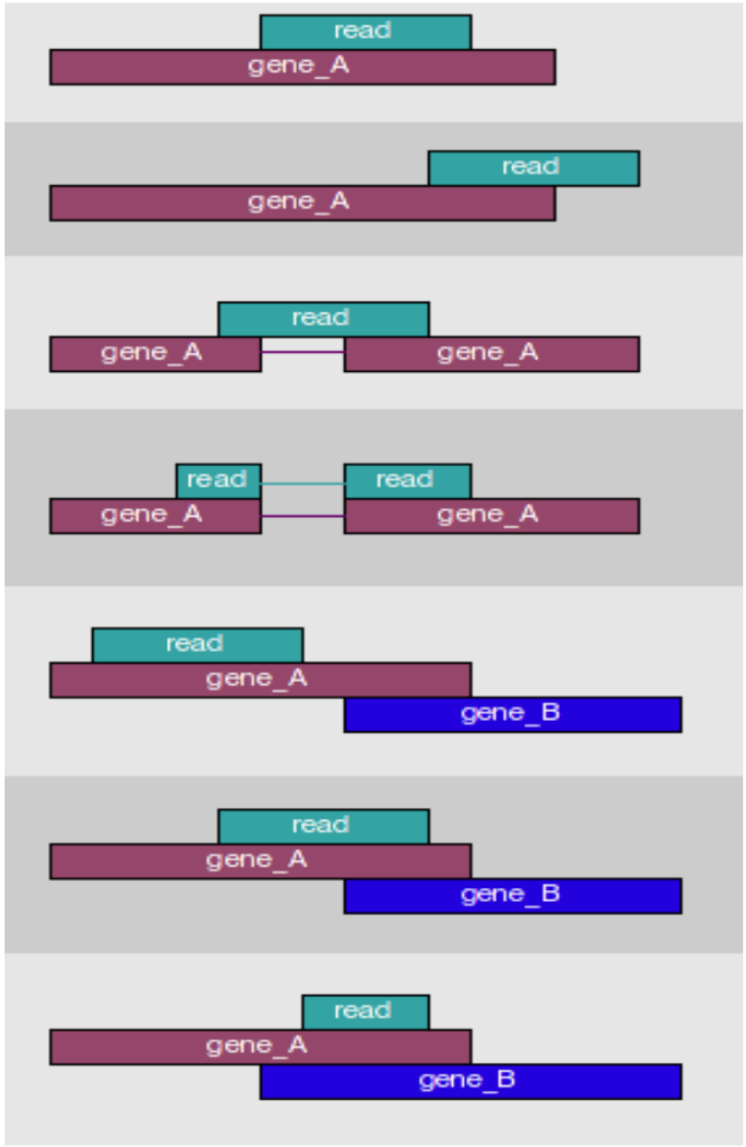


Quantifying expression

genes != transcripts



Counting read–gene overlaps



`featureCounts` will use read-gene overlaps as small as 1 bp

multi-overlap reads will be discarded

http://www-huber.embl.de/users/anders/HTSeq/doc/count.html#count

Normalization of read counts

From counting reads to expression units

- **Raw counts**: number of reads (or fragments) overlapping with the union of exons of a gene X_i

raw counts != expression strength

strongly influenced by:

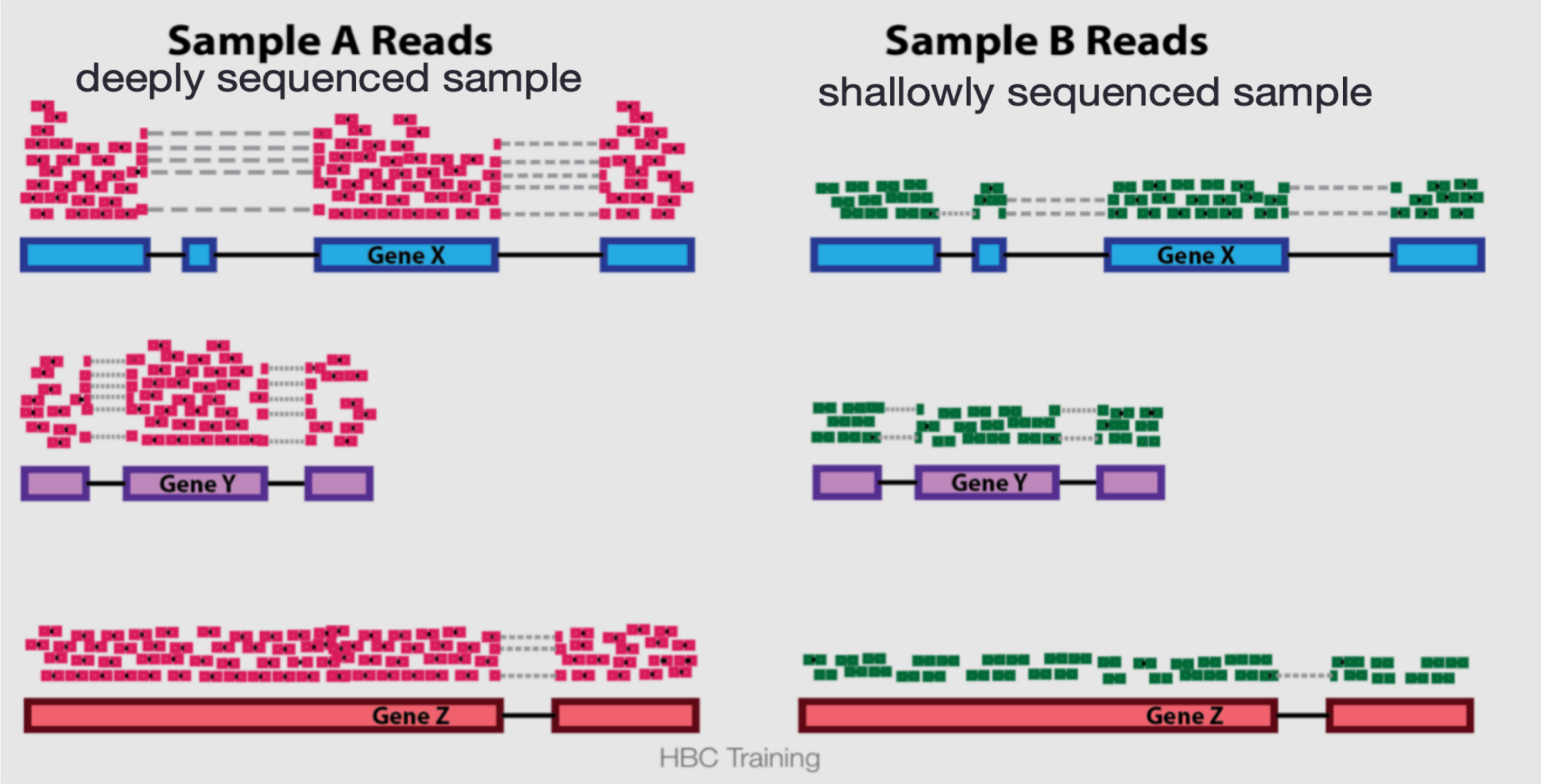
- gene length
- transcript sequence (% GC)
- sequencing depth
- expression of all other genes in the same sample

may cause variations for **different genes** expressed at the same level

may cause variations for the **same gene** in different samples

Influences on read count numbers

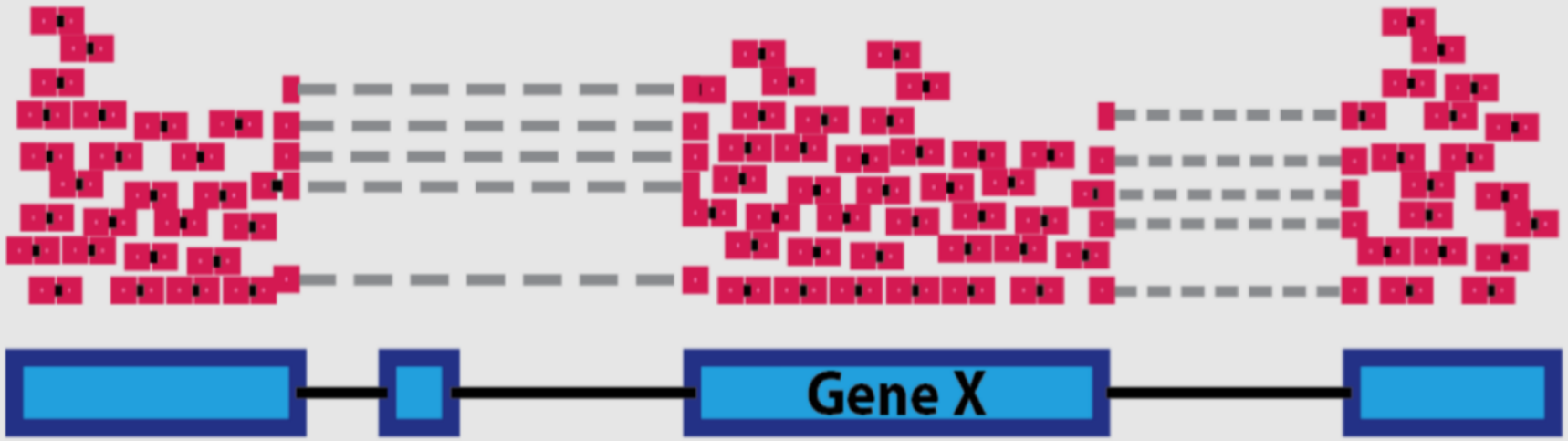
Sequencing depth, i.e. total number of reads/sample

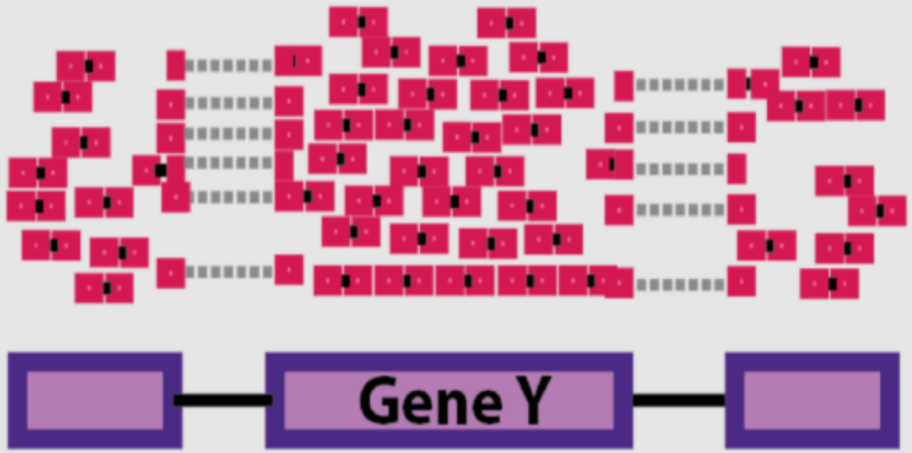


seq. depth of Sample A >> Sample B automatically leads to larger counts for the genes of Sample A even if the expression levels are the same

Influences on read count numbers

Gene lengths (and GC bias)





Gene X and Gene Y are expressed at the same levels, but the number of reads that originate off of their transcripts varies because they are of different lengths

Influences of read count numbers

Summary

GENE-SPECIFIC

- gene length
- transcript sequence (% GC)

need to be corrected when comparing different **genes**

SAMPLE-SPECIFIC

- sequencing depth
- expression of all other genes within the same sample

need to be corrected when comparing the same gene across different **samples**

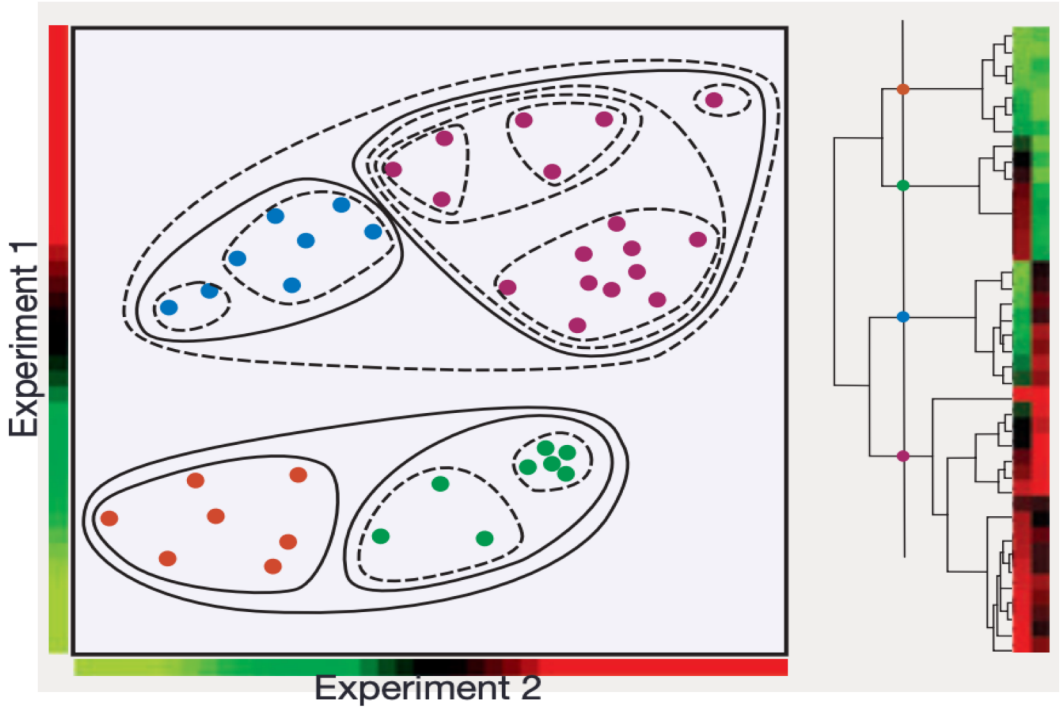
Normalisation methods for within sample gene counts

Name	Details	Comment
RPKM (reads per kilobase of exons per million mapped reads)	1. For each gene, count the number of reads mapping to it. 2. Divide that count by: the length of the gene in base pairs divided by 1,000 multiplied by the total number of mapped reads divided by 10^6. $$RPKM_i = \frac{read\,count\,of\,gene\,i}{(\frac{length\,of\,gene\,i}{10^3})(\frac{library\,size}{10^6})}$$	• introduces a bias in the per-gene variances, in particular for lowly expressed genes (Oshlack and Wakefield, 2009) • implemented in `edgeR`'s `rpkm()` function
FPKM (fragments per kilobase...)	1. Same as RPKM, but for paired-end reads: 2. The number of fragments (defined by two reads each) is used.	• implemented in `DESeq2`'s `fpkm()` function
TPM	Instead of normalizing to the total library size, TPM represents the abundance of an individual gene i in relation to the abundances of the other transcripts (e.g., j) in the sample. 1. For each gene, count the number of reads mapping to it and divide by its length in base pairs (= counts per base). 2. Multiply that value by 1 divided by the sum of all counts per base of every gene. 3. Multiply that number by 10^6. $$TPM_i = \frac{X_i}{l_i} * \frac{1}{\sum_j \frac{X_j}{l_k}}$$	• details in Wagner et al. (2012)

Some exploration

Clustering gene expression values

Goal: partition the samples into homogeneous groups such that the within-group similarities are large.



single-sample (or single-gene) clusters are successively joined

+ "unbiased"
- not very robust

- Result: dendrogram
  - clustering obtained by cutting the dendrogram at the desired level
- Similarity measures
  - Euclidean
  - Pearson correlation
- Distance measures
  - Complete: largest distance
  - Average: average distance

R function: `hclust()`

Figure from D'haeseleer P. Nat Biotechnol. 2005 Dec;23(12):1499-501. 10.1038/nbt1205-1499

PCA

starting point: matrix with expression values per gene and sample,
e.g. 7,100 genes x 10 samples

	SNF2_1	SNF2_2	SNF2_3	SNF2_4	SNF2_5	WT_1	WT_2	WT_3	WT_4	WT_5
YDL248W	109	84	100	112	62	47	65	60	95	43
YDL247W.A	0	1	1	0	3	0	0	1	0	0
YDL247W	6	6	1	3	4	2	3	4	7	9
YDL246C	6	6	1	4	4	1	3	2	4	0
YDL245C	1	6	9	5	3	6	2	5	5	6
YDL244W	79	59	49	60	37	9	8	12	30	14

If we want to understand the main differences between SNF2 and WT samples, the most detailed view (with the most "dimensions") would entail all 7,100 genes.

However, it is probably enough to focus on the genes that are actually different.
In fact, it'll be even better if we could somehow identify entire groups of genes that capture the majority of the differences.

PCA does exactly that ("grouping genes") using the correlation amongst each other.

	PC1	PC2
SNF2_1	-9.322866	0.8929154
SNF2_2	-9.390920	-0.6478100
SNF2_3	-9.176814	0.3460428
SNF2_4	-9.693035	1.2174519
SNF2_5	-9.450847	-0.3668670
WT_1	8.378671	-6.3321623
WT_2	10.421518	4.6749399
WT_3	8.486379	-1.1793146
WT_4	8.517490	-4.5814481
WT_5	11.230425	5.9762519

2 PCs (or more) x 10 samples

http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Principal component analysis

Goal: Reduce the dataset to fewer dimensions yet approx. preserve the distance between the individual samples

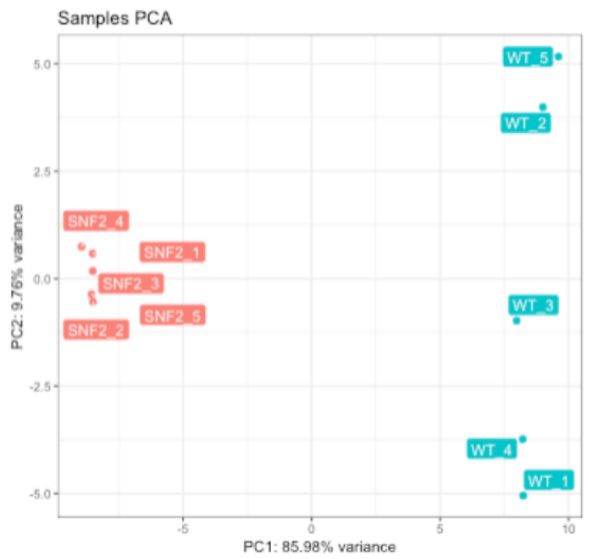
starting point: matrix with expression values per gene and sample,
e.g. 7,100 genes x 10 samples

	SNF2_1	SNF2_2	SNF2_3	SNF2_4	SNF2_5	WT_1	WT_2	WT_3	WT_4	WT_5
YDL248W	109	84	100	112	62	47	65	60	95	43
YDL247W.A	0	1	1	0	3	0	0	1	0	0
YDL247W	6	6	1	3	4	2	3	4	7	9
YDL246C	6	6	1	4	4	1	3	2	4	0
YDL245C	1	6	9	5	3	6	2	5	5	6
YDL244W	79	59	49	60	37	9	8	12	30	14

7,100 principal components x 10 samples

- vectors along which the variation between samples is maximal
- PC1-3 usually sufficient to capture the major trends!

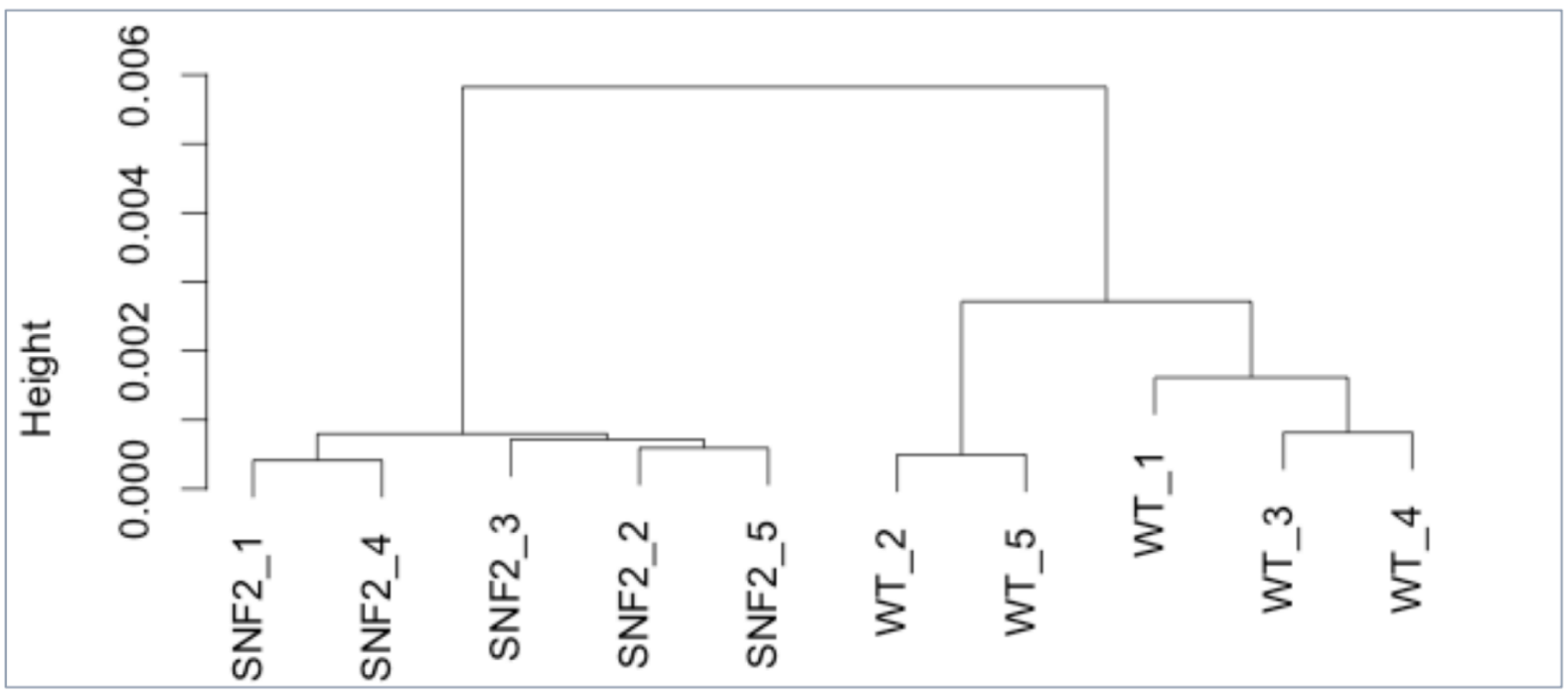
	PC1	PC2
SNF2_1	-9.322866	0.8929154
SNF2_2	-9.390920	-0.6478100
SNF2_3	-9.176814	0.3460428
SNF2_4	-9.693035	1.2174519
SNF2_5	-9.450847	-0.3668670
WT_1	8.378671	-6.3321623
WT_2	10.421518	4.6749399
WT_3	8.486379	-1.1793146
WT_4	8.517490	-4.5814481
WT_5	11.230425	5.9762519



http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Read count table

	SNF2_1	SNF2_2	SNF2_3	SNF2_4	SNF2_5	WT_1	WT_2	WT_3	WT_4	WT_5
YAL012W	7347	7170	7643	8111	5943	4309	3769	3034	5601	4164
YAL068C	2	2	2	1	0	0	0	0	2	2
YAL067C	103	51	44	90	53	12	23	21	30	29
YAL066W	2	0	0	0	0	0	0	0	0	0
YAL065C	5	9	6	3	1	10	5	2	4	3
YAL064W-B	13	9	10	9	6	9	12	4	4	8



List of relevant packages

CRAN

- data.table: https://cran.r-project.org/package=data.table
- dplyr: https://cran.r-project.org/package=dplyr
- reshape2: https://cran.r-project.org/package=reshape2
- ggplot2: https://cran.r-project.org/package=ggplot2
- ggrepel: https://cran.r-project.org/package=ggrepel
- ggbeeswarm: https://cran.r-project.org/package=ggbeeswarm

Bioconductor

- GenomicFeatures: https://bioconductor.org/packages/GenomicFeatures/
- tximport: https://bioconductor.org/packages/tximport/
- DESeq2: https://bioconductor.org/packages/DESeq2/
- apeglm: https://bioconductor.org/packages/apeglm/
- EnhancedVolcano: https://bioconductor.org/packages/EnhancedVolcano/
- DRIMSeq: https://bioconductor.org/packages/DRIMSeq/
- DEXSeq: https://bioconductor.org/packages/DEXSeq/
- stageR: https://www.bioconductor.org/packages/stageR/
- edgeR: https://www.bioconductor.org/packages/edgeR/

GitHub

- wasabi: https://github.com/COMBINE-lab/wasabi

Resources used

- https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/
- https://galaxyproject.org/tutorials/rb_rnaseq/
- https://chagall.med.cornell.edu/RNASEQcourse/.
- Love *et al.* (2015) RNA-Seq workflow: gene-level exploratory analysis and differential expression, F1000R 4 – 1070