TOPIC 6:

RNAseq and analysis of differential gene expression

BIOL525D - Bioinformatics for Evolutionary Biology 2021

Outline

1. Introduction and background
2. Overview of the methods and workflow
3. Quantifying expression levels
4. Analyzing patterns of expression
5. Technical considerations

Learning outcomes

Explain how RNAseq is generated and used

Identify the basic steps to align and analyze RNAseq data

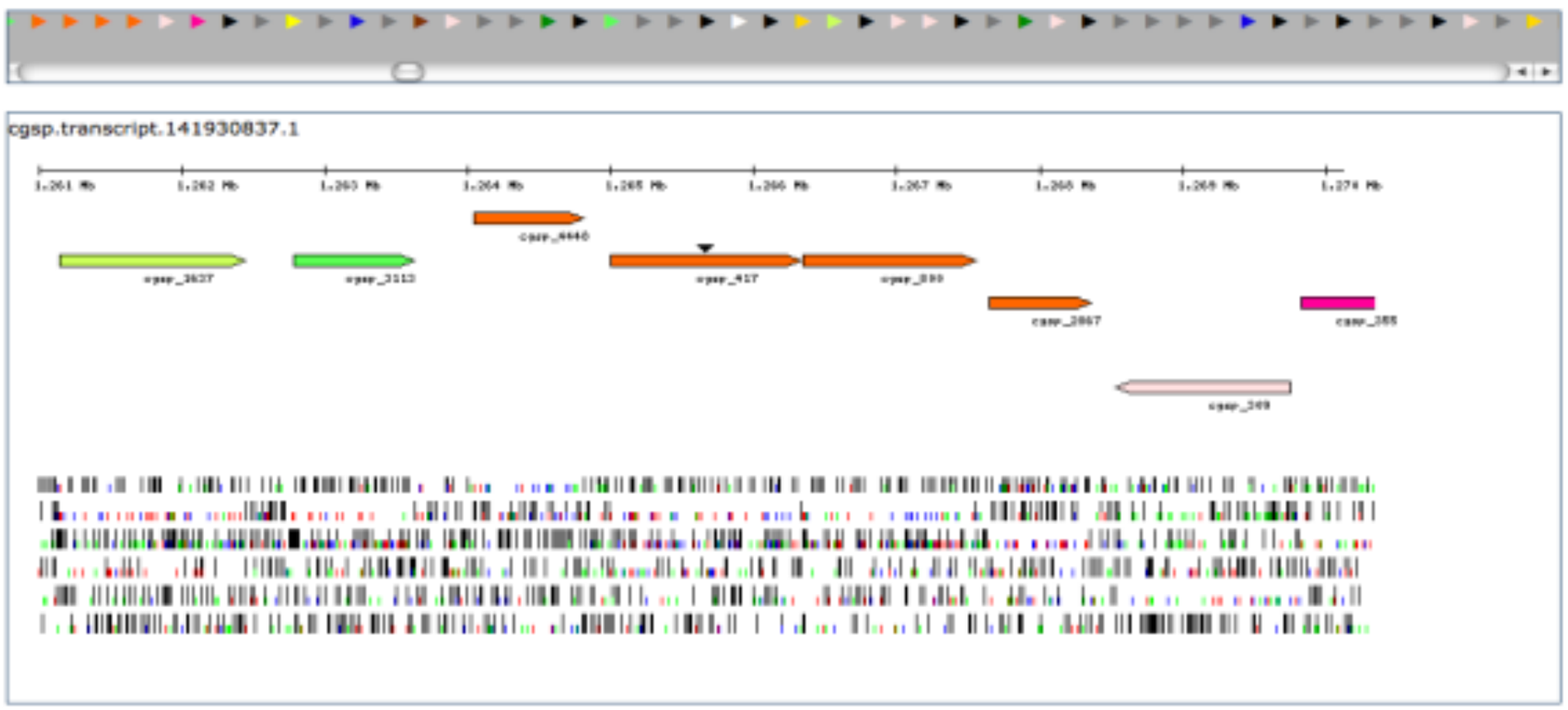
Introduction and background

Why use RNA-seq?

Can you think of some uses for RNA-seq?

Introduction and background

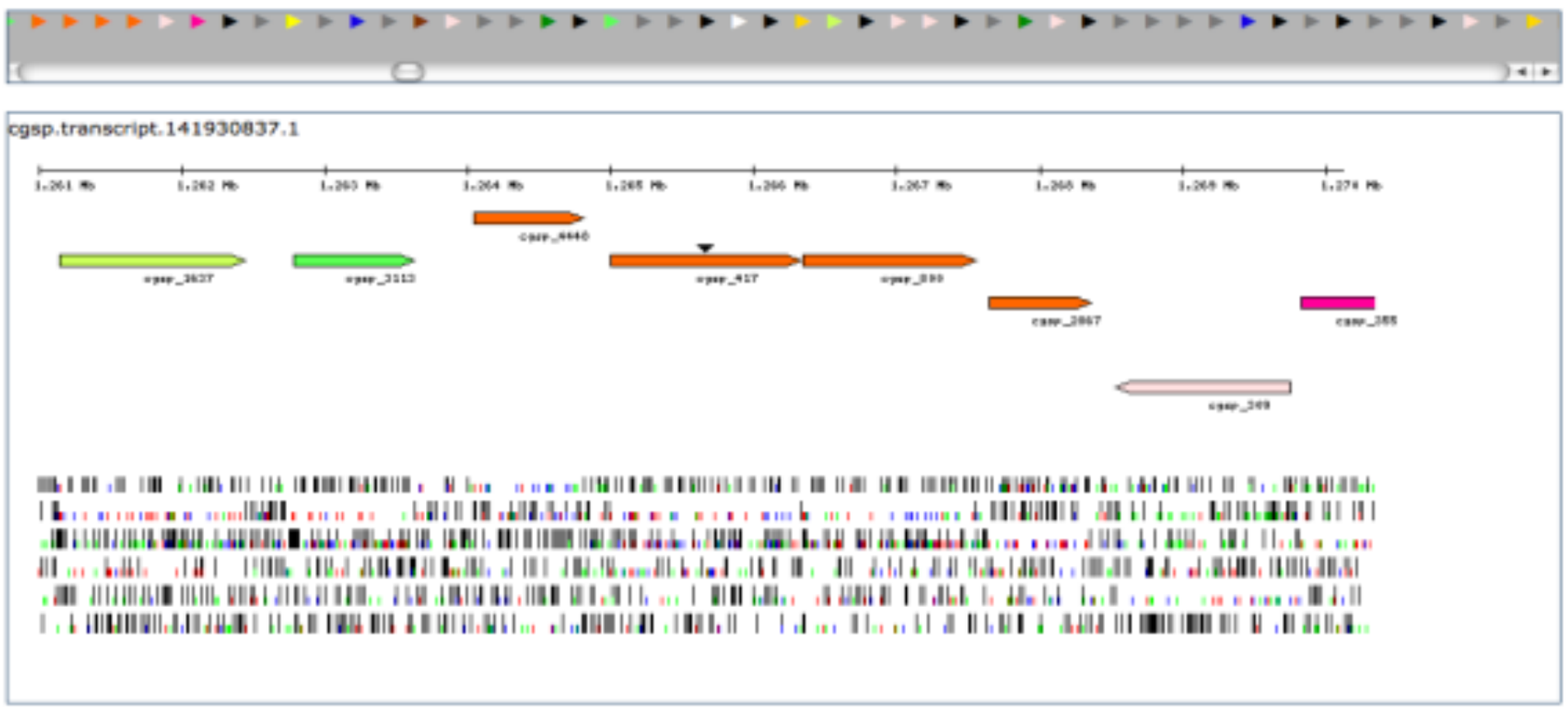
Why use RNA-seq?



Identifying transcribed regions of the genome - annotation

Introduction and background

Why use RNA-seq?

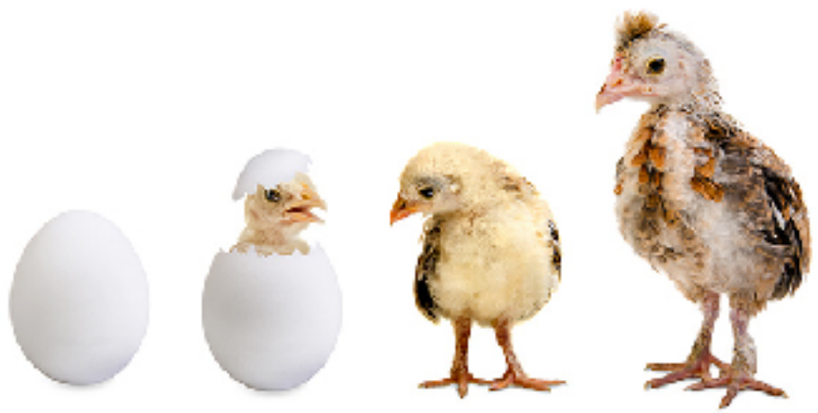


Identifying transcribed regions of the genome - annotation

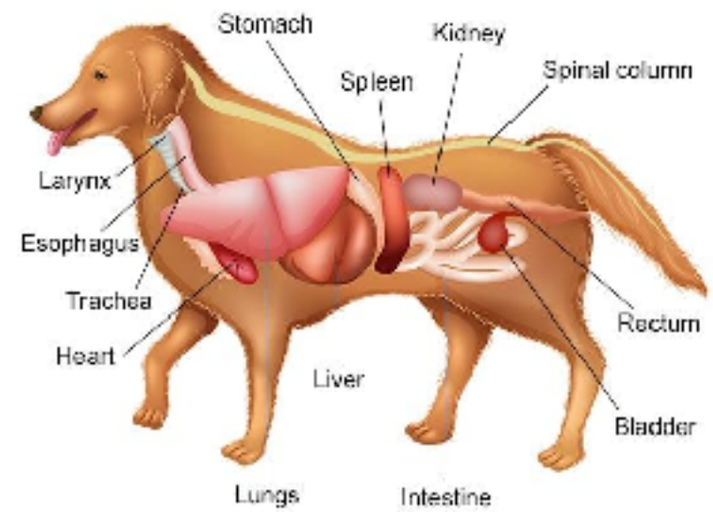
Genotyping transcribed regions of the genome

Introduction and background

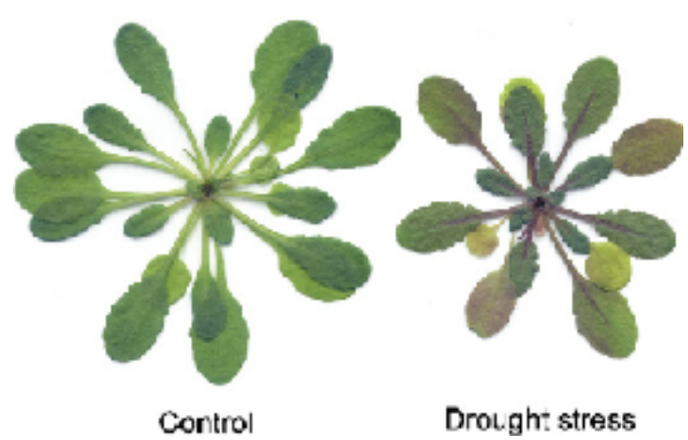
Quantifying expression differences



Developmental timepoints



Different organ, tissue or cell types



Experimental treatments

How is RNAseq data generated?

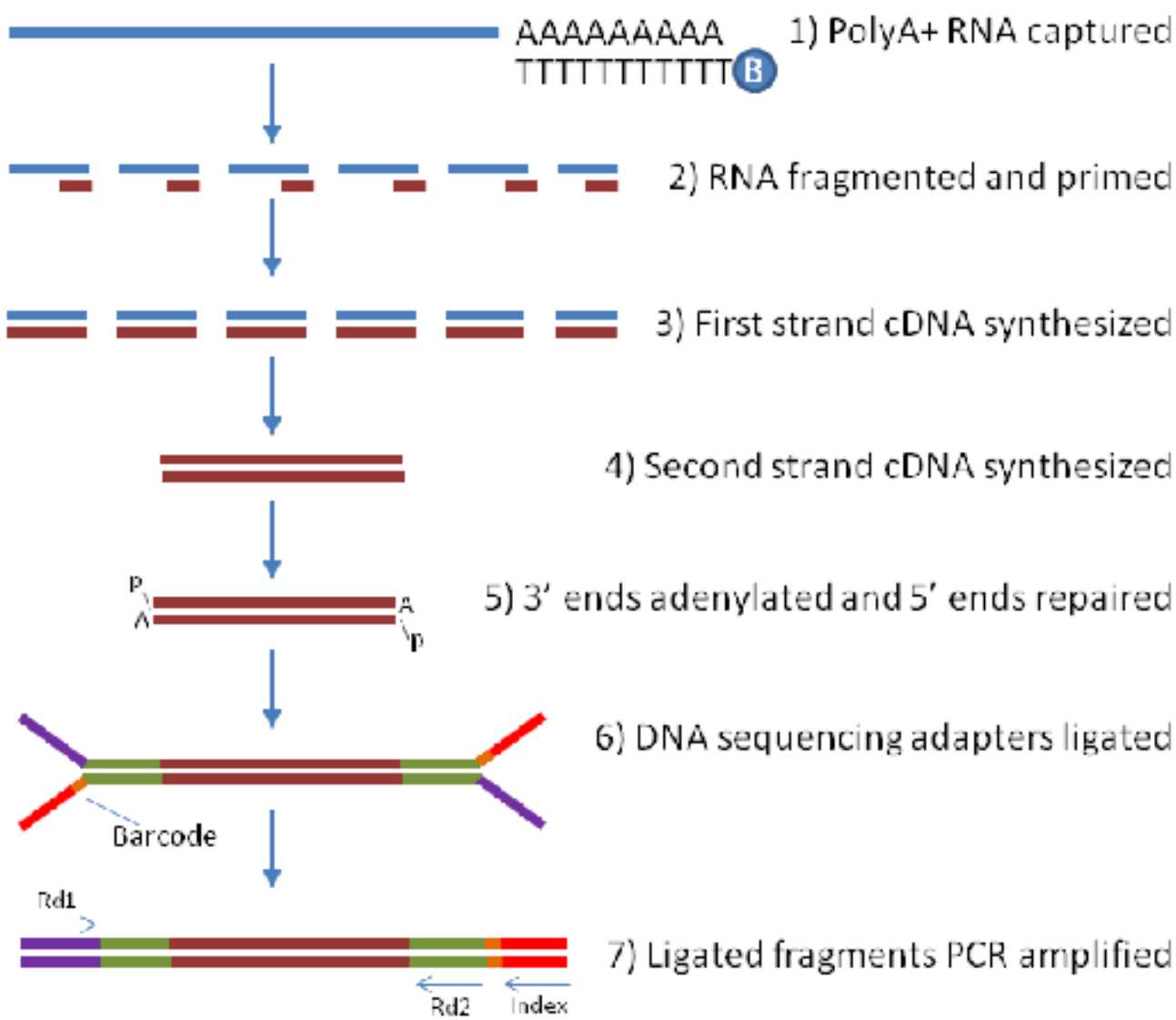
Overview of the methods

1. **RNA extraction protocol and sequencing**
2. Clean and filter reads
3. Map reads to a reference (genome or transcriptome)
4. Quantifying gene expression
5. Statistical analysis of differences in read counts

1. RNA extraction protocol and sequencing

mRNA is isolated, fragmented, and cDNA is synthesized and sequenced

Standard Illumina paired-end data will thus represent a snapshot of the mRNA present in your sample



Can you tell that I'm a computational biologist?

Figure from http://aws.labome.com/

How is RNAseq data generated?

<u>Overview of the methods</u>

1. RNA extraction protocol and sequencing
2. **Clean and filter reads**
3. Map reads to a reference (genome or transcriptome)
4. Quantifying gene expression
5. Statistical analysis of differences in read counts

2. Clean and filter reads

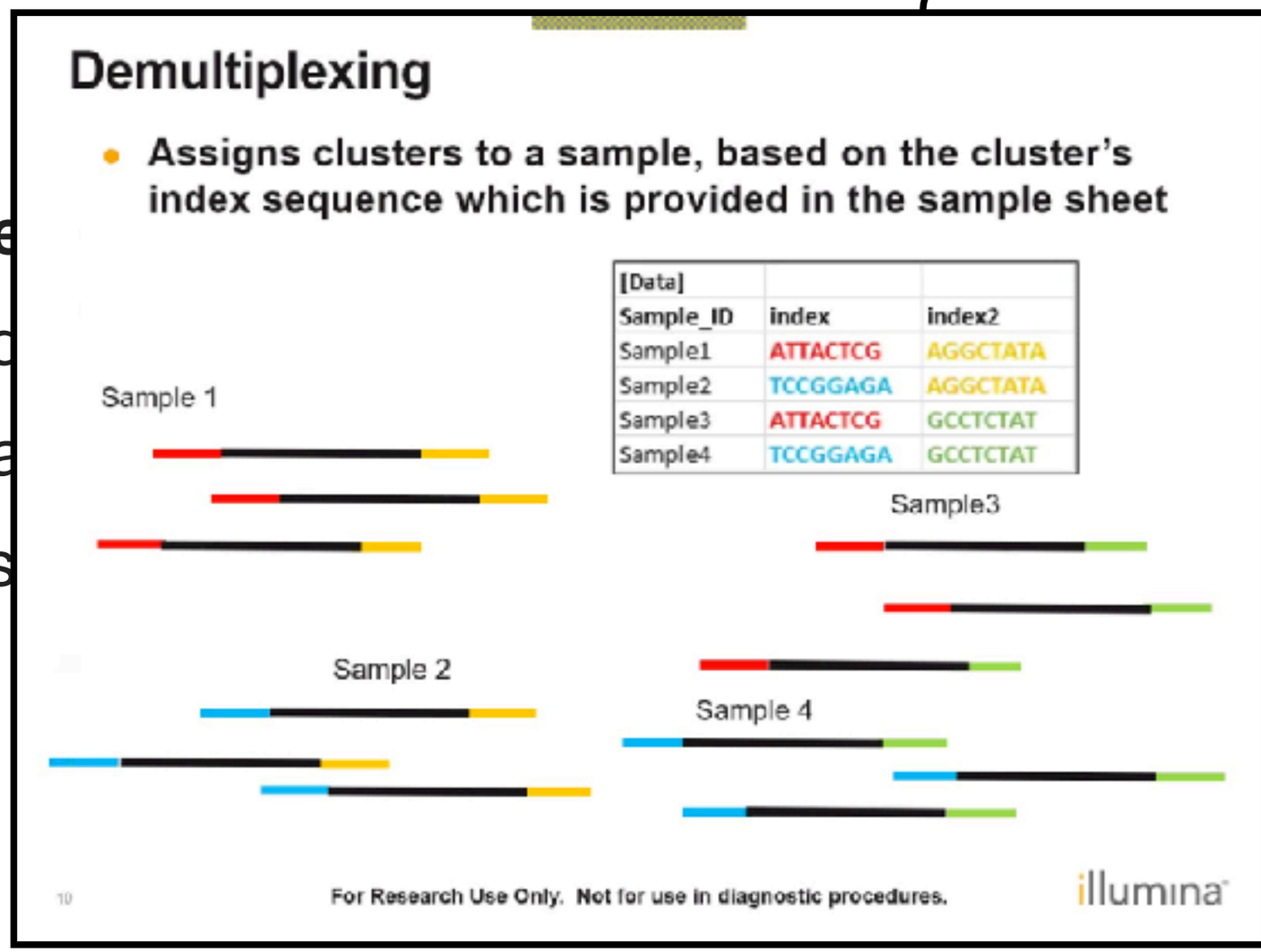
A. **Demultiplex by index or barcode**
B. Remove adapter sequences
C. Discard reads by quality/ambiguity
D. Filter reads by k-mer coverage

Samples that have been pooled onto the same sequencing lane need to be separated

Samples are distinguished using specific identifying DNA tags - those need to be removed before analysis

2. Clean and filter reads

A. **Demultiple**
B. Remove ac
C. Discard rea
D. Filter reads



t have been
o the same
ne need to be
rated

distinguished
ic identifying
ose need to be
fore analysis

2. Clean and filter reads

A. Demultiplex by index or barcode
B. Remove adapter sequences
C. Discard reads by quality/ambiguity
D. Filter reads by k-mer coverage

When fragments are sufficiently long, adapter sequences will not be sequenced, but some RNA reads may be fairly short, so adapters can end up being sequenced

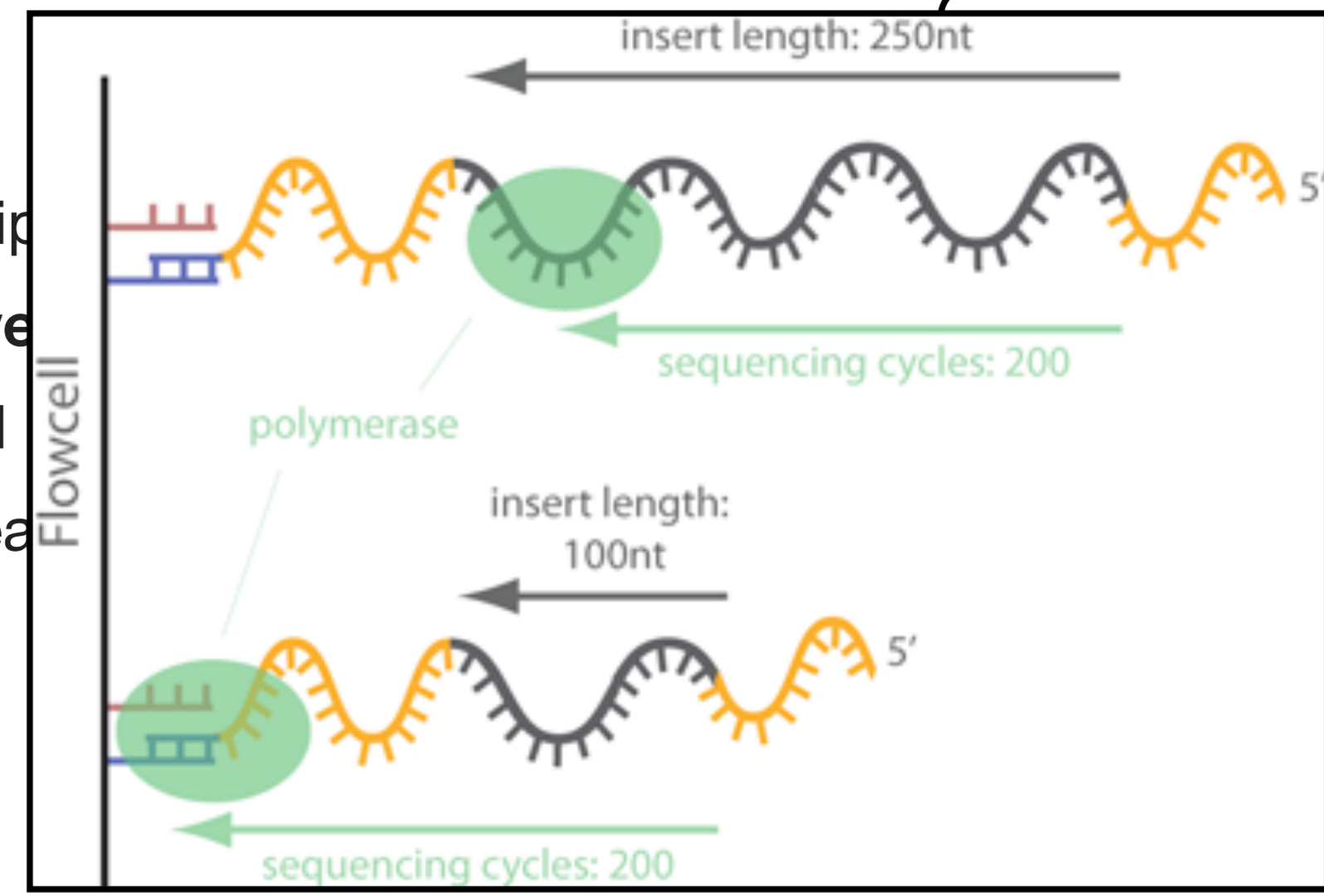
2. Clean and filter reads

A. Demultip

B. Remove

C. Discard

D. Filter rea



nts are
, adapter
l not be
some RNA
ly short, so
d up being
ed

2. Clean and filter reads

A. Demultiplex by index or barcode
B. Remove adapter sequences
C. Discard reads by quality/ambiguity
D. Filter reads by k-mer coverage

Remove reads with evidence of poor quality - particularly important when building a transcriptome

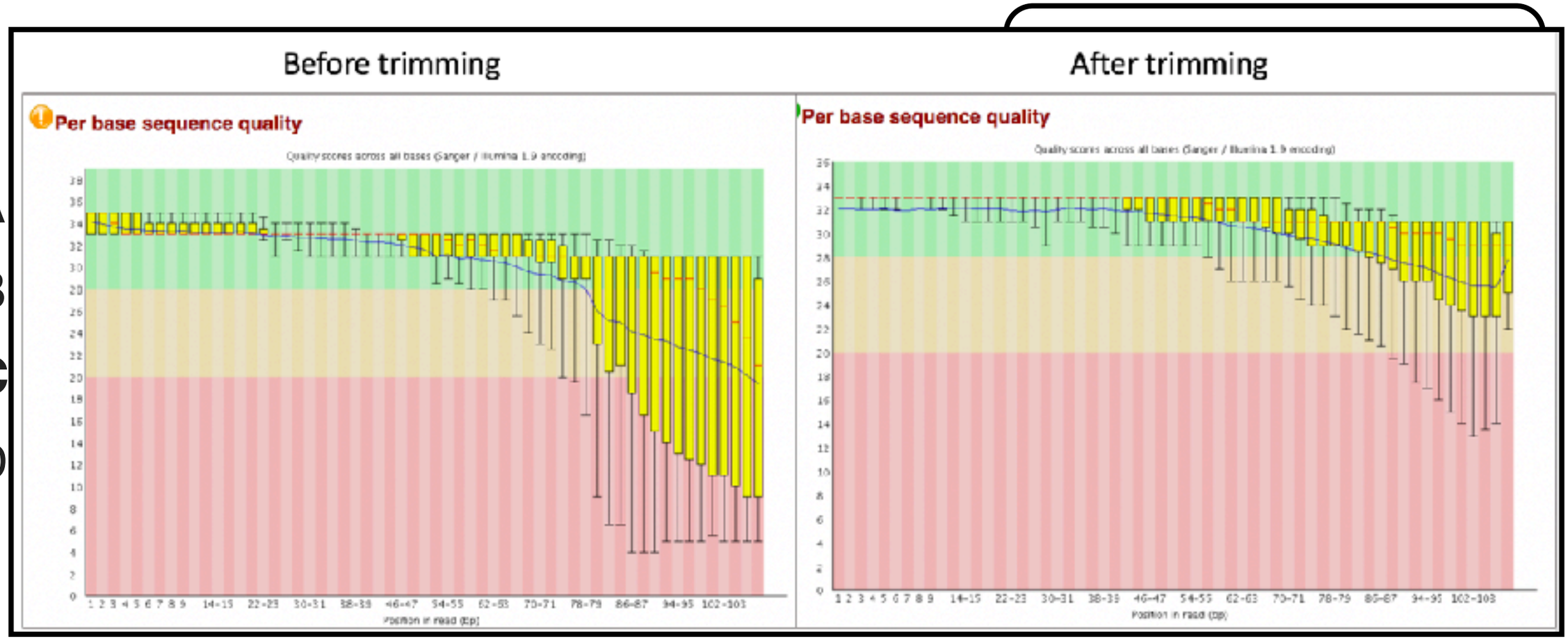
2. Clean and filter reads

A

B

C

D



2. Clean and filter reads

A. Demultiplex by index or barcode
B. Remove adapter sequences
C. Discard reads by quality/ambiguity
D. Filter reads by k-mer coverage

Gene sequences have characteristic distribution of k-mers

Deviations in distribution of k-mers can indicate sequencing errors

Sequencing errors can be very bad news when assembling transcriptomes

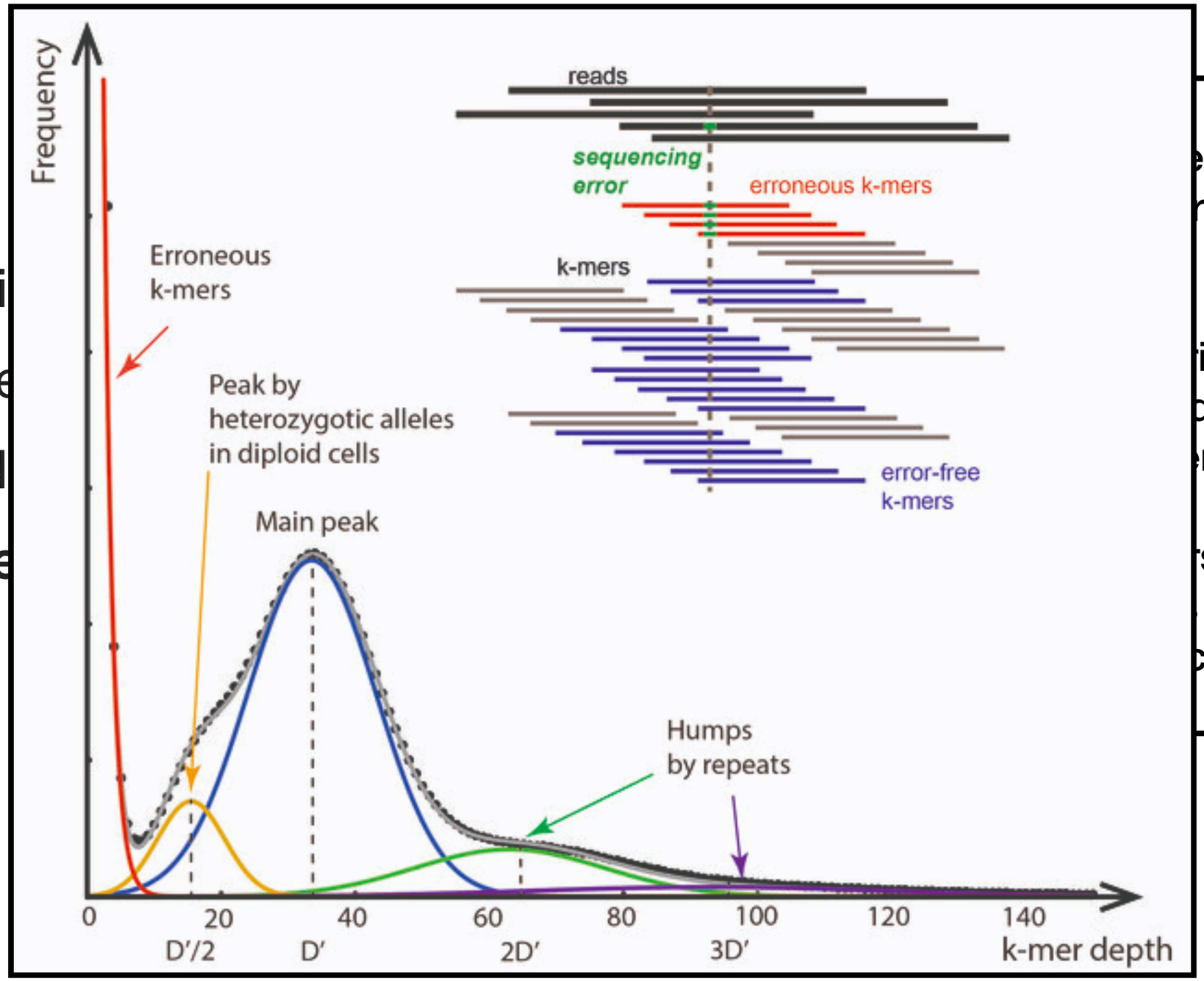
2. Clean and filter reads

A. Demulti

B. Remove

C. Discard

D. Filter re



es have
ribution of

ibution of
dicate
errors

s can be
when
criptomes

2. Clean and filter reads

Table 5.1 Read Processing Software

Software	De-mulitplexing	Adaptor Trimming	Quality Filtering/ Trimming	K-mer Filtering	K-mer Normalization
FASTX-Toolkit	✔	✔	✔		
Goby	✔	✔			
khmer				✔	✔
NGS_backbone		✔	✔		
Stacks	✔	✔	✔	✔	✔
trimmomatic		✔	✔		
biopieces	✔	✔	✔		

I'd recommend you take a look at the following for a more detailed overview of how/when/why to clean up your reads

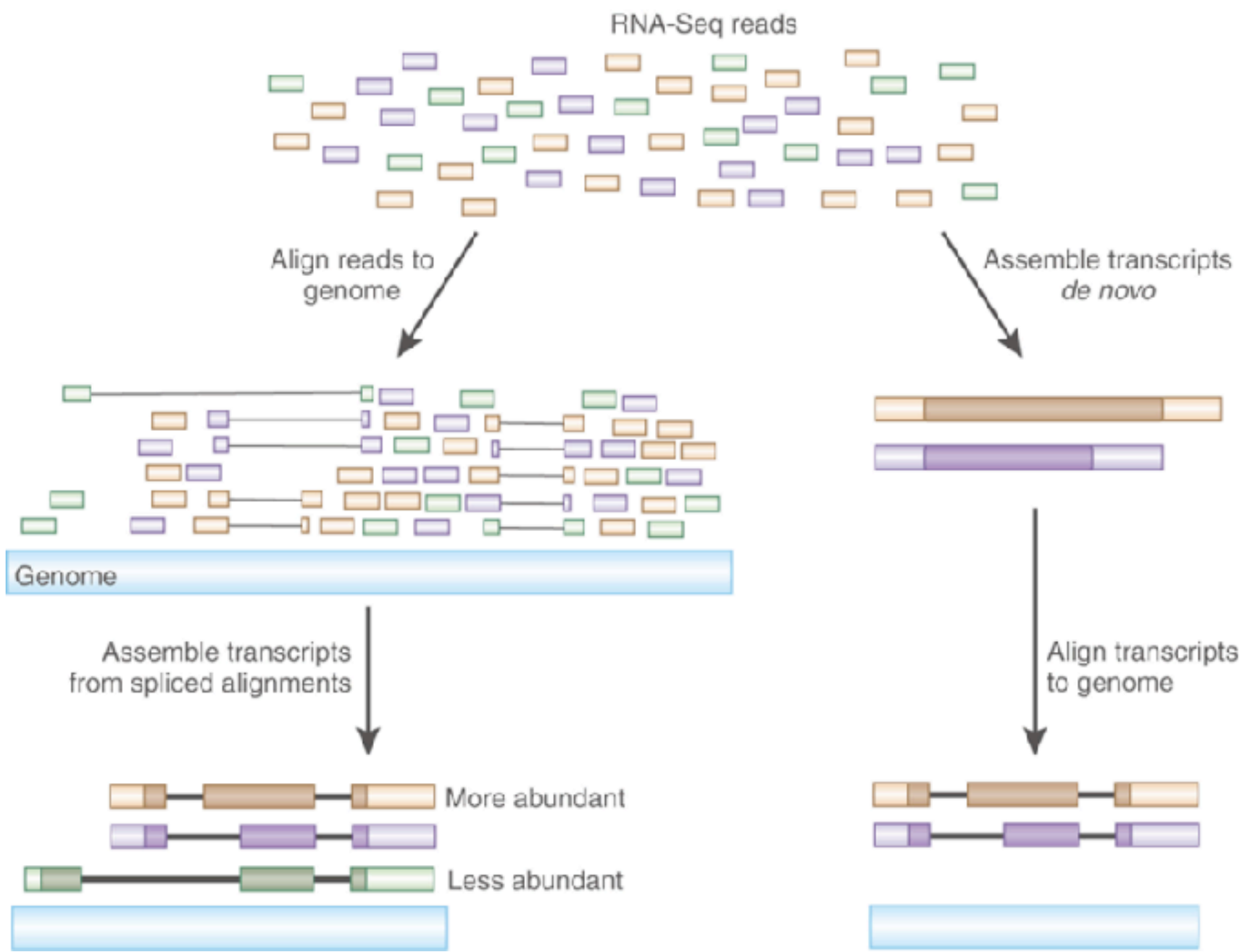
https://rnaseq.uoregon.edu/#analysis-initial-processing

How is RNAseq data generated?

<u>Overview of the methods</u>

1. RNA extraction protocol and sequencing
2. Clean and filter reads
3. **Map reads to a reference (genome or transcriptome)**
4. Quantifying gene expression
5. Statistical analysis of differences in read counts

3. Map reads to a reference (genome or transcriptome)



Assembling and Aligning

Haas & Zody 2010 Nature Biotechnology

3. Map reads to a reference (genome or transcriptome)

What difficulties arise when mapping RNA seq reads?

3. Map reads to a reference (genome or transcriptome)

What difficulties arise when mapping RNA seq reads?

A. Reads that map across intron/exon boundaries

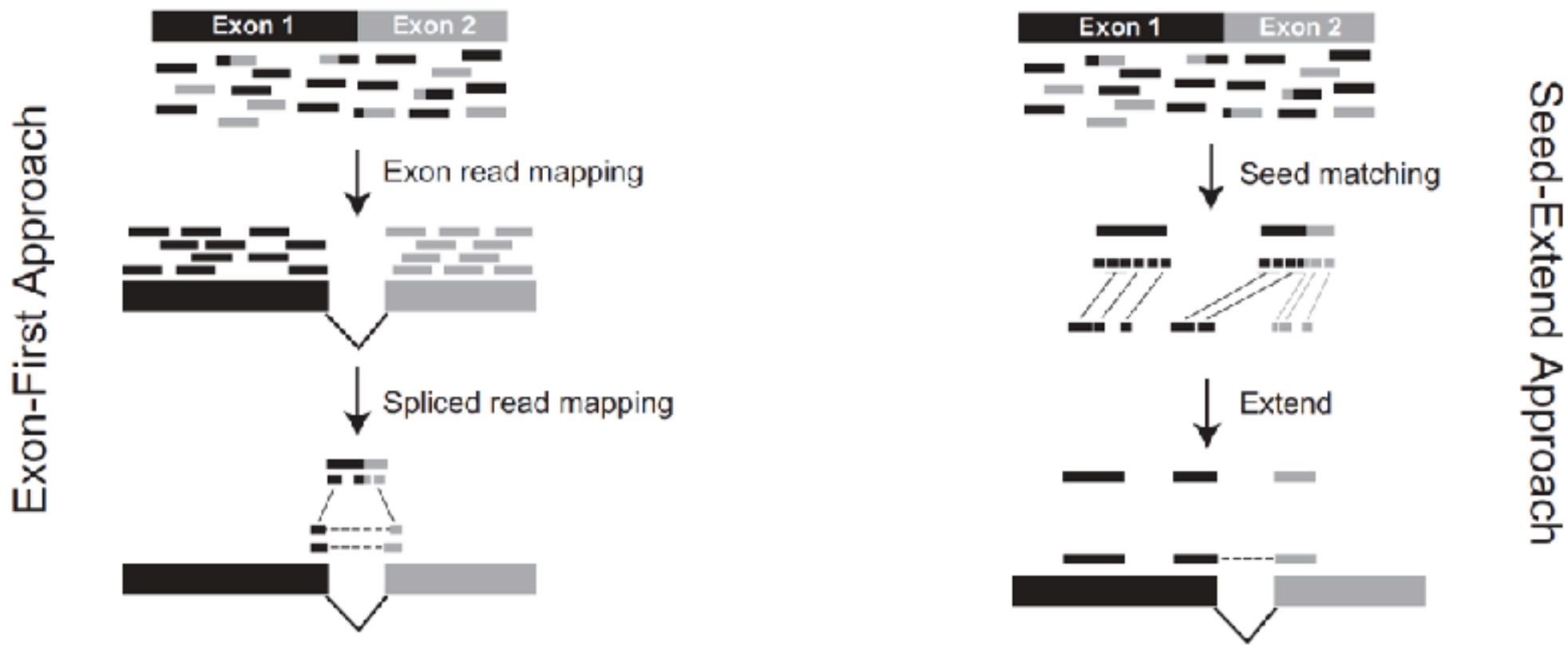
B. Identifying abundance of alternatively spliced transcripts

C. Dealing with multi-reads

3. Map reads to a reference (genome or transcriptome)

A. Reads that map across intron/exon boundaries

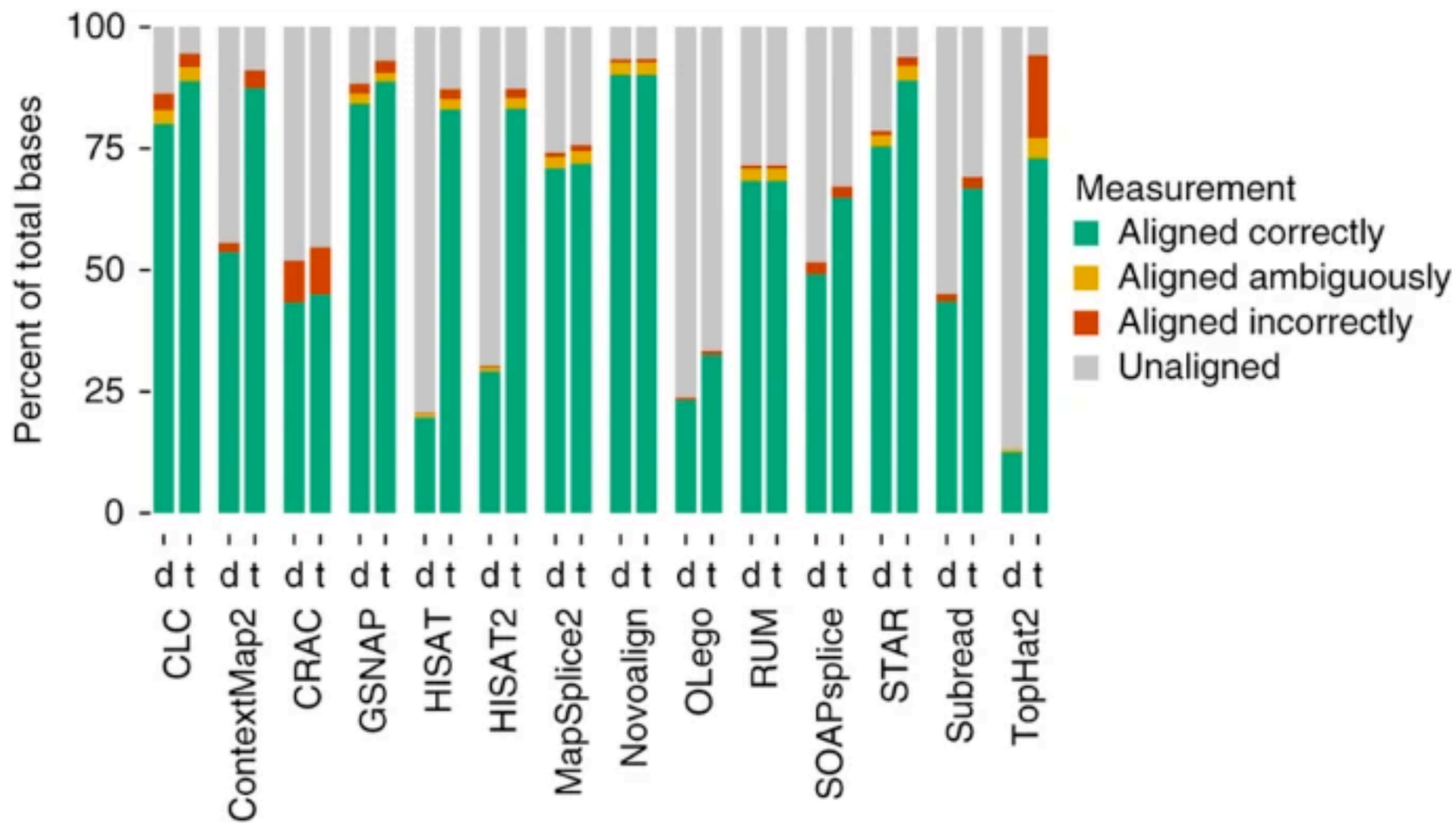
Specific algorithms have been developed for mapping RNA-seq reads to genomes



3. Map reads to a reference (genome or transcriptome)

A. Reads that map across intron/exon boundaries

Specific algorithms have been developed for mapping RNA-seq reads to genomes

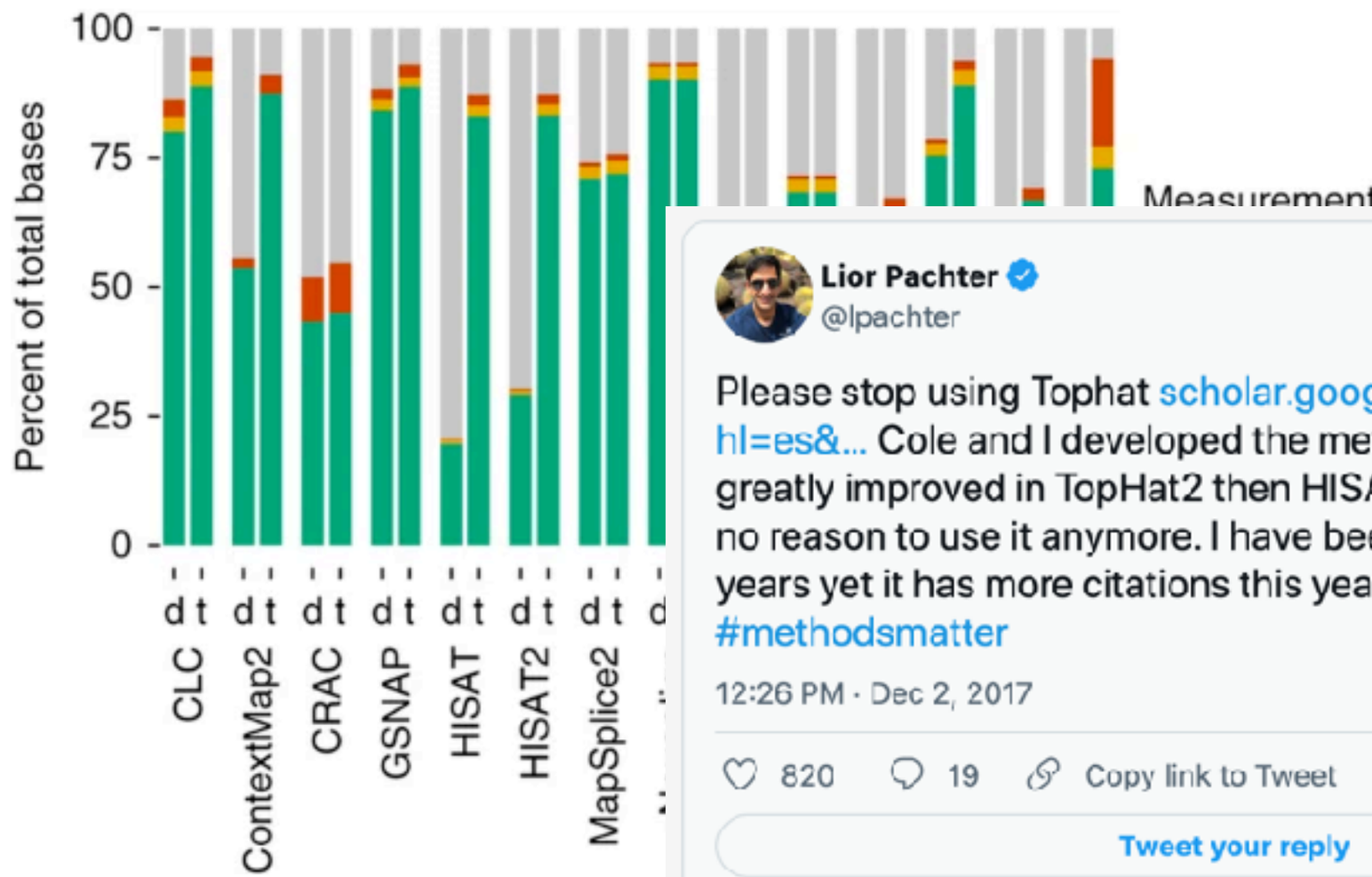


Baruzzo et al 2017

3. Map reads to a reference (genome or transcriptome)

A. Reads that map across intron/exon boundaries

Specific algorithms have been developed for mapping RNA-seq reads to genomes



3. Map reads to a reference (genome or transcriptome)

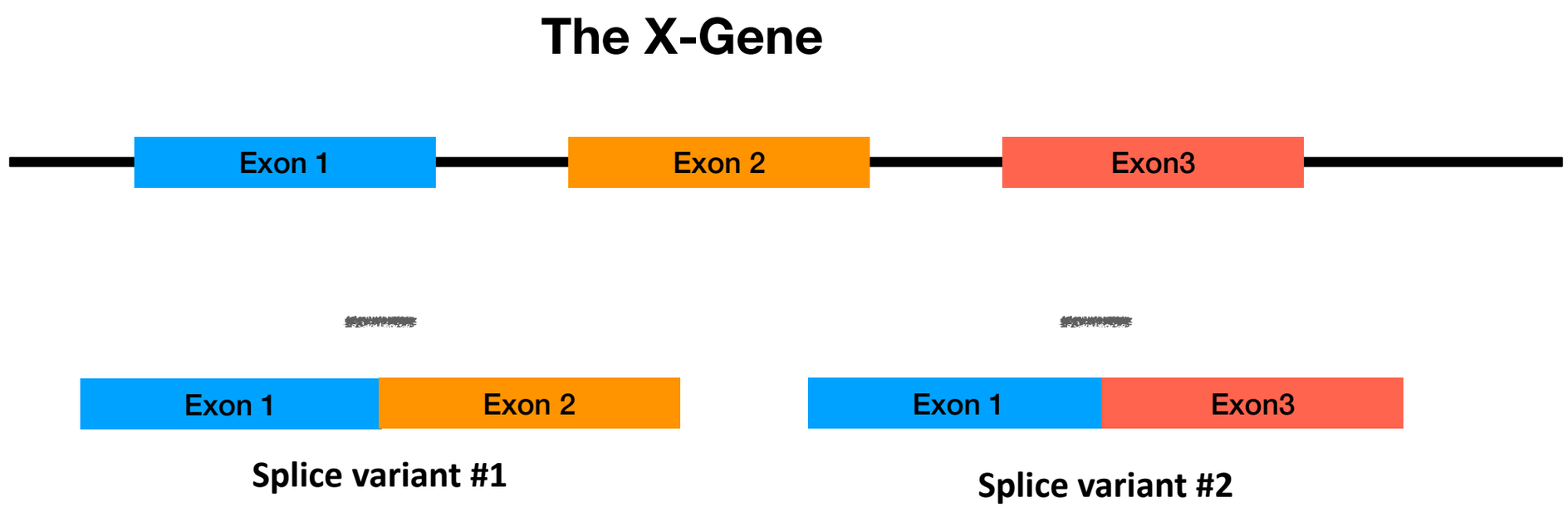
A. Reads that map across intron/exon boundaries

Alternatively, you can map reads directly to a transcriptome (e.g. RSEM)

A consensus has not yet been reached about the optimal approach, in practice what you do will likely be informed by the data you have

3. Map reads to a reference (genome or transcriptome)

B. Identifying abundance of alternatively spliced transcripts



If there are two known splice variants, a read spanning exon **1** & **2** or **1** & **3** will identify which variant is present

If a read aligned to either exon **2** or **3** then differential expression of isoforms can be inferred, relative to the expression levels of other isoforms

3. Map reads to a reference (genome or transcriptome)

C. Dealing with multi-reads

Both paralogs and alternatively spliced transcripts (isoforms) can give the problem of “multireads”: a read that maps with high score to several places

Li et al. (2010) found that 17% (mouse) or 52% (maize) of reads were multireads!!

3. Map reads to a reference (genome or transcriptome)

C. Dealing with multi-reads

A

B

Approach to handle multireads	Read distribution representation	Counts
Ignore		G1: 10 reads G2: 6 reads
Count once per alignment		G1: 18 reads G2: 14 reads
Split them equally		G1: 14 reads G2: 10 reads
Rescue based on uniquely mapped reads		G1: 15 reads G2: 9 reads
Expectation-maximization	(n)	G1: 15 reads G2: 9 reads
Read coverage based methods		G1: 15 reads G2: 9 reads
Cluster methods		G1:10 reads G2:6 reads Cluster G1/G2: 8 reads

Deschamps-Francoeur et al
Comp Struc. Biochem.

How is RNAseq data generated?

<u>Overview of the methods</u>

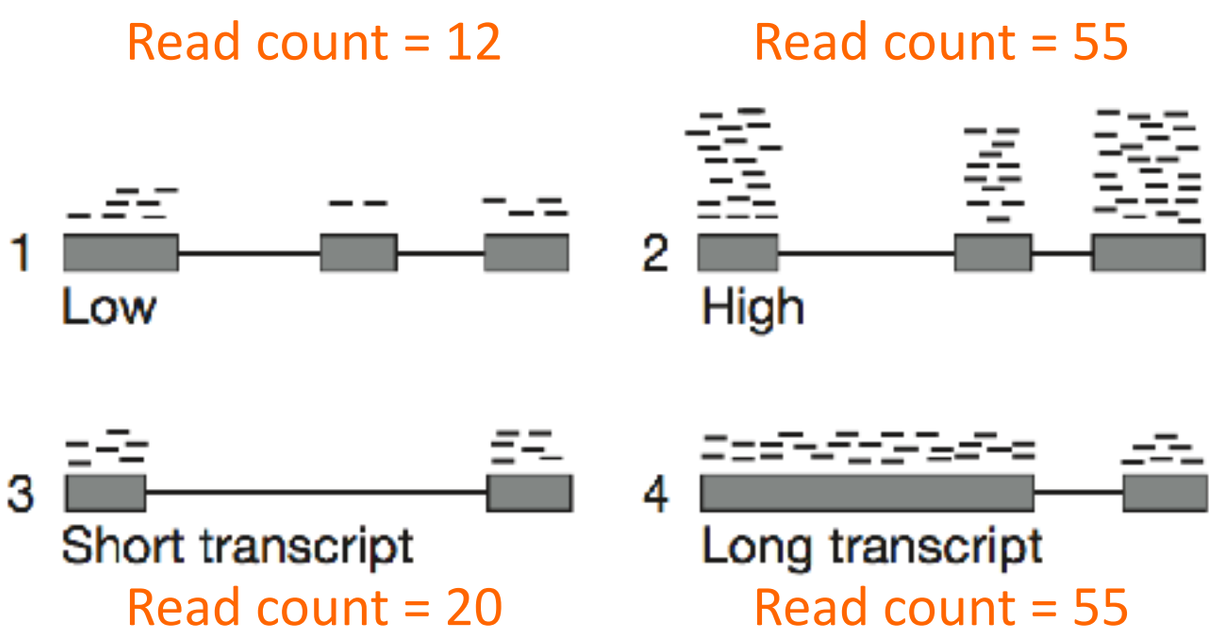
1. RNA extraction protocol and sequencing
2. Clean and filter reads
3. Map reads to a reference (genome or transcriptome)
4. **Quantifying gene expression**
5. Statistical analysis of differences in read counts

4. Quantifying gene expression

RNAseq normalization needed due to two systematic causes of variation:

1) Differences in the amount sequenced among individuals

2) More reads from a long transcript than from a short transcript



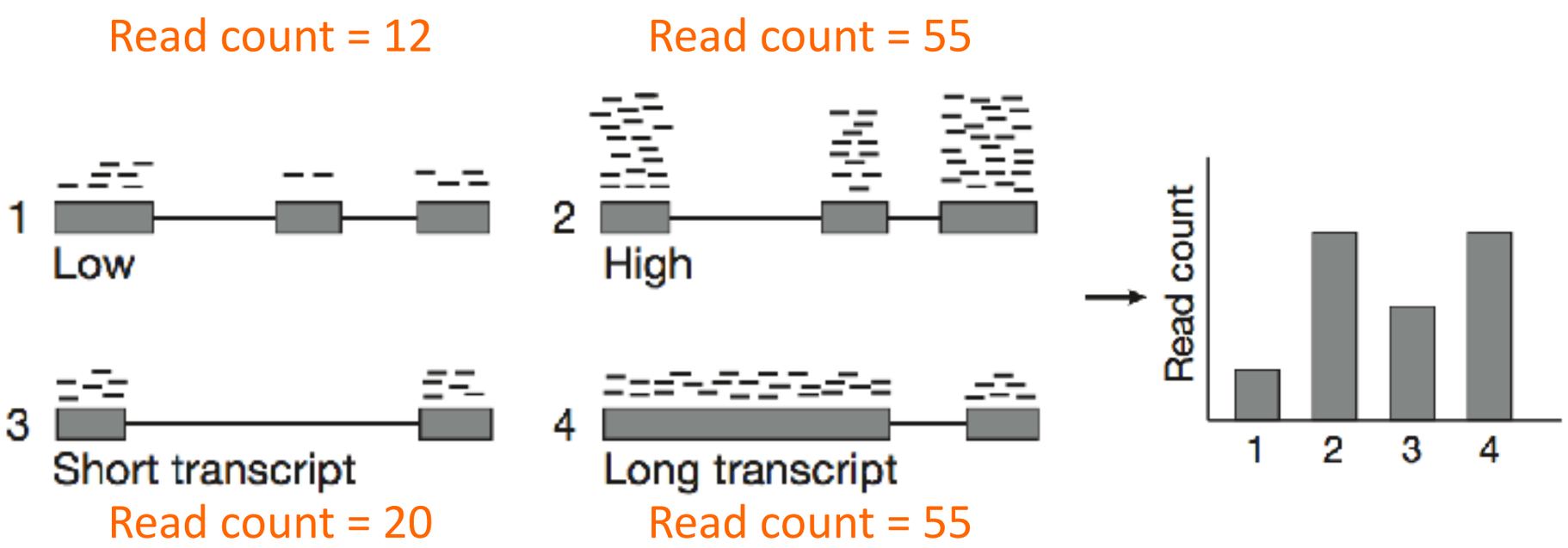
Garber et al. 2011

4. Quantifying gene expression

RNAseq normalization needed due to two systematic causes of variation:

1) Differences in the amount sequenced among individuals

2) More reads from a long transcript than from a short transcript

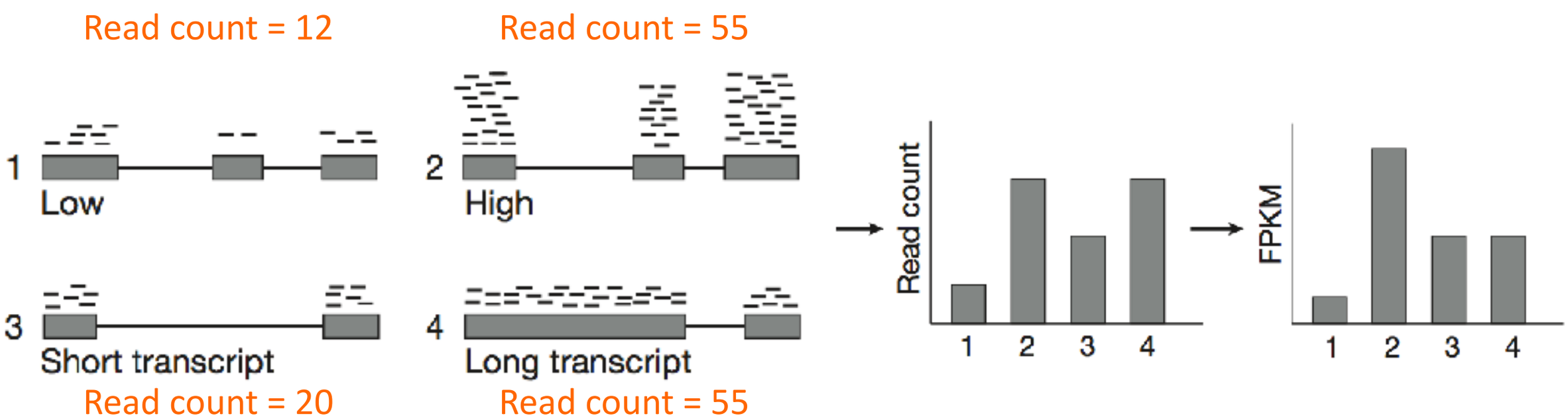


Garber et al. 2011

4. Quantifying gene expression

FPKM: **F**ragments **P**er **K**ilobase of transcript per **M**illion reads mapped

- Normalizes by transcript length and the total size of the mapped library
- Corrects both issues
- BUT - not to be used for differential expression analysis!



Garber et al. 2011

4. Quantifying gene expression

FPKM: **F**ragments **P**er **K**ilobase of transcript per **M**illion reads mapped

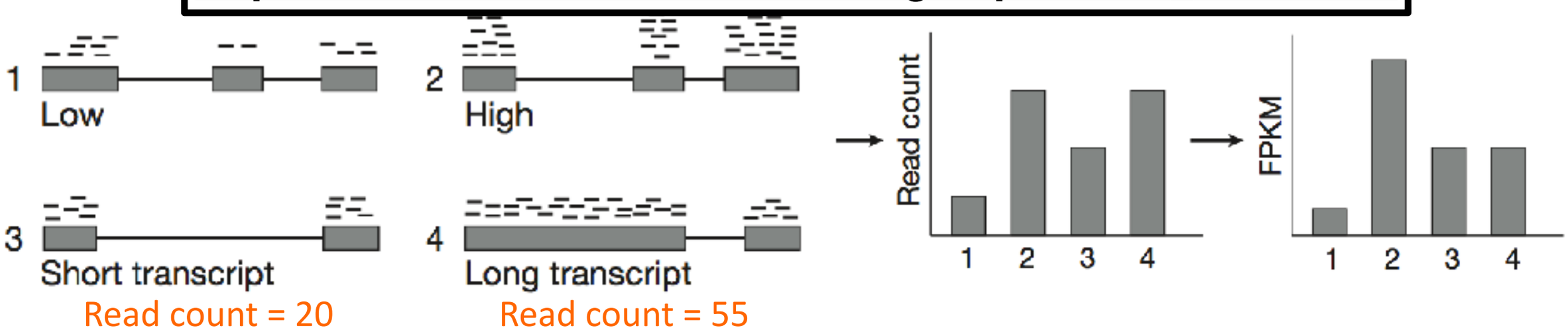
- N
  n
- C

Read c

BUT! FPKM is not good for comparing across samples

There are lots of units that have been proposed for measuring gene expression

Here's a good overview of the commonly used ones:
https://www.reneshbedre.com/blog/expression_units.html



Garber et al. 2011

4. Quantifying gene expression

FPKM: **F**ragments **P**er **K**ilobase of transcript per **M**illion reads mapped

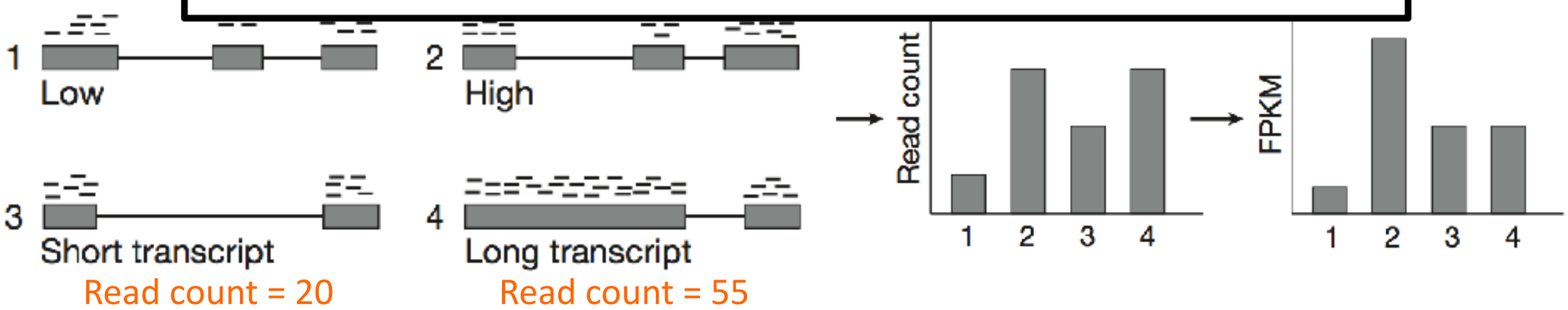
- N
  n
- C

DESeq/DESeq2 propose a method for obtaining normalised counts

A nice walkthrough of the DESeq2 method is available here:

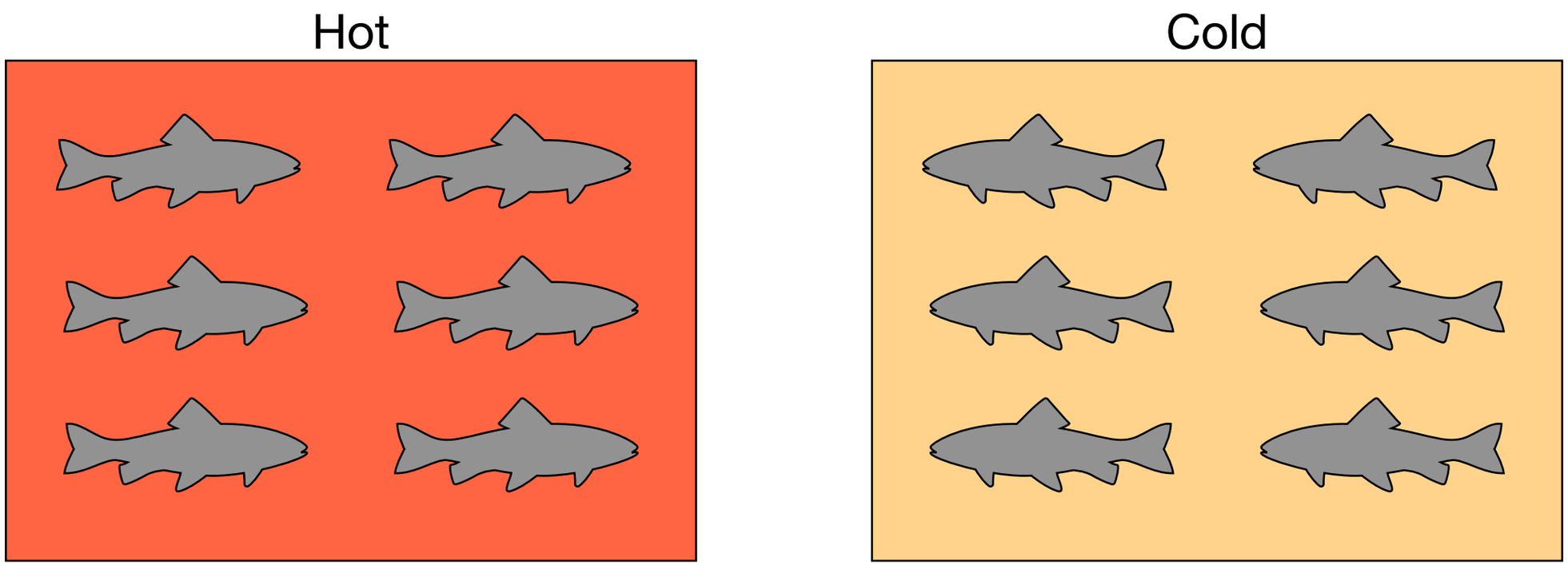
https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Read c



Garber et al. 2011

Tutorial: Align reads and measure gene expression for fish from the two environments



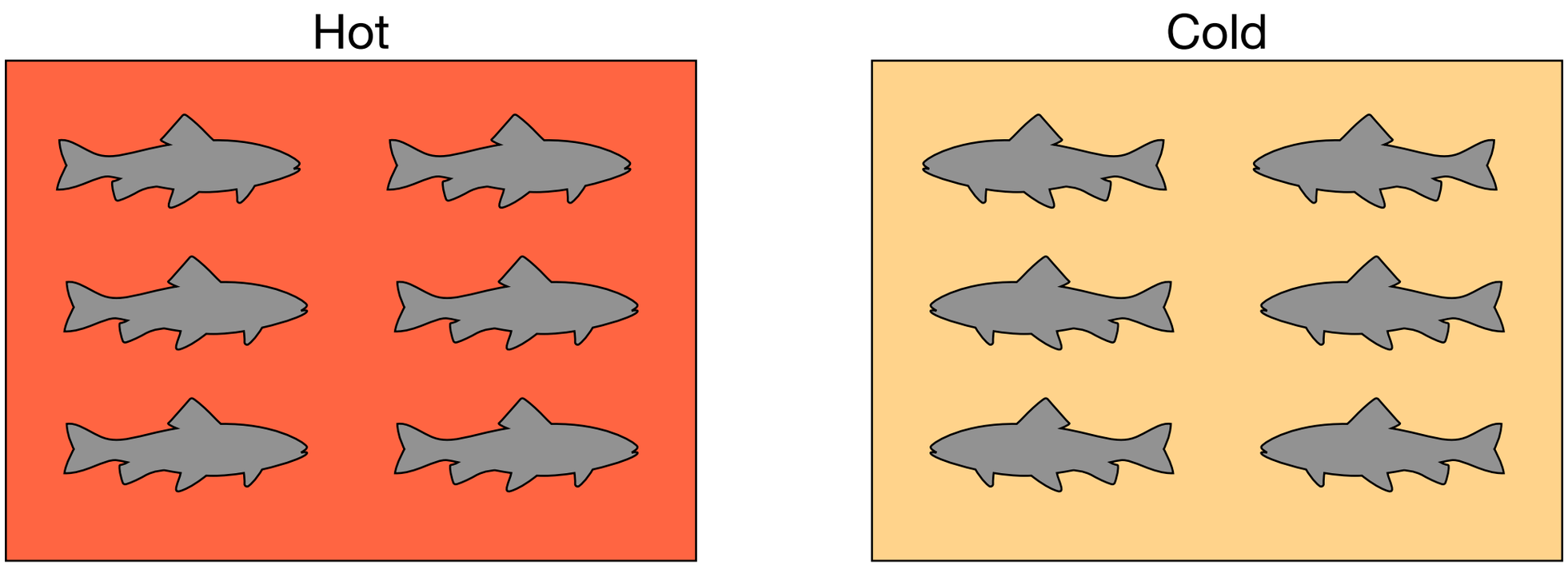
6 individuals per treatment (1 library/ind)

How is RNAseq data generated?

<u>Overview of the methods</u>

1. RNA extraction protocol and sequencing
2. Clean and filter reads
3. Map reads to a reference (genome or transcriptome)
4. Quantifying gene expression
5. **Statistical analysis of differences in read counts**

Tutorial: Align reads and measure gene expression for fish from the two environments



6 individuals per treatment (1 library/ind)

What genes are differentially expressed in response to temperature?

Analyzing patterns of expression

How to go from expression counts

comp10109_c2	0.00	0.00	0.00	0.00
comp10109_c20	0.00	0.00	0.00	0.00
comp10109_c22	176.00	13.00	5.00	9.00
comp10109_c23	0.00	0.00	0.00	0.00
comp10109_c25	0.00	0.00	2.00	2.00
comp10109_c31	0.00	0.00	0.00	0.00
comp10109_c32	0.00	0.00	0.00	0.00
comp10109_c33	1.00	0.00	0.00	0.00
comp10109_c35	148.00	403.87	327.20	117.14
comp10109_c36	0.00	0.00	0.00	0.00
comp10109_c37	0.00	0.00	0.00	0.00
comp10109_c38	1.00	1.00	0.00	0.00
comp10109_c40	0.00	0.00	0.00	0.00
comp10109_c41	96.00	51.00	61.00	24.00
comp10109_c42	15.00	0.00	0.00	1.00
comp10109_c7	0.00	0.00	0.00	0.00
comp1010_c0	483.00	2125.91	2397.11	526.00

To biologically meaningful results?

Analyzing patterns of expression

Approaches to analysis:

1. Differential gene expression on gene-by-gene basis (e.g. DESeq, EdgeR, limma)
   - Examine how each gene is affected by a factor (e.g. treatment)
   - Use glms to identify genes with significant expression differences among groups

2. Patterns of gene co-expression
   - Identify clusters of genes that are regulated together

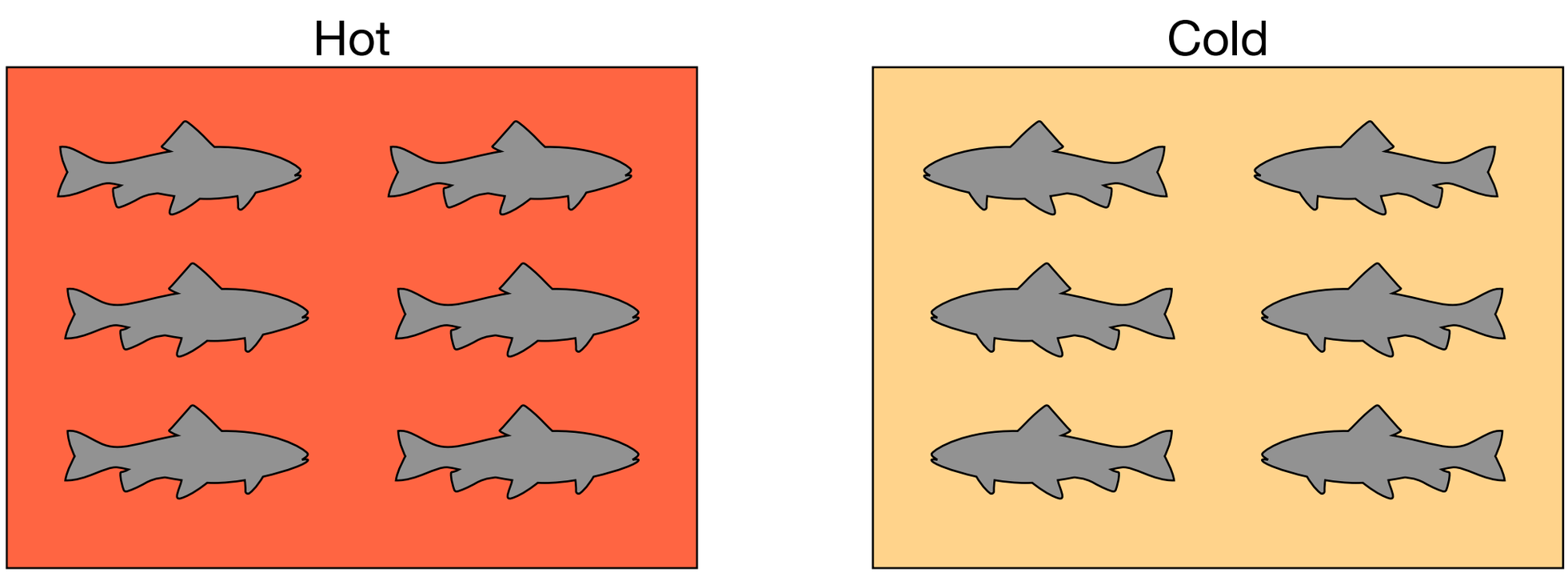
Analyzing patterns of expression

Biological variation

Real differences between samples due to:

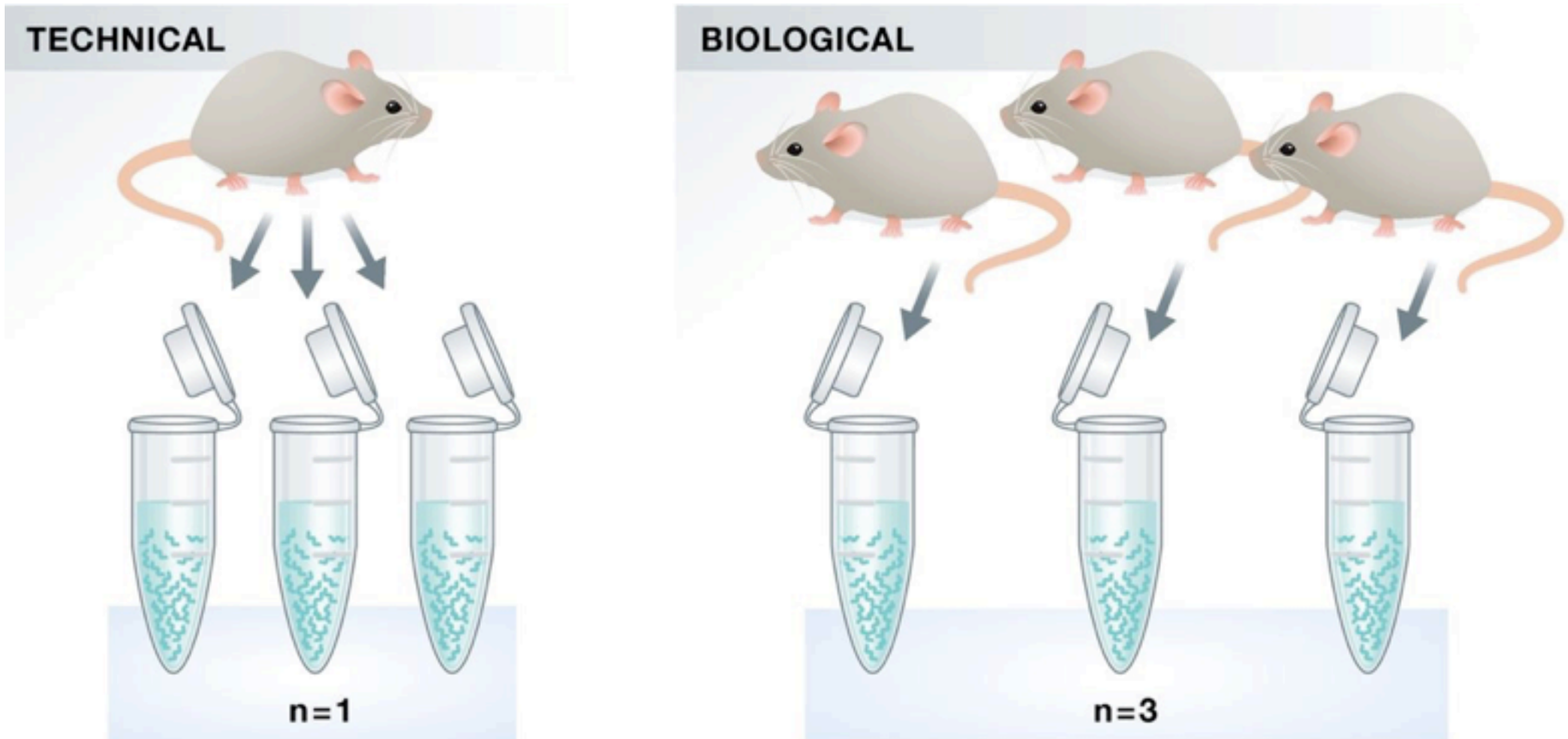
1) Uncontrolled sources (e.g. genetic background and/or cell type) - hopefully homogenous across treatments

2) Controlled sources that arise from experimental treatment/ design (e.g. hot v. cold below)



Analyzing patterns of expression

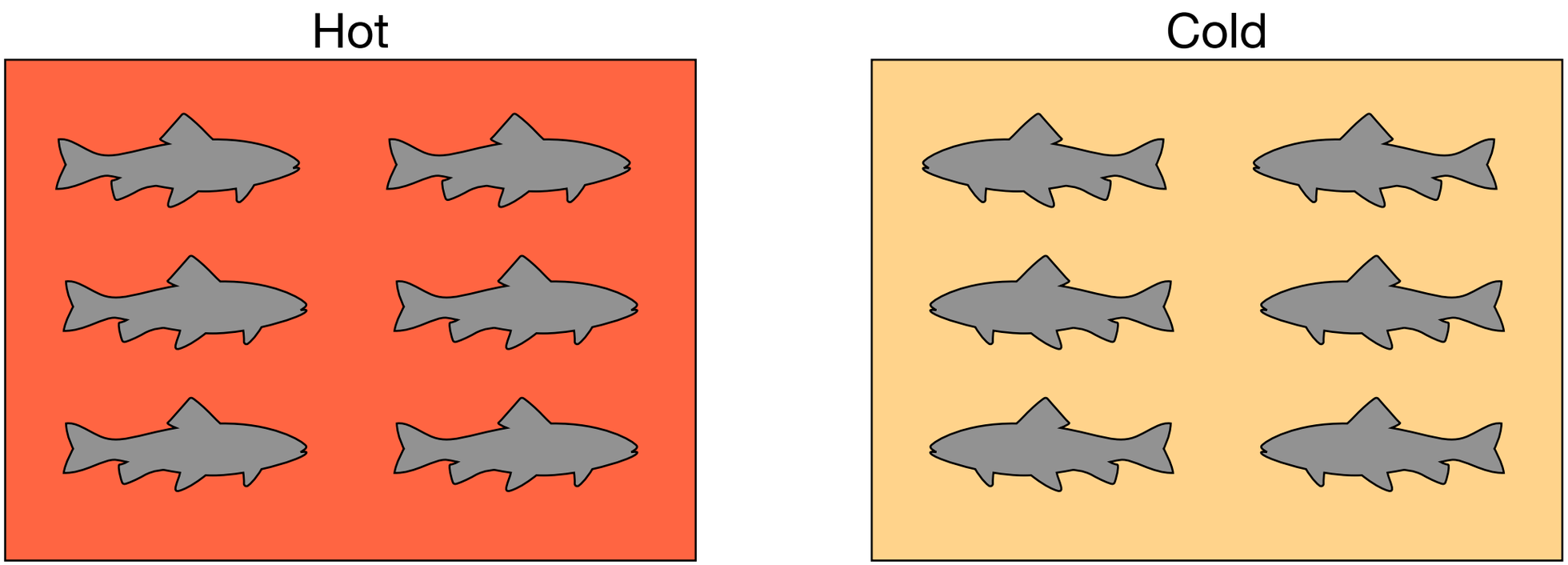
Technical variation



Technical variation is less important than biological variation in RNA seq, but still something to be aware of

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

Analyzing patterns of expression



- Biological replication (6 individuals per treatment)
- Technical replication (here, there is no technical replication)

Regression of normalized counts on variable(s) of interest

- fold-change in expression among factor levels (log2(Hot/Cold))
- estimates of significance

Who were the best batters?

The worst players in history?

Name	Home Runs	At Bats	Average
Frank Abercrombie	0	4	0.0
Horace Allen	0	7	0.0
Pete Allen	0	4	0.0
Walter Alston	0	1	0.0
Bill Andrus	0	9	0.0

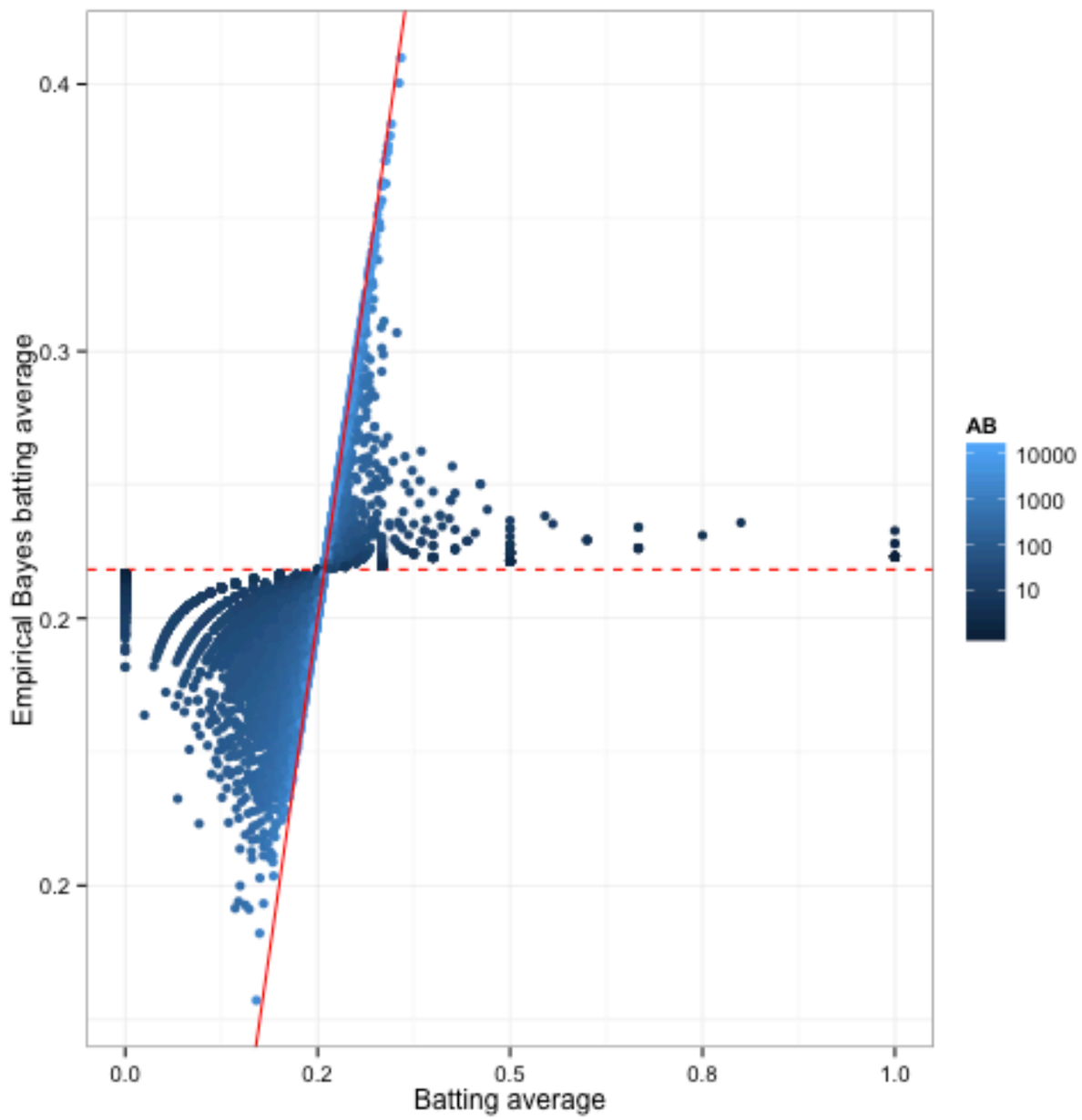
The best players in history?

Name	Home Runs	At Bats	Average
Jeff Banister	1	1	1.0
Doc Bass	1	1	1.0
Steve Biras	2	2	1.0
C. B. Burns	1	1	1.0
Jackie Gallagher	1	1	1.0

I know less about baseball than I do about working in a lab

Who were the best batters?

In empirical Bayes analyses, you use the data itself to generate a prior



Points close to the 1:1 line have lots of data

Lots of data = a better estimate of the batting average

http://varianceexplained.org/r/empirical_bayes_baseball/

Analyzing patterns of expression

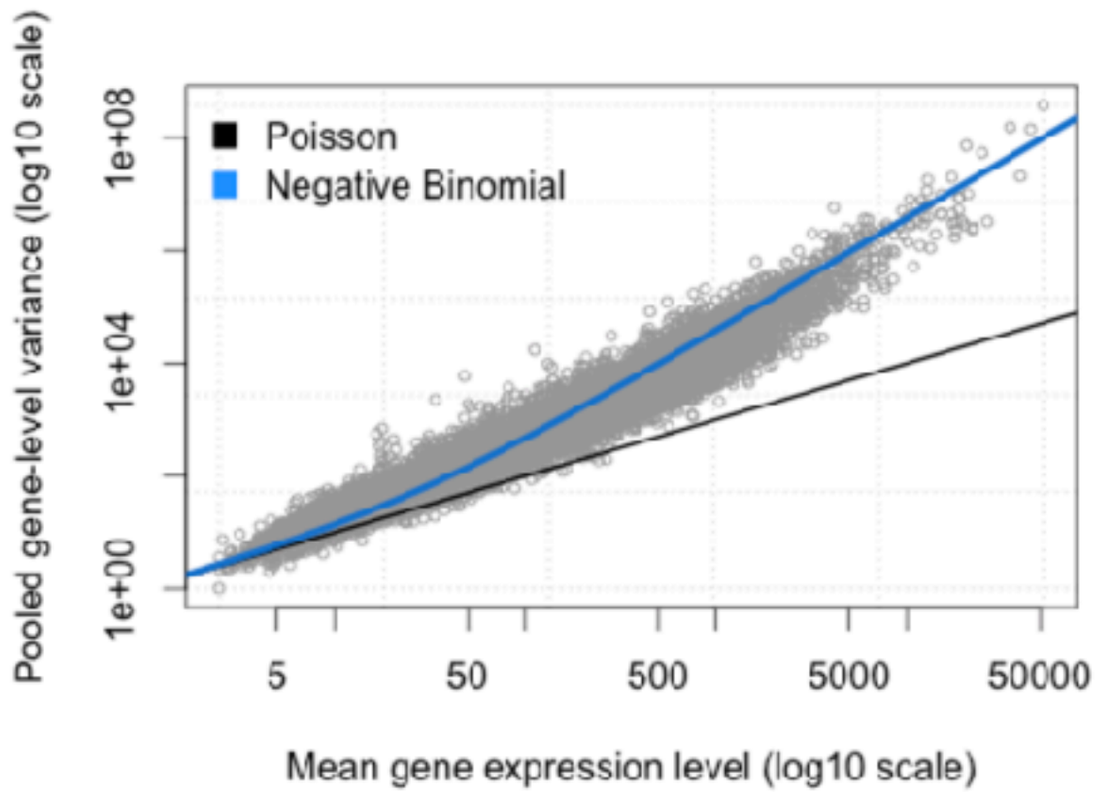
Gene	Treatment 1		Treatment 2	
	Sample 1	Sample 2	Sample 3	Sample 4
gene_A	10	20	16	14
gene_B	0	3	1	5
gene_C	32	41	11	8
gene_D	1	1	0	0

Analyzing patterns of expression

Gene	Treatment 1		Treatment 2	
	Sample 1	Sample 2	Sample 3	Sample 4
gene_A	10	20	16	14
gene_B	0	3	1	5
gene_C	32	41	11	8
gene_D	1	1	0	0

Analyzing patterns of expression

Read count data could potentially be modelled using the Poisson distribution (where mean=variance)



Biological variance creates over-dispersion so the mean does not equal the variance

The negative binomial is often used to model gene expression

Analyzing patterns of expression

An overview of one particularly common differential expression method
DESeq2 - (>20,000 citations)

Start with a set of normalised counts for each sample

Gene	Treatment 1		Treatment 2		Mean of normalised counts
	Sample 1	Sample 2	Sample 3	Sample 4	
gene_A	10	20	16	14	15
gene_B	0	3	1	5	2.25
gene_C	32	41	11	8	23
gene_D	1	1	0	0	0.5

These normalised counts are calculated from the raw read counts

See the following link for a detailed walkthrough:

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Analyzing patterns of expression

An overview of one particularly common differential expression method
DESeq2 - (>20,000 citations)

Then use a GLM of reads counts per gene on treatment and estimate dispersion

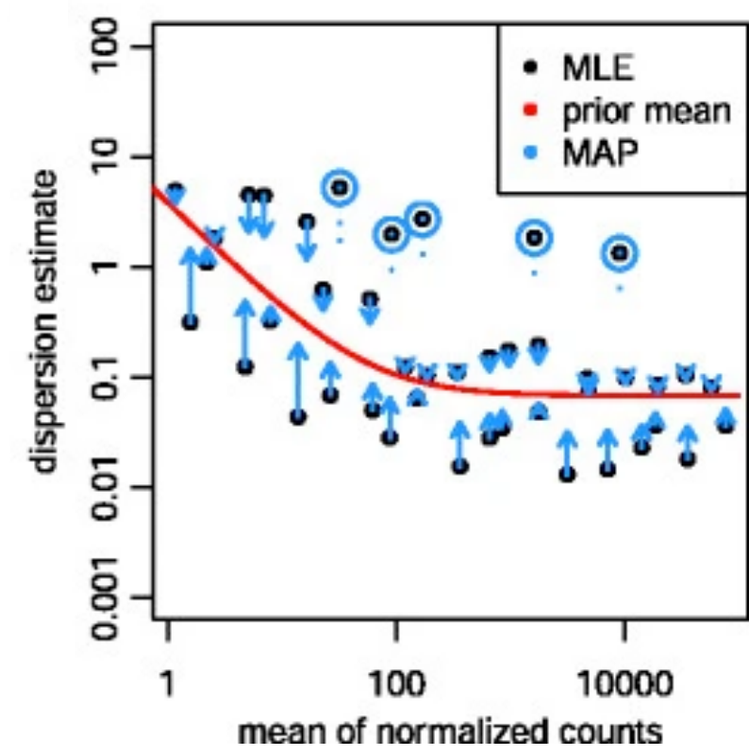
Gene	Treatment 1		Treatment 2		Mean of normalised counts	MLE of dispersion
	Sample 1	Sample 2	Sample 3	Sample 4		
gene_A	10	20	16	14	15	0.01
gene_B	0	3	1	5	2.25	0.1
gene_C	32	41	11	8	23	0.01
gene_D	1	1	0	0	0.5	1

Analyzing patterns of expression

An overview of one particularly common differential expression method
DESeq2 - (>20,000 citations)

Use an empirical Bayes approach to “shrink" dispersion estimates back to the *prior**

Gene	Treatment 1		Treatment 2		Mean of normalised counts	MLE of dispersion
	Sample 1	Sample 2	Sample 3	Sample 4		
gene_A	10	20	16	14	15	0.01
gene_B	0	3	1	5	2.25	0.1
gene_C	32	41	11	8	23	0.01
gene_D	1	1	0	0	0.5	1



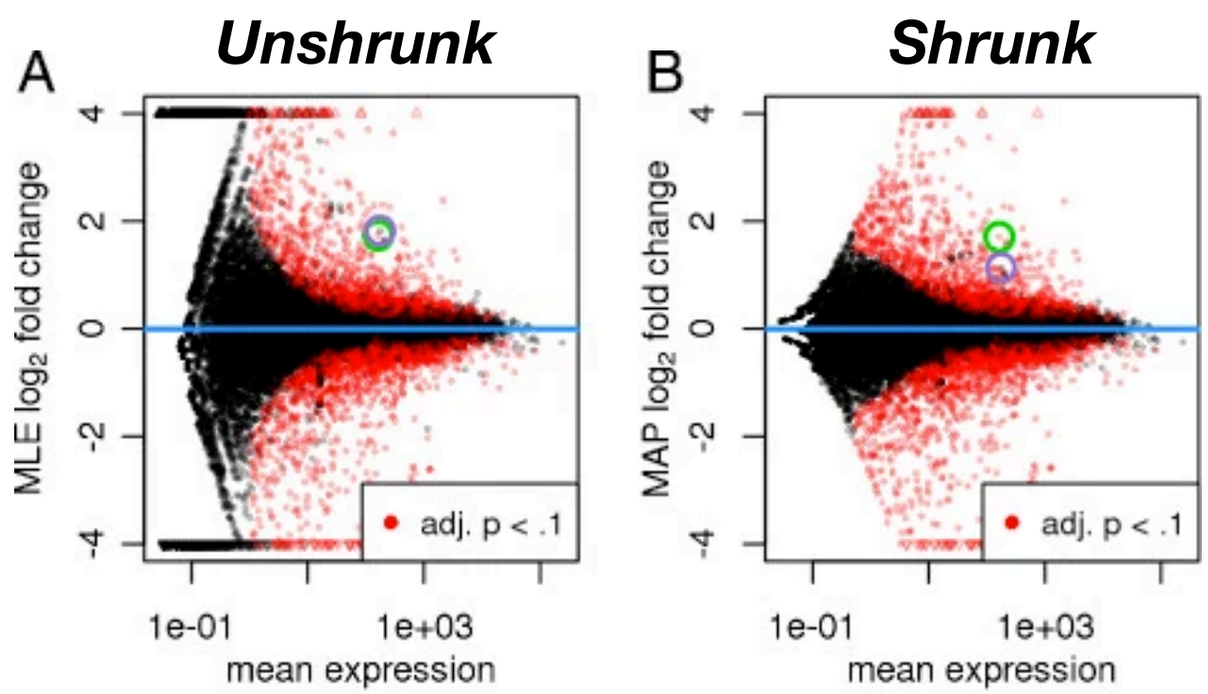
*** as inferred from all data**

Love et al 2014

Analyzing patterns of expression

An overview of one particularly common differential expression method
DESeq2 - (>20,000 citations)

The shrunken dispersion estimates for each gene are used to assess the evidence for differences in expression between treatment

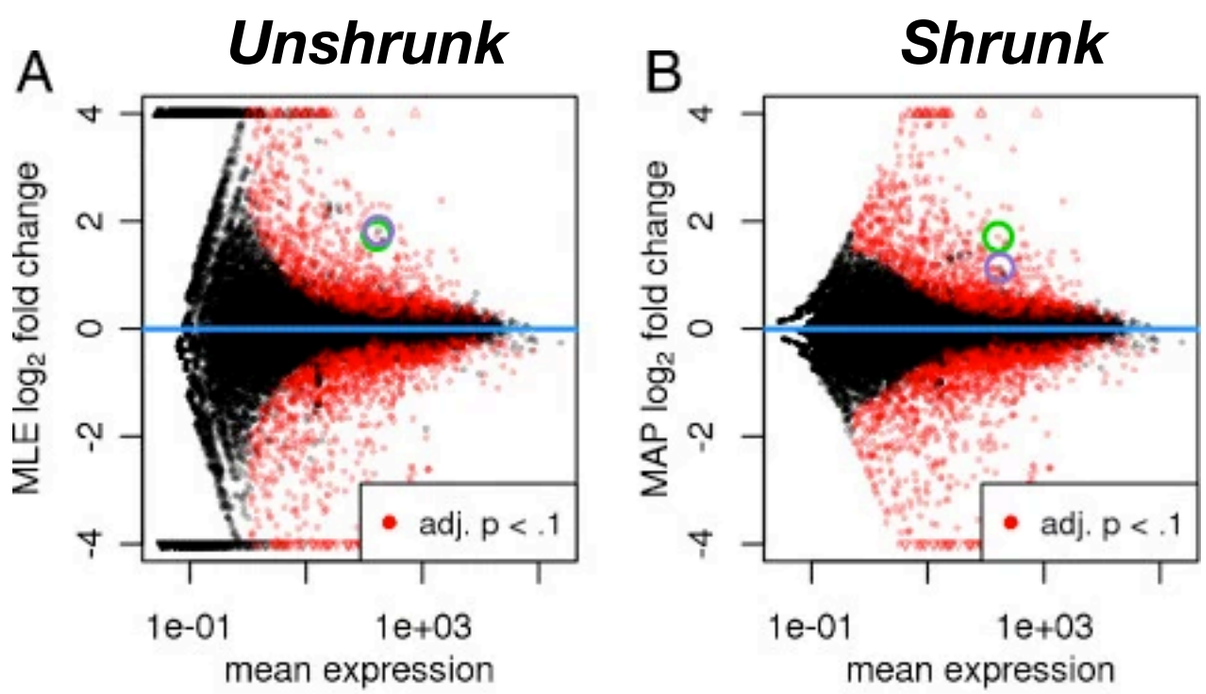


Love et al 2014

Analyzing patterns of expression

An overview of one particularly common differential expression method
DESeq2 - (>20,000 citations)

The shrunken dispersion estimates for each gene are used to assess the evidence for differences in expression between treatment



Can then identify genes with significant differences in expression

Love et al 2014

Outline

1. Introduction and background
2. Overview of the methods and workflow
3. Quantifying expression levels
4. Analyzing patterns of expression
5. **Technical considerations**

Technical considerations

Depth of coverage?

Dependent on:

1. Study organism
2. Transcriptome size
3. Purpose of your study

Low power if < 50 counts per million per gene

10 million reads per sample is a benchmark from which to start for most eukaryotes

Biological replication is often more valuable than higher depth of coverage per individual

Too many individuals per lane can increase your technical variation

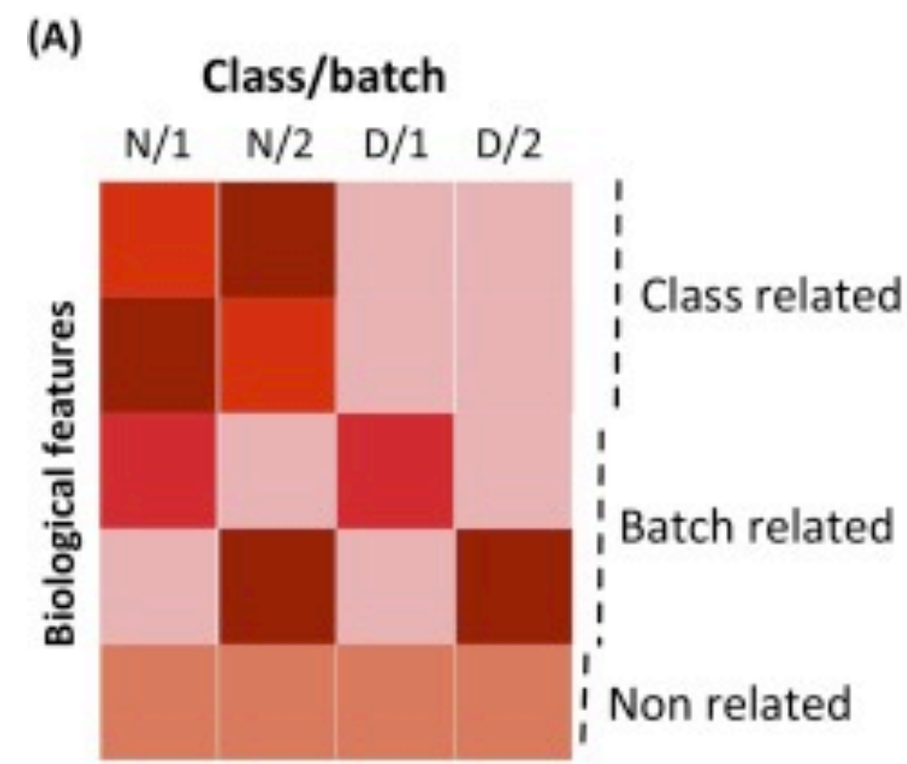
Table 1 Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

Conesa et al 2016 Genome Biology

Technical considerations

- Variation among cells of the same type sampled at the same time (single-cell sequencing)
- Variation among cell types of the same tissue (micro-dissection)
- Important that replicates be randomized during sample prep and sequencing due to batch effects (RNA extraction, library prep and sequencing).



Goh et al 2017

Technical considerations

<u>Transcriptome assembly</u>

De novo assembly from short reads needs large amounts of RAM

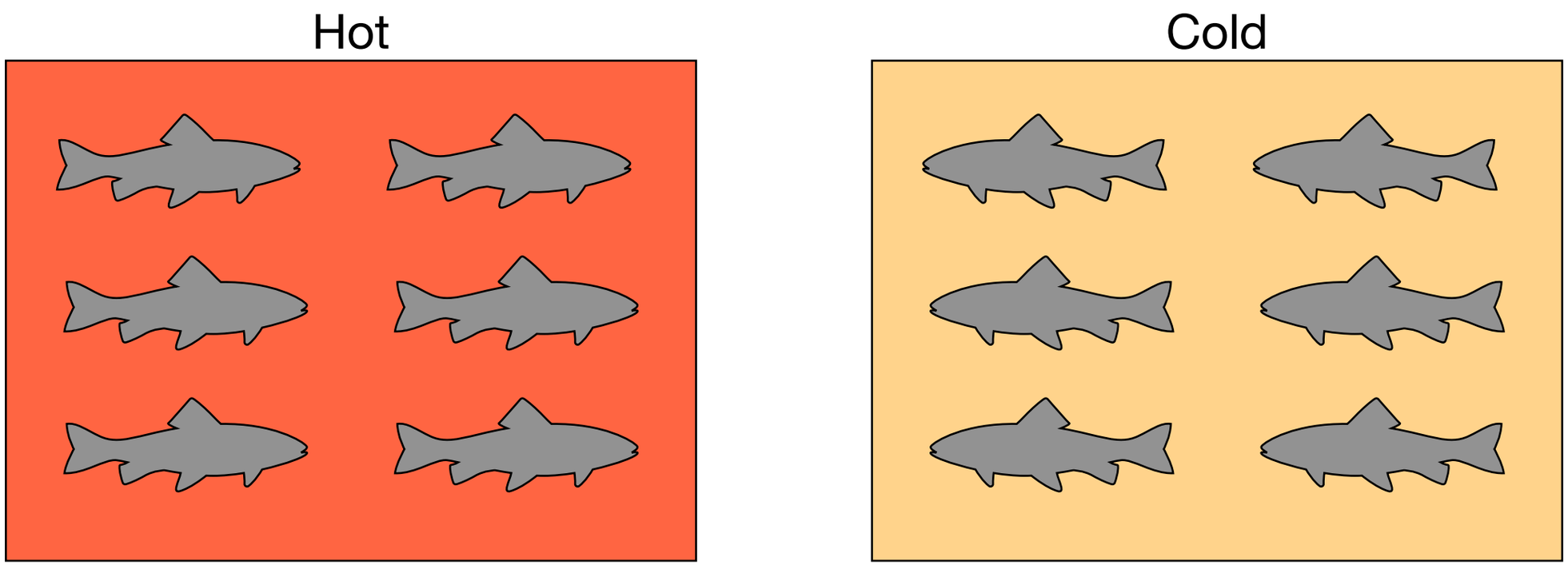
Haploid tissue from a single individual is best - no heterozygotes

- Feasible to pool data from multiple individuals but difficult to know whether putative isoforms are real or just different genotypes

Pooling from multiple tissues, treatments, developmental time points

Long read transcriptome sequencing (e.g., PacBio) is an alternative (no assembly required)

Tutorial: Analyse read counts from the fish using DESeq2



6 individuals per treatment (1 library/ind)

What genes are differentially expressed in response to temperature?

Further Reading

Baruzzo, G., Hayer, K., Kim, E. *et al.* Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat Methods* **14,** 135–139 (2017).

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, et al. 2016. A survey of best practices for RNA-seq data analysis. Genome Biology 17: 1–19.

Garber et al. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. Nature Methods. 8:469-477.

Marinov et al. 2014. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. Genome Research. 24:496–510.

Rapaport et al. 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biology. 14:R95.

Seyednasrollah et al. 2013. Comparison of software packages for detecting differential expression in RNA-seq studies. Briefings in Bioinformatics.

Tarazona et al. 2011. Differential expression in RNA-seq: A matter of depth. Genome Res. 21: 2213-2223

http://www.labome.com/method/RNA-seq-Using-Next-Generation-Sequencing.html

http://deweylab.biostat.wisc.edu/rsem/

http://www.mi.fu-berlin.de/wiki/pub/ABI/GenomicsLecture12Materials/rnaseq1.pdf

http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

http://rnaseq.uoregon.edu/