TOPIC 6:

RNA-seq and analysis of differential gene expression

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

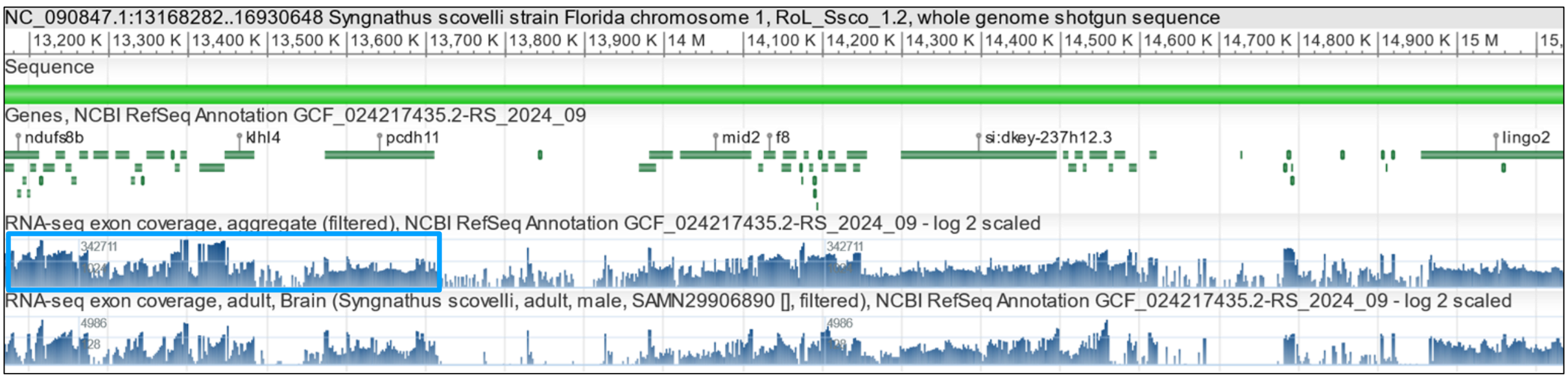
Why use RNA-seq?

```
>NC_090847.1:15049400-15049553 Syngnathus scovelli strain Florida chromosome 1, whole genome shotgun sequence
AAACAAGGAATTTGACTTCGGTAAATCACAGCCTCTGTTCAACATTTAGGTGACTAACAACAACACTCAGGACATGTGAA
GAACGAAAGATATTCTCAAACACCCCCTGATCTTAAACTCCCAAGAGGGCAAGGAAAAACTCAAAACTCCAGCT
```

Introduction and background

Why use RNA-seq?

>NC_090847.1:15049400-15049553 Syngnathus scovelli strain Florida chromosome 1, whole genome shotgun sequence
AAACAAGGAATTTGACTTCGGTAAATCACAGCCTCTGTTCAACATTTAGGTGACTAACAACAACACTCAGGACATGTGAA
GAACGAAAGATATTCTCAAACACCCCCTGATCTTAAACTCCCAAGAGGGCAAGGAAAAACTCAAAACTCCAGCT

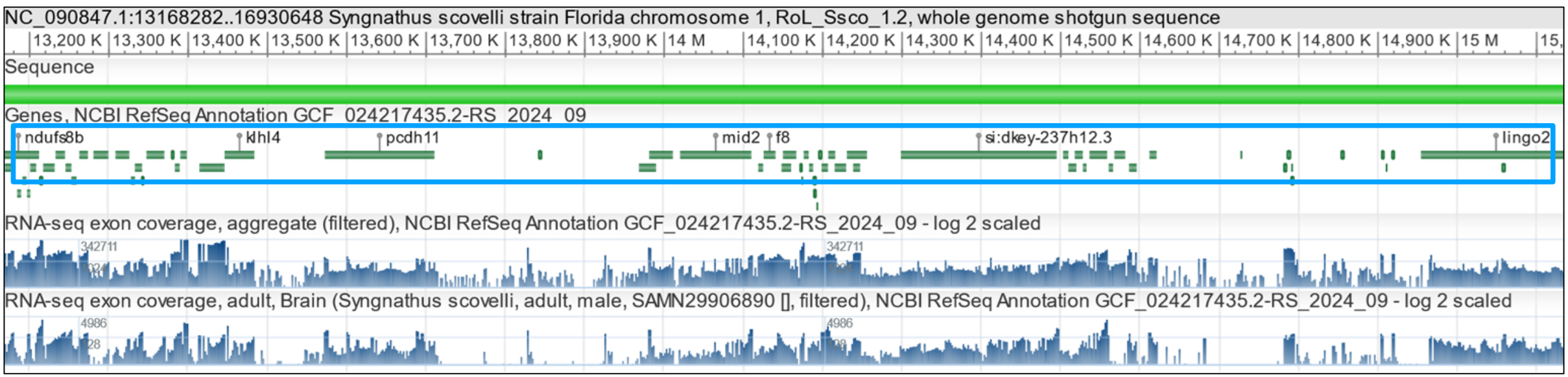


Genome annotation: Identifying transcribed regions of the genome

Introduction and background

Why use RNA-seq?

>NC_090847.1:15049400-15049553 Syngnathus scovelli strain Florida chromosome 1, whole genome shotgun sequence
AAACAAGGAATTTGACTTCGGTAAATCACAGCCTCTGTTCAACATTTAGGTGACTAACAACAACACTCAGGACATGTGAA
GAACGAAAGATATTCTCAAACACCCCCTGATCTTAAACTCCCAAGAGGGCAAGGAAAAACTCAAAACTCCAGCT

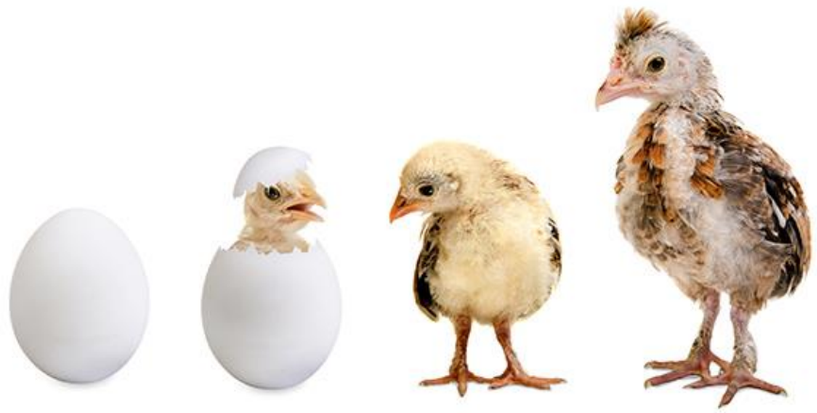


Genome annotation: Identifying transcribed regions of the genome

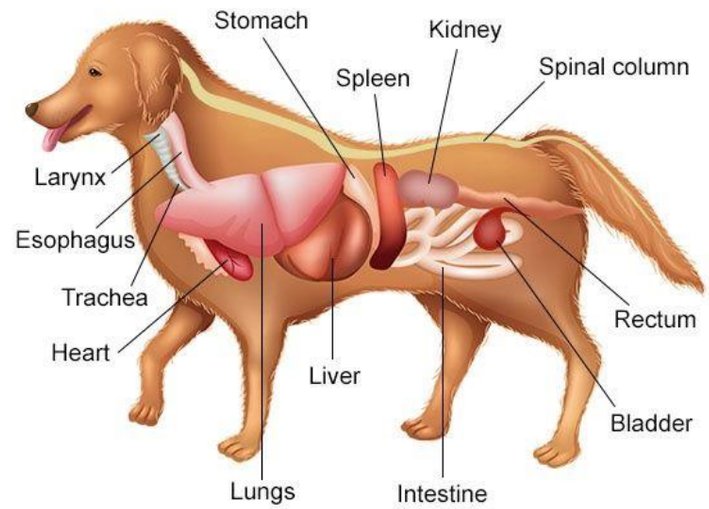
and designating locations of exon-intron boundaries (splice junctions)

Introduction and background

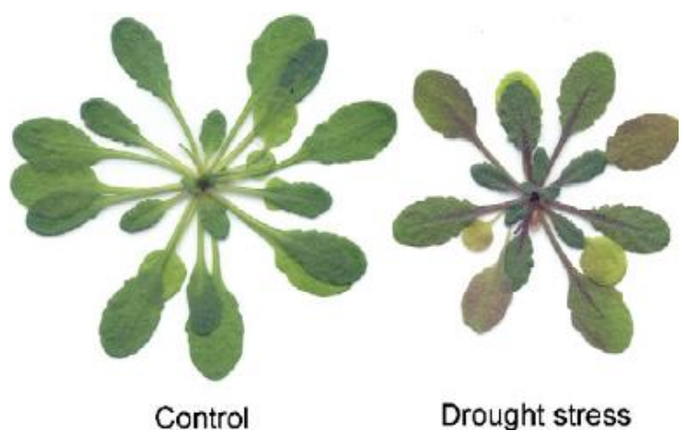
Quantifying differences in gene expression



Developmental timepoints



Different organs, tissues, or cell types



Experimental treatments

Between sexes, ecotypes, morphs

How is RNAseq data generated?

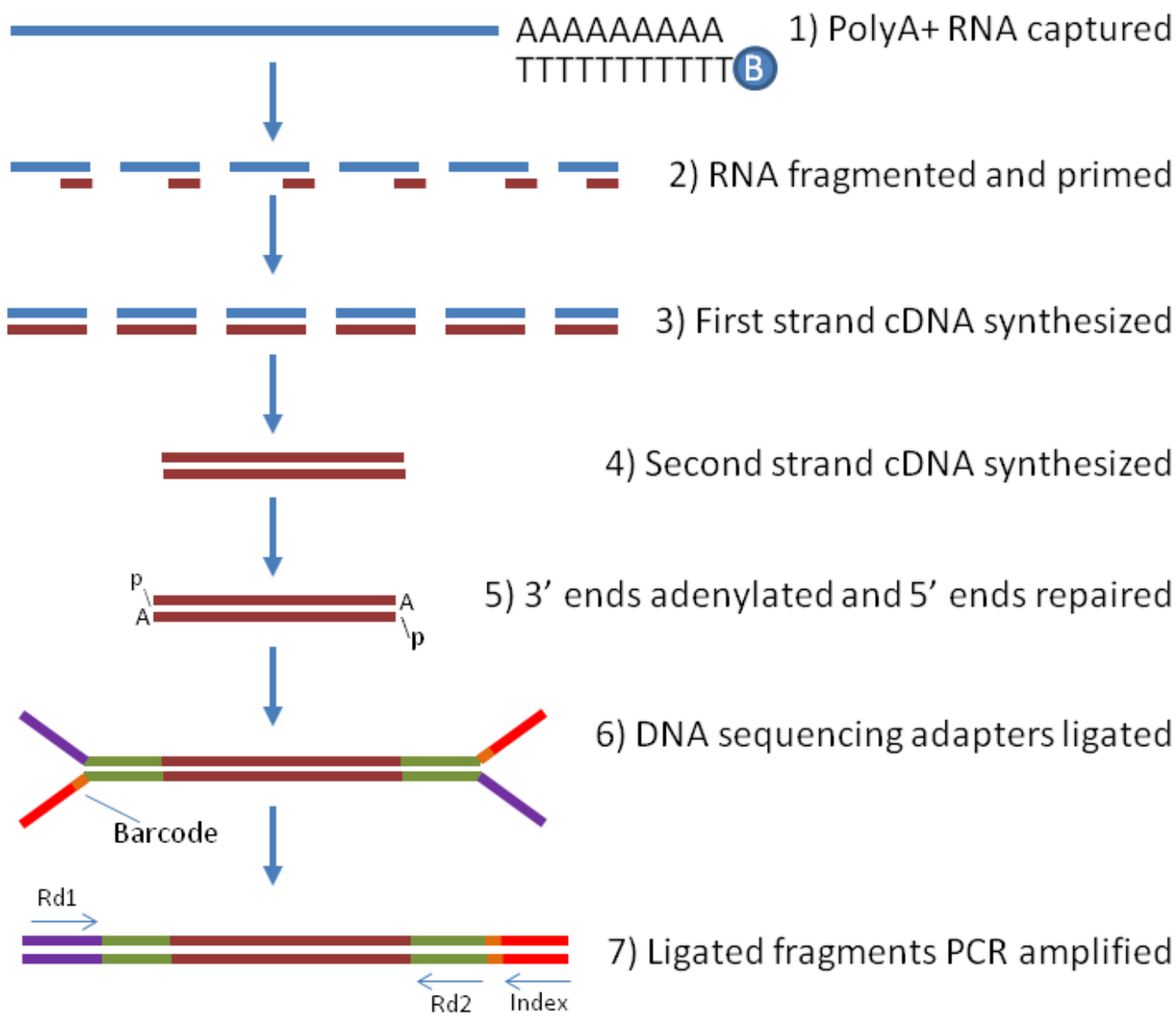
Overview of the methods

1. **RNA extraction 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 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 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

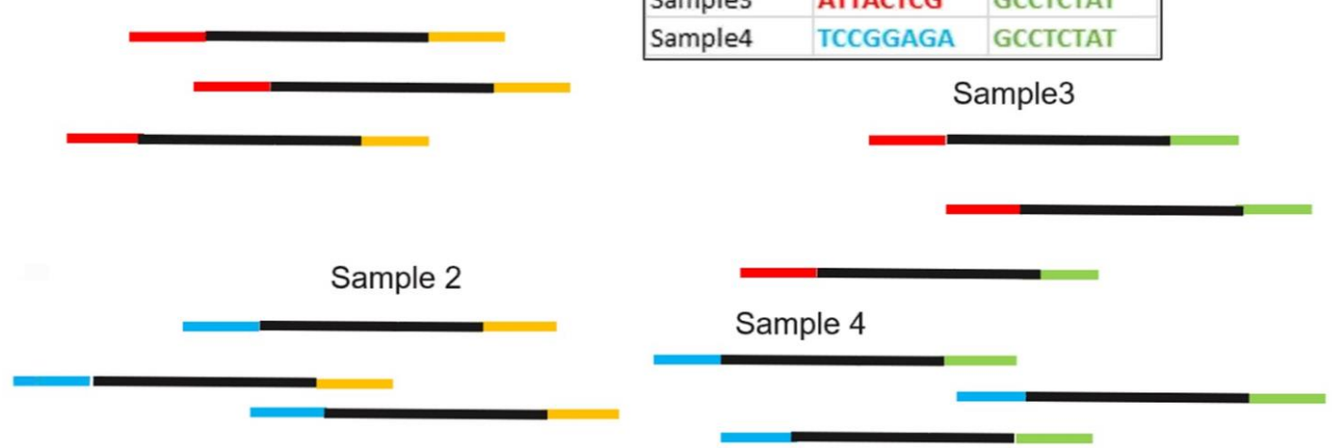
Samples that have
ed onto the
encing lane
separated

es are
hed using
dentifying
s - those
e removed
analysis

Demultiplexing

- **Assigns clusters to a sample, based on the cluster's index sequence which is provided in the sample sheet**

[Data]		
Sample_ID	index	index2
Sample1	ATTACTCG	AGGCTATA
Sample2	TCCGGAGA	AGGCTATA
Sample3	ATTACTCG	GCCTCTAT
Sample4	TCCGGAGA	GCCTCTAT



10

For Research Use Only. Not for use in diagnostic procedures.

illumina

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.

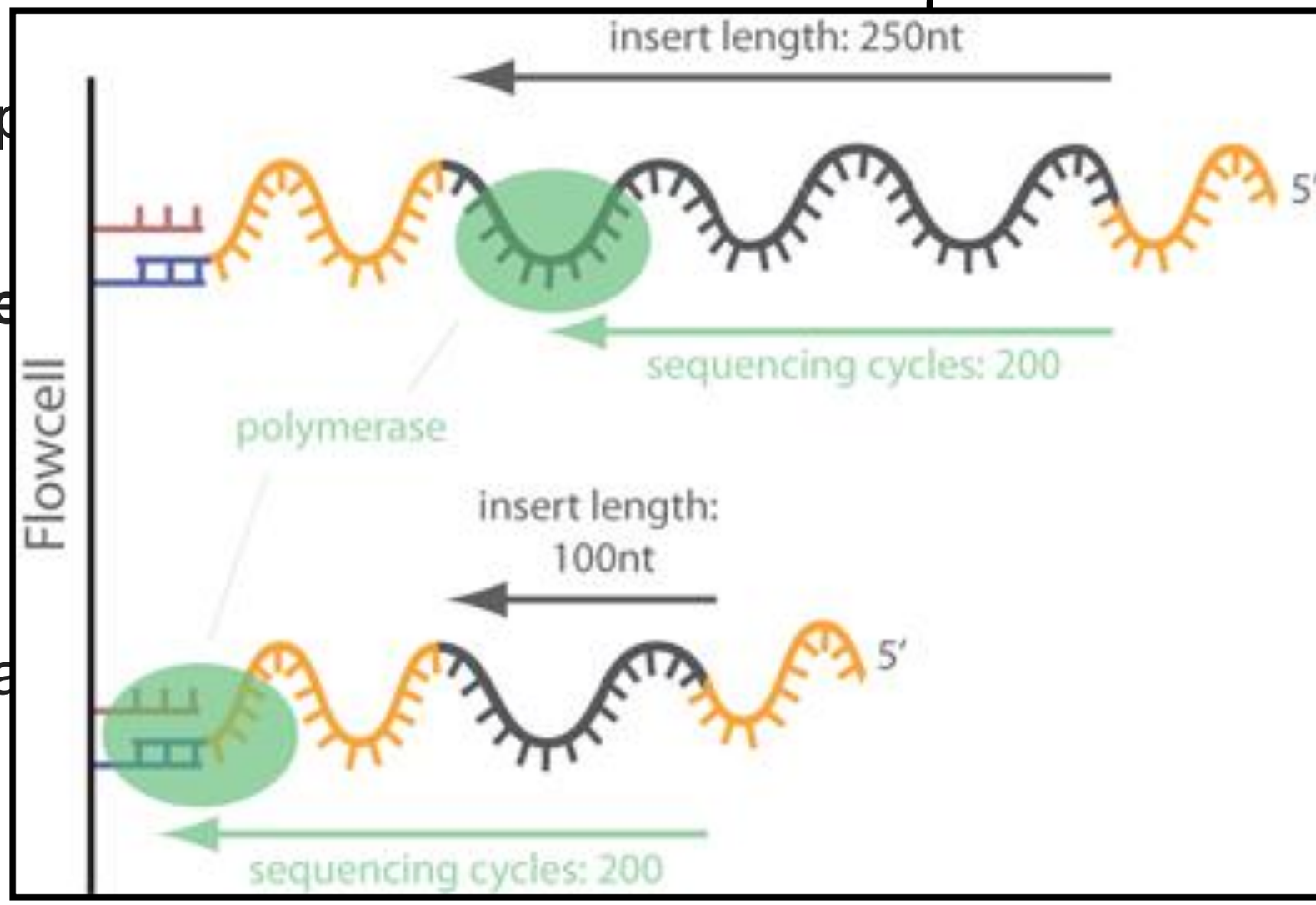
2. Clean and filter reads

A. Demultip

B. Remove

C. Discard

D. Filter rea



nts are
long,
uences
be
, but
reads
short,
s can
eing
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.

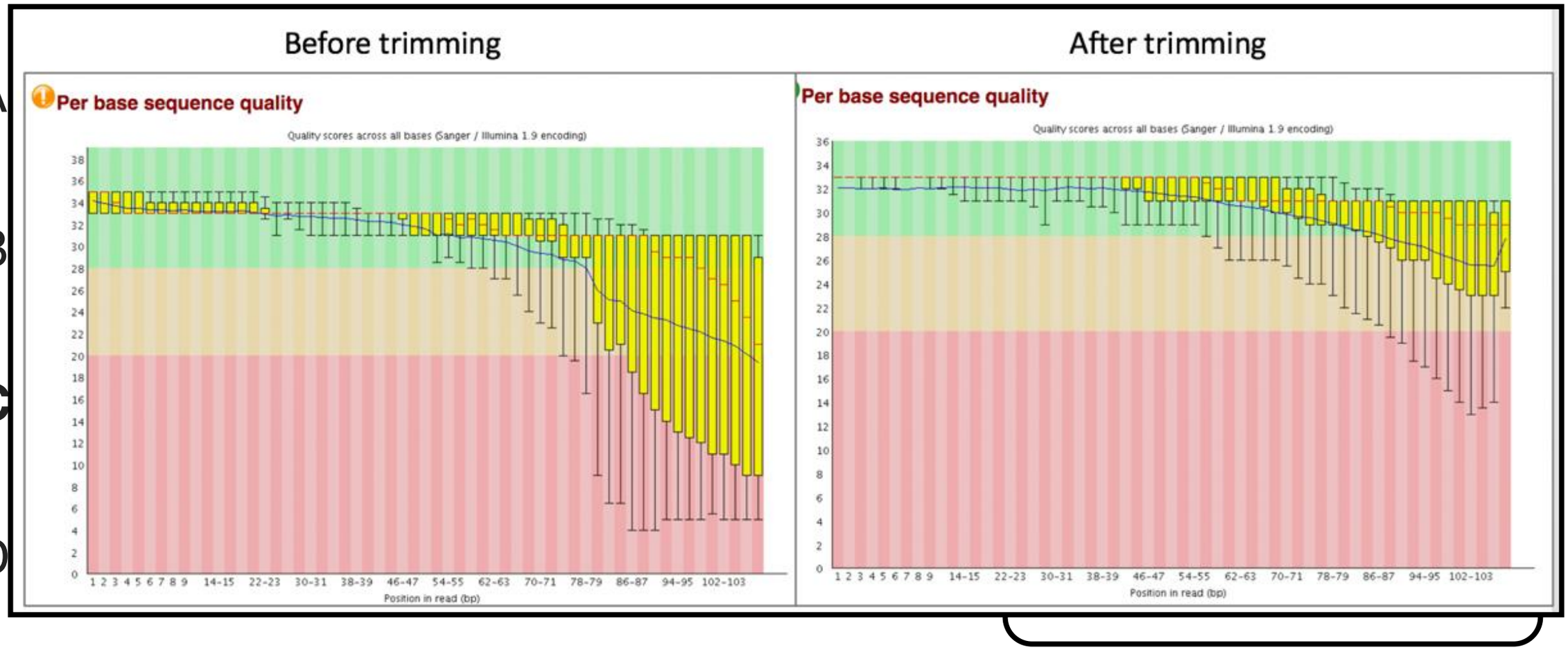
2. Clean and filter reads

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.

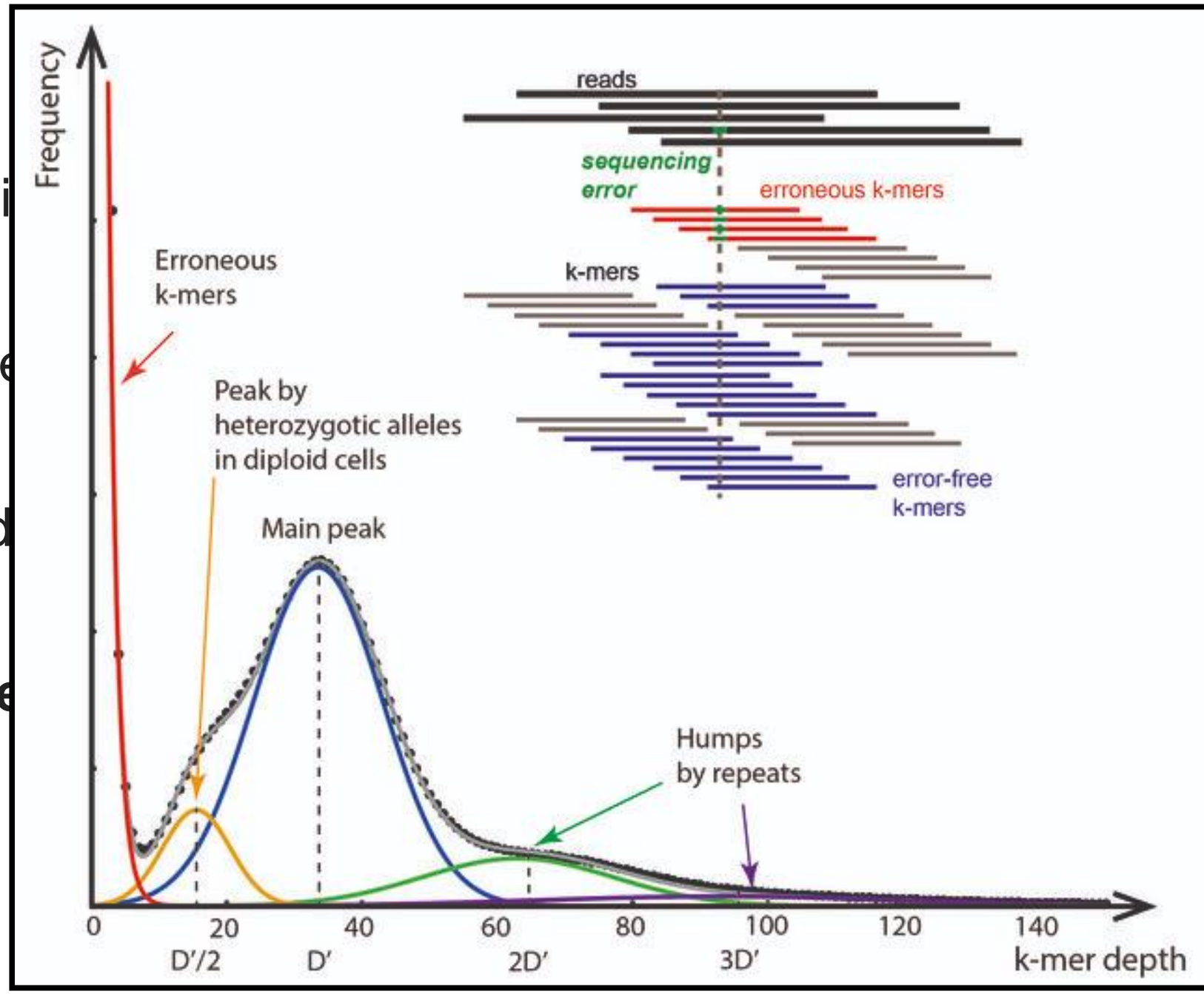
2. Clean and filter reads

A. Demulti

B. Remove

C. Discard

D. Filter re



es have
ribution of

ibution of
dicate
rrors.

rs 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	✔	✔	✔		

Which is the best?

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	✔	✔	✔		

Some trimmers rely on downstream assemblers to filter and normalize by k-mers.

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	✔	✔	✔		

Some programs are better suited for different methods. For example, Stacks was primarily designed for restriction enzyme-based data, such as RAD-seq, used in population genomics.

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	✔	✔	✔		

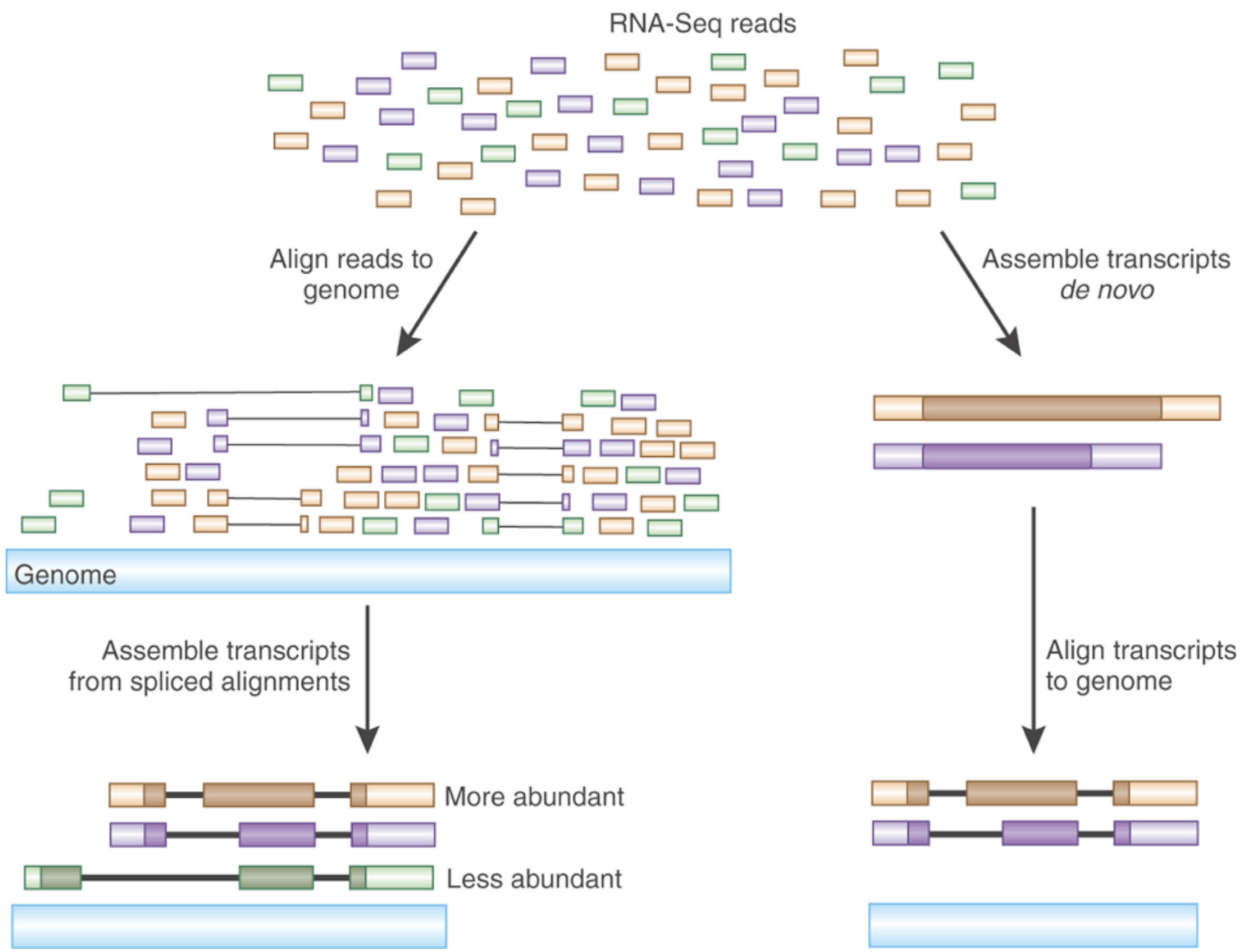
Trimmomatic is designed to handle paired-end reads generated from Illumina sequencing and is preferably implemented in RNA seq pipelines.

How is RNAseq data generated?

Overview of the methods

1. RNA extraction 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

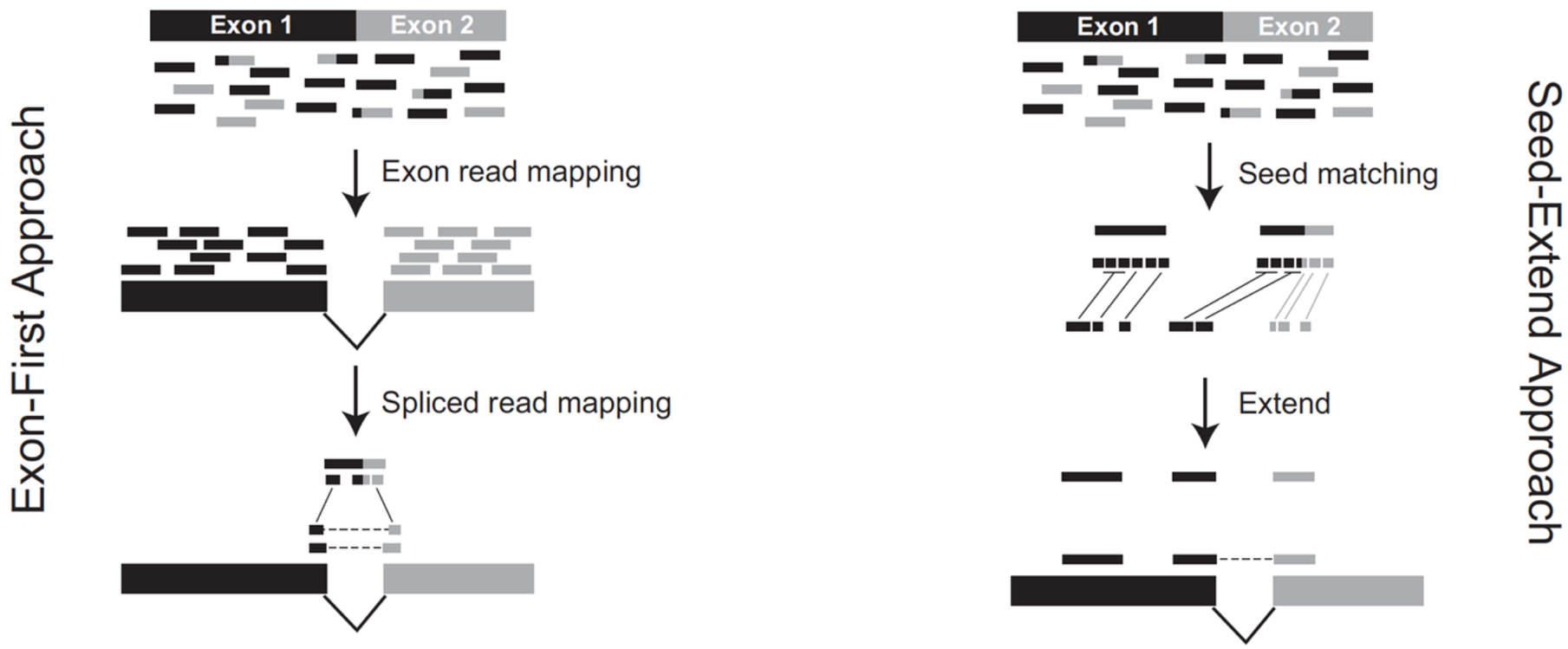
C. Dealing with multi-mapped reads

D. No reference available - *de novo* assembly

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

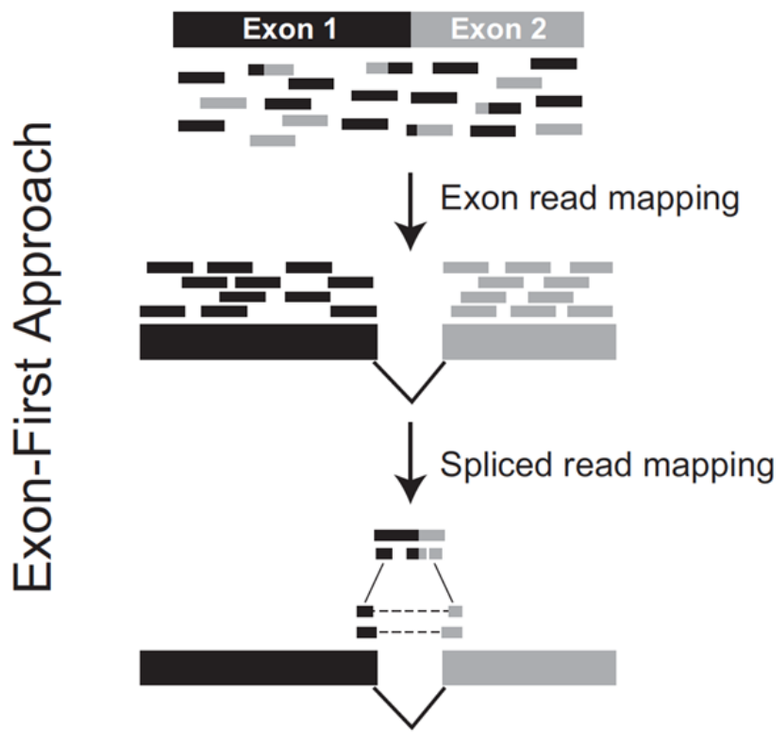


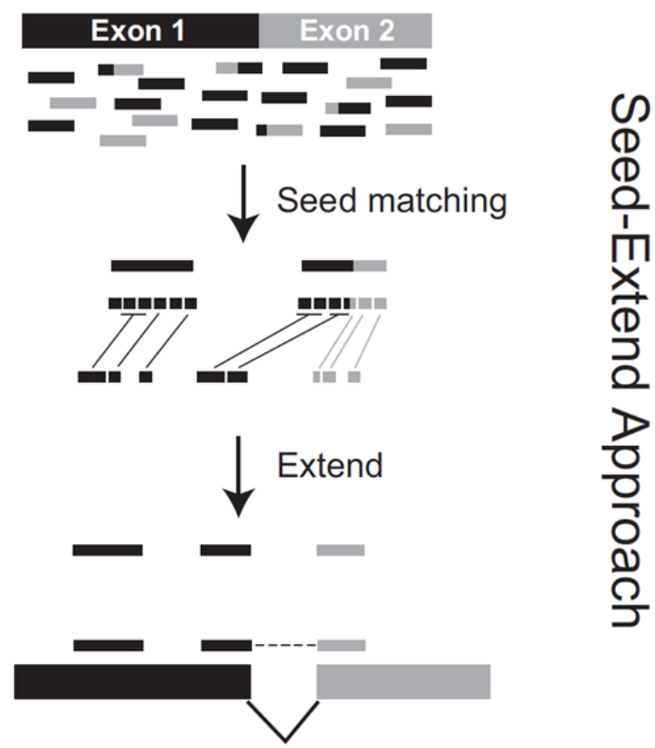
Garber et al. 2011

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





- Works well if a comprehensive transcriptome is already available = needs known transcript structures
- Less computational power
- Outdated (Ex. TopHat)

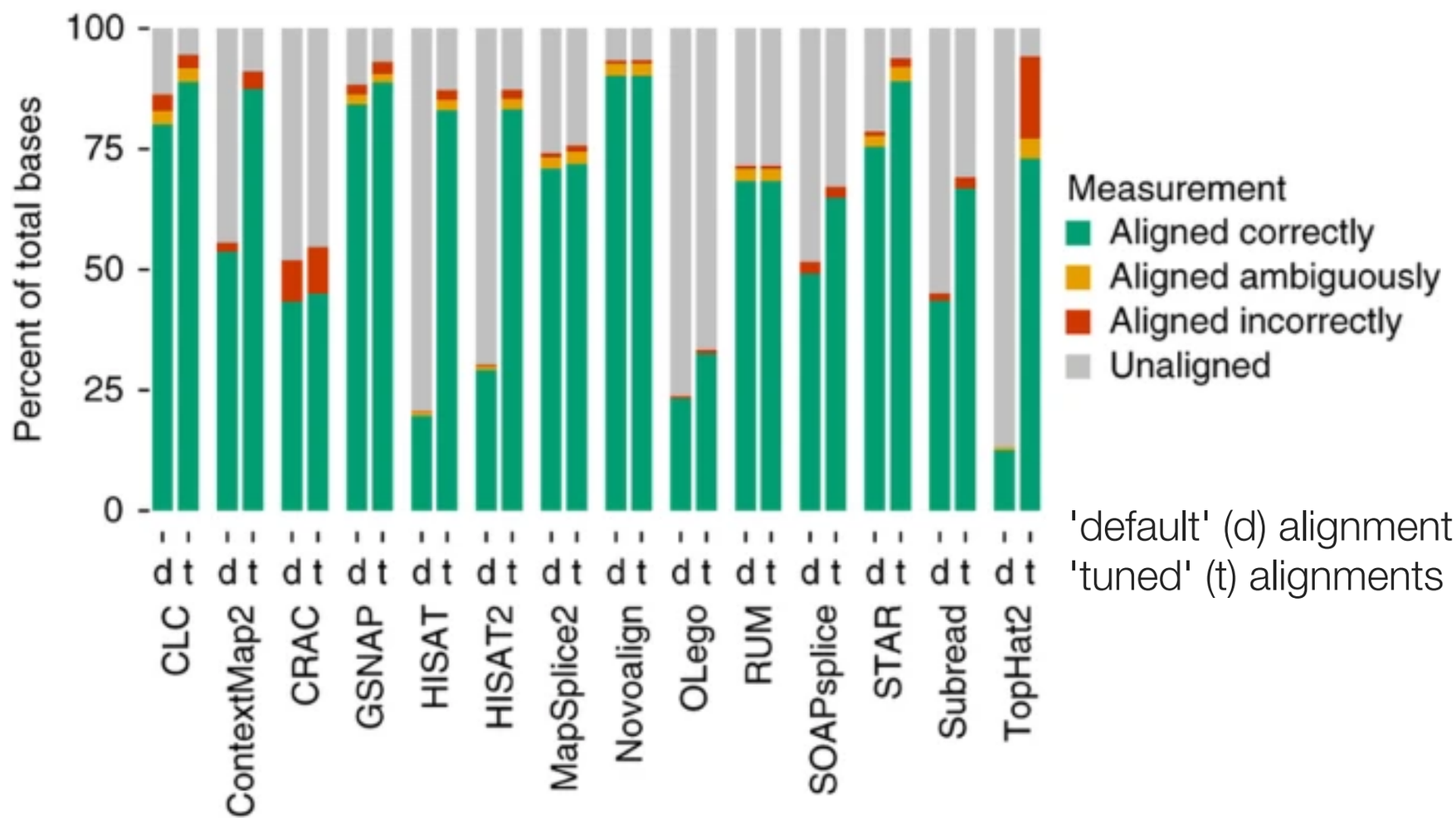
- Bridges over exon junctions
- Robust against indels
- Can be incorporated into splicing alignment algorithms (Ex. STAR)

Garber et al. 2011

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

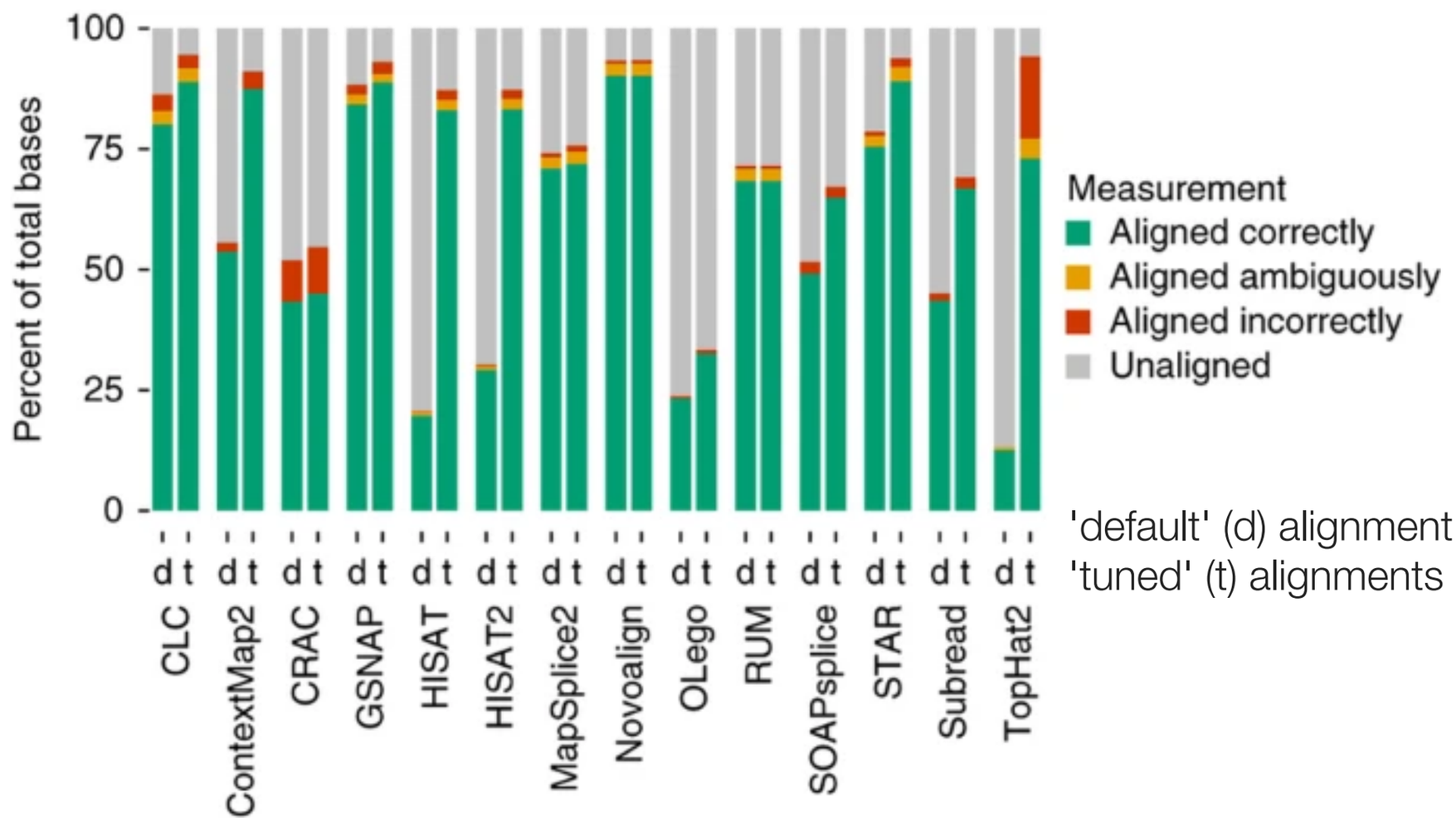


Baruzzo et al 2017

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



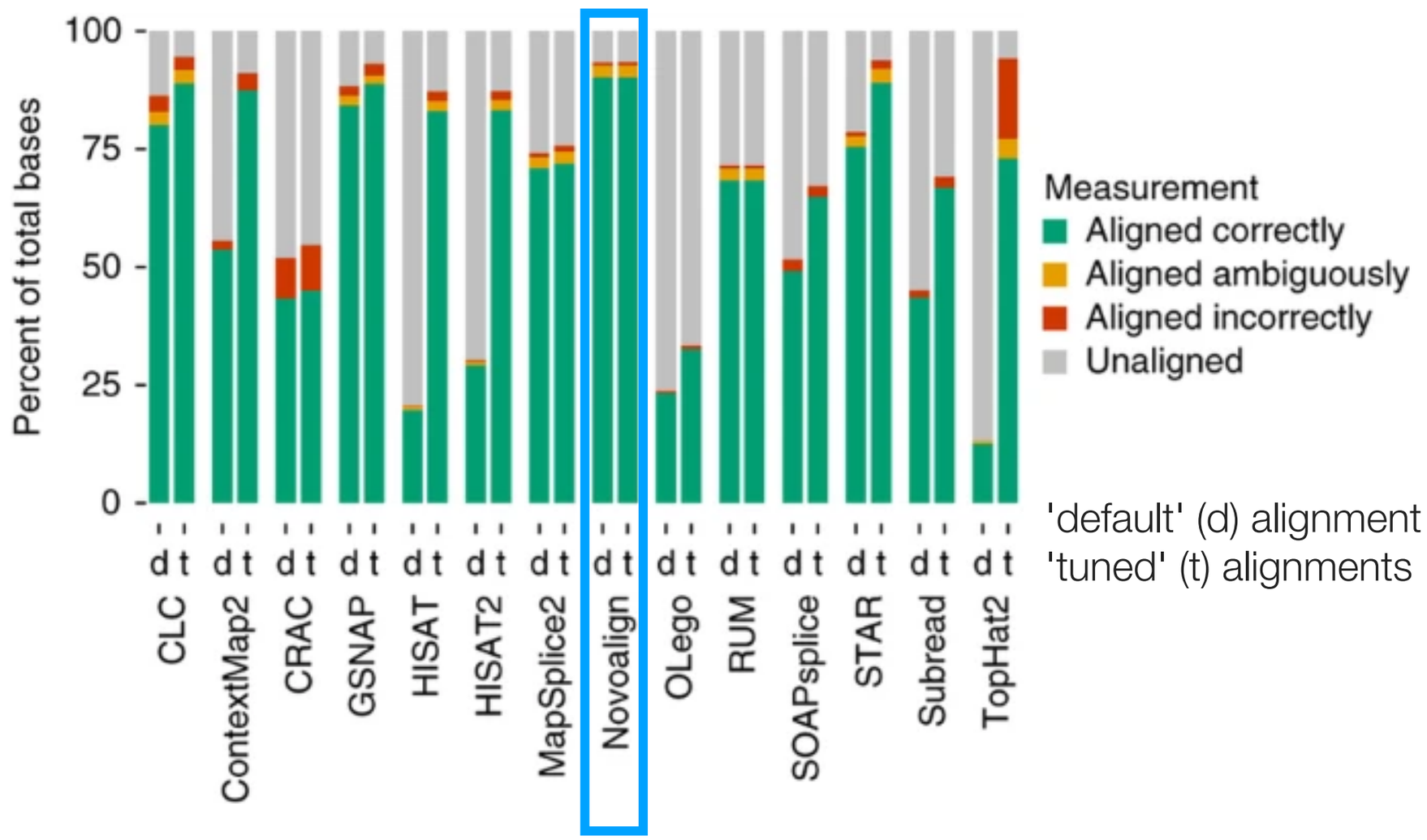
Other consideration factors: cost, compatibility with sequencing technology, ease of use, flexibility, maintained/updated, computational power needed, speed, popularity.

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



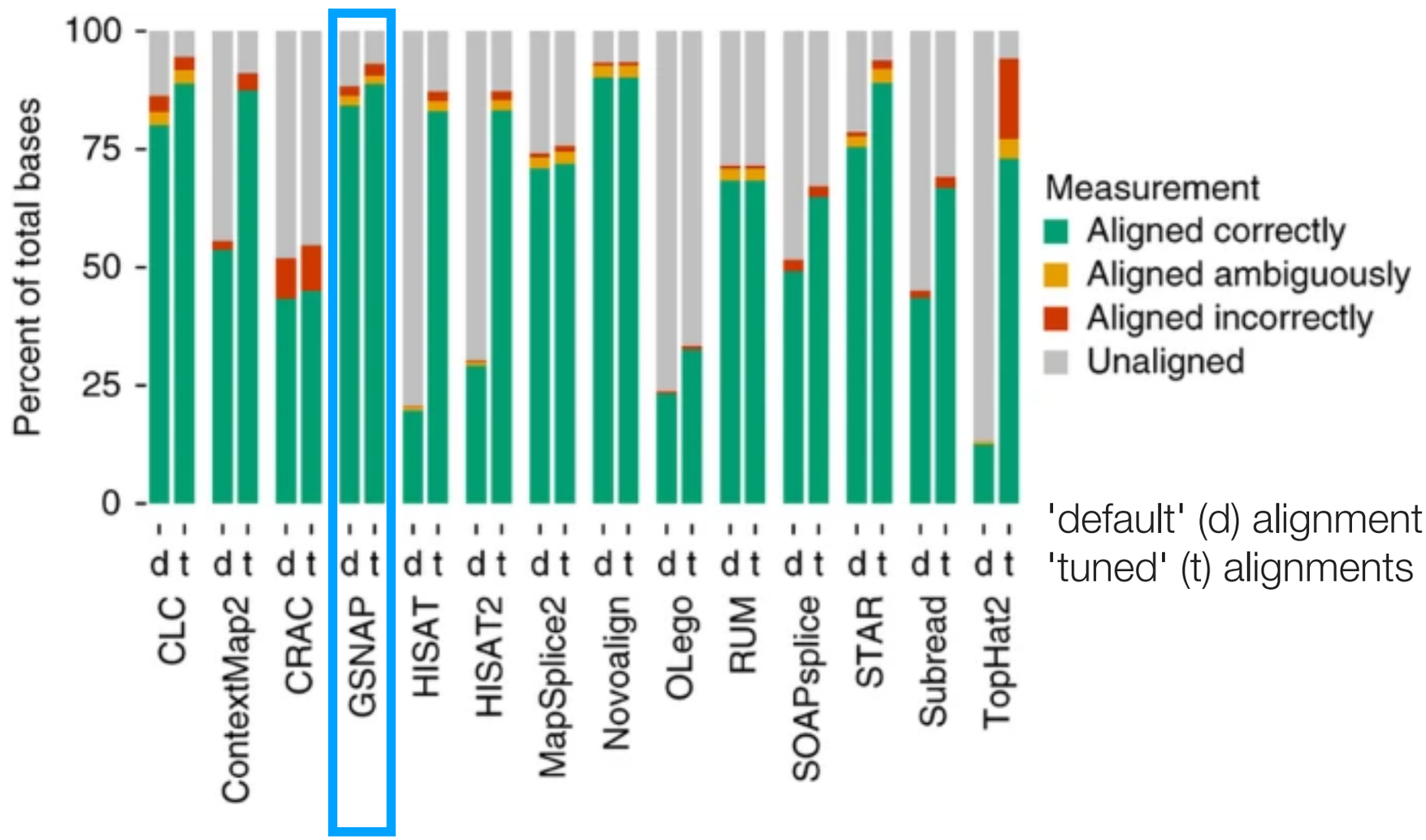
Great, but more than $1000 USD/year.

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



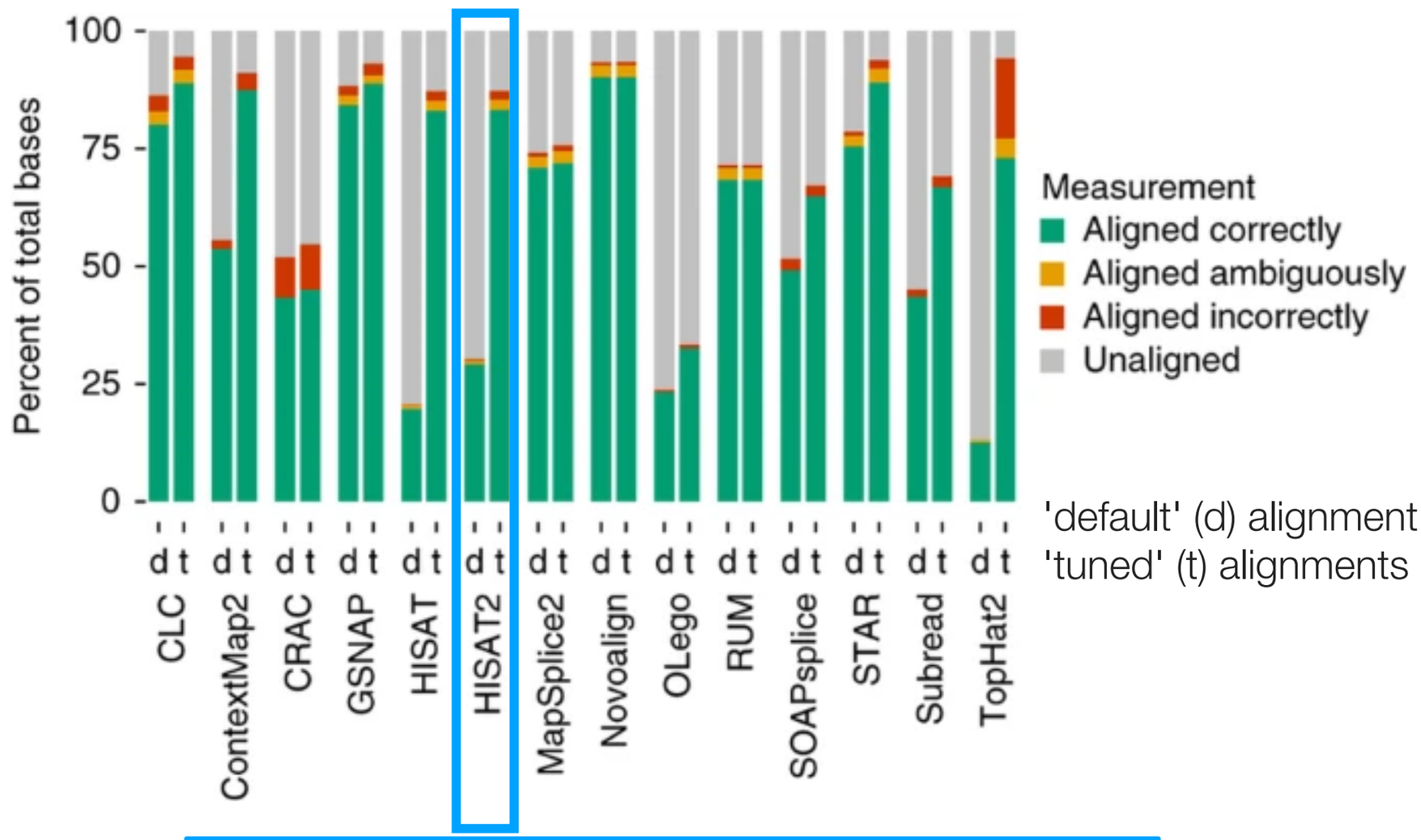
Really accurate, but slow! Not ideal for large datasets.

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



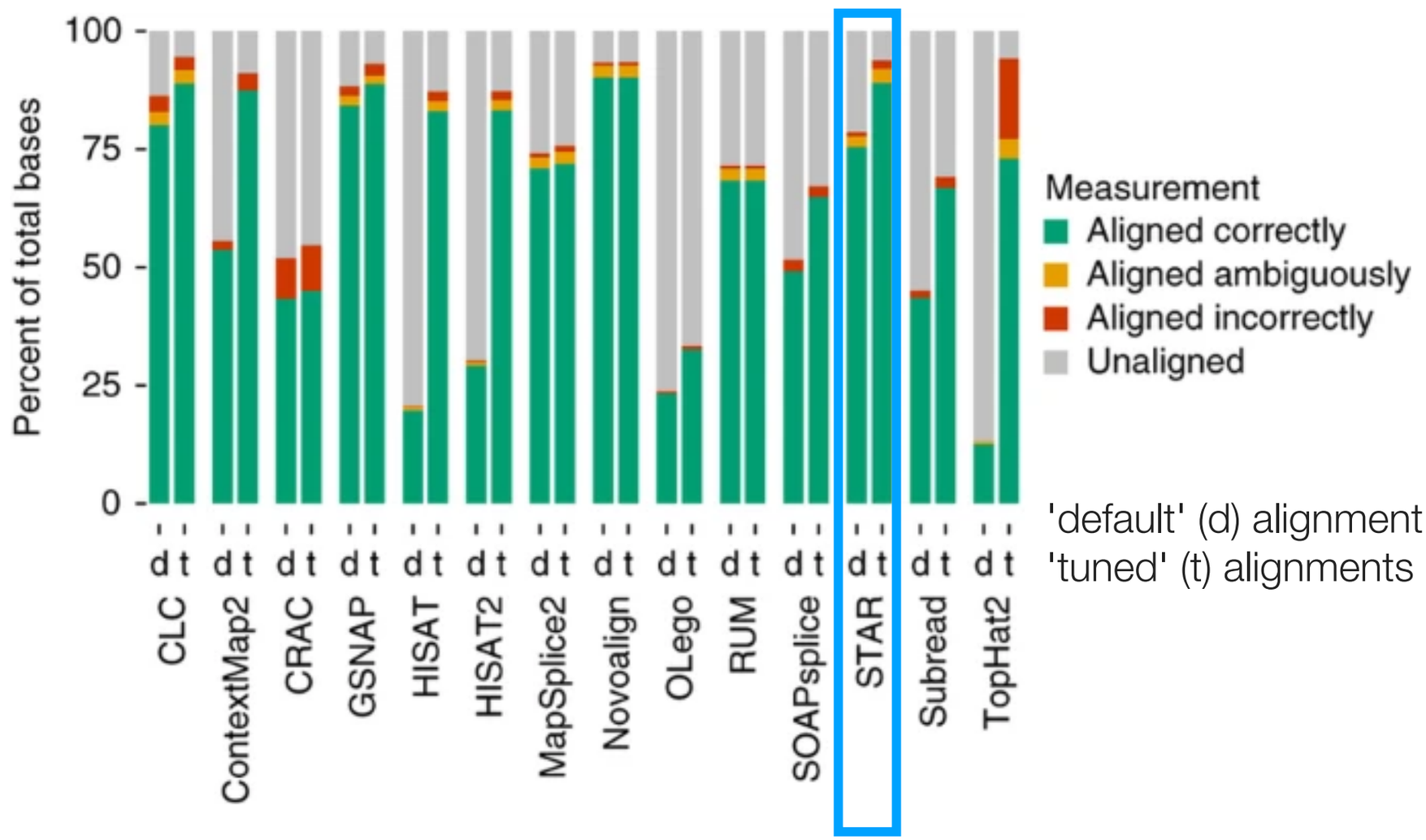
Accurate when parameters are tuned, but also really fast and memory-efficient for large datasets. Popular implementation in analysis pipelines.

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



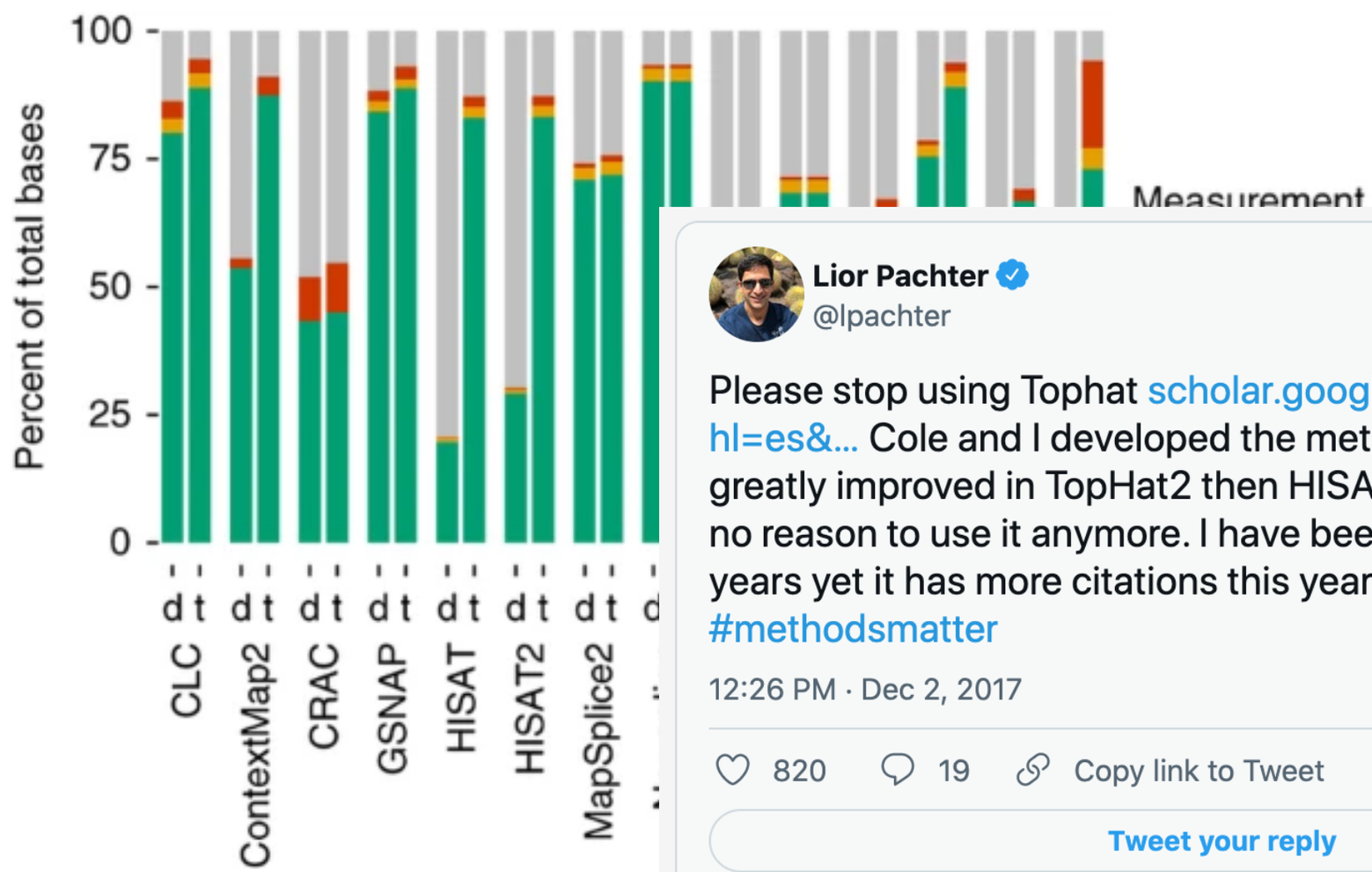
Accurate when tuned, decently fast, and good for large datasets. Handles splicing patterns well. Popular as well.

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, if there is a high quality and complete one available (e.g. RSEM, Salmon).

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

A. Reads that map across intron/exon boundaries

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.

Explore bioinformatic tools (including mappers):
https://australianbiocommons.github.io/toolfinder/

How to install and run popular RNA-seq analysis programs:
https://bernadettebiology.weebly.com/protocols--tutorials.html

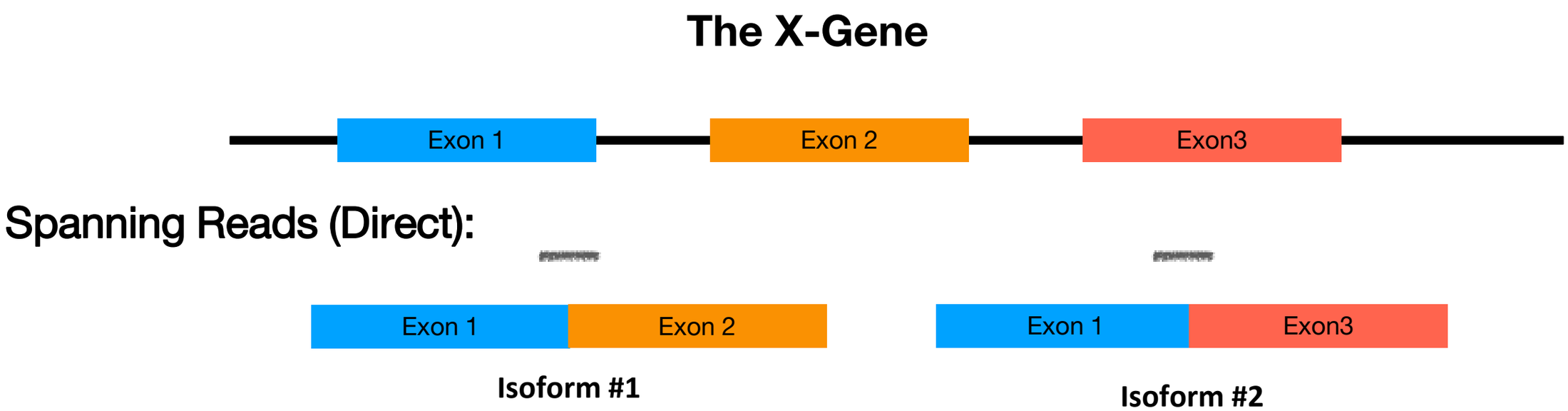
3. Map reads to a reference (genome or transcriptome)

B. Identifying abundance of alternatively spliced transcripts

The X-Gene

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.

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

B. Identifying abundance of alternatively spliced transcripts

The X-Gene

Spanning Reads (Direct):

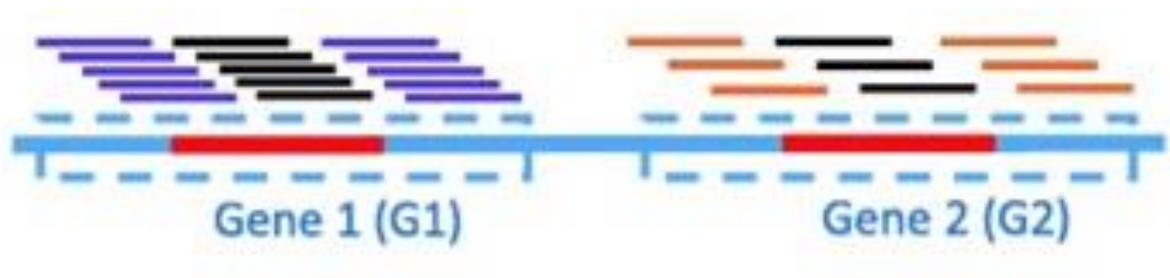
If there are two known splice variants, a read spanning exon **1** & **2** or **1** & **3** will identify which variant is present.

Exon-Specific Reads (Inferred):

If a read is 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-mapping reads



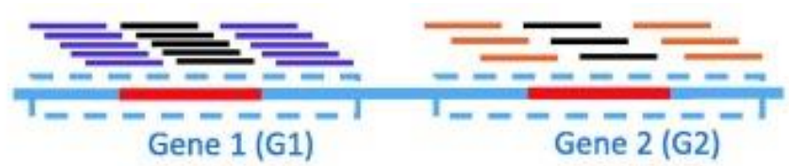
Gene duplications (paralogs), and alternatively spliced transcripts (isoforms) can give the problem of “multireads”: a read that maps with high score to several places

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-mapping 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

Ex. STAR (sort of)

Ex. STAR (sort of)

Ex. RSEM, Salmon

Deschamps-Francoeur et al
Comp Struc. Biotech..

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

D. No reference available - *de novo* assembly

De novo assembly from short reads

Programs: TRINITY

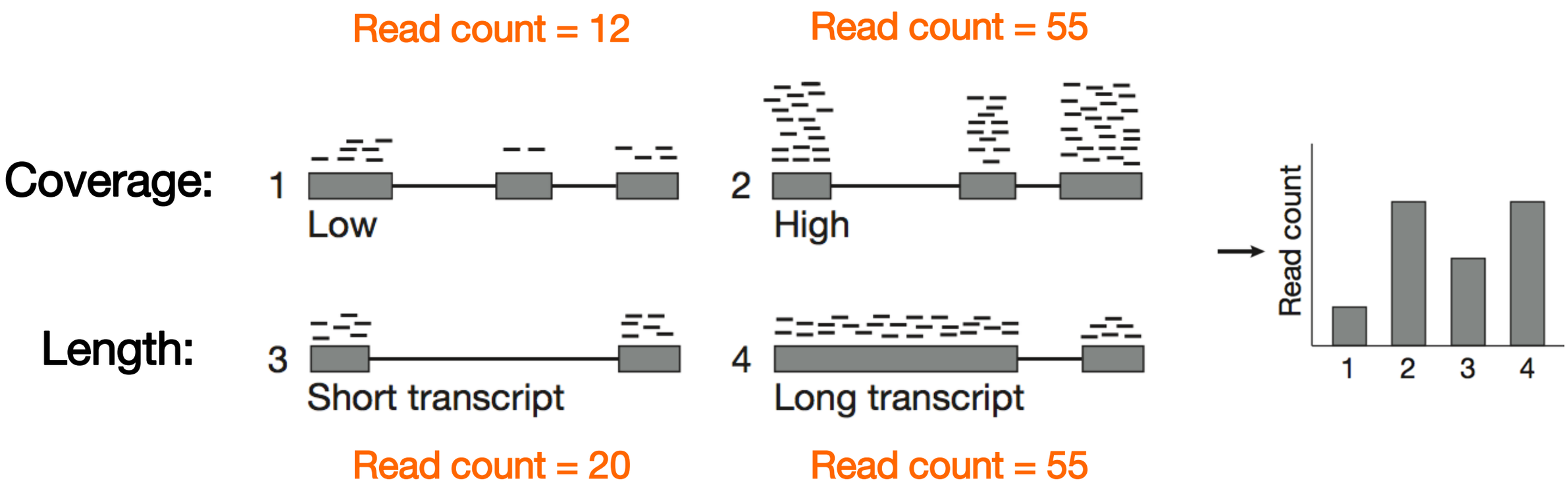
Biggest Issues: needs lots of RAM, inflates unique transcript counts, more susceptible to including contamination or sequencing errors

How is RNAseq data generated?

<u>Overview of the methods</u>

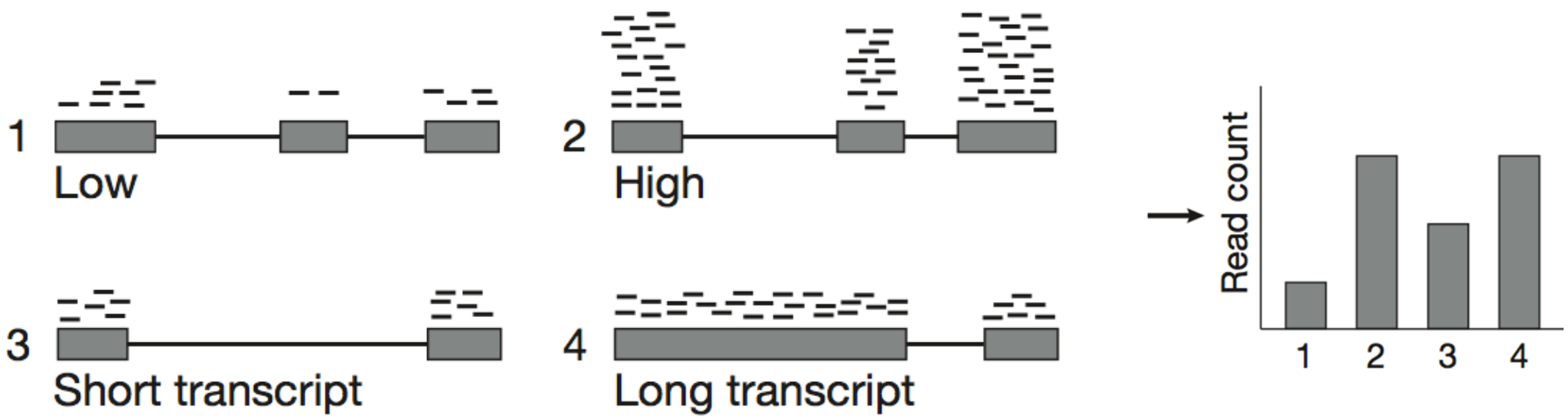
1. RNA extraction and sequencing
2. Clean and filter reads
3. Map reads to a reference (genome or transcriptome)
4. **Quantifying gene expression**
5. Statistical analysis of differences in read counts

4. Quantifying gene expression



Garber et al. 2011

4. Quantifying gene expression



Why is it important to normalize?

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

Normalizing read counts! Most common methods:

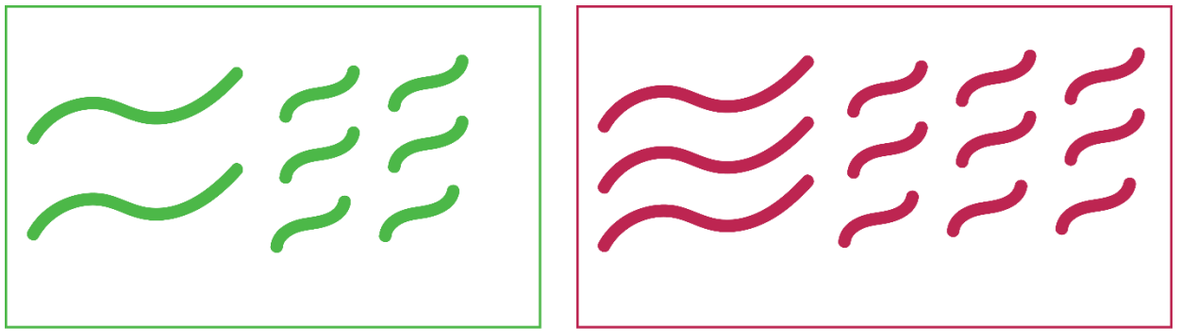
Method	Description	Accounted Factors
RPKM	**R**eads **P**er **K**ilobase per **M**illion reads mapped	• Sequencing depth • Gene length
FPKM	**F**ragments **P**er **K**ilobase per **M**illion reads mapped	• Sequencing depth • Gene length
CPM	**C**ounts **P**er **M**illion	• Sequencing depth
TPM	**T**ranscripts **P**er kilobase **M**illion	• Gene length • Sequencing depth

4. Quantifying gene expression

Normalizing read counts! Most common methods:

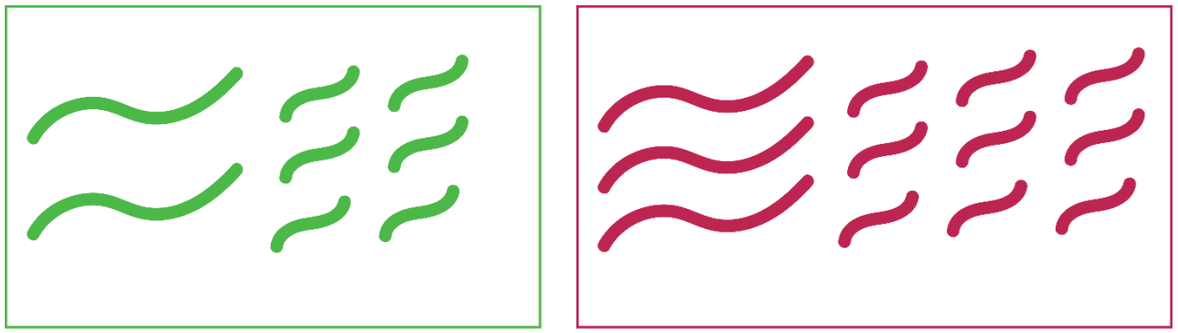
Method	Description	Accounted Factors	Usage
RPKM	**R**eads **P**er **K**ilobase per **M**illion reads mapped	• Sequencing depth • Gene length	• For single-end reads, NOT for paired-end reads • NOT for between sample comparisons
FPKM	**F**ragments **P**er **K**ilobase per **M**illion reads mapped	• Sequencing depth • Gene length	• For paired-end reads • NOT for between sample comparisons
CPM	**C**ounts **P**er **M**illion	• Sequencing depth	• NOT for within sample comparisons
TPM	**T**ranscripts **P**er kilobase **M**illion	• Gene length • Sequencing depth	• Good for within and between sample comparisons • Good for paired-end reads

4. Quantifying gene expression



Sample 1 Sample 2

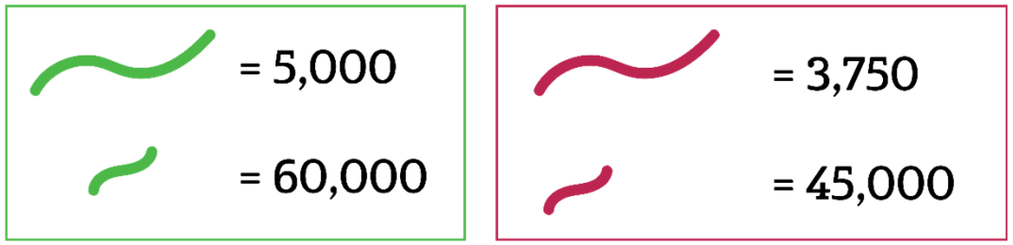
4. Quantifying gene expression



Sample 1 Sample 2

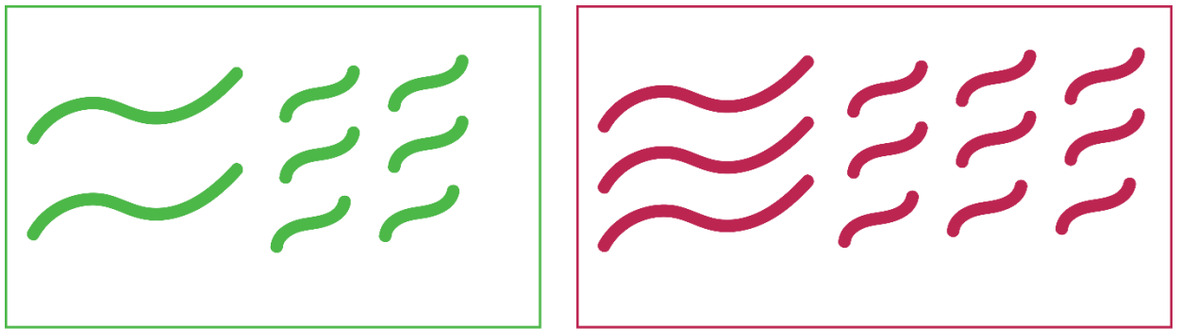
FPKM

1. Normalize for sequencing depth
2. Normalize each gene's read count by length



Sample 1 Sample 2

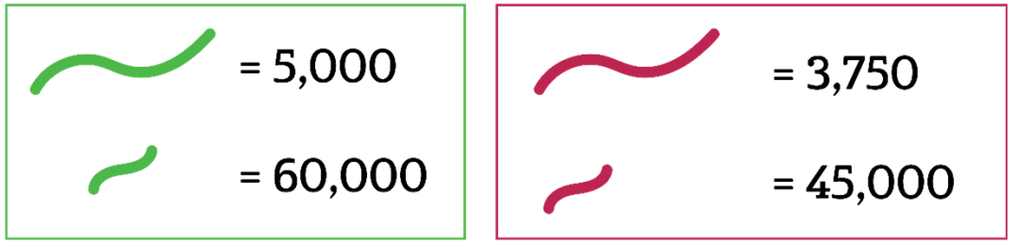
4. Quantifying gene expression



Sample 1 Sample 2

FPKM

1. Normalize for sequencing depth
2. Normalize each gene's read count by length

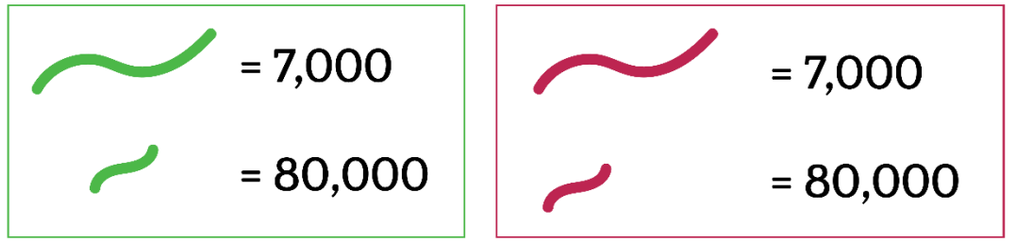


Sample 1 Sample 2

= the sum of all FPKMs in each sample are not the same

TPM

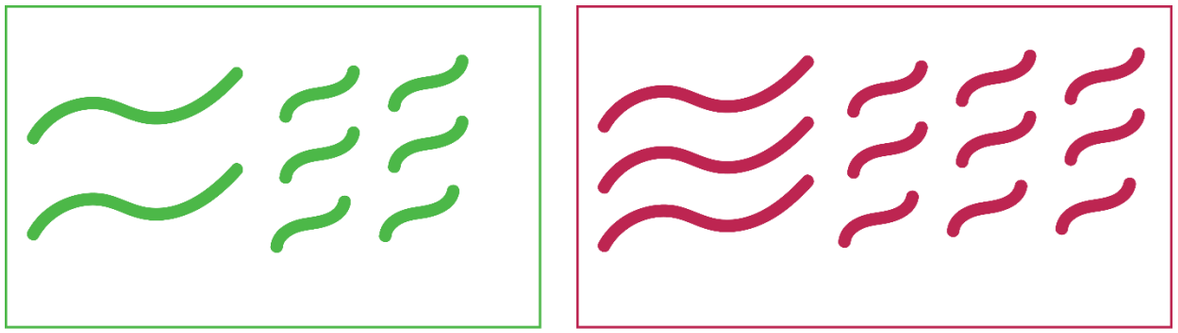
1. Normalize for read length (= Reads per Kilobase)
2. Sum all RPK values in a sample and divide by 1 M
3. Divide RPK value by per M scaling factor



Sample 1 Sample 2

= the sum of all TPMs in each sample are the same

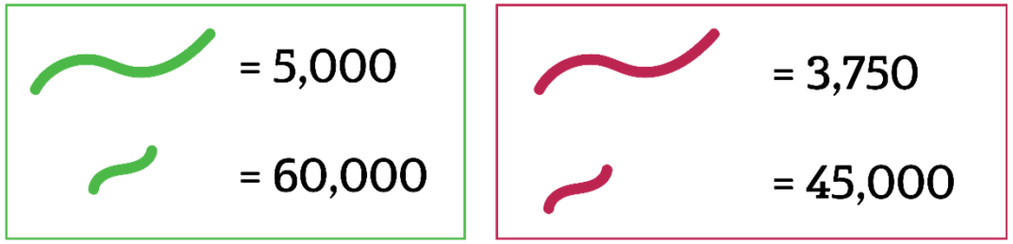
4. Quantifying gene expression



Sample 1 Sample 2

FPKM

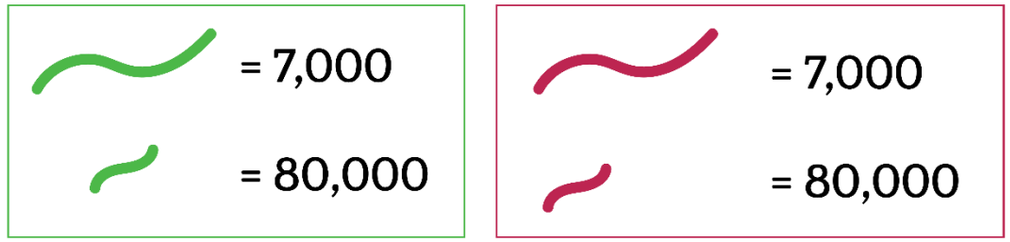
1. Normalize for sequencing depth
2. Normalize each gene's read count by length



Sample 1 Sample 2

TPM

1. Normalize for read length (= Reads per Kilobase)
2. Sum all RPK values in a sample and divide by 1 M
3. Divide RPK value by per M scaling factor



Sample 1 Sample 2

4. Quantifying gene expression

Should I use normalized reads before inputting them into a differential expression program?

4. Quantifying gene expression

Should I use normalized reads before inputting them into a differential expression program?

NO!

Why? Because differential expression programs already implement normalization strategies.

4. Quantifying gene expression

- Here are good overviews of commonly used expression units:

https://www.reneshbedre.com/blog/expression_units.html

https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

- A nice walkthrough of the DESeq2 method (for differential expression) is available here:

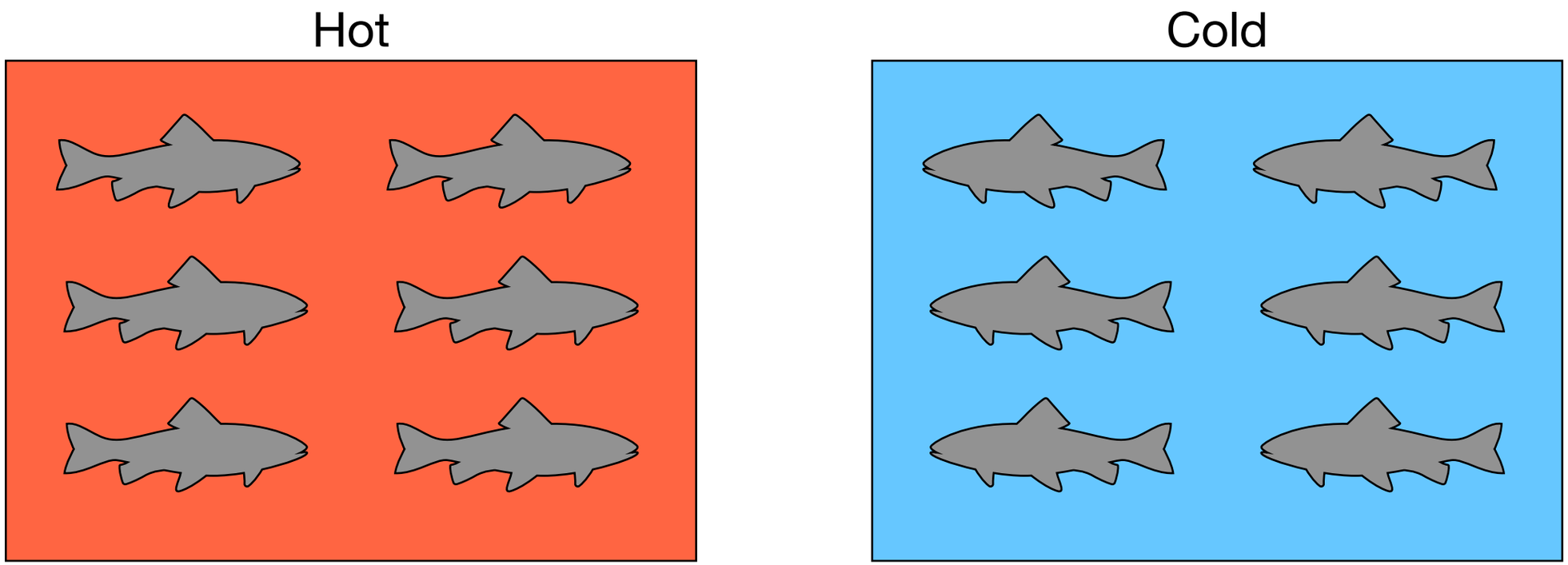
https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

How is RNAseq data generated?

<u>Overview of the methods</u>

1. RNA extraction 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/individual)

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:

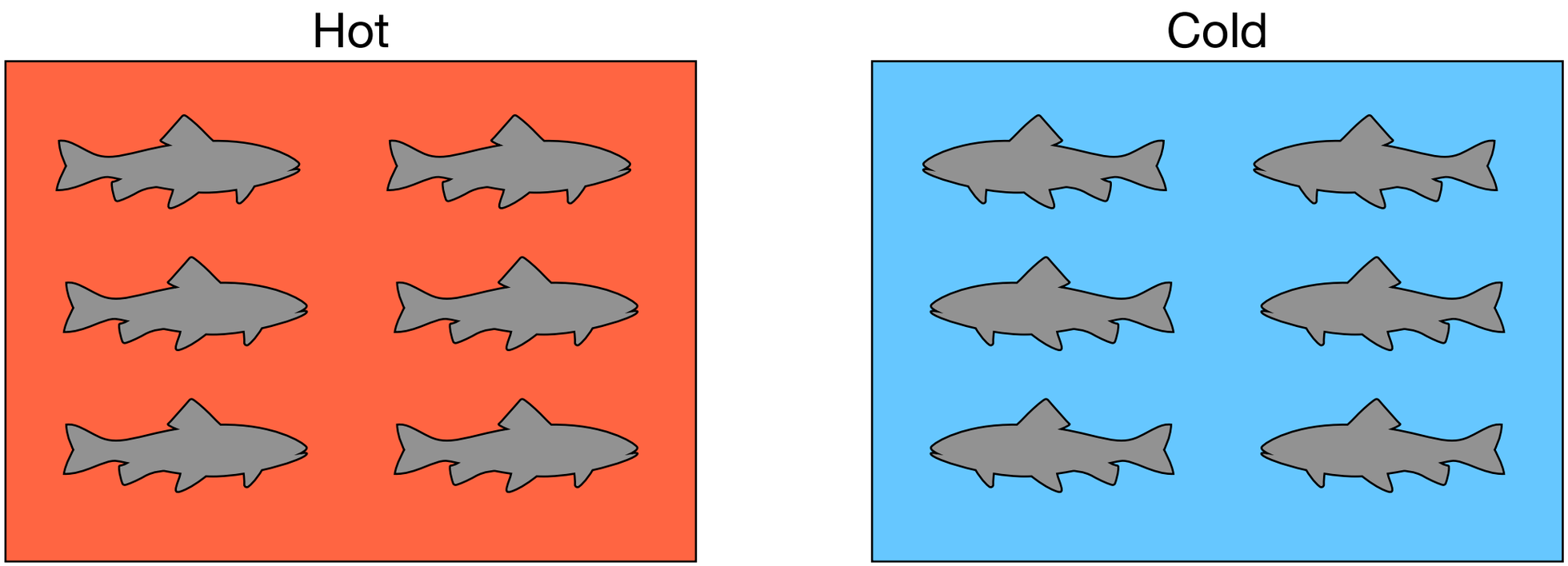
1. **Differential gene expression** on gene-by-gene basis (e.g. DESeq, EdgeR, limma)
   a. Examine how each gene is affected by a factor (e.g. treatment)
   b. Many use glms to identify genes with significant expression differences among groups

2. **Patterns of gene co-expression**
   a. Identify clusters of genes that are regulated together –
   Ex. WGCNA (Weighted Gene Co-Expression Network Analysis)

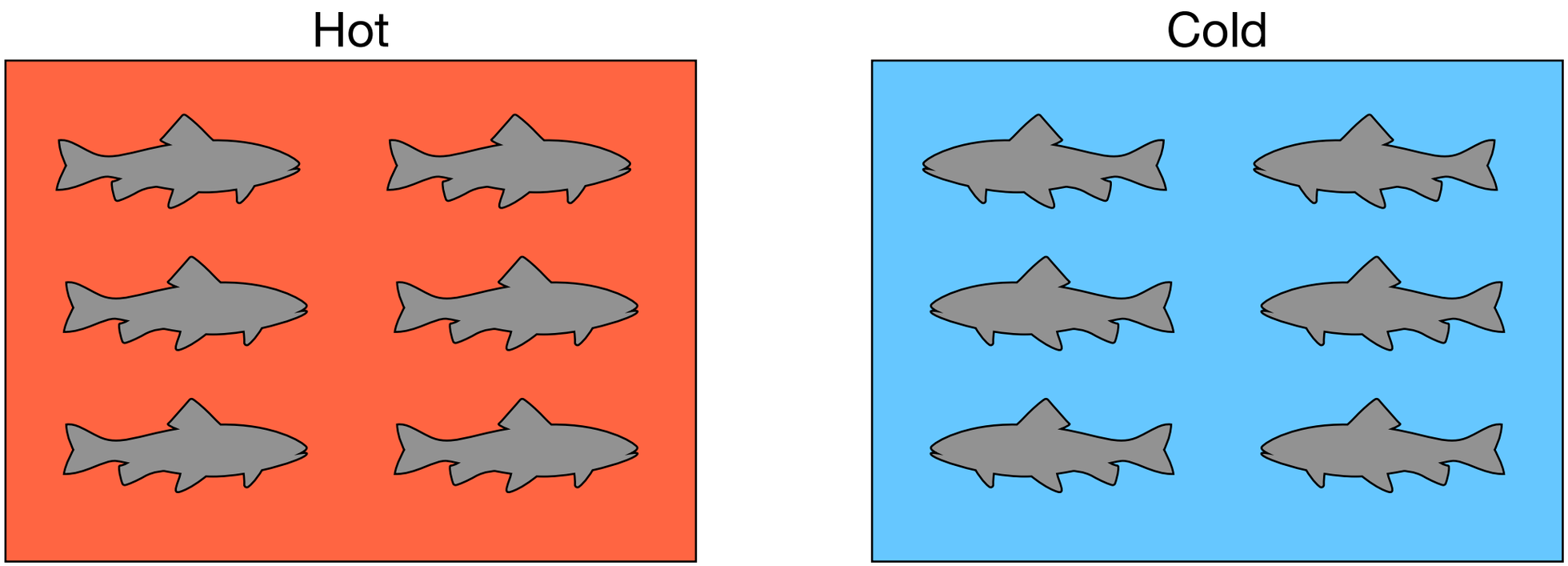
Analyzing patterns of expression

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



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?

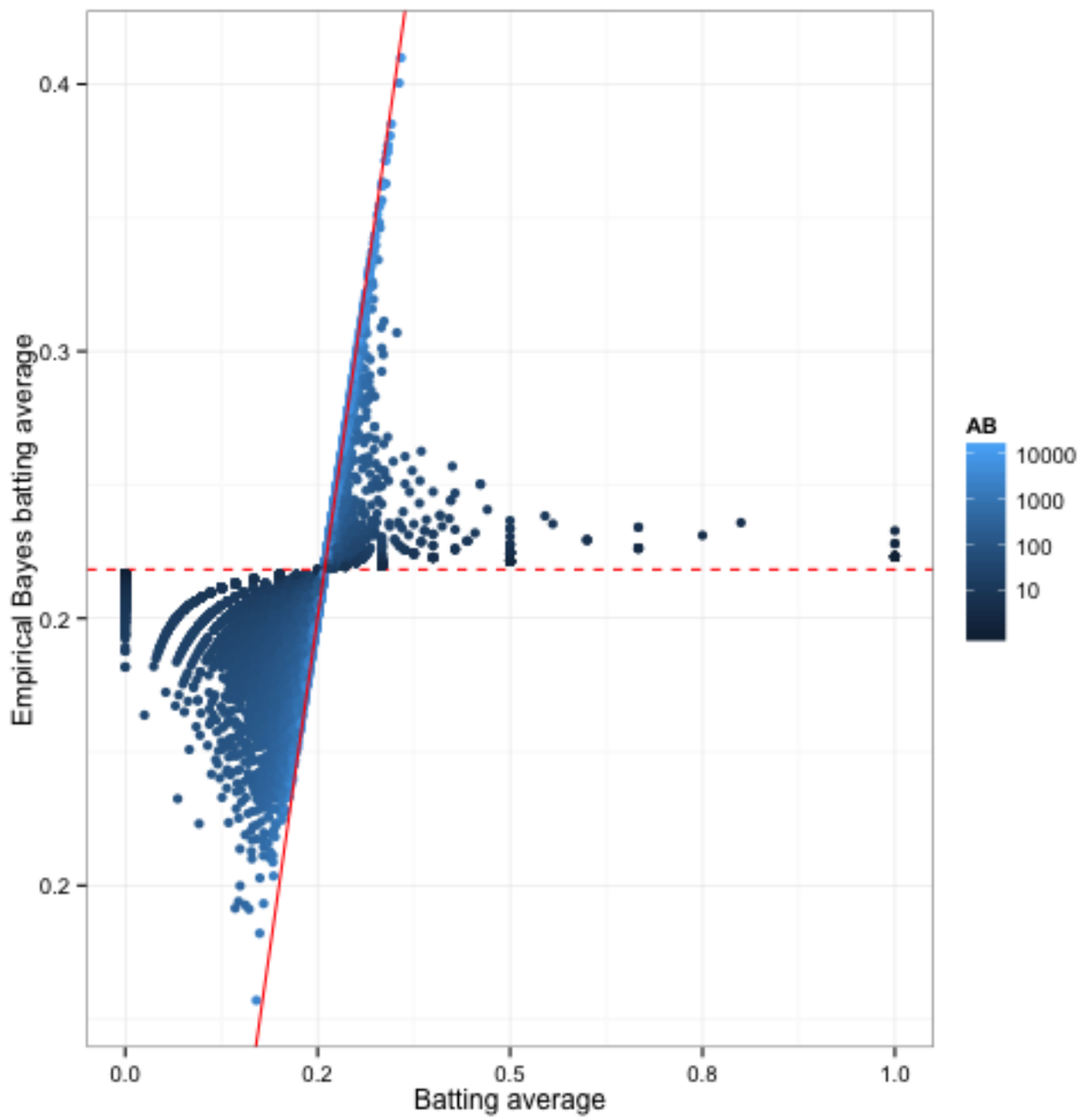
Name	Home Runs	At Bats	Average
Jeff Banister	1	1	1.0
Doc Bass	1	1	1.0
Steve Biras	2	2	1.0
C. B. Burns	1	1	1.0
Jackie Gallagher	1	1	1.0

Is this informative?

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

Who were the best batters?

In empirical Bayes analysis, 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

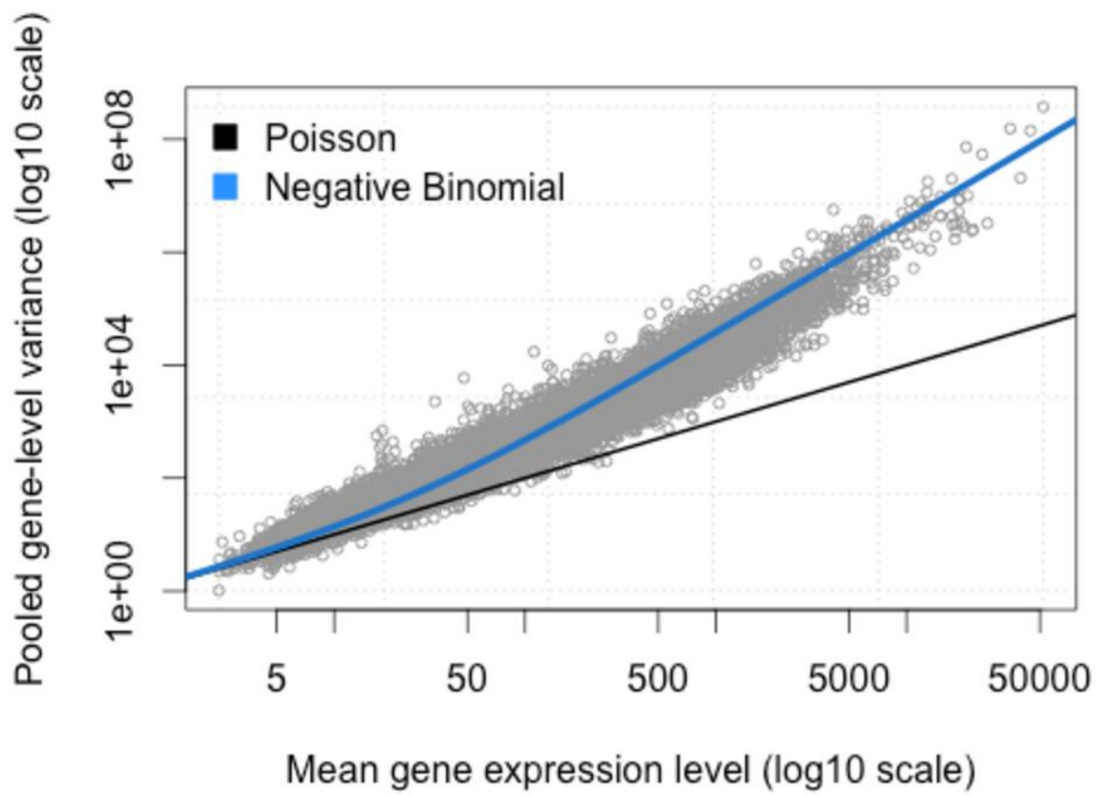
Gene	Treatment 1		Treatment 2	
	Sample 1	Sample 2	Sample 3	Sample 4
gene_A	10	20	16	14
gene_B	0	3	1	5
gene_C	32	41	11	8
gene_D	1	1	0	0

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 – DESeq2

DESeq2 is one common differential expression method

Starts with a set of normalized 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 normalized 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 – DESeq2

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

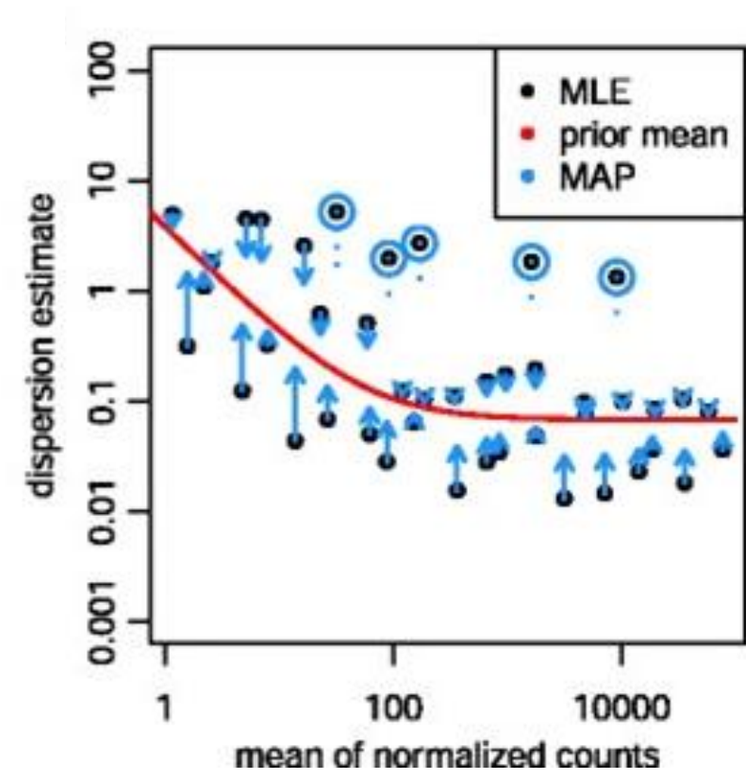
Gene	Treatment 1		Treatment 2		Mean of normalised counts	MLE of dispersion
	Sample 1	Sample 2	Sample 3	Sample 4		
gene_A	10	20	16	14	15	0.01
gene_B	0	3	1	5	2.25	0.1
gene_C	32	41	11	8	23	0.01
gene_D	1	1	0	0	0.5	1

Analyzing patterns of expression – DESeq2

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

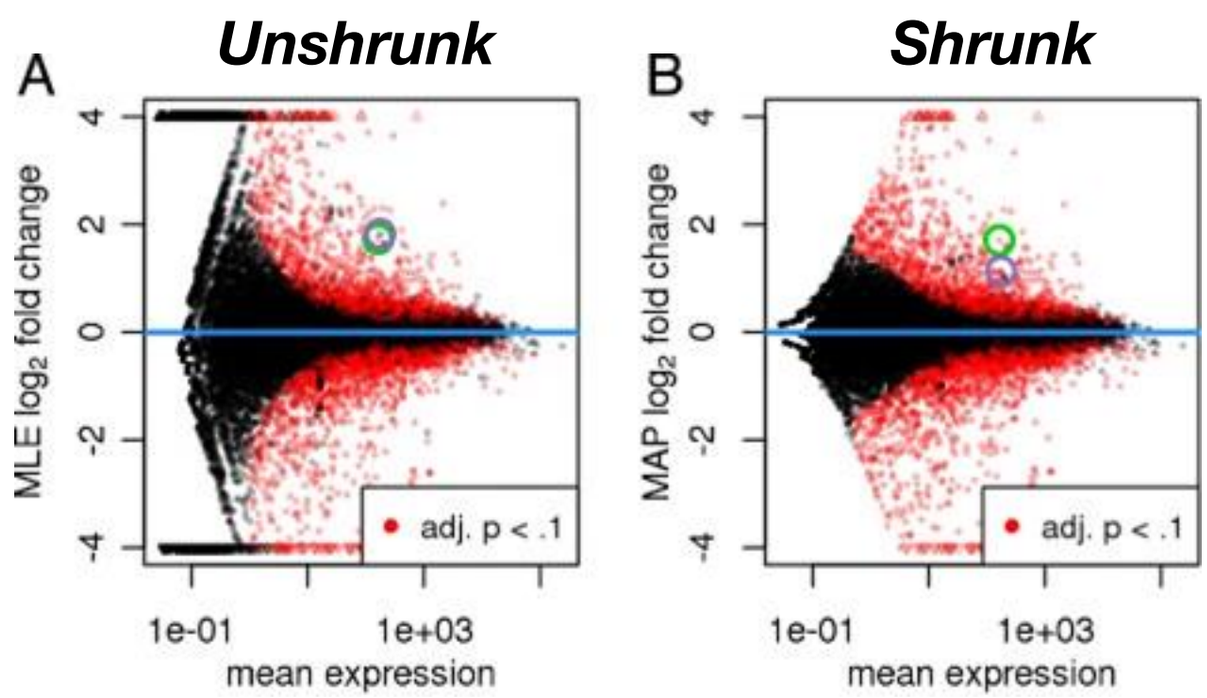
*** as inferred from all data**



Love et al 2014

Analyzing patterns of expression – DESeq2

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



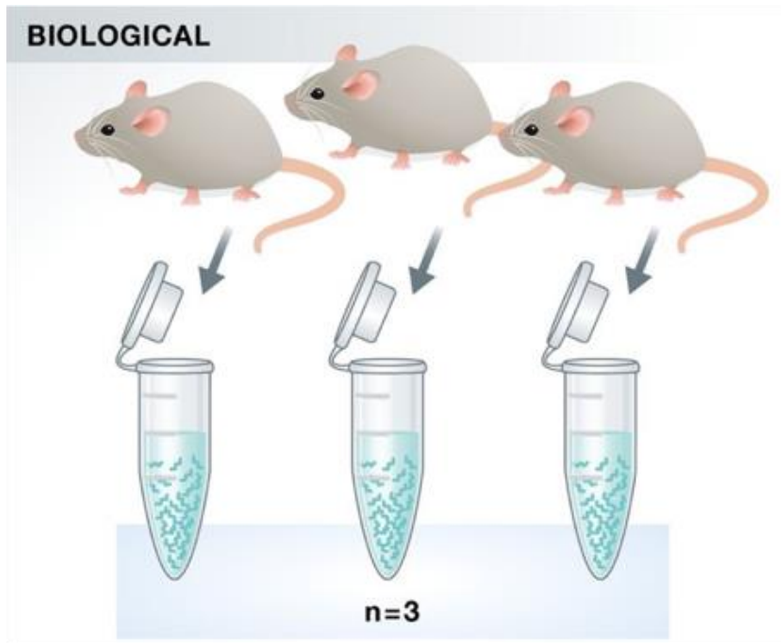
Can then identify genes with significant differences in expression

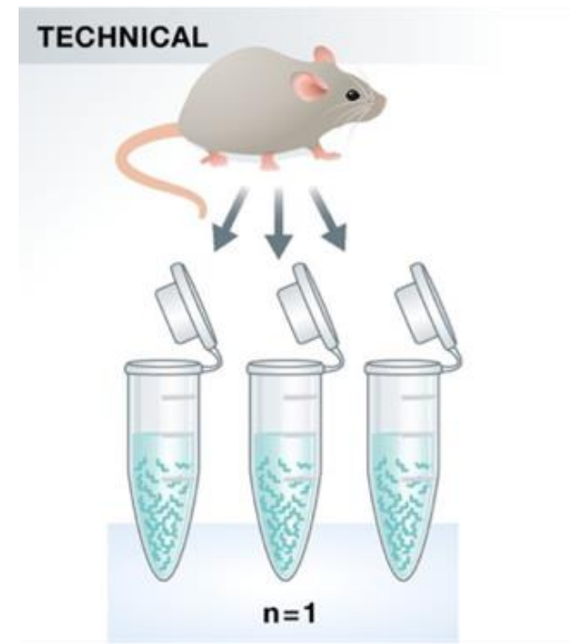
Love et al 2014

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





- Multiple samples capture biological variation.
- Usually more important than technical sampling.

- Multiple of the same sample.
- Increased sequencing depth.

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

Technical considerations

Sample Number vs Sequencing Depth

- **Sequencing Depth**: how many times a gene has been sequenced.
- Depends on transcriptome size (protein coding regions of genome), purpose of study, & known species characteristics.
- For RNA seq, depth is typically more useful than coverage (the proportion of the genome has been sequenced).

Technical considerations

Sample Number vs Sequencing Depth

- Increasing sample size results in identification of more unique reads, and generally more robust results than increasing sequencing depth.

- Increasing sequencing depth beyond 20M reads make less of an impact.

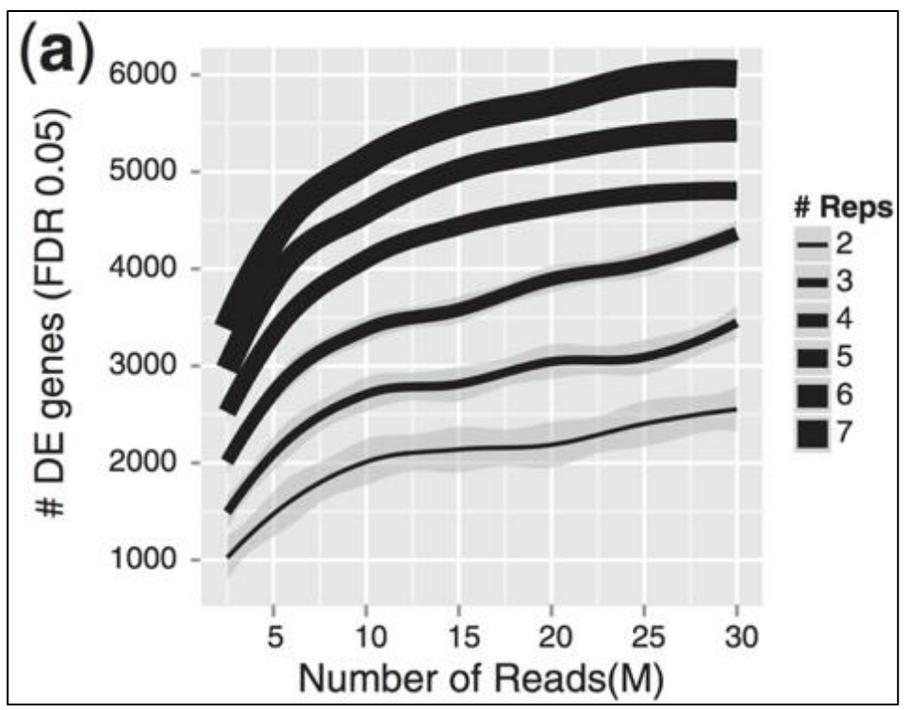


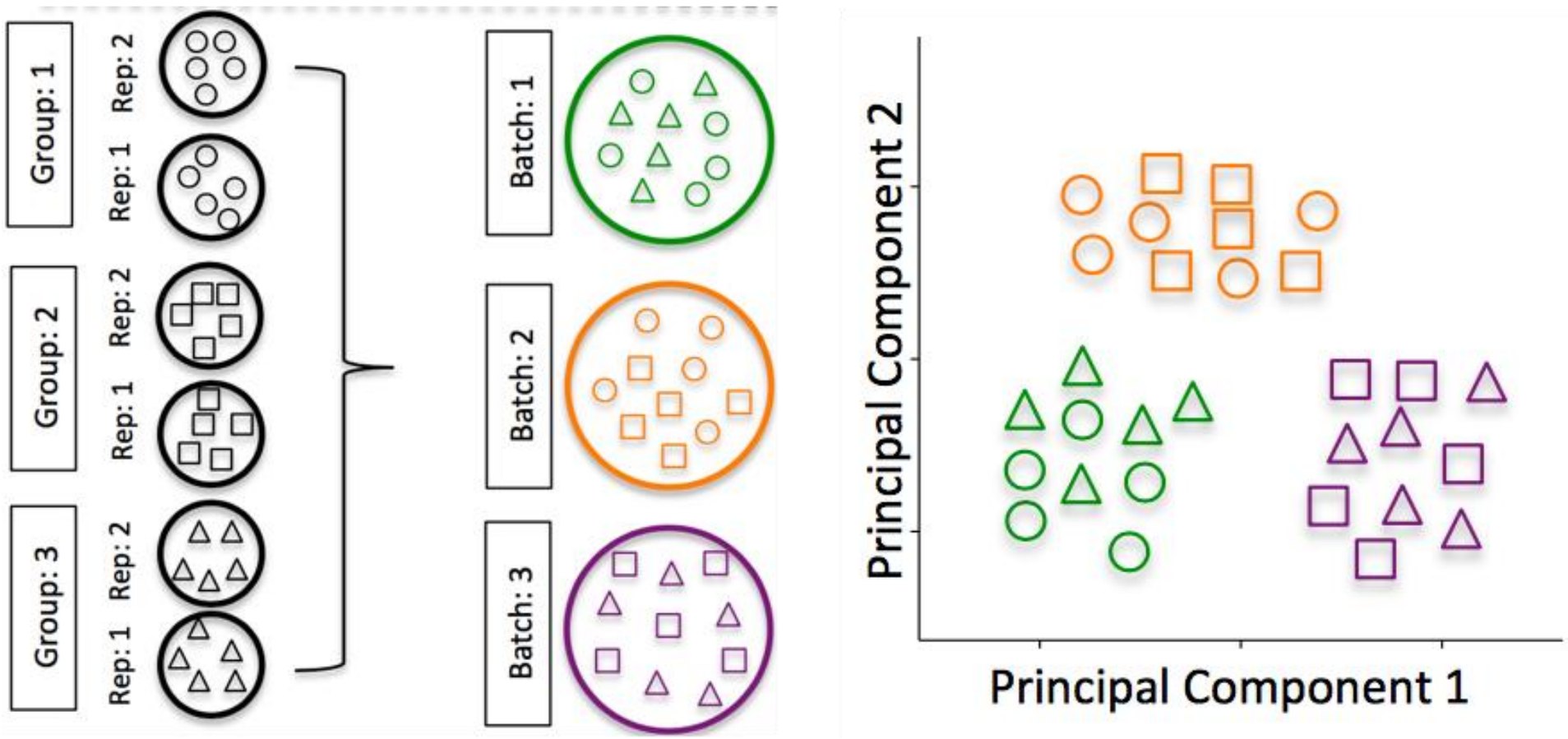
Table 1 Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	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 %

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

Technical considerations

Batch Effects: Important that replicates be randomized during sample prep and sequencing (RNA extraction, library prep and sequencing).

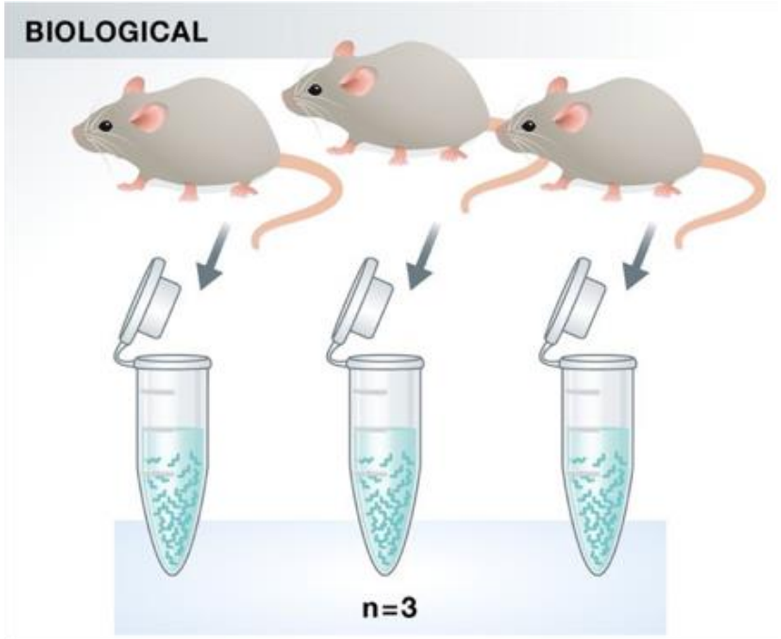


Hicks SC, et al., bioRxiv (2015)

Technical considerations

Minimize heterogeneity within a sample group:

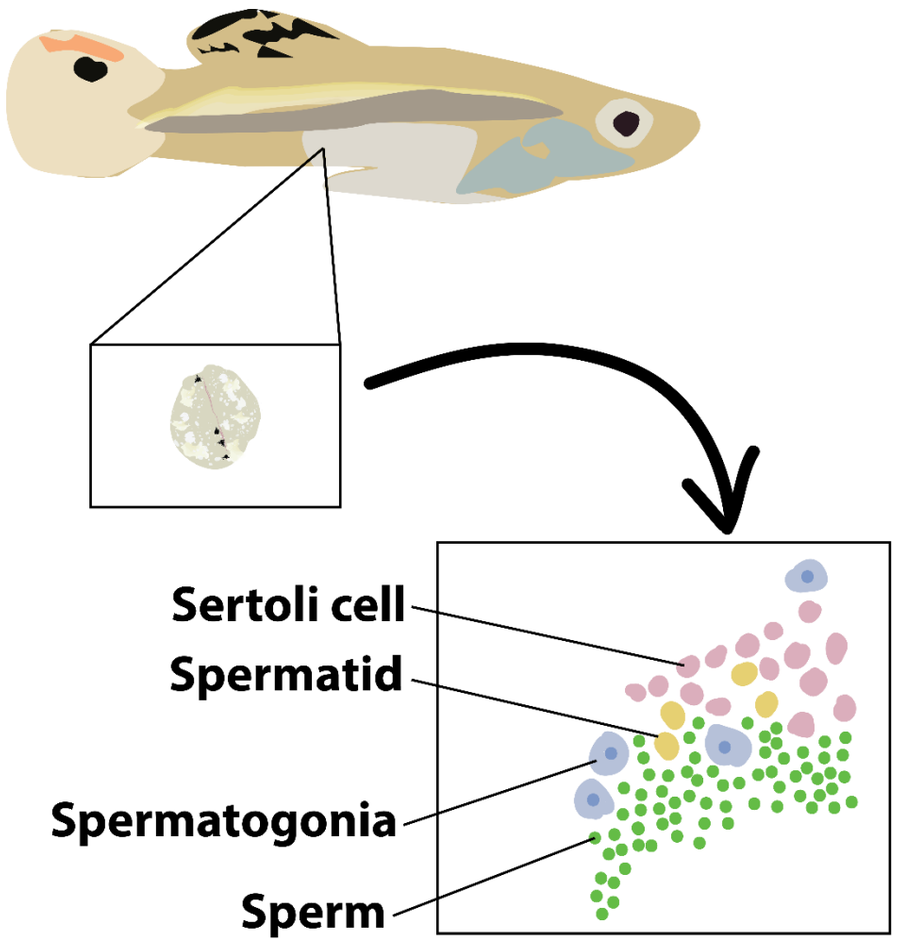
- Subsampling tissue (micro dissections)
- Sampling only one cell type
- Accounting for potential sources of variance within samples (might be species-specific)

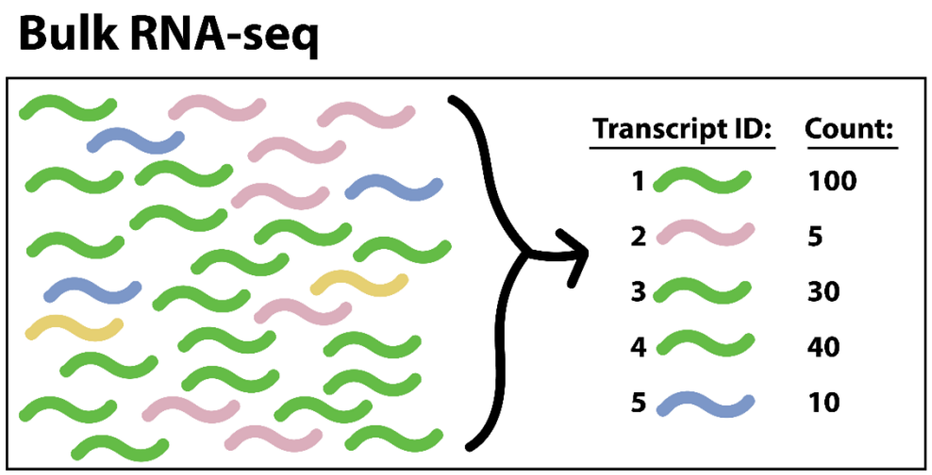


https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/experimental_planning_considerations.html

Technical considerations

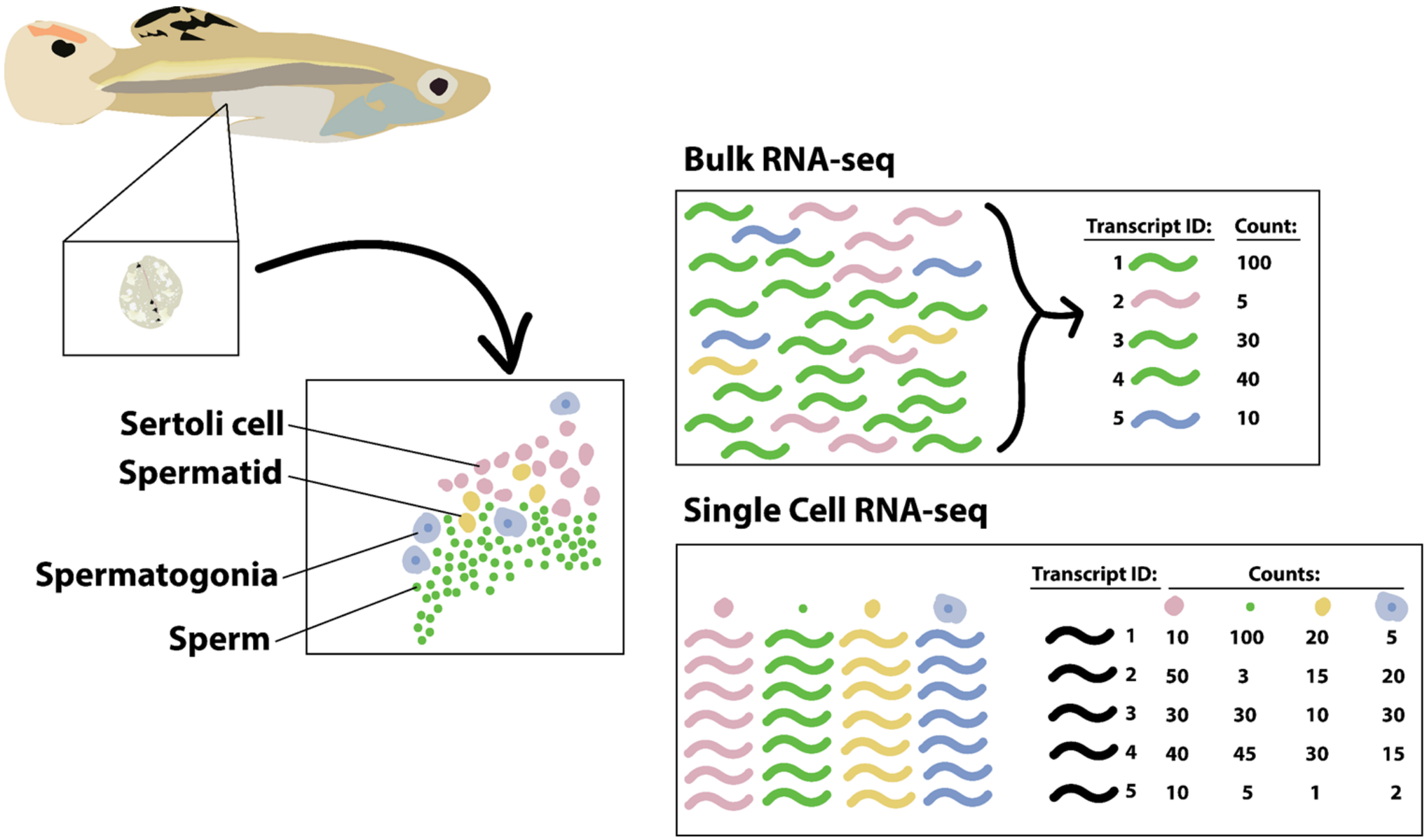
Single cell sequencing: Label transcripts on a cellular level





Technical considerations

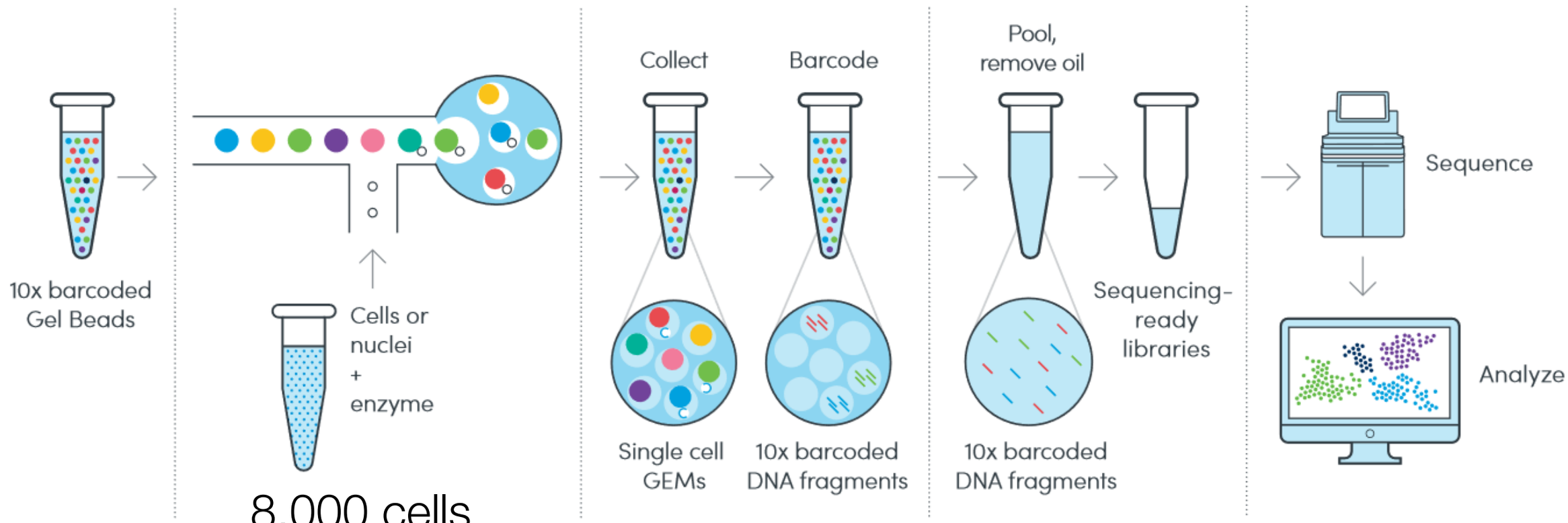
Single cell sequencing: Label transcripts on a cellular level



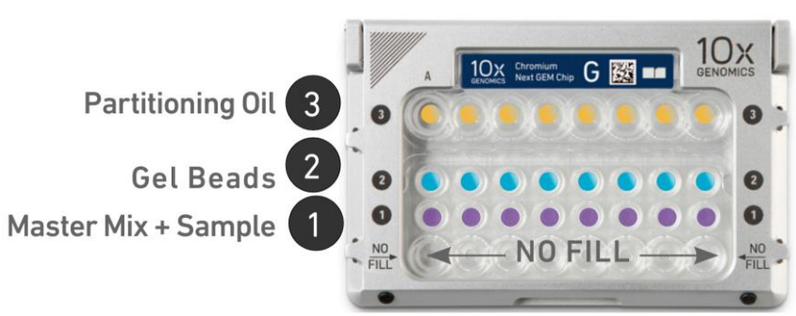
Benefit: provides cell-level transcriptome, normalizes for differences in cell abundances between samples (developmental vs expression differences).

Technical considerations

Single cell sequencing: how it works



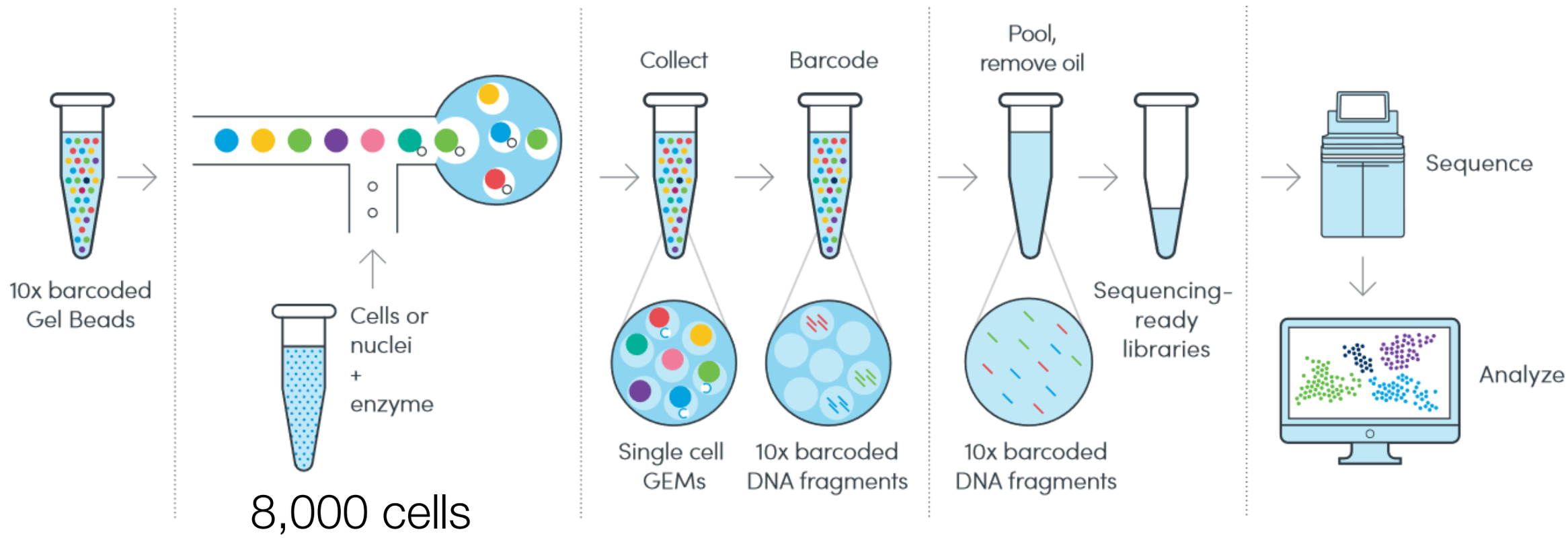
8,000 cells

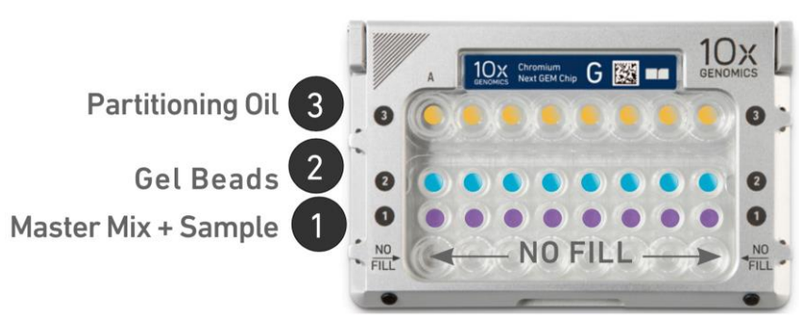


10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

Technical considerations

Single cell sequencing: how it works





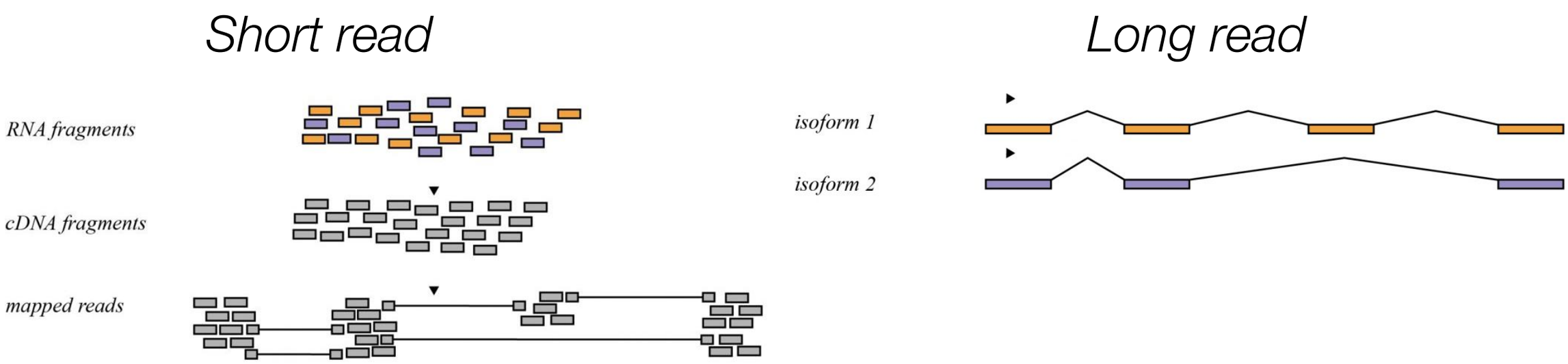
~1,600 CAD per sample
~25,000 CAD for instrument

Cons: very expensive, requires specialized equipment, very sensitive to failure

10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

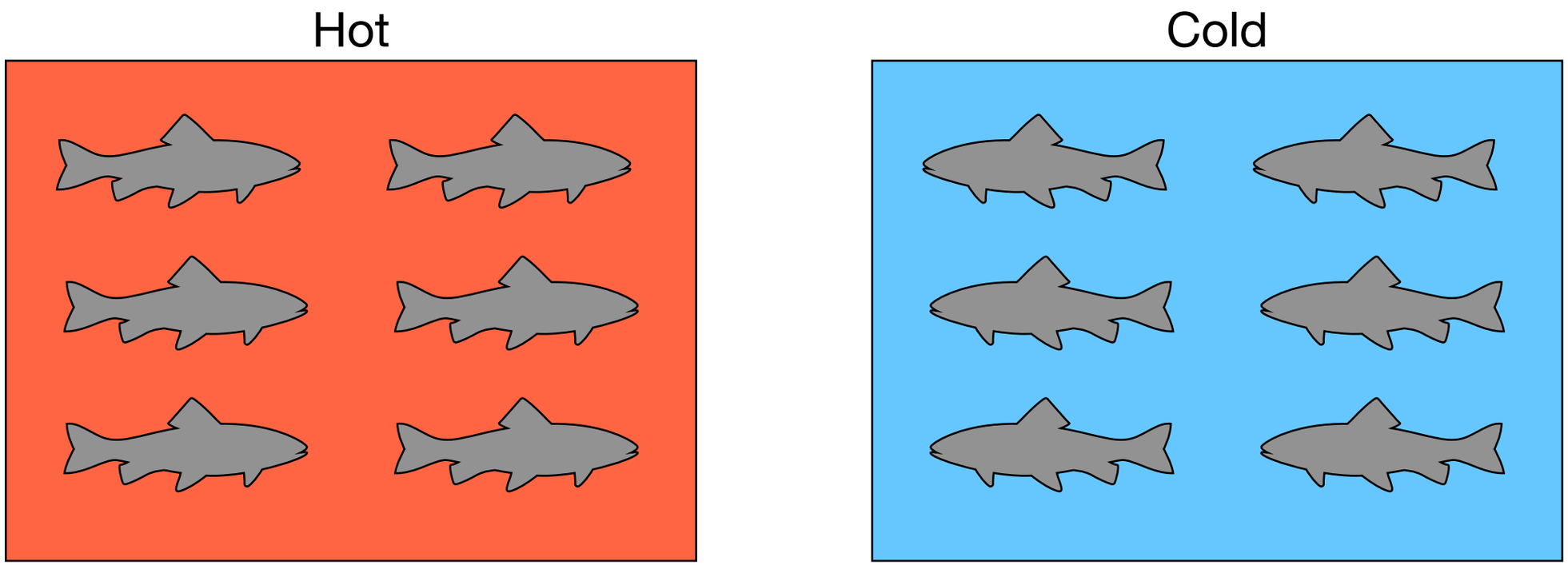
Technical considerations

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



RNA extractions are similar to bulk (somewhat easy) but can be more expensive (~4,000 CAD per sample for library prep and sequencing).

Tutorial: Analyze read counts from the fish using DESeq2



6 individuals per treatment (1 library/individual)

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

https://www.labome.com/method/RNA-seq.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