Introduction to RNA-seq for differential gene expression (DGE) analysis

Owen M. Wilkins, PhD

Bioinformatics scientist

Center for Quantitative Biology, Geisel School of Medicine at Dartmouth

Email: DataAnalyticsCore@groups.dartmouth.edu

Website: (https://sites.dartmouth.edu/cqb/projects-and-cores/data-analytics-core/

08/23/21

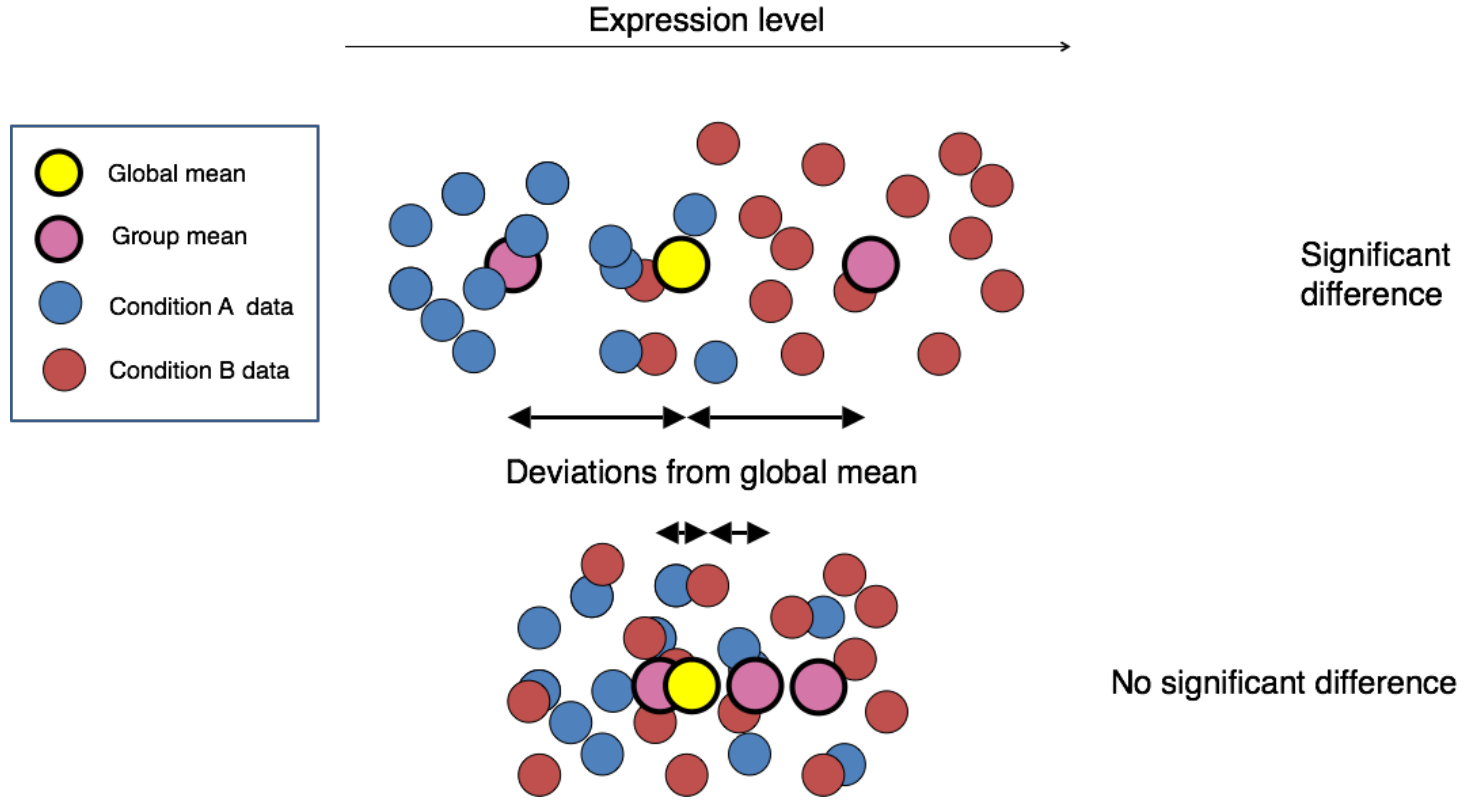
Differential expression analysis

- **Aim: Test quantitative expression changes between two or more experimental groups**

- Perform separate statistical test for read counts of each gene

Sample metadata:

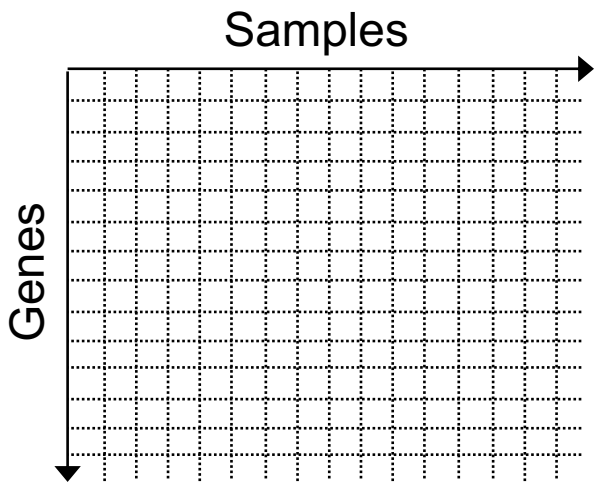
	Sex	Age	Tx-group
Sample_1	F	33	vehicle
Sample_1	F	21	vehicle
Sample_1	M	22	tx
Sample_1	F	29	tx
….	….	….	….
Sample_X	M	35	tx



https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html

Differential expression analysis

Gene expression matrix
(read counts)



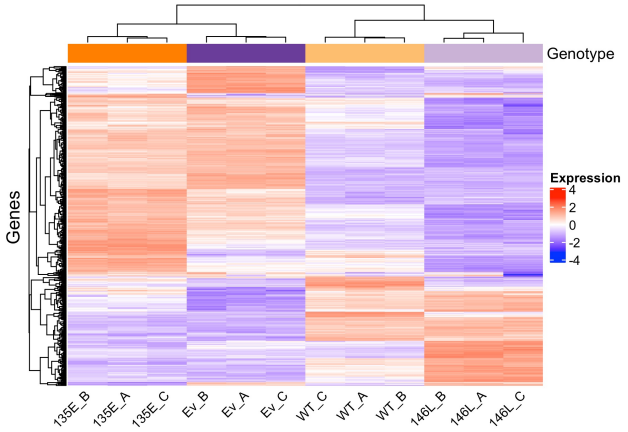
+

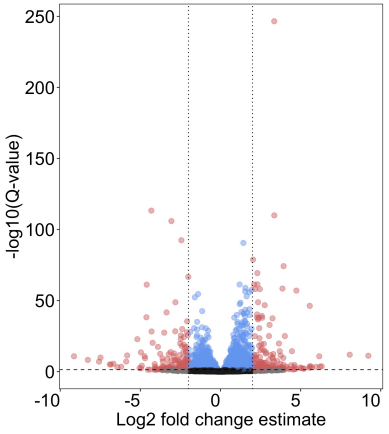
Sample metadata

- Normalization
- Exploratory analysis
- Statistical modeling
- Results prioritization
- Visualization

Differential expression results

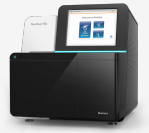
ID	lfc	lfcSE	stat	pvalue	padj	Gene
ENSG032219	2.89	0.12	-25.4	1.6E-251	1.7E-249	ARID4A
ENSG012951	-3.44	0.15	-24.5	1.2E-132	1.4E-122	KLK10
ENSG016754	-3.14	0.12	-24.1	2.9E-123	1.4E-118	KLK5
ENSG080097	-2.78	0.11	-20.7	5.7E-115	2.4E-109	LGALS1
ENSG006277	5.92	0.28	19.8	1.9E-80	3.5E-92	UCHL1
…..	…..	…..	…..	…..	…..	…..





Overview of a typical RNA-seq experiment

Sample/library preparation



Experimental Design
Sample groups
Replicate #
Sequencing strategy

Conduct experiment

Library preparation
Choose library type
RNA isolation procedure
Spike-ins

Sequencing
Paired or single-end
Choose read length
Read depth

Reads in FASTQ format

Data Pre-processing

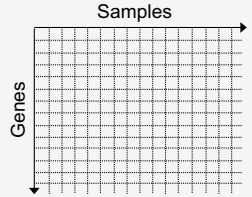
Quality control
Sequence quality
Adapter content
GC content

Read trimming
PolyA
Low quality bases
Adapters

Read mapping
Mapped %
rRNA %
Duplicate rate

Feature quantification
Read strand
Select genome annotation

Read counts



Downstream analysis

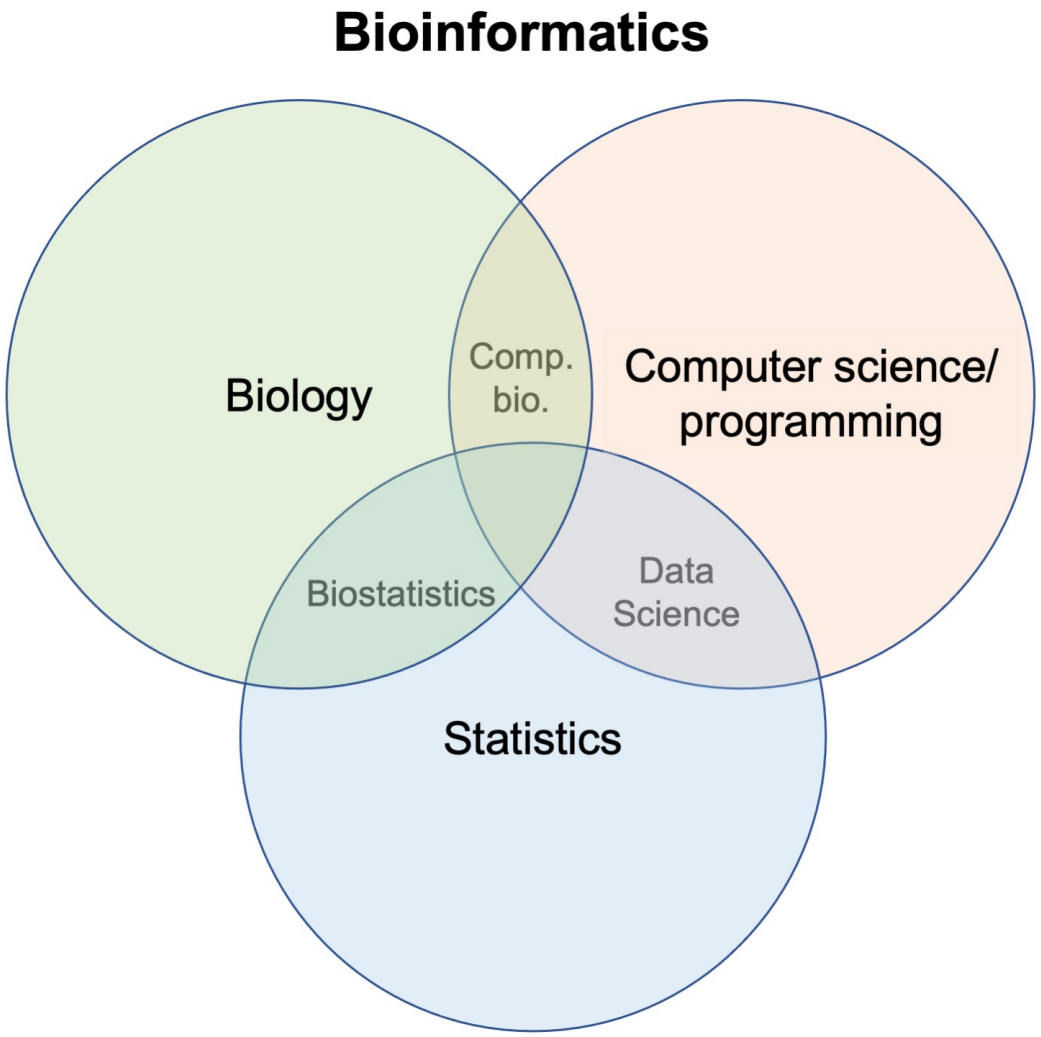
Normalization
RPKM/FPKM
TPM
Size factors/TMM

Differential expression
Statistical modeling
Visualization

Functional analysis
Pathway analysis
Gene ontology analysis

Integrative genomics
RNA-seq + ChIP-seq
RNA-seq + WGS/WES

Technical skills for RNA-seq analysis



Data pre-processing:

- Requires understanding of library preparation and sequencing technology
- More dependent on programming & computational skill set
- Often performed using established pipelines

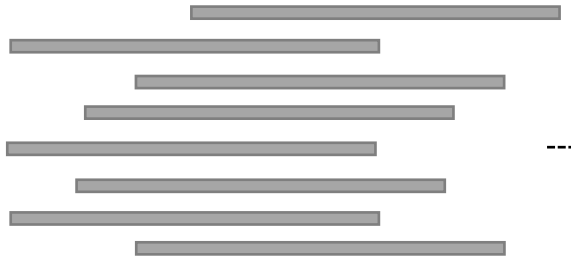
Down-stream analysis

- Requires basic knowledge of statistical concepts
- More dependent on statistical programming & data visualization skill set

Talk outline:

- **Library preparation**
  - Basic protocol
  - Types of libraries
- **Data pre-processing**
  - Quality control
  - Read trimming
  - Read mapping
  - Read counting/expression quantification
- **Sample preparation & experimental design**
  - Replicates
  - Sequencing depth
  - Sequencing configuration (read type & length)

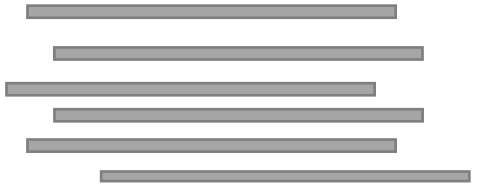
Library preparation for RNA-seq



Total RNA

Enrichment

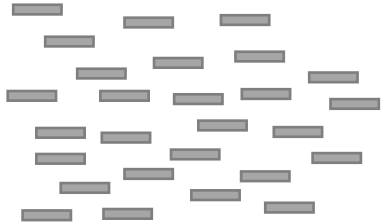
(e.g. PolyA capture, ribodepletion)



RNAs of interest

(e.g. mRNAs)

Fragmentation



Fragmented RNA

Reverse transcription (RT)

cDNA:RNA hybrid

5’

First strand synthesis

Special DNA polymerase used to RT cDNA using RNA as the template

Double stranded cDNA (dscDNA)

Second strand synthesis

3’ ends adenylated & 5’ ends repaired before adapter ligation

Index 1

SP1

DNA insert

SP2

Index 2

Adapter ligation

Sequencing adapters & barcodes for sample multiplexing

PCR amplification

Amplified libraries

1 set for each sample

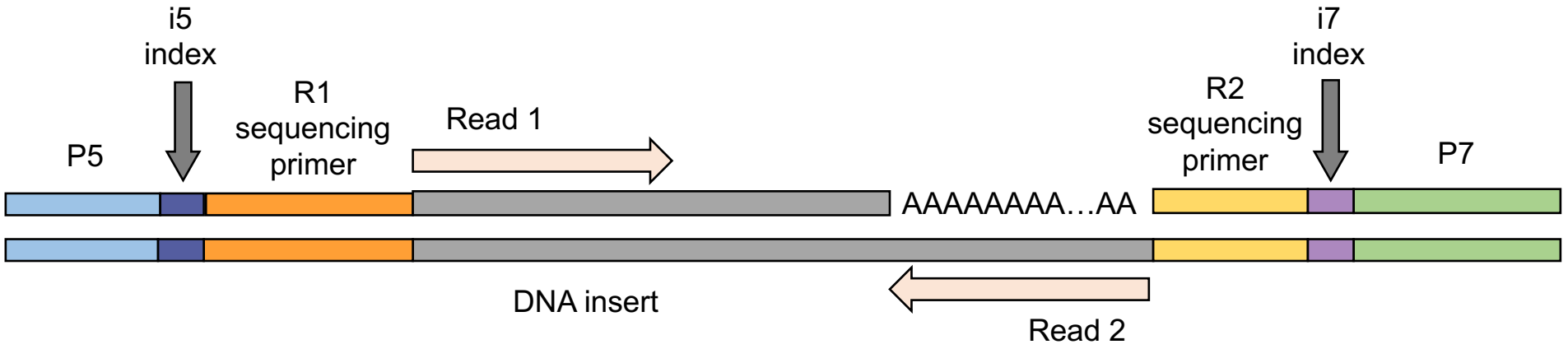
Sequencing ready cDNA pool

RNA-seq library structure

- How you choose to sequence is your choice

Barcode unique to sample, used for demultiplexing reads to produce FASTQ

Typical Illumina library:



Single-end: Collect R1 only
Paired-end: Collect R1 and R2

Choice affected by:
- Library type
- Hypothesis

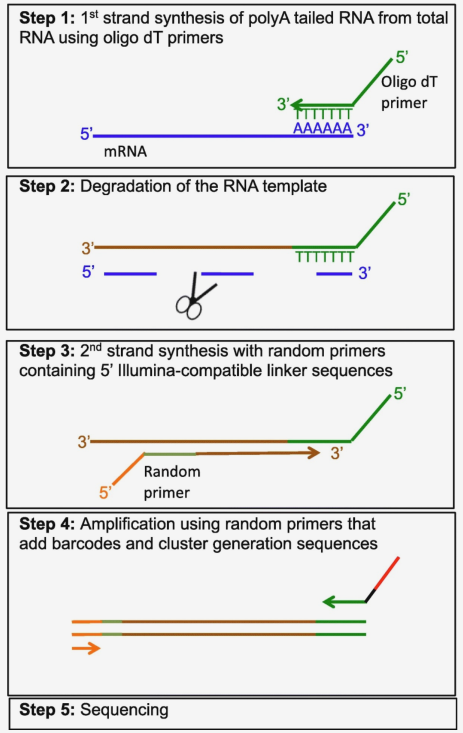
Read length: No. of seq. cycles

Seq. configuration: PE or SE + read length
e.g. PE 75bp

Applications of different library types

3’ End

3’ method (LEXO)

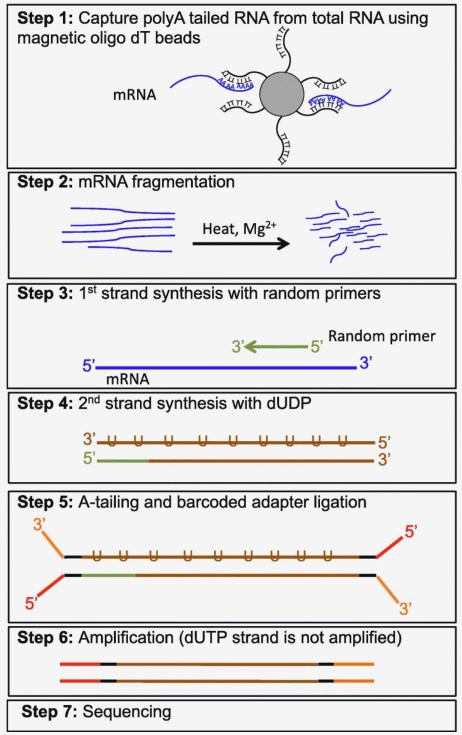


Adapted from Fukua *et al*, 2019. *Genom. Biol*

Differential Expression
Lower cost/High Throughput
Low Input and Low-Quality Samples
FFPE

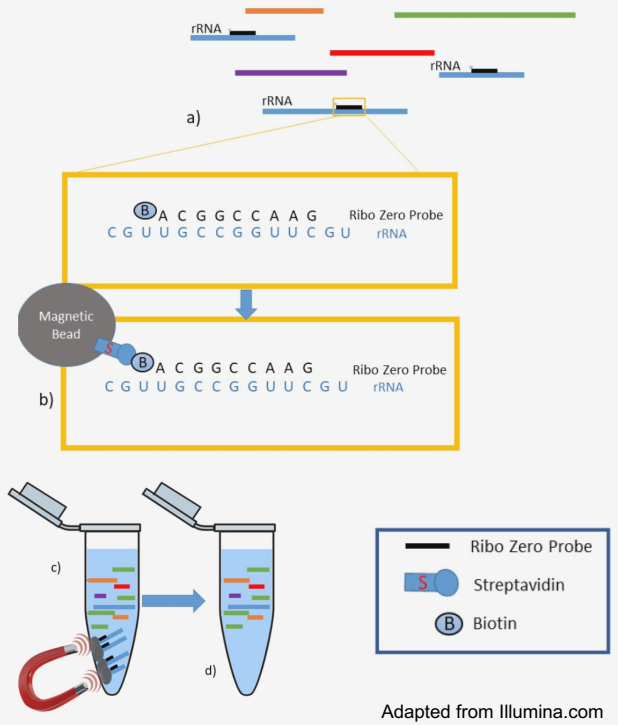
Full-length - PolyA

Traditional method (KAPA)



Full Length mRNA
Differential Expression
Splice Variants
SNV Detection
Low Input with Amplification

Ribodepletion



Adapted from Illumina.com

Full length mRNA + lincRNAs
Differential Expression
Splice Variants
SNVs
FFPE

Throughput

Cost / Data Richness

Image Credit:
Lexogen Inc

Traditional library prep. vs .3'end

Traditional method

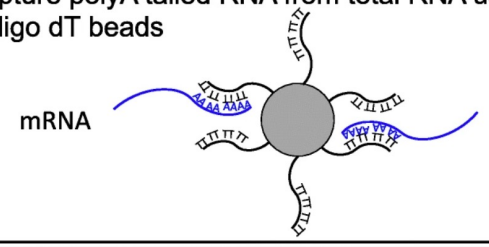
- cDNA generation using random hexamers
- Full-length transcript
- More detailed data

3' method

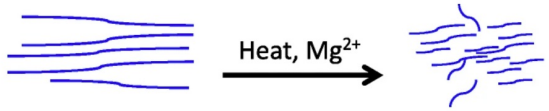
- 3' –end of transcript only
- No transcript information
- Only eukaryotic samples
- **Big cost savings**

Traditional method (KAPA)

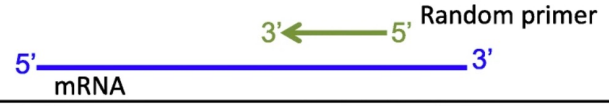
Step 1: Capture polyA tailed RNA from total RNA using magnetic oligo dT beads



Step 2: mRNA fragmentation



Step 3: 1st strand synthesis with random primers



Step 4: 2nd strand synthesis with dUDP

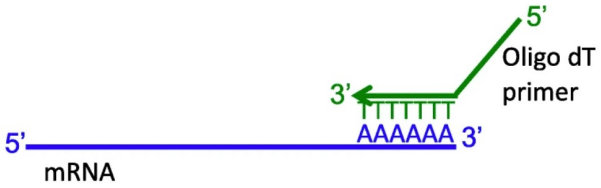
Step 5: A-tailing and barcoded adapter ligation

Step 6: Amplification (dUTP strand is not amplified)

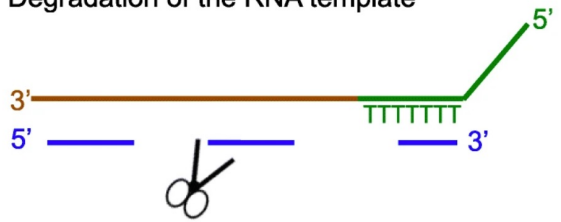
Step 7: Sequencing

3' method (LEXO)

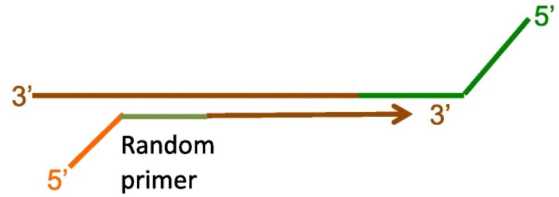
Step 1: 1st strand synthesis of polyA tailed RNA from total RNA using oligo dT primers



Step 2: Degradation of the RNA template



Step 3: 2nd strand synthesis with random primers containing 5' Illumina-compatible linker sequences

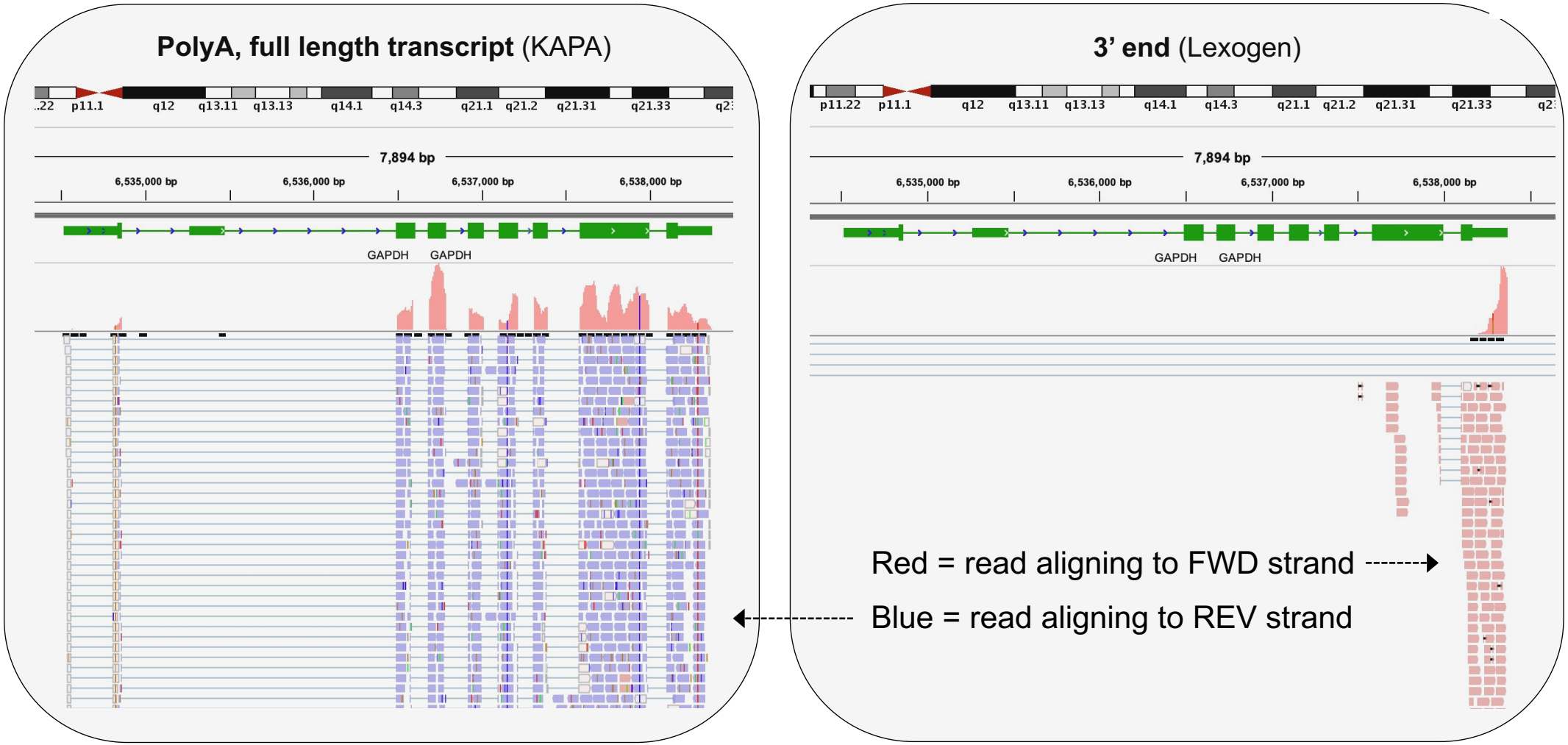


Step 4: Amplification using random primers that add barcodes and cluster generation sequences

Step 5: Sequencing

Adapted from: Fukua *et al*, 2019. *Genom. Biol*

Data from different library types looks different



Quality control, analysis pipelines, and possible hypotheses therefore inherently differ

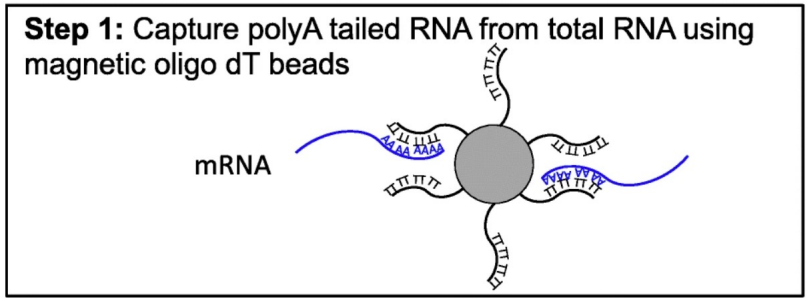
RNA selection/enrichment

- **Most of RNA in cell is ribosomal, which we don't usually care about..**

- **Oligo-d(T) selection:**
  - Uses oligos of dT attached to magnetic beads to capture polyadenylated RNAs (mostly mRNA)
  - Magnet used to retain RNAs with polyA sequences

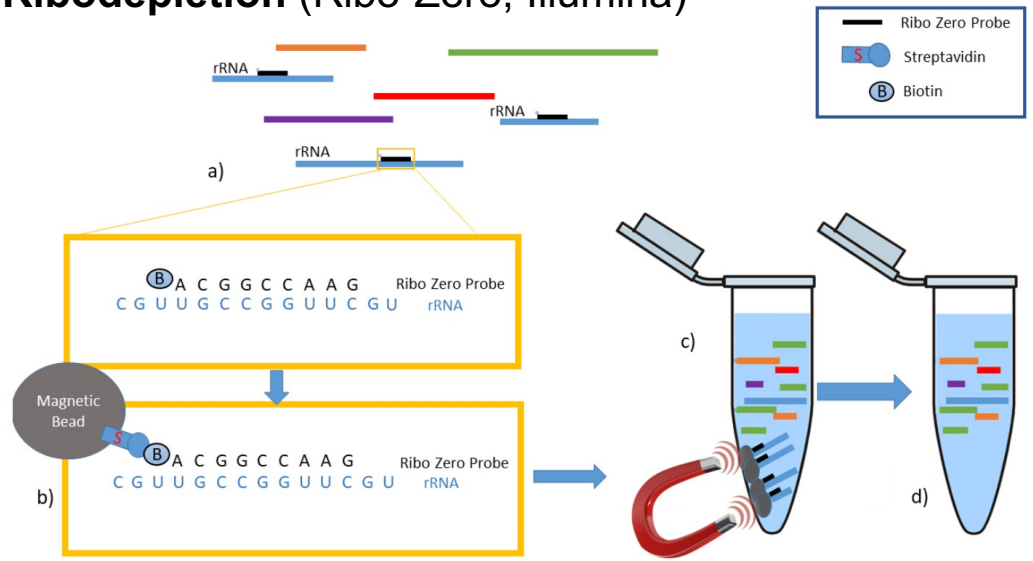
- **Ribodepletion:**
  - Enables enrichment of polyA mRNA & non-polyA ncRNAs (e.g. lncRNAs)
  - Several methods exist, e.g. Ribo-zero:
    - Oligos complementary to highly conserved rRNA sequences used to capture rRNA
    - Streptavidin-bound magnetic beads used to capture and remove hybridized sequences

Oilgo-d(T) selection



Adapted from Fukua *et al*, 2019. *Genom. Biol*

Ribodepletion (Ribo-Zero, Illumina)



Adapted from Illumina.com

Stranded libraries

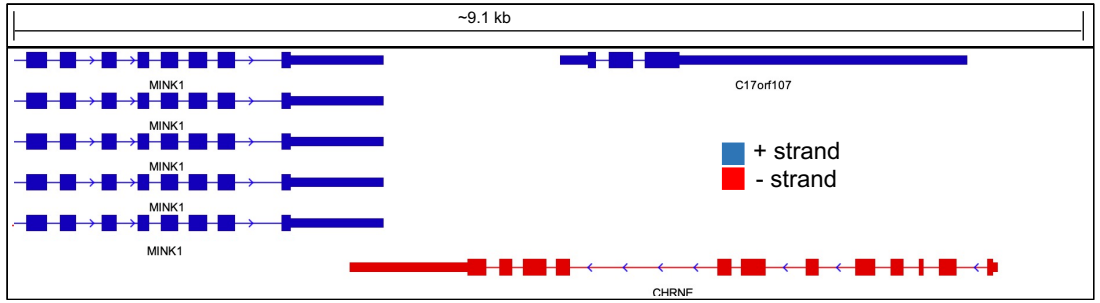
Problem:

- Unstranded protocols contain 2 cDNA populations:
  - Sequence corresponds to RNA ***'sense'*** strand
  - Sequence corresponds to RNA ***'anti-sense'*** strand

Why do we care?

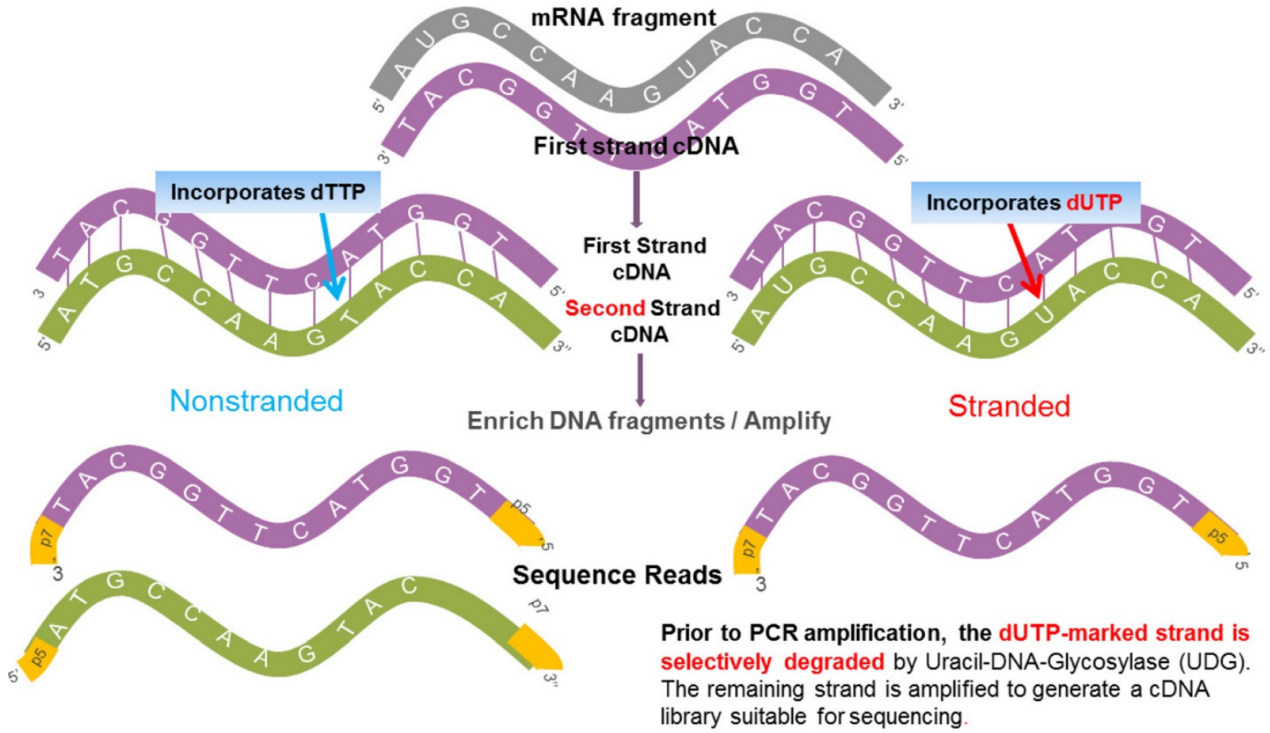
- Strand knowledge critical to assign reads to overlapping features e.g. overlapping genes, anti-sense transcripts

chr17:4,894,452-4,903,779 (GRCh38/hg38)



Stranded library preparation:

- Stranded protocols maintain information of which RNA strand the read came from

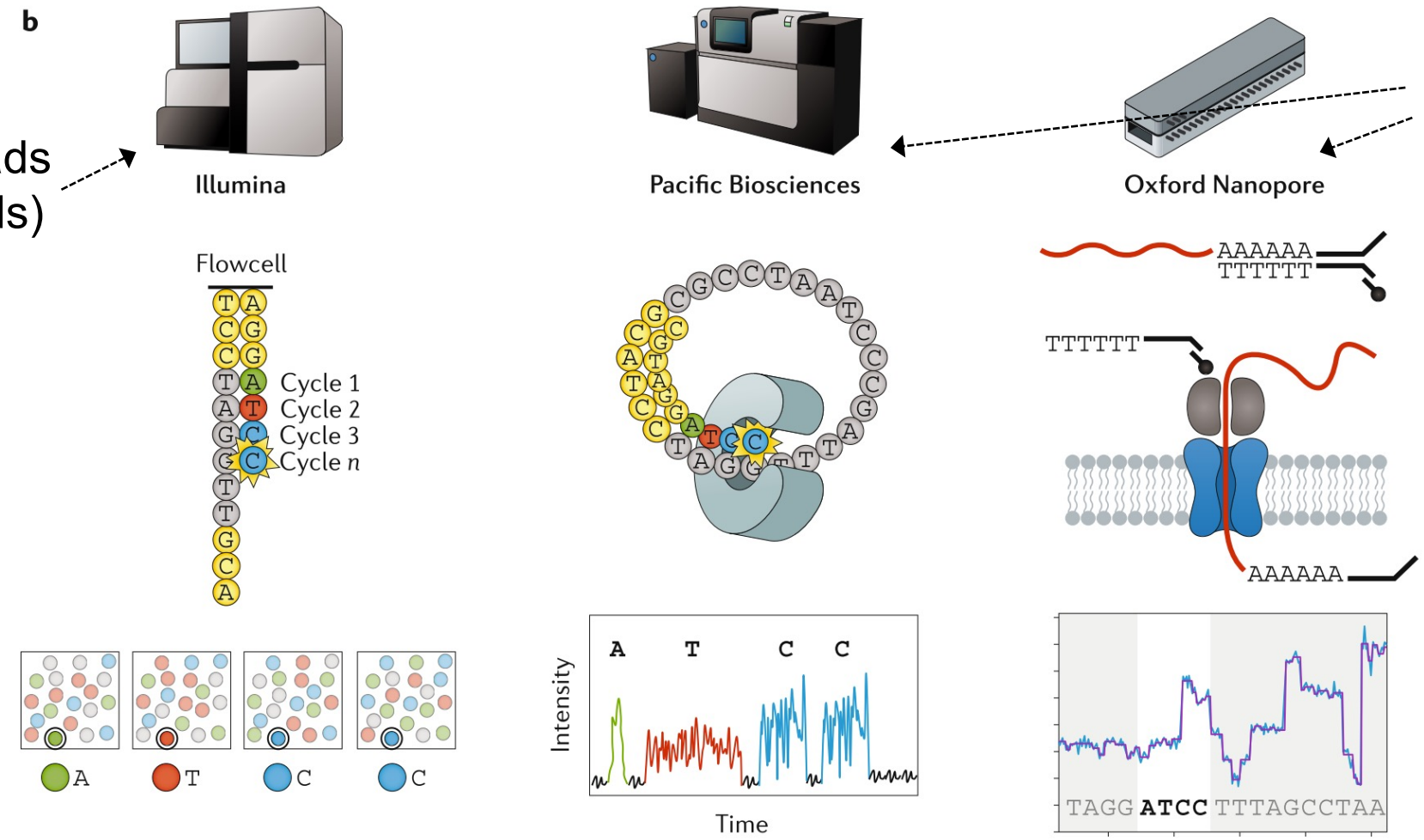


Zhao *et al,* Genome Biol. 2015.

Sequencing technologies for RNA-seq

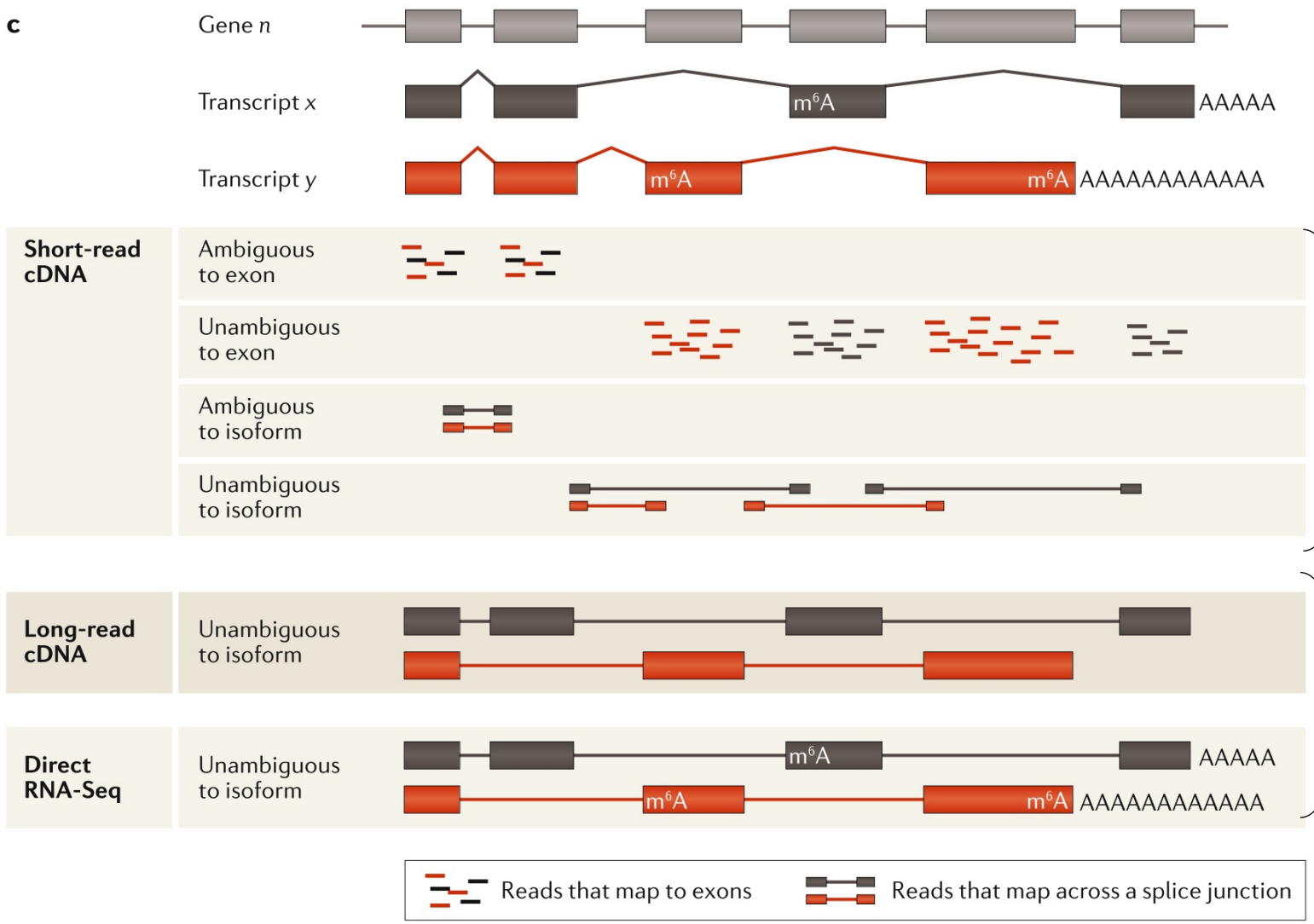
Short reads (hundreds)

Very long reads (kb)



Stark *et al,* 2019, *Nat Reviews genetics*

Different tech., different data., different applications



Best for standard differential gene/transcript expression analysis

- Isoform discovery
- De novo transcriptome
- Fusion transcripts

Stark *et al,* 2019, *Nat Reviews genetics*

FASTQ files

- 4 lines per record
- Single file or 1 per direction for paired-end reads (R1.fastq vs. R2.fastq)

Single FASTQ entry:

```
@SRR1039508.1 HWI-ST177:290:C0TECACXX:1:1101:1225:2130 length=63
CATTGCTGATACCAANNNNNNNNGCATTCCTCAAGGTCTTCCTCCTTCCCTTACGGAATTACA
+
HJJJJJJJJJJJJJ########00?GHIJJJJJJJIJJJJJJJJJJJJJJJJJHHHFFFFFD
```

- Row 1: Header line (starts with @)
- Row 2: Base calls
- Row 3: '+' or line 1
- Row 4: Base qualities

In bash:

```
> zcat $sample.fastq.gz | head –n 12
```

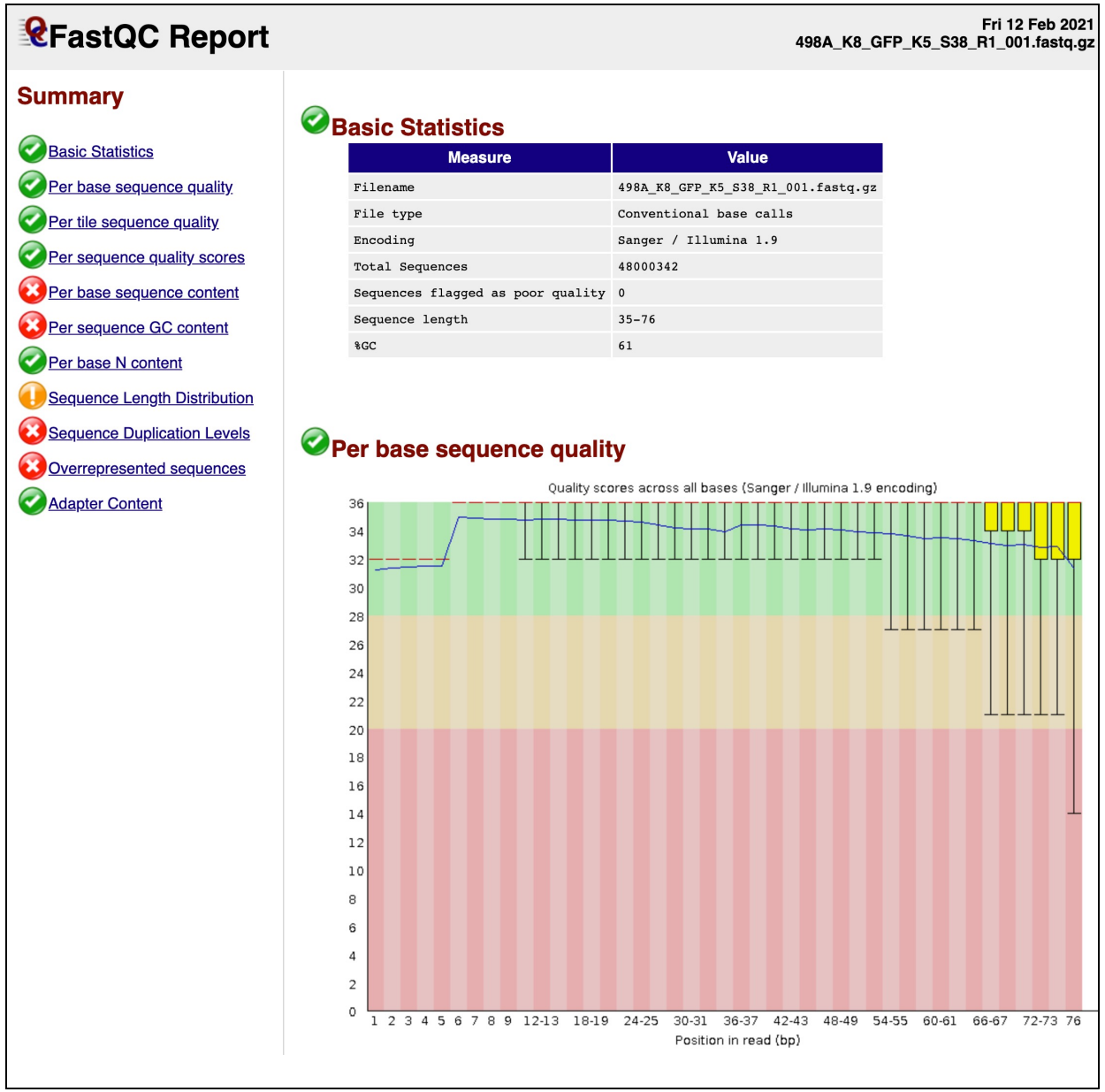
```
[d41294d@discovery7 ATACseq_6-4-19]$ zcat 648-PKA-Cre-C3Tag_S4_R2_001.fastq.gz | head -n12
@NB501031:325:HK52YBGXB:1:11101:9675:1055 2:N:0:ACCACTGT
CTGGGCAGCTGGGGGGTGGGGTGGGGAAGACACAGGAGCCACACAGAGAGAGACATCCCATCCTGTCTCCATTGCT
+
AAAAAEEEEAEEEEEE/E/EE6EEEEEEEAEEEEEEA//EEEEEAEEEAEEE/E/EEEA<EEEEEEEEEEE/EAEE
@NB501031:325:HK52YBGXB:1:11101:7613:1055 2:N:0:ACCACTGT
GTGTTAGGGACTTCTCAAGGAAGTTCTATAGATAGAGCCCAGTACTCTTGAGTGACAGGGGGAACATCTGTGTAAA
+
6AA6AEAAE/EEEEEEEEEEEE/AEAEEEEEE6EEEAE/EEEE//EE/AEE/E/<EEEEEEE/EEEAEE<EAAEE
@NB501031:325:HK52YBGXB:1:11101:16664:1055 2:N:0:ACCACTGT
ATGCTGGAGTTCTGTGGCCACCACCTTCTCAAACACTCATTCATCCATTGAATGGGACATAGGGGACAATAGTGAT
+
AAAAAEEAEAEEEEEEEA</A<AEAEEEEEEA/E//EEEEEAEEEEEA/EE/EEEEEEEAA6/EA/EE/</EEE/E
```

Phred Score	Probability of incorrect call	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Raw data quality control

- Metrics can be calculated from the raw sequences (FASTQs) that inform on quality of RNA-seq data
  - Seq/base quality
  - Per seq. GC content
  - Adapter content
  - GC content
  - Over-represented sequences (e.g. rRNA)
- QC metrics inform how (and if) you analyze the data
- Software exists that will calculate these metrics for you
- Or you could calculate it yourself…

HTML report:

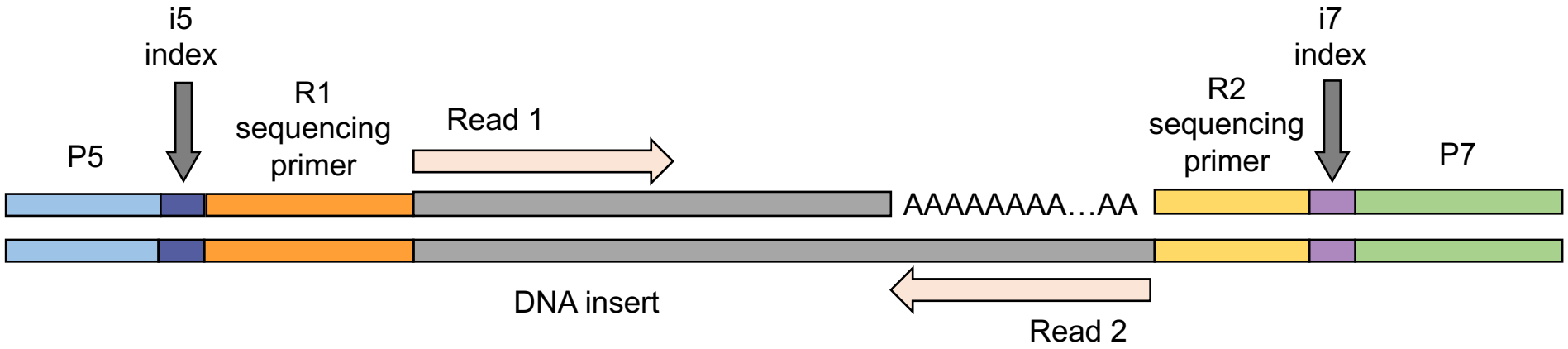


Read trimming (pre-processing)

Commonly trimmed sequences:
- Sequencing adapters
- PolyA tails
- Low quality bases
- 'N' base calls (no call)

- Process of 'cleaning' NGS reads to improve **speed** & **quality** of downstream analysis
- NGS reads contain sequences not wanted for downstream steps:
- Trimming algorithms search reads for specified sequences and remove matches based on user specified behavior

Typical Illumina RNA-seq library

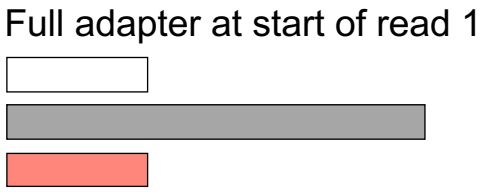


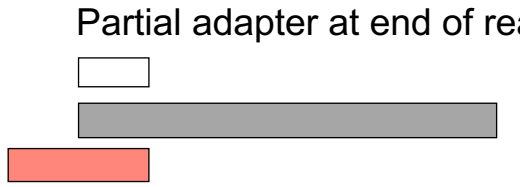
Legend:

DNA insert

Adapter

Trimmed



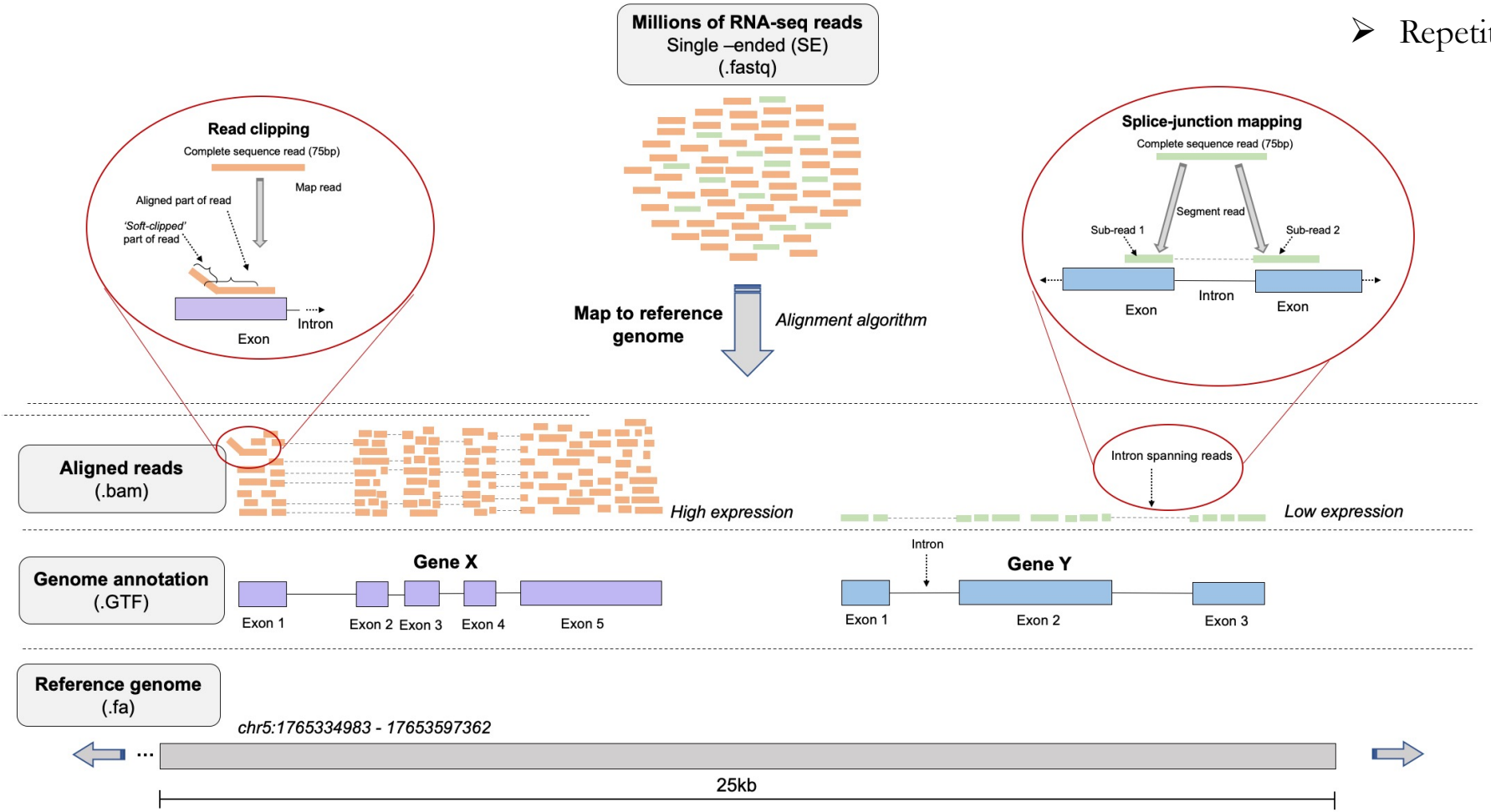


Read mapping to a reference genome

- Reads can be aligned to a reference genome for well studied organisms. e.g. hg38, mm10
- **Where is the most likely origin for a read in a reference genome?**

Challenges:

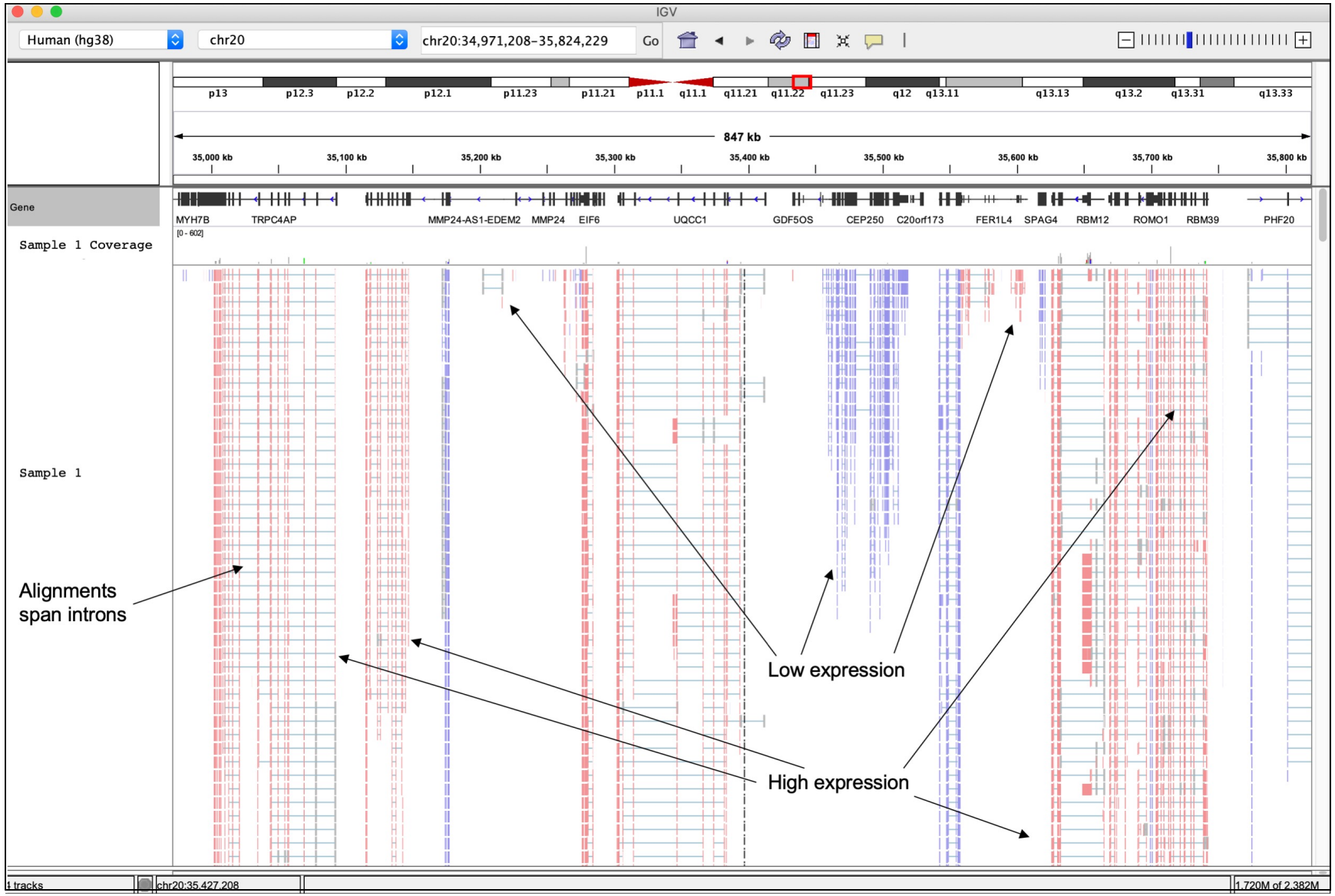
- Enormous search space
- Computationally intensive
- Genome vs transcriptome
- Genetic variation
- Repetitive sequences



Conesa et al, Genome Biol, 2016, DOI 10.1186/s13059-016-0881-8

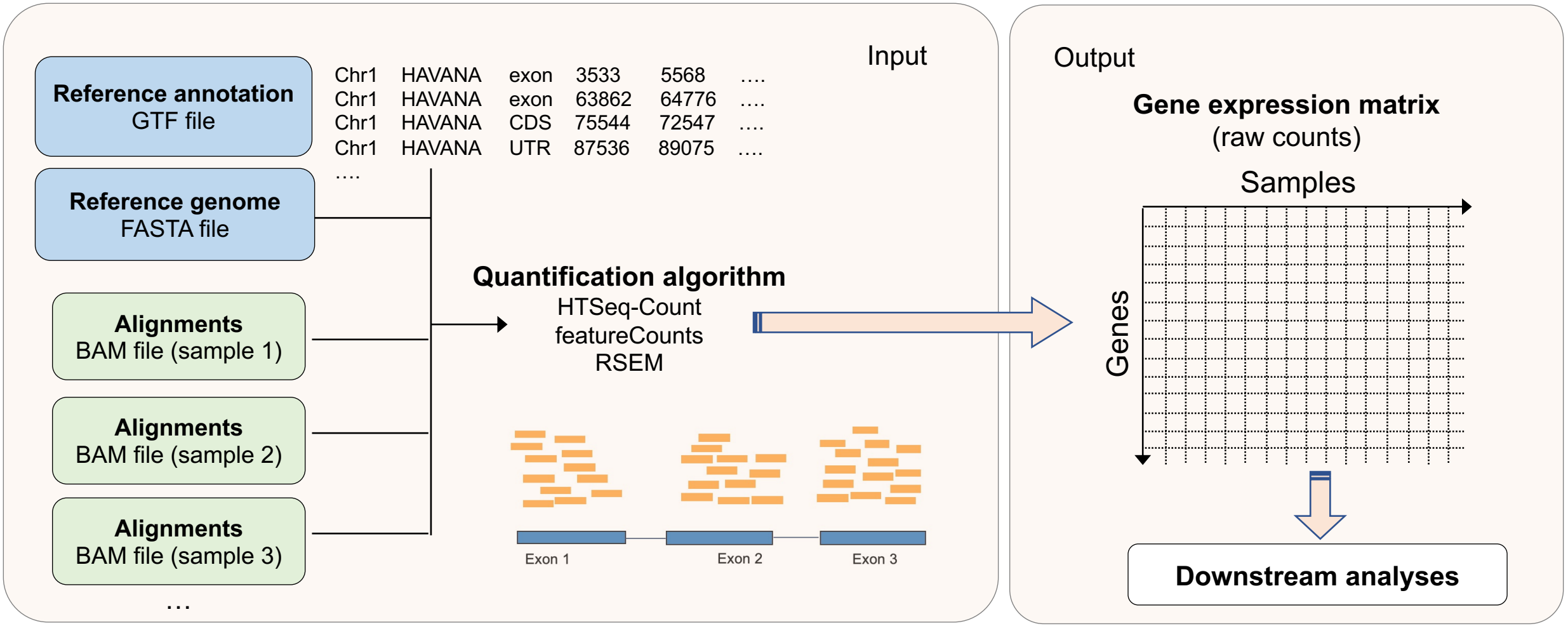
Aligned reads in browser

- Real data example using Integrative Genomic Viewer (IGV)
- This is just <u>1 sample</u>

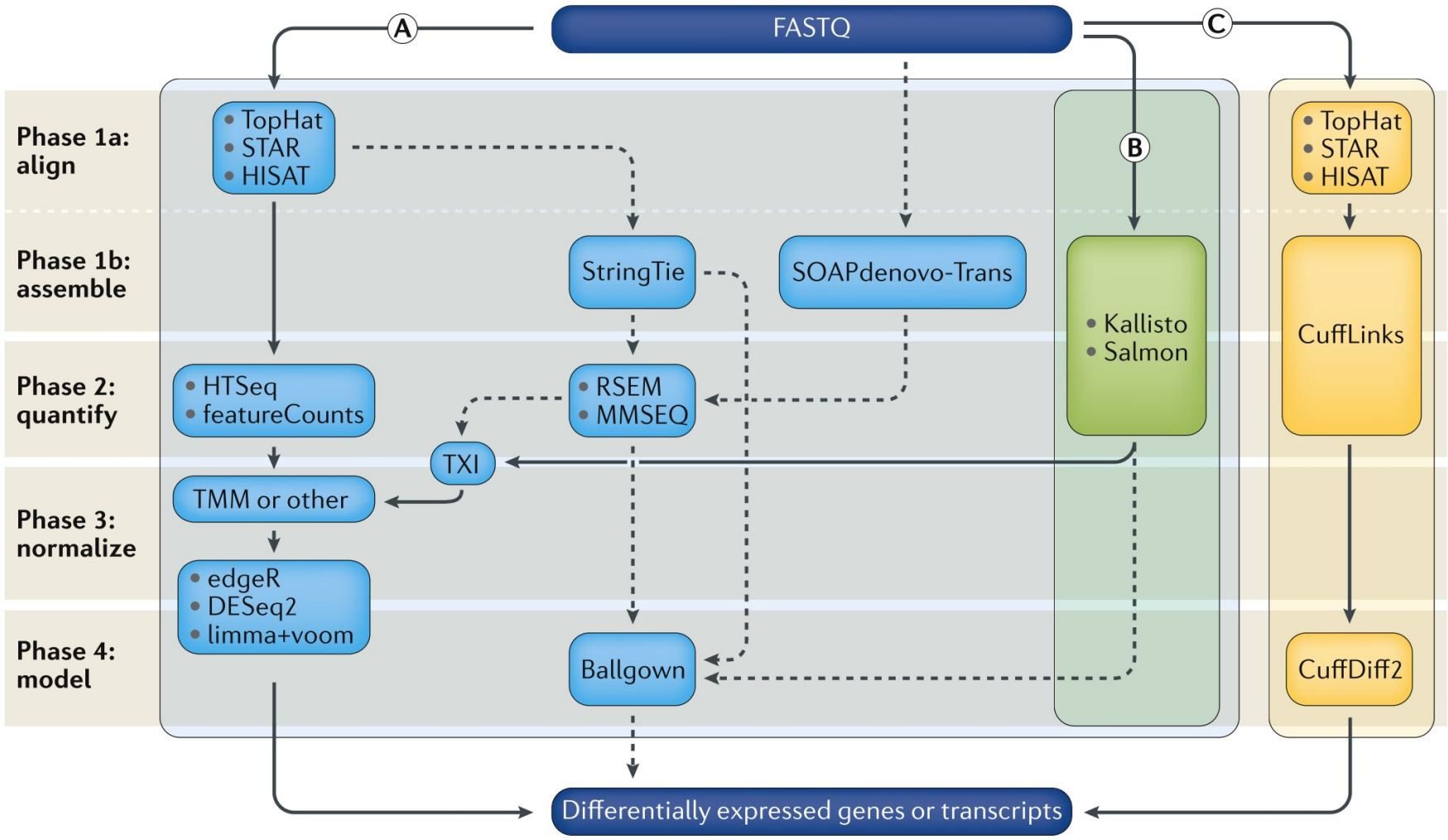


Quantification (read counting)

- Count reads at each genomic position for desired feature of interest (usually exons for RNA-seq)



Differential expression workflow(s)



- Several tools available for each step
- Each have various strengths, weaknesses, applications

Stark *et al,* 2019, *Nat Reviews genetics*

Sample preparation

- **Be consistent with sample prep**
  - Practice protocol 1st, don't get better over course of collecting more samples
- **Minimize batches**
  - 1 batch is ideal, otherwise, smallest number possible
  - MUST randomly distribute samples from experimental conditions across batches
  - Treat each batch EXACTLY the same
- **Collect replicates**
  - Statistics cannot be done on one sample! (statistics is study of populations)
  - The more you collect, the more power you have to discover DEGs
  - Make each replicate as similar as possible e.g. same passage number of cell line
- **Work with your genomics core (they do this a lot)**
- **Pilot experiments can be valuable**

Its well worth spending time to create a high-quality dataset upfront, rather than trying to improve & rescue it later

Replicates

- Arguably more important than read depth or length for DE

How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?

NICHOLAS J. SCHURCH,[1,6] PIETÀ SCHOFIELD,[1,2,6] MAREK GIERLIŃSKI,[1,2,6] CHRISTIAN COLE,[1,6] ALEXANDER SHERSTNEV,[1,6] VIJENDER SINGH,[2] NICOLA WROBEL,[3] KARIM GHARBI,[3] GORDON G. SIMPSON,[4] TOM OWEN-HUGHES,[2] MARK BLAXTER,[3] and GEOFFREY J. BARTON[1,2,5]

[1]Division of Computational Biology, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom
[2]Division of Gene Regulation and Expression, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom
[3]Edinburgh Genomics, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom
[4]Division of Plant Sciences, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom
[5]Division of Biological Chemistry and Drug Discovery, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom

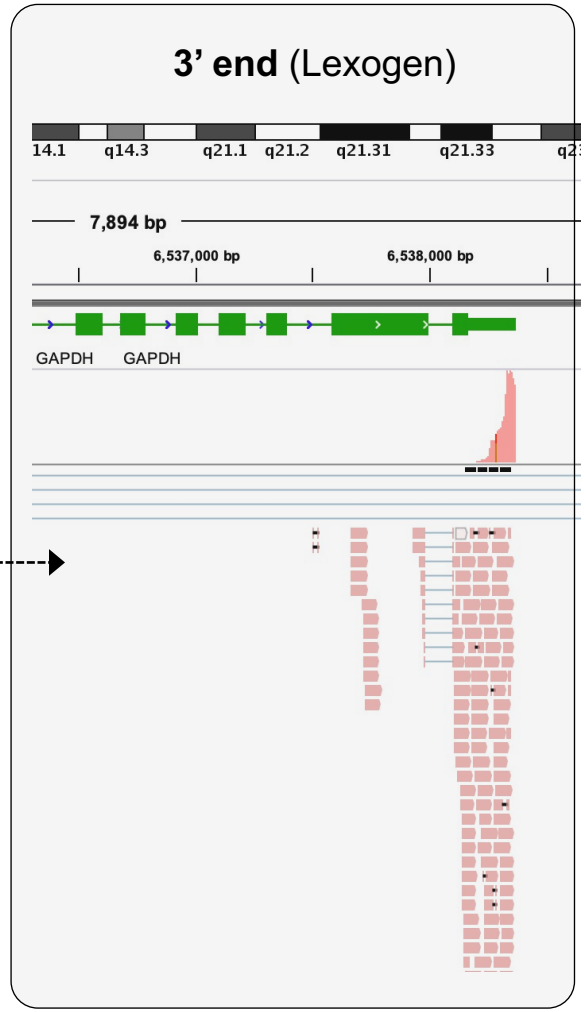
"With 3 biological replicates, 9 of the 11 tools evaluated found only 20%–40% of the significantly differentially expressed (SDE) genes"

- Schurch *et al,* RNA, 2016

- Suggested minimum no. of replicates should = 6
- More heterogeneity = more replicates needed (e.g. human tissues vs. cultured cell lines)

Sequencing depth (or coverage?)

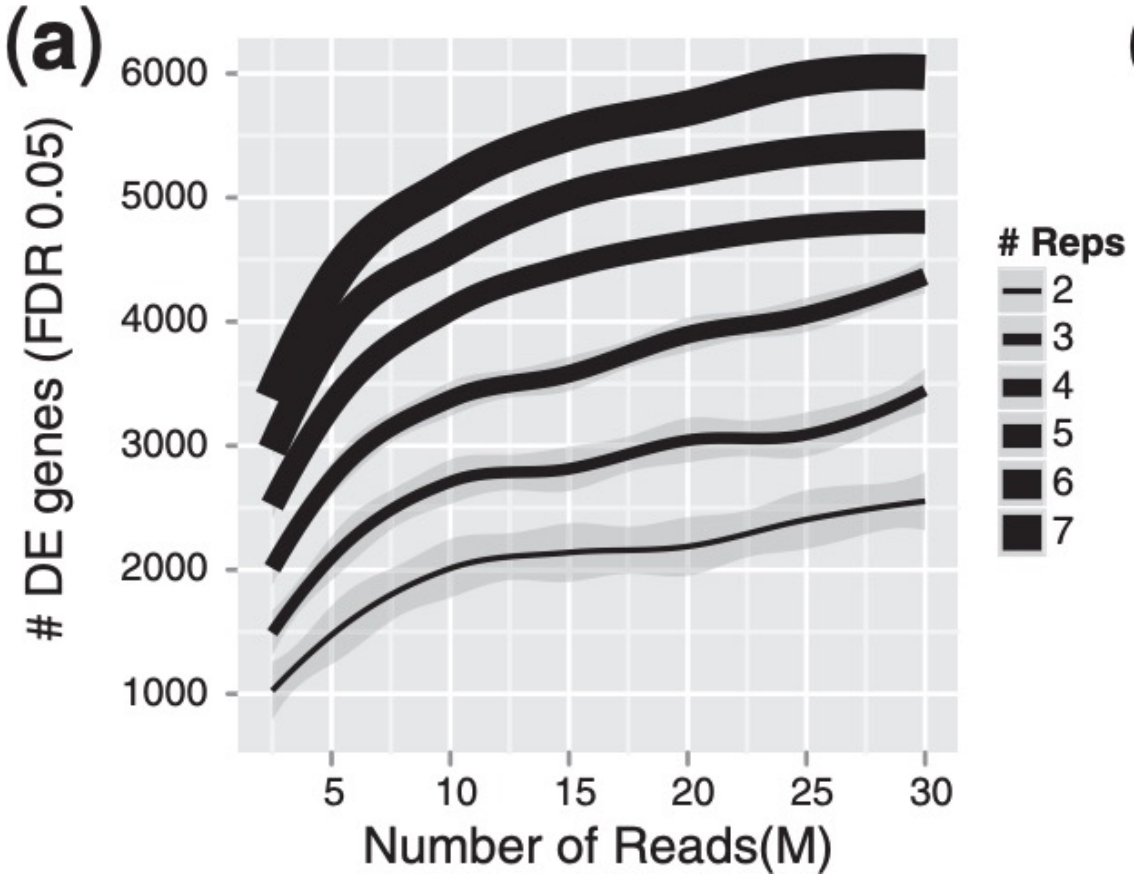
- '*Coverage*' doesn't have much meaning for transcriptome data
  - We are less concerned with how many times we cover a specific locus w/ a read..
- Generally total of 10-30 million reads for DGE of eukaryotic genomes
- Some species require many fewer than this
- Technology also affects required read number (3'-end data needs fewer)
- Checking saturation can help you assess if you've sequenced enough (next slide)
- Try to avoid generating libraries of differing complexity (vastly different reads nos.)



Depth vs. Replicates

Which is more important?

- No. of DEGs increases w/ replicate no.
- Diminishing returns after 10-15M reads (for this dataset)
- Additional replicates are more valuable than sequencing really deeply (for DE analysis)

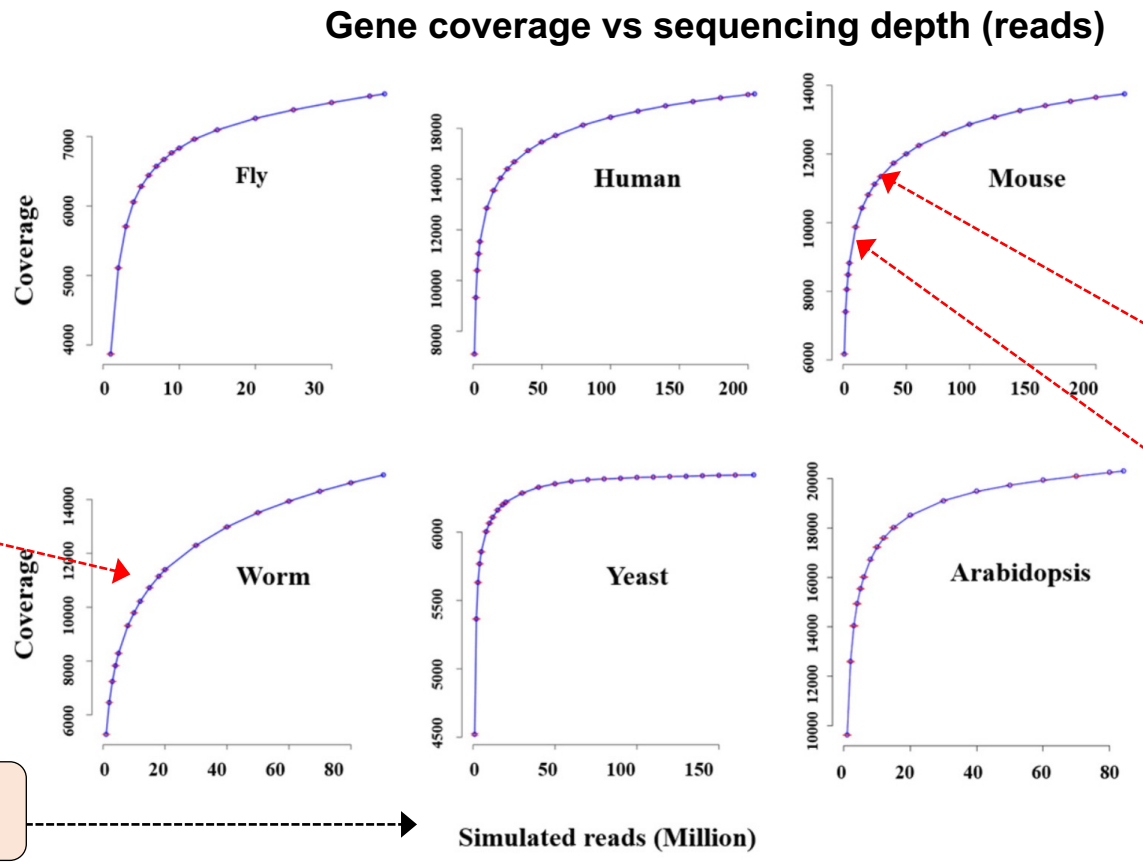


Liu *et al*, 2014, *Bioinformatics*

Sequencing depth

- Saturation curves can help you figure out if more sequencing will improve power

No. of features (genes) detected with at least 10 reads



- If you don't reach saturation, you can always do more sequencing!

Get data points by subsampling reads

Try to avoid generating 1 sample here, and another here

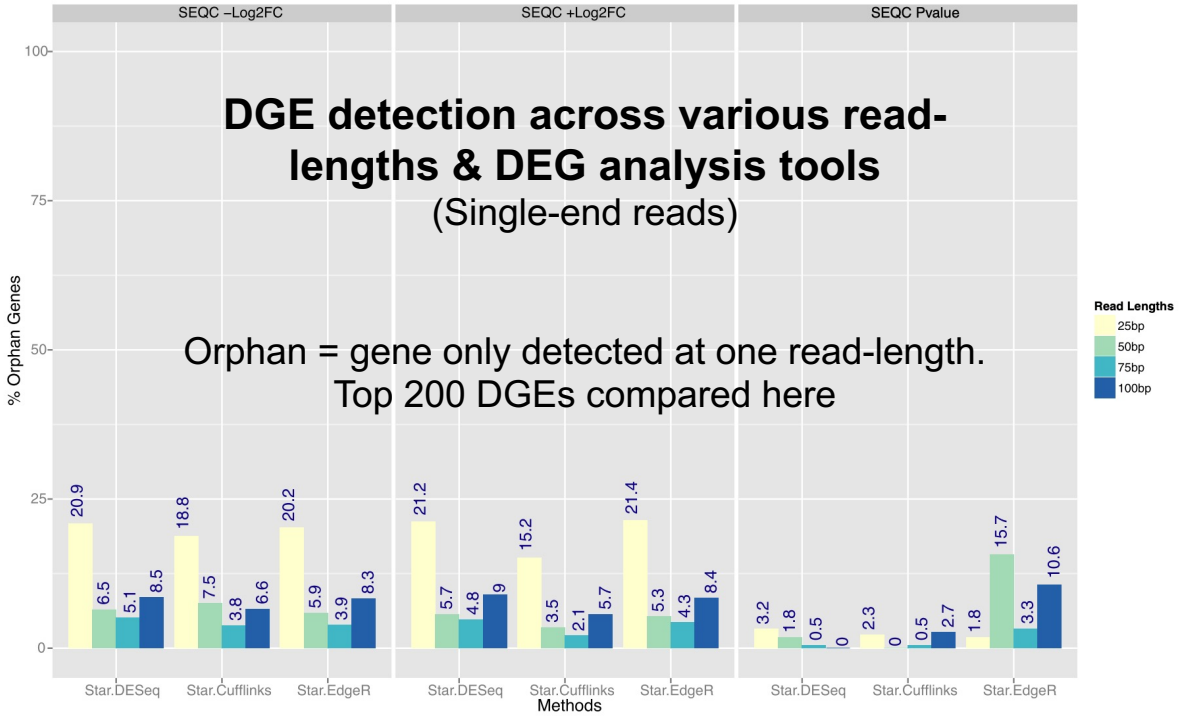
Different complexities can introduce bias!

Lei *et al*, 2015, *Gene*

Read length

- For DGE, we want the **MINIMUM read length required to accurately map** a read to a gene

Shorter vs. longer-reads

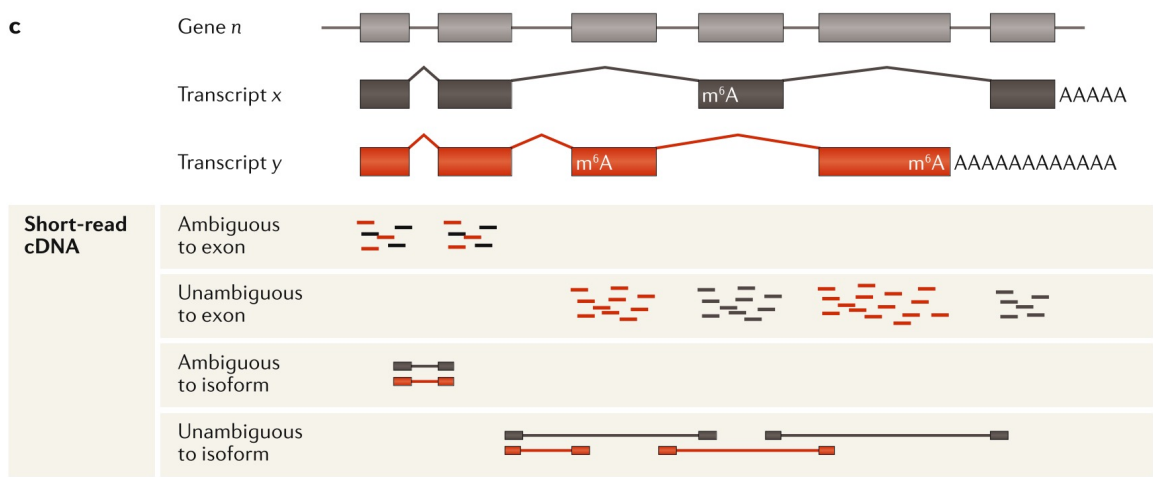


Chhangawala *et al*, 2015, *Genom. Biol.*

- For DGE, invest in **more replicates** & **more reads**

- For other applications, we need **longer & PE reads** to unambiguously map to transcripts

Paired-end vs. single-end



Stark *et al,* 2019, *Nat Reviews genetics*

- For isoform-detection, alternative exon usage, & SNV detection, get **paired-end**

Take home messages

- Work with your genomics core to design an experiment that will address your hypothesis
- Involve your analysis team early (during experimental design)
- You should always know and understand the type of library used to generate your data
- If you don’t perform data pre-processing, you should understand the basics steps involved
- Replicates are critical for differential expression analysis

Questions?