



Introduction to RNA-seq

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/>)

07/08/20



Dartmouth
GEISEL SCHOOL OF MEDICINE

Outline



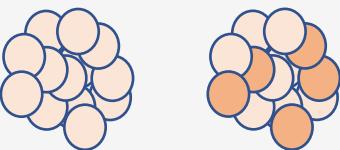
- **RNA-seq overview**
 - Basics of an RNA-seq experiment
 - Sequencing technologies for RNA-seq
- **Library types for RNA-seq**
 - Poly-A, 3'-end, Ribodepletion
 - What hypotheses can be tested with each?
- **Sample preparation & experimental design**
 - Replicates
 - Sequencing depth & configuration
- **Data analysis**
 - Where does the Bioinformatician fit in?
 - Basic analytical steps for differential expression (DE)

RNA-seq: Overview

Sample/library preparation

Biological sample(s)

Sample 1
Control
Sample 1
Treated



RNA isolation

Purification &
selection

Library preparation

Fragmentation, reverse
transcription, adapters, PCR

Sequencing

cDNA pool from samples



**Reads in
FASTQ format**

Demultiplexing

Separate FASTQs obtained for each
sample using '**demultiplexing**'

```
[d41294d@discovery7 ATACseq_6-4-19]$ zcat 648-PKA-Cre-C3Tag_S4_R2_001.fastq.gz | head -n12
@NB501631:325:HK52YBGBX:1:11101:9675:1055 2:N:0:ACCACTG
CTGGGCAGCTGGGGGTGGGGAGACACAGGAGCCACAGAGAGACATCCCATTCTGTCTCATTGCT
+
AAAAAEEAEAAAA//E/EEEEEEEEAEAAAA//EEEEAEAAEEE/E/EEA<EEEEEEEEE/EAAE
@NB501631:325:HK52YBGBX:1:11101:7613:1055 2:N:0:ACCACTG
GTGTTAGGGACTTCTCAAGGAAGTTCTATAGATAGAGGCCAGTACTCTTGAGTGACAGGGGAACATCTGTAAA
+
GAA6AEAAE/EEEEEEEEE/AEAEAAA6EEEAE/EEE//EE/AEE/E/<EEEEEE/EEEAE<EAEE
@NB501631:325:HK52YBGBX:1:11101:16664:1855 2:N:0:ACCACTG
ATGCTGGAGTTCTGTGCCACCACCTCTAACACTCATTCATCCATTGAATGGGACATAGGGACAATAGTGAT
+
AAAAAEEAEAAAA//A<AEAEAAA/E//EEEEAEAAA/EE/EEEEEEAA5/EA/EE/<EEE/E
```

Data analysis

Downstream analysis

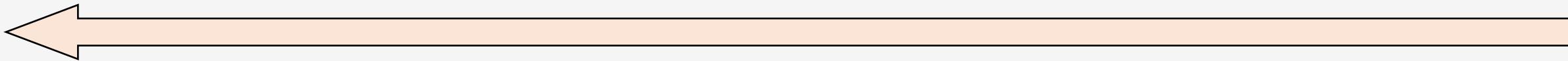
Differential expression
Isoform discovery
Annotation
Variant calling

Data normalization

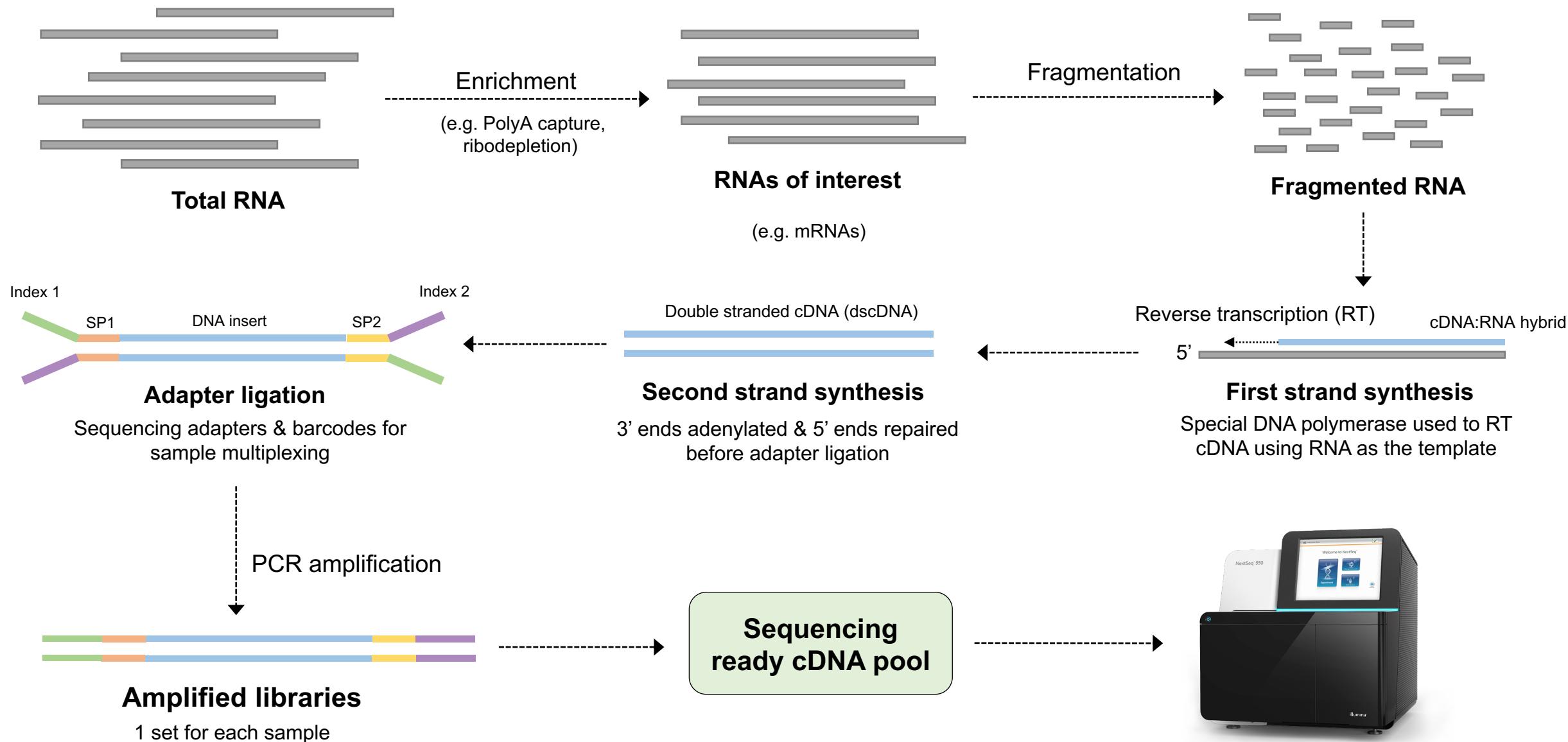
Feature quantification

Genome mapping

Quality control



Library preparation for RNA-seq



RNA selection/enrichment

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

➤ Oligo-d(T) selection:

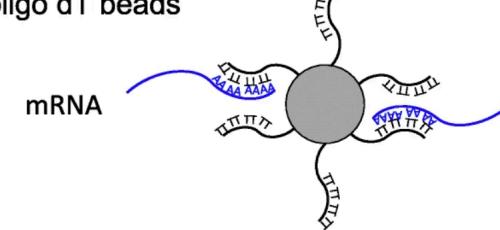
- Uses oligos of dT attached to magnetic beads to capture polyadenylated RNAs (mostly mRNA)
- Magnet used to retain RNAs with polyA sequences

➤ Ribodepletion:

- Oligos complementary to highly conserved rRNA sequences used to capture rRNA
- Streptavidin-bound magnetic beads used to capture and remove hybridized sequences
- Enables enrichment of polyA mRNA & non-polyA ncRNAs (e.g. lncRNAs)

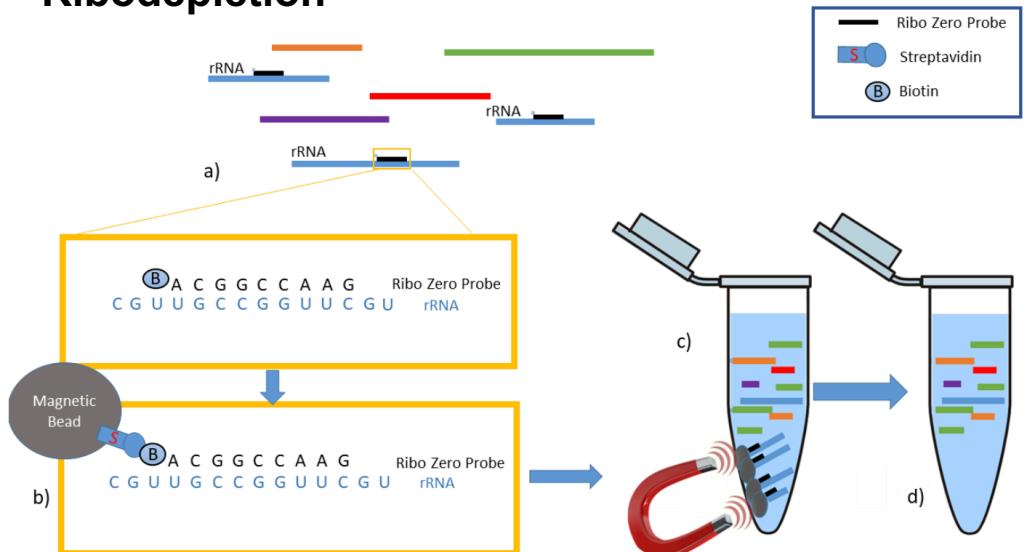
Oilgo-d(T) selection

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



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

Ribodepletion



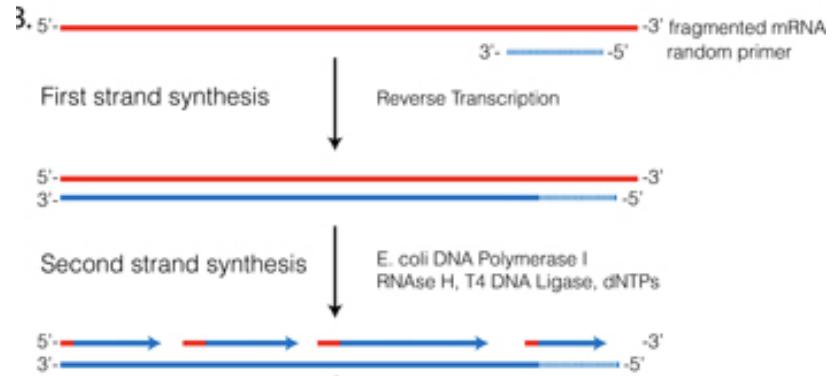
Adapted from Illumina.com

1st & 2nd-strand synthesis

➤ Random priming:

- Random hexamers anneal along length of transcript to facilitate cDNA synthesis
- Generate fragments along full-length of transcript

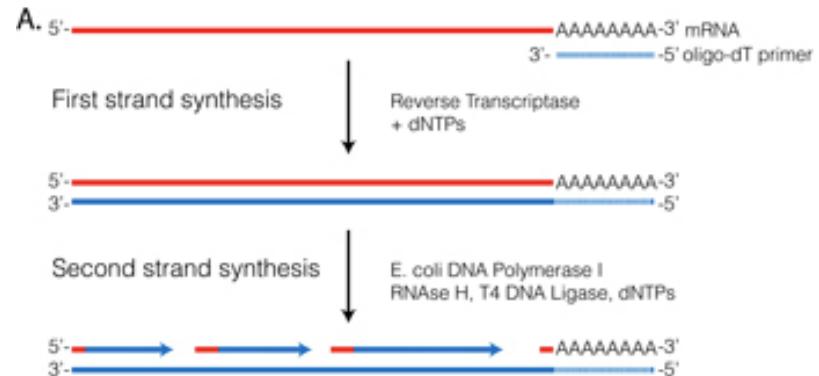
Random priming



➤ Oligo-d(T) priming:

- Oligo-d(T) used directly as a primer
- Enriches for polyA RNAs at same time
- Library fragments will be concentrated at 3'-end only

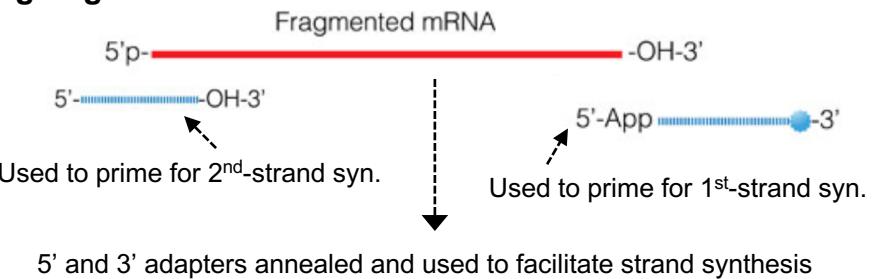
Oligo-d(T) priming



➤ Oligo-ligation & priming

- Oligos containing primer are directly ligated to fragmented RNA

Oligo-ligation



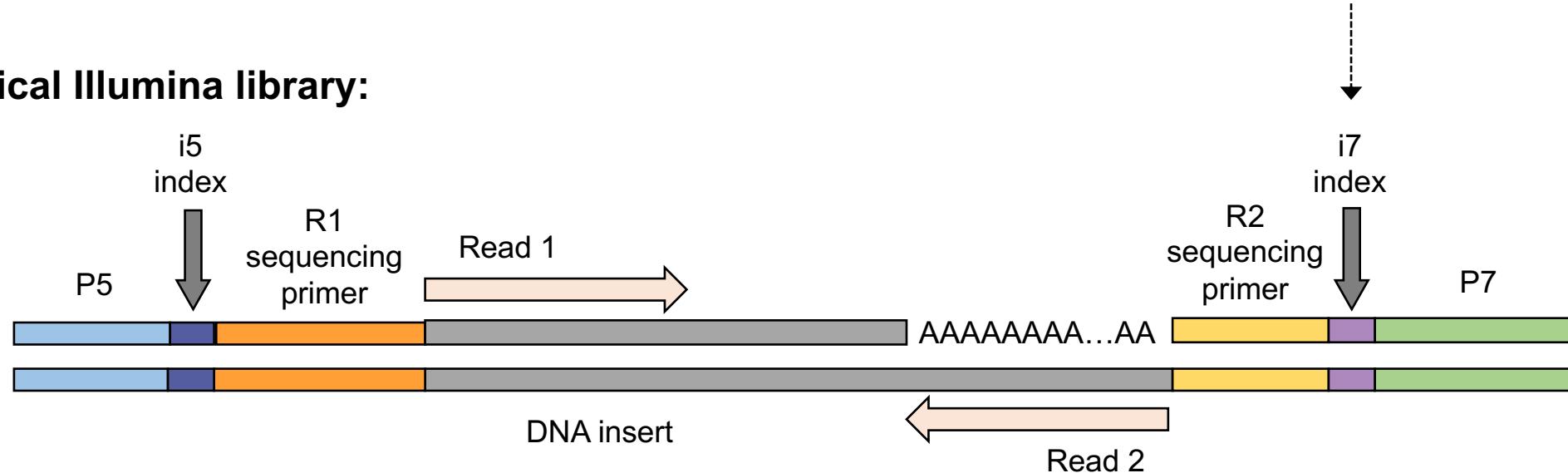
Images adapted from RNA-seqlopedia: UOregon

RNA-seq library structure

- How you choose to sequence is your choice



Typical Illumina library:

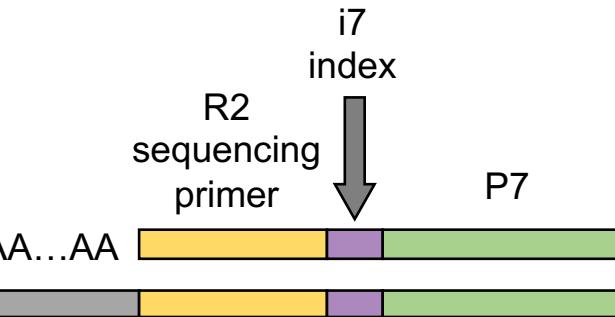


Single-end: Collect R1 only

Paired-end: Collect R1 and R2

- **Choice affected by:**
- Library type
 - Hypothesis

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



Read length: No. of seq. cycles

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

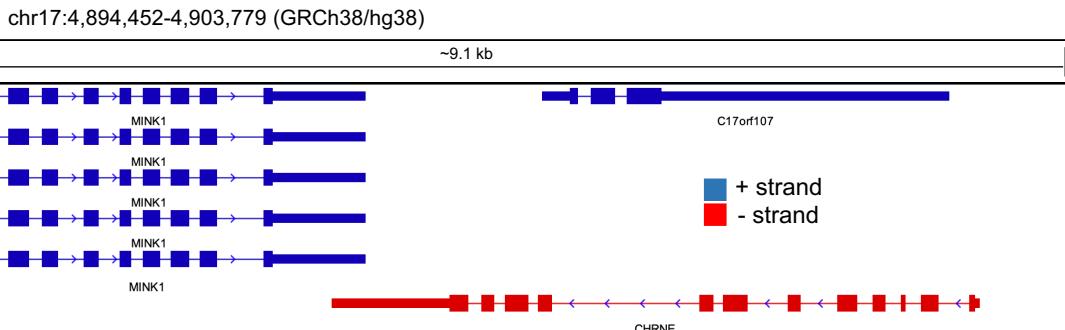
Stranded libraries

Problem:

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

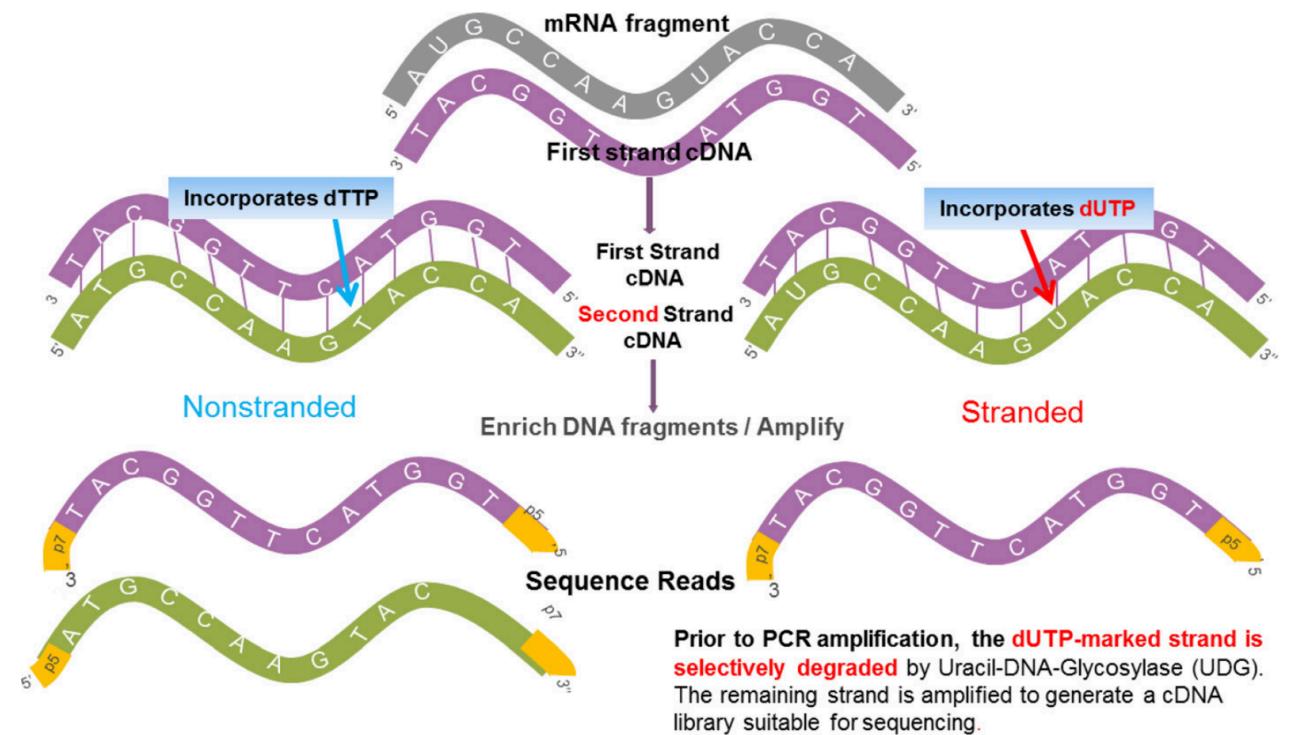
Why do we care?

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



Stranded library preparation:

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



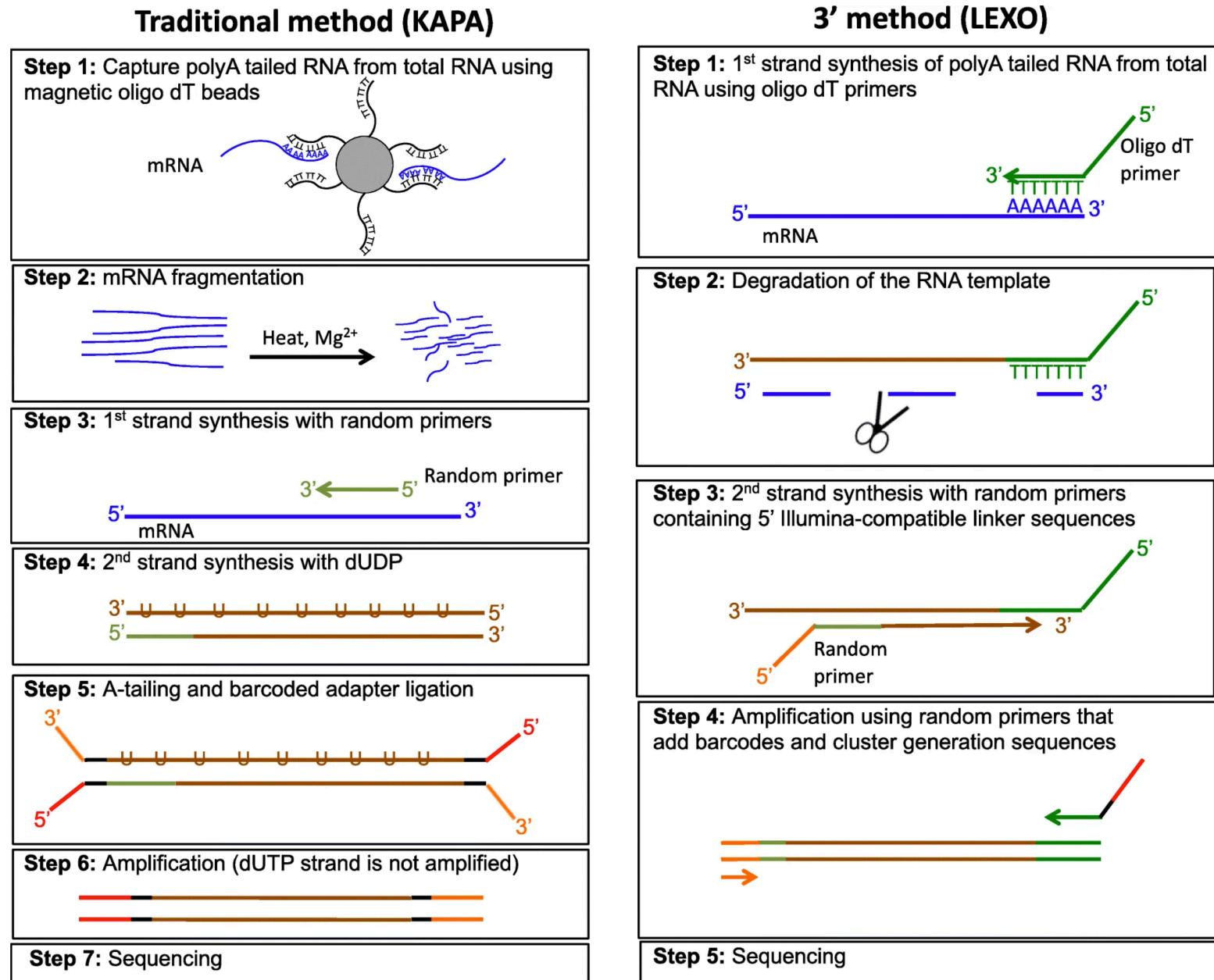
Traditional library prep. vs .3'end

Traditional method

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

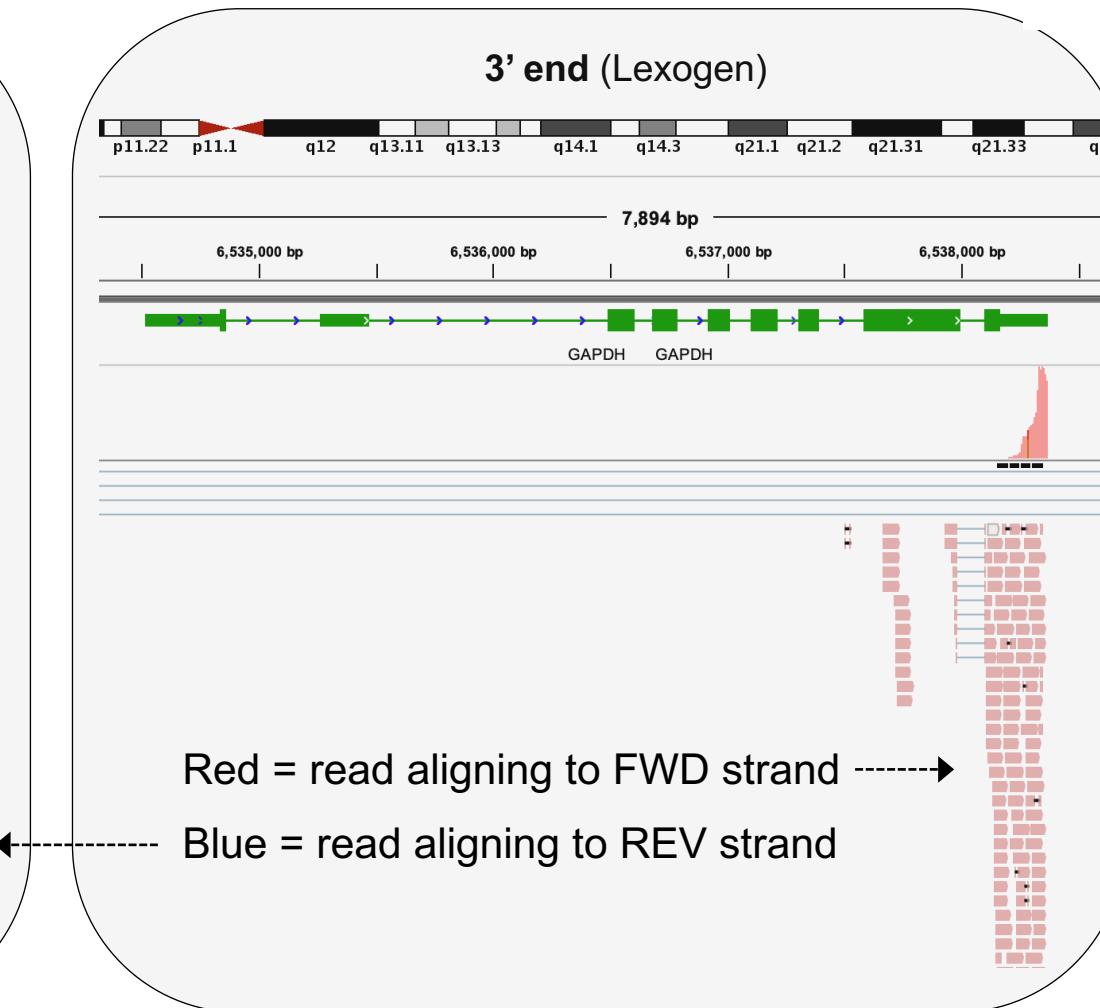
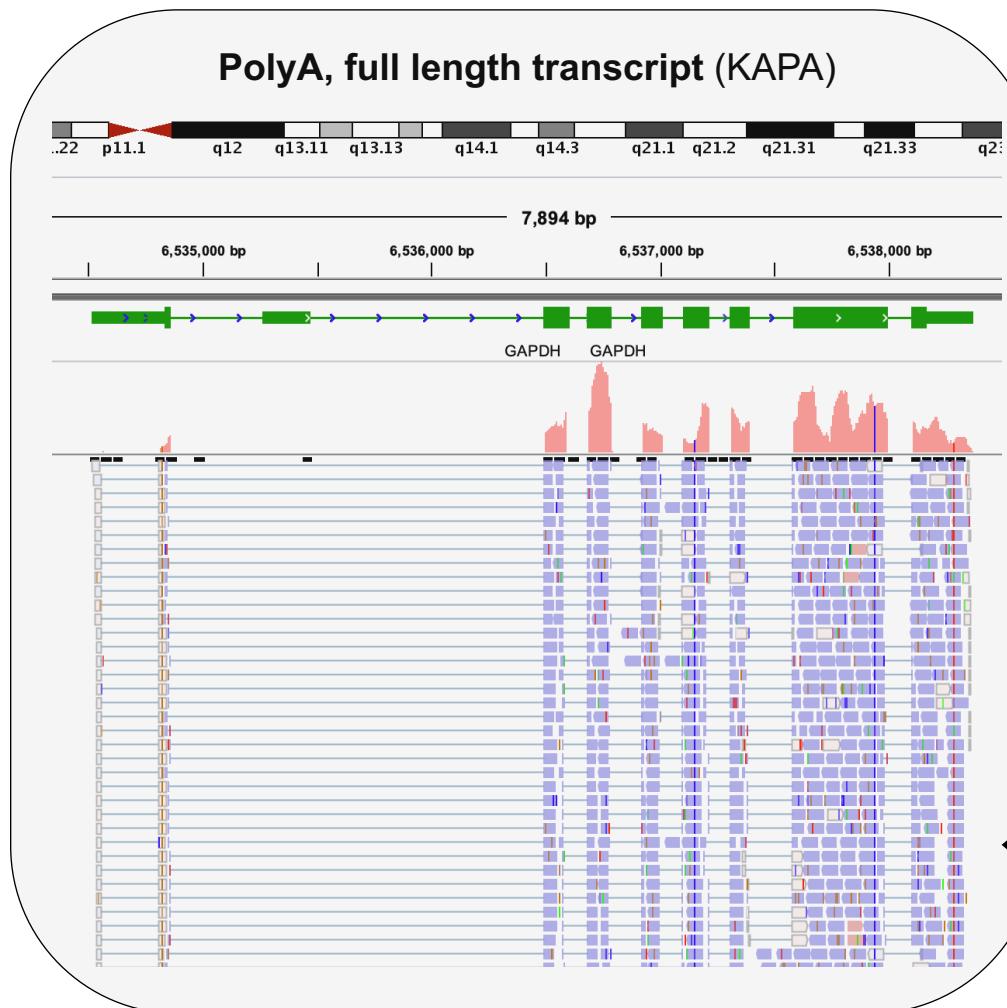
3' method

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



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

Data from different library types looks different



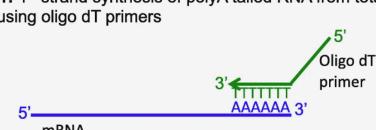
Quality control, analysis pipelines, and possible hypotheses therefore inherently differ

Applications of different library types

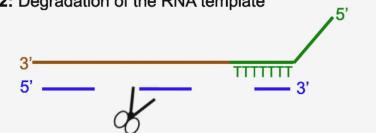
3' End

3' method (LEXO)

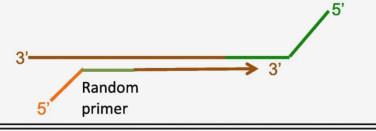
Step 1: 1st strand synthesis of polyA tailed RNA from total RNA using oligo dT primers



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.*

Differential Expression

Lower cost/High Throughput

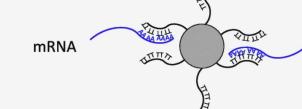
Low Input and Low-Quality Samples

FFPE

Full-length - PolyA

Traditional method (KAPA)

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



Step 2: mRNA fragmentation



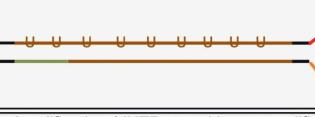
Step 3: 1st strand synthesis with random primers



Step 4: 2nd strand synthesis with dUDP



Step 5: A-tailing and barcoded adapter ligation



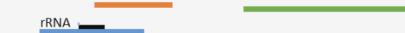
Step 6: Amplification (dUTP strand is not amplified)



Step 7: Sequencing

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

Ribodepletion

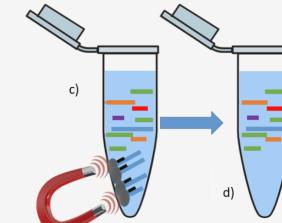


a)

(B) A C G G C C A A G Ribo Zero Probe
C G U U G C C G G U U C G U rRNA

b)

(B) A C G G C C A A G Ribo Zero Probe
C G U U G C C G G U U C G U rRNA



c)

Ribo Zero Probe
Streptavidin
Biotin

d)

Adapted from Illumina.com

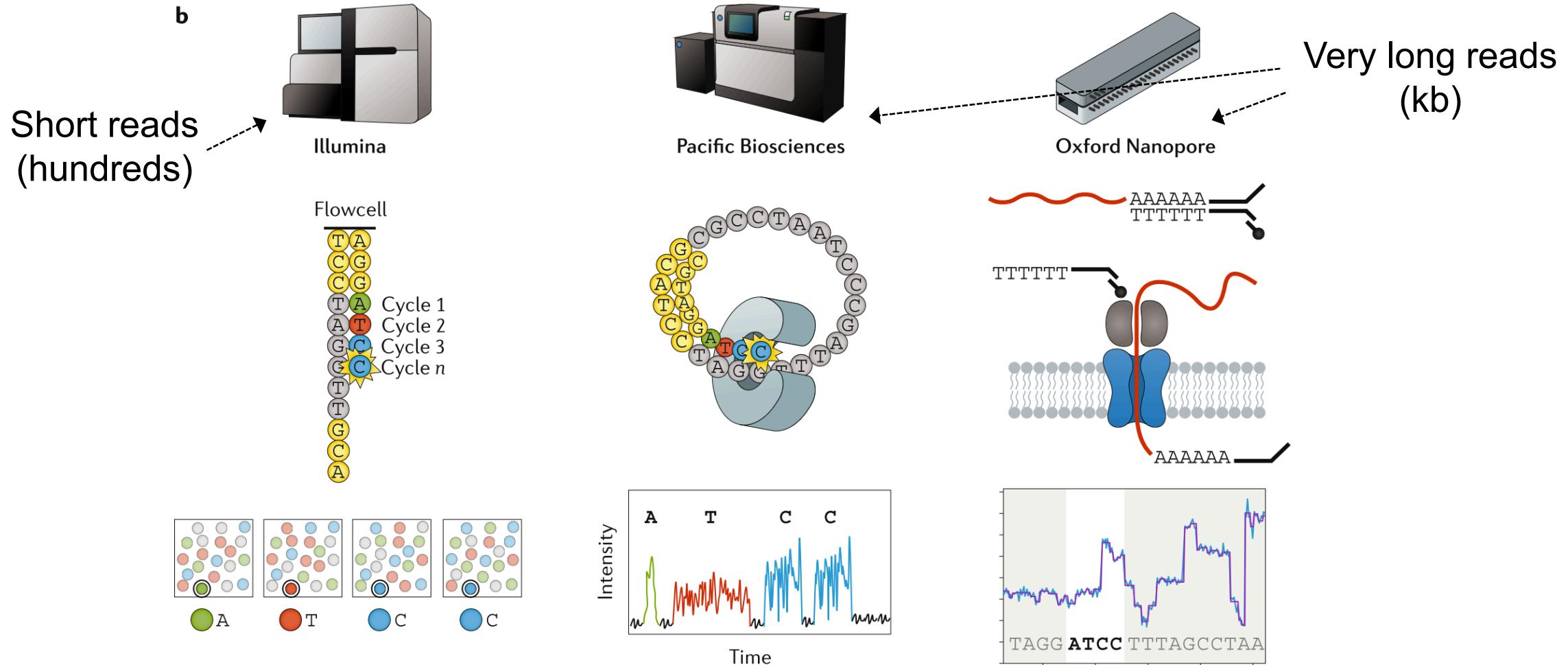
Full length mRNA + lincRNAs
Differential Expression
Splice Variants
SNVs
FFPE

Throughput

Cost / Data Richness

Image Credit:
Lexogen Inc

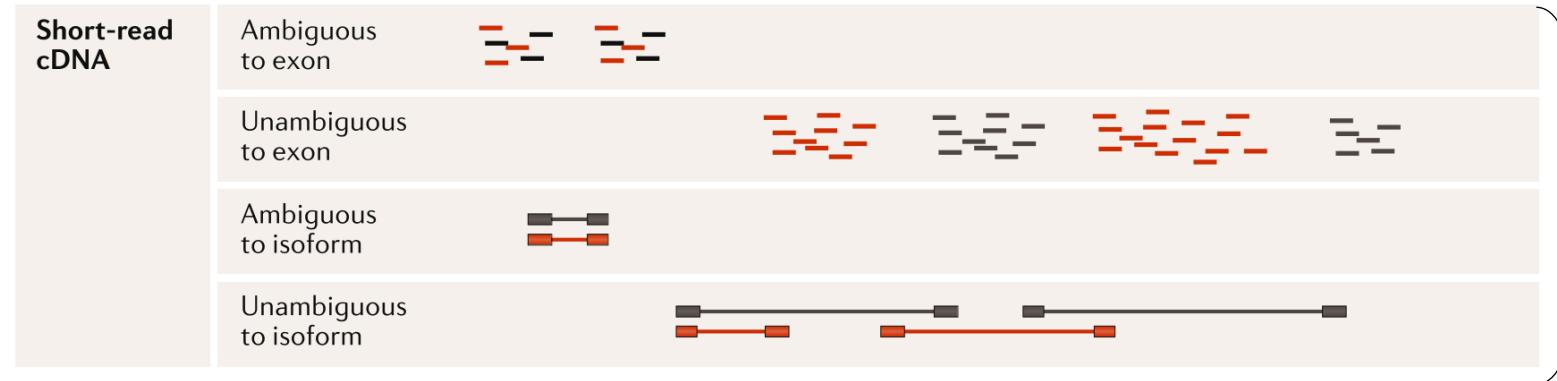
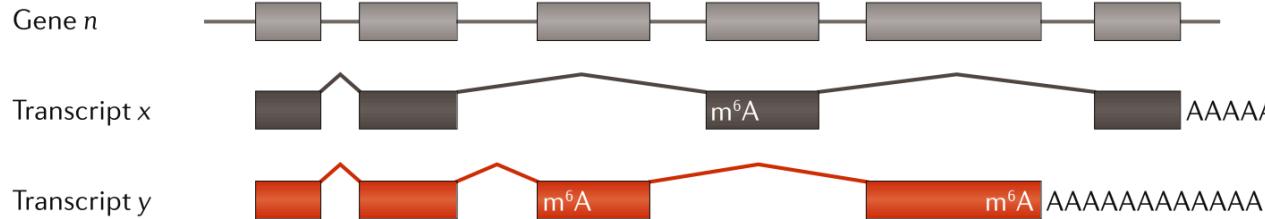
Sequencing technologies for RNA-seq



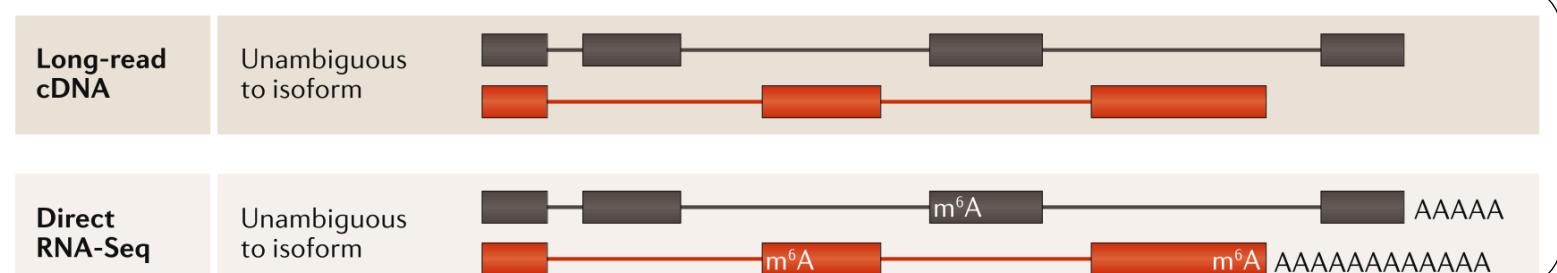
Stark et al, 2019, *Nat Reviews genetics*

Different tech., different data., different applications

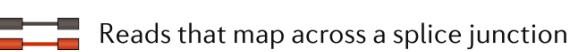
c



Best for standard differential gene/transcript expression analysis



- Isoform discovery
- De novo transcriptome
- Fusion transcripts

 Reads that map to exons  Reads that map across a splice junction

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}

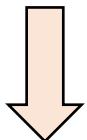
¹Division of Computational Biology, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom

²Division of Gene Regulation and Expression, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom

³Edinburgh Genomics, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

⁴Division of Plant Sciences, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom

⁵Division 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"

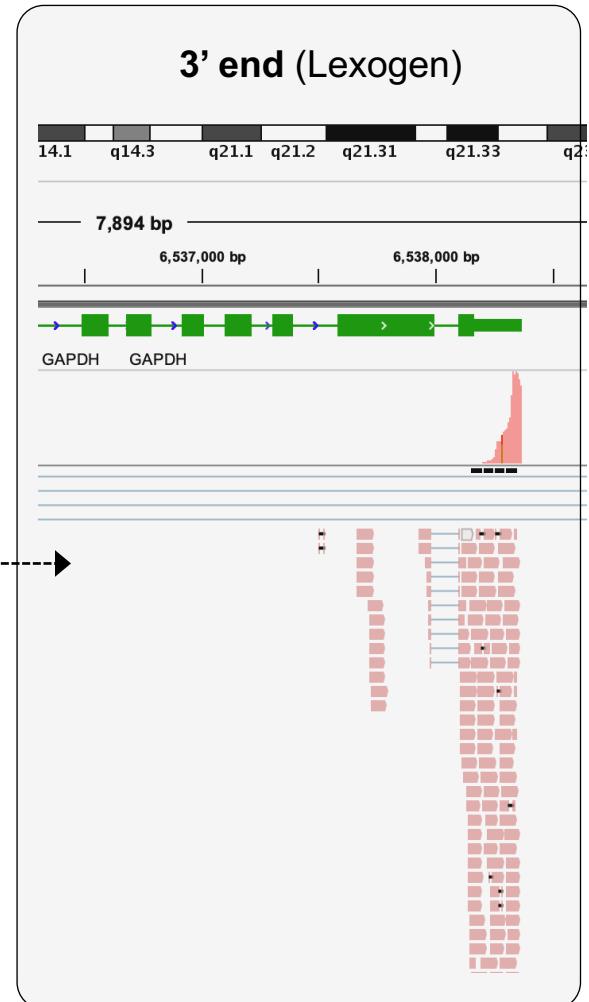
- Church et al, RNA, 2016

- 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.)

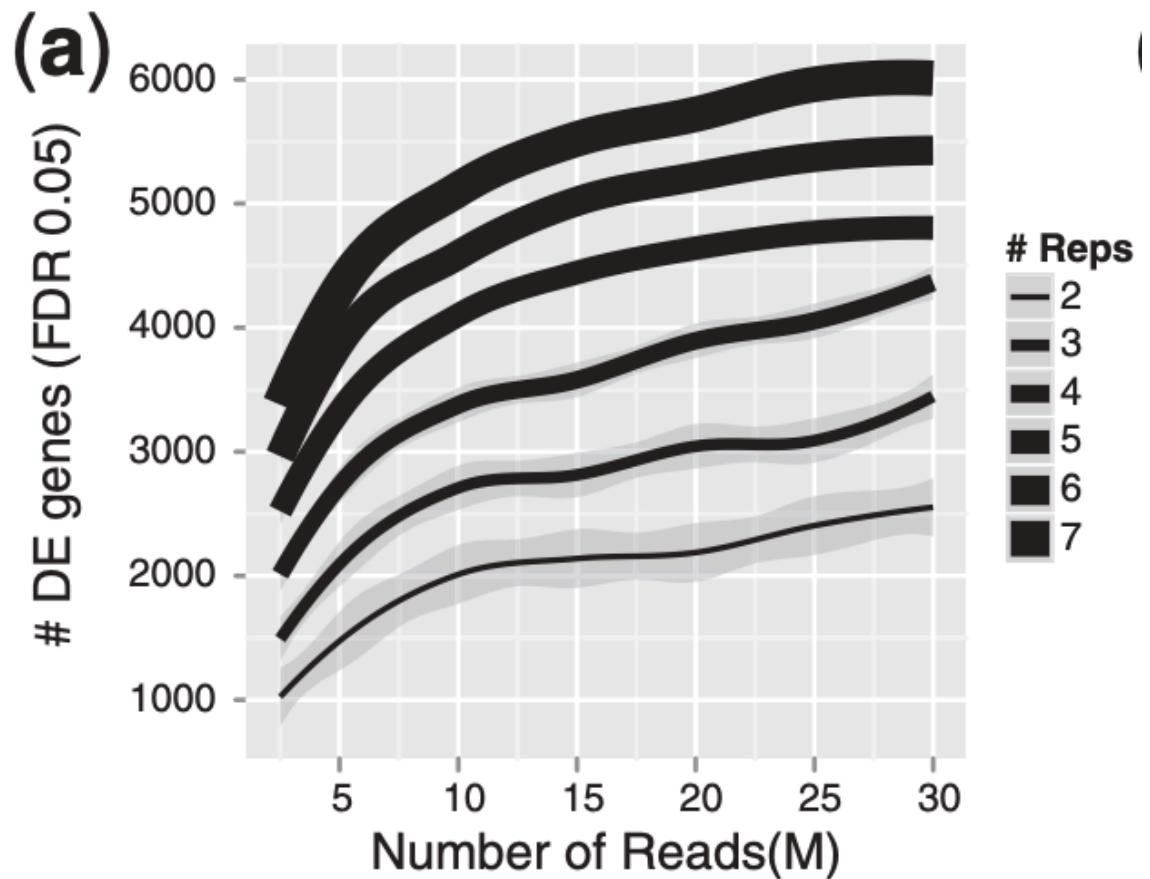


Depth vs. Replicates



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)



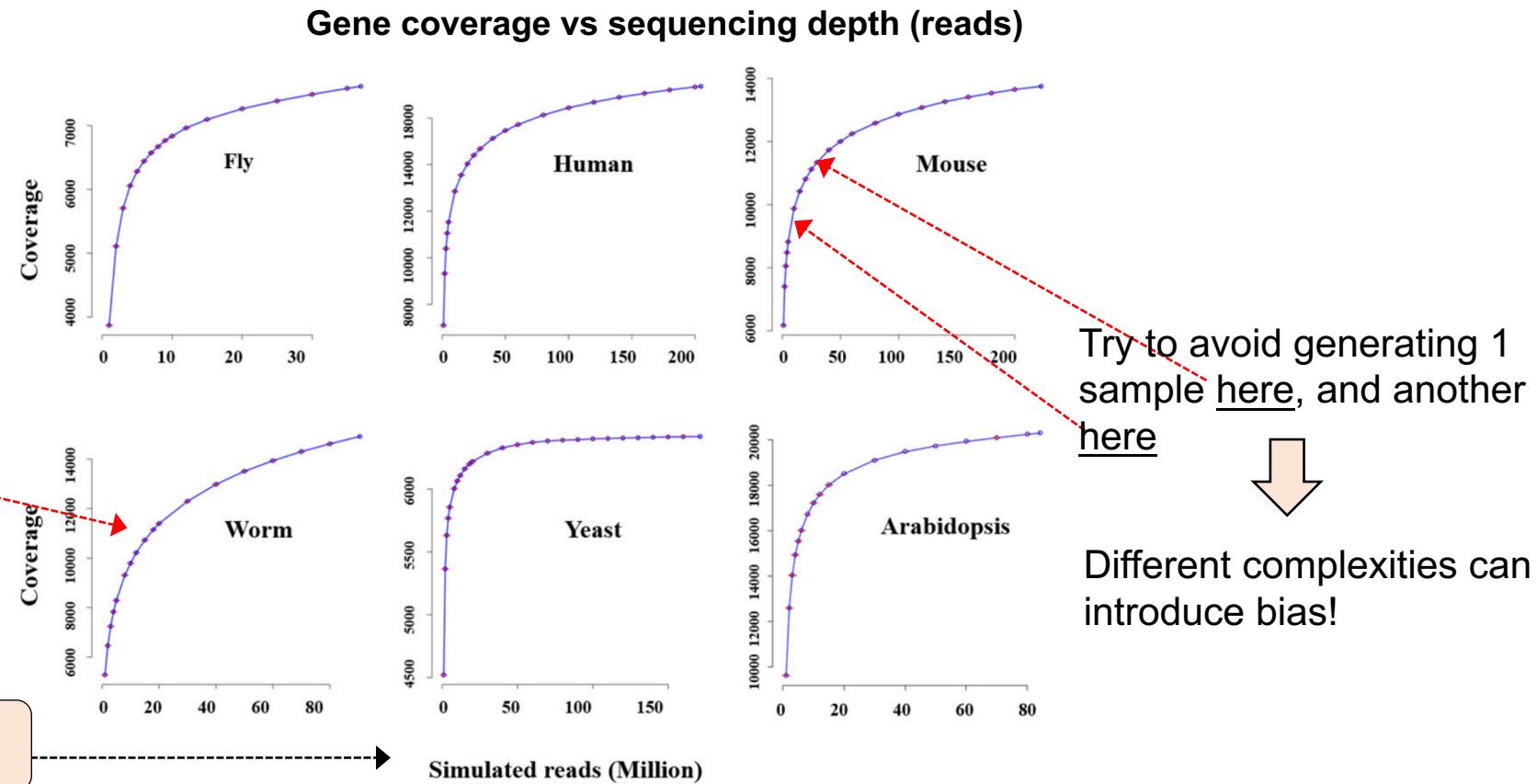
Liu *et al*, 2014, *Bioinformatics*

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

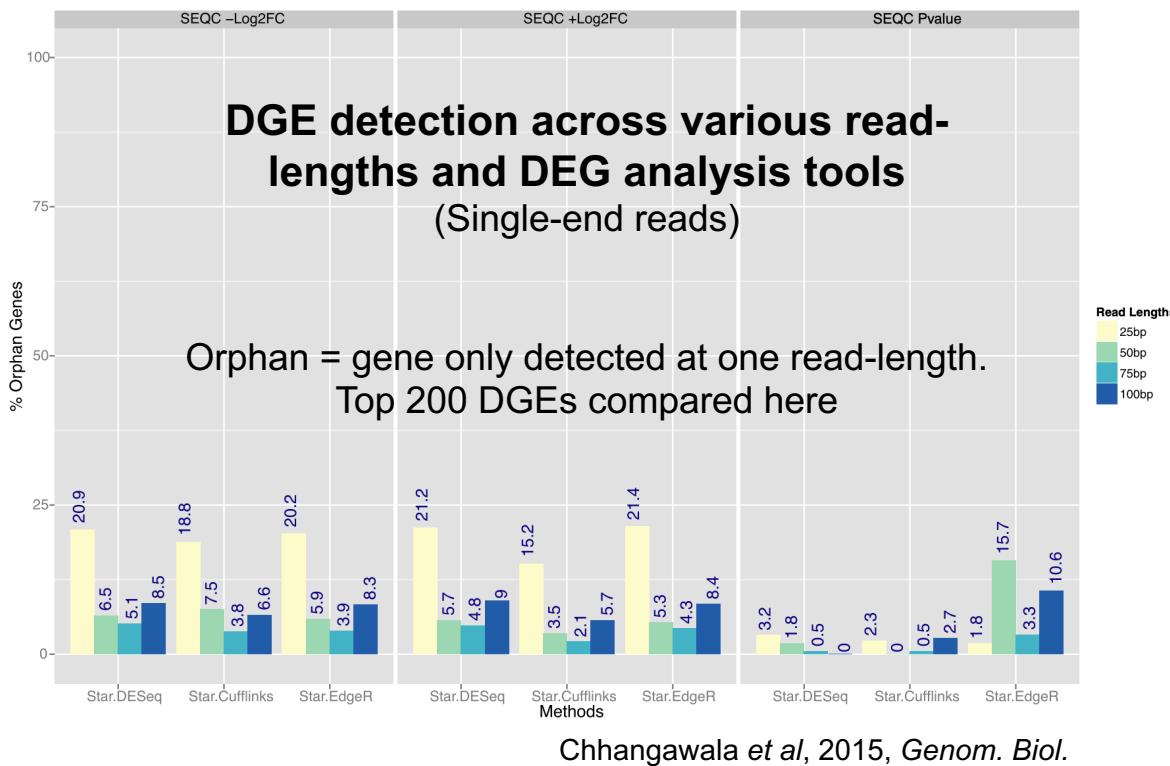


Get data points by subsampling reads

Read length

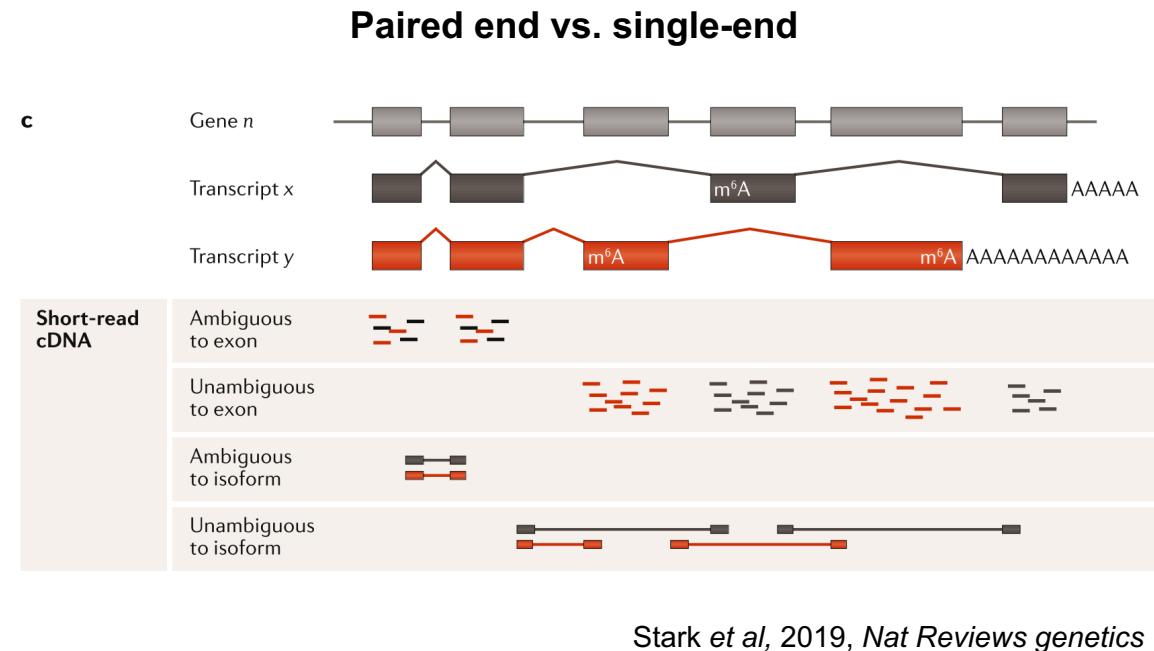
- For DGE, we want the **MINIMUM** read length required to accurately map a read to a gene

Shorter vs. end vs. longer-reads



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

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

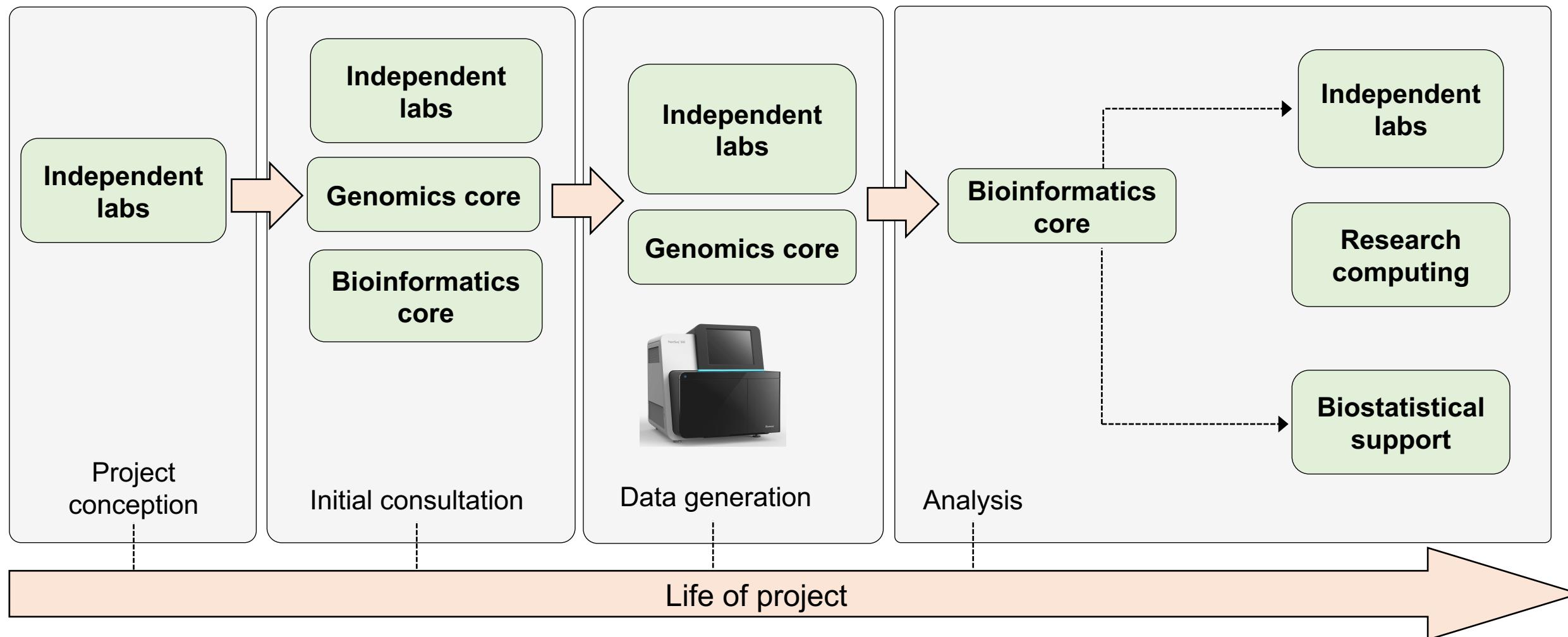


- For isoform-detection, alternative exon usage, & SNV detection, get **paired-end**

Where does the Bioinformatician fit in?



- Ideally, before data generation..

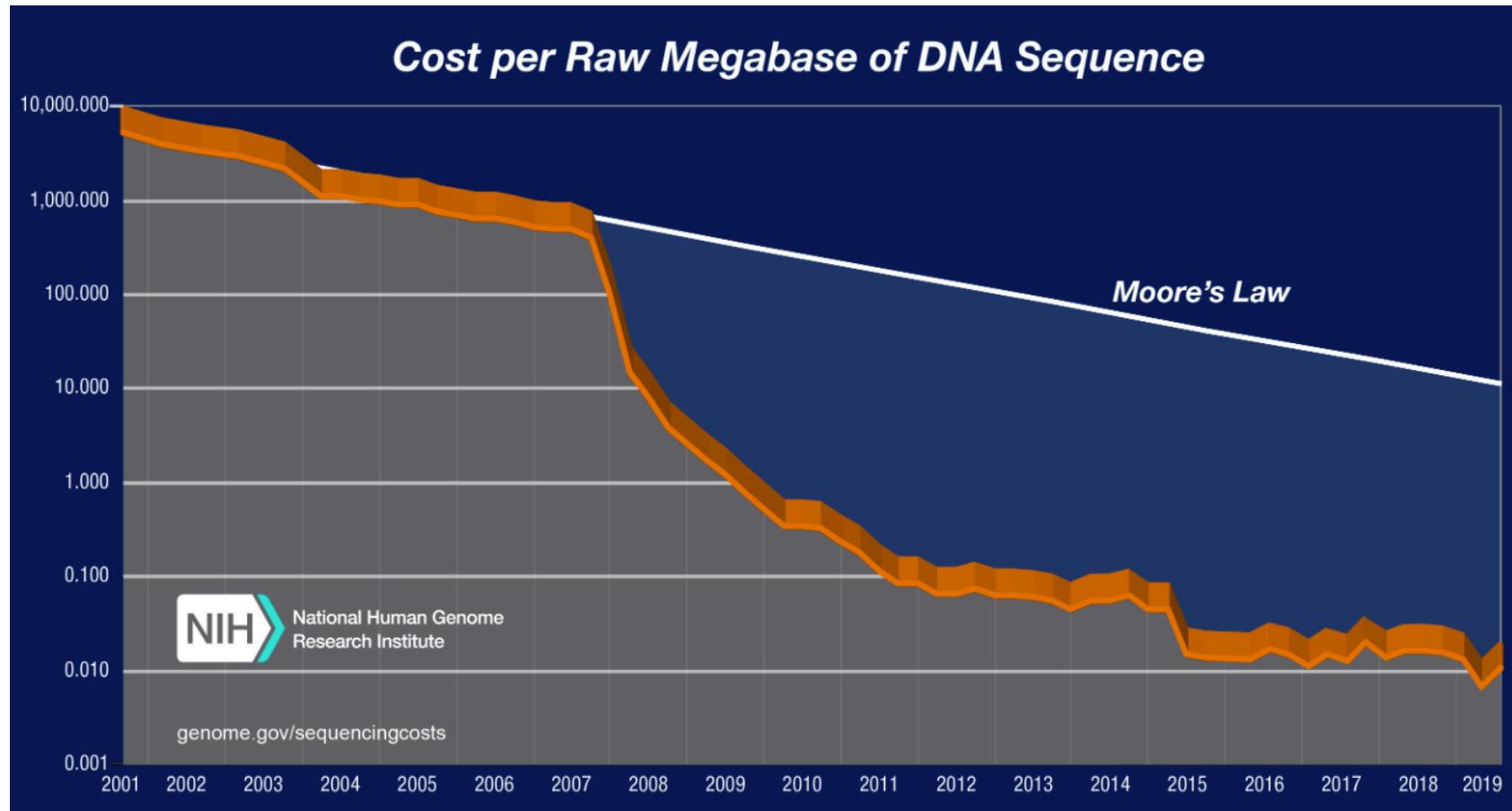


Common perception of bioinformatics..



Reality..

Data is getting cheaper & bigger



<https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data>

Complex analytics is playing a larger role



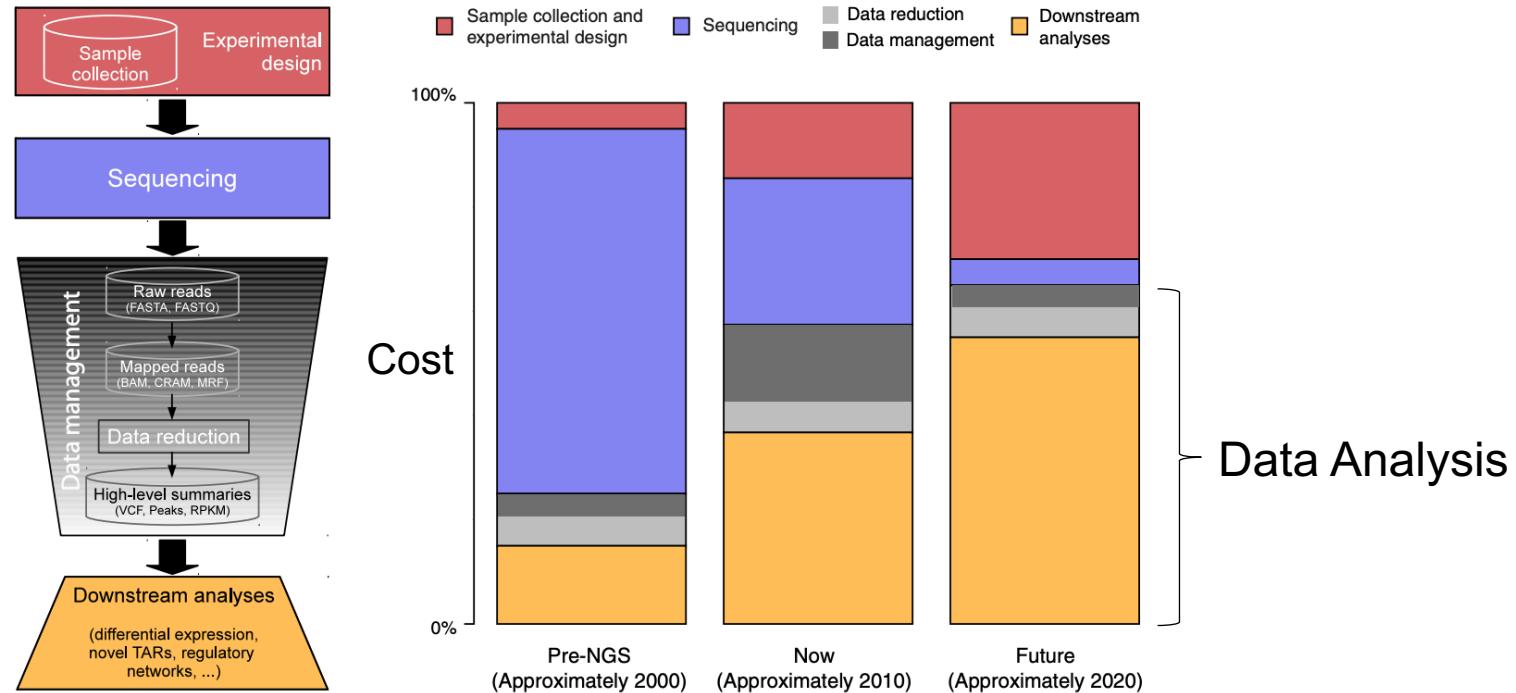
Sboner et al. *Genome Biology* 2011, 12:125
<http://genomebiology.com/2011/12/8/125>



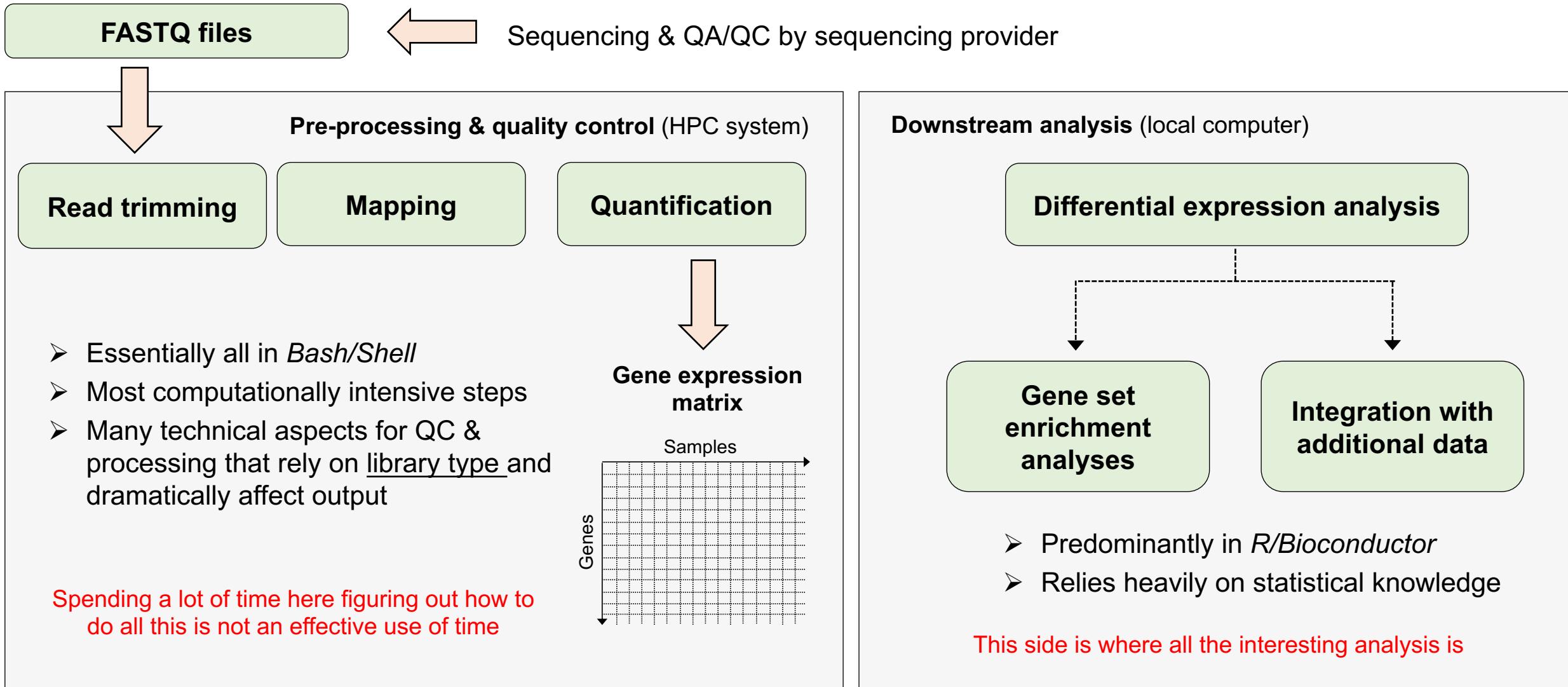
OPINION

The real cost of sequencing: higher than you think!

Andrea Sboner^{1,2}, Xinmeng Jasmine Mu¹, Dov Greenbaum^{1,2,3,4,5}, Raymond K Auerbach¹ and Mark B Gerstein*^{1,2,6}

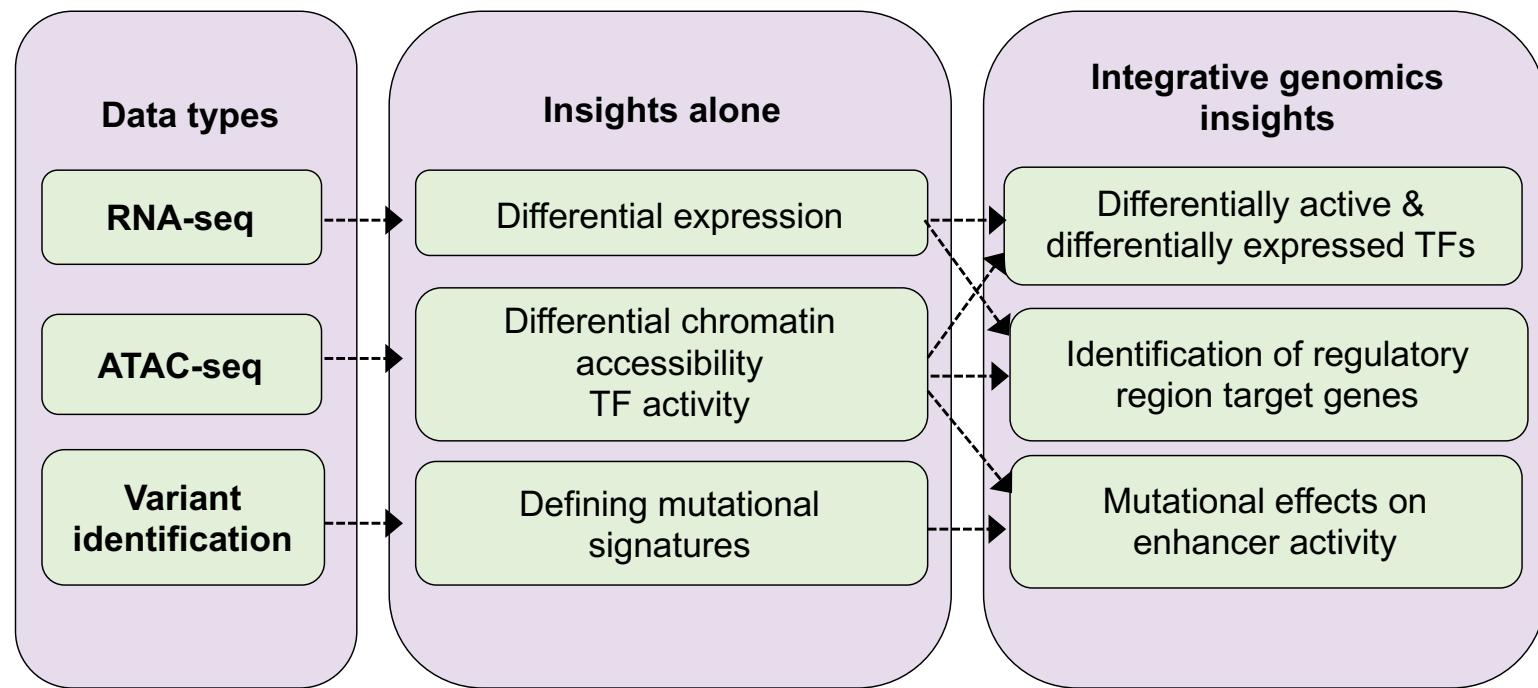


RNA-seq differential expression analysis pipeline



Beyond differential expression: Integrative genomics

- Leveraging data integration across more than one ‘omics platform, to reveal insights not possible with each data type alone



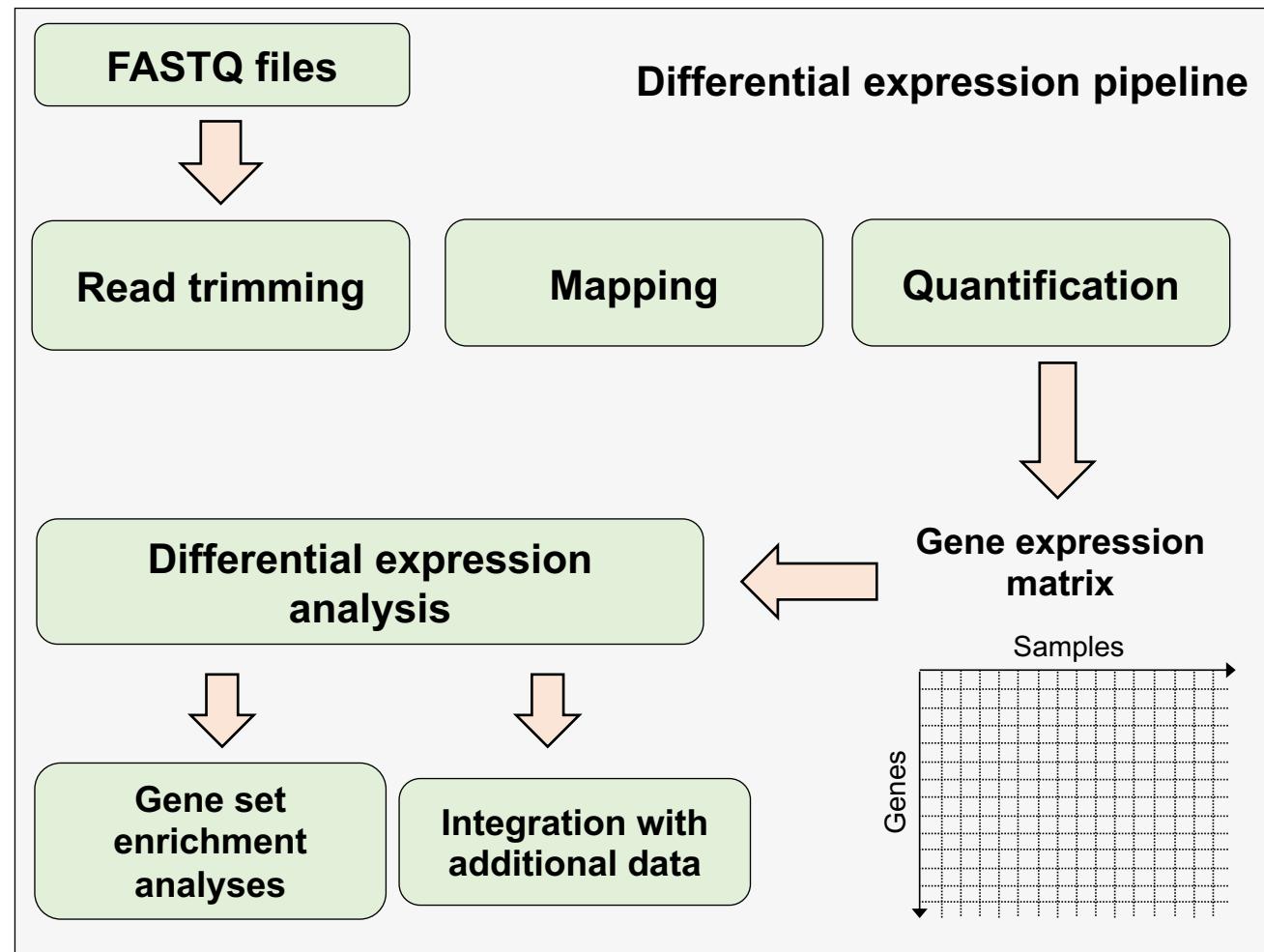
- You may not have generated each dataset in-house (e.g. combine in house & public data)

Summary

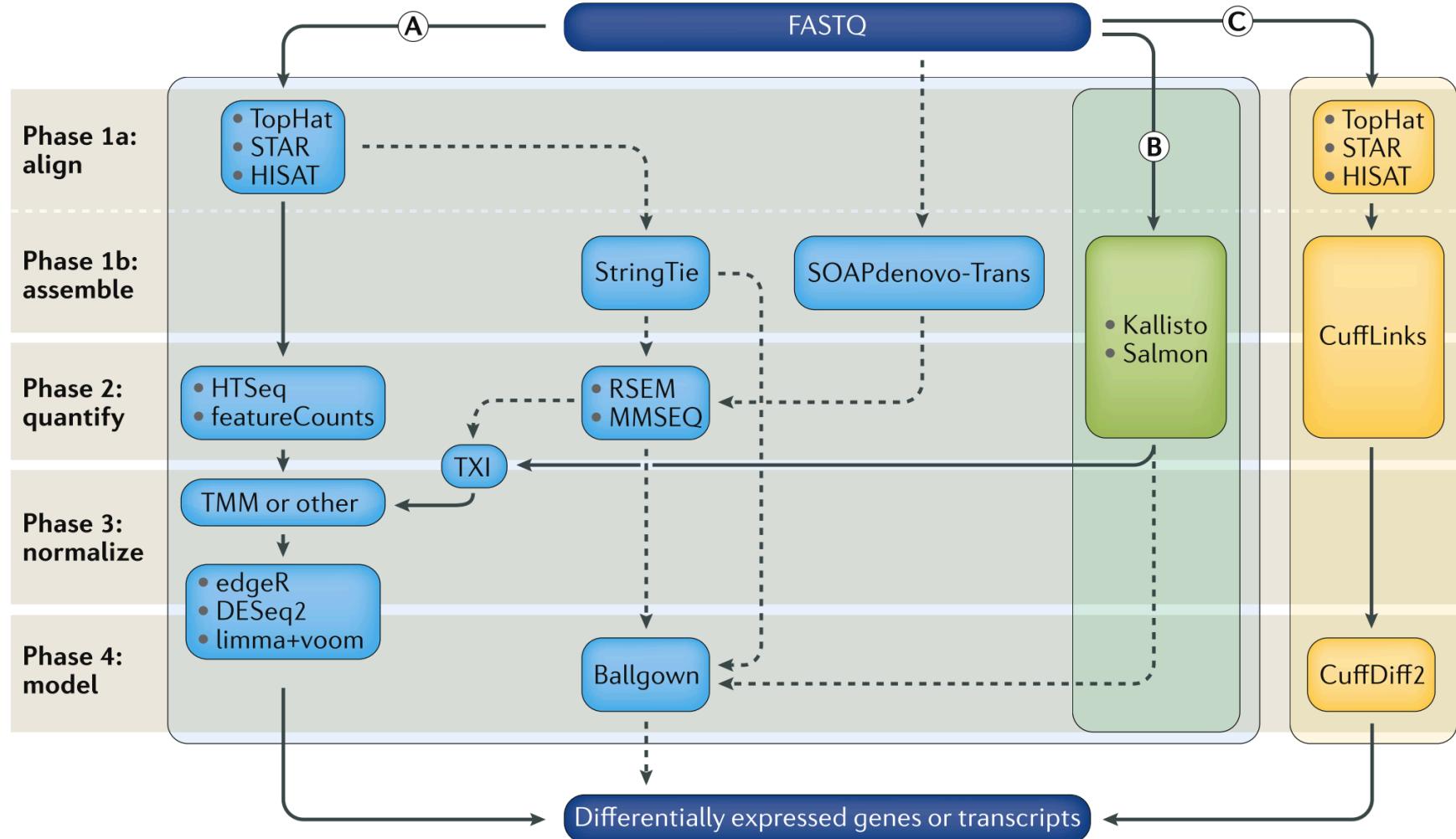


- **RNA-seq overview**
 - Basics of an RNA-seq experiment
 - Sequencing technologies for RNA-seq
- **Library types for RNA-seq**
 - Poly-A, 3'-end, Ribodepletion
 - What hypotheses can be tested with each?
- **Sample preparation & experimental design**
 - Replicates
 - Sequencing depth & configuration
- **Data analysis**
 - Overview of analysis pipeline(s) for differential expression (DE)
 - Where does the Bioinformatician fit in?
 - Integrative genomics & DE
- **Bulk RNA-seq vs. single-cell RNA-seq**

Questions?



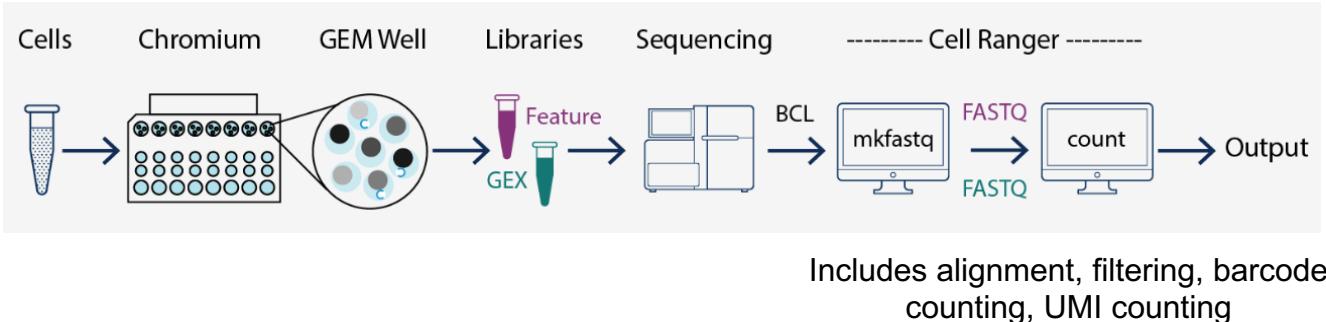
Differential expression workflow(s)



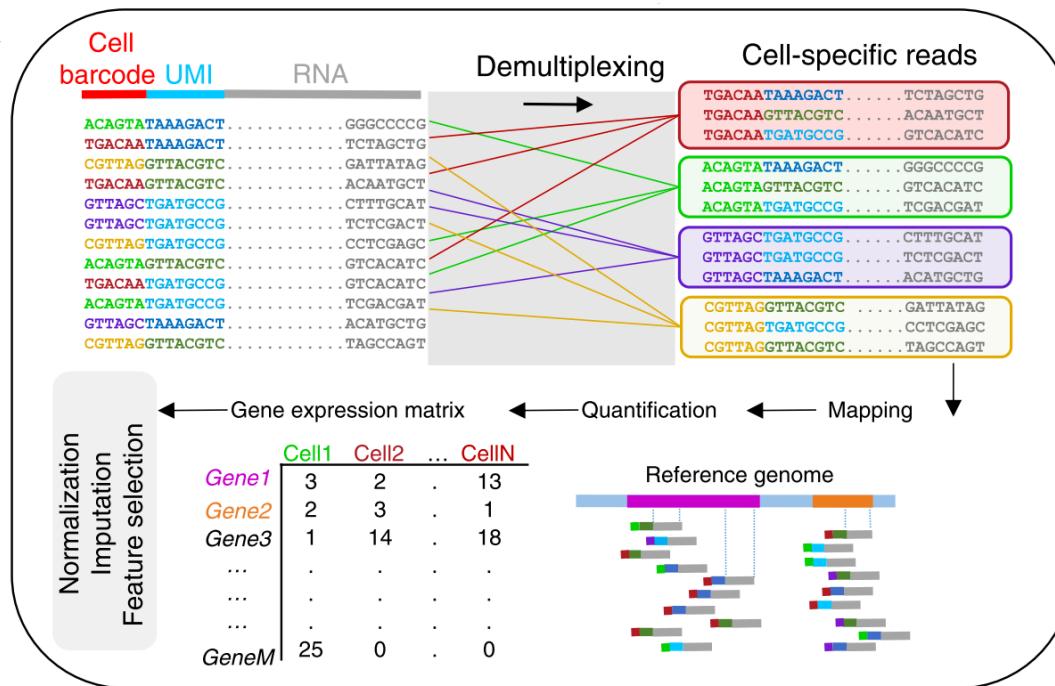
- Several tools available for each step
- Each have various strengths, weaknesses, applications

Single cell RNA-seq

10X Genomics – Cell Ranger pipeline



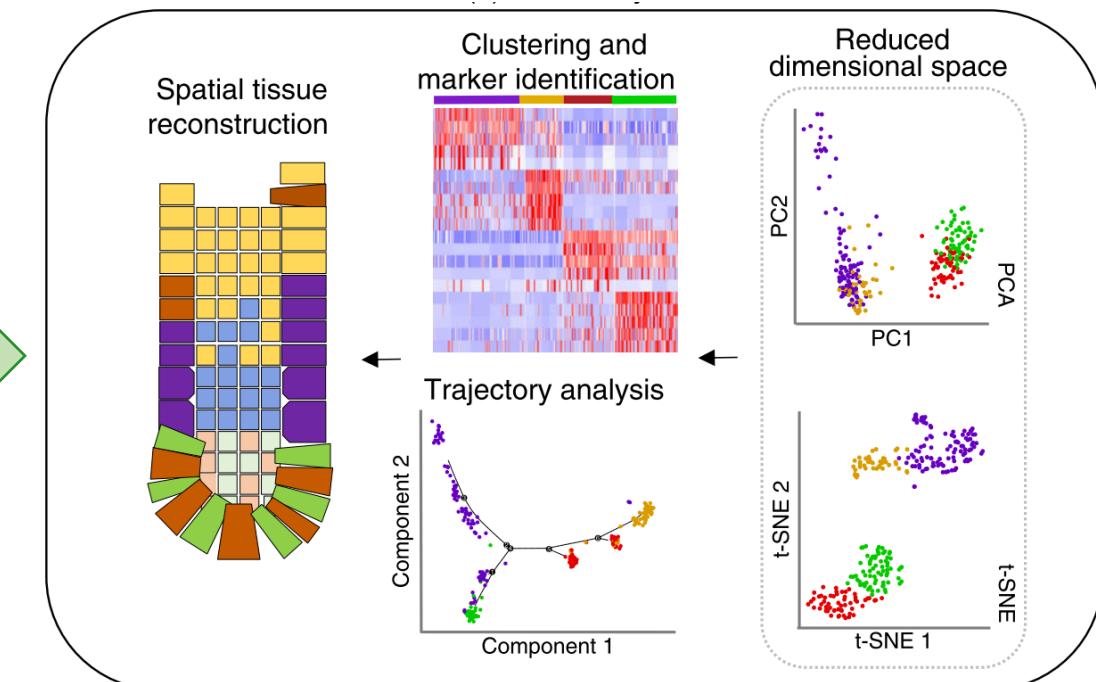
Data processing



Bulk & single-cell RNA-seq are distinct and generally address different questions

- Do different cell types exist?
- What genes define these cell types?
- How do gene expression profiles vary within cell types across conditions or cancer subtypes

Data analysis



Adapted from, Lafzi, 2018, *Nat Protocols*.