Introduction to RNA-seq

Owen M. Wilkins, PhD

Bioinformatics scientist

Data Analytics Core (DAC)

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/

06/14/21

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)

RNA isolation
Purification & selection

Library preparation
Fragmentation, reverse transcription, adapters, PCR

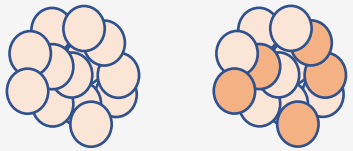
Sequencing
cDNA pool from samples

Reads in FASTQ format

Demultiplexing

Sample 1
Control

Sample 1
Treated



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
@NB501031:325:HK52YBGXB:1:11101:9675:1055 2:N:0:ACCACTGT
CTGGGCAGCTGGGGGGTGGGGTGGGGAAGACACAGGAGCCACACAGAGAGAGACATCCCATCCTGTCTCCATTGCT
+
AAAAAEEEEAEEEEEE/E/EE6EEEEEEEAEEEEEEA//EEEEEAEEEAEEE/E/EEEA<EEEEEEEEEE/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/EEEEEEEAA5/EA/EE/</EEE/E
```

Data analysis

Downstream analysis
Differential expression
Isoform discovery
Annotation
Variant calling

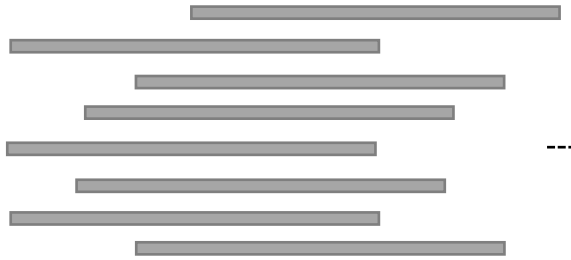
Data normalization

Feature quantification

Genome mapping

Quality control

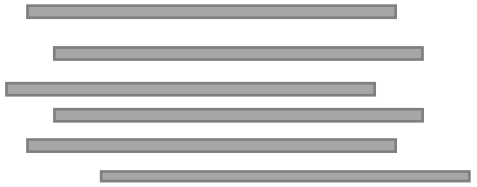
Library preparation for RNA-seq



Total RNA

Enrichment

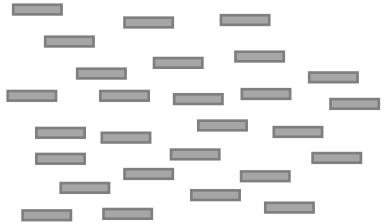
(e.g. PolyA capture, ribodepletion)



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

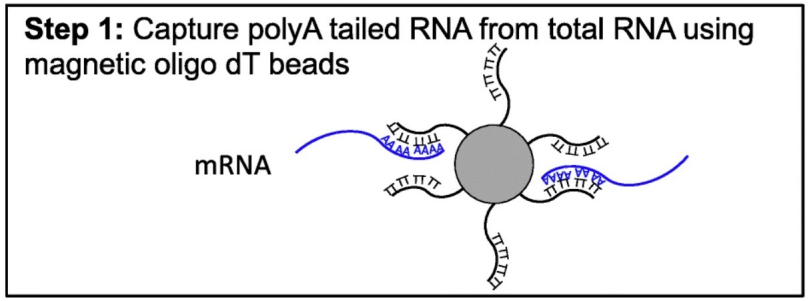
RNA selection/enrichment

- **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

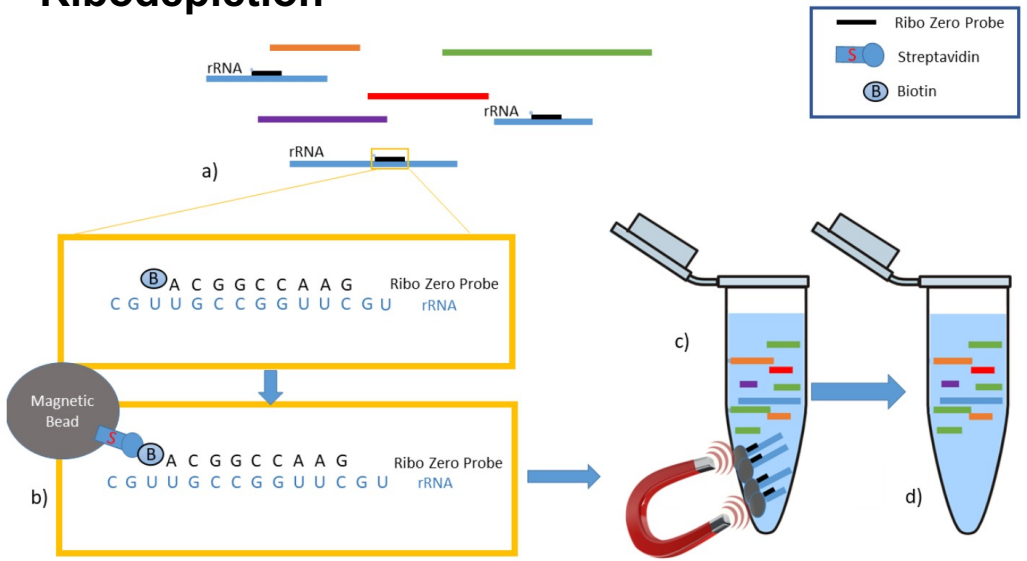
- **Ribodepletion:**
  - Oligos complementary to highly conserved rRNA sequences used to capture rRNA
  - Streptavidin-bound magnetic beads used to capture and remove hybridized sequences
  - Enables enrichment of polyA mRNA & non-polyA ncRNAs (e.g. lncRNAs)

Oilgo-d(T) selection



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

Ribodepletion



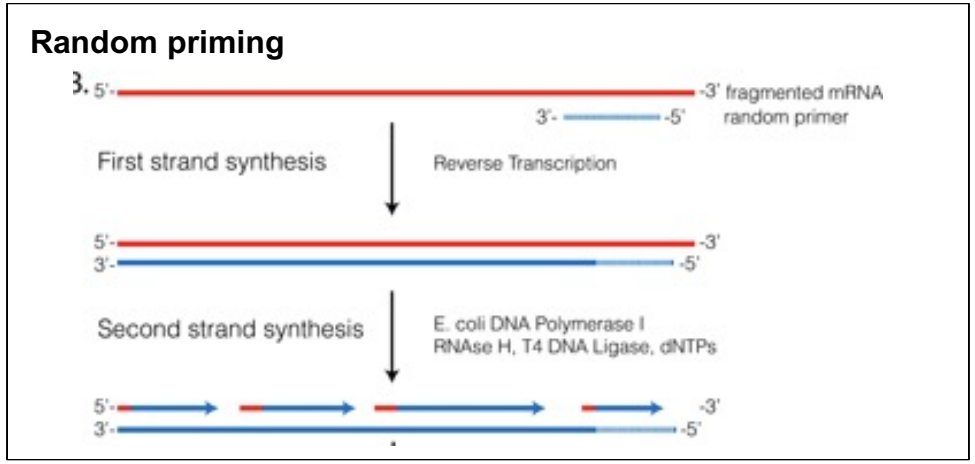
Adapted from Illumina.com

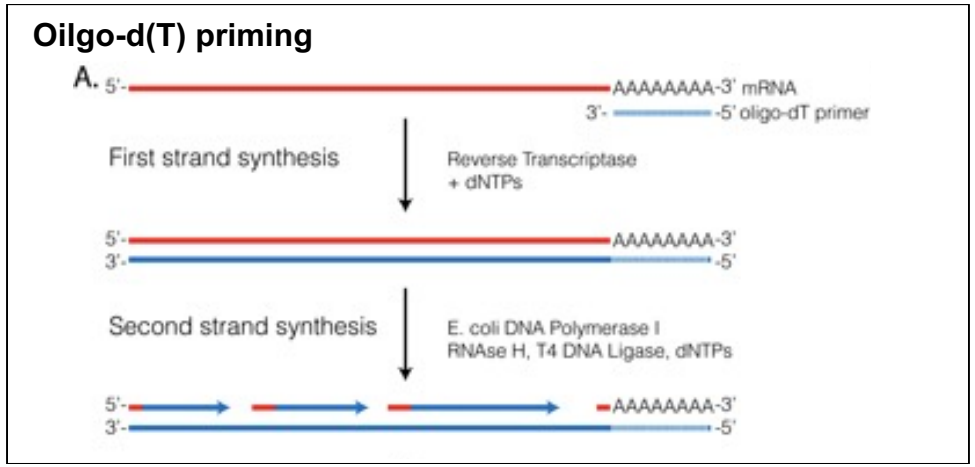
1st & 2nd-strand synthesis

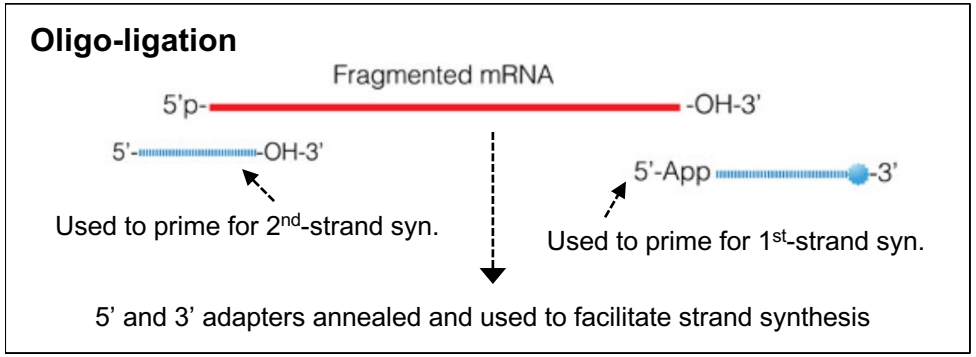
- **Random priming:**
  - Random hexamers anneal along length of transcript to facilitate cDNA synthesis
  - Generate fragments along full-length of transcript

- **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-ligation & priming**
  - Oligos containing primer are directly ligated to fragmented RNA







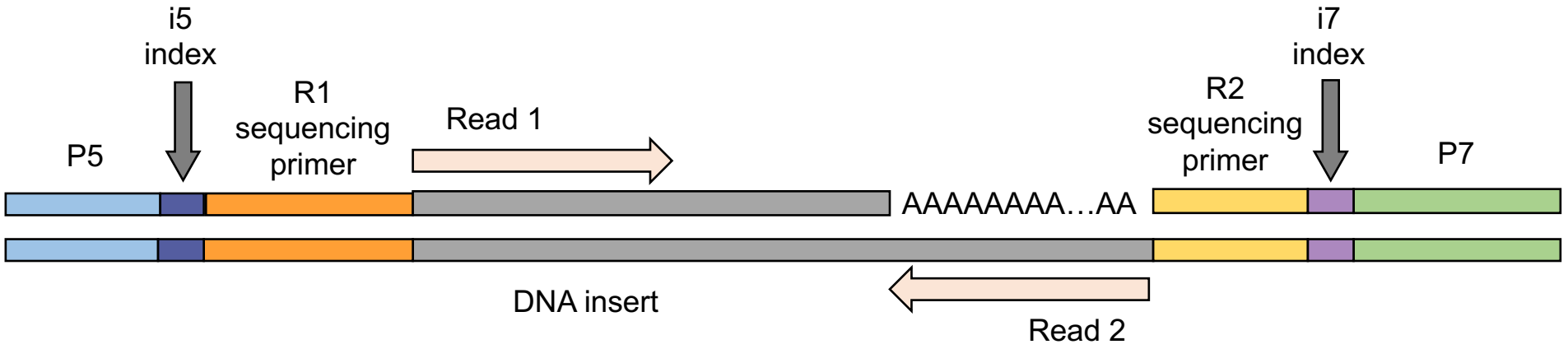
Images adapted from RNA-seqlopedia: UOregan

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

Stranded libraries

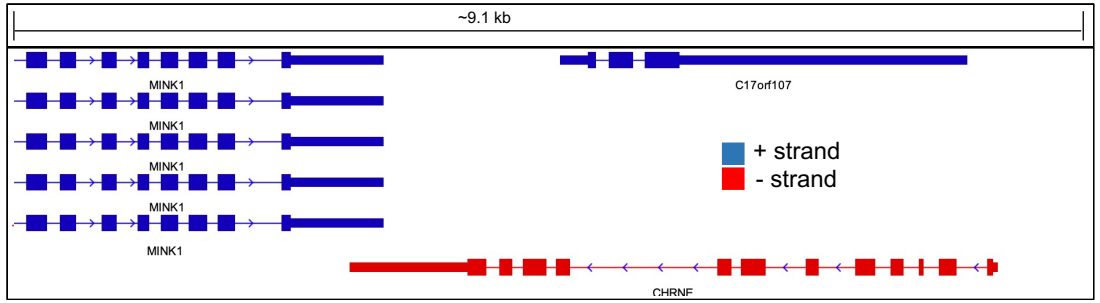
Problem:

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

Why do we care?

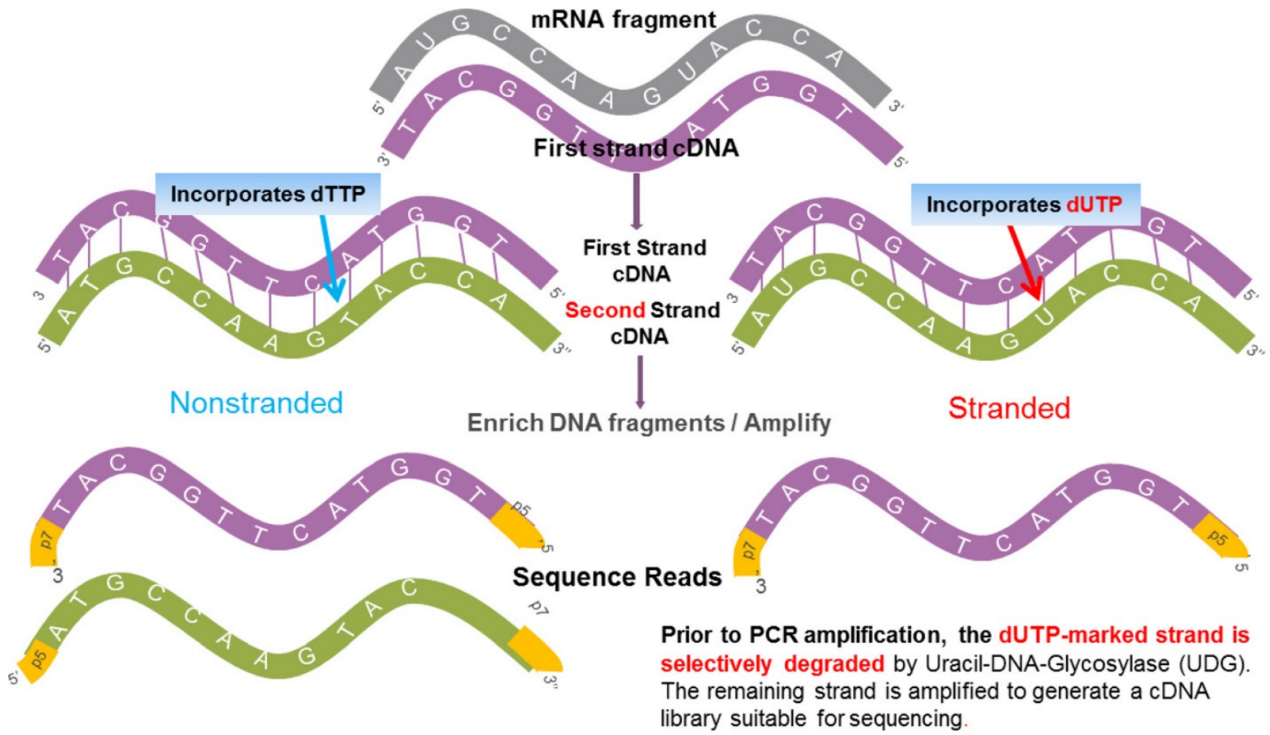
- Strand knowledge critical to assign reads to overlapping features e.g. overlapping genes, anti-sense transcripts

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.

Traditional library prep. vs .3'end

Traditional method

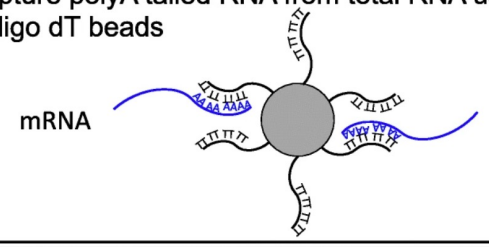
- cDNA generation using random hexamers
- Full-length transcript
- More detailed data

3' method

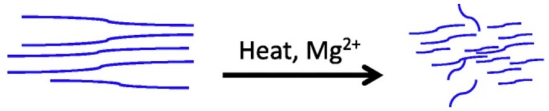
- 3' –end of transcript only
- No transcript information
- Only eukaryotic samples
- **Big cost savings**

Traditional method (KAPA)

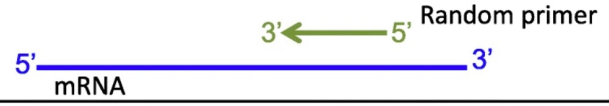
Step 1: Capture polyA tailed RNA from total RNA using magnetic oligo dT beads



Step 2: mRNA fragmentation



Step 3: 1st strand synthesis with random primers



Step 4: 2nd strand synthesis with dUDP

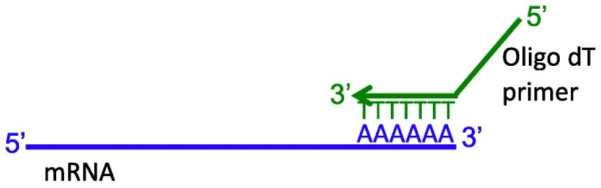
Step 5: A-tailing and barcoded adapter ligation

Step 6: Amplification (dUTP strand is not amplified)

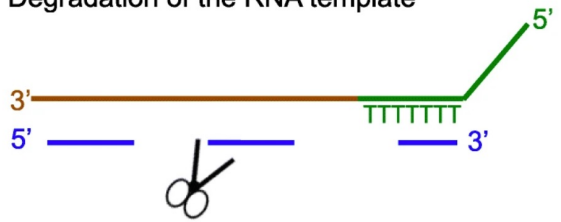
Step 7: Sequencing

3' method (LEXO)

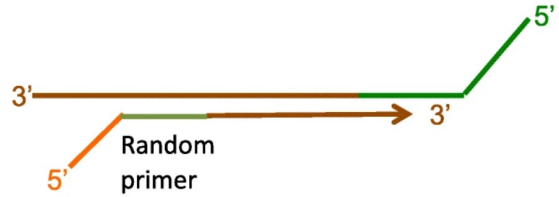
Step 1: 1st strand synthesis of polyA tailed RNA from total RNA using oligo dT primers



Step 2: Degradation of the RNA template



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

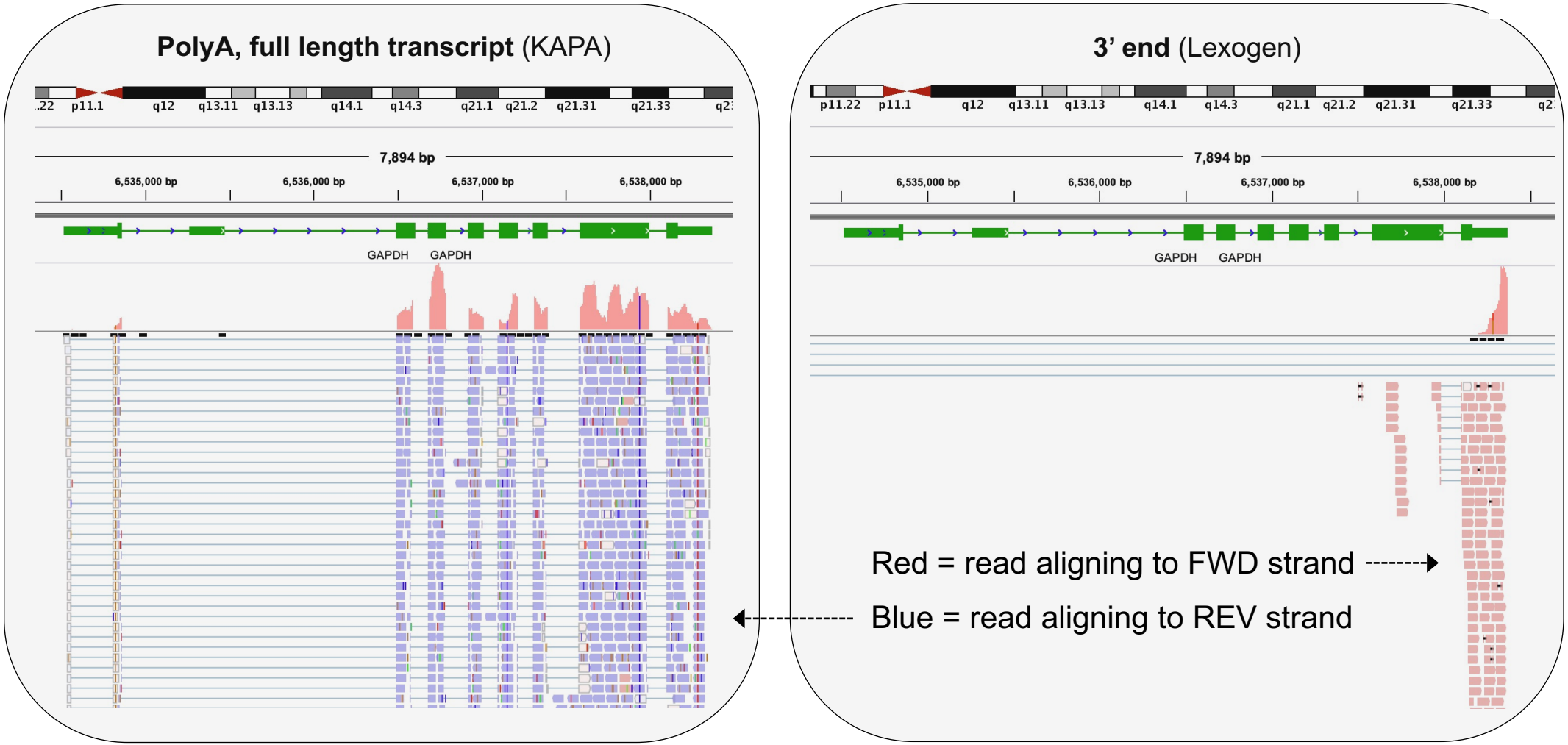


Step 4: Amplification using random primers that add barcodes and cluster generation sequences

Step 5: Sequencing

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

Data from different library types looks different

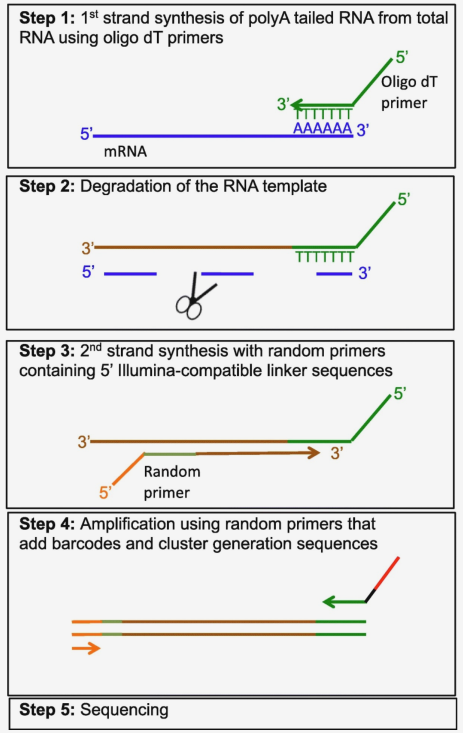


Quality control, analysis pipelines, and possible hypotheses therefore inherently differ

Applications of different library types

3’ End

3’ method (LEXO)

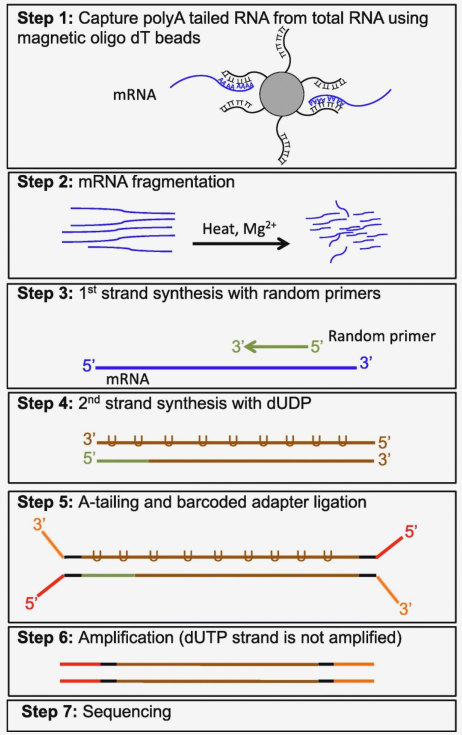


Adapted from Fukua *et al*, 2019. *Genom. Biol*

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

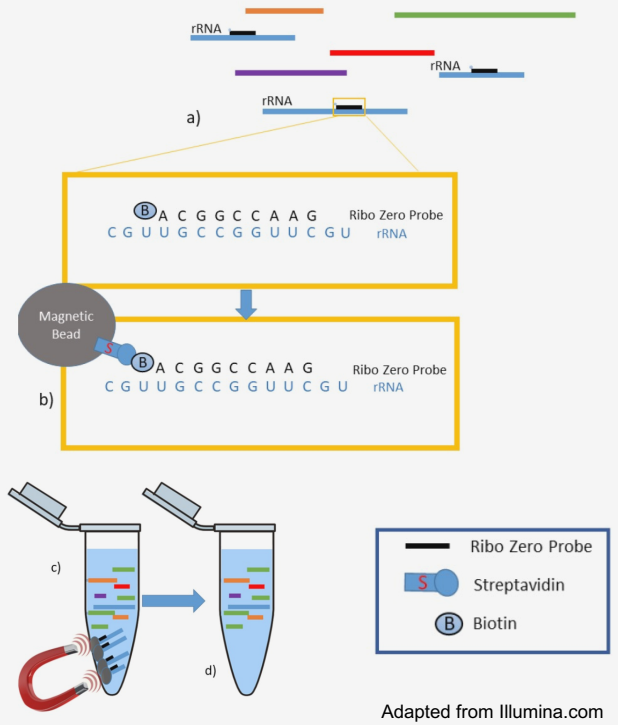
Full-length - PolyA

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

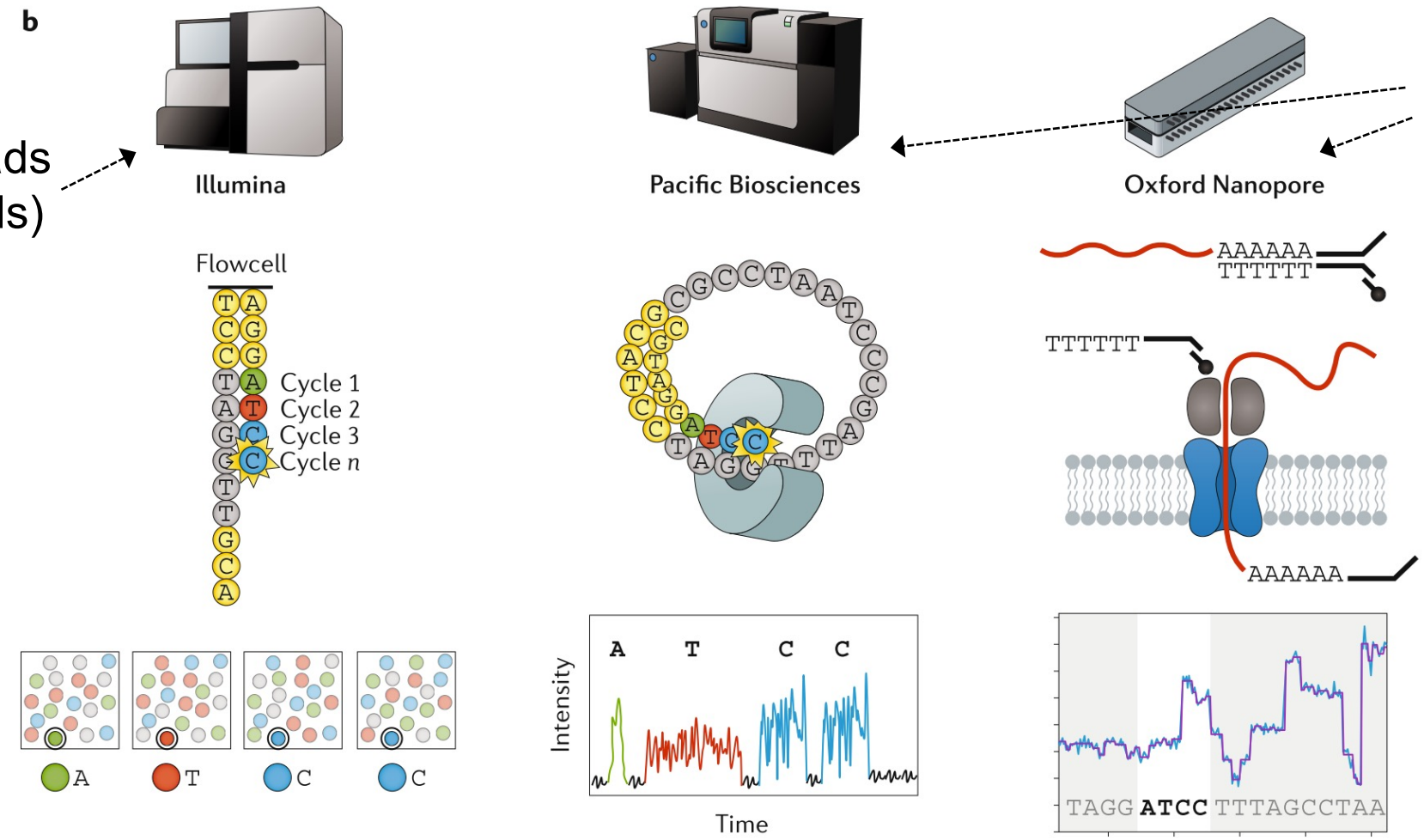
Cost / Data Richness

Image Credit:
Lexogen Inc

Sequencing technologies for RNA-seq

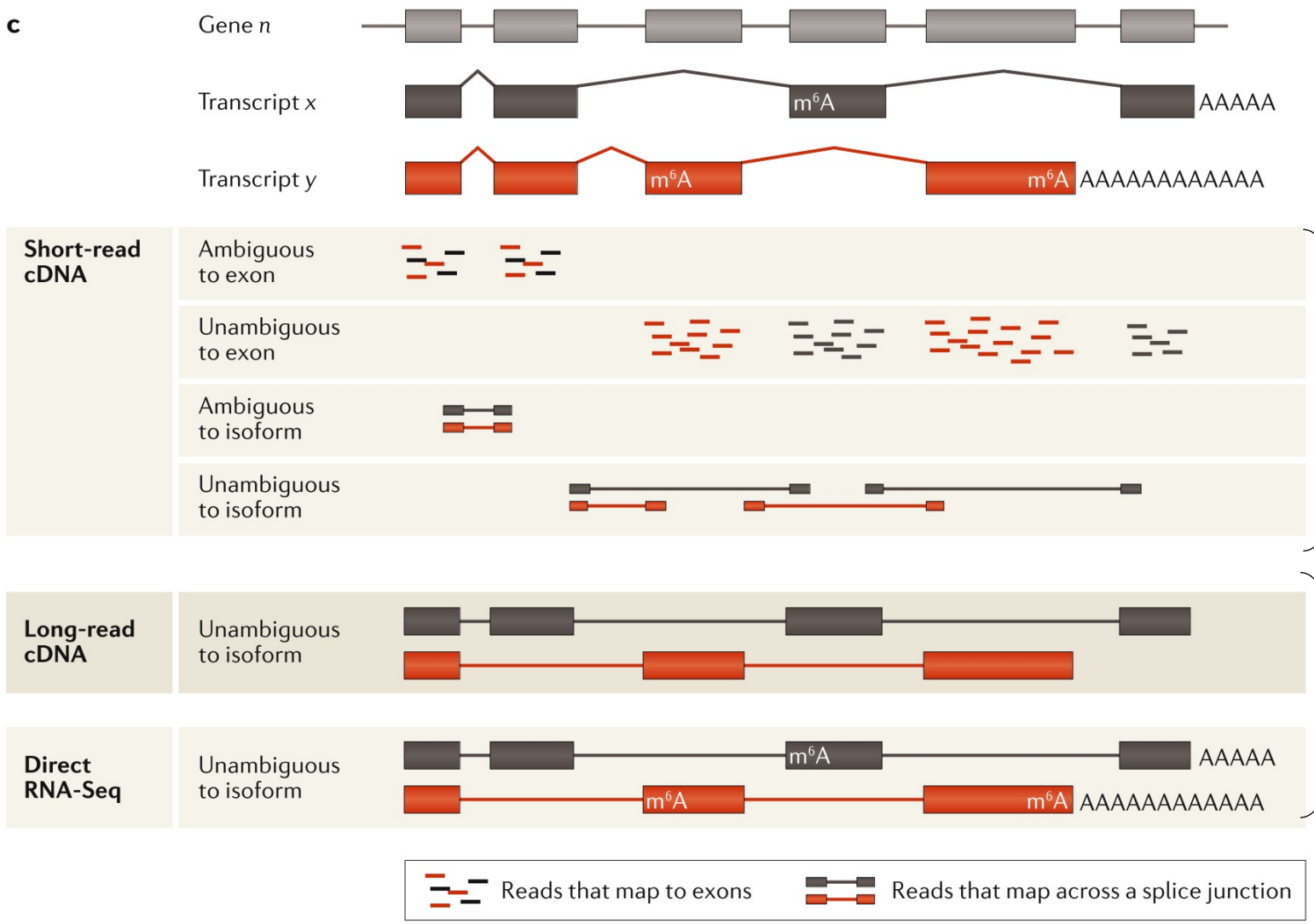
Short reads (hundreds)

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*

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"

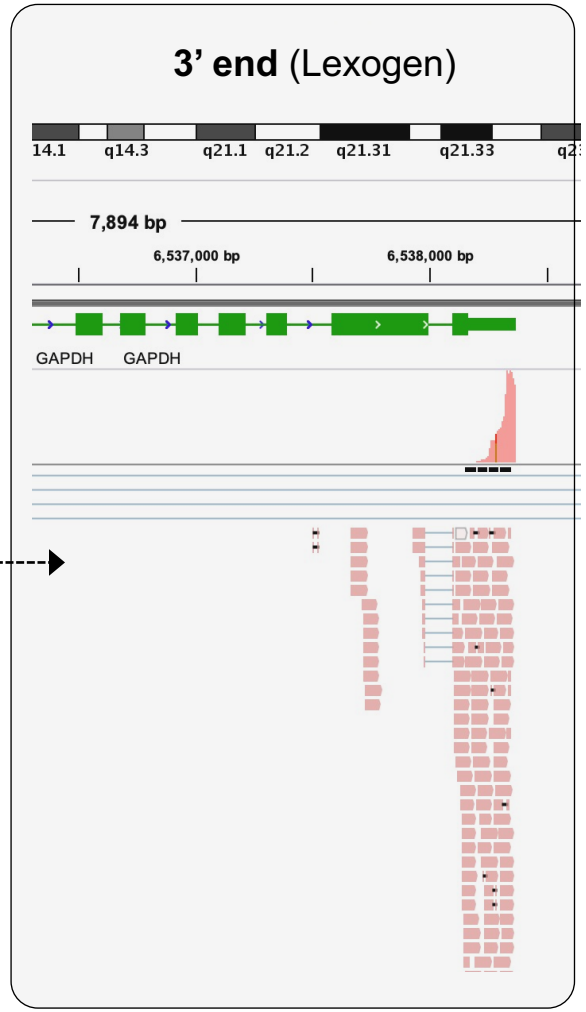
- Church *et al,* RNA, 2016

- Suggested minimum no. of replicates should = 6

- More heterogeneity = more replicates needed (e.g. human tissues vs. cultured cell lines)

Sequencing depth (or coverage?)

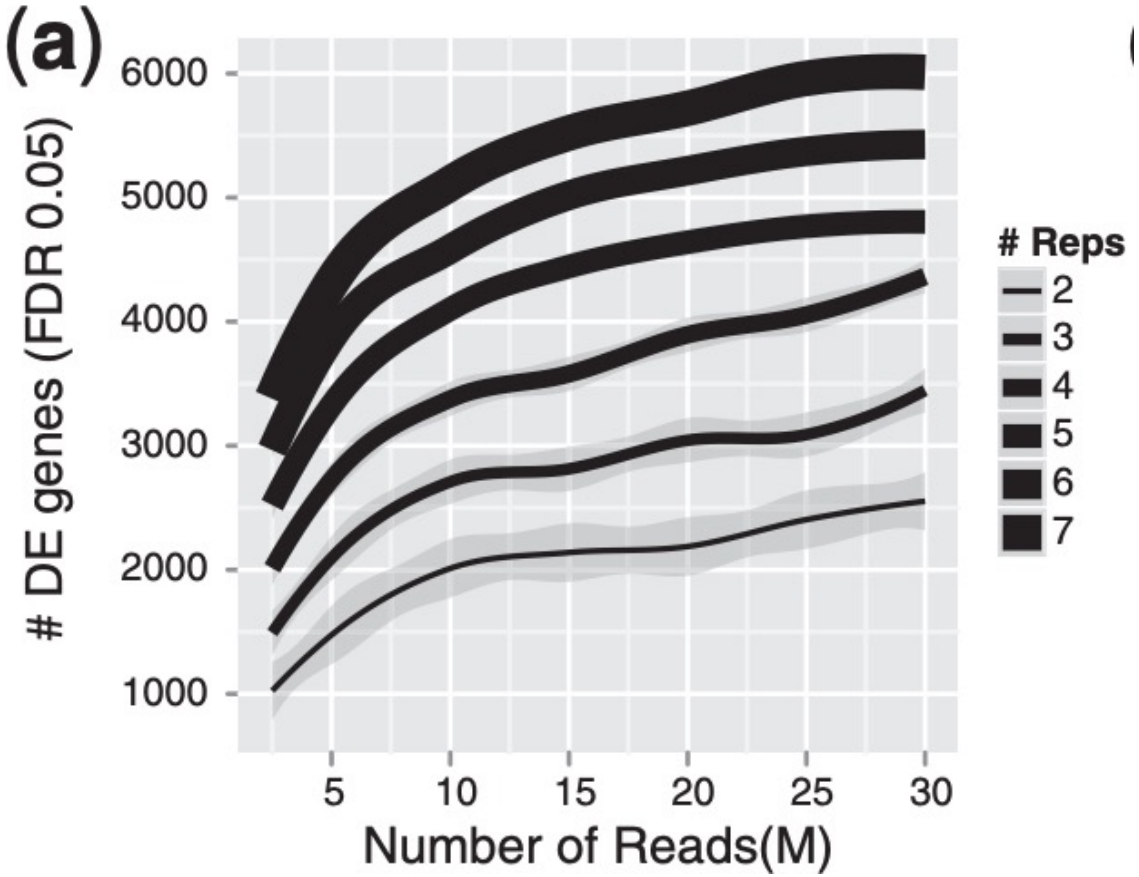
- '*Coverage*' doesn't have much meaning for transcriptome data
  - We are less concerned with how many times we cover a specific locus w/ a read..
- Generally total of 10-30 million reads for DGE of eukaryotic genomes
- Some species require many fewer than this
- Technology also affects required read 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)

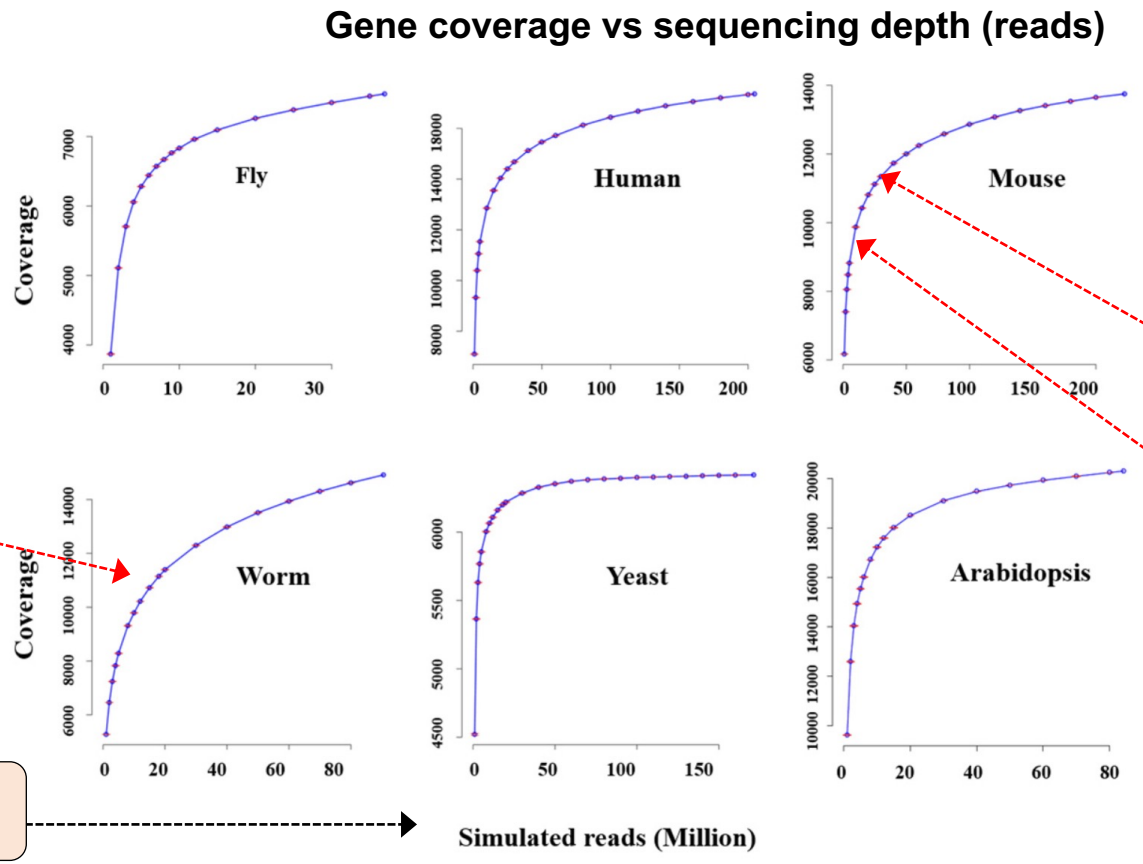


Liu *et al*, 2014, *Bioinformatics*

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



- 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

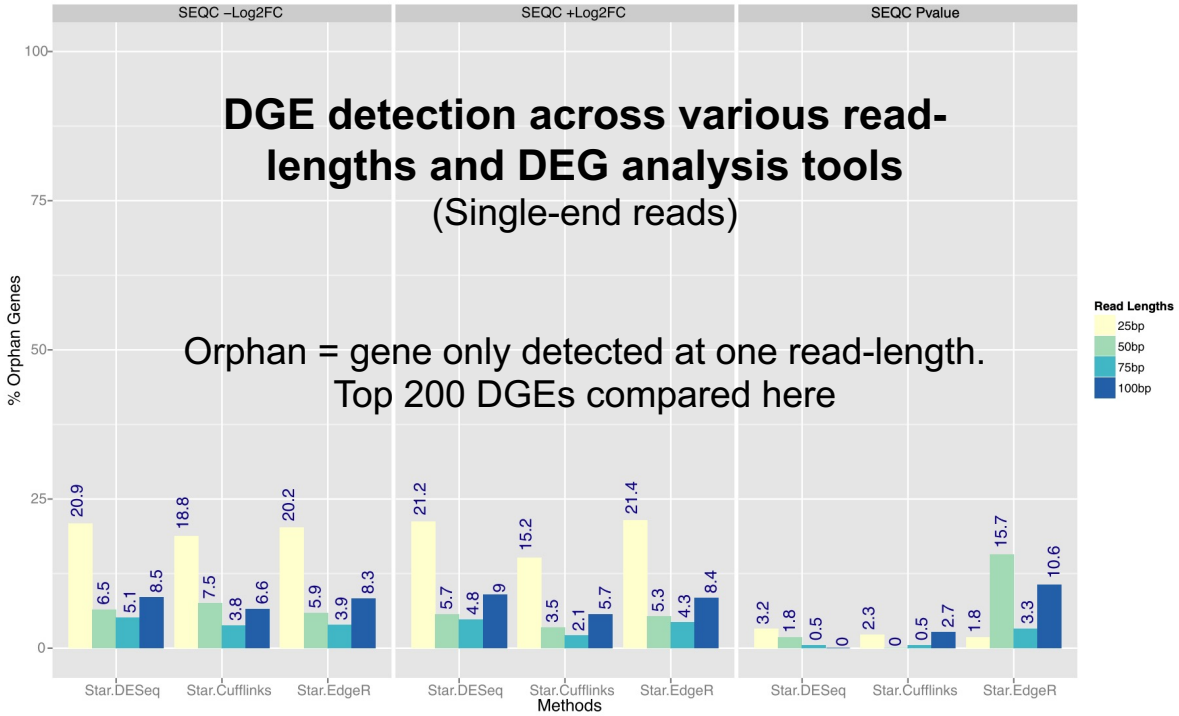
Different complexities can introduce bias!

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. end vs. longer-reads

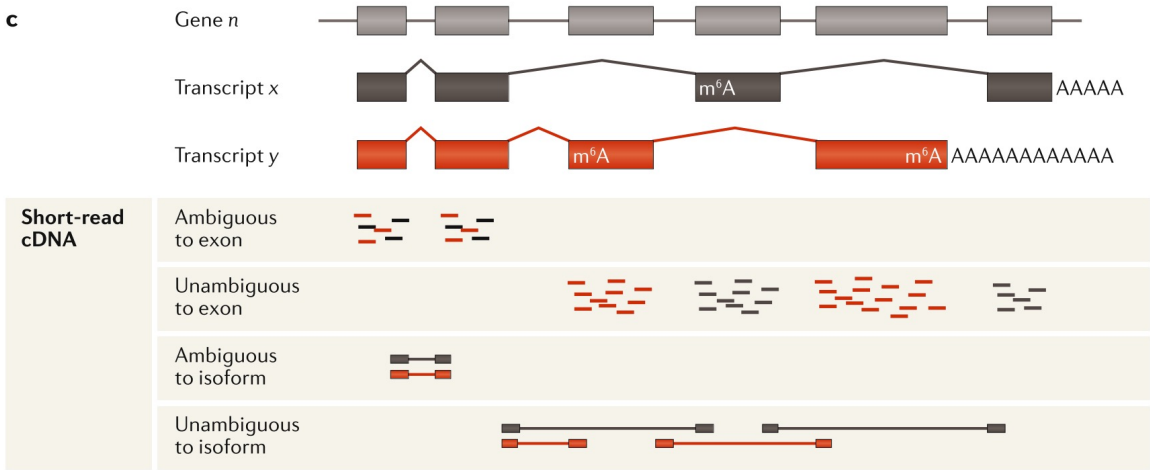


Chhangawala *et al*, 2015, *Genom. Biol.*

- 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

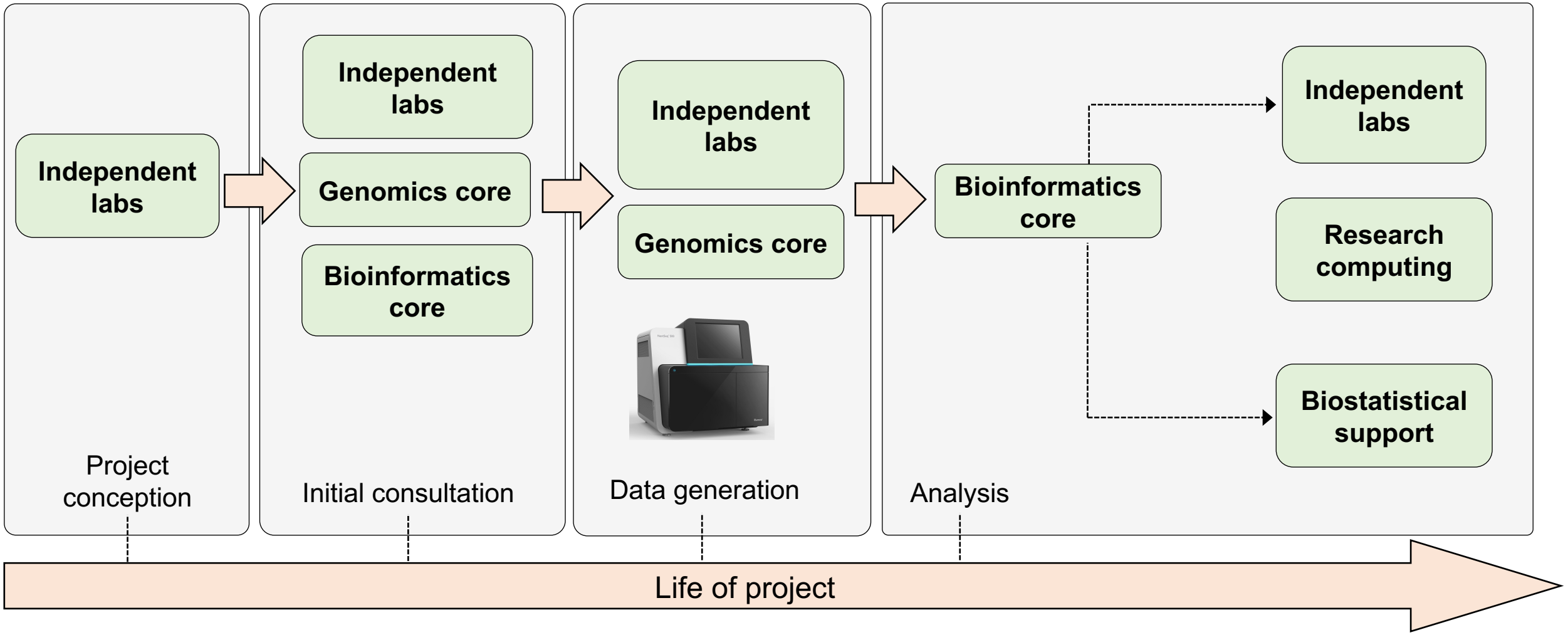


Stark *et al,* 2019, *Nat Reviews genetics*

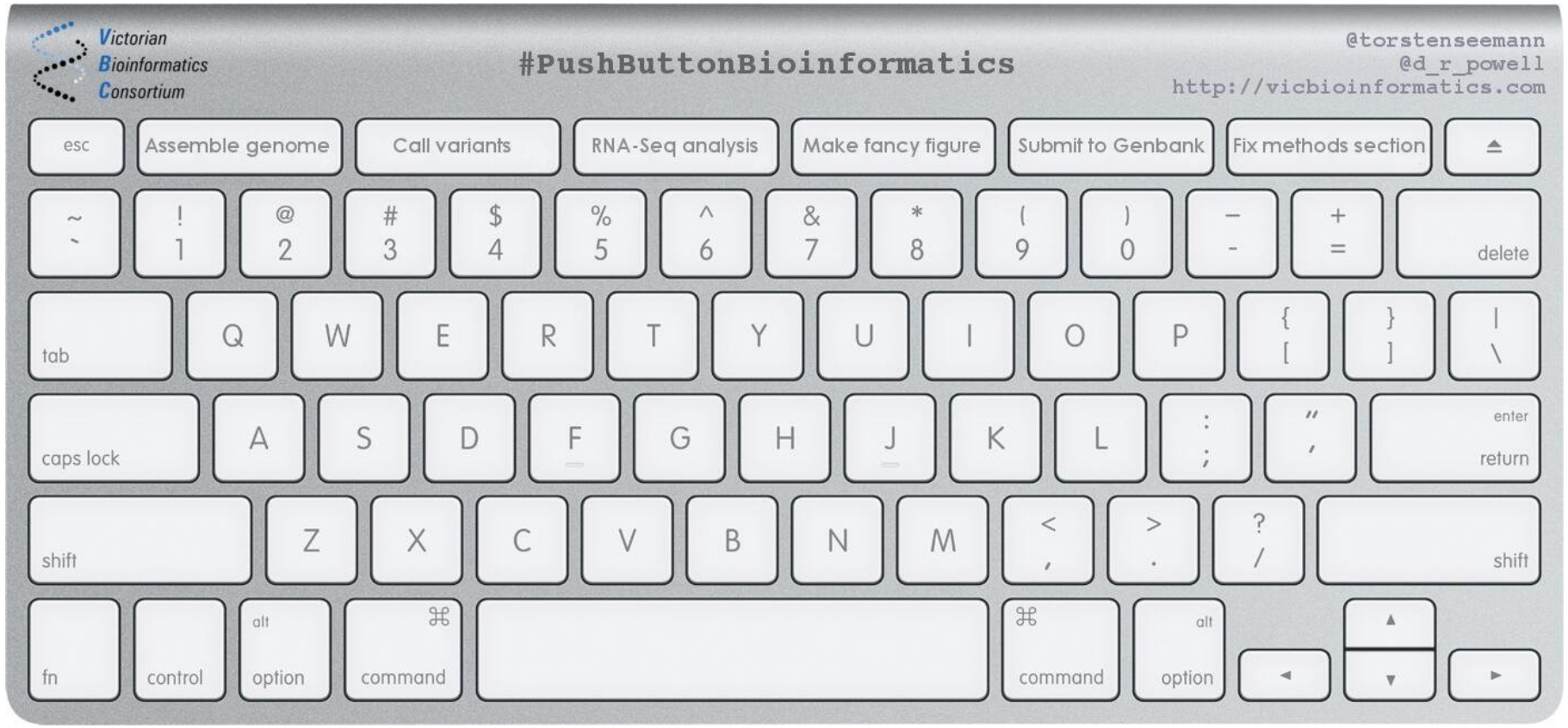
- For isoform-detection, alternative exon usage, & SNV detection, get **paired-end**

Where does the Bioinformatician fit in?

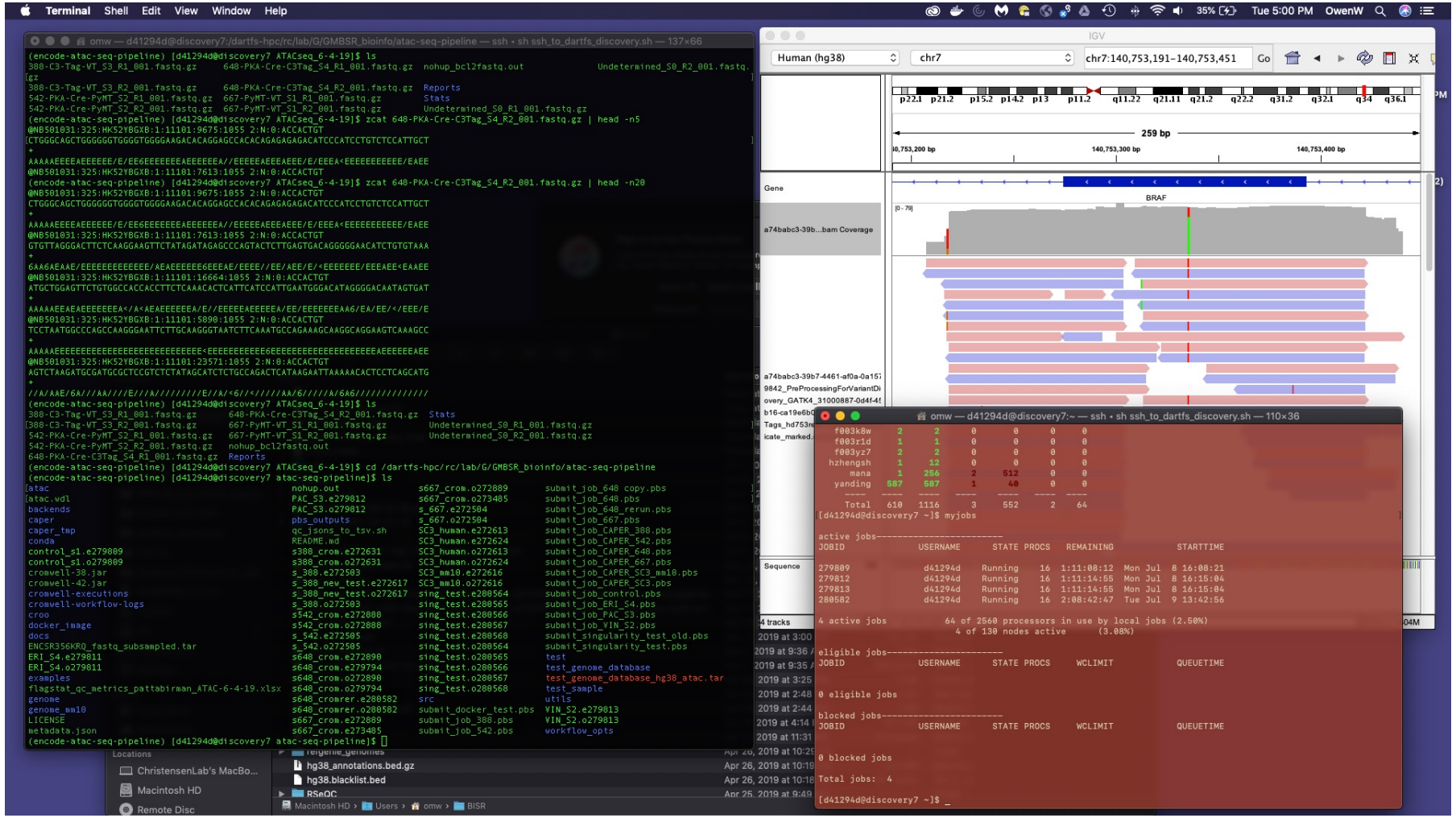
- Ideally, before data generation..



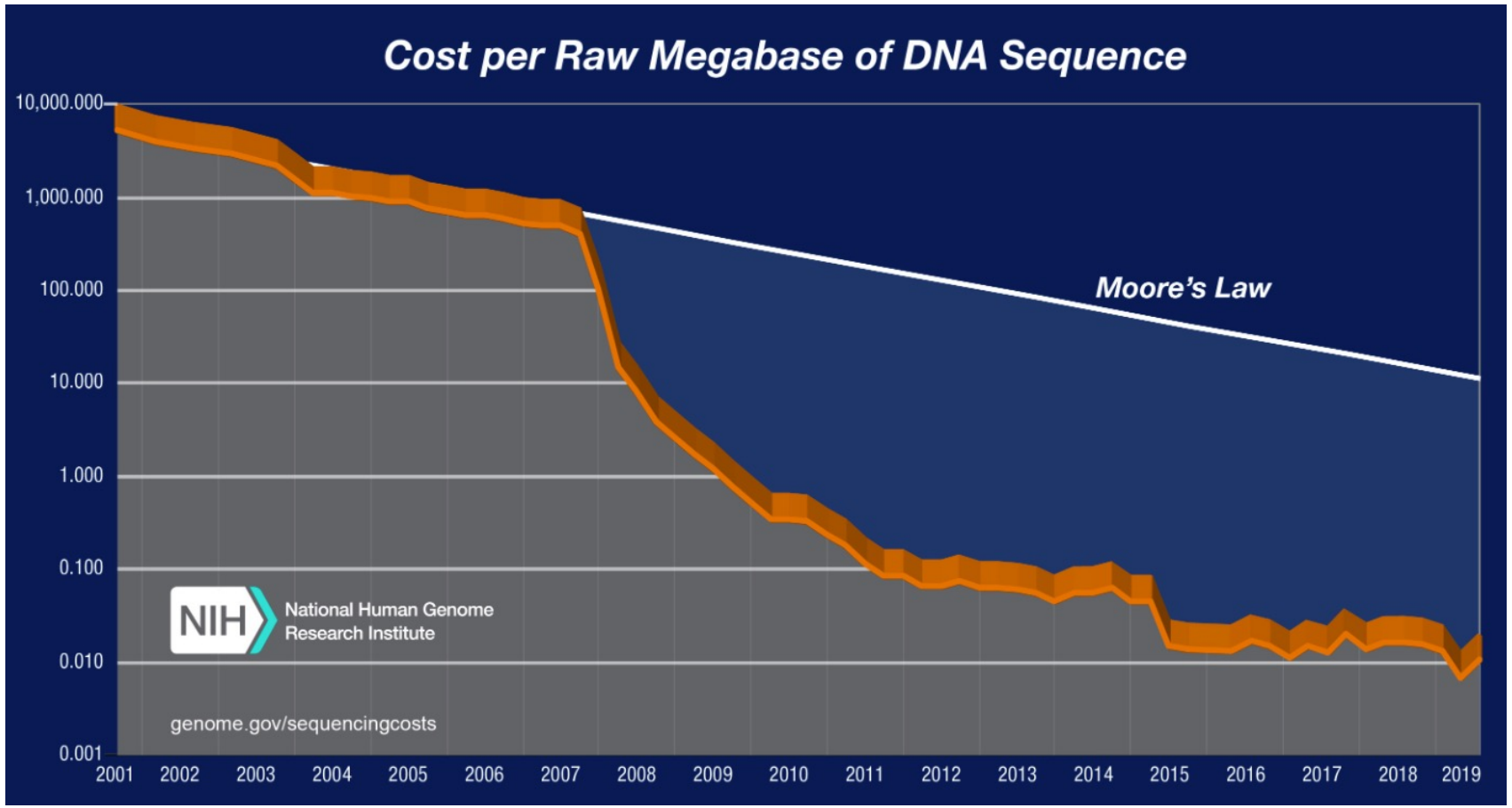
Common perception of bioinformatics..



Reality..


Terminal Shell Edit View Window Help
IGV
Human (hg38)
chr7
chr7:140,753,191-140,753,451
Go
259 bp
BRAF
Gene
Sequence
omw — d41294d@discovery7:~ — ssh • sh ssh_to_dartfs_discovery.sh — 110×36
myjobs
active jobs
eligible jobs
blocked jobs
Total jobs: 4

Data is getting cheaper & bigger



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

Complex analytics is playing a larger role

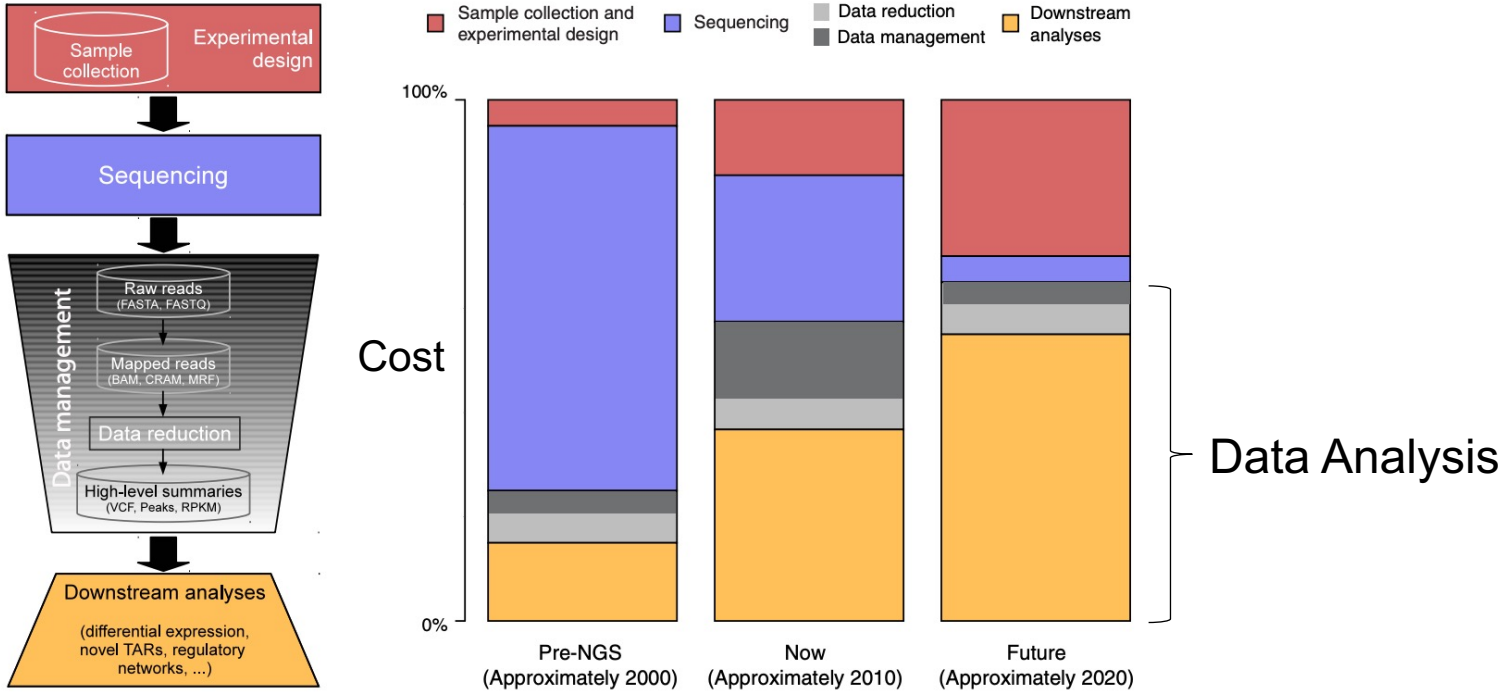
Sboner *et al. Genome Biology* 2011, **12**:125
http://genomebiology.com/2011/12/8/125

Genome Biology

OPINION

The real cost of sequencing: higher than you think!

Andrea Sboner[1,2], Xinmeng Jasmine Mu[1], Dov Greenbaum[1,2,3,4,5], Raymond K Auerbach[1] and Mark B Gerstein*[1,2,6]

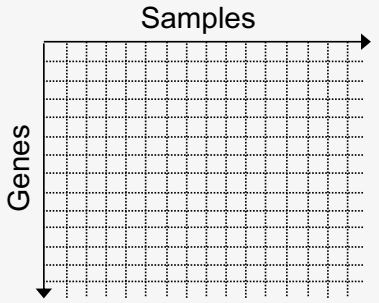


RNA-seq differential expression analysis pipeline

FASTQ files ⟸ Sequencing & QA/QC by sequencing provider

Pre-processing & quality control (HPC system)

Read trimming → **Mapping** → **Quantification** → **Gene expression matrix**



- Essentially all in *Bash/Shell*
- Most computationally intensive steps
- Many technical aspects for QC & processing that rely on library type and dramatically affect output

Spending a lot of time here figuring out how to do all this is not an effective use of time

Downstream analysis (local computer)

Differential expression analysis

Gene set enrichment analyses

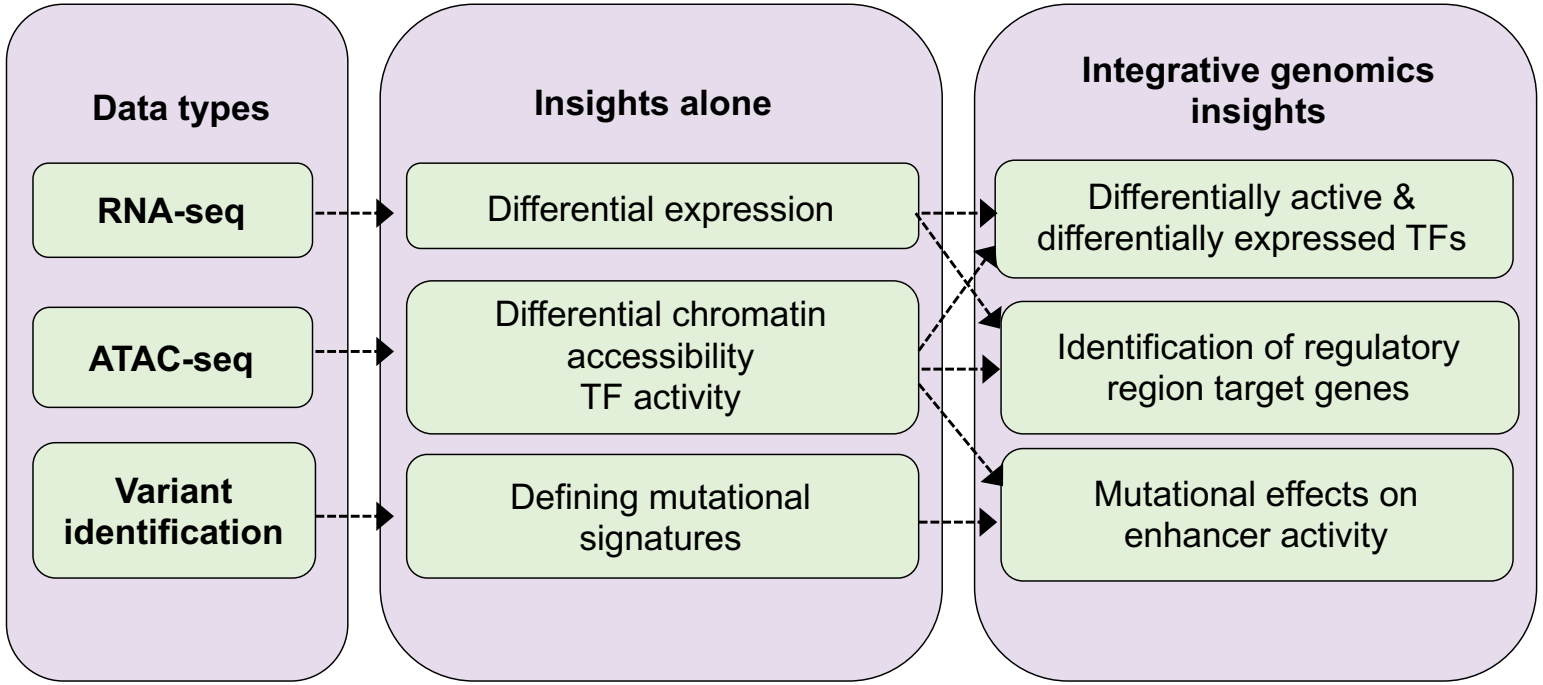
Integration with additional data

- Predominantly in *R/Bioconductor*
- Relies heavily on statistical knowledge

This side is where all the interesting analysis is

Beyond differential expression:
Integrative genomics

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



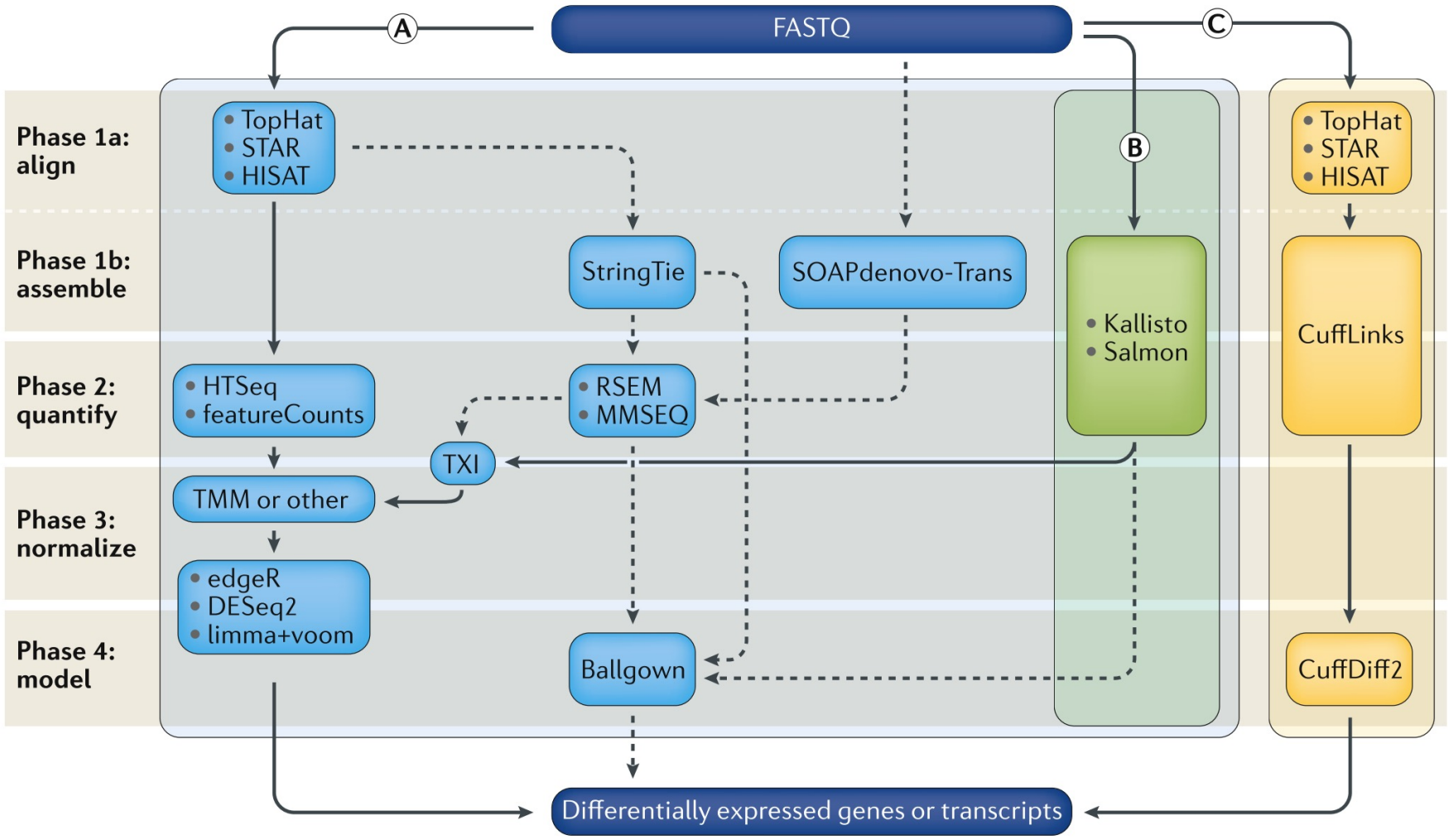
- You may not have generated each dataset in-house (e.g. combine in house & public data)

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)

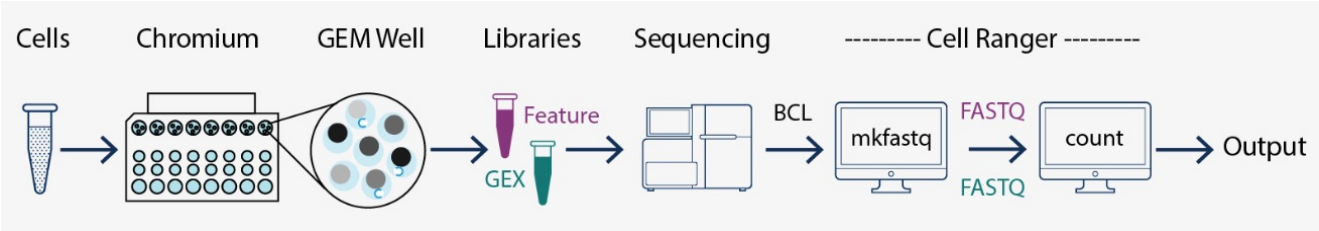


- Several tools available for each step
- Each have various strengths, weaknesses, applications

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

Single cell RNA-seq

10X Genomics – Cell Ranger pipeline

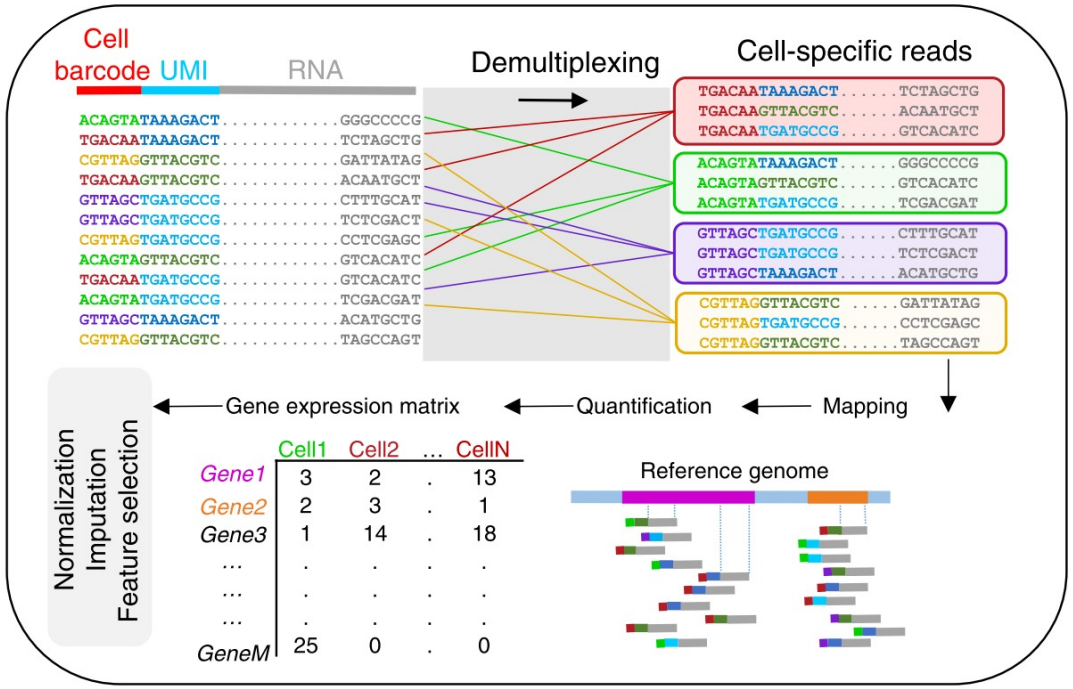


Includes alignment, filtering, barcode counting, UMI counting

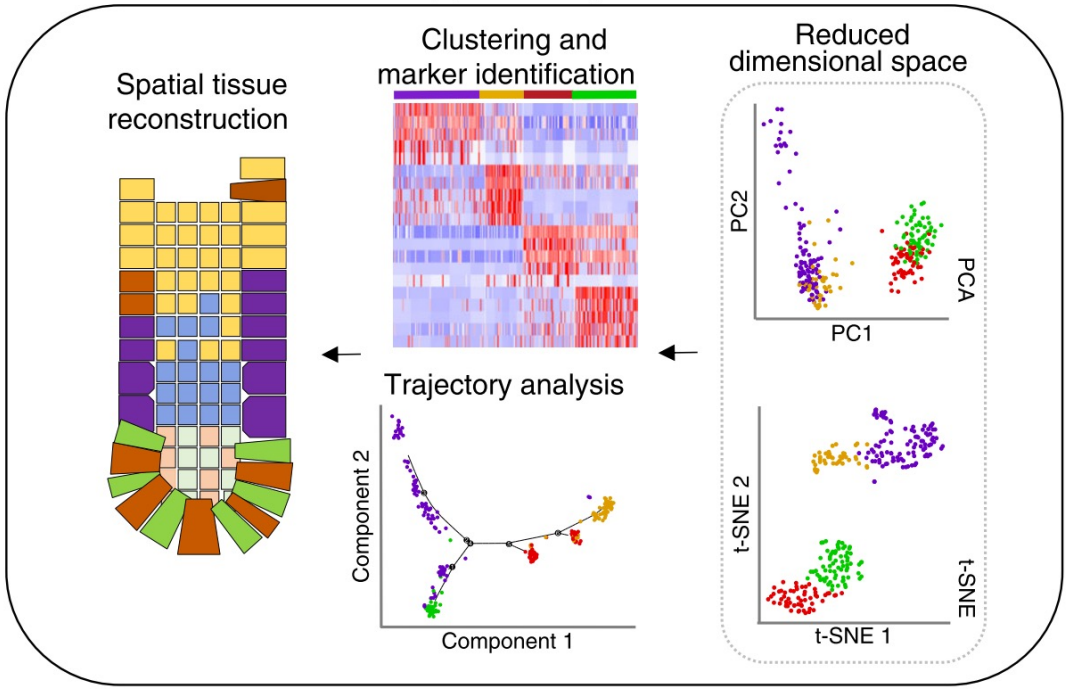
Bulk & single-cell RNA-seq are distinct and generally address different questions

- 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 processing



Data analysis



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