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

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**
  - 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**

RNA-seq: Overview

Sample/library preparation

Biological sample(s)

RNA isolation
Purification & selection

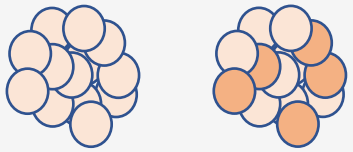
Library preparation
Fragmentation, reverse transcription, adapters, PCR

Sequencing
cDNA pool from samples

Reads in FASTQ format

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/AEAEEEEE6EEEAE/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
```

Demultiplexing

Data analysis

Downstream analysis
Differential expression
Isoform discovery
Annotation
Variant calling

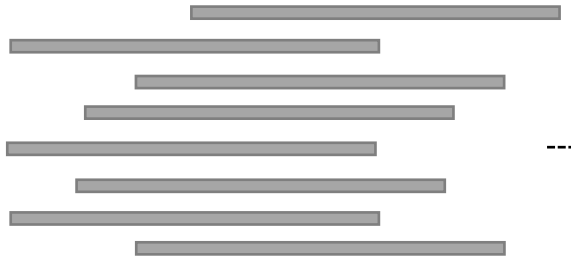
Data normalization
Library size & composition correction

Feature quantification
Simple counting
Statistical modeling

Genome mapping
Reference genome
De novo assembly

Quality control
Quality & adapter trimming

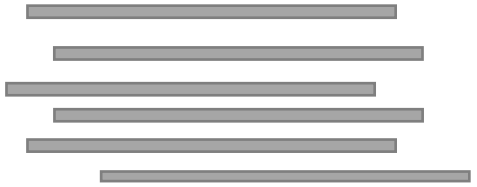
Library preparation for RNA-seq



Total RNA

Enrichment

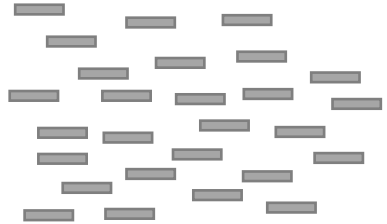
(e.g. PolyA capture, ribodepletion)



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

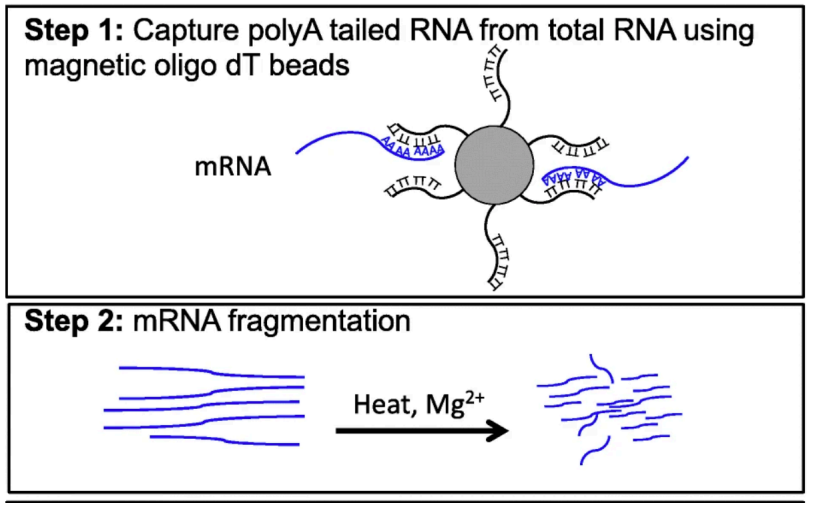
1 set for each sample

Sequencing ready cDNA pool

RNA selection/enrichment

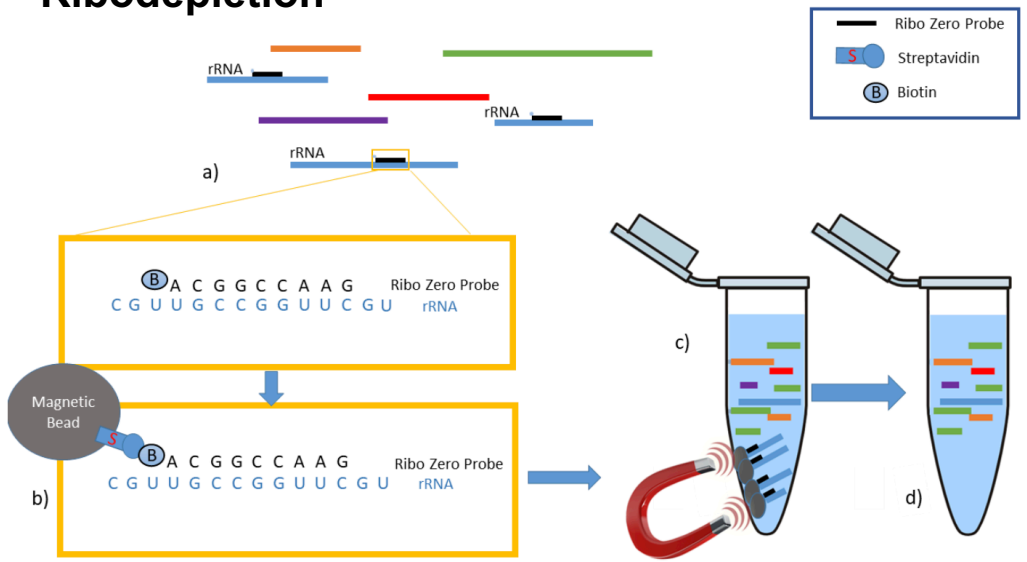
- **Multiple ways exist to enrich for the RNA you want to study**
- **Most of RNA in cell is ribosomal, and we don't want to waste sequencing all of this and little else**
- **Oligo-d(T) selection:**
  - Uses oligos of dT attached to magnetic beads to capture polyadenylated RNAs (mostly mRNA)
  - Magnet 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)
- **Size selection:**
  - Used to collect small non-coding RNAs, such as miRNAs

Oilgo-d(T) selection



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

Ribodepletion



Adapted from Illumina.com

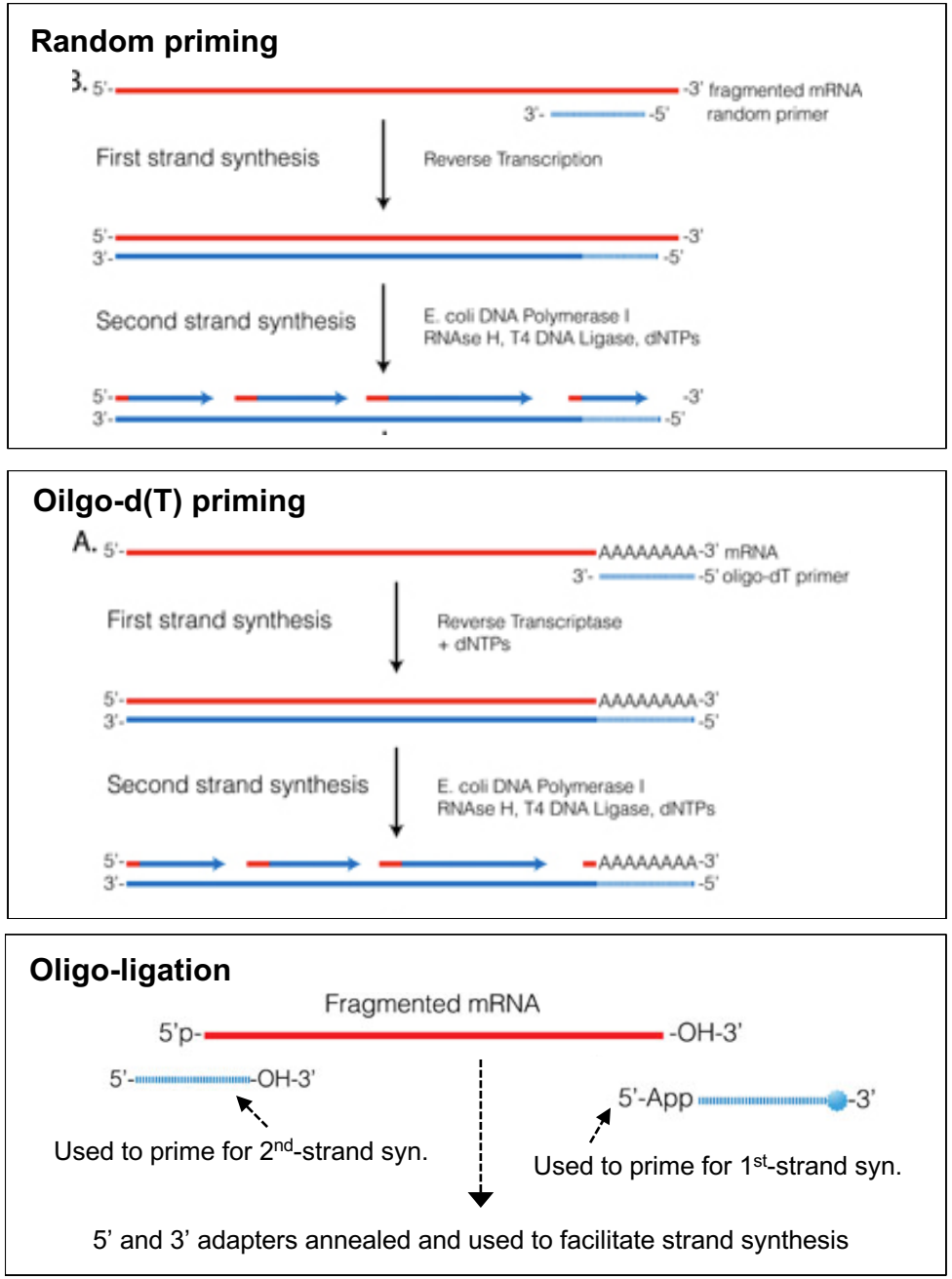
1st & 2nd-strand synthesis

- **Various approaches used to facilitate 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



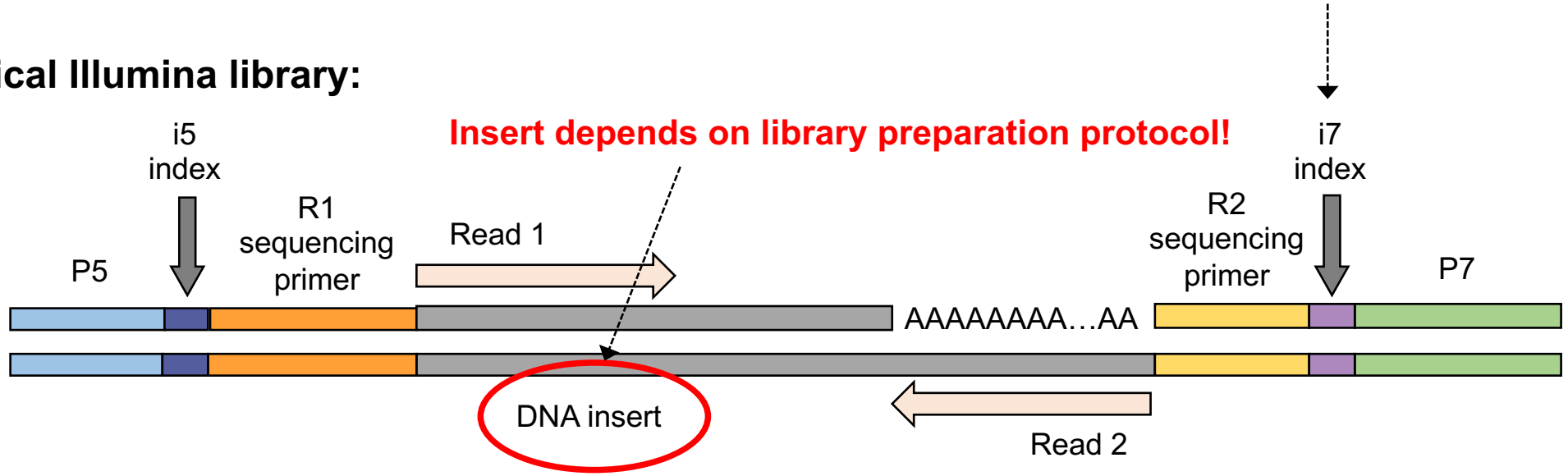
Images adapted from RNA-seqlopedia: UOregan

RNA-seq libraries and sequencing

- 1 library represents all the RNA fragments prepared from 1 sample
- Samples generally pooled for sequencing
- How you choose to sequence is your choice

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

Typical Illumina library:



Paired-end: Collect R1 and R2

Single-end: Collect R1 only

Choice affected by: Library type, desired hypothesis

Read length: No. of seq. cycles

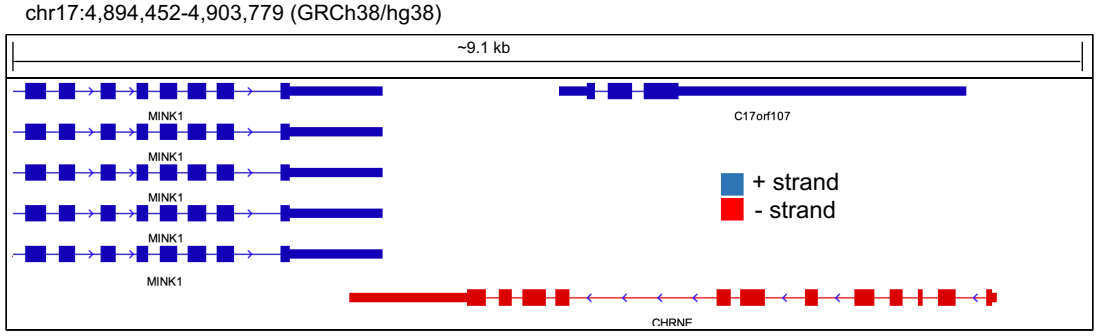
Seq. configuration: PE or SE + read length
e.g. PE 75bp

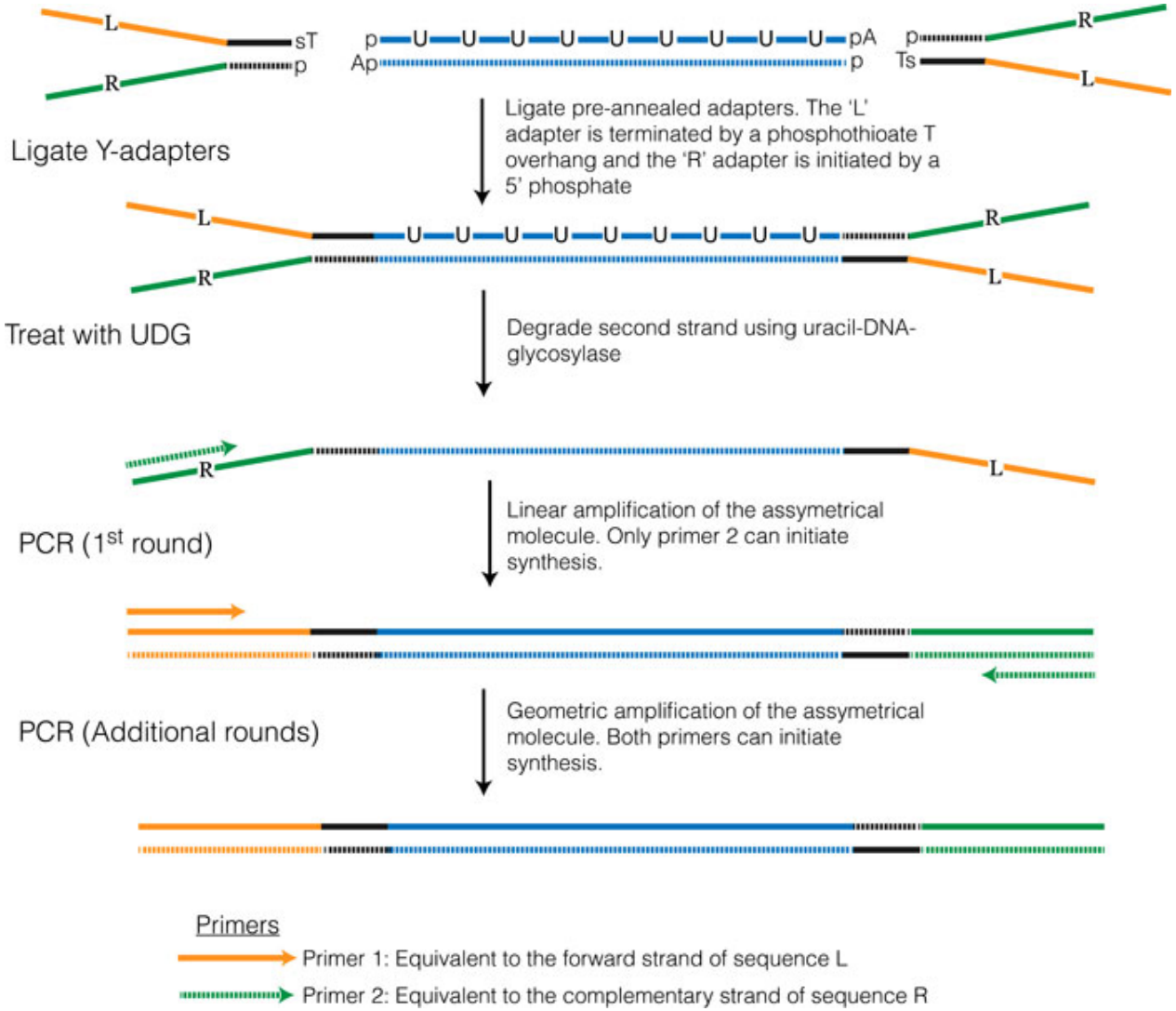
Stranded libraries

- Library prep. protocols that maintain which strand RNA template came from
- Standard protocols contain cDNA corresponding to *'sense strand'* AND *'anti-sense'* strand
- Most popular approach uses a DNA polymerase that can synthesize w/ dUTP or dTTP

Why do we care?

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





RNA-seqlopedia: https://rnaseq.uoregon.edu/

RNA-seq Library types

Full-length transcript (KAPA)

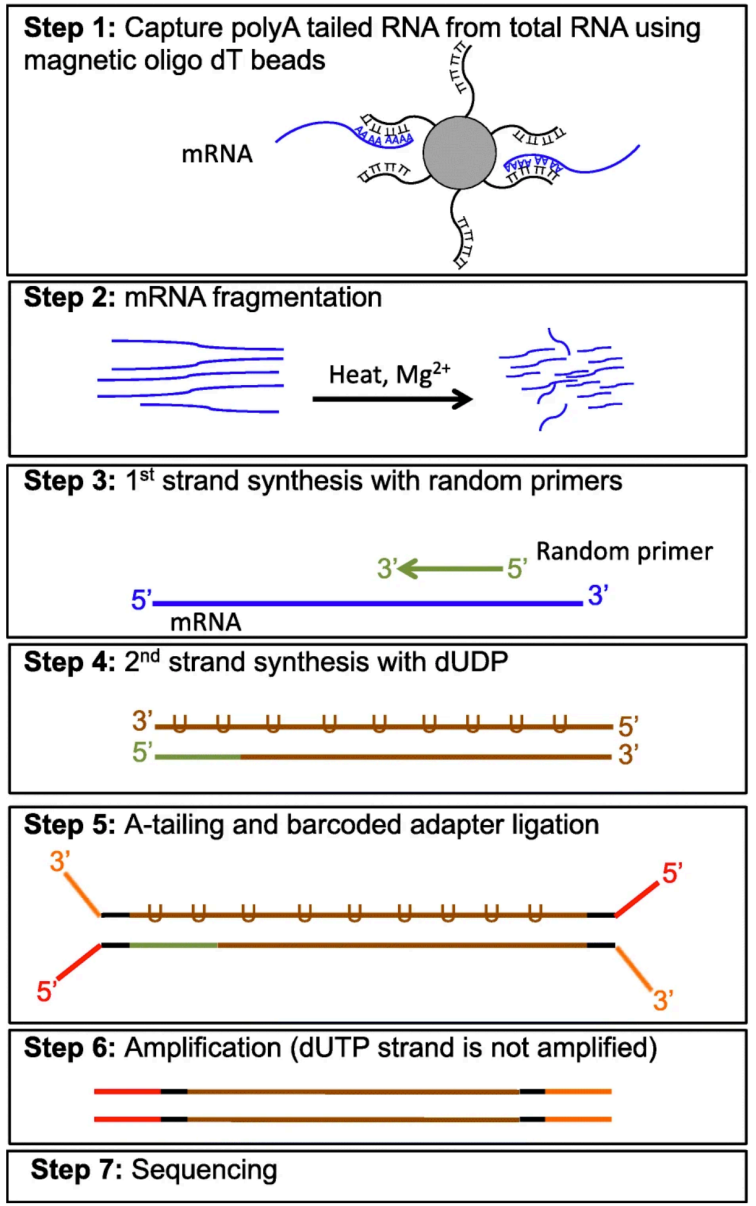
- cDNA generation using random hexamers
- Full-length transcript
- PE or SE
- Can call variants (PE)

3’ method (Lexogen, QuantSeq)

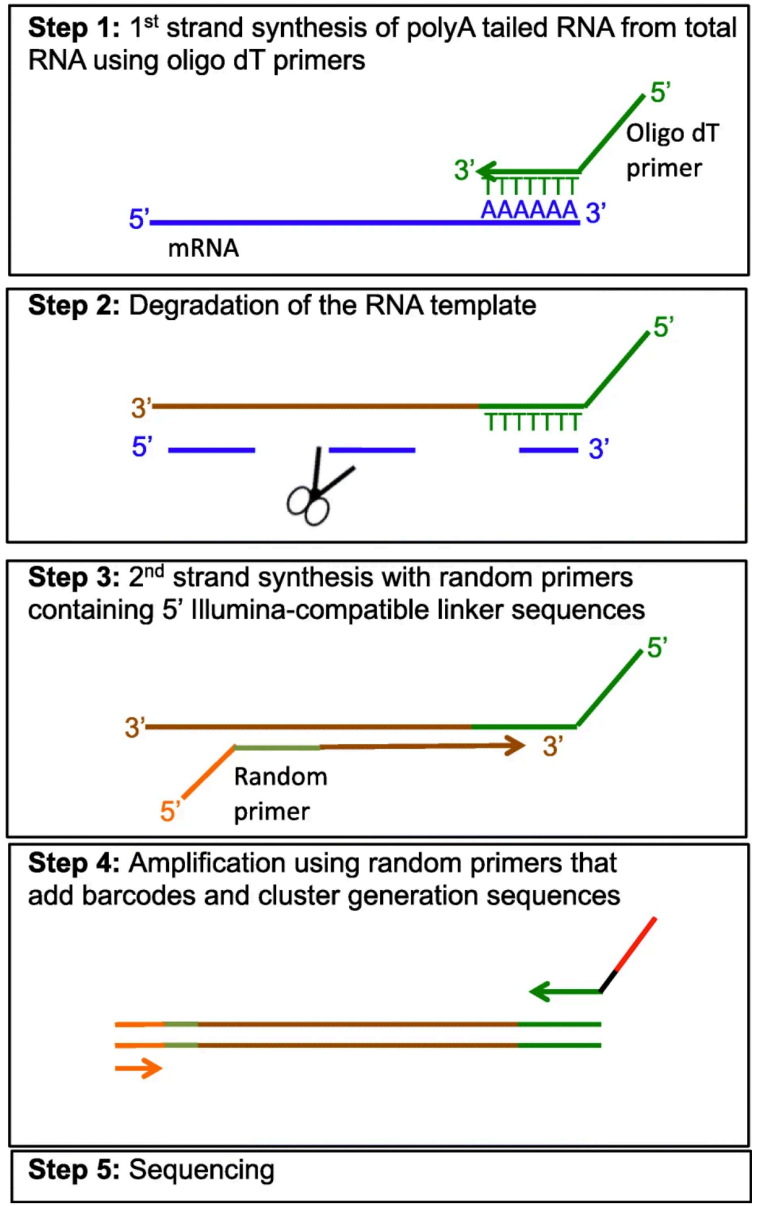
- 3’ –end of transcript only
- No transcript information
- High PCR duplicate no.
- Only eukaryotic samples
- Cannot be used to identify variants
- Paired-end is pointless
- **Big cost savings**

Both produce stranded libraries
(only sequence transcribed strand)

Traditional method (KAPA)

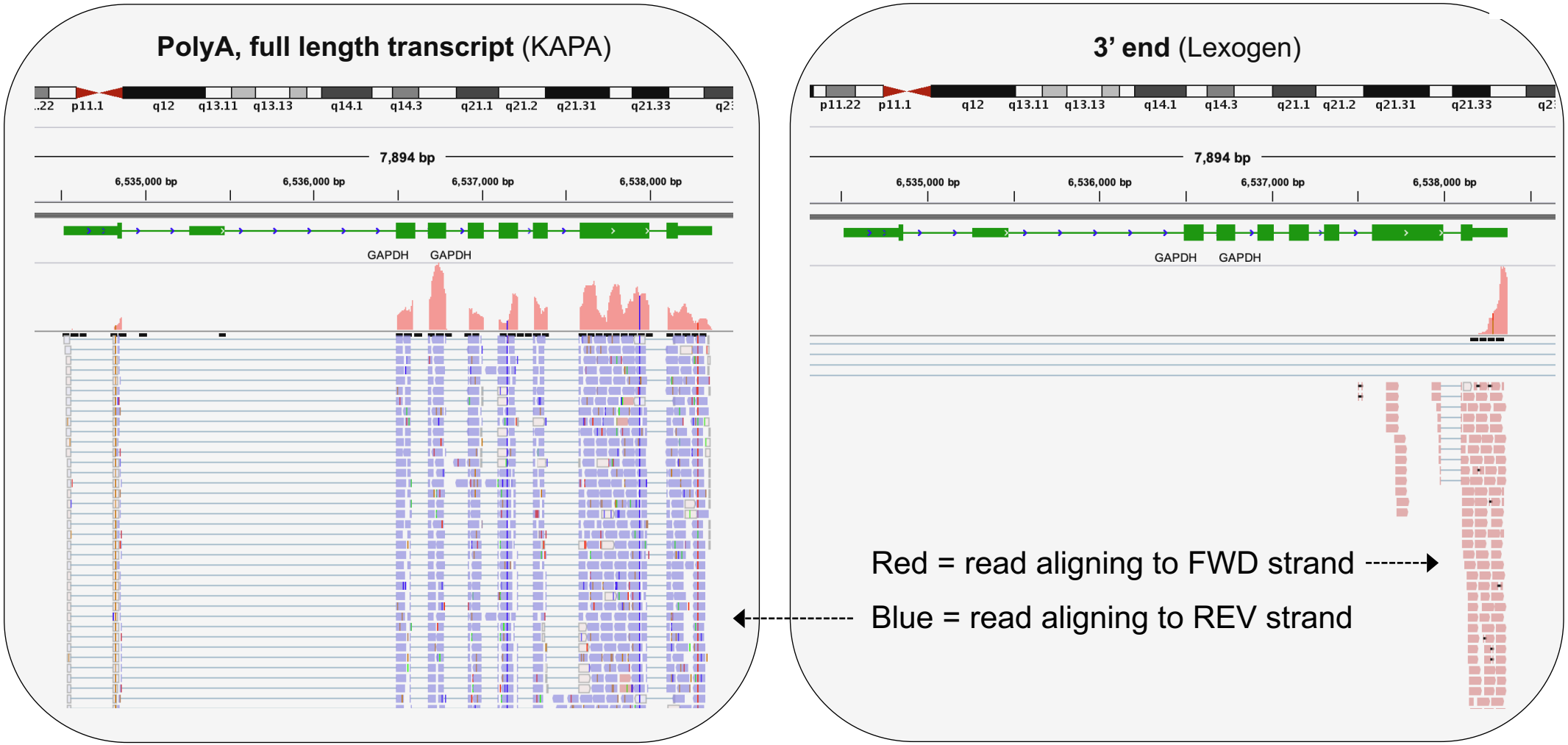


3’ method (LEXO)



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

Data from each workflow looks different

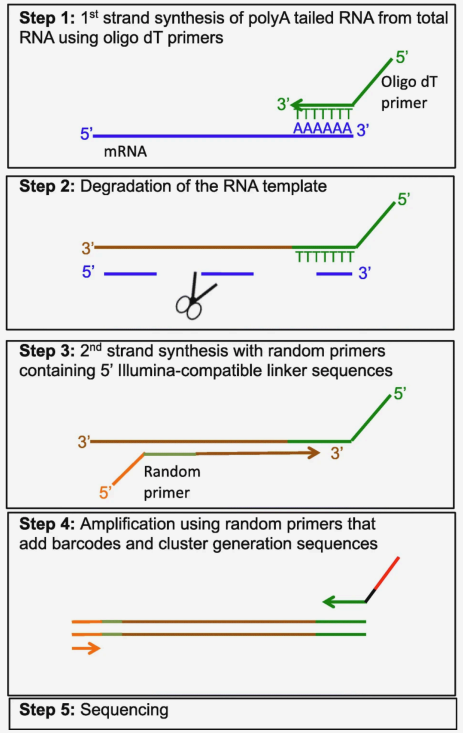


Quality control, analysis pipelines, and possible hypotheses therefore inherently differ

RNA-seq library types

3’ End

3’ method (LEXO)

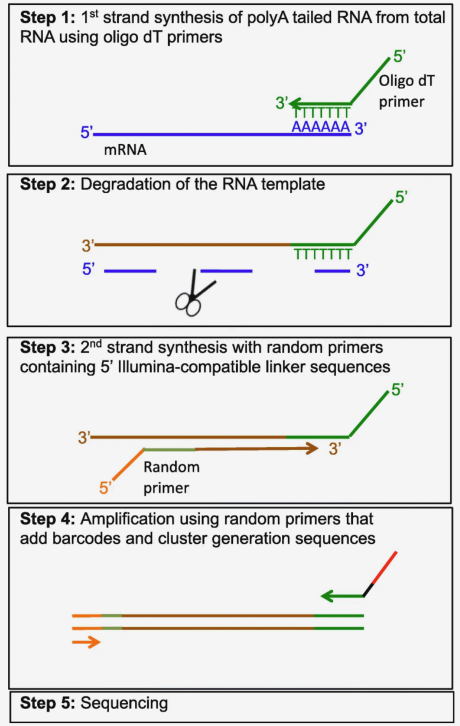


Adapted from Fukua *et al*, 2019. *Genom. Biol*

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

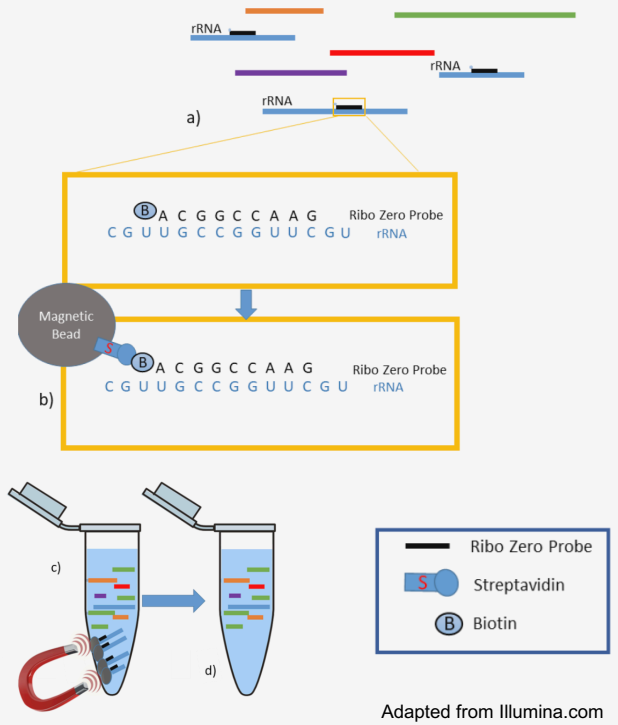
Full-length - PolyA

3’ method (LEXO)



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

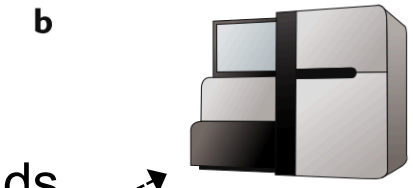
Cost / Data Richness

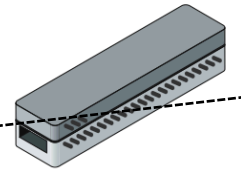
Image Credit:
Lexogen Inc

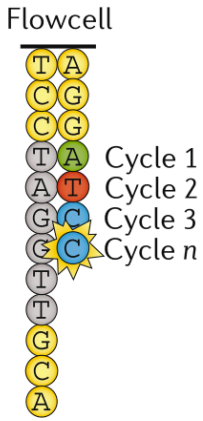
Sequencing technologies for RNA-seq

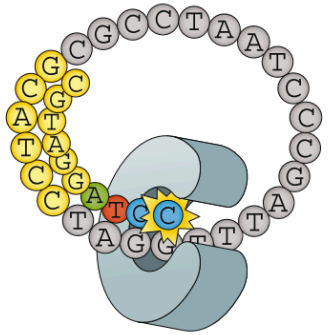
Short reads (hundreds)

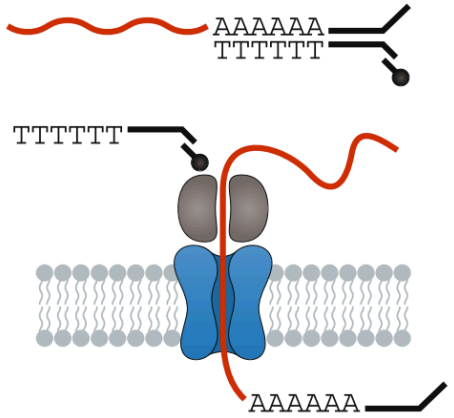
Very long reads (kb)



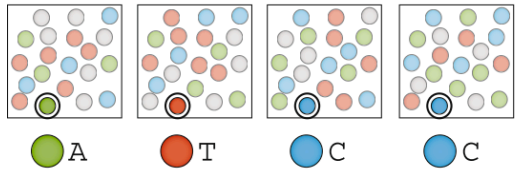


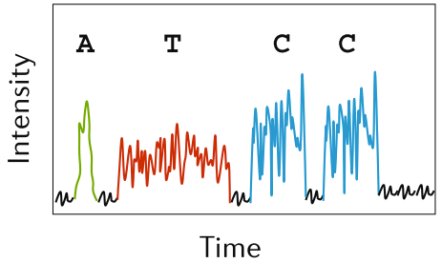


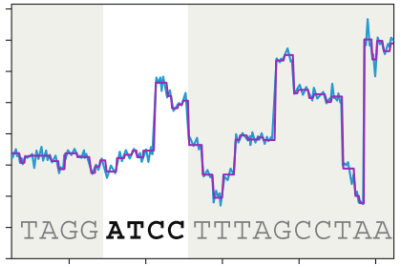




SmidgION



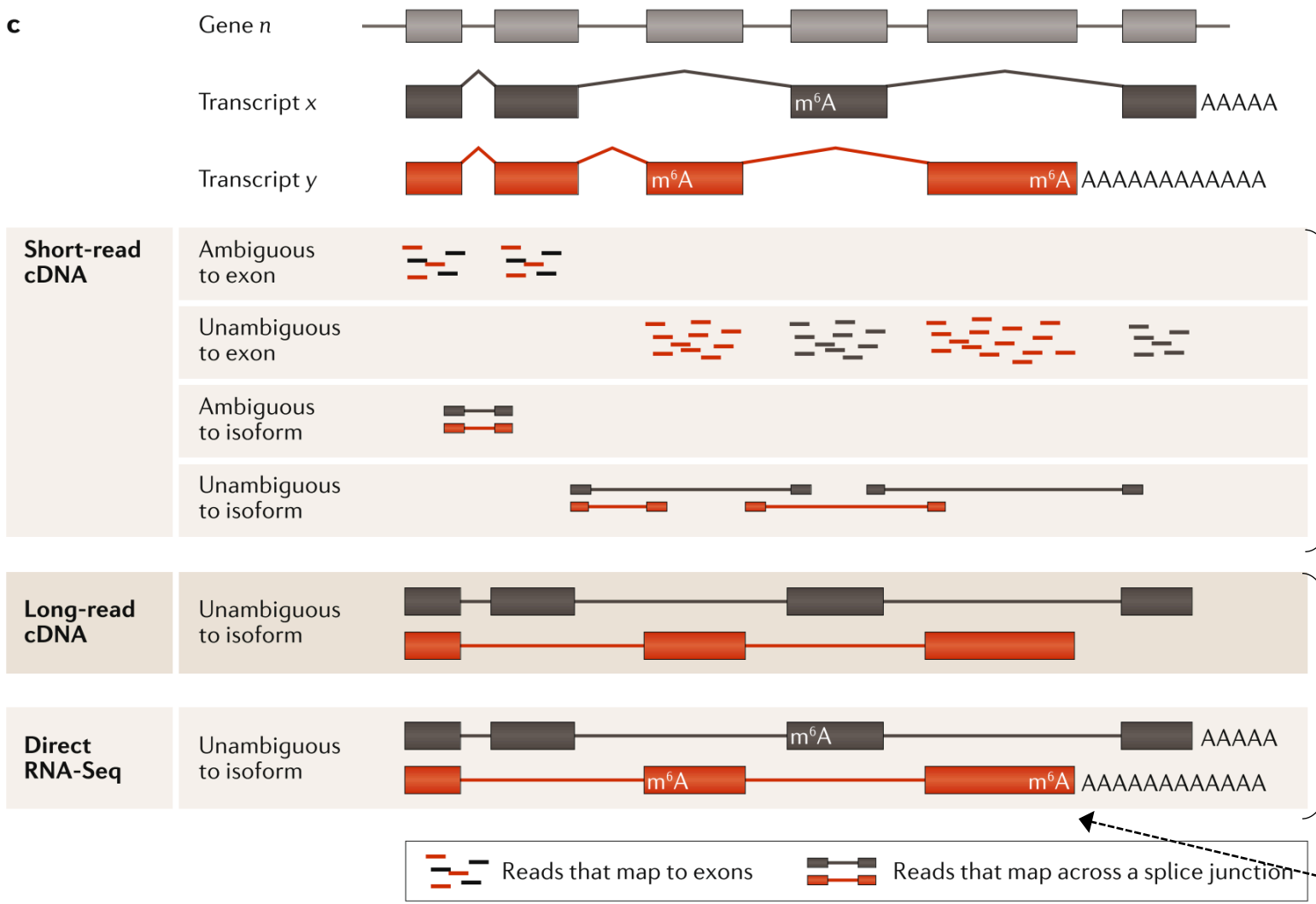




MinION

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

Different tech., different data., different applications



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

Best for standard differential gene/transcript expression analysis

- Isoform discovery
- De novo transcriptome
- Fusion transcripts
- MHC, HLA transcripts

+

- Directly measure RNA modifications

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
- Number needed replies on multiple factors, including true effect size, and within-group variation

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

From DESeq2 documentation:

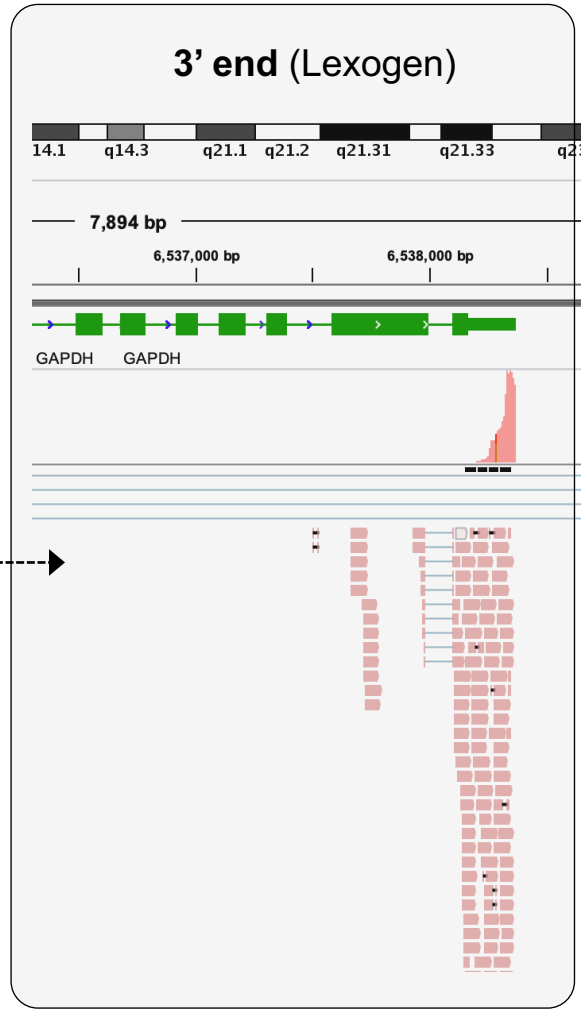
Can I use DESeq2 to analyze a dataset without replicates?

No. This analysis is not possible in *DESeq2*.

- Suggested minimum no. of replicates should = 6
- The more heterogenous your sample become, the more replicates you will need to achieve adequate statistical power (e.g. human tissues have higher within group variance than cultured cell lines)

Sequencing depth (or coverage?)

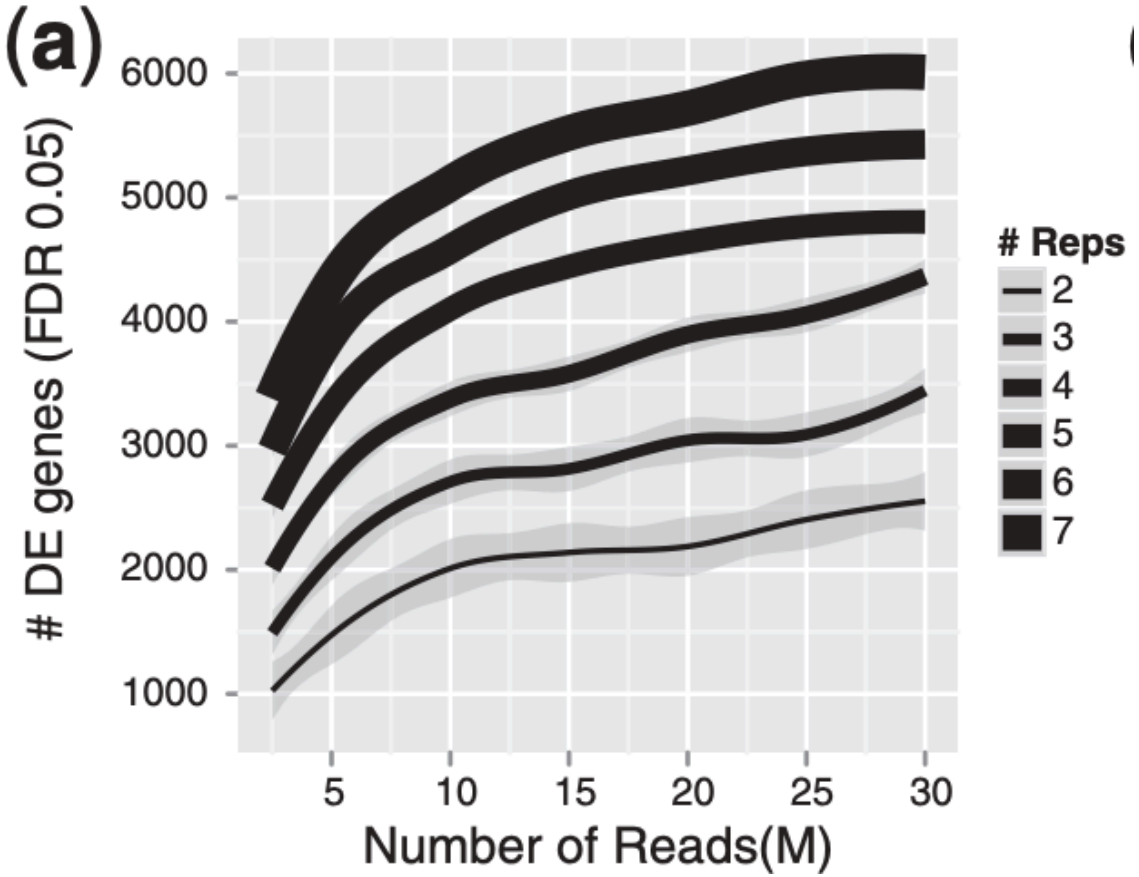
- '*Coverage*' doesn't have much meaning for transcriptome data
  - We are less concerned with how many times we cover a specific locus w/ a read..
- Generally total of <u>10-30 million reads</u> for DGE of eukaryotic genomes
- Some species require many fewer than this
- Technology also affects required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

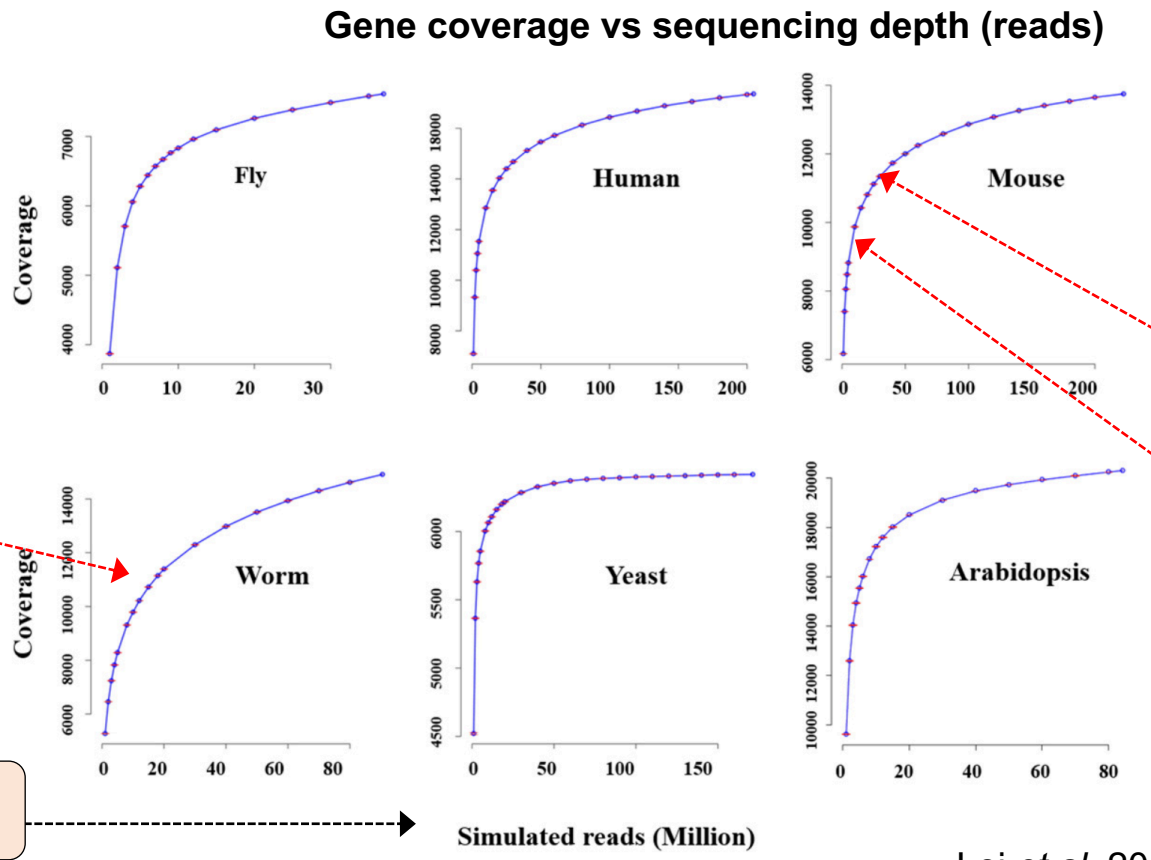


Liu *et al*, 2014, *Bioinformatics*

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

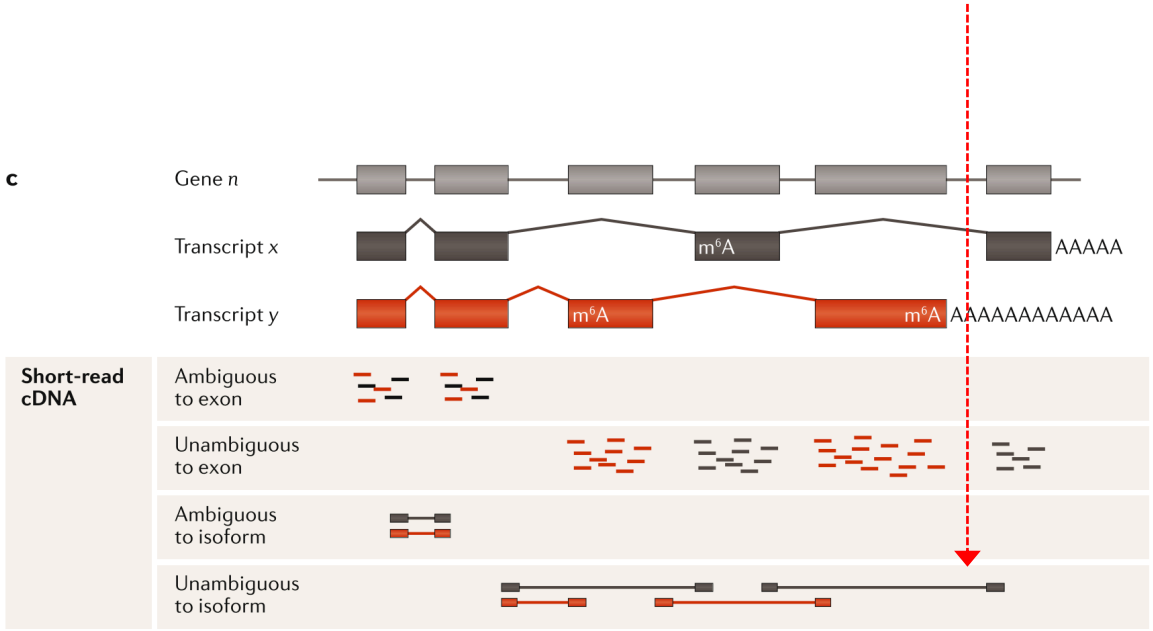
Different complexities can introduce bias!

Lei *et al*, 2015, *Gene*

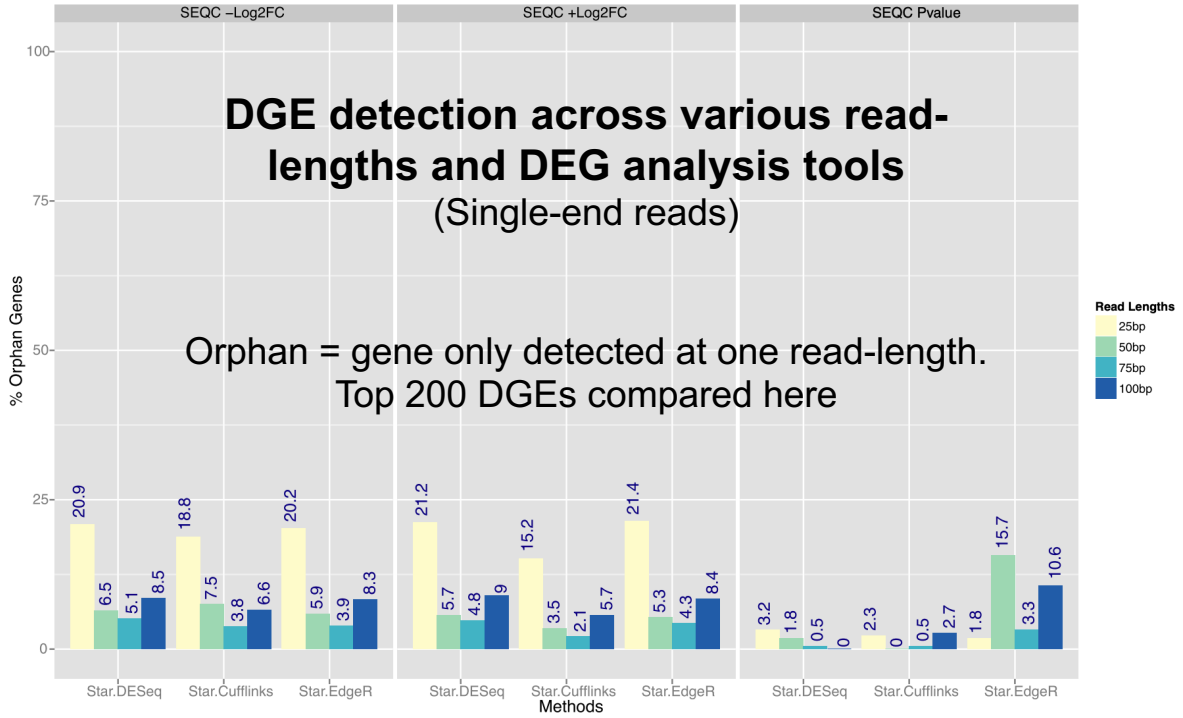
- **Pilot studies can help identify replicate no. & sequencing depth needed**

Read configuration

- For DGE, the minimum read length that is useful is the length required to accurately map a read to a gene
- For isoform detection, longer reads help resolve transcripts



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

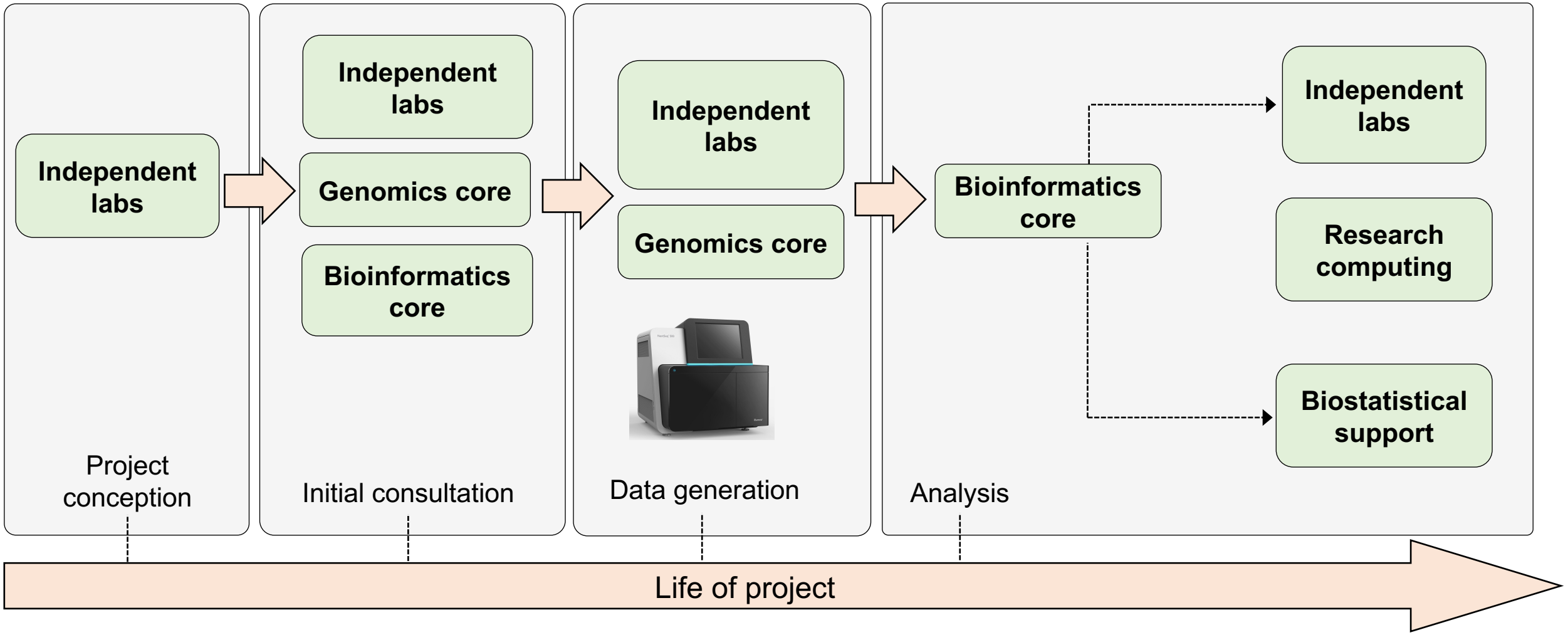


Chhangawala *et al*, 2015, *Genom. Biol.*

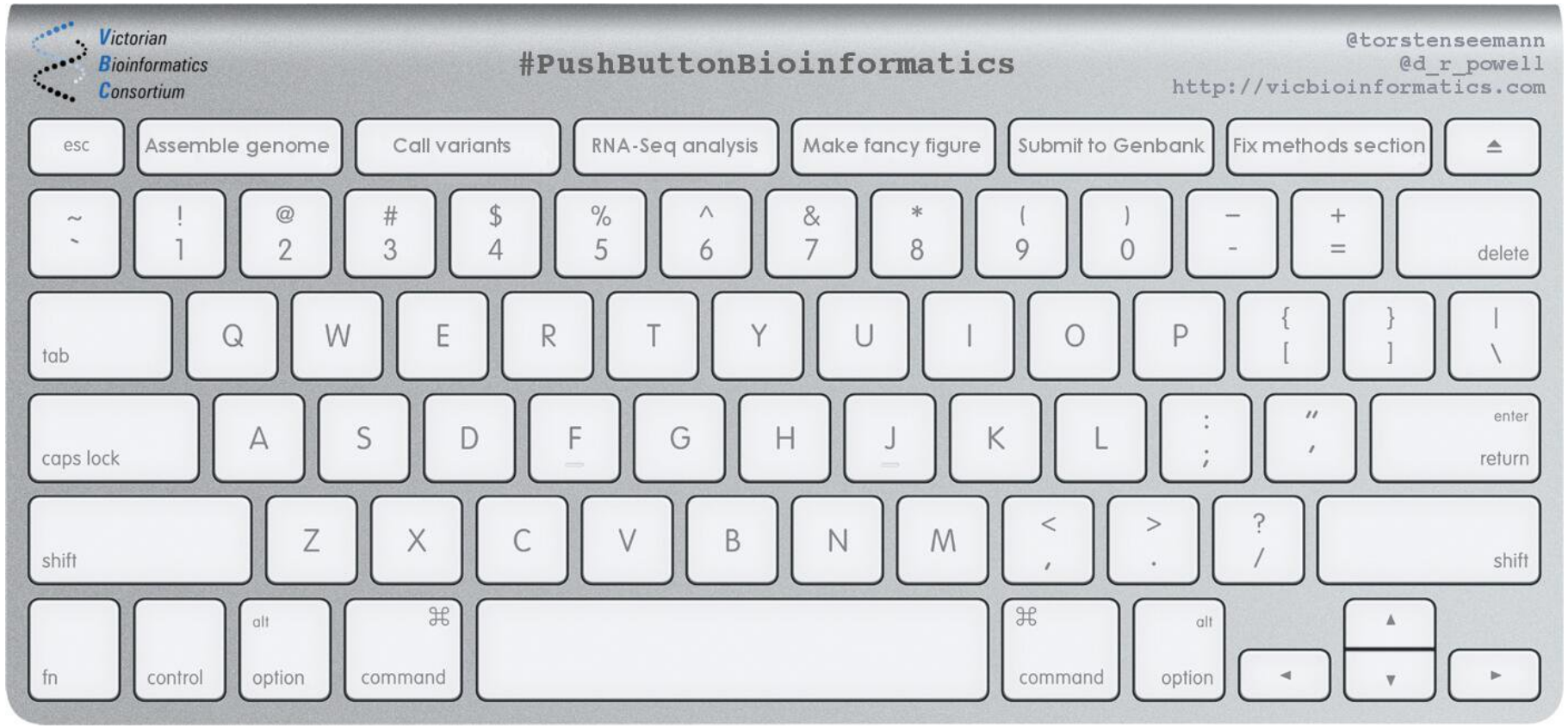
- Paired-end (PE) useful for isoform-detection, alternative exon usage, & variant calling from RNA-seq
- If you only care about DGE, invest in more replicates and more reads (and then more replicates again), instead of long or PE reads

Where does the Bioinformatician fit in?

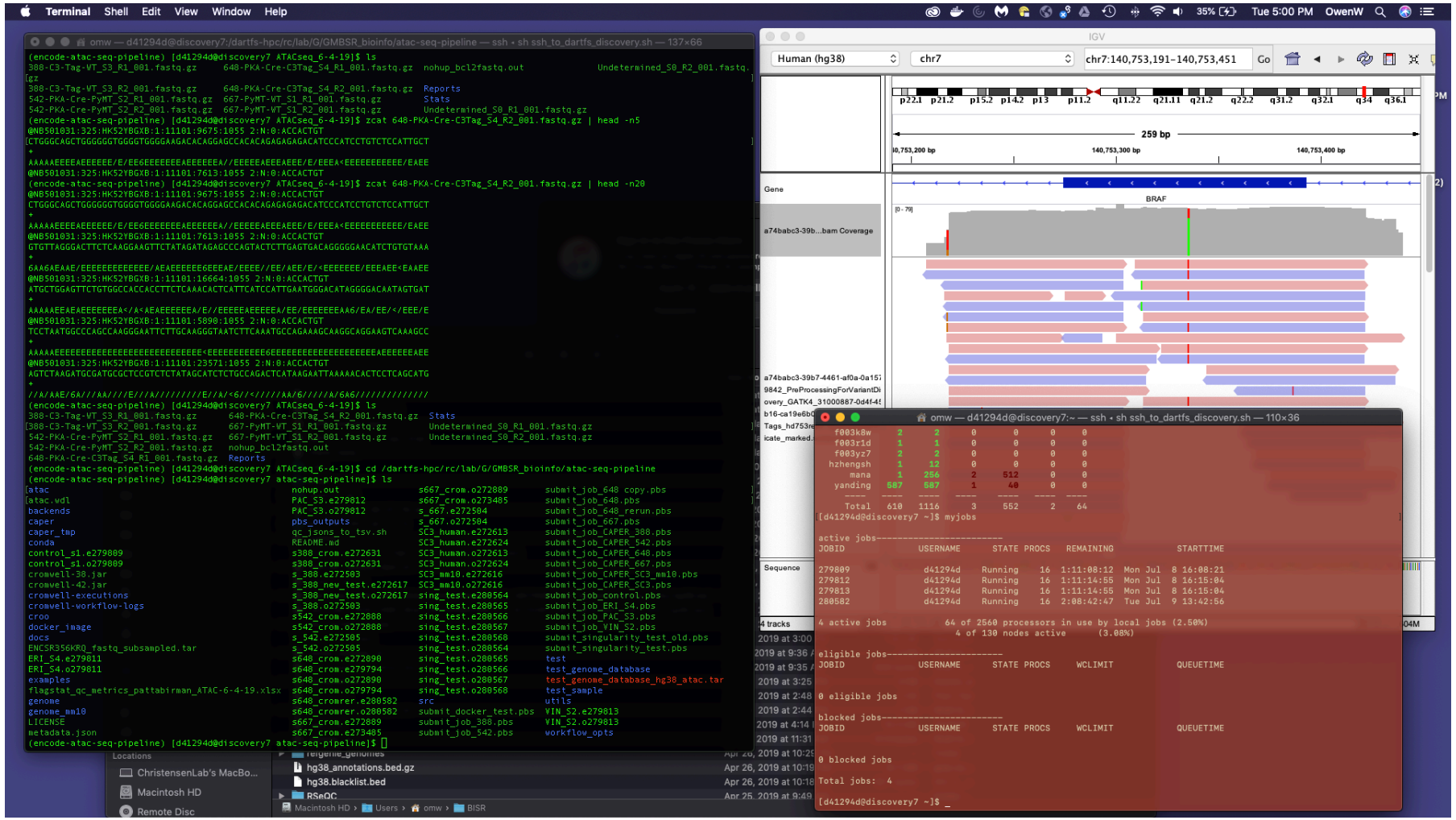
- Ideally, before data generation..



Common perception of bioinformatics..

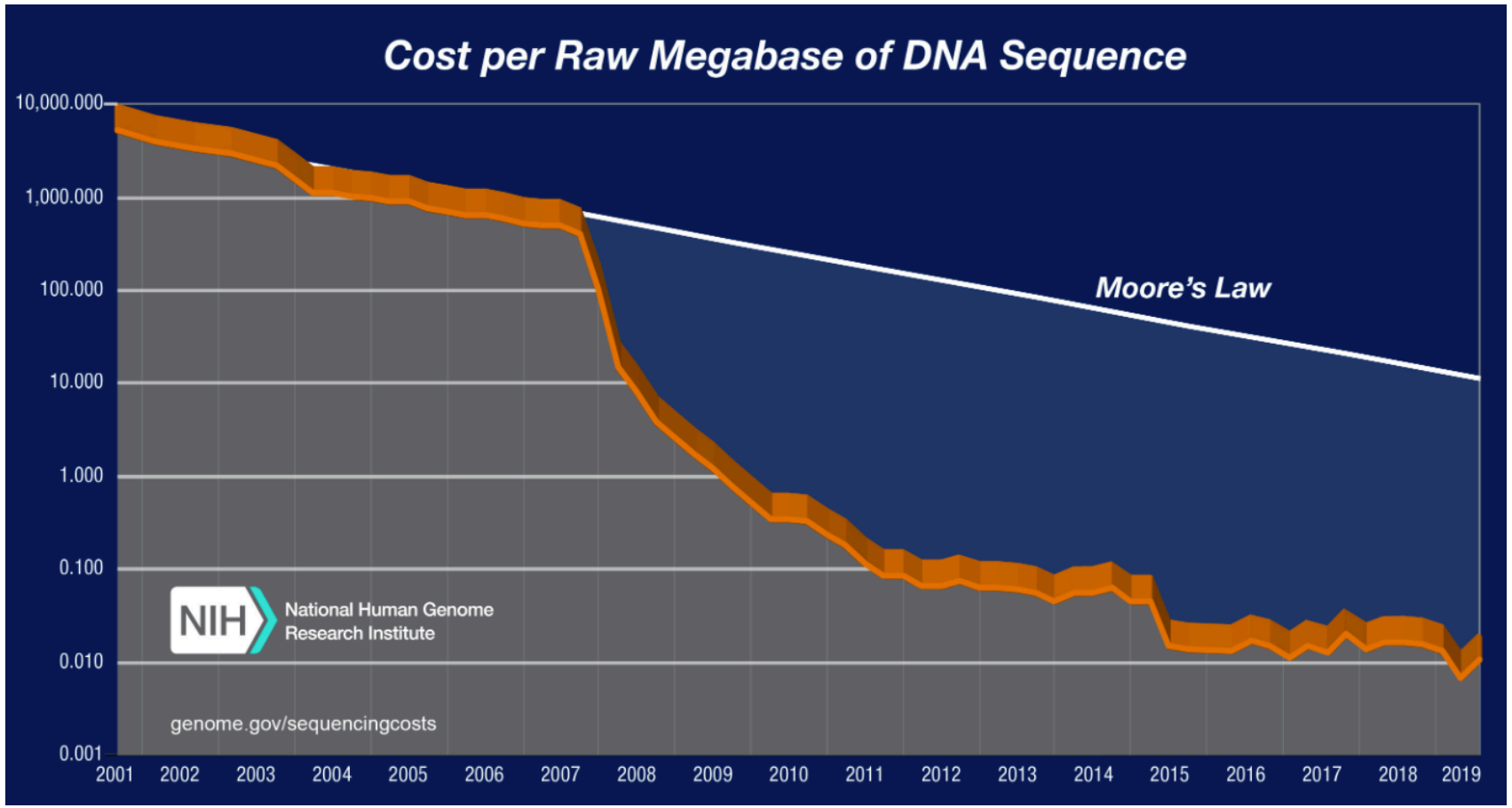


Reality..



Bad bioinformatics looks really easy..

Data is getting cheaper & bigger



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

Complex analytics is playing a larger role

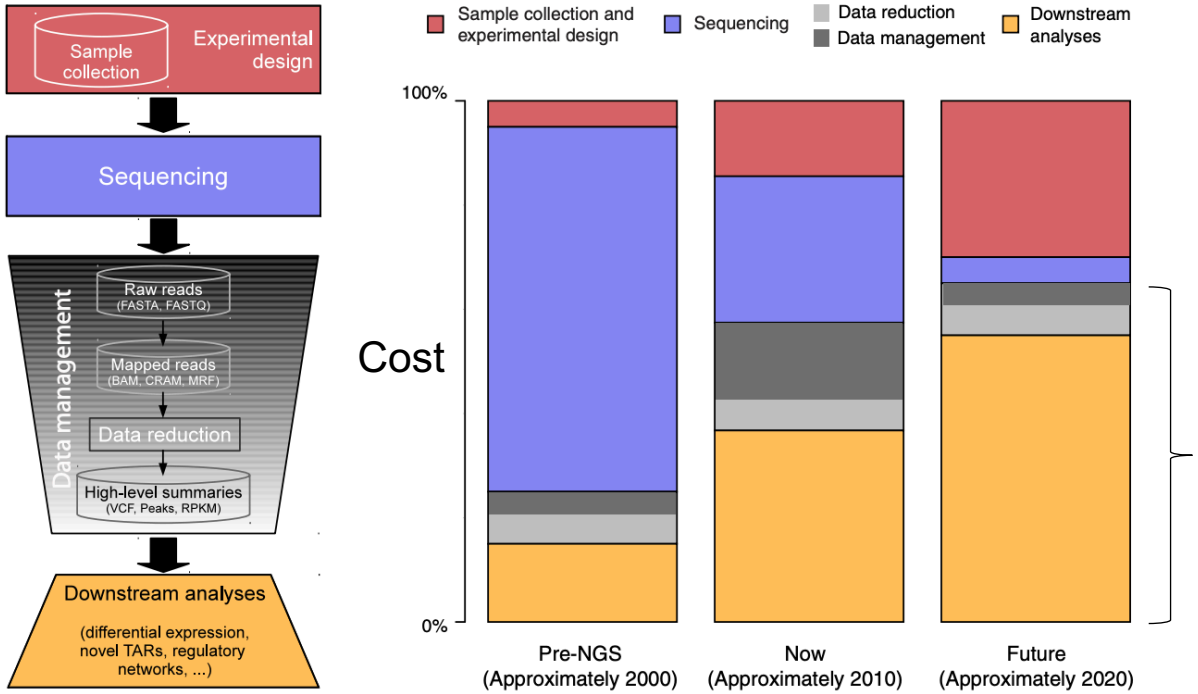
Sboner *et al. Genome Biology* 2011, **12**:125
http://genomebiology.com/2011/12/8/125

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]



"The amount spent on bioinformatics will be at least as much spent on sequencing and closer to double the cost."

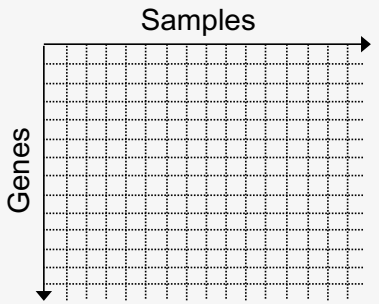
Iowa State Bioinformatics core

RNA-seq differential expression analysis pipeline

FASTQ files ⟸ Sequencing & QA/QC by sequencing provider

Pre-processing & quality control

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

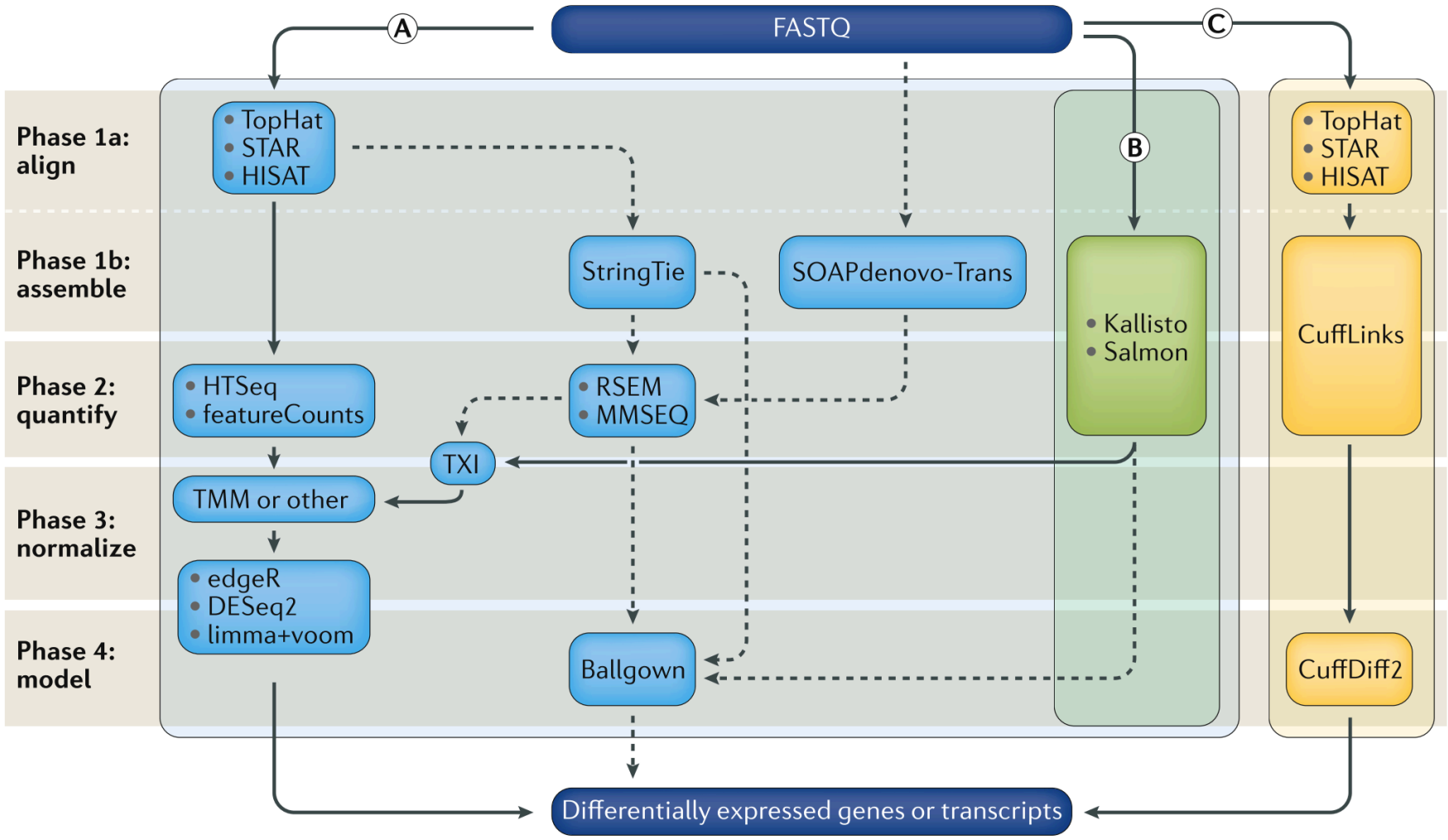
Differential expression analysis → **Gene set enrichment analyses**, **Integration with additional data**

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

This side is where all the interesting analysis is

Many ways to do each step..

Differential expression workflow(s)



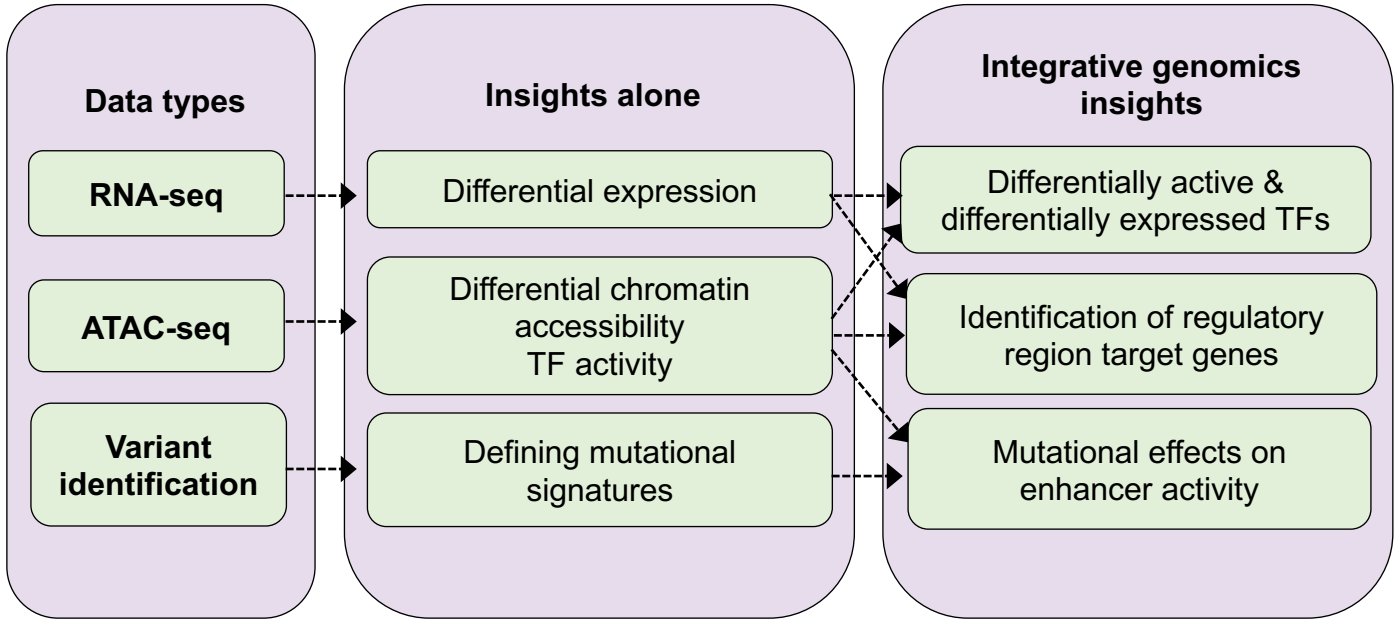
- Several tools available for each step
- Each have various strengths, weaknesses, applications

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

Beyond differential expression:
Integrative genomics

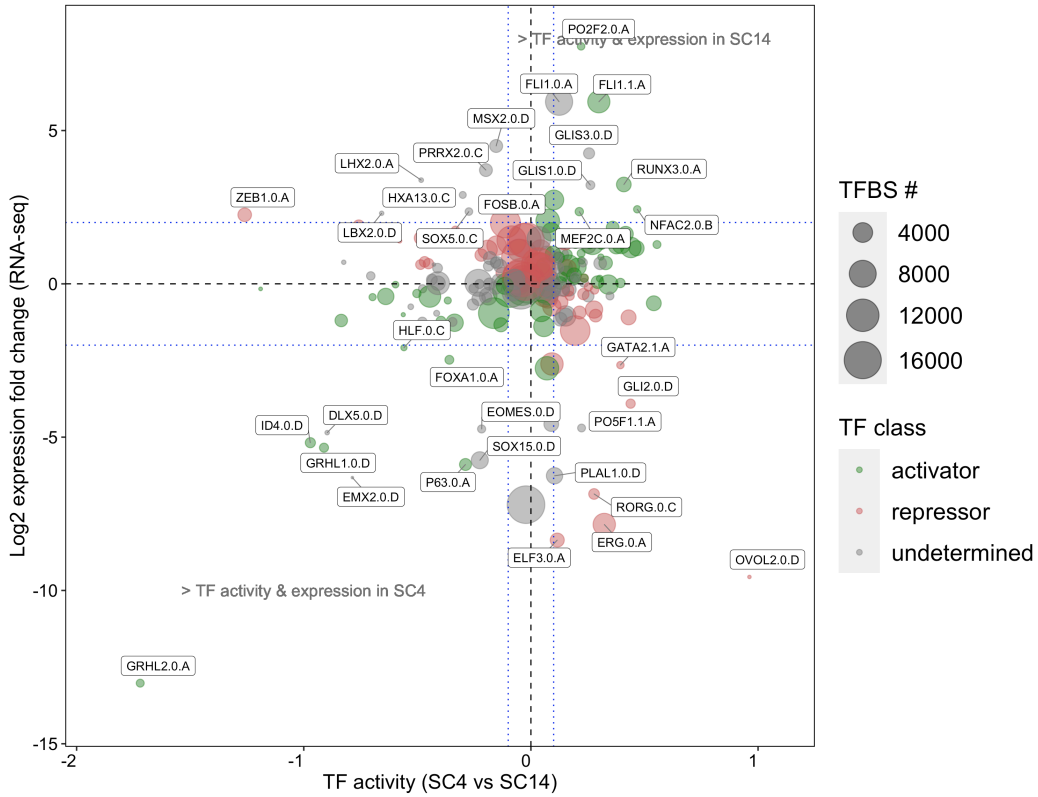
- *Leveraging data integration across more than one 'omics platform, to reveal insights not possible with each data type alone*

Example:



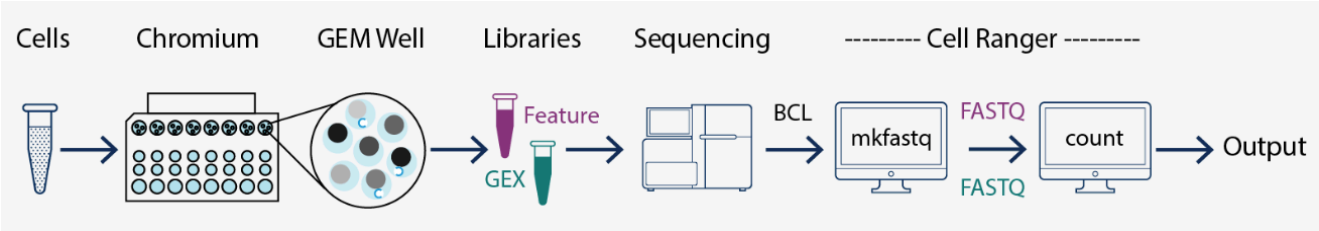
- You may not have generated each dataset in-house

Diff. expression (RNA-seq) vs diff. TF activity (ATAC-seq)



Single cell RNA-seq

10X Genomics – Cell Ranger pipeline

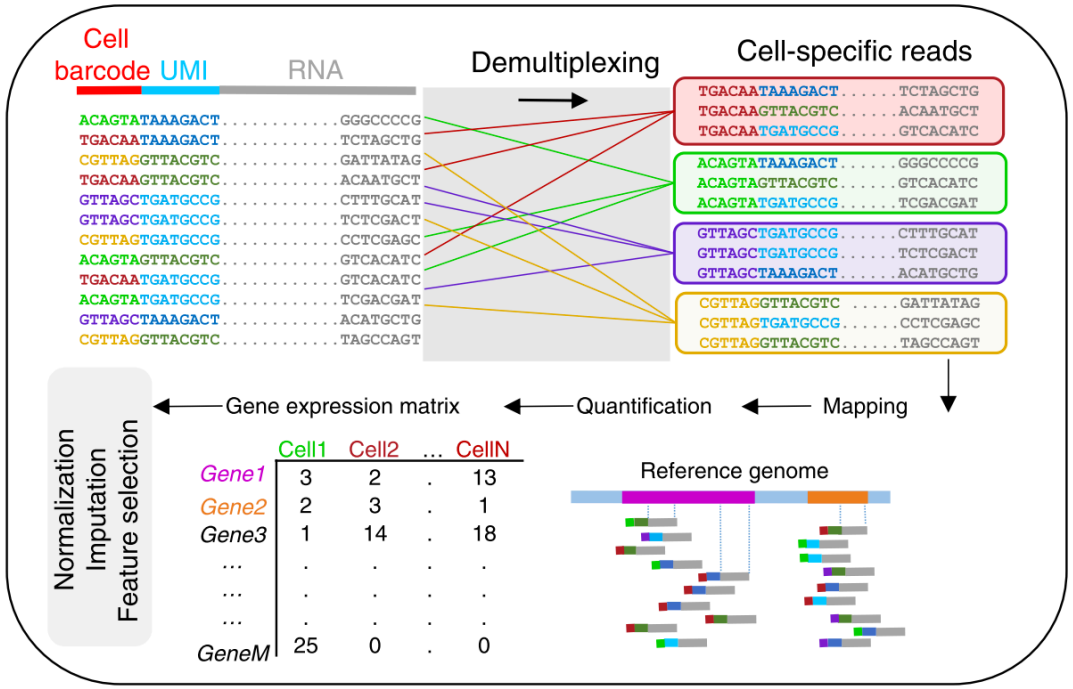


Includes alignment, filtering, barcode counting, UMI counting

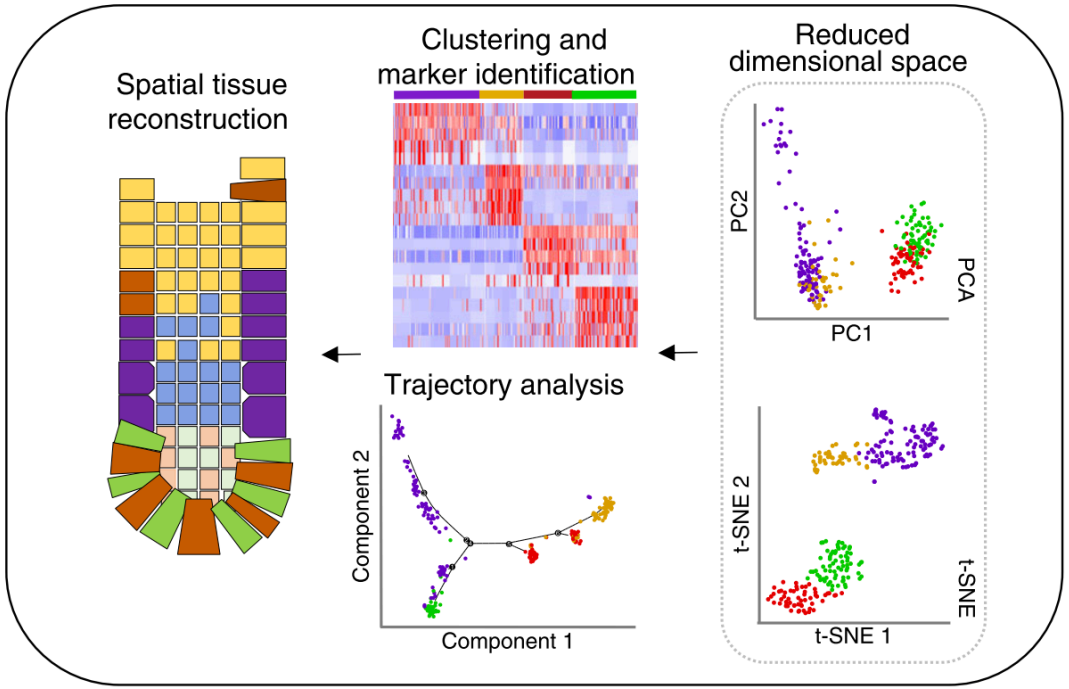
Bulk & single-cell RNA-seq are distinct and generally address different questions

- 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

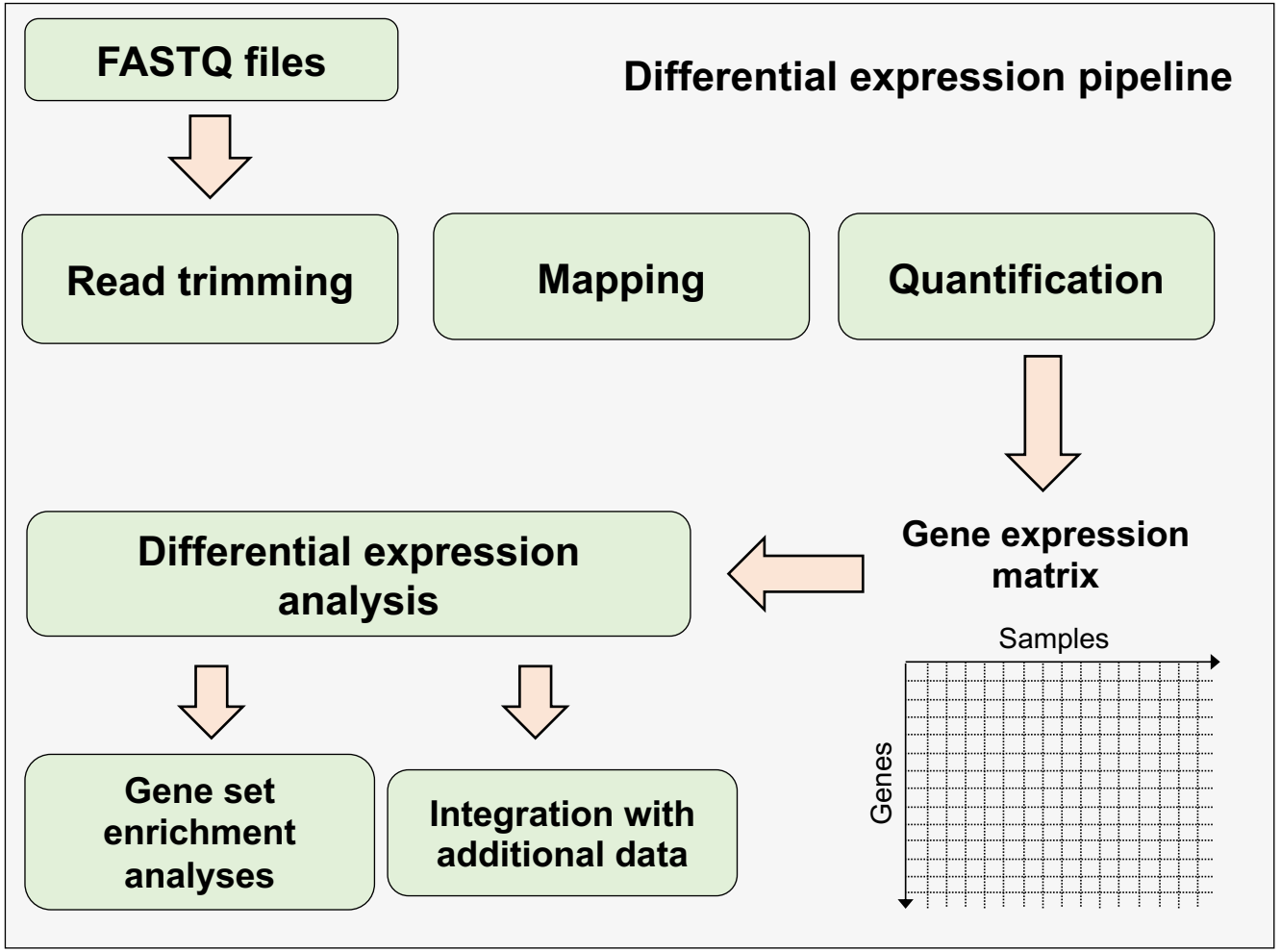


Adapted from, Lafzi, 2018, *Nat Protocols.*

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?


FASTQ files
Differential expression pipeline
Read trimming
Mapping
Quantification
Gene expression matrix
Samples
Genes
Differential expression analysis
Gene set enrichment analyses
Integration with additional data