RNA-seq

Introduction

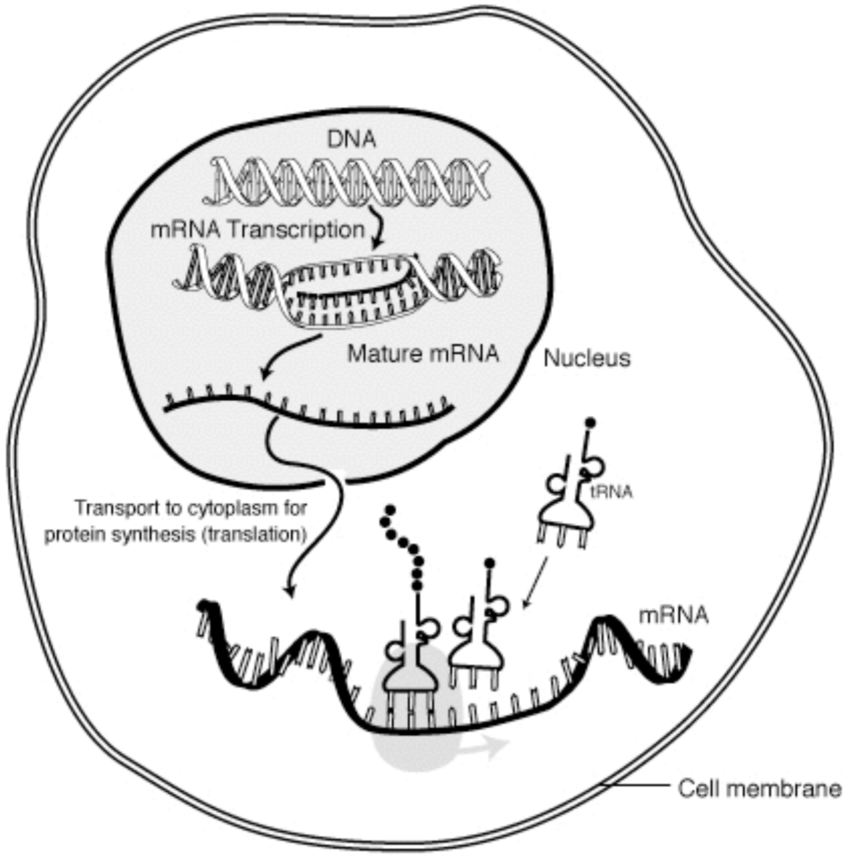
- Associate Bioinformatics Scientist at Coriell Institute for Medical Research
- Help other people with their data analysis
- Independent data analysis
- All materials for my lectures will be online at https://kelseykeith.github.io/2020_bmsc8203_bioinformatics_lectures/

CORIELL INSTITUTE
FOR MEDICAL RESEARCH

What is RNA-seq and how does it work?

What is RNA-seq?

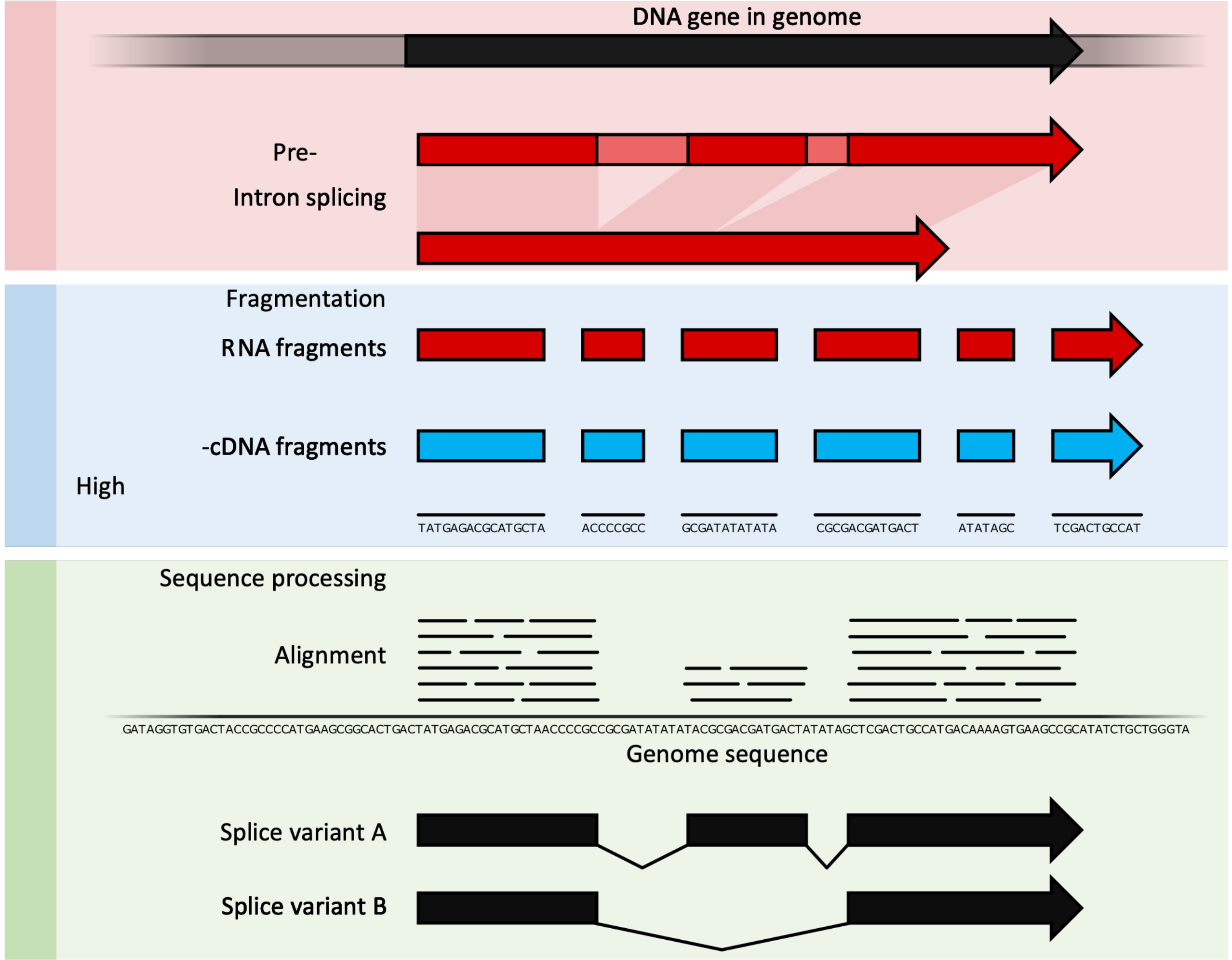
- RNA-seq = RNA sequencing
- Use next generation sequencing to quantify RNA in a cell
- Mainly when people do RNA-seq they’re concerned with mature messenger RNA, because they want to know what genes are expressed



Stark 2019, Image: Wikipedia

How are RNA-seq libraries prepared?

1. Capture RNA
   - Poly-A Capture
   - Ribo-Zero
2. Fragment
3. Reverse transcribe to make cDNA
4. PCR amplification
   - Sequencing adapters
   - PCR bias
5. Illumina sequencing
6. Align to a reference genome
7. Count number of reads that correspond to each gene



Stark 2019, Image: Wikipedia

Designing an RNA-seq Experiment

- Minimum experiment is 3 control samples vs 3 experimental samples
  - Examples:
    1. 3 control samples vs 3 samples treated with a drug
    2. 3 samples in young mouse intestine vs 3 samples in old mouse intestine
    3. 3 samples from germ free (no microbiome) mice vs 3 samples from normal mice
- You need 3 replicates per condition MINIMUM
  - Statistically need at least replicates in order to calculate variance
- Can have multiple conditions
  - Limited by budget
  - Increases the complexity of analysis

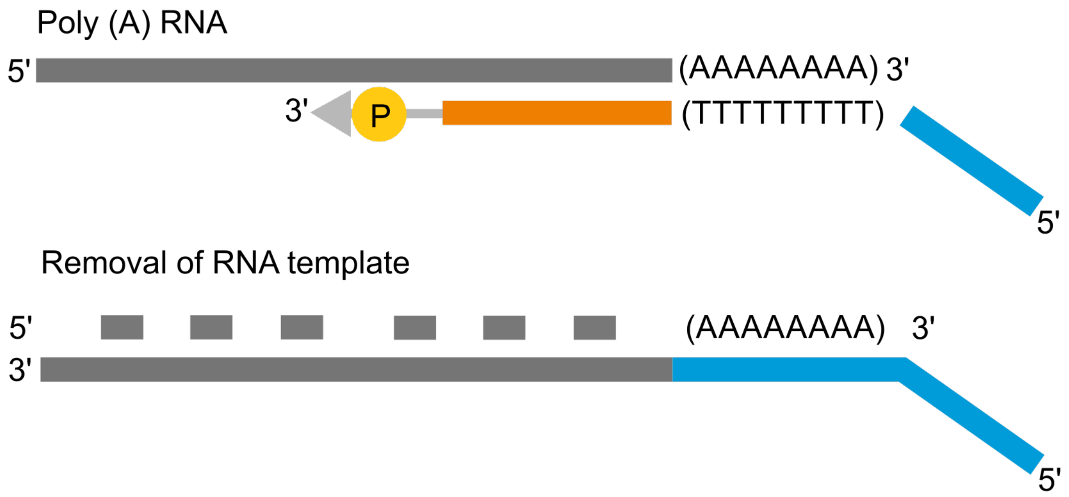
Stark 2019

Designing an RNA-seq Experiment: Isolating RNA

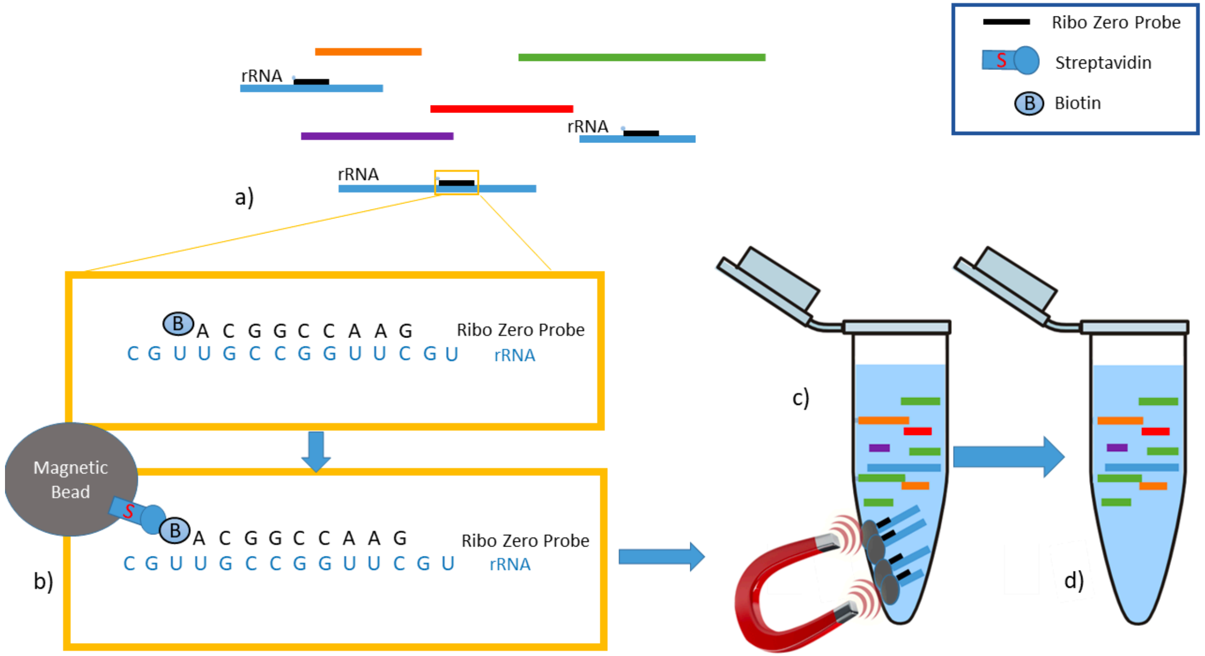
ISSUE: Most RNA (80-90%) is ribosomal RNA that we’re not interested in

- Poly-A Capture
  - Use an oligo-dT primer complimentary to the poly-A tail of mature mRNA
  - Better at capturing only RNAs that correspond to expressed genes
- Ribosomal Depletion
  - Have sequences complimentary to ribosomal RNA on magnetic beads.
  - Sequence more non-coding RNAs, but some experiments want that
- The strategy you choose DOES effect your results

Poly-A Capture



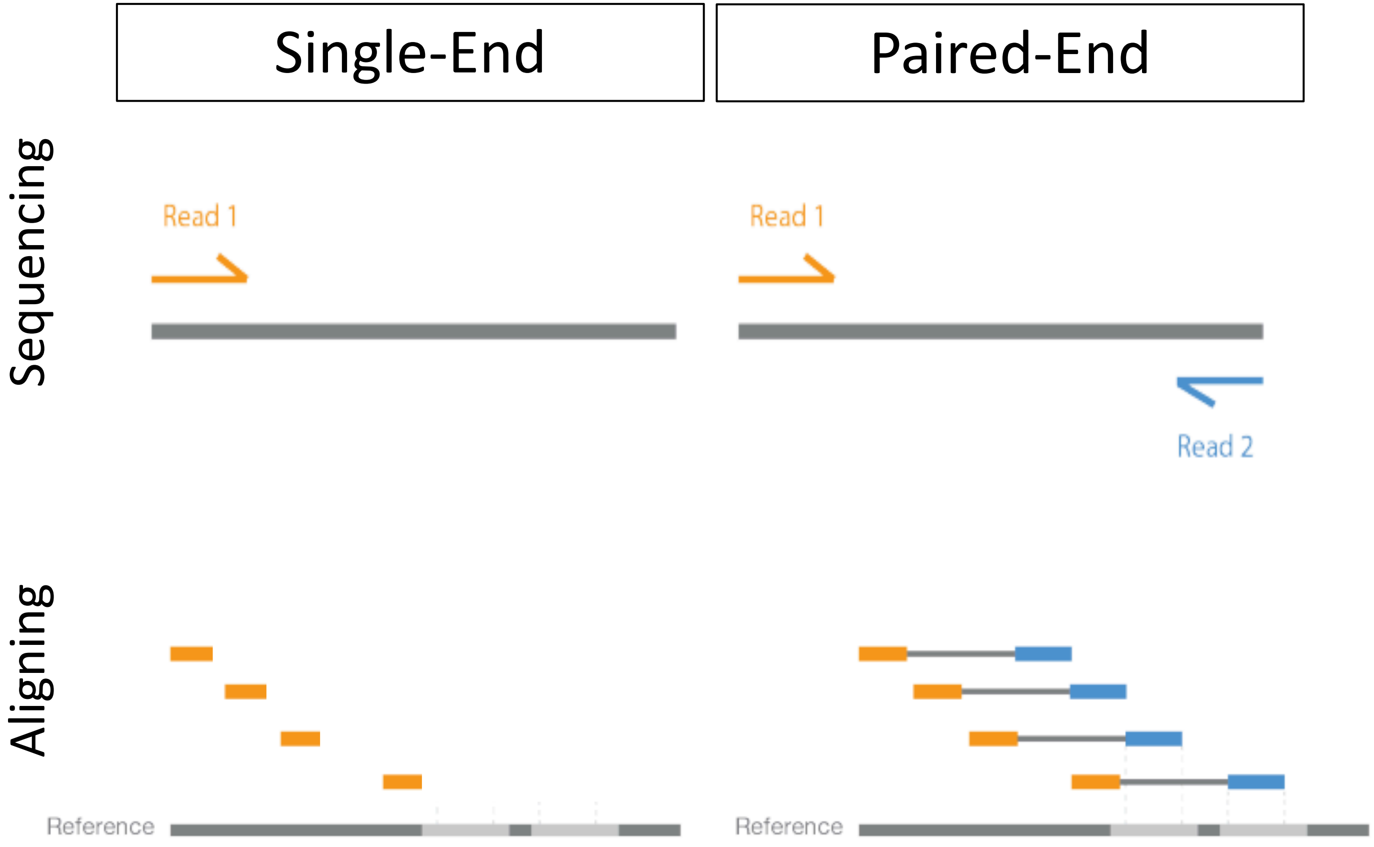
Ribosomal Depletion



Zhao 2018, top image: Thermo Fisher, bottom image: Illumina

Designing an RNA-seq Experiment: Picking the Sequencing Type

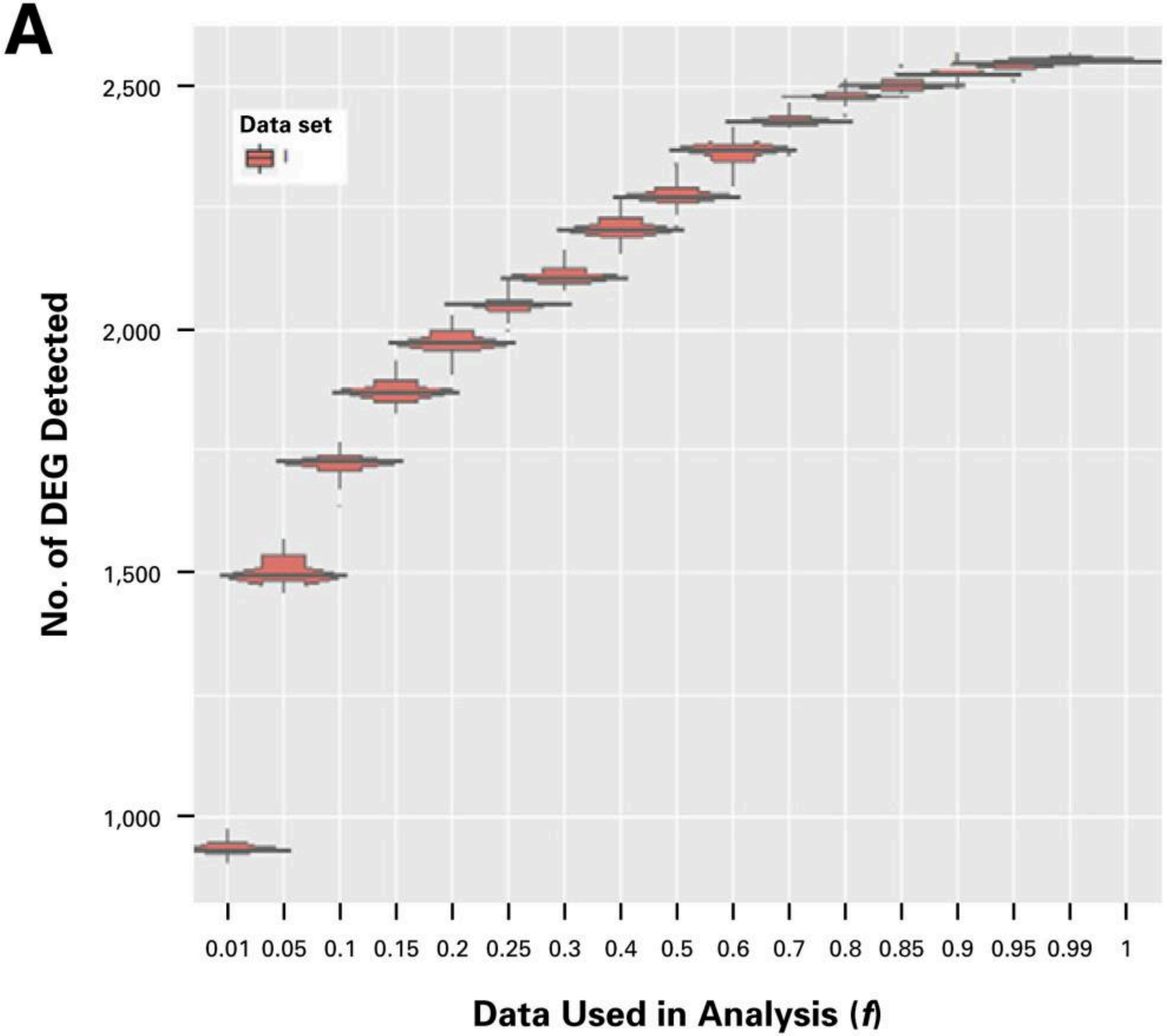
- Read Length
  - Longer the read length the more sequence overlap / coverage
  - More expensive as read length increases
    - Sequencing company charges more
    - Also more like to read into adapters which is a waste of money
- Single vs Paired-End Reads
  - Single-end is cheap and simple
  - Paired-end
    - Get 2 or more times the information from the same DNA
    - Good for repetitive RNA
    - Necessary for splicing



Illumina

Designing an RNA-seq Experiment: Sequencing Depth

- Sequencing depth
    - Standard depth is ~30 million reads per sample
    - If you’re looking for very lowly expressed genes, may need to sequence more, but this should be fine for most experiments
- For plot - 50 million reads total. 1 = all 50 millions reads used



Stupnikov 2018

Analyzing RNA-seq

Processing Sequencing Data

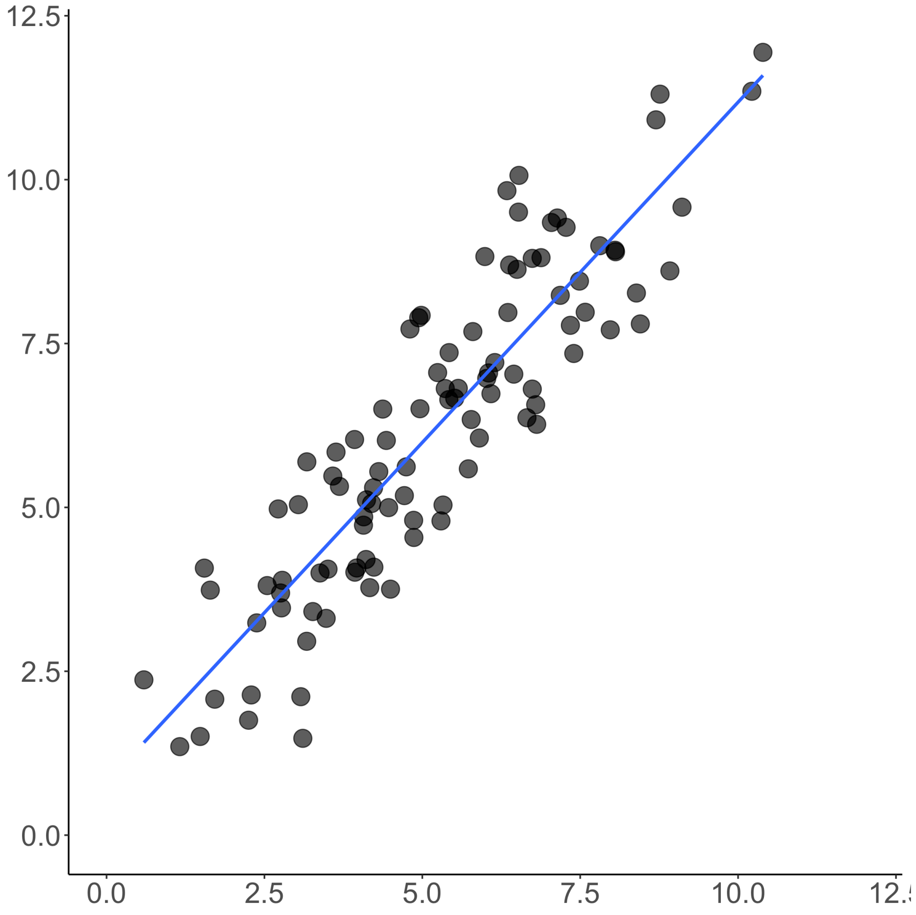
1. Remove unwanted sequences
2. Align to reference genome
3. Count feature of interest
4. Filter and normalize.
   - Remove low counts
   - Remove features with low read depth
   - Compensate for differences in library size

Stark 2019

Differential Gene Expression

Is there is difference in expression of this gene between my conditions?

- Can use a variety of statistical tests, but most common practice is a linear model
- Test association of each gene with phenotype or condition
- Correct for multiple testing



Stark 2019

Pathway Analysis

- Do I see a bunch of genes that are known to work together?
- over-representation analysis (ORA)
  - Do I see more genes from this pathway than I expect by chance?
- functional class scoring methods (FCS)
  - Same thing, but incorporates expression values. Idea is coordinated, but not necessarily larges changes in expression could be important.
- Great free webtool is the Consensus Path Database http://cpdb.molgen.mpg.de/

Pathway	Score	P-value	Q-value
Binding and Uptake of Ligands by Scavenger Receptors	0.88	0.0001	0.001
Beta oxidation of lauroyl-CoA to decanoyl-CoA-CoA	0.72	0.0078	0.0156
mitochondrial fatty acid beta-oxidation of unsaturated fatty acids	0.65	0.0050	0.0125
NOTCH1 Intracellular Domain Regulates Transcription	0.63	0.0034	0.0113
Creatine metabolism	0.45	0.0022	0.0110
Abasic sugar-phosphate removal via the single-nucleotide replacement pathway	0.33	0.0100	0.0167
ARMS-mediated activation	-0.37	0.0200	0.0286
Estrogen-dependent nuclear events downstream of ESR-membrane signaling	-0.55	0.0300	0.0375
Axonal growth inhibition (RHOA activation)	-0.78	0.0500	0.0500
E3 ubiquitin ligases ubiquitinate target proteins	-0.79	0.0500	0.0500

Nyguen 2019

What else can you look at with RNA-seq?

- Transcript level quantification
- Alternative splicing
- Small RNAs like miRNA, etc…
- Transposable elements
- Ribosomal RNA
- Mutations like single nucleotide variants (SNVs), small insertions or deletions (indels), or copy number variations (CNVs)

Stark 2019

PLAY WITH DATA

https://infinityloop.shinyapps.io/TCC-GUI/

Single Cell RNA-seq

What is single cell RNA-seq and why do we need it?

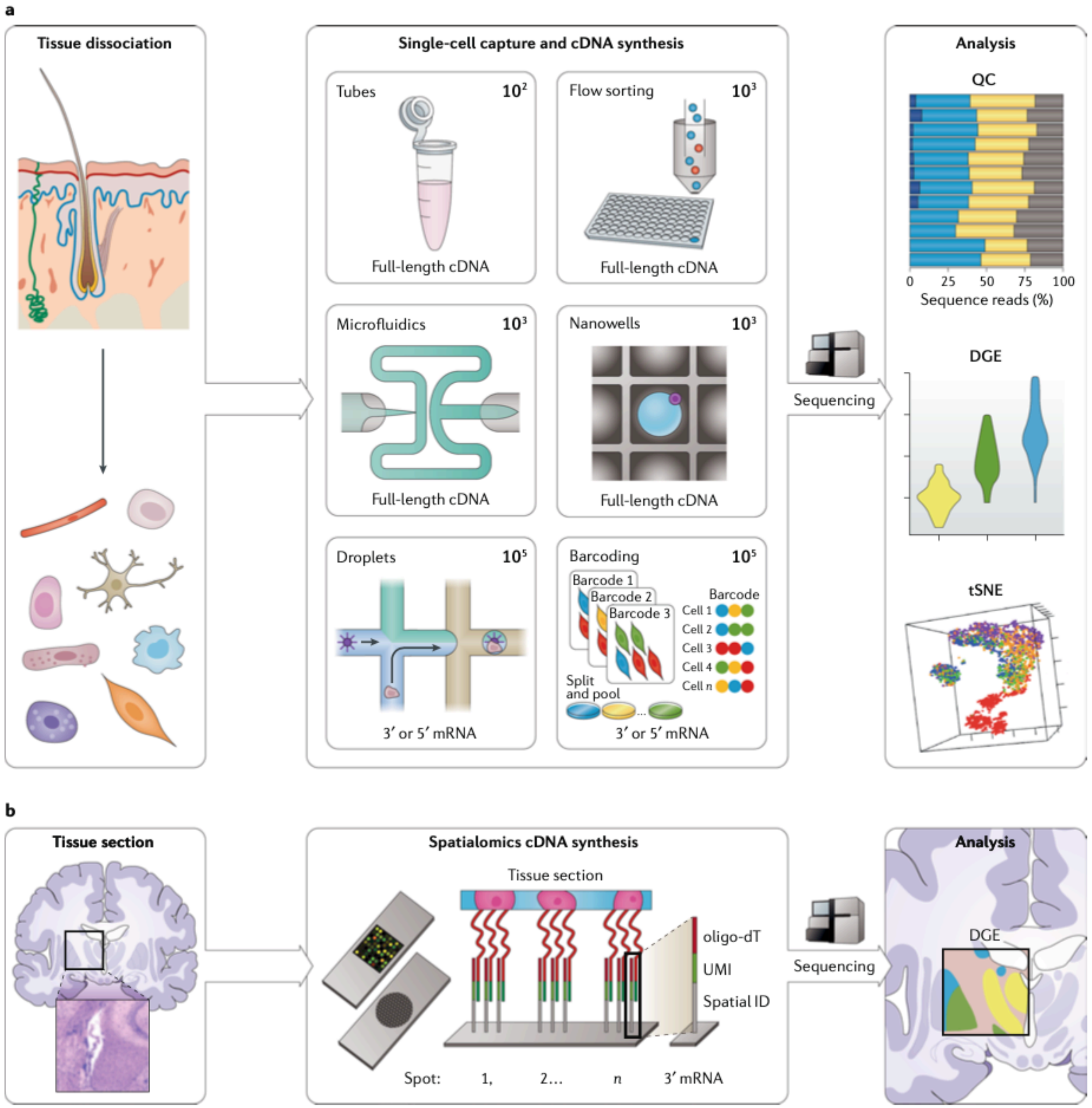
- Issues with bulk RNA-seq that single-cell RNA-seq overcomes
    - Average of multiple cell types
    - Diversity in expression
    - Where in the tissue was it?
- Challenges that still need to be addressed
    - What cell type is this? (cell atlases are working on it)
    - Data sparsity and measurement uncertainty
    - Cost, ~10x more expensive than bulk RNA-seq
- When should I use single cell sequencing? When you want to test something in **multiple cell types**

Stark 2019

Single Cell Analysis

Overall strategy: Isolate single cells and attach a barcode to them.

1. Isolation
   - Flow sorting
   - Microfluidics / droplet encapsulation
2. Barcoding
   - Capture oligos either on a slide or a bead
   - Indexed PCR primers
3. Sequence everything together
4. Computationally sort cells apart when processing data



Stark 2019

References

1. Nguyen, T., Shafi, A., Nguyen, T. *et al.* Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol* **20,** 203 (2019). https://doi.org/10.1186/s13059-019-1790-4
2. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat Rev Genet* **20,** 631–656 (2019). https://doi.org/10.1038/s41576-019-0150-2
3. Stupnikov, A. et al. Impact of Variable RNA-Sequencing Depth on Gene Expression Signatures and Target Compound Robustness: Case Study Examining Brain Tumor (Glioma) Disease Progression. JCO Precision Oncology 1–17 (2018) doi:10.1200/po.18.00014
4. Zhao, S., Zhang, Y., Gamini, R. *et al.* Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci Rep* **8,** 4781 (2018). https://doi.org/10.1038/s41598-018-23226-4

How to can I get started in bioinformatics on my own? (A collection of free online resources)

1. Learn a scripting language
   - R https://r4ds.had.co.nz/
   - Python https://jakevdp.github.io/PythonDataScienceHandbook/
2. Take a free online genomics, biostatistics, bioinformatics classes
   - Biomedical Data Science http://genomicsclass.github.io/book/
   - Modern Statistics for Modern Biology http://web.stanford.edu/class/bios221/book/index.html
   - Bioinformatics Data Skills https://vincebuffalo.com/book/ (book not free sorry!)
3. Go to a MeetUp
   - RLadies Philly https://www.meetup.com/rladies-philly/
   - Philadelphia Python Users Group https://vincebuffalo.com/book/