ILLINOIS
Roy J. Carver Biotechnology Center

Bulk RNA Sequencing Analyses

Jessica Holmes
High Performance Computing in Biology,
Roy J. Carver Biotechnology Center

June 24, 2025

ILLINOIS

Today's Topics

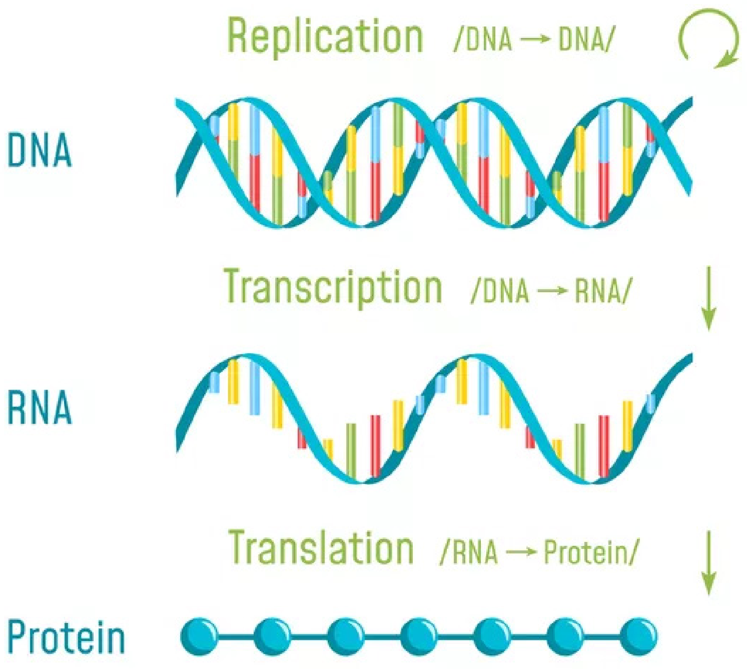
- Transcriptomics
- Traditional RNA-Seq Methods
  - Sequencing & experimental considerations
  - Traditional gene counting
  - Gene quantification
  - Statistics
- Where to find help

Today's Topics

- Transcriptomics
- Traditional RNA-Seq Methods
  - Sequencing & experimental considerations
  - Traditional gene counting
  - Gene quantification
  - Statistics
- Where to find help

Transcriptome

- Includes all transcripts expressed in a sample at a given time point
- Unlike the genome, it is actively changing all the time
- Which transcripts are present depends on:
  - Environment
  - Developmental stage
  - Tissue type
  - And more!



FancyTapis / Getty Images

What can we do with RNA sequences?

Differential Gene Expression

- Quantitative evaluation
- Comparison of transcript levels, usually between different groups
- Vast majority of RNA-Seq is for DGE

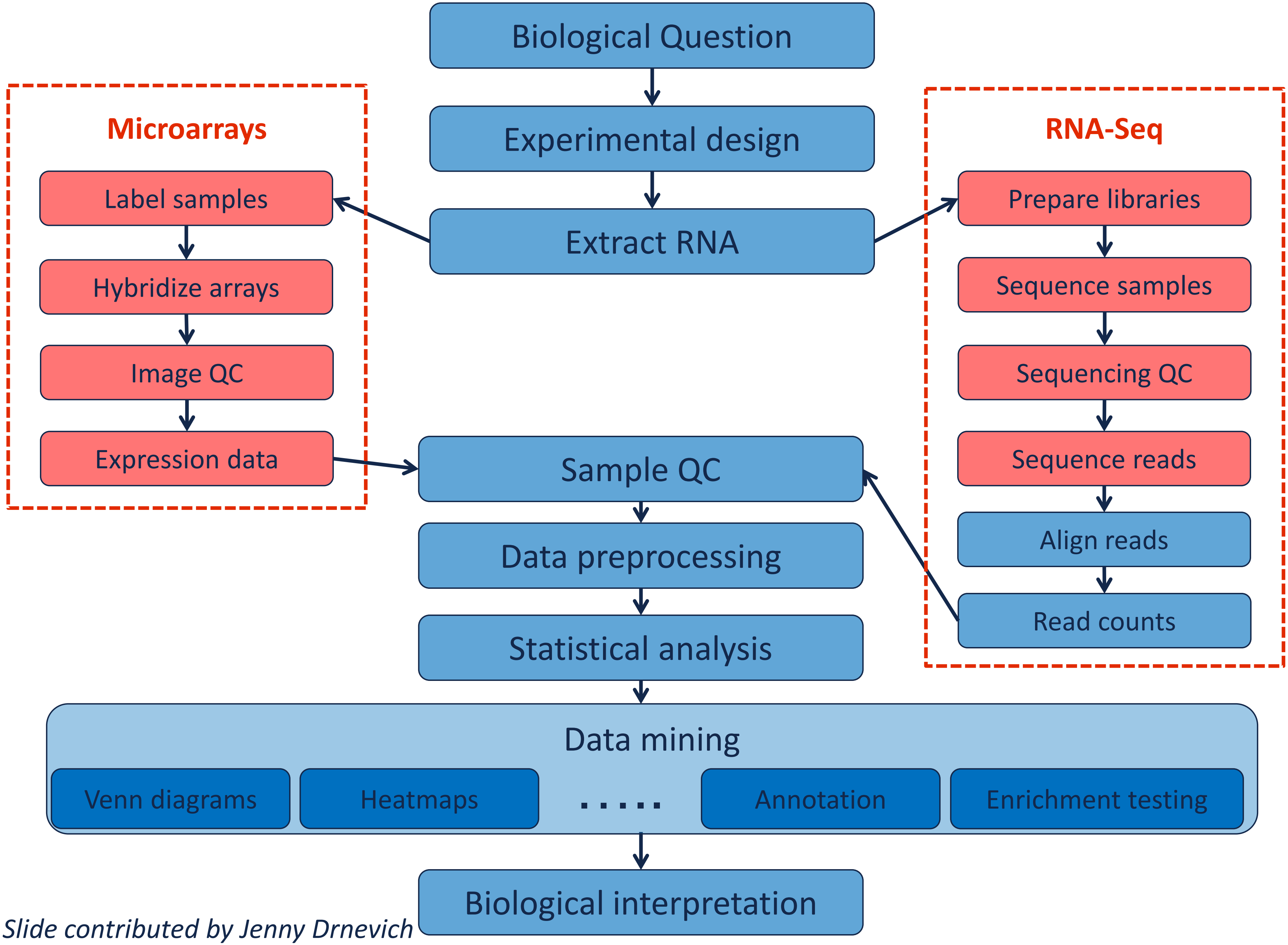
Transcriptome Assembly

- Build new or improved profile of transcribed regions (“gene models”) of the genome
- Can then be used for DGE

Metatranscriptomics

- Transcriptome analysis of a community of different species (e.g., gut bacteria, hot springs, soil)
- Gain insights on the functioning and activity rather than just who is present

Content adapted from Radhika Khetani

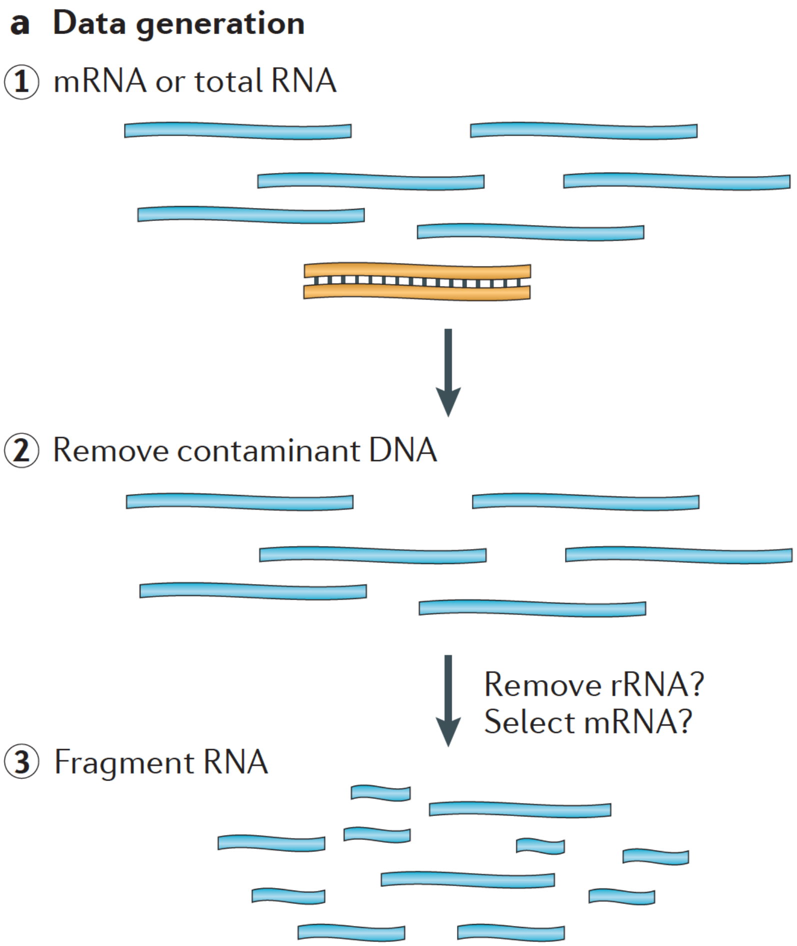


Slide contributed by Jenny Drnevich

Today's Topics

- Transcriptomics
- Traditional RNA-Seq Methods
  - Sequencing & experimental considerations
  - Traditional gene counting
  - Gene quantification
  - Statistics
- Where to find help

From RNA -> sequence data

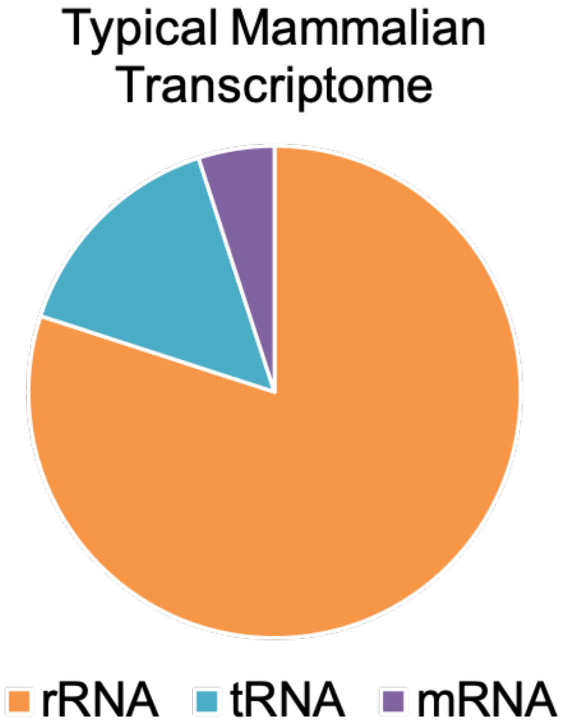


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

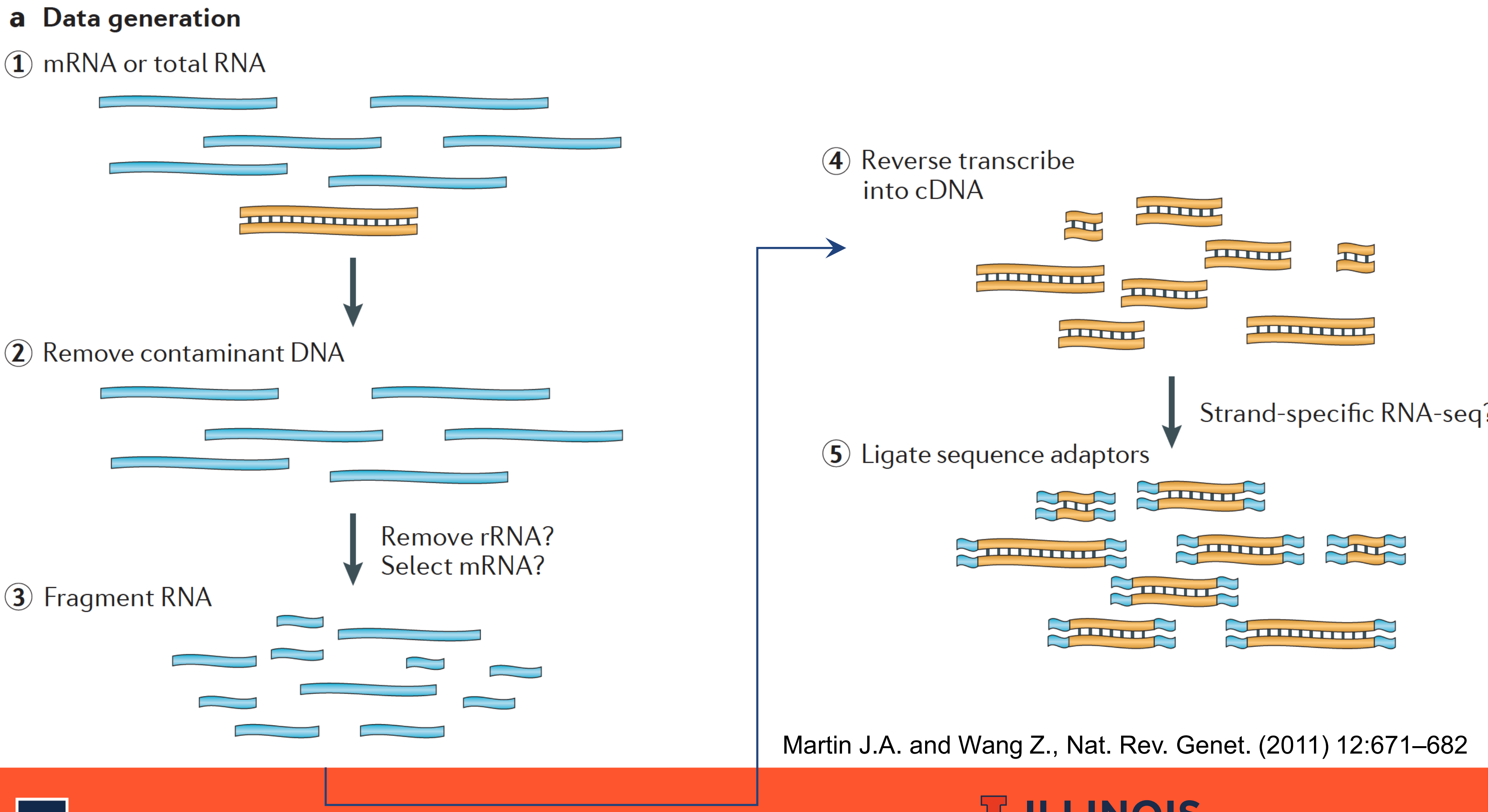
Removal of rRNA is almost always recommended

Removal Methods:

- poly-A selection (eukaryotes only)
- ribosomal depletion
- Size selection



From RNA -> sequence data



a Data generation

1. mRNA or total RNA
2. Remove contaminant DNA

Remove rRNA?
Select mRNA?

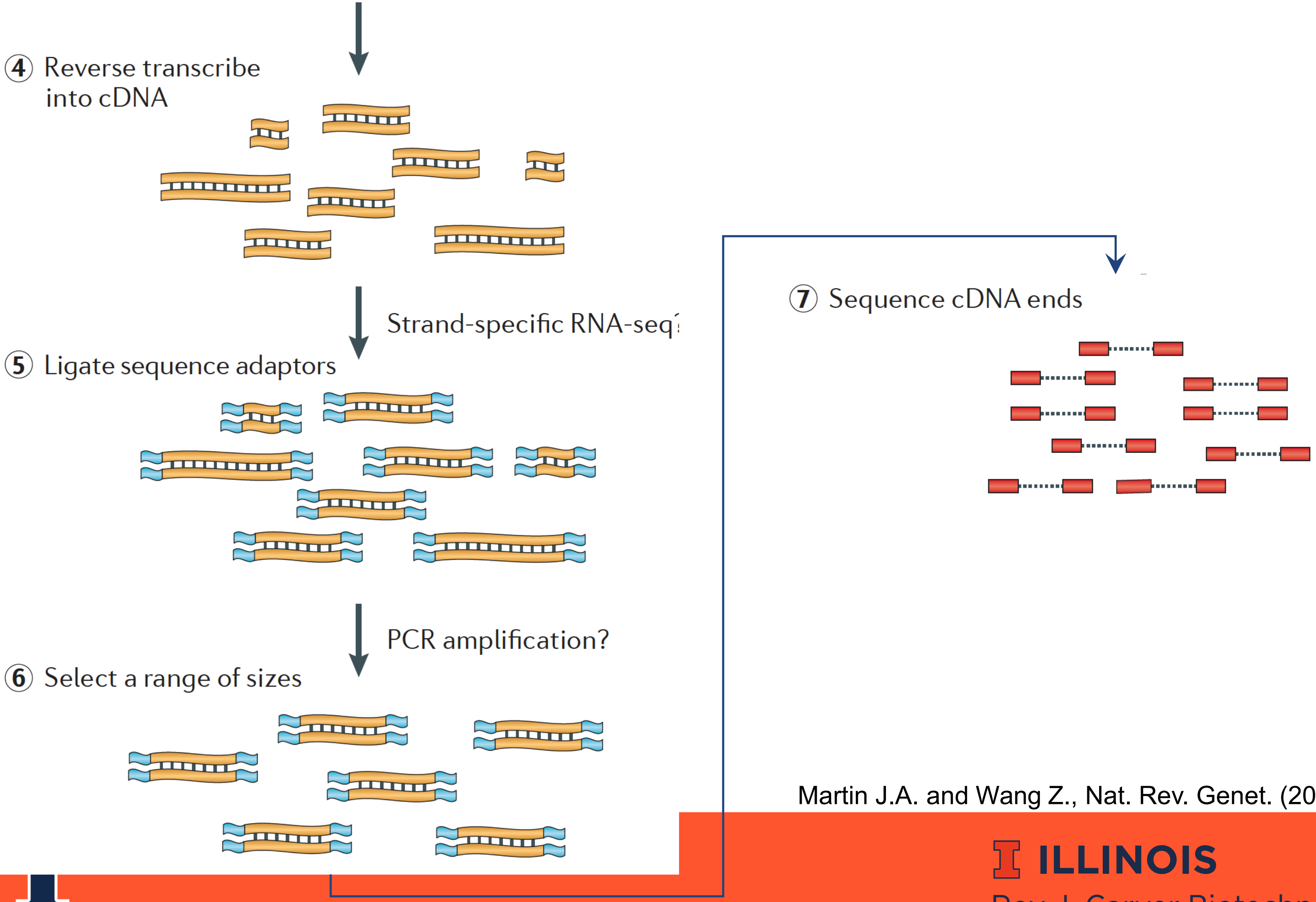
3. Fragment RNA
4. Reverse transcribe into cDNA

Strand-specific RNA-seq?

5. Ligate sequence adaptors

Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

From RNA -> sequence data



4 Reverse transcribe into cDNA

Strand-specific RNA-seq?

5 Ligate sequence adaptors

PCR amplification?

6 Select a range of sizes

7 Sequence cDNA ends

Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

How do we sequence DNA?

1st generation: **Sanger** method (1987)

2nd generation ("next generation"; 2005):

- **454** - pyrosequencing
- **SOLiD** – sequencing by ligation
- **Illumina** – sequencing by synthesis
- **Ion Torrent** – ion semiconductor
- **Pac Bio** – Single Molecule Real-Time (SMRT) sequencing, 1000 bp

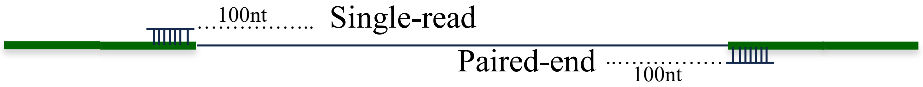
3rd generation (2015)

- **Pac Bio** – SMRT sequencing, but now 20,000+ bp
- **Oxford Nanopore** – ion current detection
- **10X Genomics** – novel library prep for Illumina

Slide contributed by Jenny Drnevich, HPCBio

Illumina – "short read" sequencing

- 300bp reads at lower throughput
- 100-150bp reads at highest throughput
- Many different types of sequencers for various applications.
- Can also "flip" a longer DNA strand and sequence from the other end to get **paired-end reads**



- **Accuracy**: 99.99% **Biases**: yes
- Most common platform for transcriptome sequencing
- New NovaSeq X may be able to perform whole transcriptome sequencing!

Content adapted from Radhika Khetani

Considerations for…
Differential Gene Expression

- Illumina short read still the most cost-effective method
    - Poly-A enrichment vs. rRNA removal?
    - 100 single end (SE) good enough for animal species with good genome/gene references
    - 2 X 150 paired end (PE) better for complex genomes (plants!) or to also measure splice variants
- Keep biological replicates separate
- Quantitative long reads just starting to be possible...

Considerations for…
Transcriptome Assembly

- Long reads PacBio give full-length sequences - strongly recommended!
- Short reads will need to be assembled to full-length transcripts; do 2 X 150 PE
- Collect RNA from many various sources for a robust transcriptome
- Poly-A enrichment is optional depending on your focus

Considerations for… Metatranscriptomics

- Long reads PacBio give full-length sequences - strongly recommended!
- Short reads will need to be assembled to full-length transcripts; do 2 X 150 PE
- Keep biological replicates separate
- Remove ribosomal RNA (rRNA) - bacteria not poly-A'd
- May need to remove host mRNA computationally downstream
  - e.g. removing human mRNA from gut samples

Experimental Design Issues

(or Why you need to think about how you will analyze the data ***before*** you do the experiment)

- Poorly designed experiments (especially with confounding factors) can lead to lower power to detect differences, ambiguous results, or even a waste of time and money!
- What to consider:
  - How many factors do you have?
  - How many levels per factor?
  - How many independent replicates should you do? (3 minimum, 5 is better, and put 5 more in the -70 if you can)
- The more complex the experiment, the more difficult the statistical analysis will be.

Slide courtesy of Jenny Drnevich, HPCBio

How many independent biological replicates (N)?

- A power analysis is recommended: https://pubmed.ncbi.nlm.nih.gov/36830591/
- Realistically, the most-used formula is:

$$N = \frac{(\$\text{ you have})}{(\$ / \text{ measurement})}$$

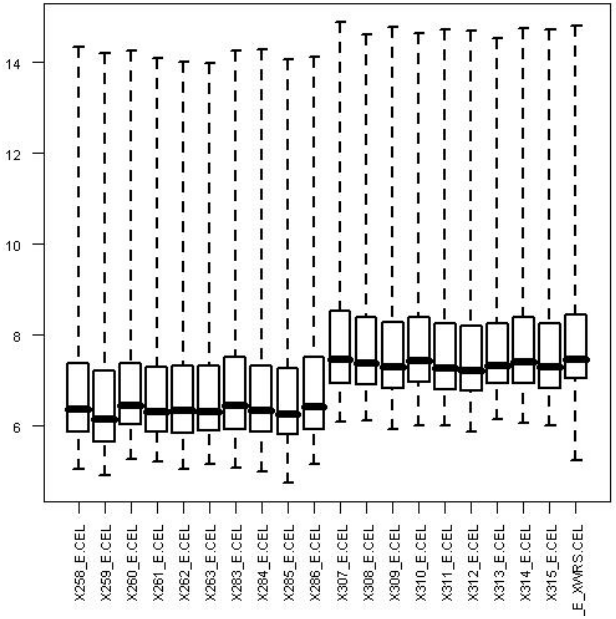
Inspiration and graphic from Jeff Leek's Statistics for Genomic Data Science course on Coursera.org

https://docs.google.com/presentation/d/1tOuTVvnIpNm_QaEpaFBvD04z2y06sFqFWBqwO6GfJes/edit#slide=id.gc69a1ad99_0_46

Edited from original slide, courtesy of Jenny Drnevich (HPCBio)

Beware of confounding factors! (aka batch effects)

- In good experimental design, you compare two groups that **only differ in one factor**.
- Batch effect can occur when subsets of the replicates are handled separately at any stage of the process; handling group becomes in effect another factor. **Avoid processing all or most of one factor level together** if you can't do all the samples at once.

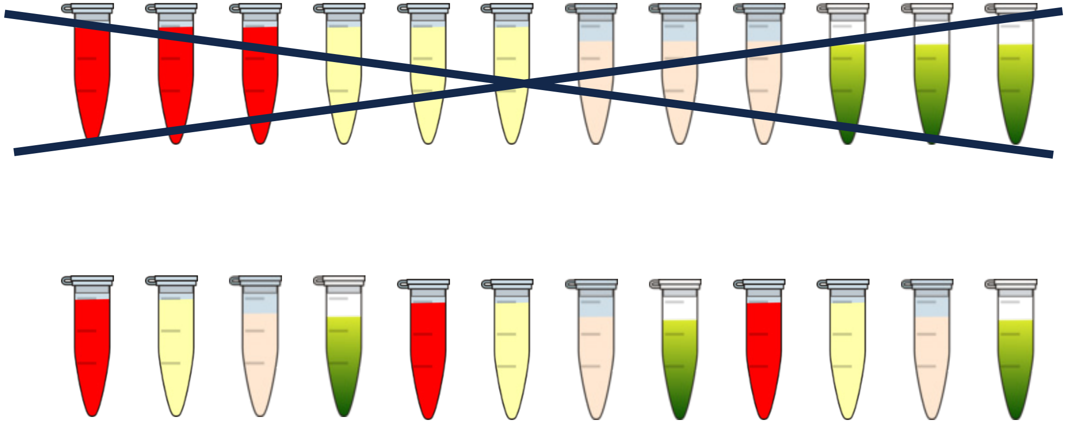


If batch effects are spread evenly over factor levels, they can be accounted for statistically

Slide courtesy of Jenny Drnevich, HPCBio

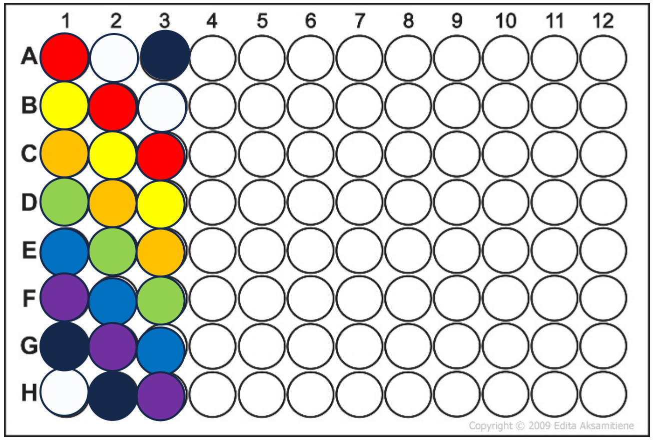
Beware of systematic biases!

- Avoid systematic biases in the arrangement of replicates
  - **Don’t** do all of one factor level first (circadian rhythms, experimenter experience, time-on-ice effects)
  - **Don’t** send samples to a sequencing center in order



http://www.clker.com/clipart-eppendorf-tube-closed.html

Have one rep in each row and each column!



http://www.cellsignet.com/media/templ.html

Slide courtesy of Jenny Drnevich, HPCBio

A word on technical replication…

Technical replication is seen by many statisticians as a waste of time and resources because they do not substantially increase your power to detect differences... **biological replicates do**!

If you cannot increase the number of biological replicates but want to get extra certainty for the samples you do have, then you could do technical replicates if you have the $$ to spend.

Slide courtesy of Jenny Drnevich, HPCBio

Today’s Topics

- Transcriptomics
- Traditional RNA-Seq Methods
  - Sequencing & experimental considerations
  - Traditional gene counting
  - Gene quantification
  - Statistics
- Where to find help

ILLINOIS
Roy J. Carver Biotechnology Center

Traditional Gene Counting Steps

1. Download data

2. Quality control steps

2. Align reads to a reference genome with splice aware software (unless bacterial)

3. Use a gene counting software to obtain the number of read counts per known gene.

1. Download sequence data

It depends on the center, but common methods include:

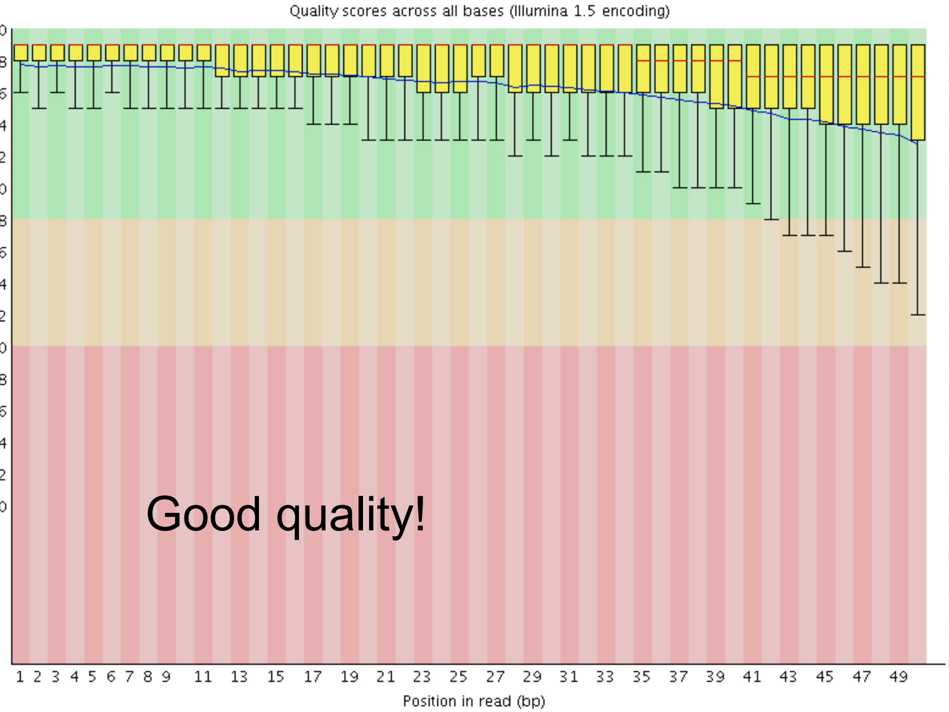
1. Globus which allows you to transfer from one endpoint to another using their webpage
2. Download data to a computer and upload to destination using an SFTP client
   - ✧ Cyberduck, WinSCP…
3. Utilize linux commands such as to perform the same steps
   - ✧ scp, rsync, wget, curl …

2. So how can we check the quality of our raw sequences?

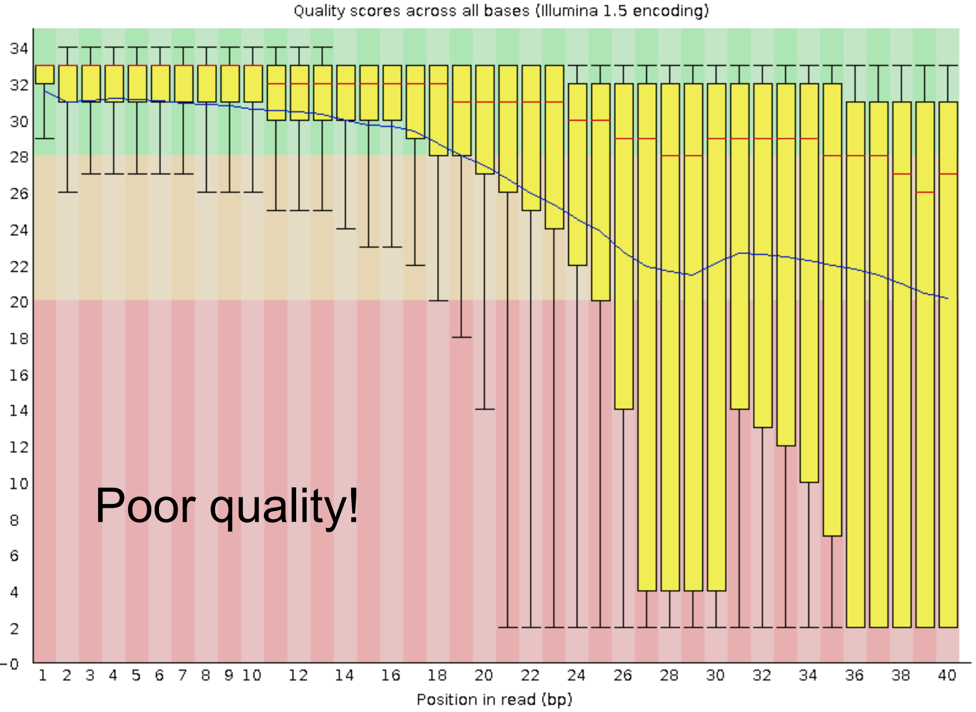
Software called **FASTQC**

- Name is a play on FASTQ format and QC (Quality Control)
- Checks quality by several metrics, and creates a visual report

FASTQC: Quality Scores



Good quality!



Poor quality!

FASTQC cont…

Additional metrics

- Presence of, and abundance of contaminating sequences
- Average read length
- GC content
- And more!

Assumes that your data is:

- WGS (i.e. evenish sampling of the whole genome)
- Derived from DNA
- Derived from one species

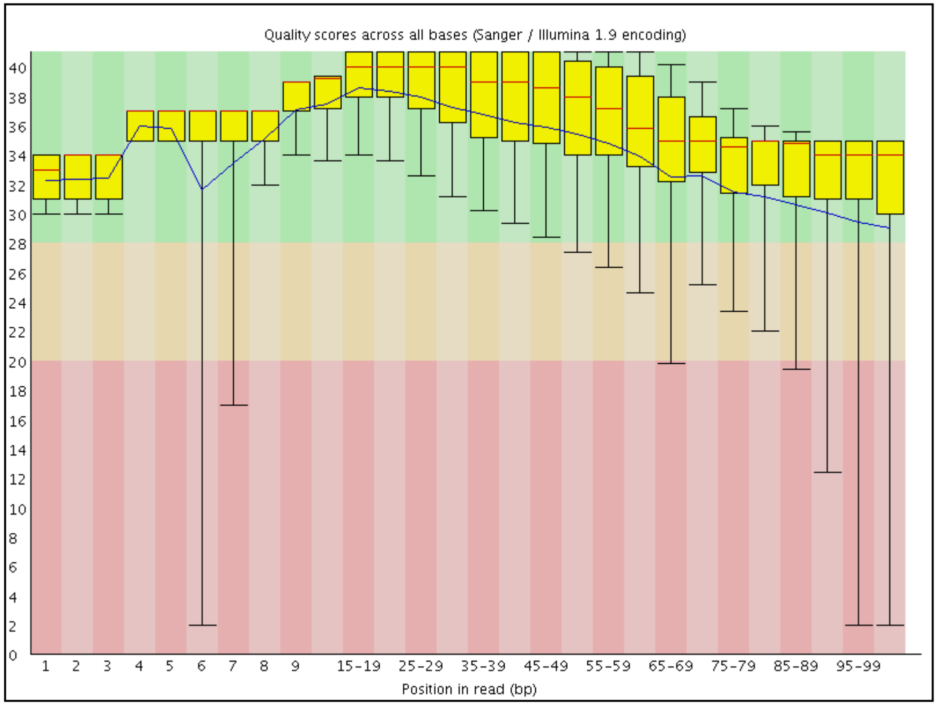
So keep this in mind when interpreting results

2. What do I do when FastQC calls my data poor?

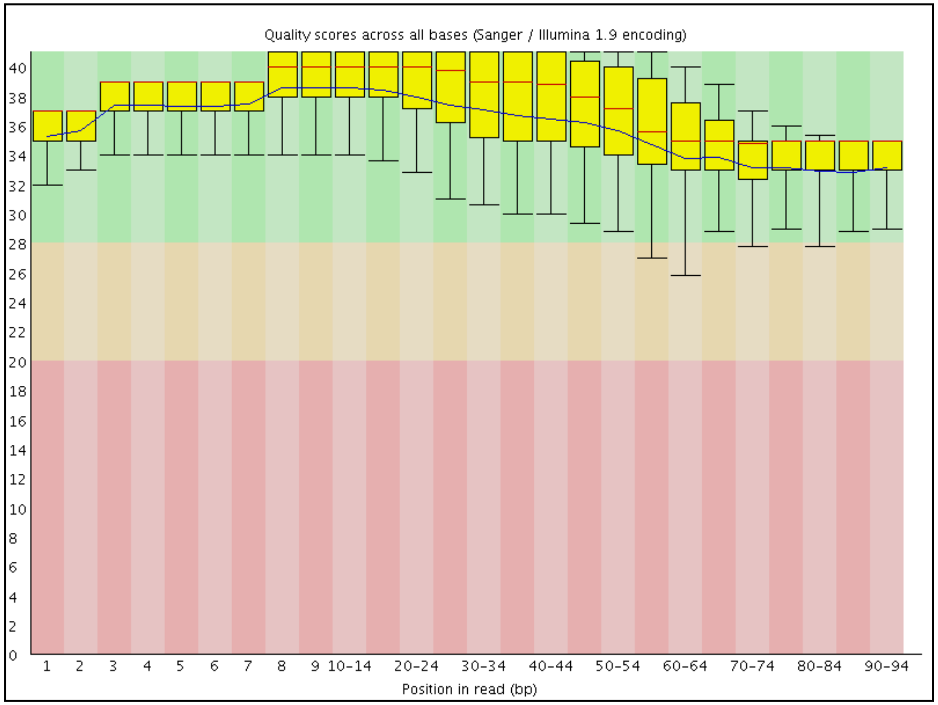
- Poor quality at the ends can be remedied
- Left-over adapter sequences in the reads can be removed
  - Always trim adapters as a matter of routine
- We need to amend these issues so we get the best possible alignment
- After trimming, it is best to rerun the data through FastQC to check the resulting data

Quality Before & After

Before quality trimming



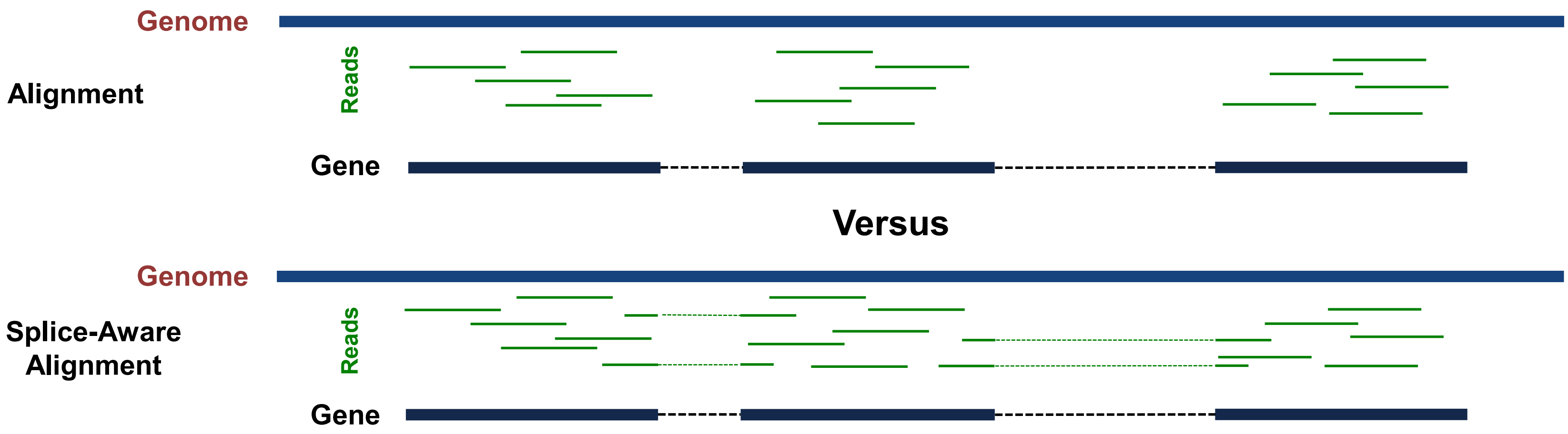
After quality trimming



3. Traditional Gene Counting: Sequence Alignment

We need to align the sequence data to our genome of interest

- If aligning RNASeq data to the genome, almost always pick a splice-aware aligner



3. Traditional Gene Counting: Sequence Alignment

Software choices:

- Splice-aware aligners: recommended for most applications
  - STAR, HiSat2, Novoalign (not free), MapSplice2, GSNAP, …
- Non-splice aware aligners: ideal for bacterial genomes
  - BWA, Novoalign (not free), Bowtie2, HiSat2

Software inputs:

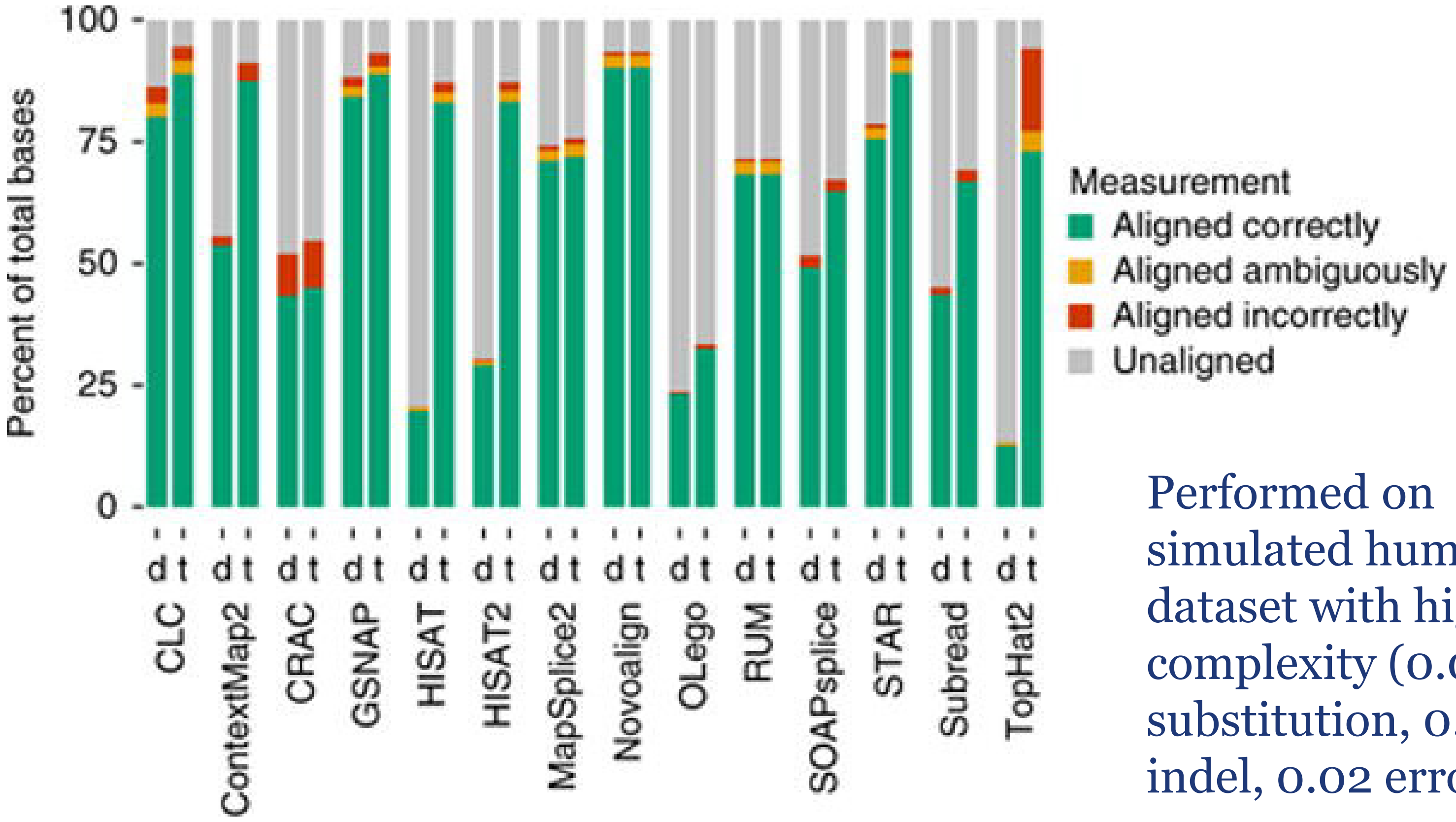
1. Trimmed sequences in FASTQ format
2. Complete reference genome FASTA (or transcriptome)
3. Reference annotation file in GTF or GFF3 format (not required for non-splice aware aligners)

3. Traditional Gene Counting: Sequence Alignment

Other considerations when performing alignment:

- How does it deal with reads that map to **multiple locations**? Will that be compatible with downstream software?
- How does it deal with **paired-end versus single-end** data? Are there extra parameters that need to be added?
- How many **mismatches** will it allow between the genome and the reads?
- What **assumptions** does it make about my genome, and can I change these assumptions, if needed?

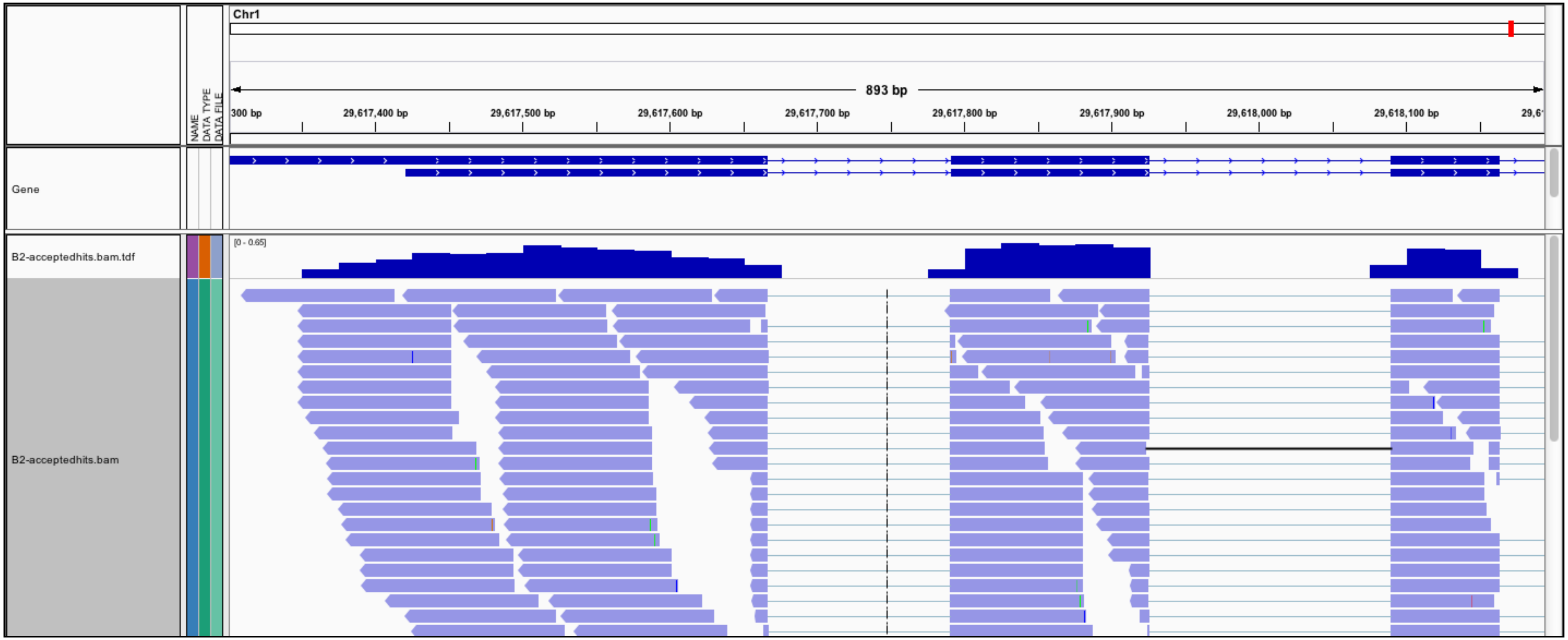
Always check the default settings of any software you use!!!



Performed on simulated human dataset with high complexity (0.03 substitution, 0.005 indel, 0.02 error)

Baruzzo et. al, 2017, doi: 10.1038/nmeth.4106

Optional: Alignment Visualization



IGV is the visualization tool used for this snapshot

3. Traditional Gene Counting: # of sequences in genes

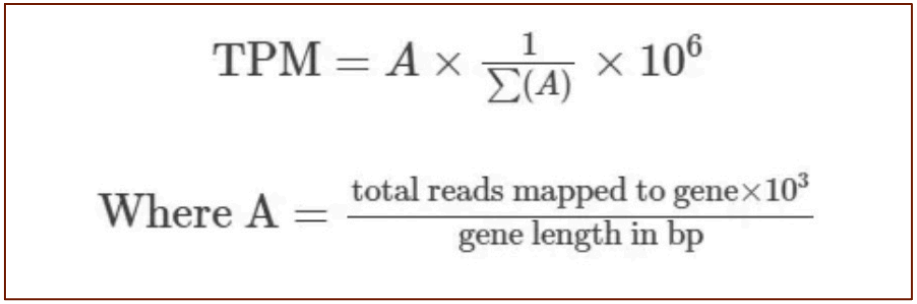
When selecting software consider whether you want to obtain:

- raw read counts or normalized read counts

Gene counting software:

- Software inputs: alignment file (e.g. SAM, BAM or CRAM files) and annotation file (e.g. GTF, GFF3)
- feature-counts & htseq return raw read counts
    - Required for R packages like DESeq, limma & EdgeR
- StringTie returns FPKM or TPM normalized counts for each gene
    - Required for R package Ballgown
- RSEM returns TPM normalized counts

TPM: Transcripts Per Million



$$\mathrm{TPM} = A \times \frac{1}{\sum(A)} \times 10^6$$

$$\text{Where A} = \frac{\text{total reads mapped to gene} \times 10^3}{\text{gene length in bp}}$$

https://www.reneshbedre.com/blog/expression_units.html

- TPM is a normalized read count that corrects for gene/transcript length bias

- What goes into calculating this?
  - Number of reads per sample (i.e. library size)
  - Transcript length adjusted for biases that affect the number of reads (i.e. effective length)

BREAK

Take break now, if needed