ILLINOIS
Roy J. Carver Biotechnology Center

RNA Sequencing Analyses

Jessica Holmes
Research & Instructional Specialist,
High Performance Computing in Biology

June 21, 2023

ILLINOIS

Today's Topics

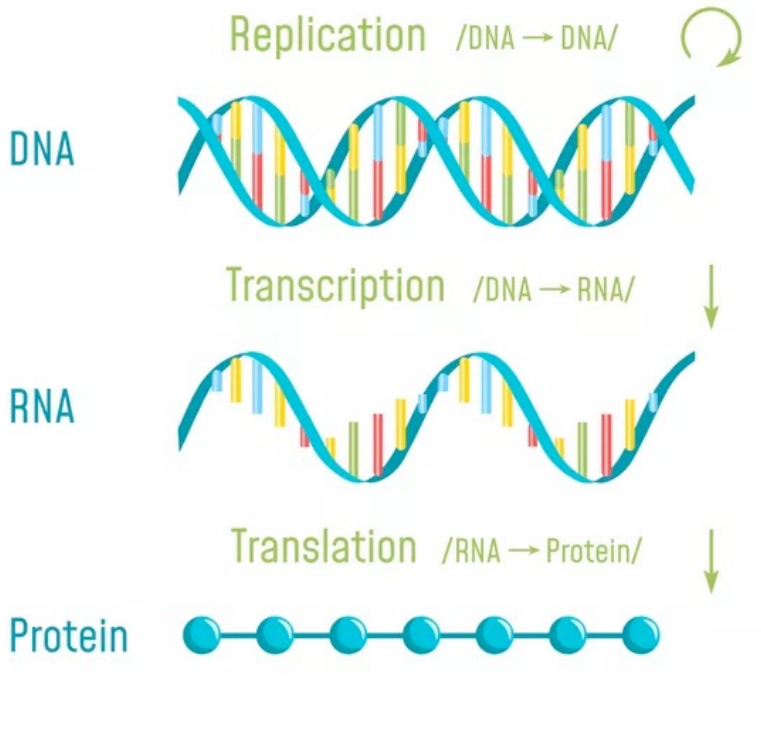
- Transcriptomics
- Traditional RNA-Seq Methods
  - Sequencing & experimental considerations
  - Gene counting
  - Statistics
- Single Cell RNA-Seq Methods
  - Sequencing
  - UMI counting & statistics
- Where to find help

Today's Topics

- Transcriptomics
- Traditional RNA-Seq Methods
  - Sequencing & experimental considerations
  - Gene counting
  - Statistics
- Single Cell RNA-Seq Methods
  - Sequencing
  - UMI counting & 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

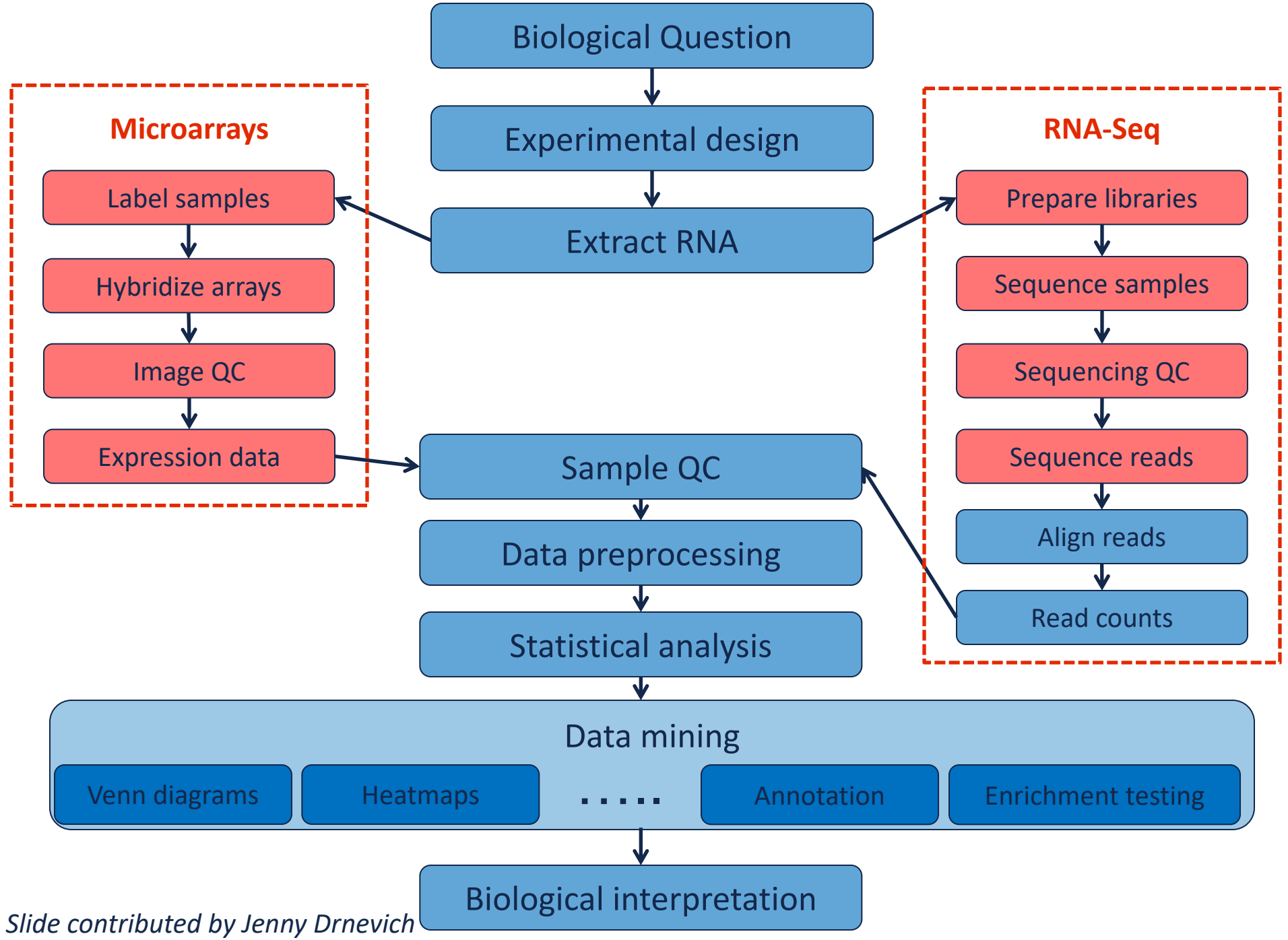
- Quantitative evaluation
- Comparison of transcript levels, usually between different groups
- Vast majority of RNA-Seq is for DGE

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

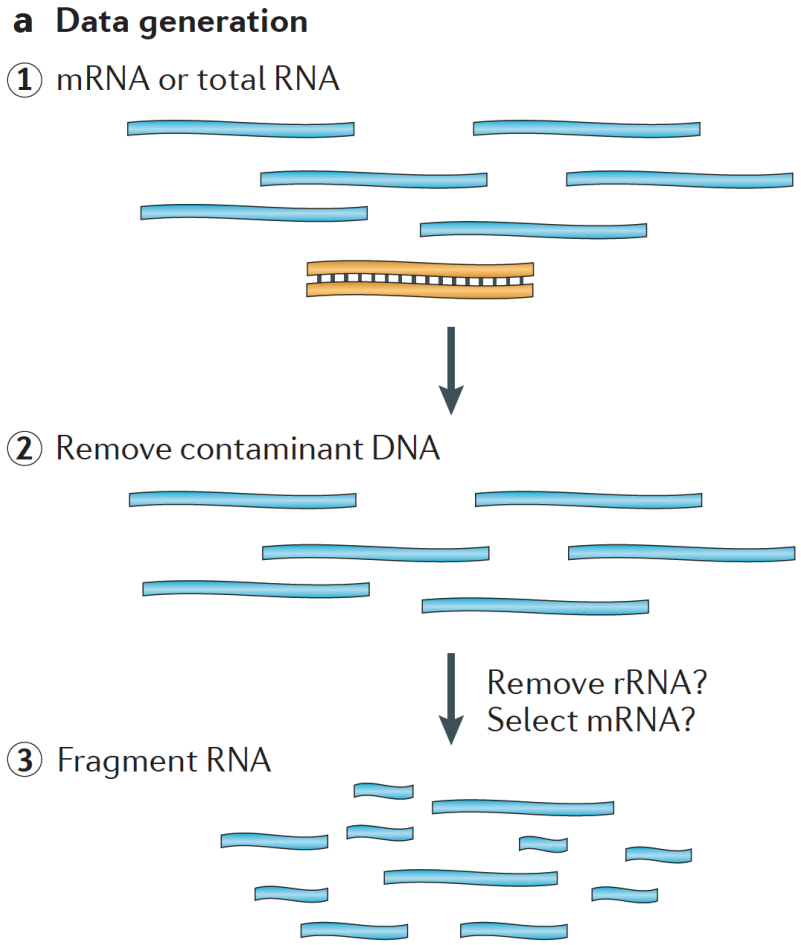


Slide contributed by Jenny Drnevich

Today's Topics

- Transcriptomics
- Traditional RNA-Seq Methods
  - Sequencing & experimental considerations
  - Gene counting
  - Statistics
- Single Cell RNA-Seq Methods
  - Sequencing
  - UMI counting & statistics
- Where to find help

From RNA -> sequence data

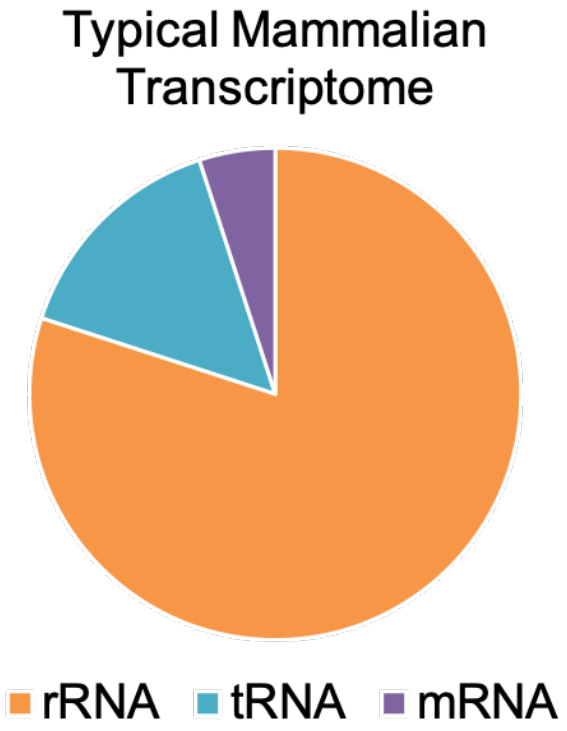


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

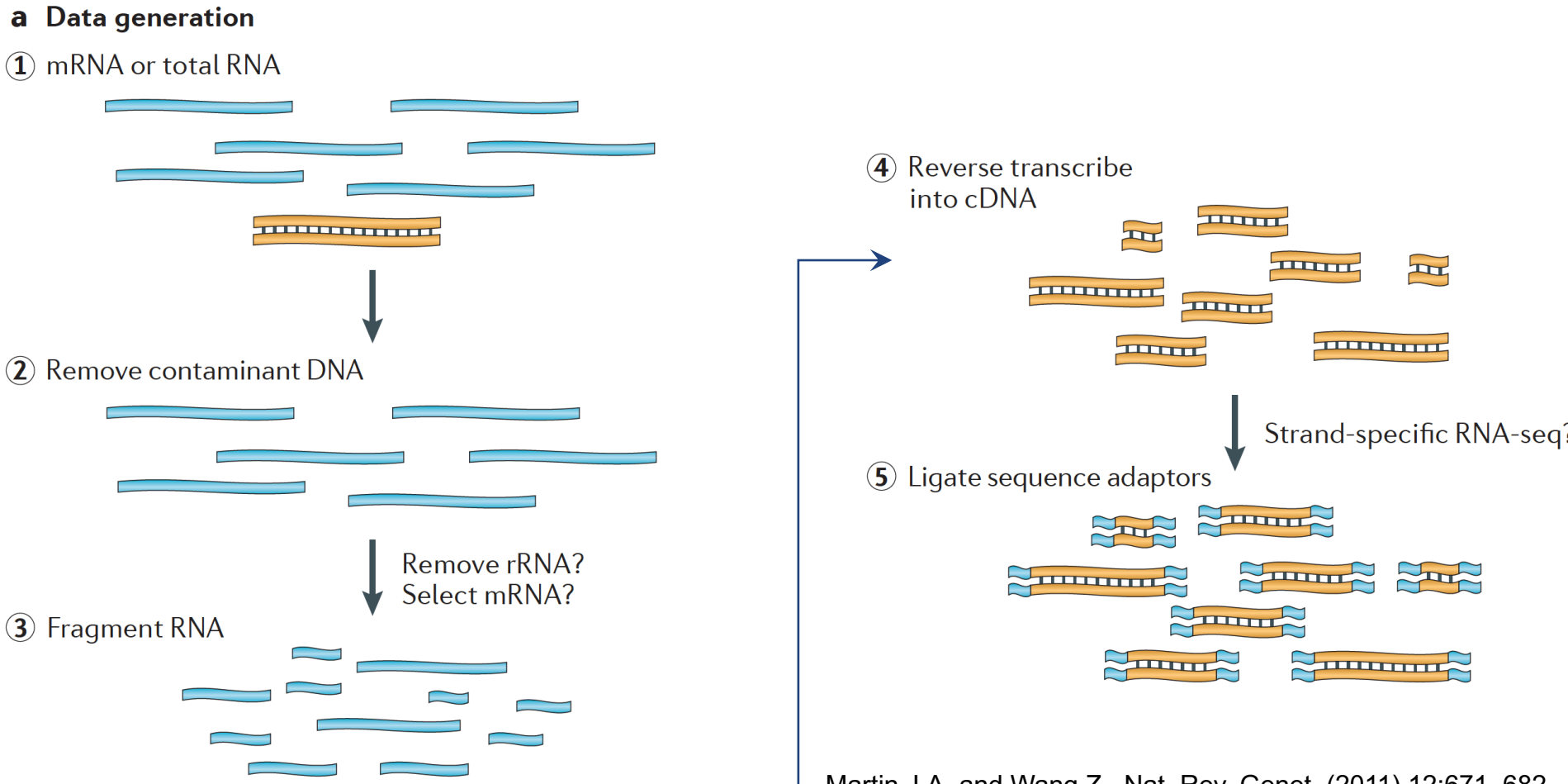
Removal of rRNA is almost always recommended

Removal Methods:

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

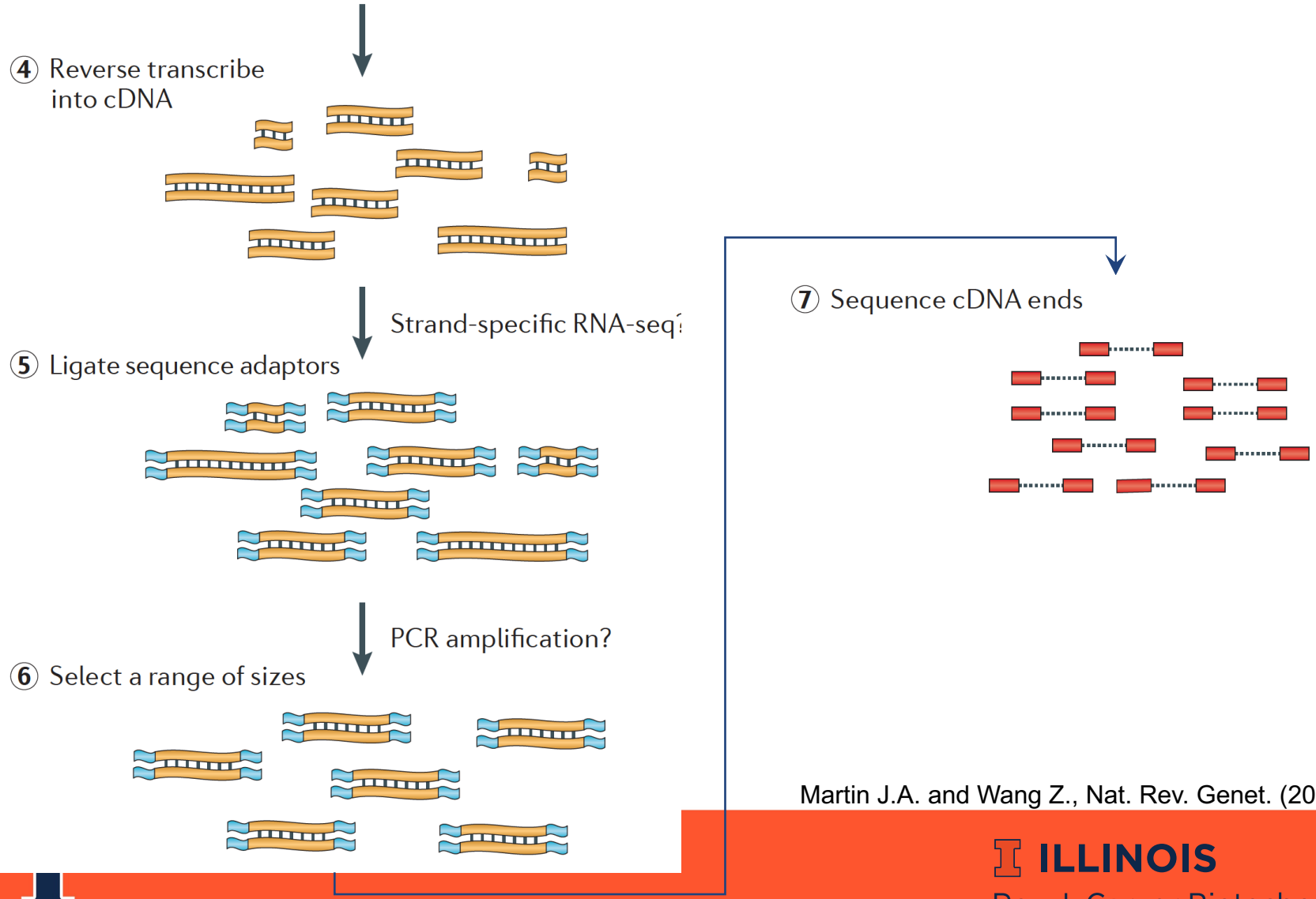


From RNA -> sequence data



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

From RNA -> sequence data



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):

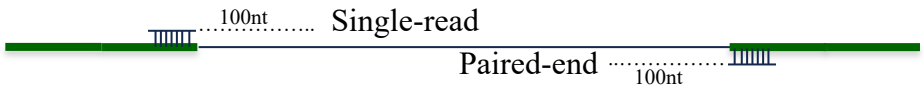
- **454** - pyrosequencing
- **SOLiD** – sequencing by ligation
- **Illumina** – sequencing by synthesis
- **Ion Torrent** – ion semiconductor
- **Pac Bio** – Single Molecule Real-Time sequencing, 1000 bp

3rd generation (2015)

- **Pac Bio** – SMRT, Sequel system, 20,000 bp
- **Nanopore** – ion current detection
- **10X Genomics** – novel library prep for Illumina

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 NovaSeqX may be able to perform whole transcriptome sequencing!

Considerations for…
Differential Gene Expression

- Keep biological replicates separate
- Poly-A enrichment is generally recommended
  - Unless you're interested in non-coding RNA!
- Remove ribosomal RNA (rRNA)
- Usually single-end (SE) is enough
  - Paired-end (PE) may be recommended for more complex genomes

Single-end read

Read1

ATGTTCCATAAGC…

Paired-end reads

Read1

ATGTTCCATAAGC…

Read2

CCGTAATGGCATG…

Considerations for…
Transcriptome Assembly

- Collect RNA from many various sources for a robust transcriptome
  - These can be pooled before or after sequencing (but before assembly)
- Poly-A enrichment is optional depending on your focus
- Remove ribosomal RNA (rRNA)
- Paired-end (PE) is recommended. The more sequence, the better.
  - Even better if you use long-read technology in addition

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.

How many independent biological replicates (N)?

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

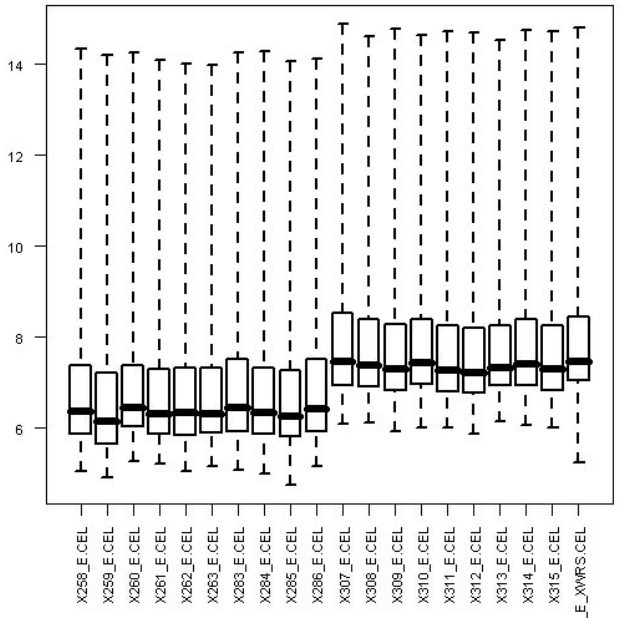
$$N = \frac{(\$ \text{ you have})}{(\$ \text{ / measurement})}$$

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

Beware of confounding factors! (aka batch effects)

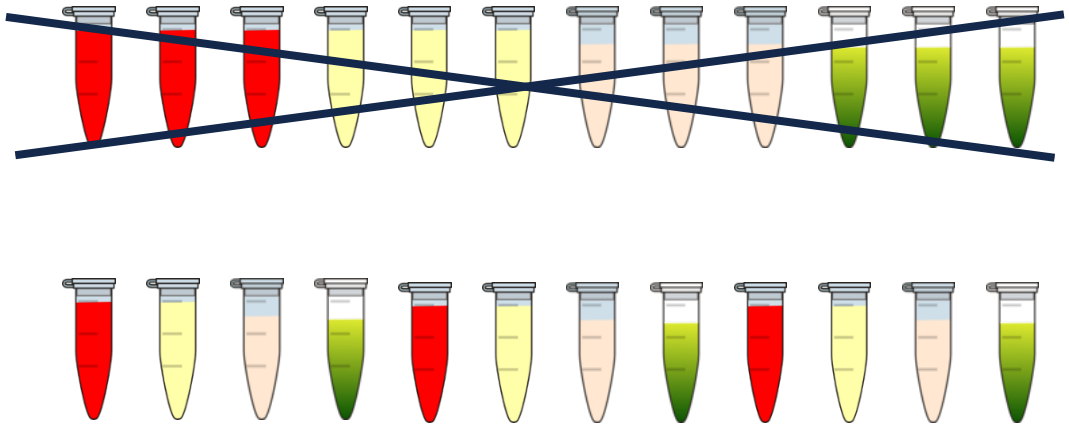
- In good experimental design, you compare two groups that **only differ in one factor**.
- Batch effect can occur when subsets of the replicates are handled separately at any stage of the process; handling group becomes in effect another factor. **Avoid processing all or most of one factor level together** 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

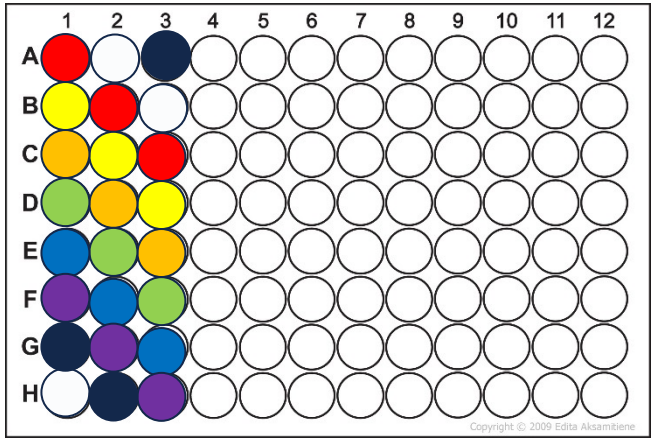
Beware of systematic biases!

- 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

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.

Today's Topics

- Transcriptomics
- Traditional RNA-Seq Methods
  - Sequencing & experimental considerations
  - Gene counting
  - Statistics
- Single Cell RNA-Seq Methods
  - Sequencing
  - UMI counting & statistics
- Where to find help

Gene Counting Steps

1. Download data

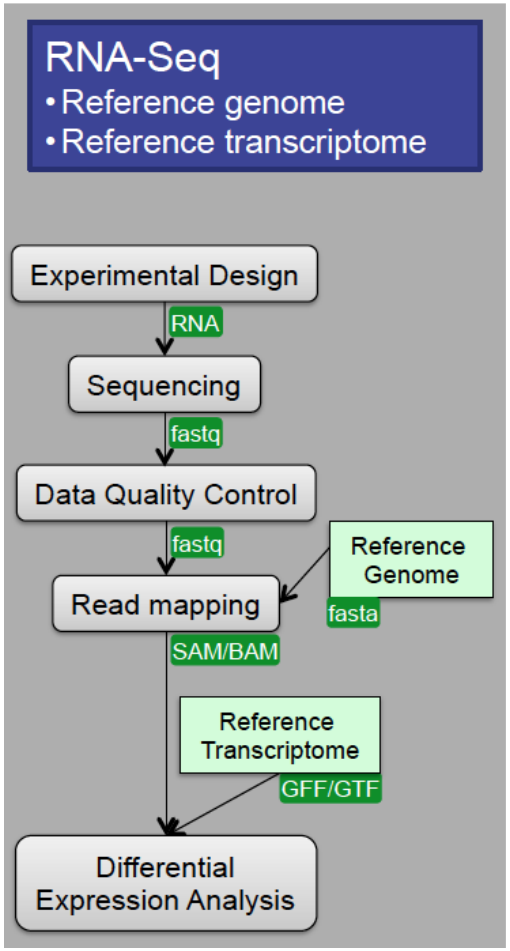
2. Quality control steps

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

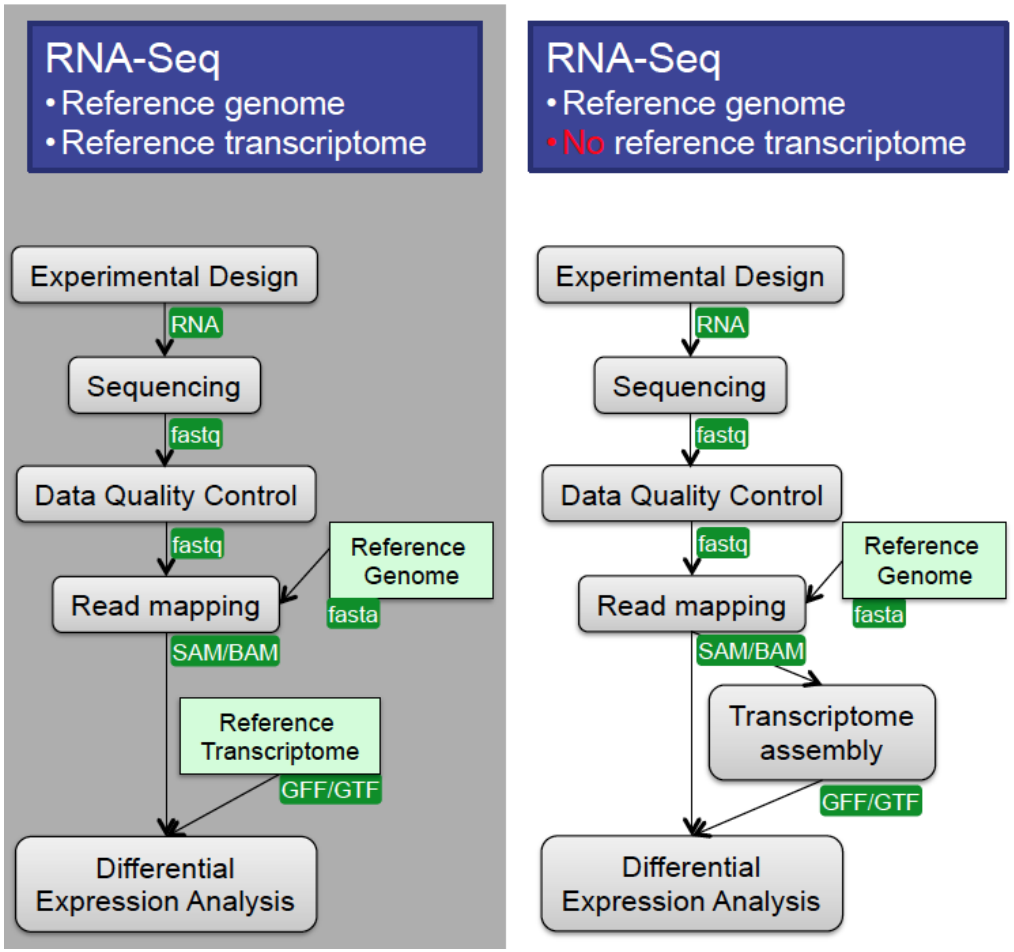
3a. Use a gene counting software to obtain the number of read counts per known gene.

3b. Alternatively, use a transcript counting software like Salmon

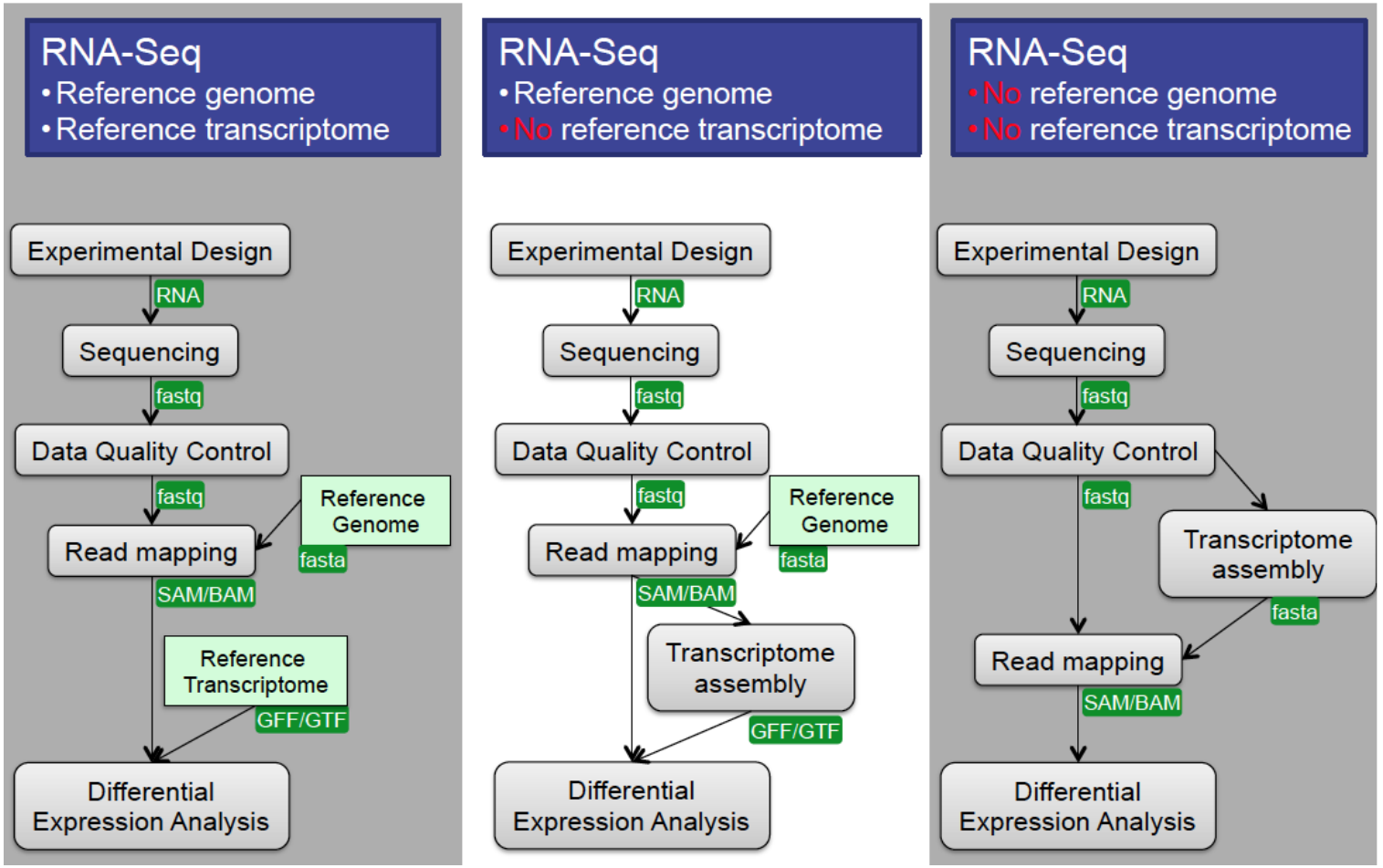
Differential Gene Expression Workflows



Differential Gene Expression Workflows



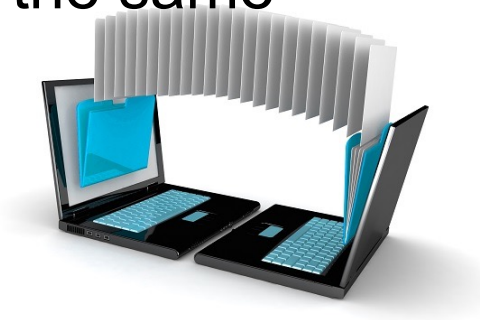
Differential Gene Expression Workflows



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 …

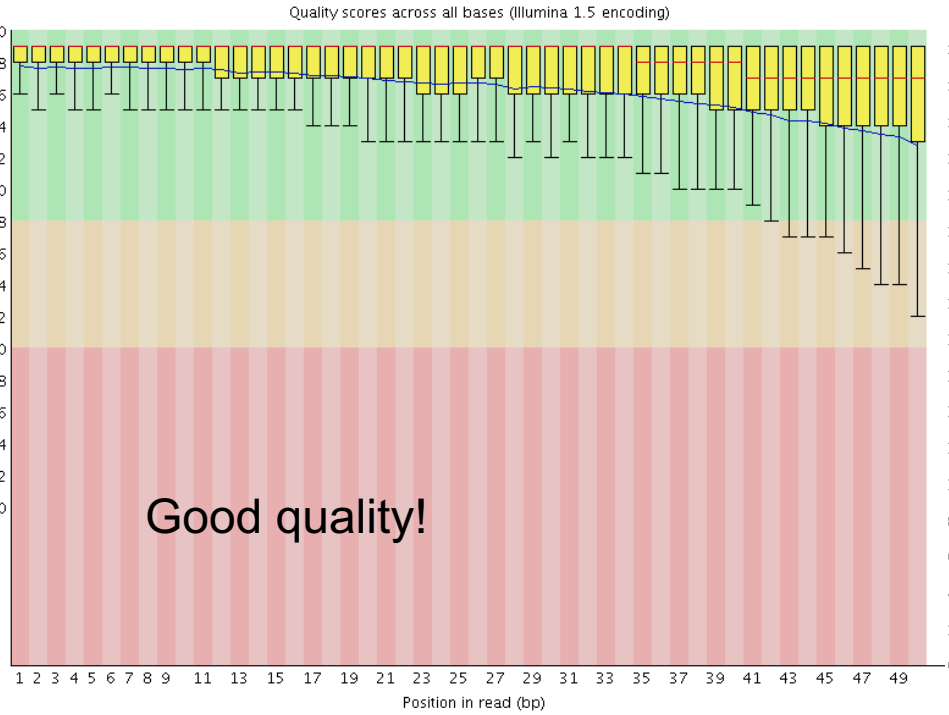


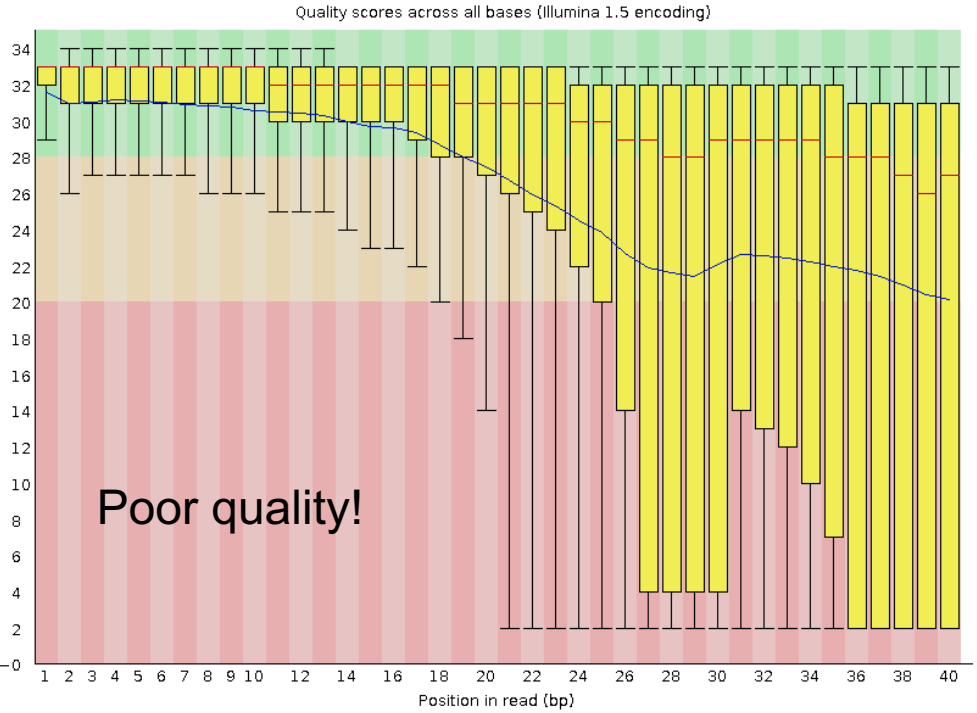
2. So how can we check the quality of our raw sequences?

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





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

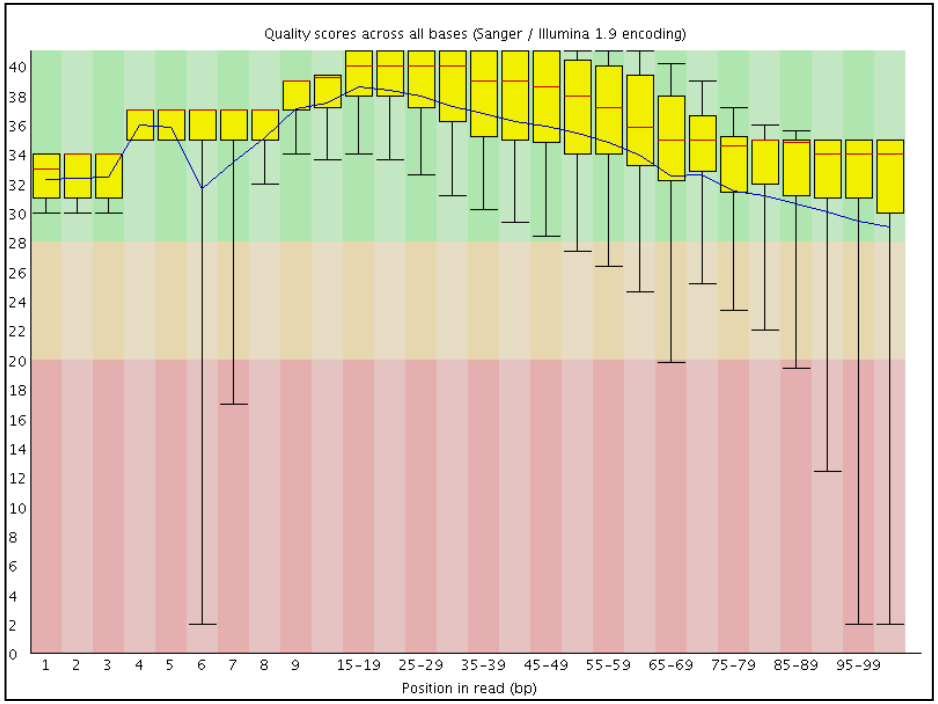
So keep this in mind when interpreting results

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

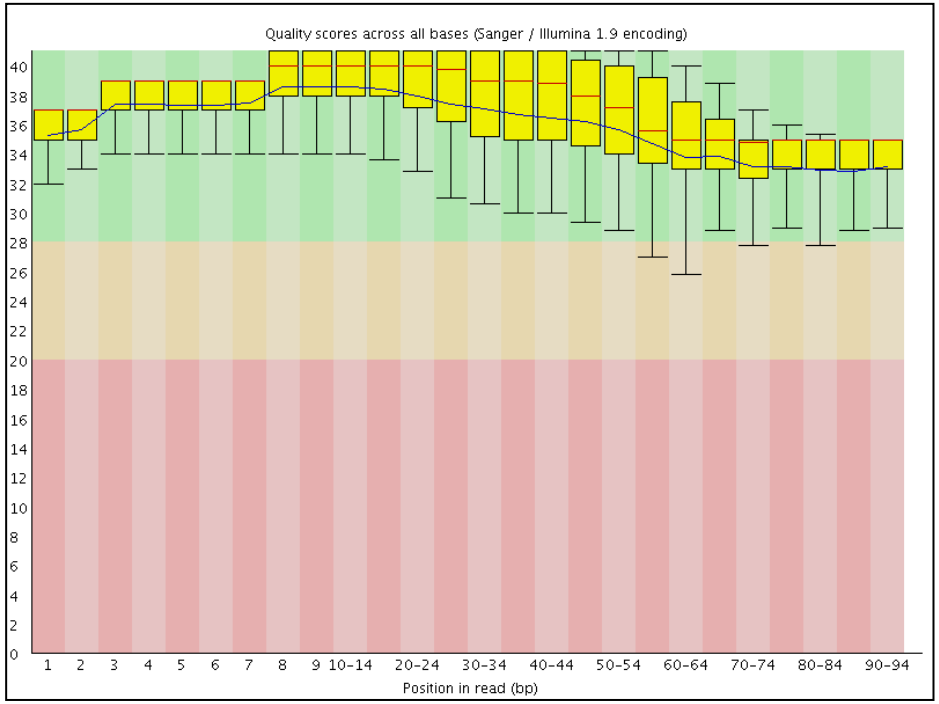
- Poor quality at the ends can be remedied
- Left-over adapter sequences in the reads can be removed
  - Always trim adapters as a matter of routine
- We need to amend these issues so we get the best possible alignment
- After trimming, it is best to rerun the data through FastQC to check the resulting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

Quality Before & After

Before quality trimming



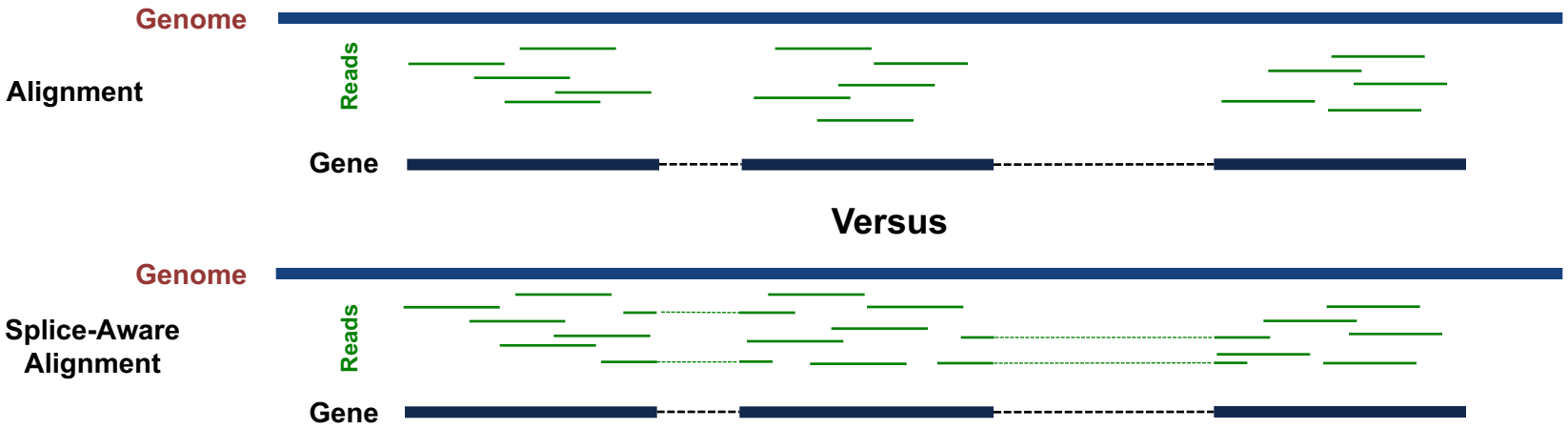
After quality trimming



3a. 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



3a. 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:

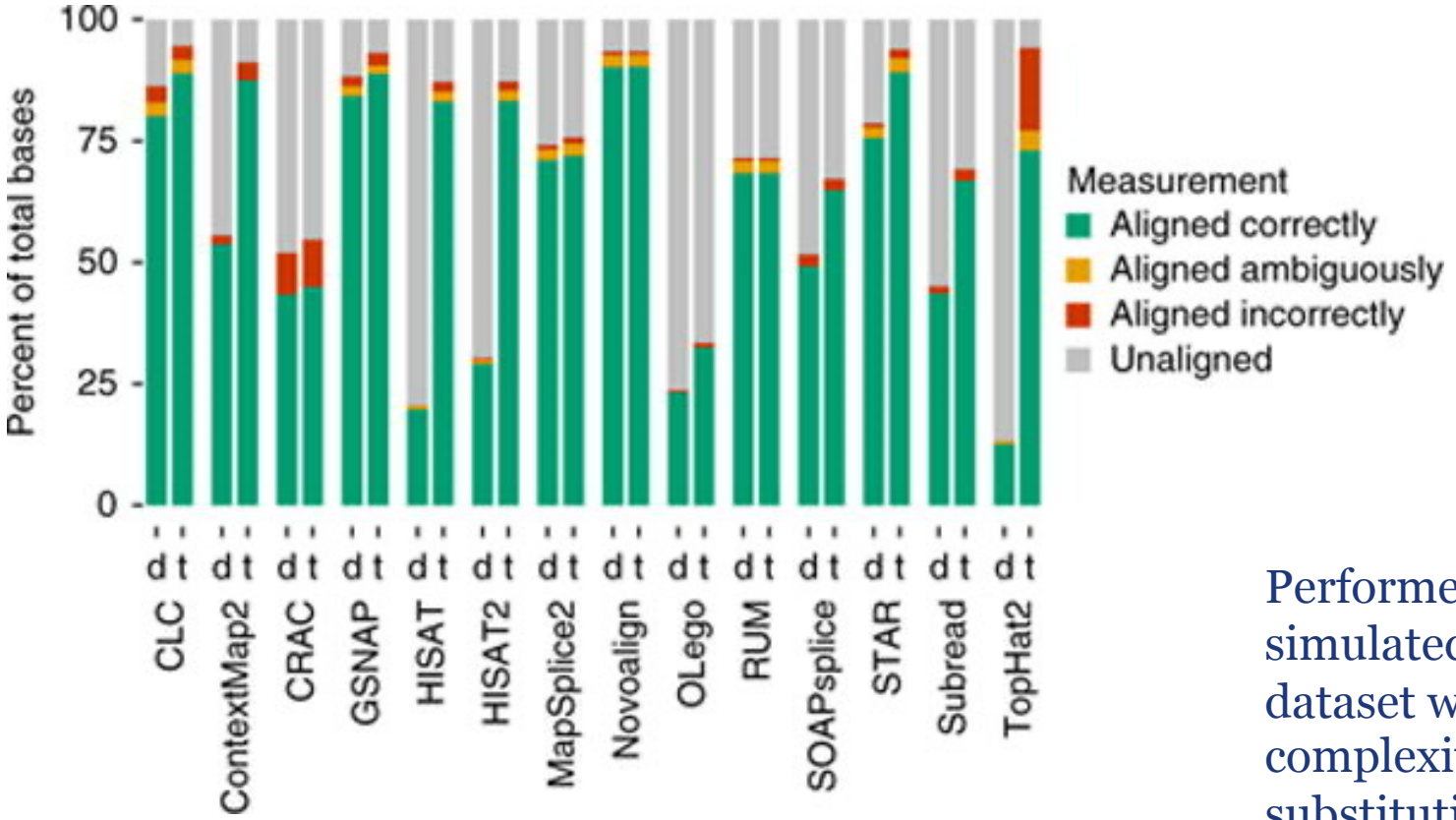
1. Trimmed sequences in FASTQ format
2. Complete reference genome FASTA (or transcriptome)
3. Reference annotation file in GTF or GFF3 format (not required for non-splice aware aligners)

3a. 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!!!



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

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

Optional: Alignment Visualization

IGV is the visualization tool used for this snapshot

3a. Traditional Gene Counting: # of sequences in genes

When selecting software consider whether you want to obtain:

- raw read counts or normalized read counts

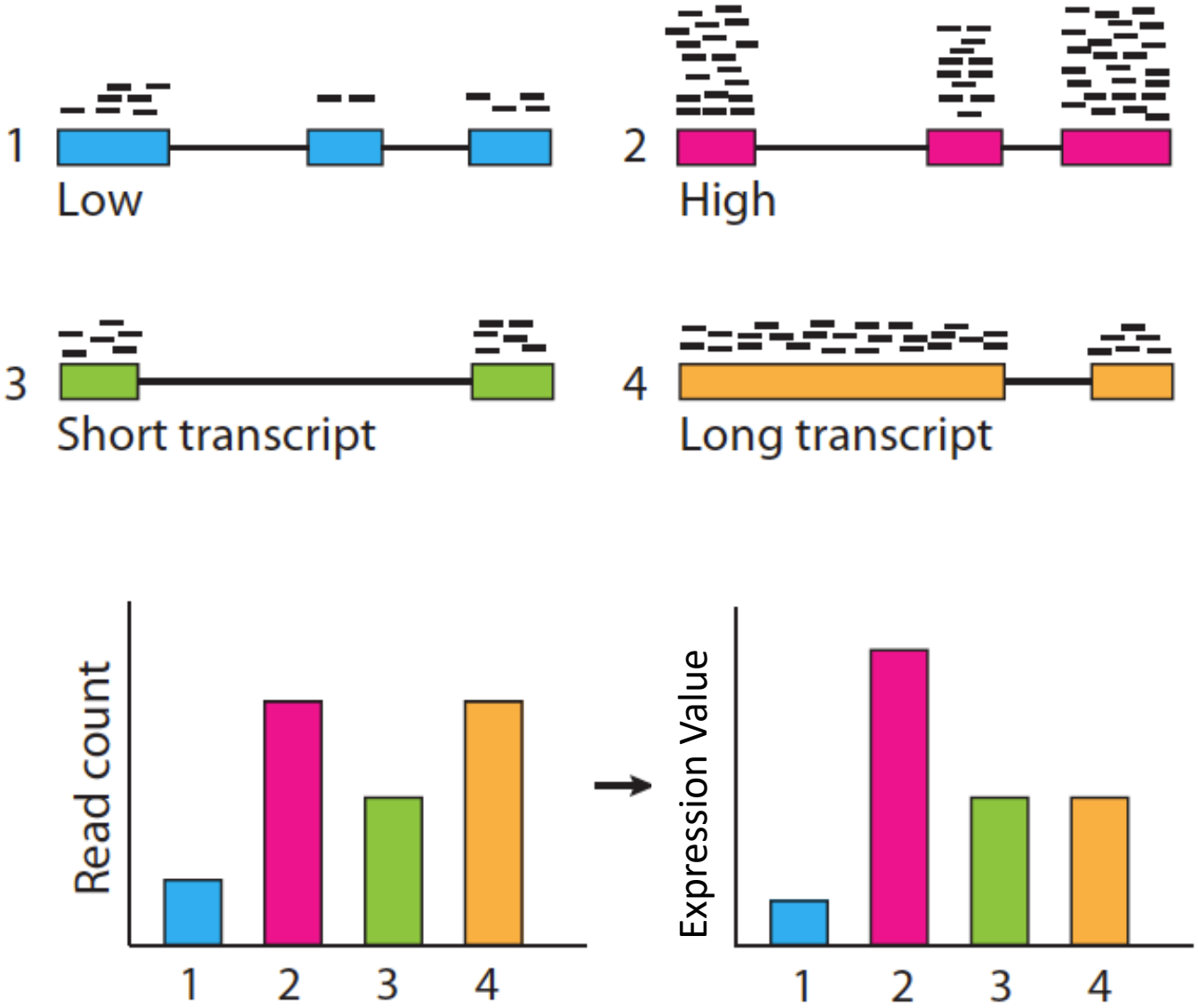
Gene counting software:

- 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

Problems with traditional gene counting software

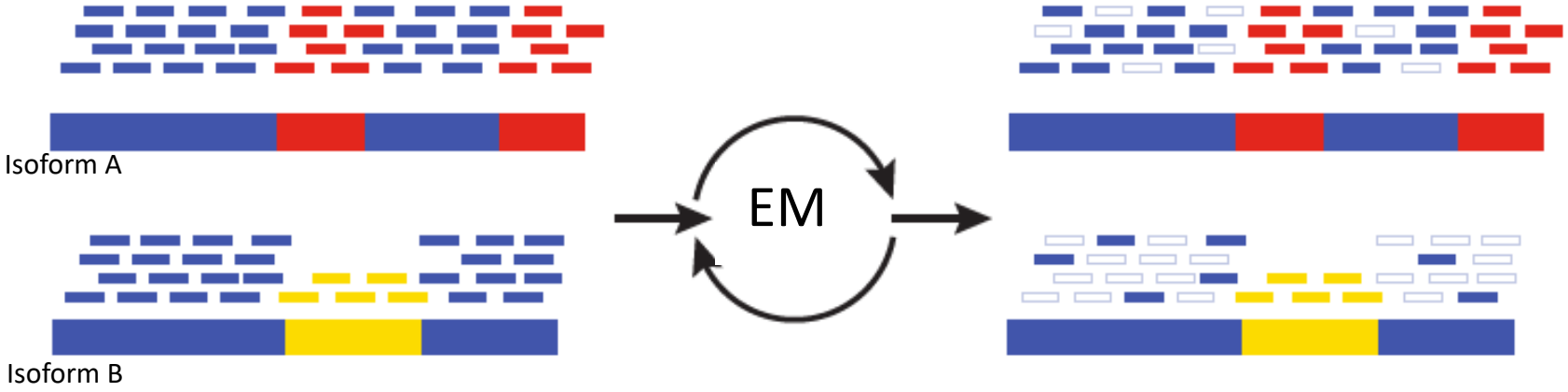
1. Multi-mapping reads not used, leading to underestimation of gene abundances, particularly for genes with more shared sequence
2. A small percentage of genes may not ever be quantifiable using this method.
3. Genes that change relative isoform usage can have erroneous results due to changes in isoform length

Calculating expression of genes and transcripts



Slide courtesy of Cole Trapnell

3b. Gene quantification
Solution: Expectation Maximization algorithms



Blue = multiply-mapped reads
Red, Yellow = uniquely-mapped reads

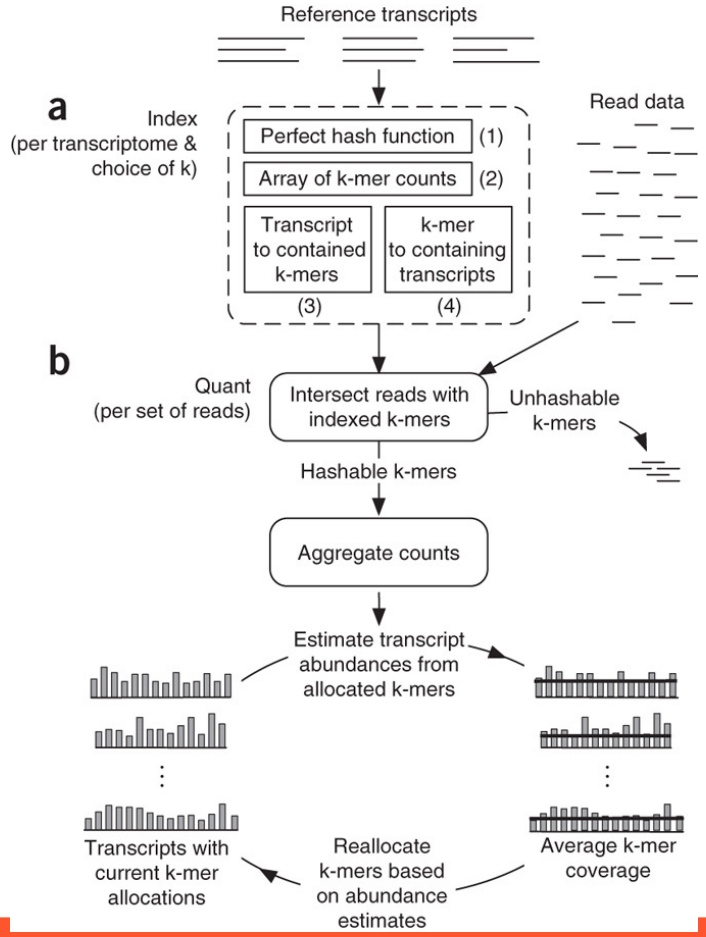
Use Expectation Maximization (EM) to find the most likely assignment of reads to transcripts.

Performed by:

- StringTie
- RSEM
- eXpress
- Salmon/kallisto

Modern transcript counting programs

- Salmon (Patro et al. 2017)
  - Estimates transcript coverage by k-mer counting approach
  - Cannot find new splice junctions/isoforms
  - Fast: 3-5 min!
- Kallisto (Bray et al. 2016)
  - Also utilizes de Bruijn graphs
  - less than 5 min on laptop computer!
- "Gene quantification"
  - Both software are more accurate when transcript-counts are grouped to the gene-level



When to use either method

Traditional Gene Counting (1st workflow)	Gene Quantification (2nd workflow)
Reads with retained introns still counted (e.g. cancer and rapidly developing tissues)	Genome duplications present
Need to find novel transcripts/splice junctions	Lots of gene families present
Want to visualize alignments on genome	When ever you have a large percentage (>15%) of multi-mapped reads

Today's Topics

- Transcriptomics
- Traditional RNA-Seq Methods
  - Sequencing & experimental considerations
  - Gene counting
  - Statistics
- Single Cell RNA-Seq Methods
  - Sequencing
  - UMI counting & statistics
- Where to find help

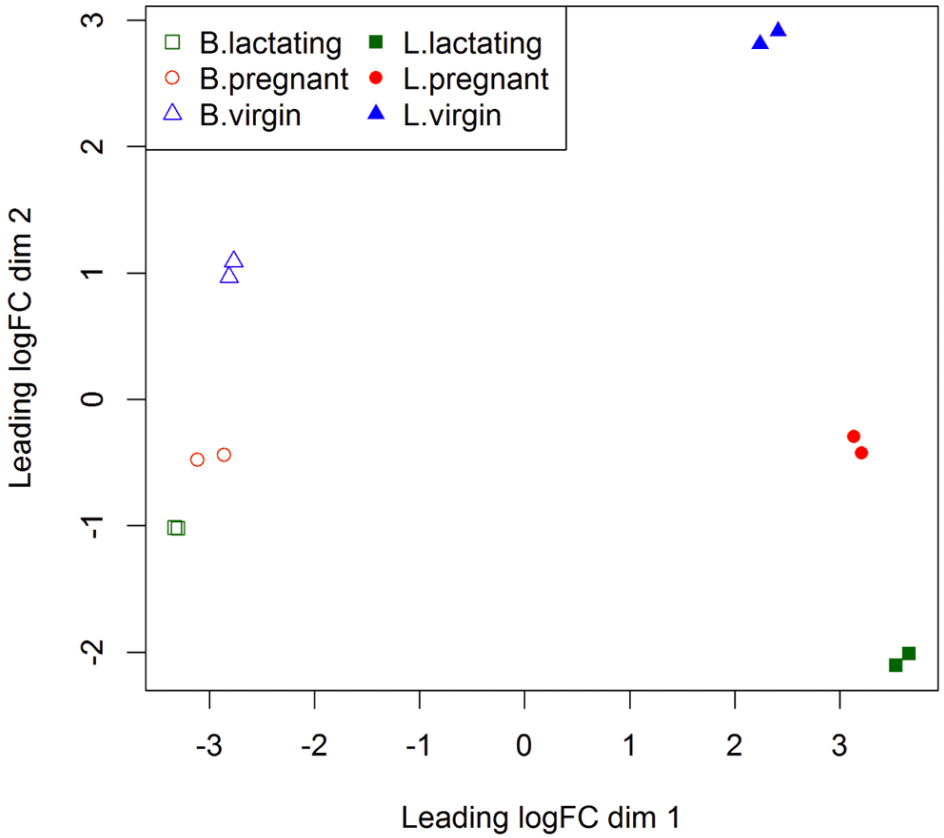
DGE Statistical Analyses

1. The first step is proper normalization of the data
   - ✧ Often the statistical package you use will have a normalization method that it prefers and uses exclusively (e.g. Voom, FPKM, TMM (used by EdgeR))
2. Is your experiment a pairwise comparison?
   - ✧ Ballgown, EdgeR, DESeq
3. Is it a more complex design?
   - ✧ EdgeR, DESeq, limma, other R/Bioconductor packages

Statistical Results

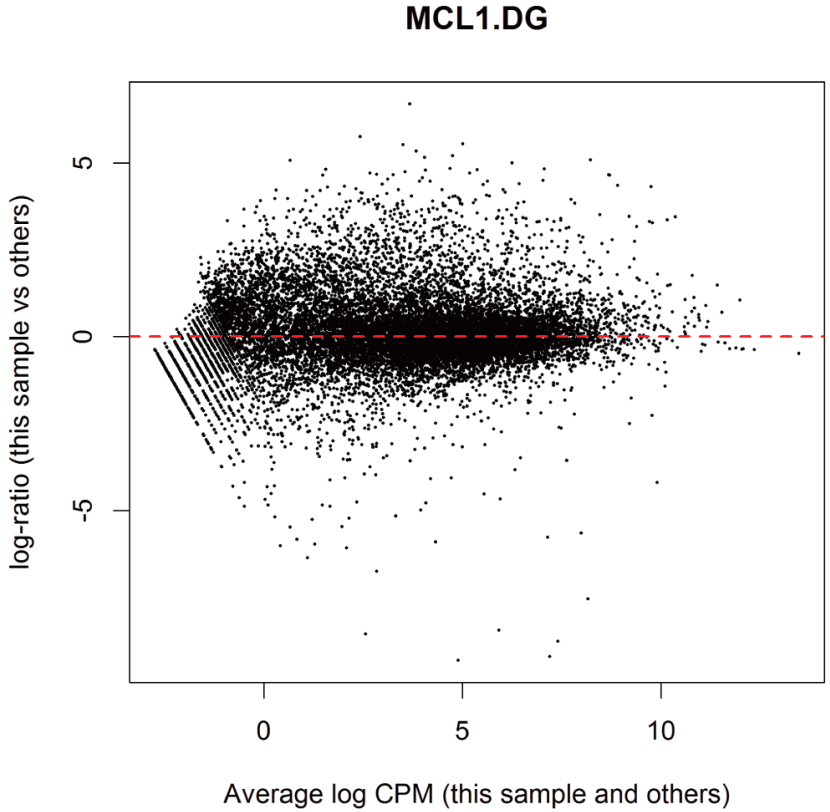
- A list of significantly differentially expressed genes
- Venn Diagrams
- Heatmaps
- WGCNA
- Advanced annotation
- ... and more!

EdgeR: MDS Plot



https://f1000research.com/articles/5-1438 (doi: 10.12688/f1000research.8987.2)

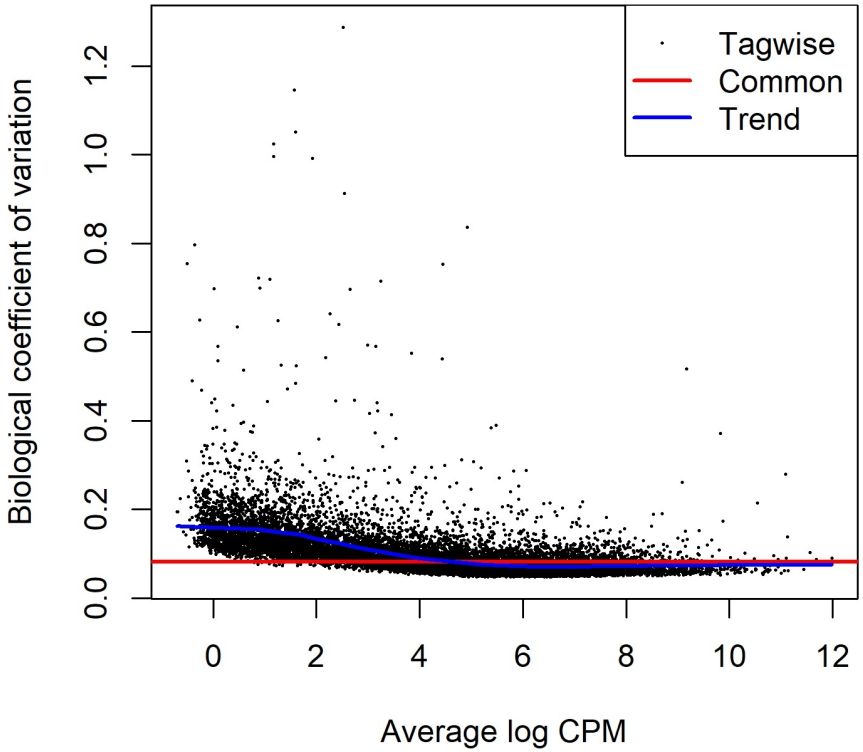
EdgeR: MD Plot



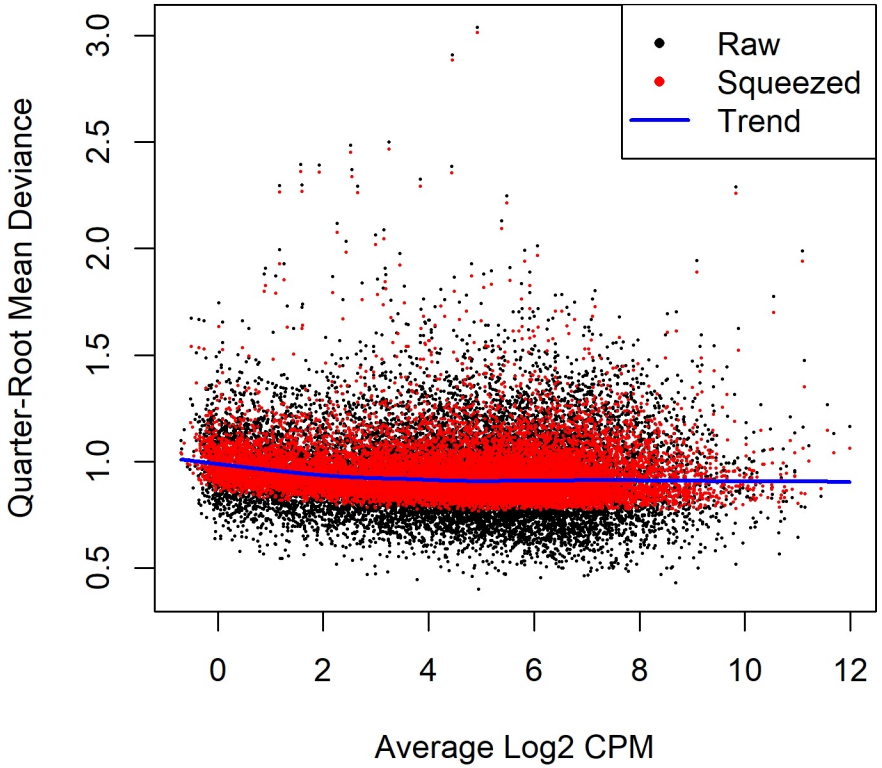
https://f1000research.com/articles/5-1438 (doi: 10.12688/f1000research.8987.2)

EdgeR Results: Dispersion Estimation

BCV Plot



QL Plot



https://f1000research.com/articles/5-1438

What does HPCBio use?

1. Quality Check - **FASTQC**
2. Trimming - **Trimmomatic**
3. Splice-aware alignment - **STAR**
   Bacterial alignment - **BWA** or **Novoalign**
4. Counting reads per gene - **featureCounts**
   Counting reads per isoform - **Salmon*** *Can also group these counts by gene for even more accuracy*
5. DGE Analysis - **edgeR** or **limma**

- Alignment visualization - **IGV**
- De novo transcriptome assembly – **Trinity**
- Reference-based transcriptome assembly – **StringTie**

Still not sure?

Recent RNA-Seq software comparison articles:

Corchete, L.A., Rojas, E.A., Alonso-López, D. *et al.* Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Sci Rep* **10,** 19737 (2020). https://doi.org/10.1038/s41598-020-76881-x

Schaarschmidt, S., Fischer, A., Zuther, E., & Hincha, D. K. (2020). Evaluation of Seven Different RNA-Seq Alignment Tools Based on Experimental Data from the Model Plant *Arabidopsis thaliana*. *International journal of molecular sciences*, *21*(5), 1720. https://doi.org/10.3390/ijms21051720

Zhang, C., Zhang, B., Lin, LL. *et al.* Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* **18,** 583 (2017). https://doi.org/10.1186/s12864-017-4002-1

Today's Topics

- Transcriptomics
- Traditional RNA-Seq Methods
  - Sequencing & experimental considerations
  - Gene counting
  - Statistics
- Single Cell RNA-Seq Methods
  - Sequencing
  - UMI counting & statistics
- Where to find help

10x Genomics RNA Applications:

RNA-Seq:

- "Bulk" Gene Expression profiling

10x **Single Cell** Transcriptomics (Single Cell RNA-Seq):

- 3' Gene Expression profiling at single cell resolution 10x

Spatial Gene Expression (Visium):

- 3' Gene Expression profiling with morphological content (1-10 cell resolution)

Bulk

Single cell

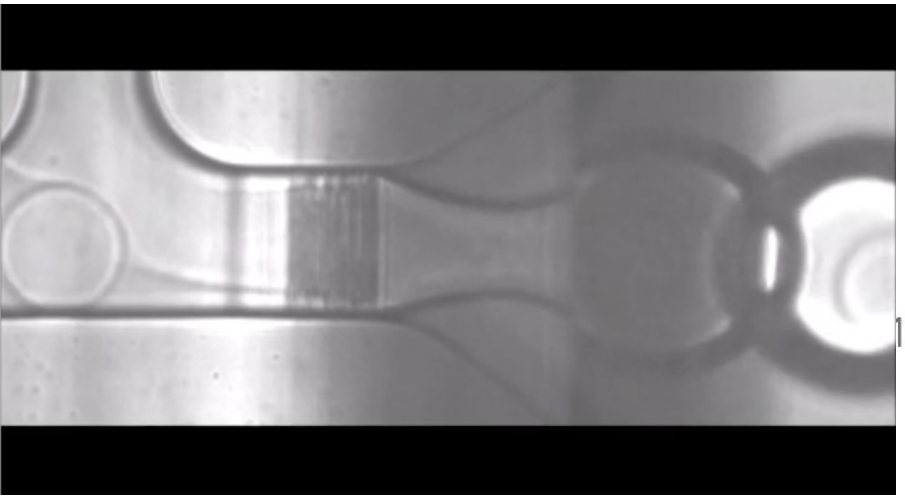
Spatial

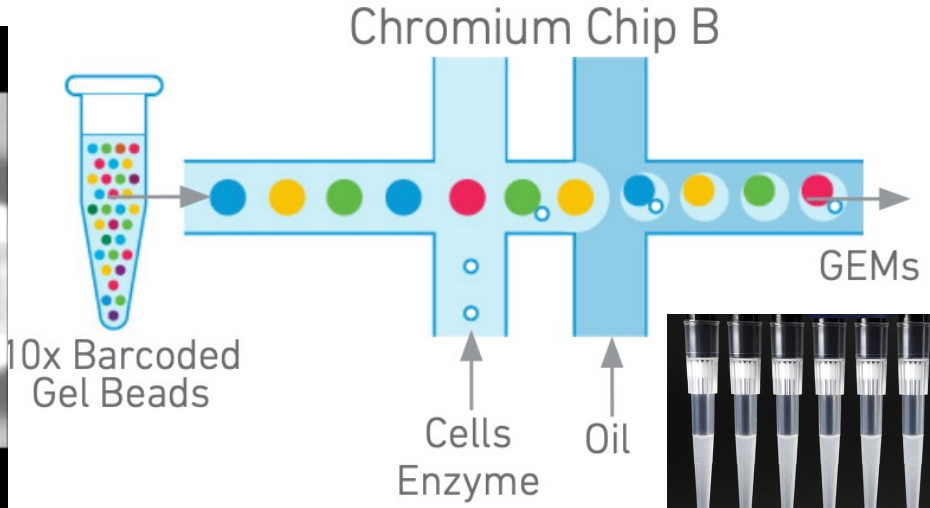
Why Single Cell RNASeq?

- Comparison of individual cell transcriptomes within a population of cells
- Analysis of cell heterogeneity/rare cell populations (average vs individual)
- Embryonic tissue / tumors
- Examining cell populations consisting of mostly unique cells (TCR, neurons)
- Transcriptome Atlas

10x Single Cell Construction:

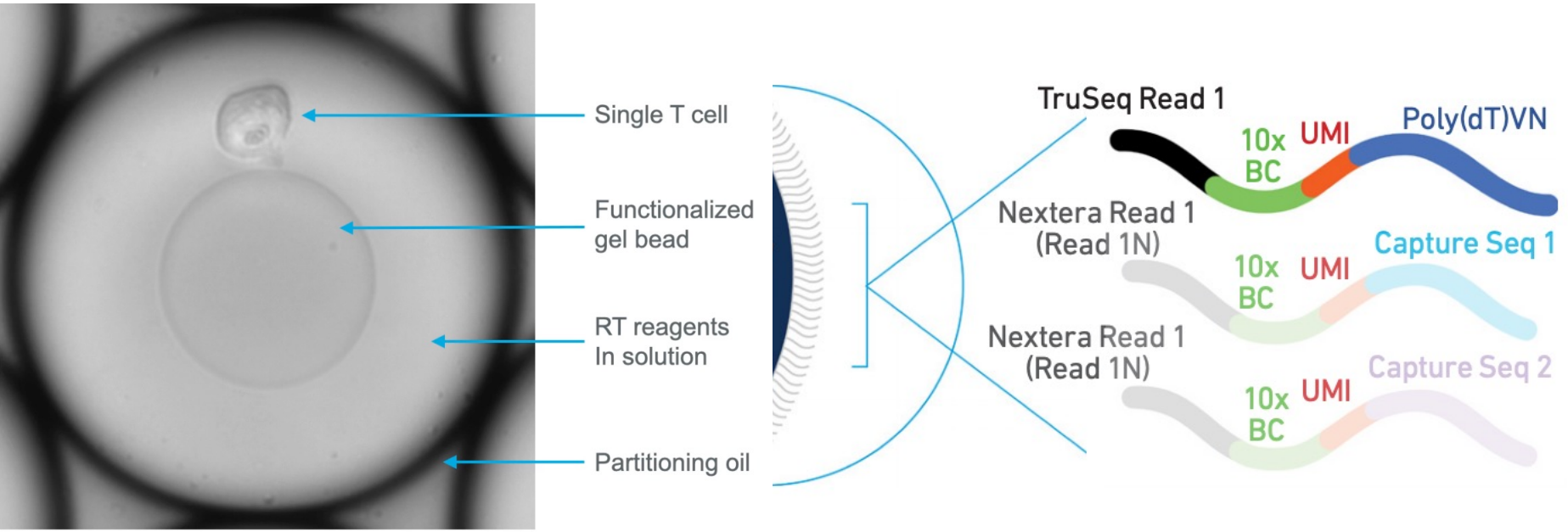
- GEM Generation: Gel Beads in Emulsion
  - 10x Chip: Add Gel Beads + Master Mix/Cells + Oil.
  - Mixing occurs on Chromium X.
  - 90-99% of GEMS do not contain a cell.





10x Single Cell Construction:

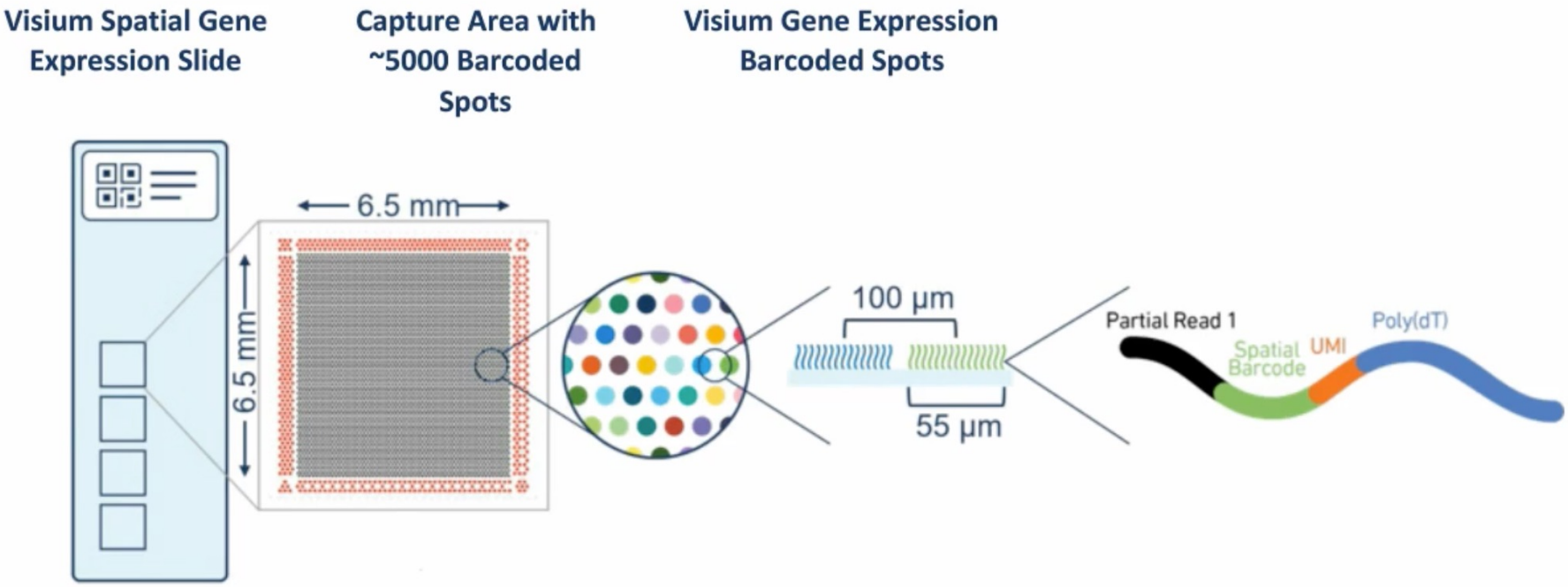
- GEM Generation and Barcoding
  - 10x Barcode: One for each cell captured (~3.5M 16bp)
  - UMI: Unique Molecular Identifier (12bp N's)



10x Visium Spatial Transcriptomics:

Unbiased Gene Expression at High Spatial Resolution

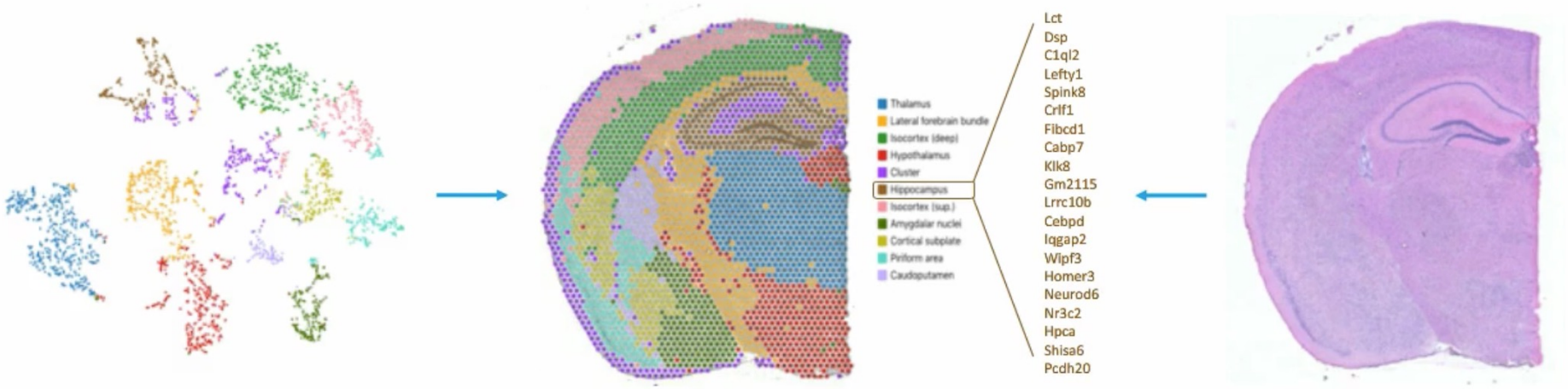
Utilizing Poly-A Capture and Unique Spatial Barcodes



10x Visium Spatial Transcriptomics:

Cluster or Image Driven Analysis of Spatial Data

Start With the Gene Expression Data or microscopy images of the same section



Clustering/classification of glass slide spots based on their gene expression profile

Annotation directly on H&E image

Captured H&E image from the microscope

Today's Topics

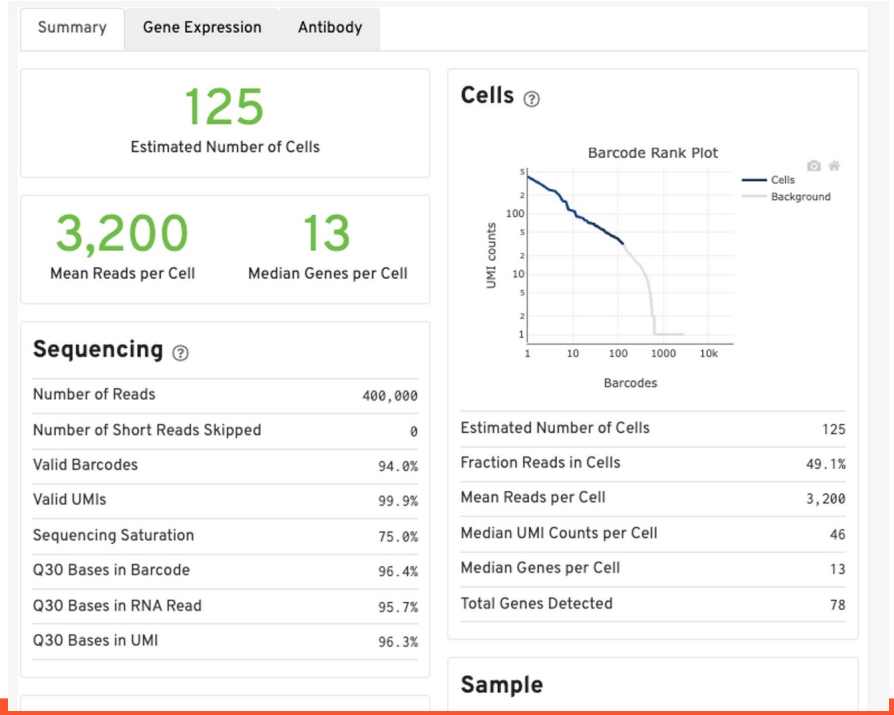
- Transcriptomics
- Traditional RNA-Seq Methods
  - Sequencing & experimental considerations
  - Gene counting
  - Statistics
- **Single Cell RNA-Seq Methods**
  - **Sequencing**
  - **UMI counting & statistics**
- **Where to find help**

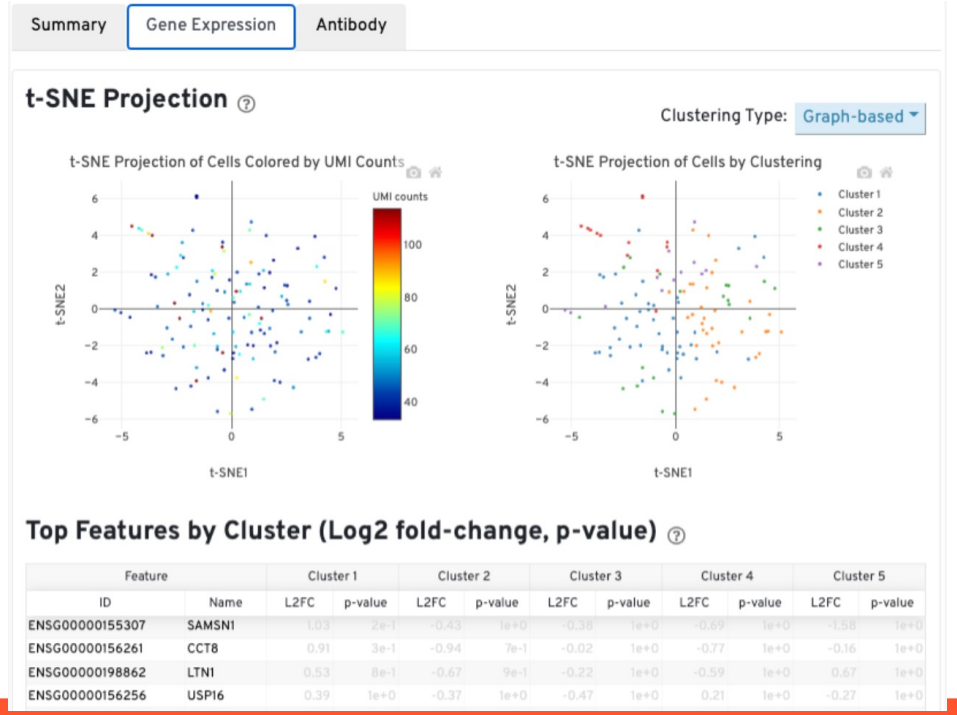
UMI counting?

- **UMI** = **U**nique **M**olecular **I**dentifier
- A UMI barcode is assigned to each **transcript**, in addition to a cell barcode (for differentiating single cells)
- We can use Cell Ranger to run differential gene expression analysis & create a visualization file
  - cellranger mkfastq
    - Demultiplexes samples (run by sequencing center)
  - cellranger count
    - Run separately for each sample
  - cellranger aggr
    - Puts > 1 sample in same .cloupe file
    - Only normalization is sub-sampling
  - cellranger reanalyze
    - Used to modify parameters of previous run

What does Cell Ranger give us?

- Cell Ranger produces:
  - web_summary.html file
  - .cloupe file for visualization (next slides)



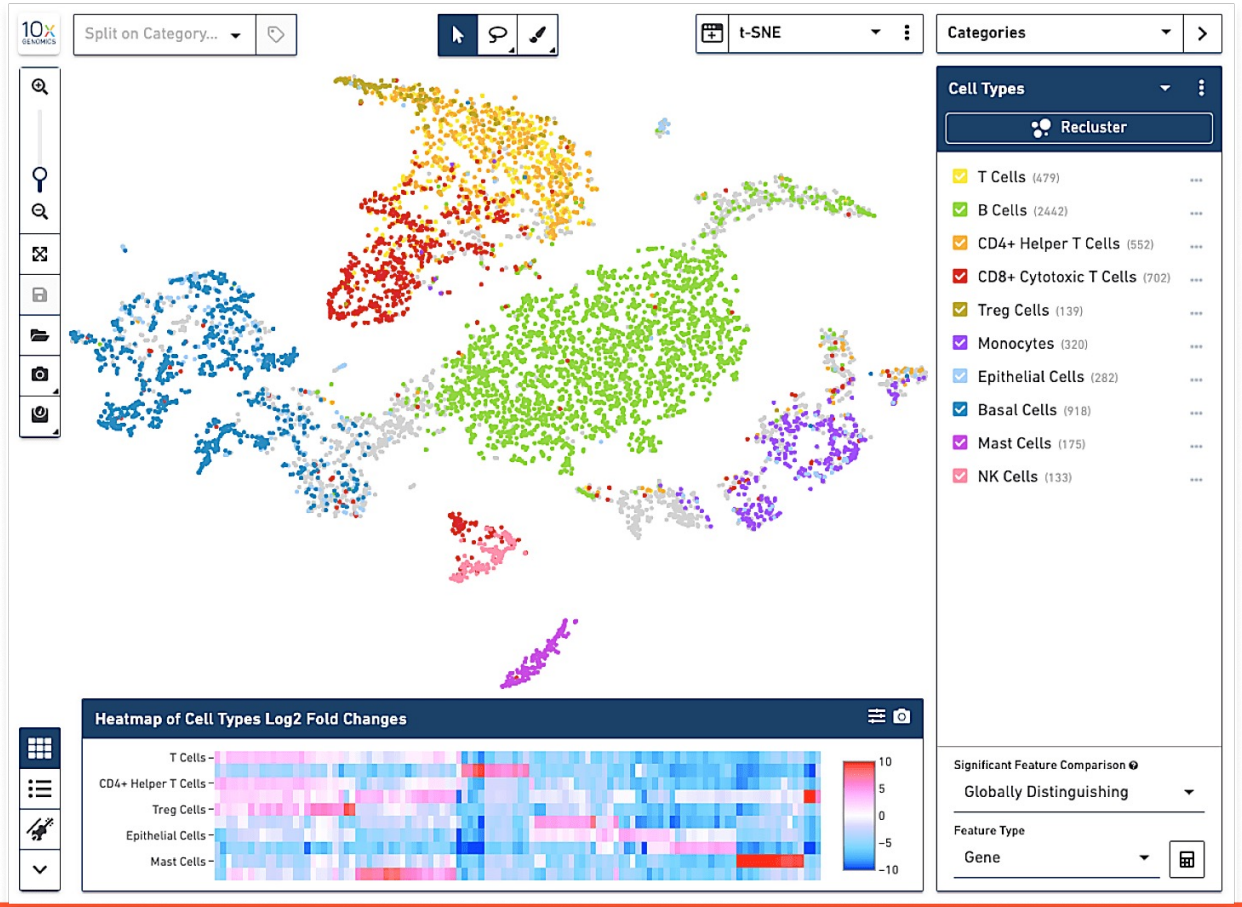


Loupe Browser

After UMI counting and clustering, you can import this data into Loupe Browser

Loupe Browser

You can also refine your clusters and assign cell types (manually)



Cell Ranger Limitations

- Excluding called cells based on other QC metrics (%MT) not easy
  - Manual selection/output in Loupe
  - Need to run cellranger reanalyze to remove cells/genes
- Only a sub-optimal normalization method available

Loupe Browser Limitations

- Limited tracking of steps performed (reproducible?) but you can save what you have done
- Manual annotation of cell types
- No trajectory analysis
- Output plots not publication-quality
- Not a substitute for a rigorous statistical analysis

Secondary analyses recommended using R

1. Filter genes
2. Filter cells
3. Select variable genes
4. Run PCA
    1. Run clustering
    2. Run t-SNE
    3. Run UMAP
5. Find marker genes per cluster
6. Annotate cell types
7. Differential analysis between cell types and/or within a cell type between samples

Recommended R Packages

- HPCBio prefers to use Seurat
- Cell annotation can be done by many programs
- Other popular R packages, especially for trajectory analysis:
  - Monocle is an R package for clustering, trajectory analysis, and differential expression
  - Scanpy is a Python tool for clustering, trajectory inference, and differential expression
- And more!
  - Link to a list of Bioconductor packages for working with single cell data
  - Link to a list of Bioconductor workflows for working with single cell data
  - Link to a list of Bioconductor single cell data packages

Today's Topics

- Transcriptomics
- Traditional RNA-Seq Methods
  - Sequencing & experimental considerations
  - Gene counting
  - Statistics
- Single Cell RNA-Seq Methods
  - Sequencing
  - UMI counting & statistics
- Where to find help

How do I learn more about these steps?

- Your lab will briefly go through some steps of a traditional RNA-Seq dataset: **alignment, gene-counting, DGE analysis, and alignment visualization**
- We do offer a longer and very detailed workshop on these methods
  - Spring semester: Bulk RNA-Seq workshop
  - Fall semester: 10X single cell RNA-Seq workshop
- Check http://hpcbio.illinois.edu/hpcbio-workshops at the beginning of the year for updates

Documentation and Support

Online resources for RNA-Seq analysis questions

- Software manuals
    - Most tools also have a dedicated lists/forums and/or github pages
- Biostar (Bioinformatics explained) - http://www.biostars.org/
- SEQanswers (the next generation sequencing community) - http://seqanswers.com/

HPCBio Bioinformatics Consulting






Contact us at:

Help desk - hpcbio@biotech.illinois.edu

Training questions - hpcbio-training@biotech.illinois.edu

HPCBio website - http://hpcbio.illinois.edu/

My email - jholmes5@illinois.edu

CNRG (Computer Network Resource Group)

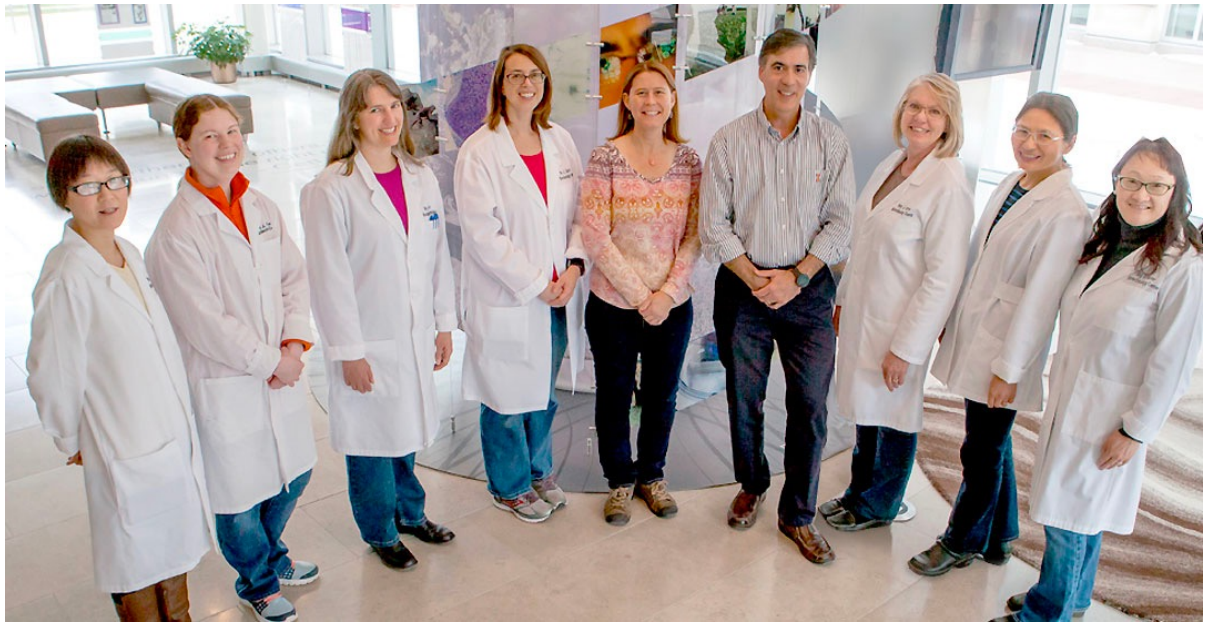
- Need help using Biocluster or need new software installed on it?

Contact CNRG at:

Help desk - help@igb.illinois.edu

CNRG website - https://help.igb.illinois.edu/

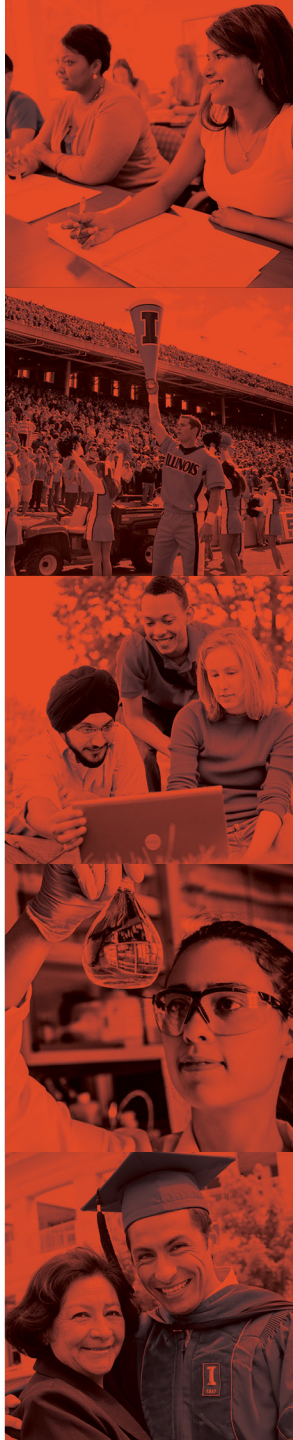
DNA Services Laboratory



Director: Alvaro Hernandez
aghernan@Illinois.edu
329 ERML
217-244-3480

Associate Director: Chris Wright
clwright@Illinois.edu
334 ERML
217-333-4372

www.biotech.Illinois.edu/htdna



New CMtO Core lab in CBC:

- Cytometry and Microscopy to Omics (CMtO)
  - Led by **Mayandi Sivaguru (Shiv), PhD., with specialist Kate Jaansen***
  - Your existing CBC Flow Cytometry Core, plus:
    - Assist with single cell / nuclei test counts (BigFoot cell sorter; K2)
    - New BSL2 Cryostat and microtome for Visium workflow
    - Custom configured Zeiss LSM980 microscopy system

Enables end-to-end CBC support, from CMtO to DNA Services to HPCBio, for 10x Visium and single cell workflows!

Thank you!