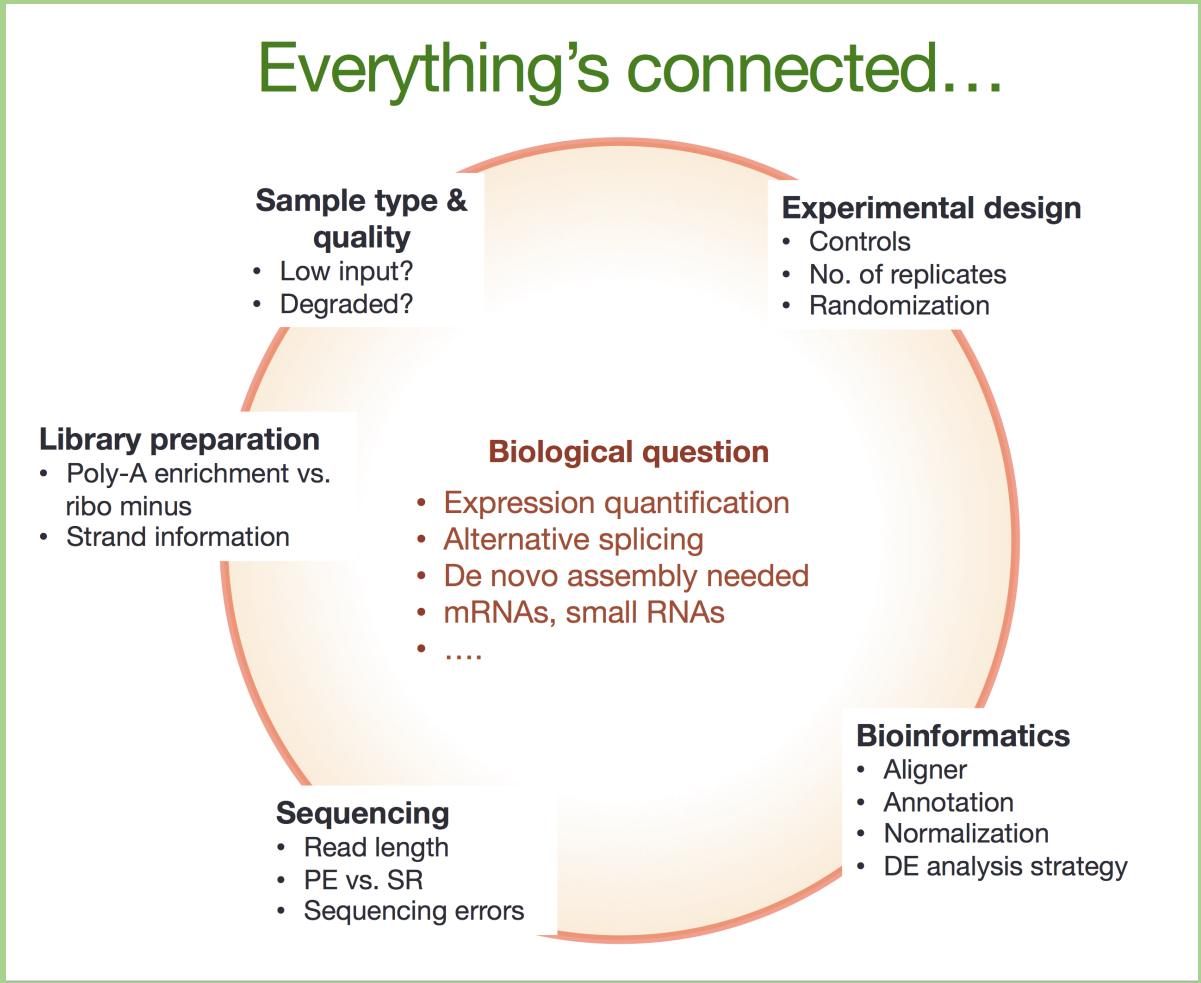


Lecture 1: Setting up an RNA-Seq experiment at UVM

Princess Rodriguez, PhD

MMG 3320/5320
Spring 2024

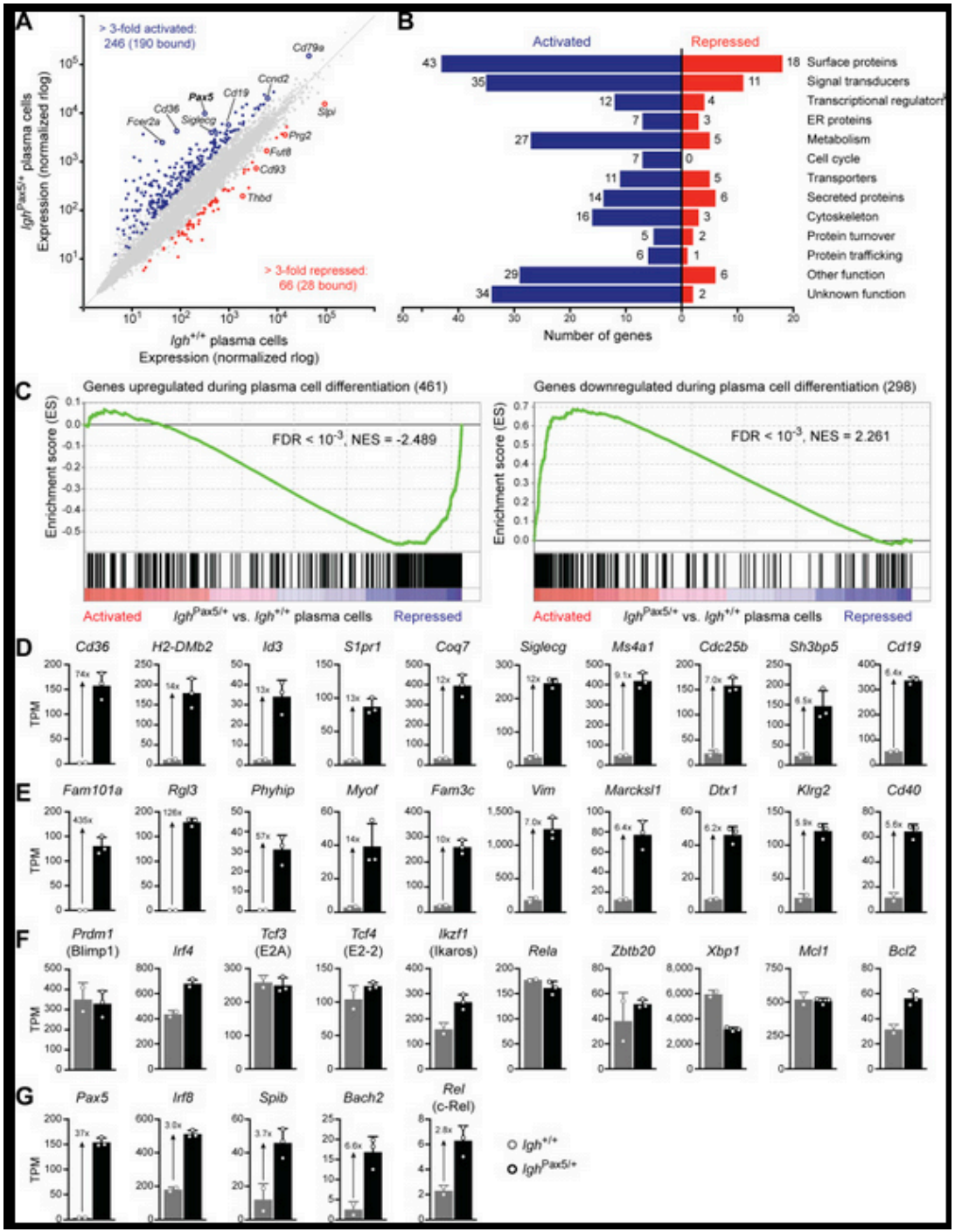
Outline of today's lecture

- Final Project Overview
- Considerations for setting up an RNA-Seq experiment


Green Mountain Trail

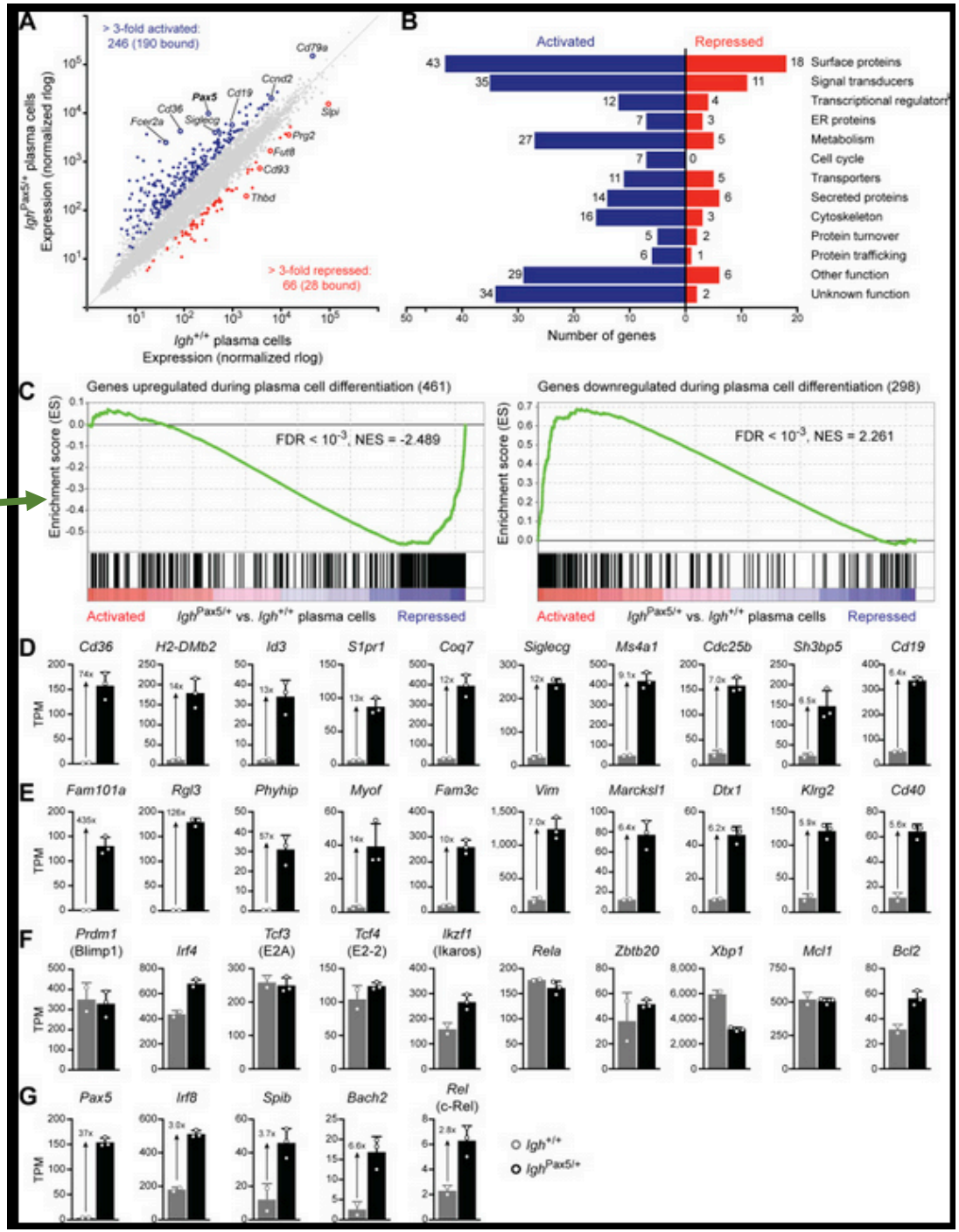
Green Mountain Trail

Your overall goal should be to replicate a Figure or Figure panel (A-C) from a peer-reviewed article



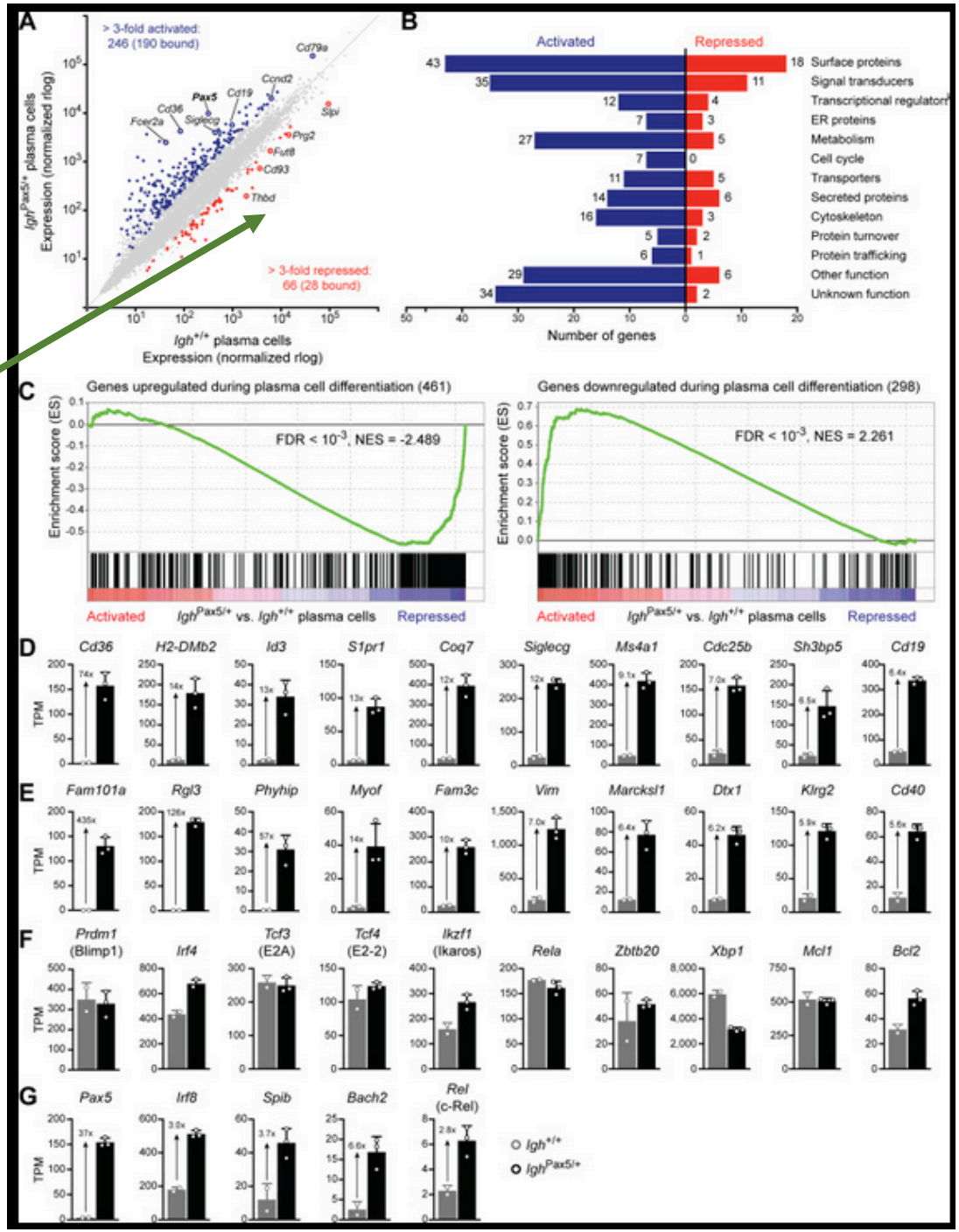
Green Mountain Trail

Perhaps you never truly understood what an Enrichment Score (ES) is?



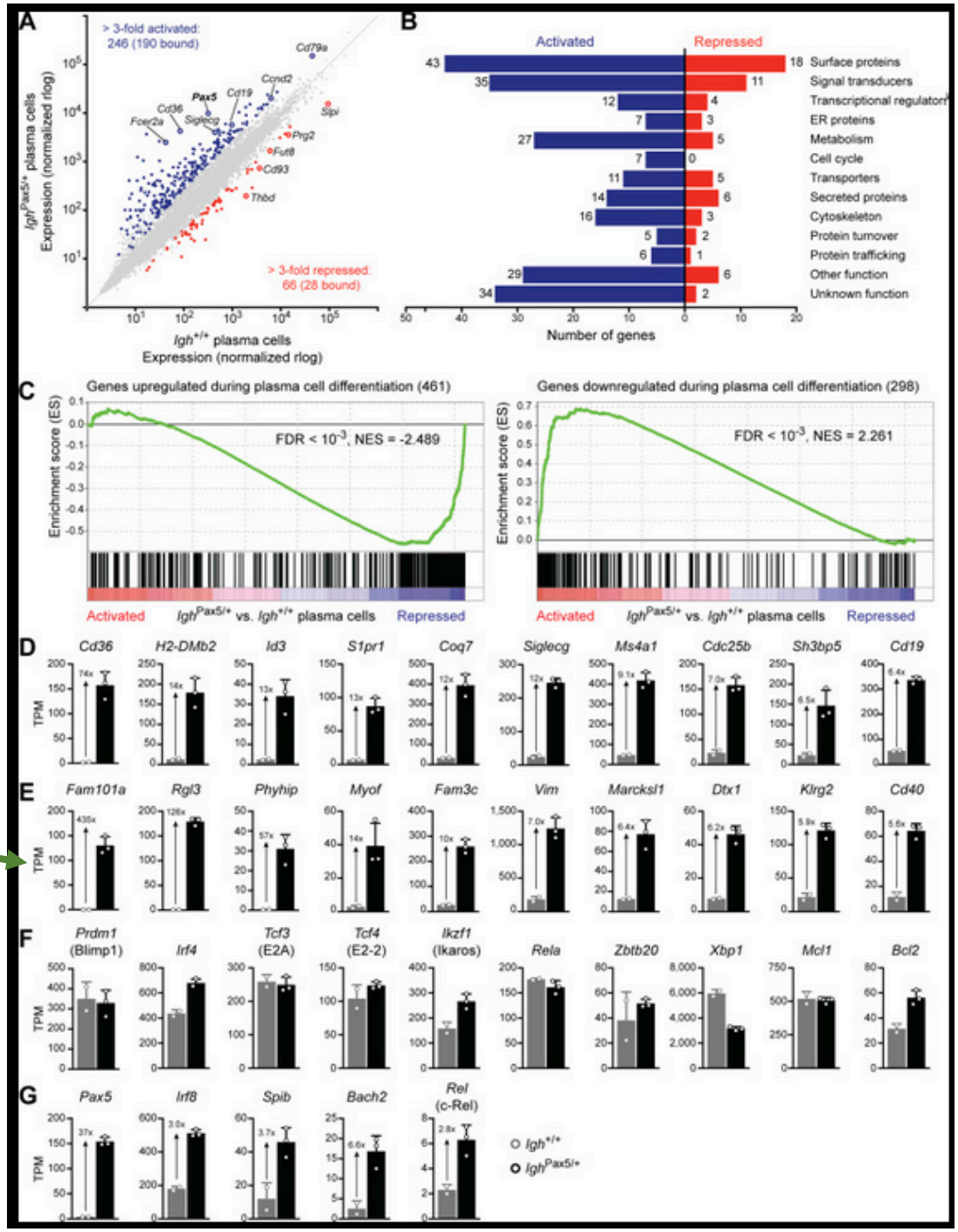
Green Mountain Trail

How would you label this MA plot?



Green Mountain Trail

Do you know TPM means?




Blue Sky Trail

The bioinformatic pipeline we will learn in class is:

MMG232	What it does…
FASTQC	Quality control FASTQC files
Trimmomatic	Trim adaptors and low quality reads
HISAT2 STAR	Alignment to Genome
SAMtools	SAM to BAM
HTSeq-count	Create counts files

RNA-Seq and bioinformatics analysis

Total RNA prepared was extracted using TRIzol Reagent (Invitrogen, Carlsbad, CA, USA) and submitted to Shanghai Personal Biotechnology, where RNA intergrity was confirmed using the Illumina HiSeq X ten system at 150 bp pair-ended. Double-strand cDNA libraries were prepared and constructed using the TruSeq RNA Sample Prep Kit (Illumina, San Diego, CA). Two replicates of the RNA-Seq experiments were performed. RNA-Seq reads were quality controlled and trimmed for adapter sequences using Trim Galore. Filtered reads were aligned to hg38 using HISAT2. Read counts for each gene were carried out using HT-Seq using the hg38 refSeq refFlat GTF file accessed on July 2015. Differentially expressed genes (DEGs) were analysed using the DESeq2 package ($|\text{fold change}| \geq 1.5$, $P < 0.05$).

FASTQC

Trimmomatic

HISAT2

STAR

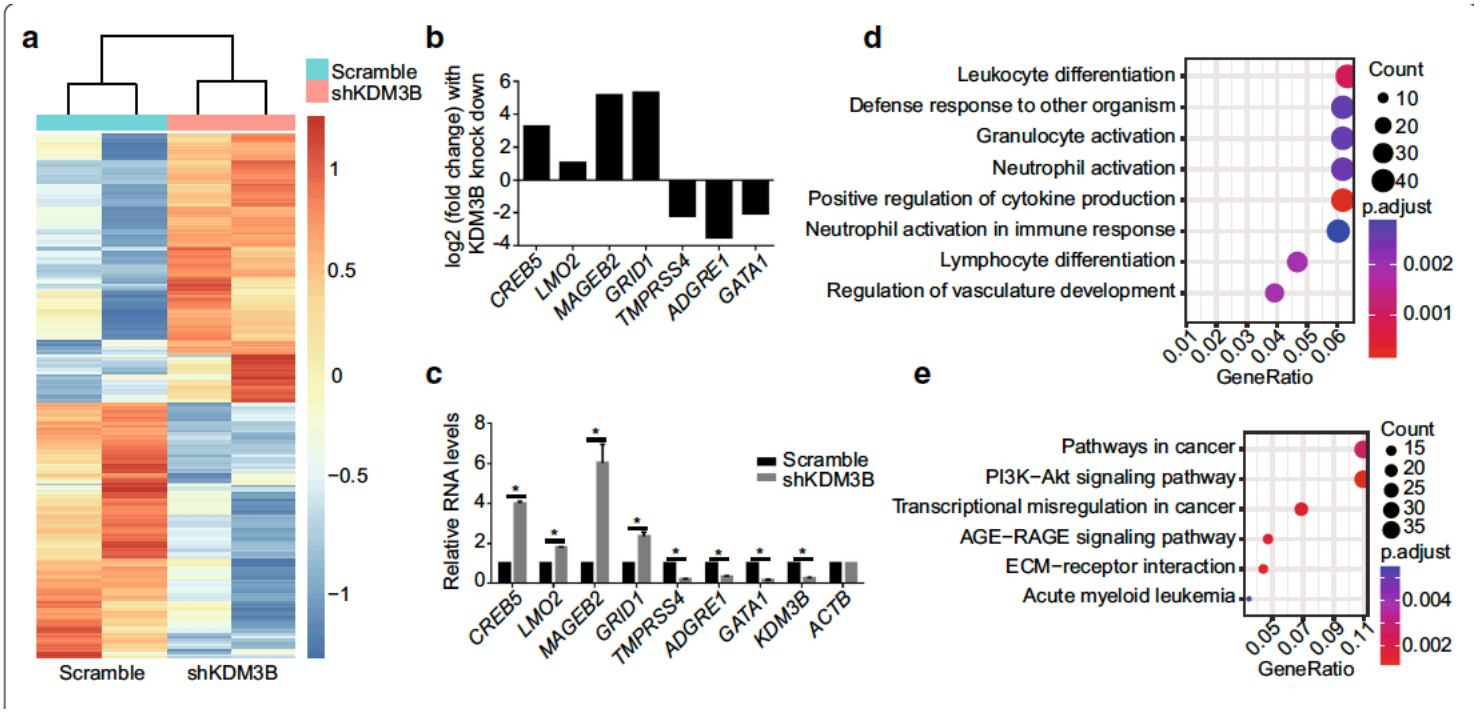
SAMtools

HTSeq-count

DeSeq2

# of DEGs

VERSUS



PAPER #2

RNA sequencing. MDMs (5×10^5/well in 24-well plates) were infected with *A. fumigatus* conidia at a 1:2 effector-to-target ratio, and pooled replicates from three different individuals were collected after 2 and 6 h. Uninfected MDMs were cultured in parallel as controls. Sample processing, sequencing, and analysis were performed at IMGM Laboratories GmbH (Germany). Briefly, the total RNA was isolated using the RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions, including on-column DNase digestion. The total RNA was eluted in 30 µL of RNase-free water. The quality of the total RNA was analyzed with the 2100 Bioanalyzer using RNA 6000 Nano and Pico LabChip kits (Agilent Technologies). Library preparation was performed using the TruSeq® Stranded mRNA HT technology, according to the manufacturer's protocol. All single libraries were pooled into a final sequencing library with an equal DNA amount per sample. The final sequencing library generated by pooling was quantified using the highly sensitive fluorescent dye-based Qubit® dsDNA HS Assay kit (Invitrogen) before sequencing at a final concentration of 1.8 pM and with a 1% PhiX v3 control library spike-in (Illumina) on the NextSeq500 sequencing system (Illumina). For the clustering and sequencing of samples, a high-output single-end 75 cycles (1×75 bp SE) run was performed under the control of the NextSeq Control Software (NCS, Illumina). Quality control was carried out using NCS and Real Time Analysis 2.4.11 softwares applying the *FastQ only* pipeline. Read data were imported into the CLC Genomics Workbench (CLC bio, Qiagen), and reads were mapped against the human reference genome (GRCh37.p13) with subsequent counting and distribution of reads across genes and transcripts. The expression values were then processed to reads per kilobase million (RPKM), a normalized measure of relative abundance of transcripts[68], followed by analysis using the EdgeR Bioconductor package[69] to identify differentially expressed genes with a fold change value ≥2 or ≤ −2 with a false discovery rate (FDR)-corrected p-value < 0.05. Heatmaps were created for the most significantly represented genes of a specific functional class using the Morpheus tool (Broad Institute; https://software.broadinstitute.org/morpheus/). For pathway analysis, the annotated hallmark gene sets from the Molecular Signatures Database (MSigDB)[33] were used and enrichment analysis was performed using the Gene Ontology Biological Processes category in the Gene Ontology Consortium software (http://www.geneontology.org/).

Very vague methods

PAPER #2

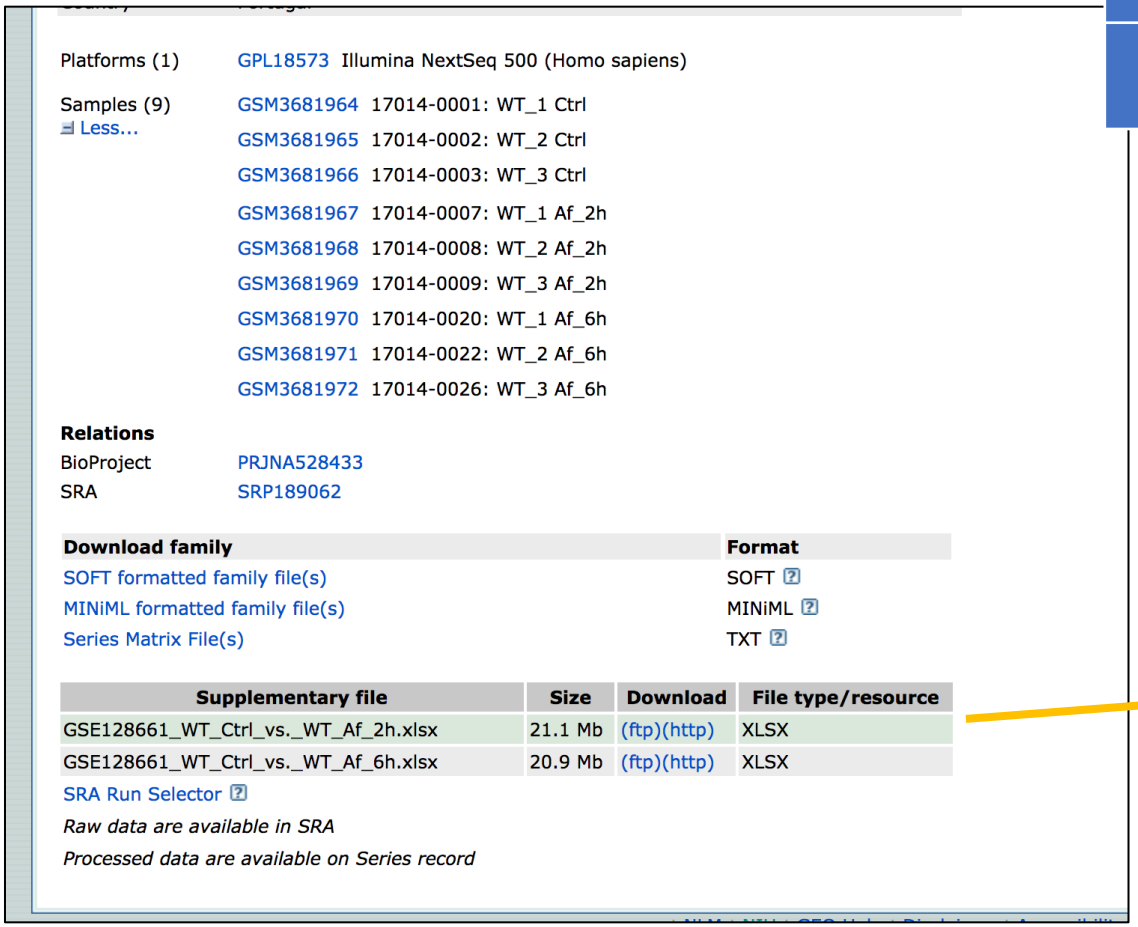
FASTQC

Trimmomatic

HISAT2

SAMtools

HTSeq-count
DeSeq2



Platforms (1) GPL18573 Illumina NextSeq 500 (Homo sapiens)

Samples (9)
Less...

GSM3681964 17014-0001: WT_1 Ctrl
GSM3681965 17014-0002: WT_2 Ctrl
GSM3681966 17014-0003: WT_3 Ctrl
GSM3681967 17014-0007: WT_1 Af_2h
GSM3681968 17014-0008: WT_2 Af_2h
GSM3681969 17014-0009: WT_3 Af_2h
GSM3681970 17014-0020: WT_1 Af_6h
GSM3681971 17014-0022: WT_2 Af_6h
GSM3681972 17014-0026: WT_3 Af_6h

Relations

BioProject PRJNA528433
SRA SRP189062

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE128661_WT_Ctrl_vs._WT_Af_2h.xlsx	21.1 Mb	(ftp)(http)	XLSX
GSE128661_WT_Ctrl_vs._WT_Af_6h.xlsx	20.9 Mb	(ftp)(http)	XLSX

SRA Run Selector

Raw data are available in SRA

Processed data are available on Series record

VERSUS

of
DEGs


Black Diamond Trail

Your overall approach will be different

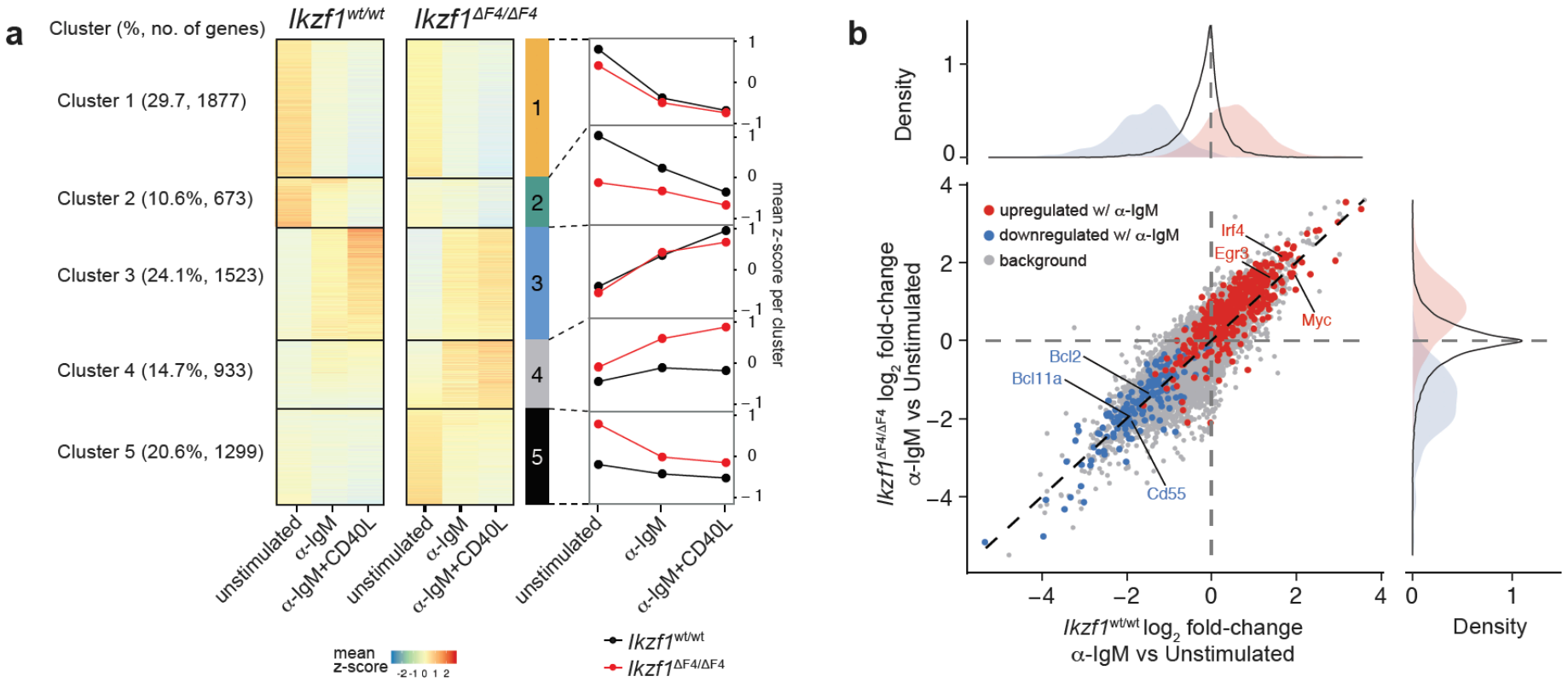
You are going in with a question and using the dataset to answer this question!

“I am interested in studying the difference in metabolic pathway regulation in macrophage versus dendritic cell activation.”

“I am interested in understanding how inhibition of GATA1 impacts the ability of naïve B cells to differentiate into plasma cells.”

"I am interested in studying the difference in metabolic pathway regulation in macrophage versus dendritic cell activation."

Dataset	# of replicates
Macrophages (control)	3
Dendritic cells (control)	3
Macrophages + LPS	3
Dendritic cells + LPS	3
Macrophages + Zeb1 KO	3
Dendritic cells + Zeb1 KO	3
Macrophages + Zeb1 KO + LPS	3
Dendritic cells + Zeb1 KO + LPS	3


a
Cluster (%, no. of genes)
Ikzf1wt/wt
Ikzf1ΔF4/ΔF4
Cluster 1 (29.7, 1877)
Cluster 2 (10.6%, 673)
Cluster 3 (24.1%, 1523)
Cluster 4 (14.7%, 933)
Cluster 5 (20.6%, 1299)
unstimulated
α-IgM
α-IgM+CD40L
mean z-score per cluster
mean z-score
-2 -1 0 1 2
Ikzf1wt/wt
Ikzf1ΔF4/ΔF4
b
Density
upregulated w/ α-IgM
downregulated w/ α-IgM
background
Irf4
Egr3
Myc
Bcl2
Bcl11a
Cd55
Ikzf1ΔF4/ΔF4 log2 fold-change α-IgM vs Unstimulated
Ikzf1wt/wt log2 fold-change α-IgM vs Unstimulated


PLEASE HAVE A GOAL WITH THE FINAL PROJECT
THIS WILL INSTILL A SENSE OF PURPOSE
DO NOT LOSE SIGHT OF THIS
THIS SHOULD BE YOUR COMPASS FORWARD.

Learning Objectives

- Understand how an RNA-Seq experiment is planned and considerations that should be taken by the experimenter and analyzer…

- The same “principals” outlined in this lecture will apply if you decide NOT to analyze an RNA-Seq dataset. You are required to be informed on data preparation all the way to data analysis. **I will ask!**

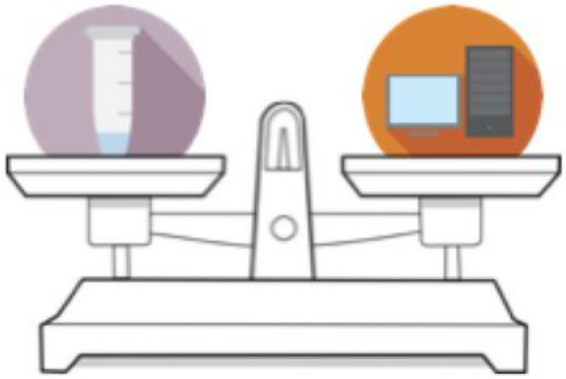
Why?

"The quality of your data is at least directly proportional to the quality of your specimen."

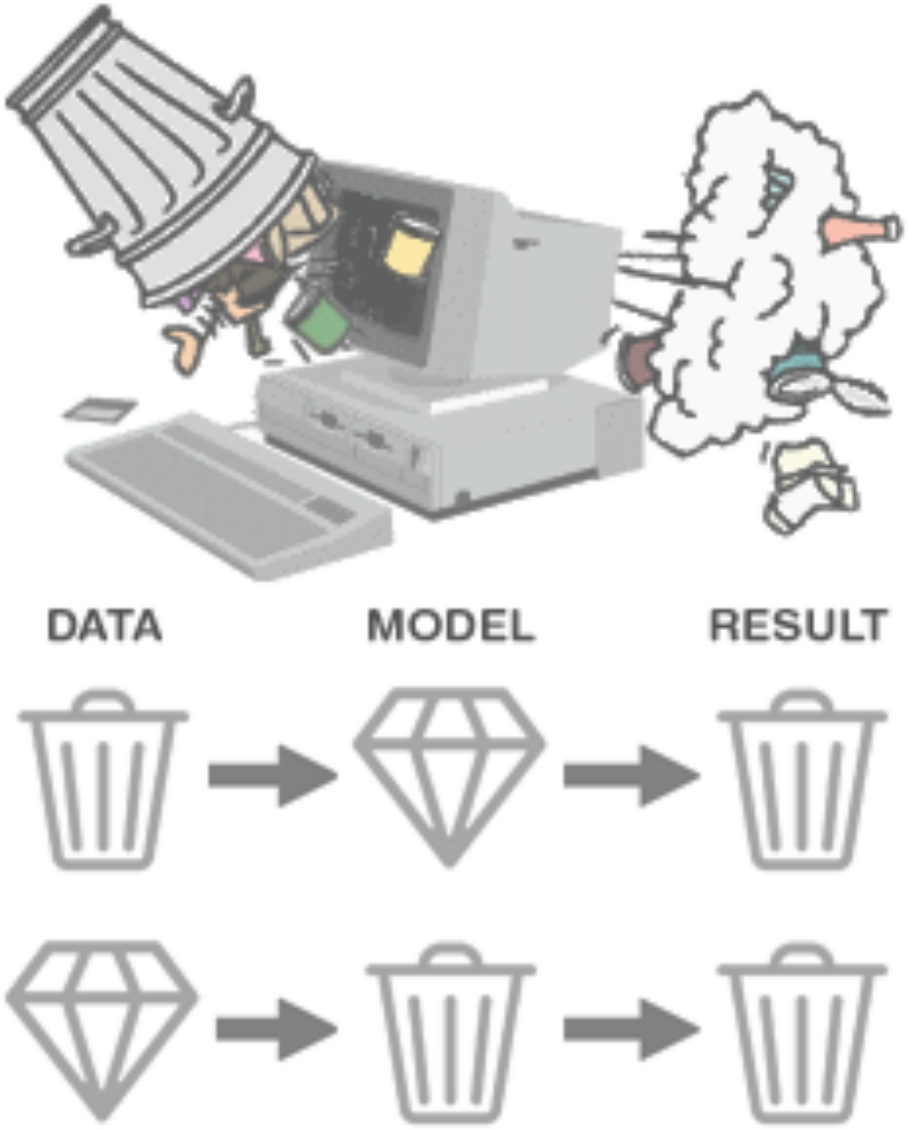
David B. Williams

Transmission Electron Microscopy: A Textbook for Materials Science

ISBN 978-0-387-76501-3



Garbage In, Garbage Out



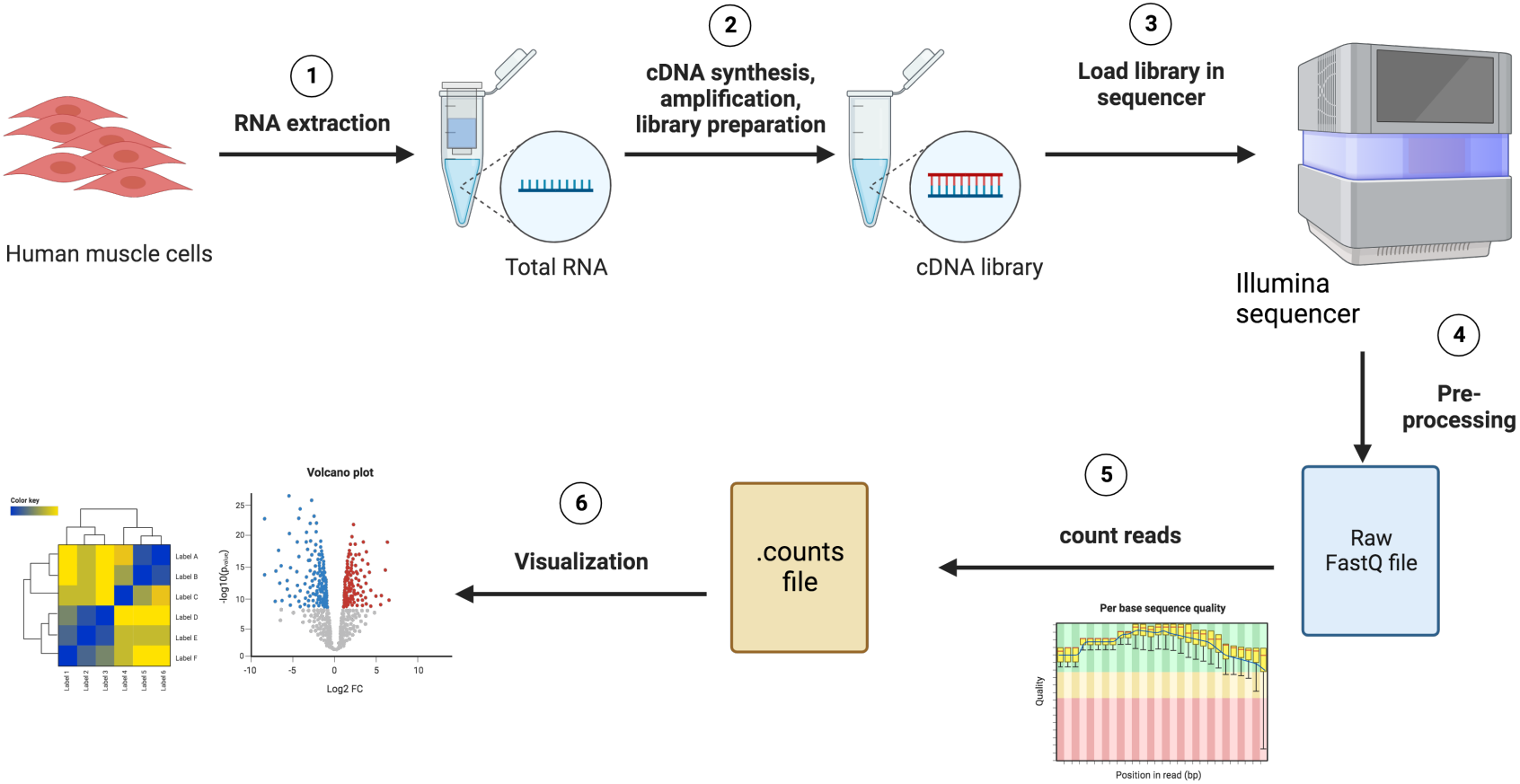
Your input will define the quality of output you get!

RNA-Sequencing (RNA-Seq)

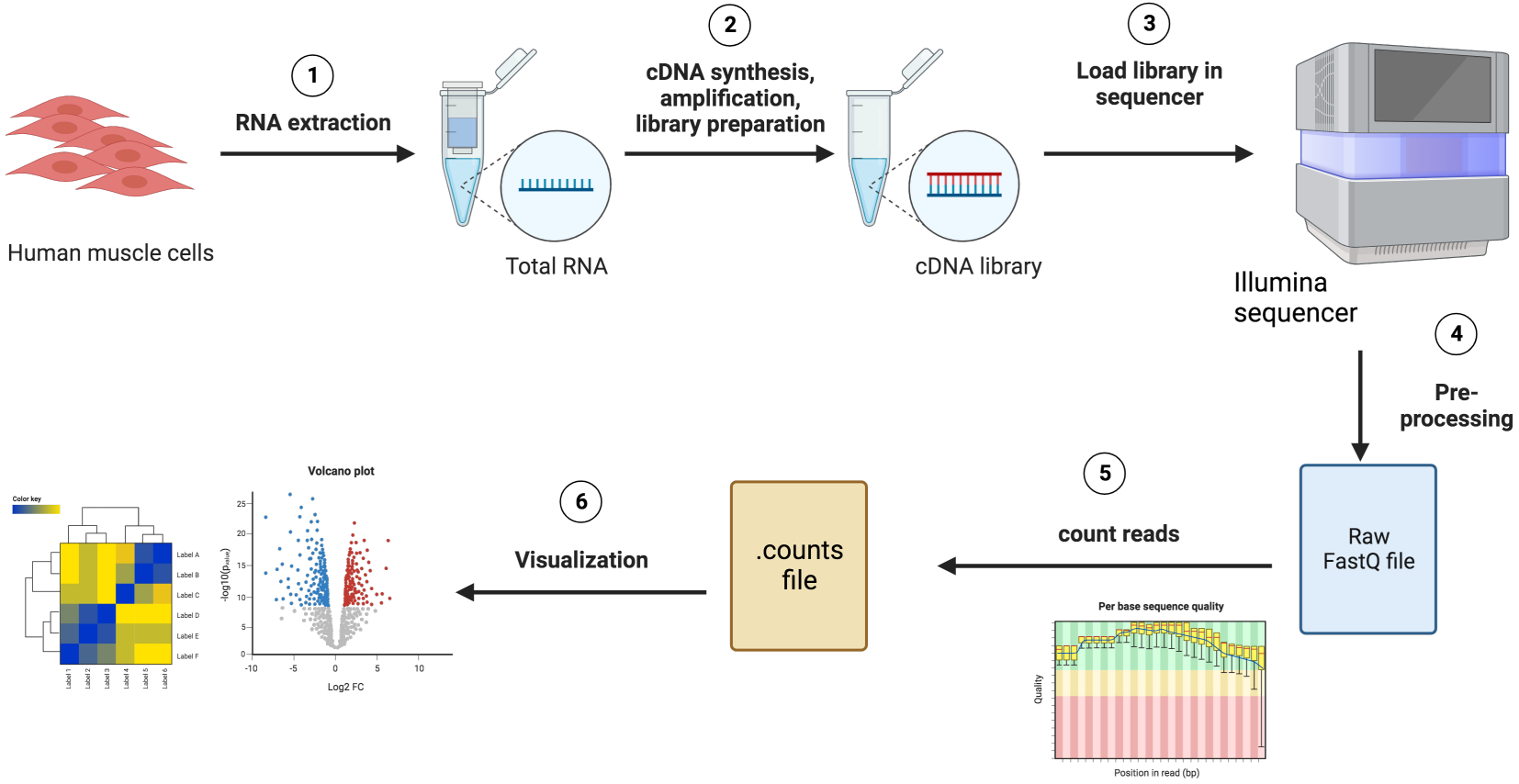
Technique used to explore and/or quantify gene expression within or between conditions of an organism

How do we get a comprehensive collection of all transcripts being transcribed within the cell?

RNA-Seq pipeline



RNA-Seq pipeline



Caveat

- Transcriptome is the collection of all transcript readouts present in a cell.

- RNA-seq can be used to explore and/or quantify the transcriptome of an organism, which can be utilized for many different ***types*** of experiments:

Caveat

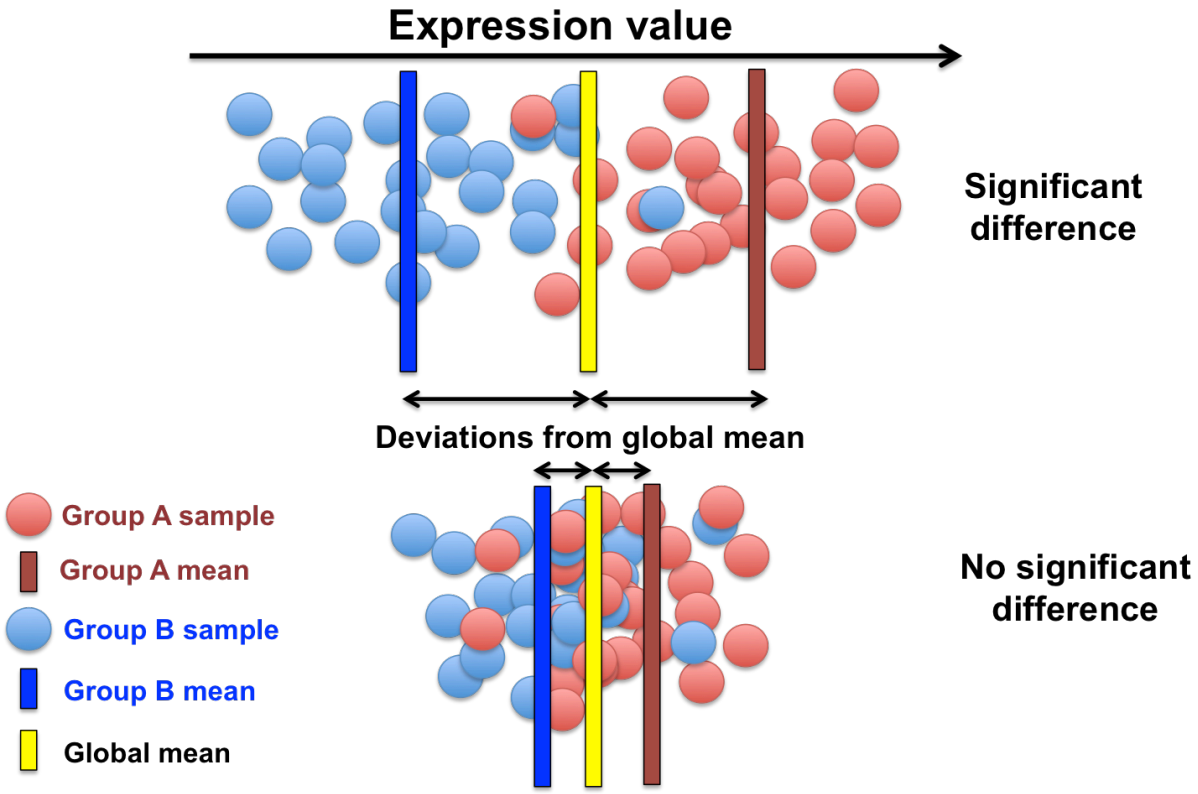
- Differential Gene Expression: quantitative evaluation and comparison of transcript levels across treatments or groups

- Transcriptome assembly: building the profile of transcribed regions of the genome

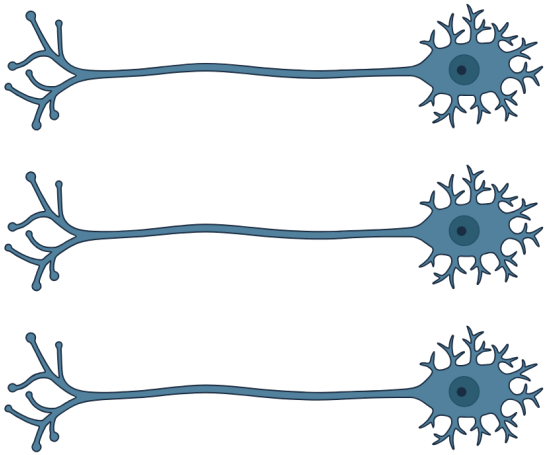
- Meta-transcriptomics

Basic types of questions answered:

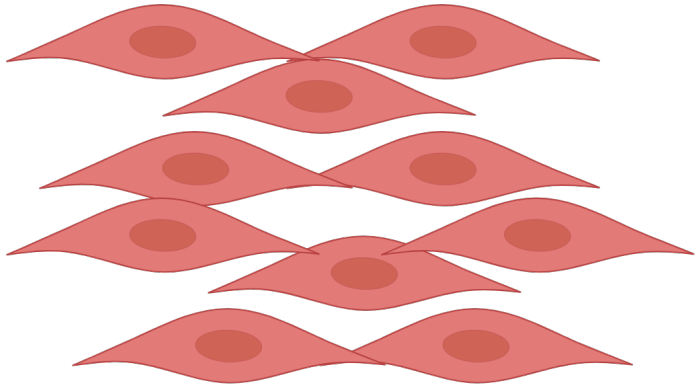
What genes are differentially expressed between conditions?



Nerve Cell



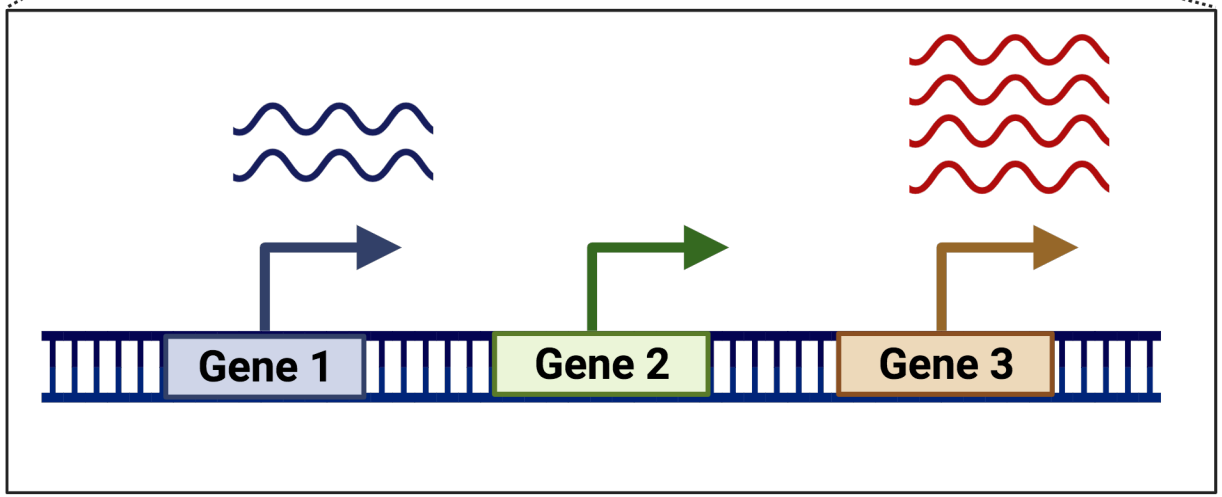
Muscle Cell



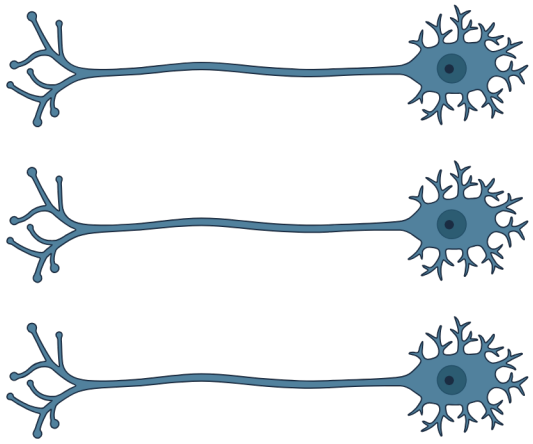
What drives the difference in cellular phenotype between nerve and muscle cells?

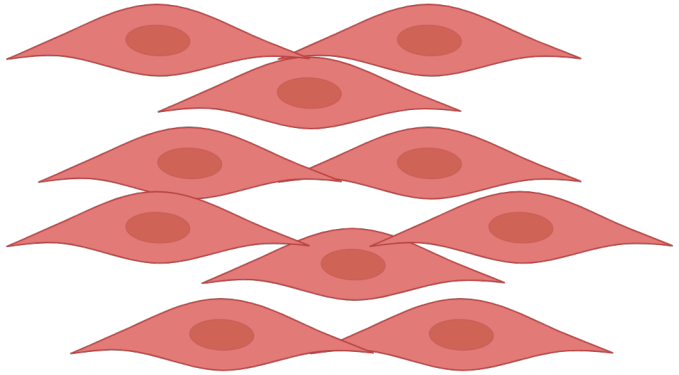
Nerve Cell

Muscle Cell



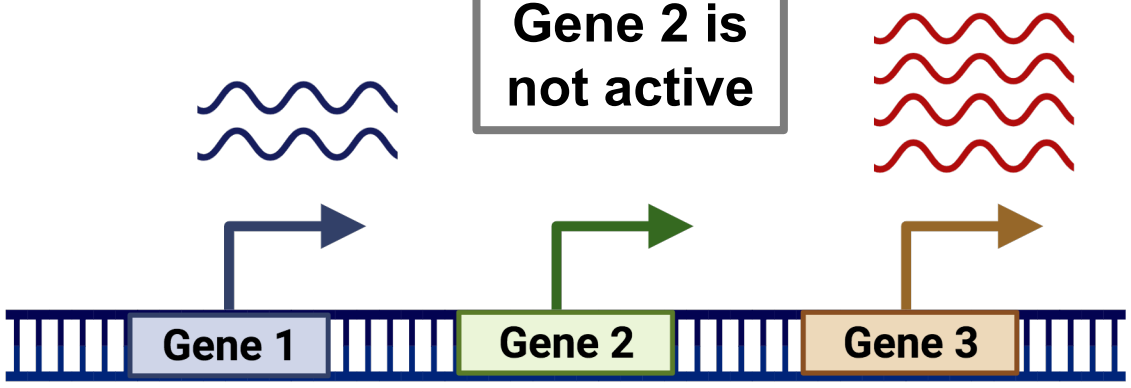
Some genes are more active than others

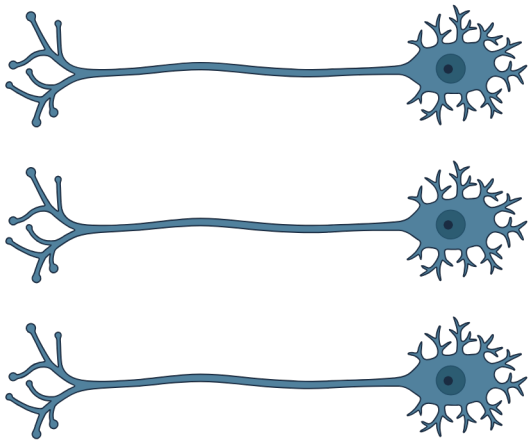


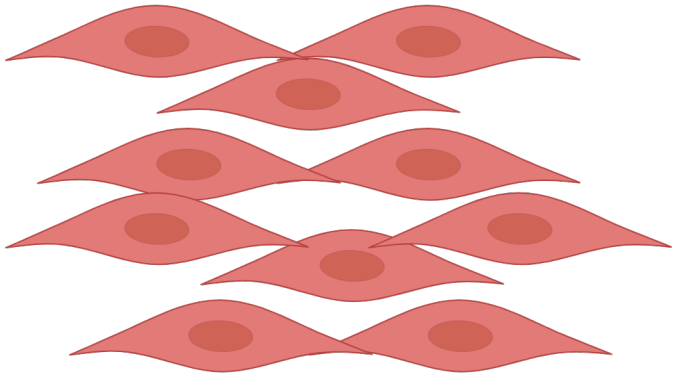


Gene 3 is most active

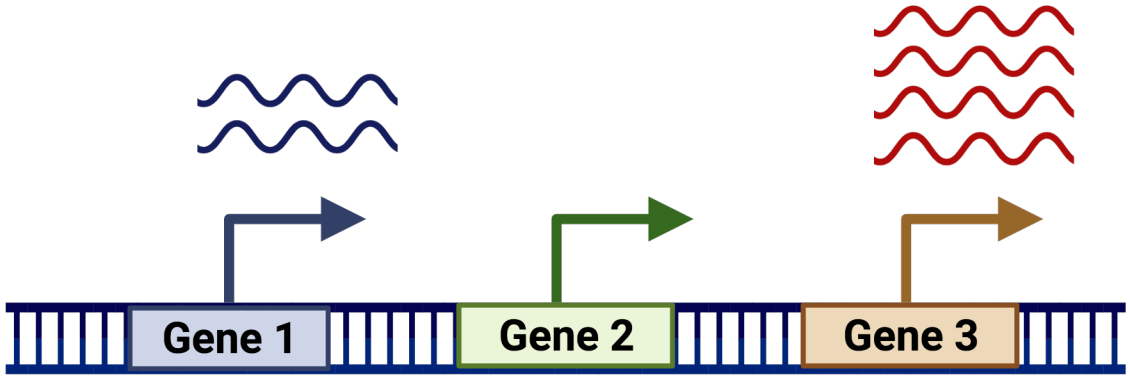
Gene 2 is not active

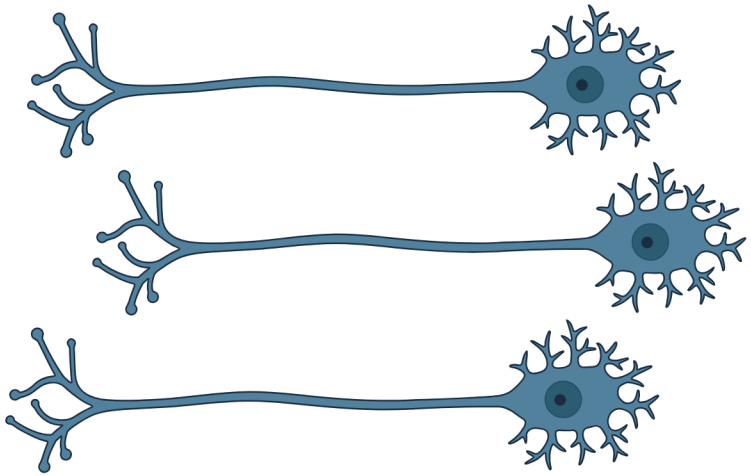


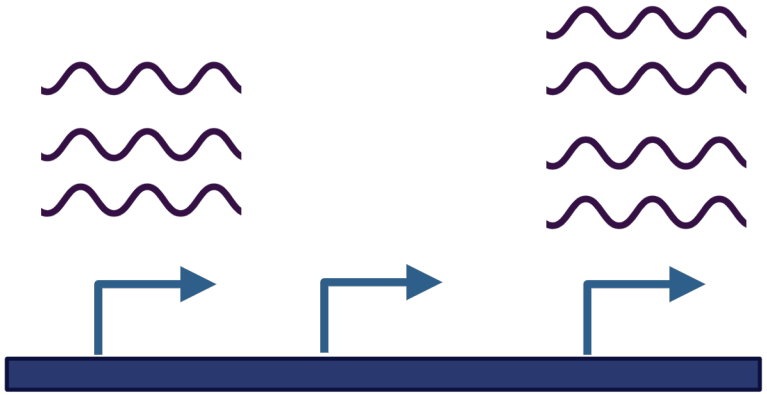




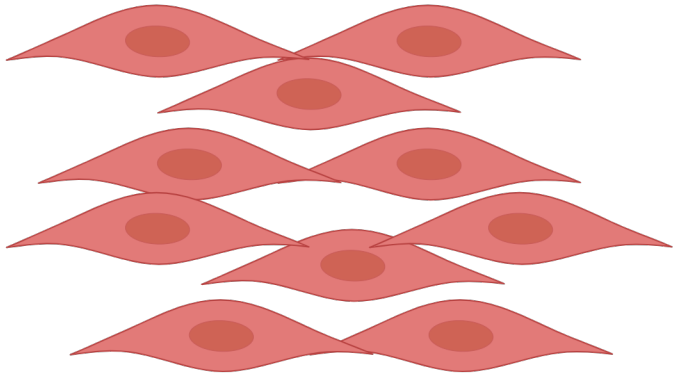
RNA-Seq tell us which genes are active and how much they are being transcribed

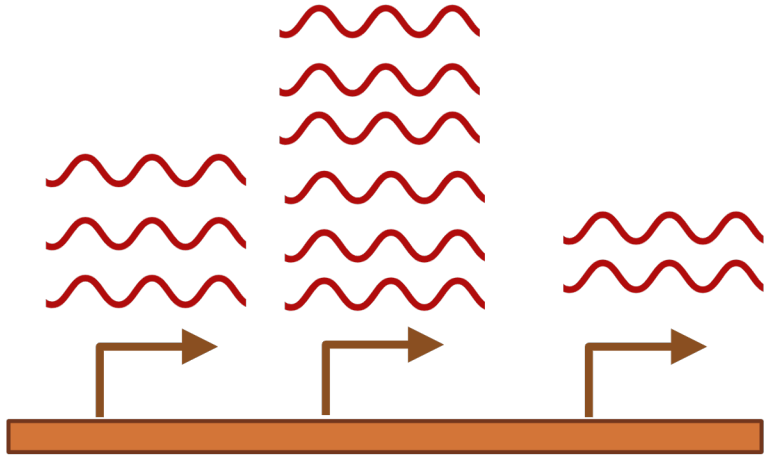




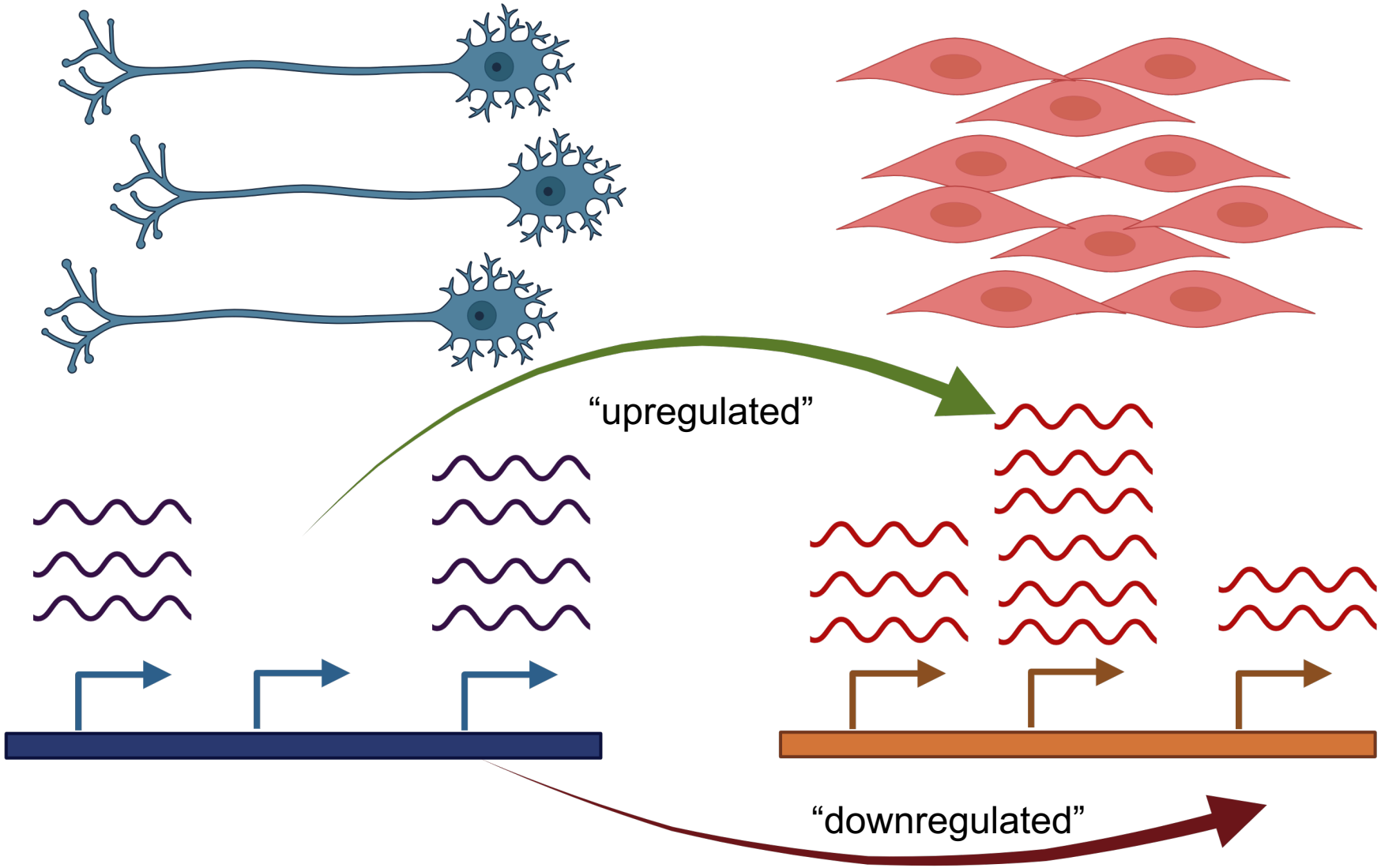


We can use RNA-Seq to measure gene expression in nerve cells



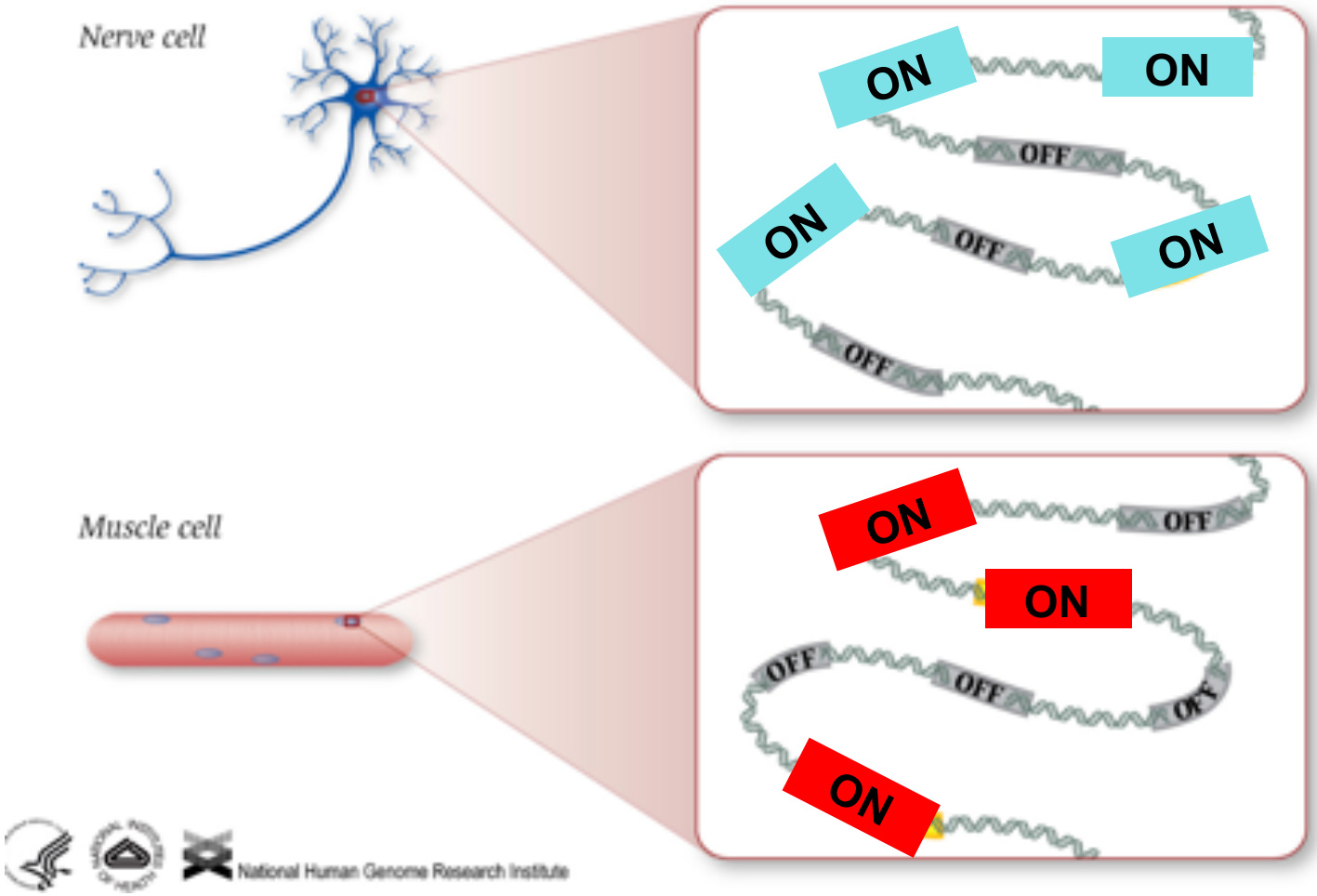


….and then use it to measure gene expression in muscle cells



Then we can compare the two cell types bioinformatically to find what is different

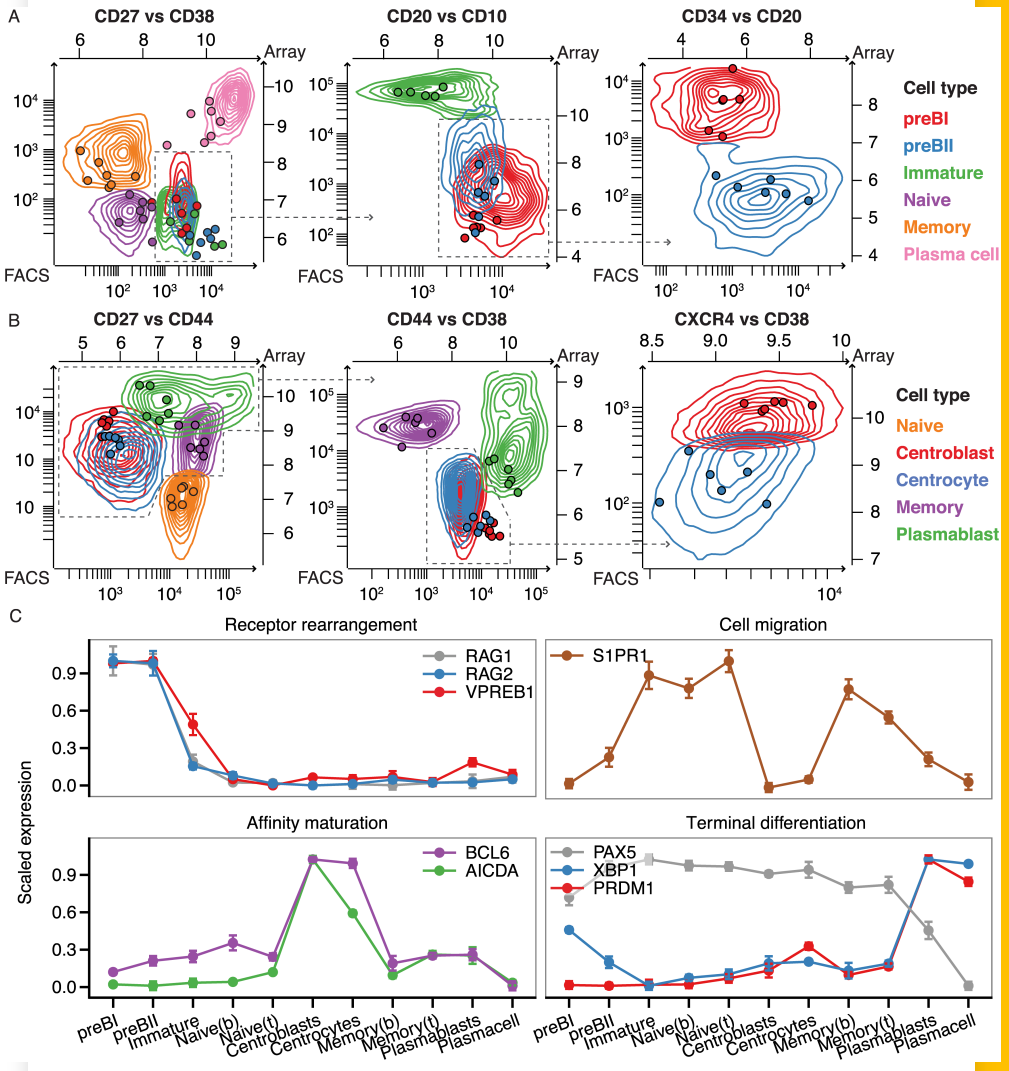
Ultimately, we would find a different set of genes are turned ON for each cell type, resulting in different RNAs and proteins produced.



Other questions answered:

Are there any trends in gene expression across development?

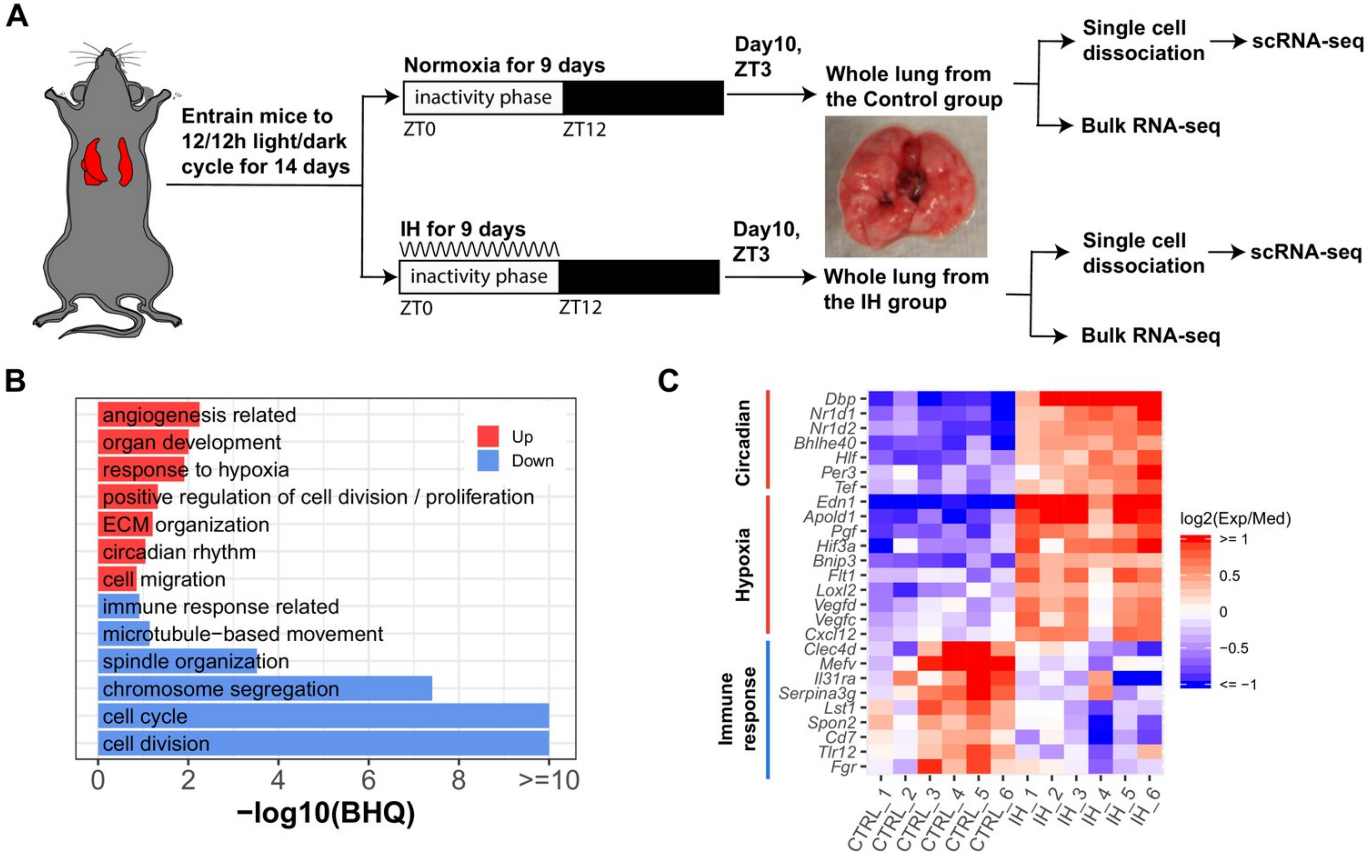
Which groups of genes change similarly over time or across conditions?



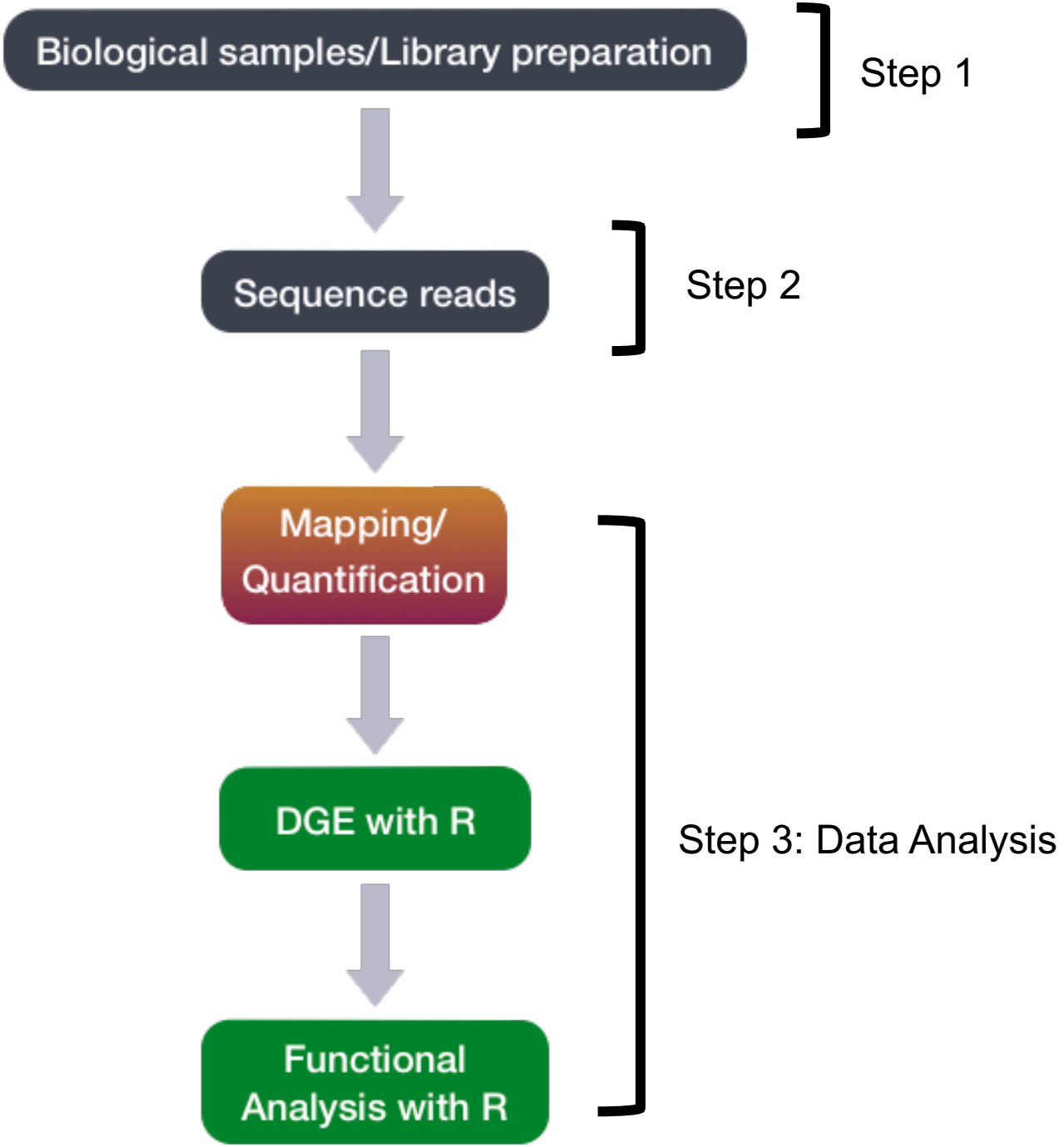
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0138236

Basic types of questions answered:

What processes or pathways are enriched in condition of interest?



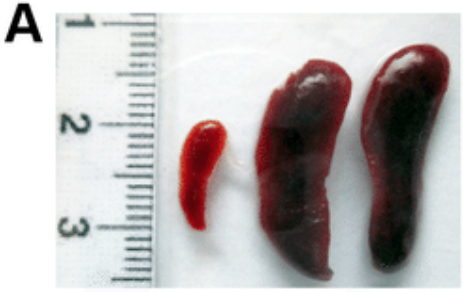
https://elifesciences.org/articles/63003


Biological samples/Library preparation
Step 1
Sequence reads
Step 2
Mapping/
Quantification
DGE with R
Functional
Analysis with R
Step 3: Data Analysis

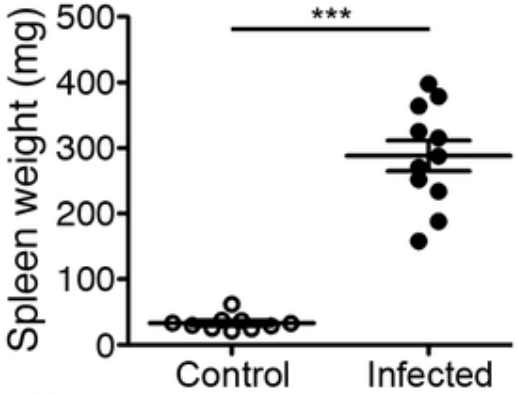
Experimental workflow

1

Samples of interest

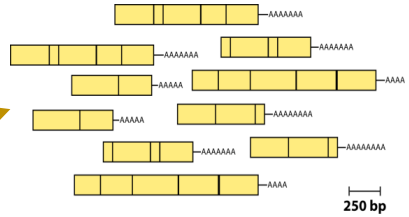


PMID: 27548618



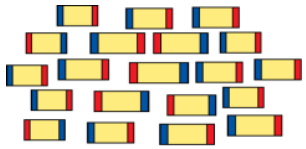
2

Isolate RNAs

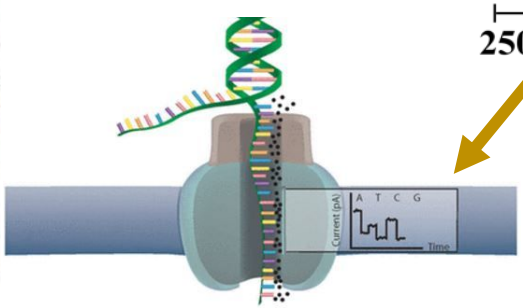


3

Library build

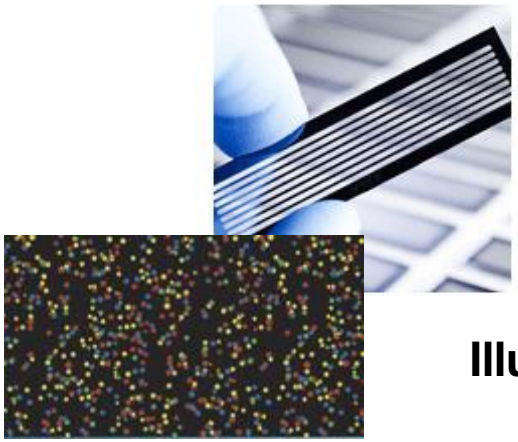


250 bp

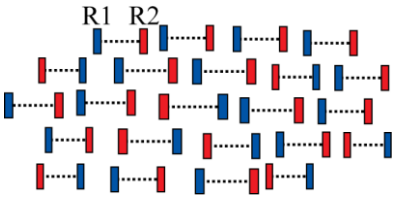


Illumina sequencing versus ONP

4

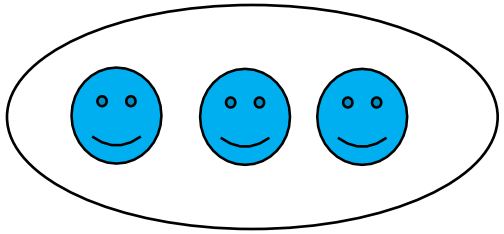


5

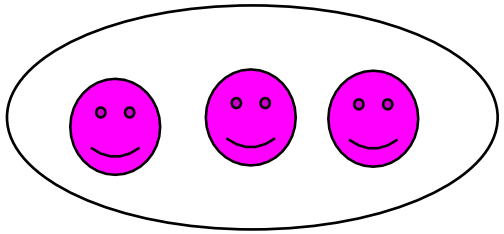


Reads (R1 and R2) generated

Biological Replicates



Condition 1

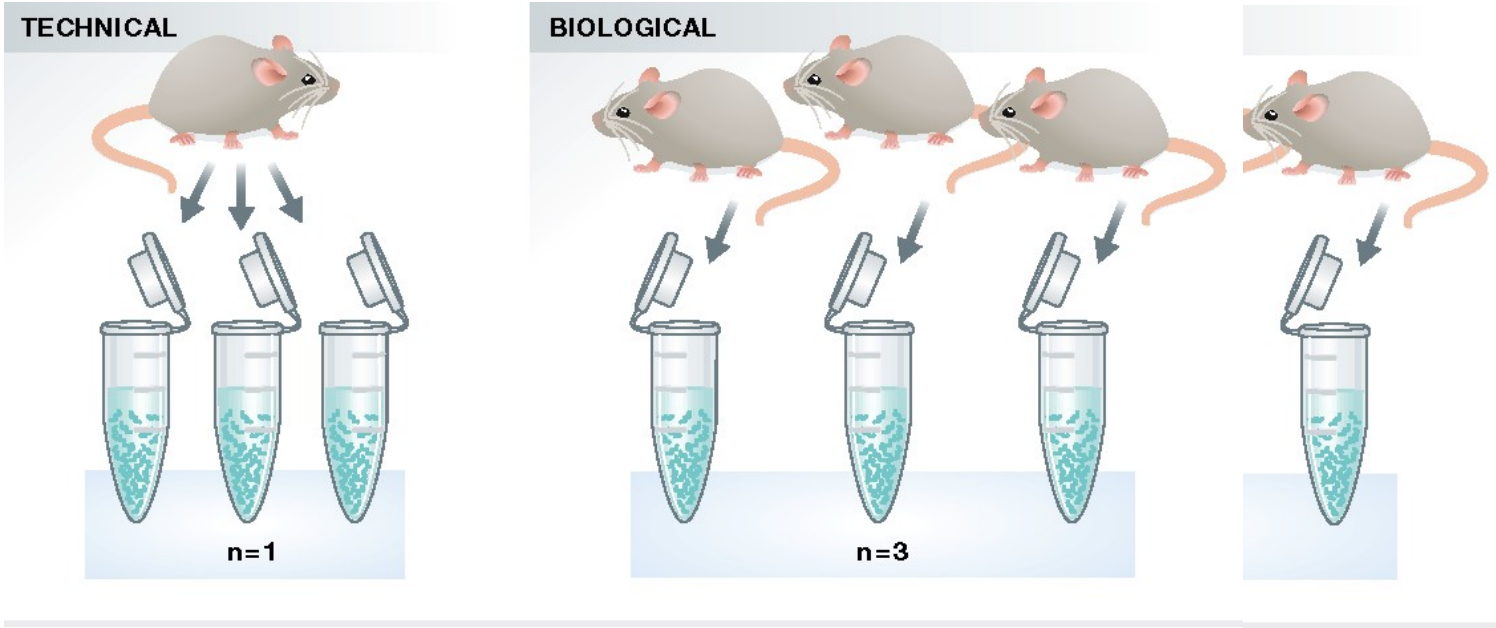


Condition 2

- To detect Differentially Expressed Genes (DEGs) between groups we should have several samples, which are also known as biological replicates

Biological Replicates

- We are not talking about technical replicates
- Assessing biological variation requires biological replicates
- Three is the standard minimum
- Yet I would recommend four or more

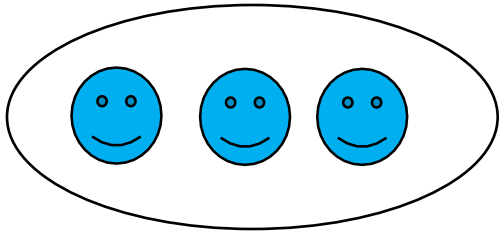


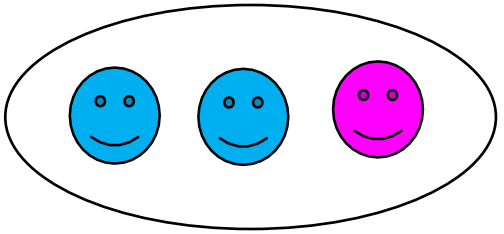
Probability of detecting DEGs

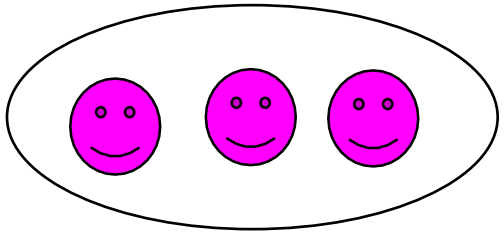
	Replicates per group		
	3	5	10
Fold change			
2	87%	98%	100%

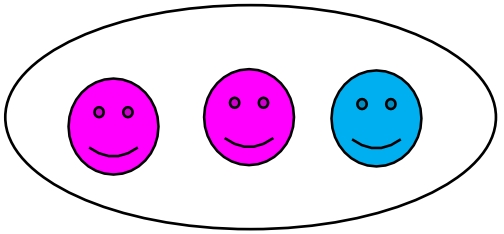
PMID: 26813401

Grouping of Replicates





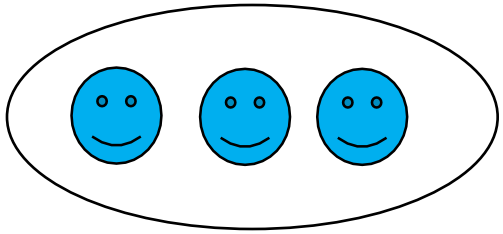


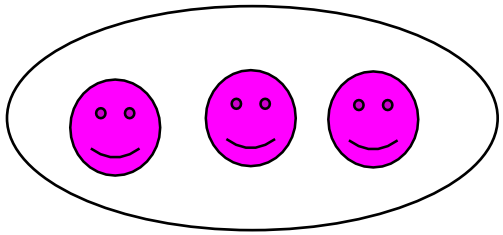


What you want

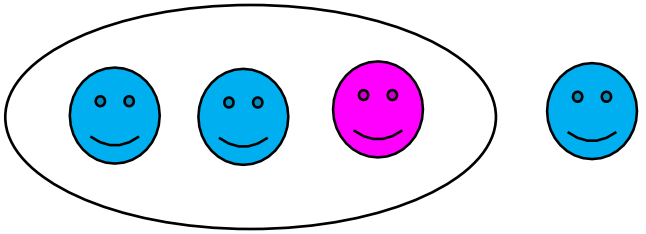
What you get

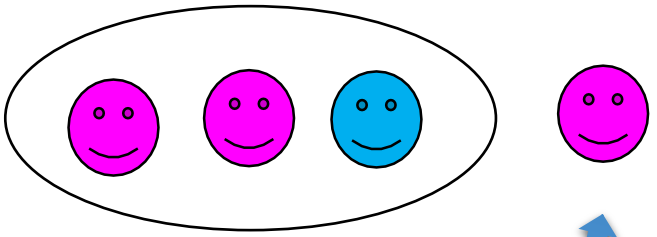
Grouping of Replicates





What you want



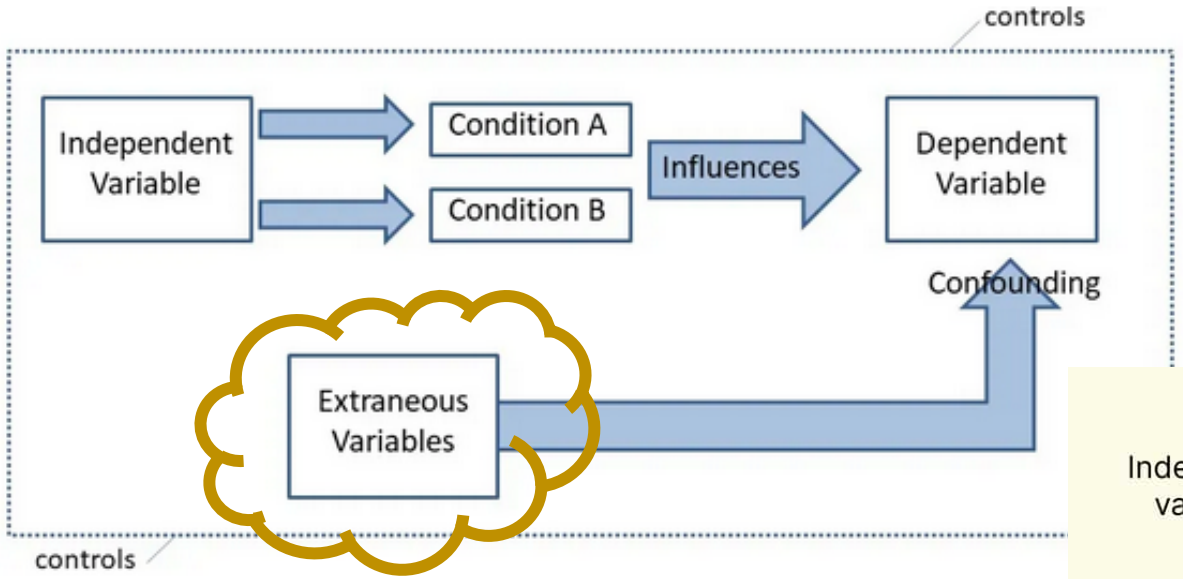


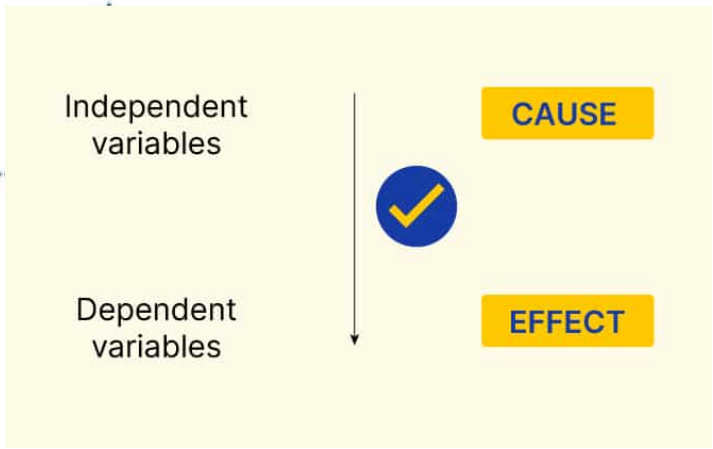
What you get

That spare comes in handy
Highly recommend especially with mice!

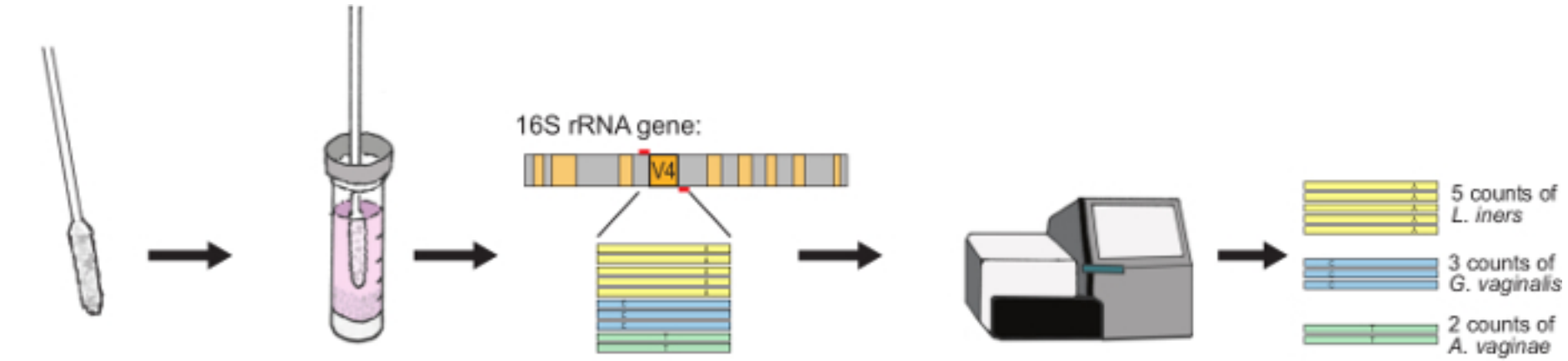
What causes this? Confounding variables

A variable that influences or *confounds* the relationship between an independent and dependent variable



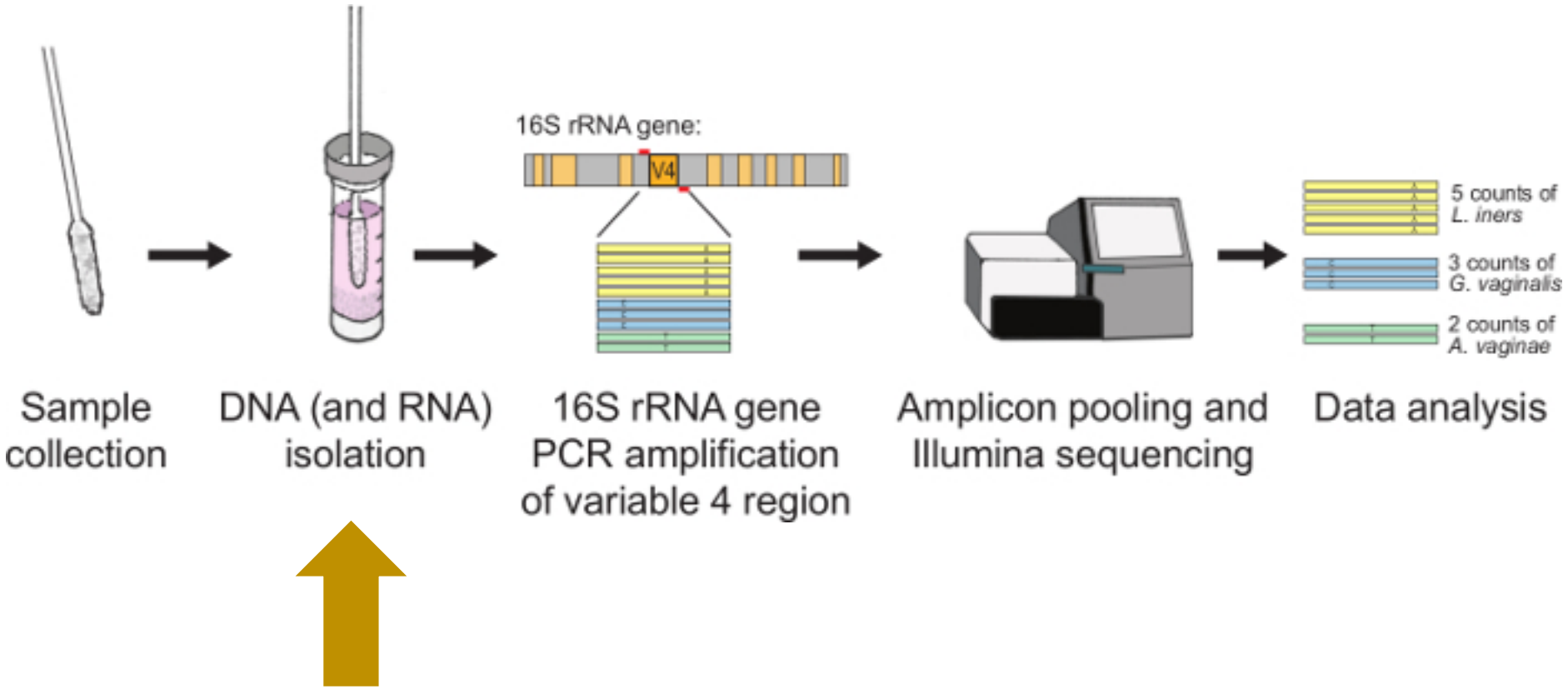


Examples of confounding variables



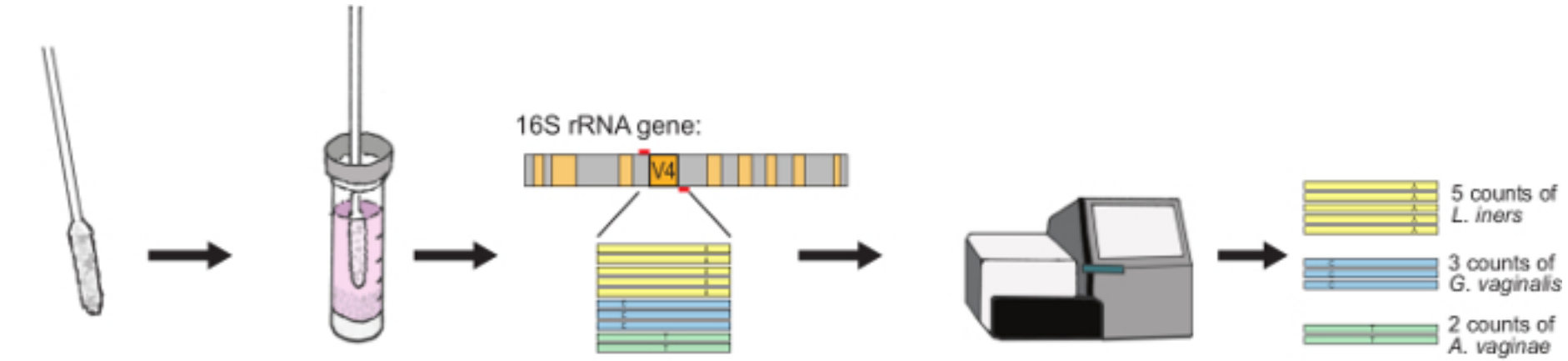
A new technician is running the sequencer

Examples of confounding variables



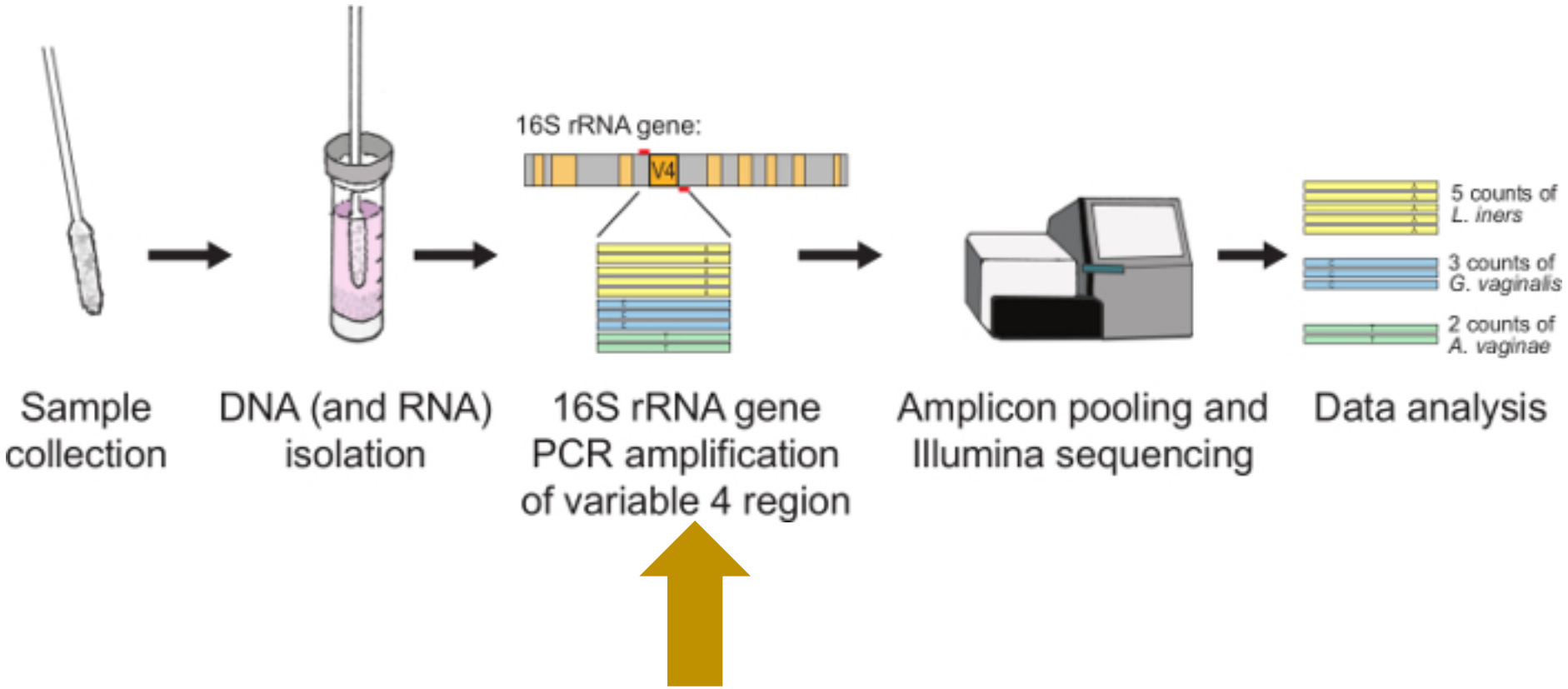
Extracting DNA/RNA with two different kits!

Examples of confounding variables



Sequencing on multiple different types of platforms

Examples of confounding variables



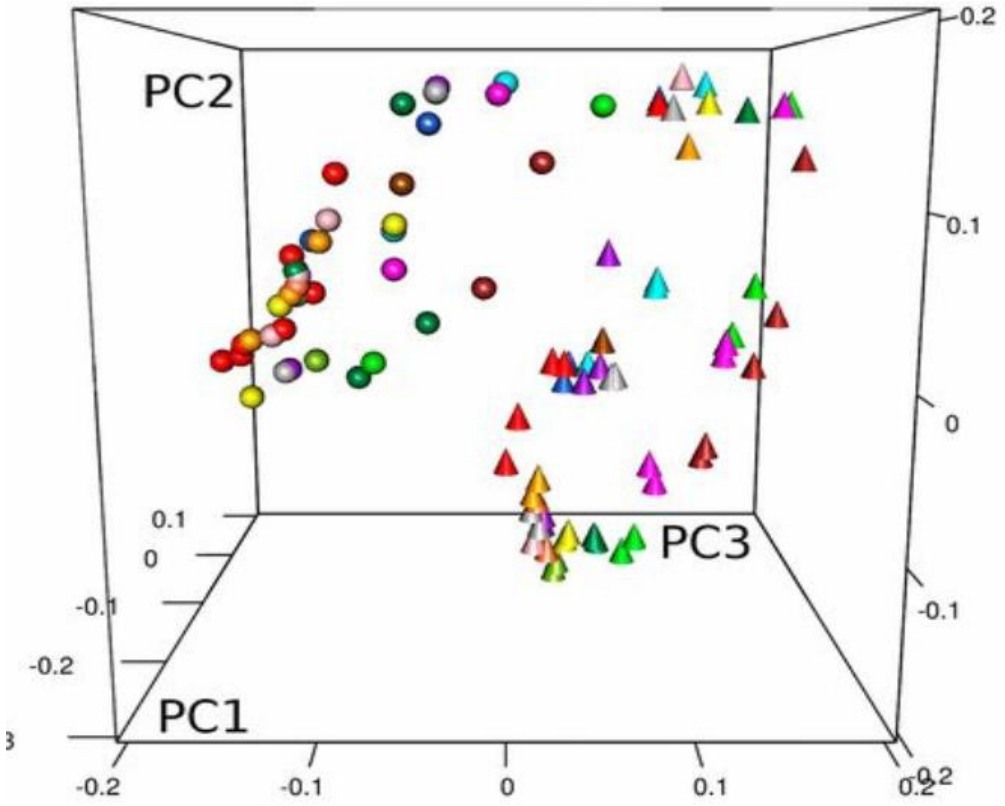
Inappropriate multiplexing strategy

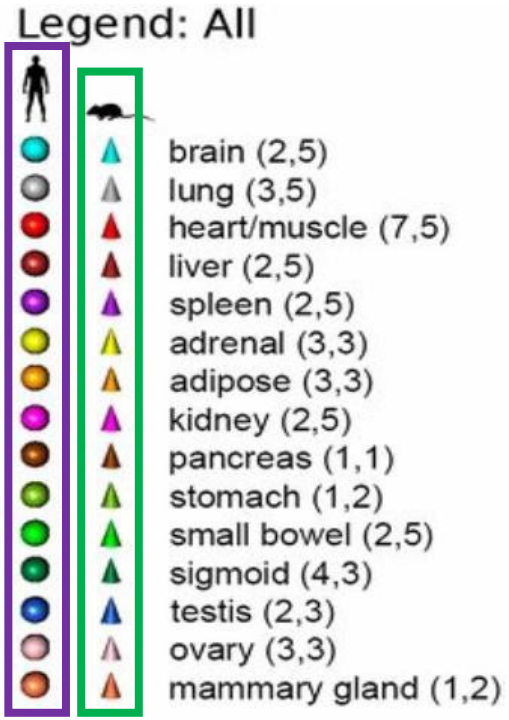
What this means?

- Its sometimes impossible for bioinformaticians to partition biological variation from technical variation, when these two sources of variation ***are confounded***.
- No amount of statistical sophistication can separate confounded factors after data have been collected.
- *…these confounding variables may or may not be in your control!*

A well-planned experiment with an additional sample, does end up saving you time and money down the road

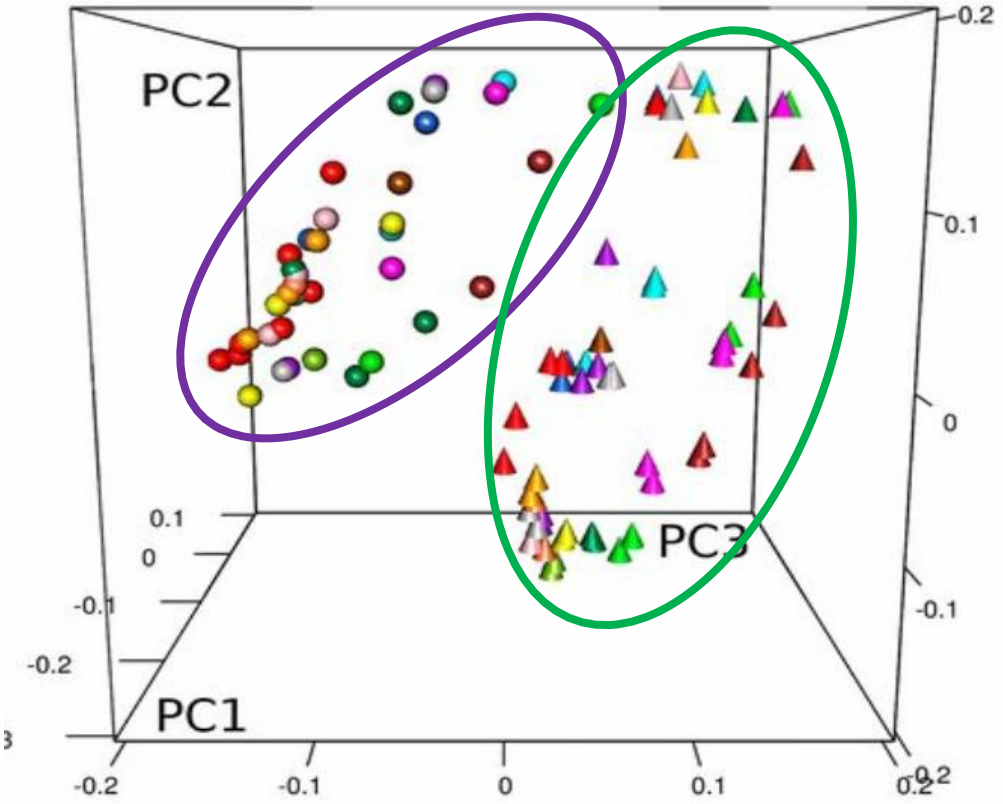
ENCODE reported that gene expression was likely to follow a species-specific rather tissue-specific pattern

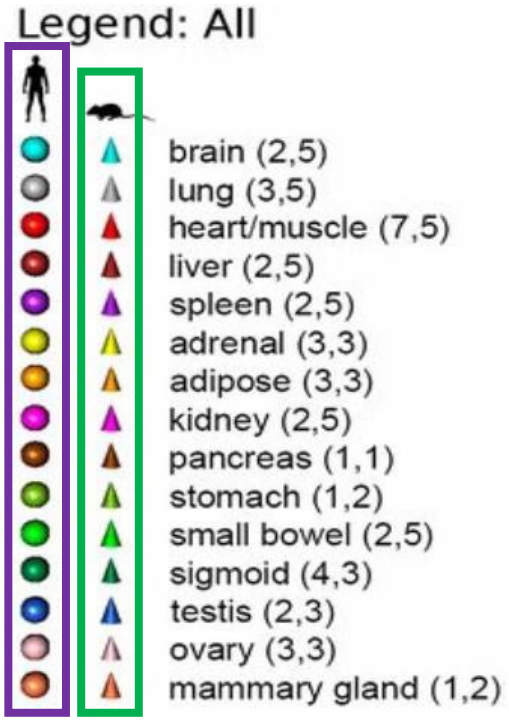




https://doi.org/10.1073/pnas.1413624111

ENCODE reported that gene expression was likely to follow a species-specific rather tissue-specific pattern



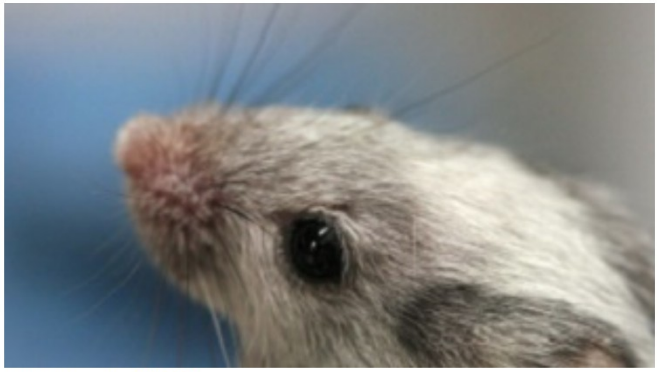


https://doi.org/10.1073/pnas.1413624111

TheScientist
EXPLORING LIFE, INSPIRING INNOVATION

Reanalysis of Mouse ENCODE data suggests mouse and human genes are expressed in tissue-specific, rather than species-specific, patterns.

May 19, 2015
JYOTI MADHUSOODANAN



WIKIMEDIA, RAMA

Late last year, members of the Mouse ENCODE consortium reported in *PNAS* that, across a wide range of tissues, gene expression was more likely to follow a species-specific rather than tissue-specific pattern. For example, genes in the mouse heart were expressed in a pattern more similar to that of other mouse tissues, such as the brain or liver, than the human heart.

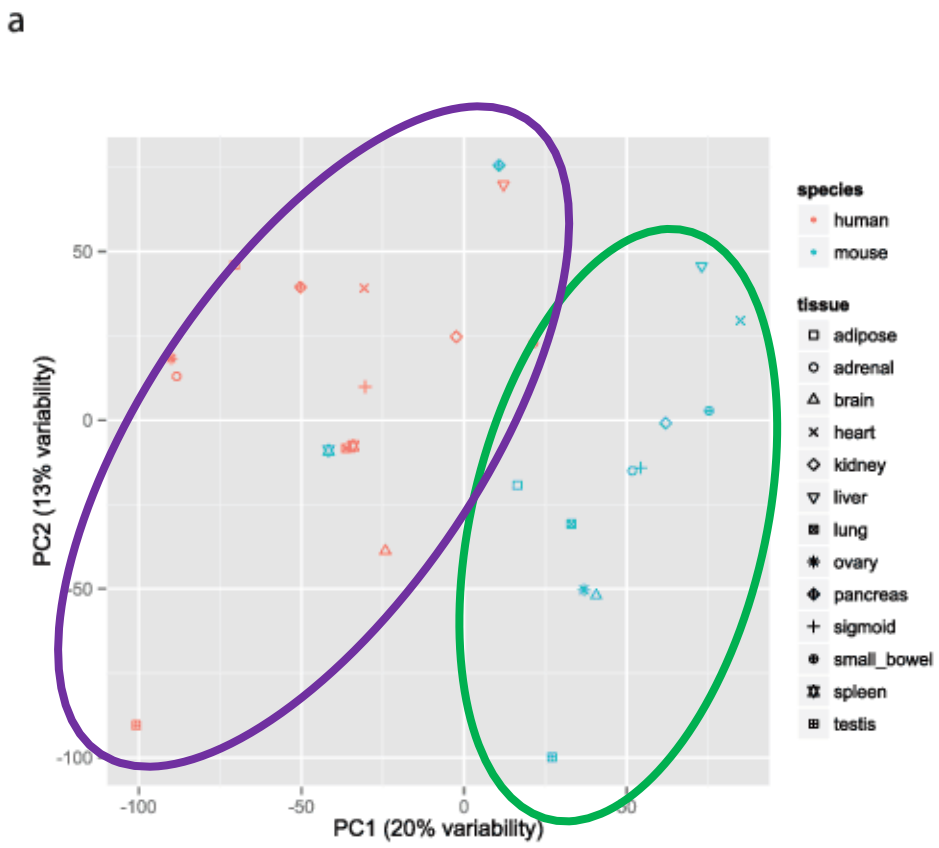
But earlier this month, Yoav Gilad of the University of Chicago called these results into question on Twitter. With a dozen or so 140-character dispatches (including three heat maps), Gilad suggested the results published in *PNAS* were an anomaly—a result of how the tissue samples were sequenced in different batches. If this "batch effect" was eliminated, he proposed, mouse and human tissues clustered in a tissue-specific manner, confirming previous results rather than supporting the conclusions reported by the Mouse ENCODE team.

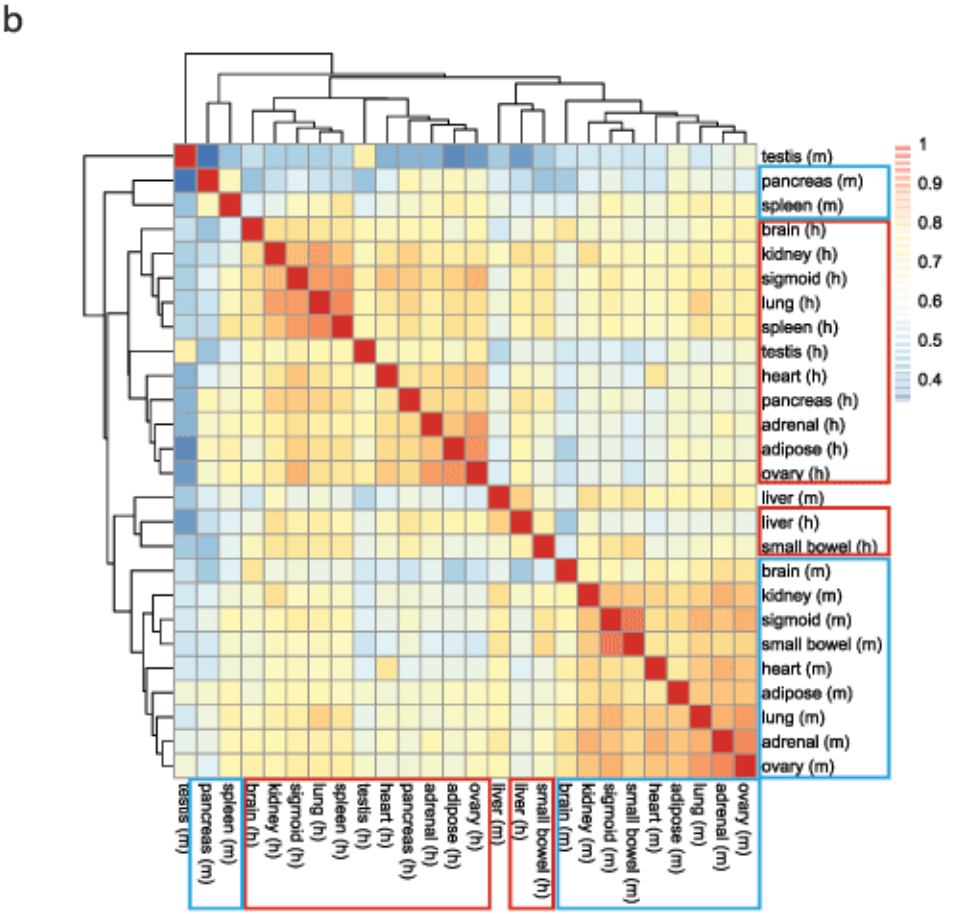
Sequence study design (sequencer ID, run ID, lane number):

D87PMJN1 (run 253, lane 7)	D87PMJN1 (run 253, lane 8)	D4LHBFN1 (run 276, lane 4)	MONK (run 312, lane 6)	HWI-ST373 (run 375, lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	● human
testis		pancreas		● mouse

Sequencing lane (a batch effect) was almost completely confounded with species in the PNAS study. From @Y_Gilad

Before accounting for batch effect

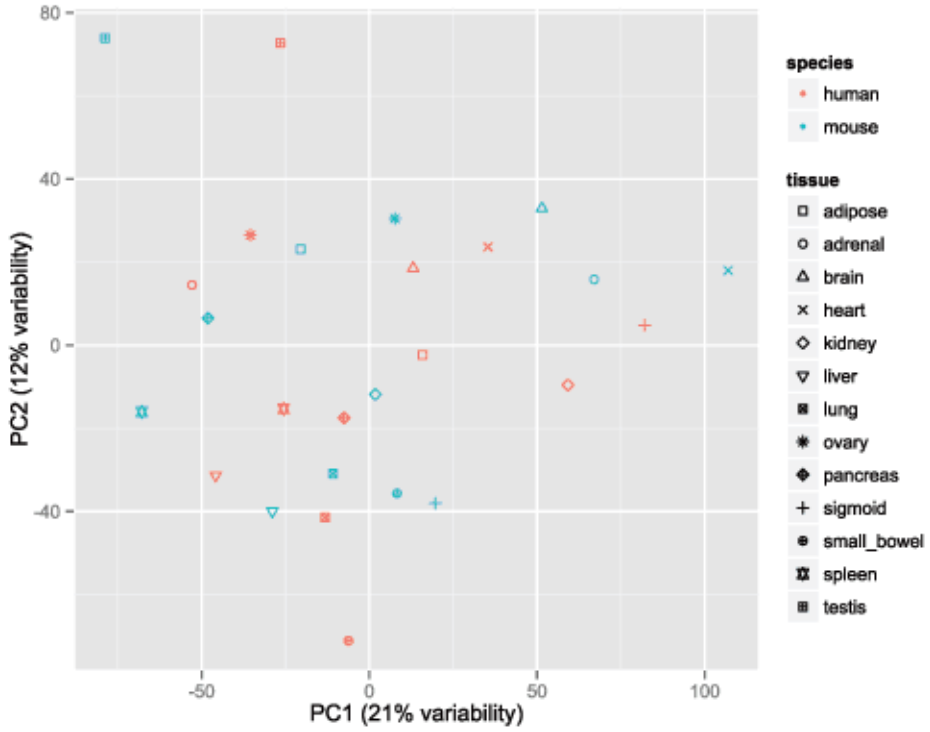




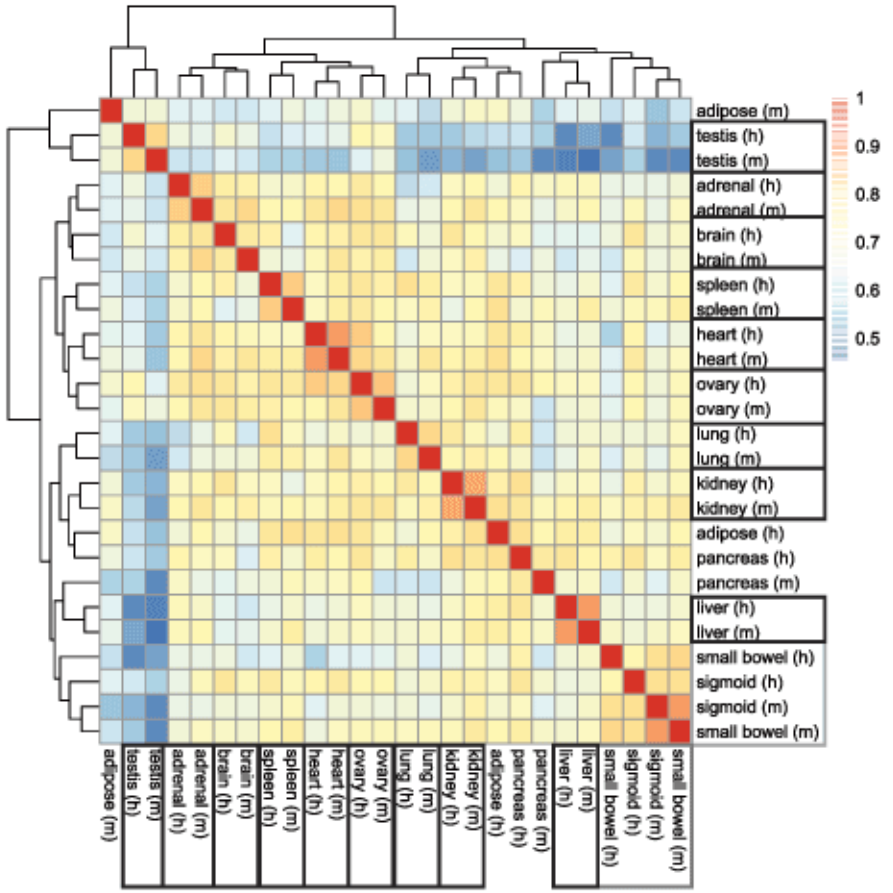
Samples grouped by animal

After accounting for batch effect

a



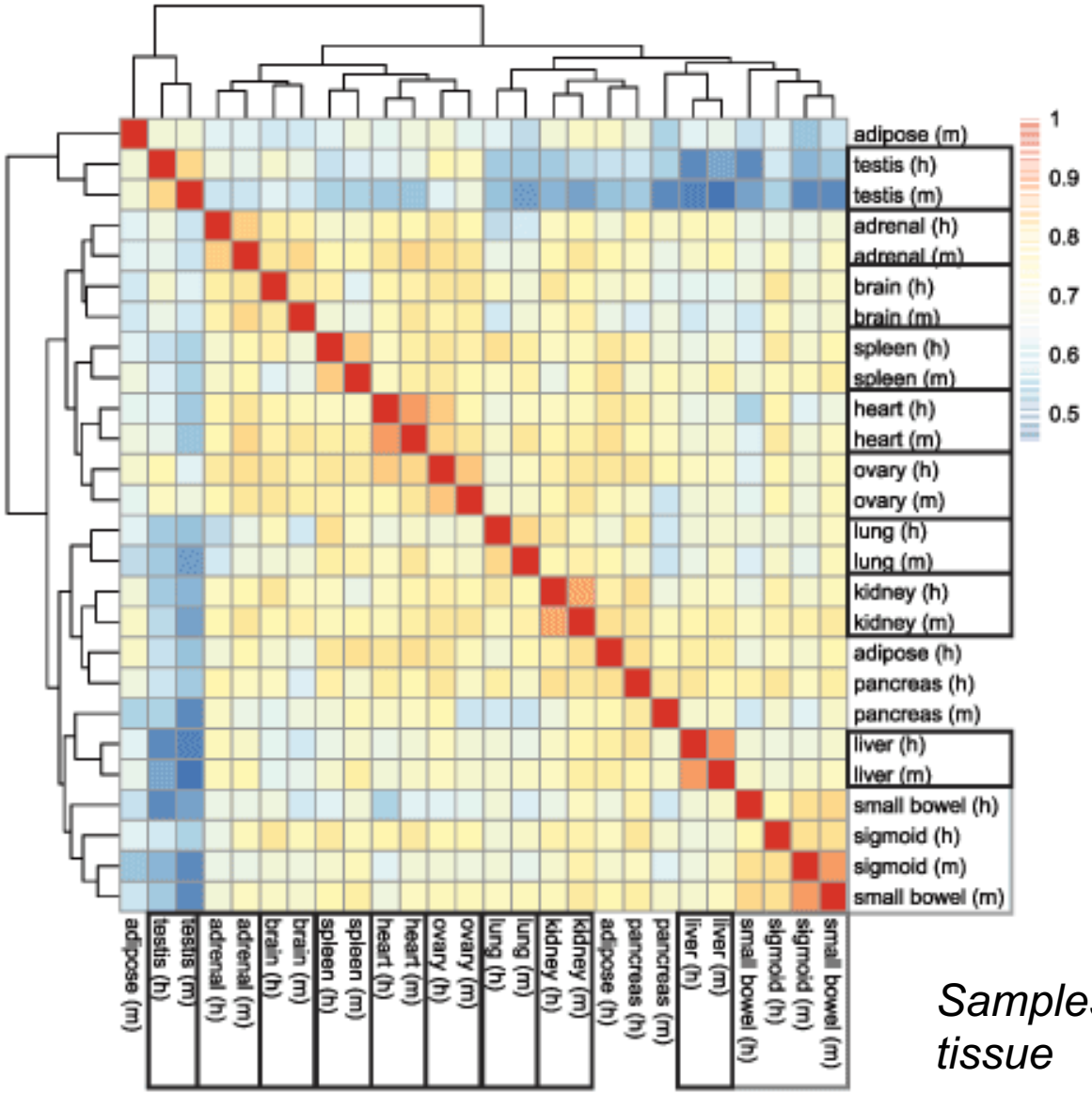
b



Samples now grouped by tissue

b

After accounting for batch effect

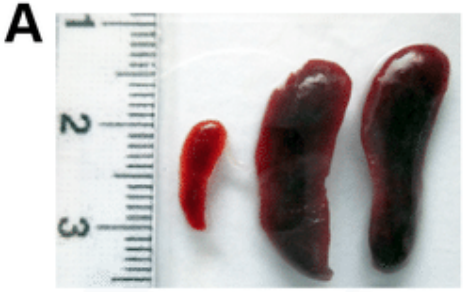


Samples now grouped by tissue

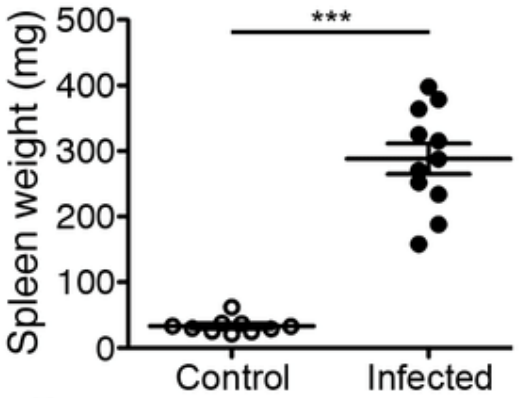
Experimental workflow

1

Samples of interest

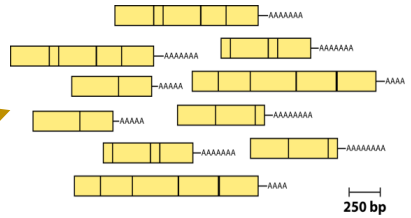


PMID: 27548618



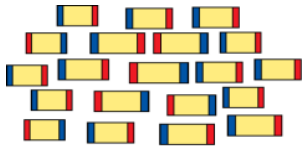
2

Isolate RNAs

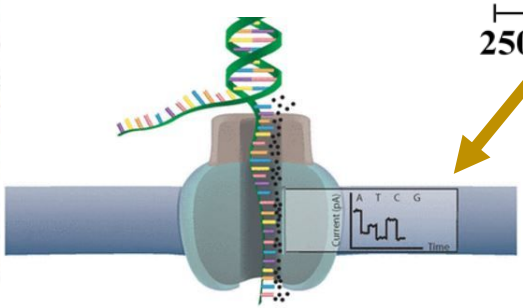


3

Library build

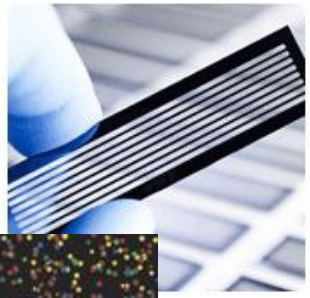


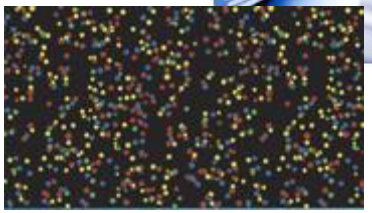
250 bp



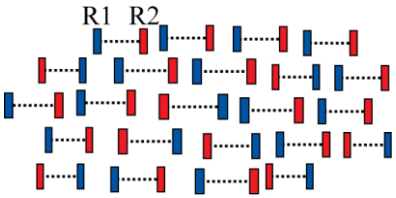
Illumina sequencing versus ONP

4





5



Reads (R1 and R2) generated

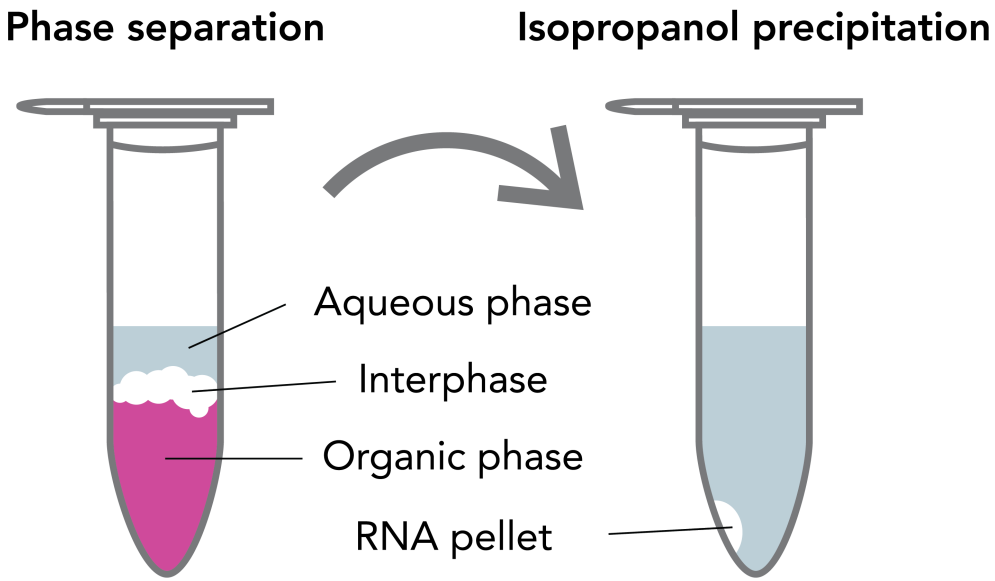
Working with RNA

- RNA-Seq is dependent on the isolation of pure RNA
- RNA is more labile than DNA so extra precautions should be taken

Isolation of RNA

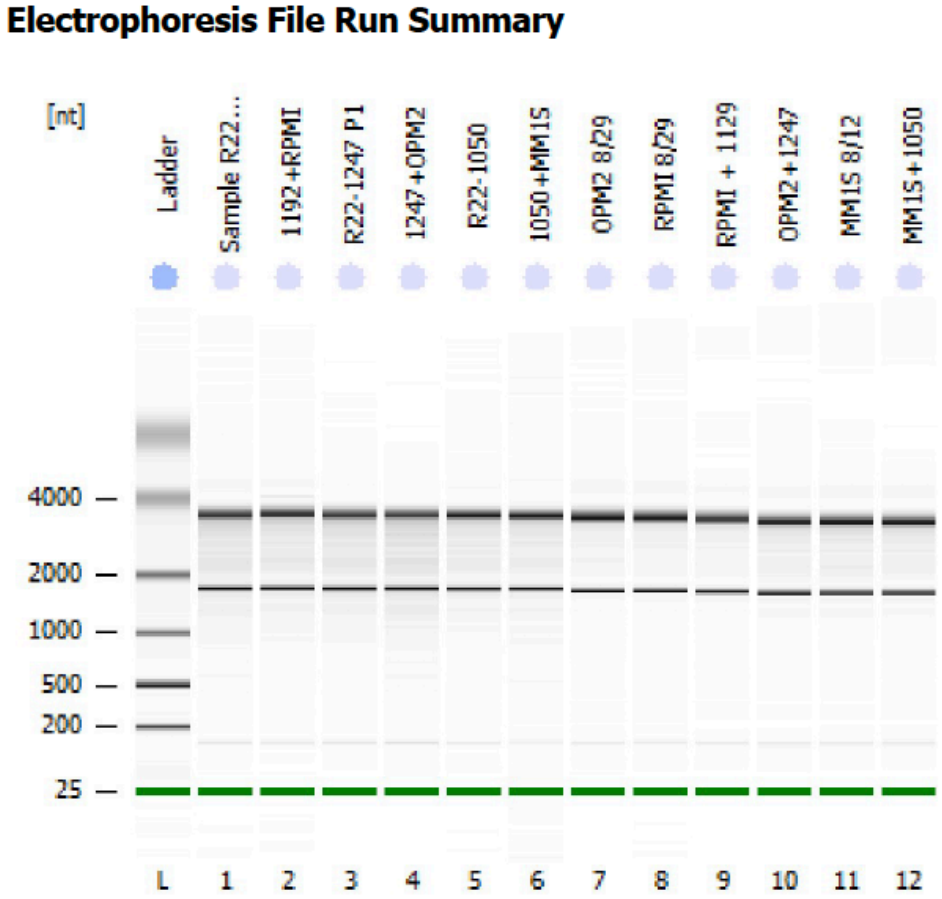
- This kit comes everything you need for total RNA isolation including the gDNA eliminator spin columns

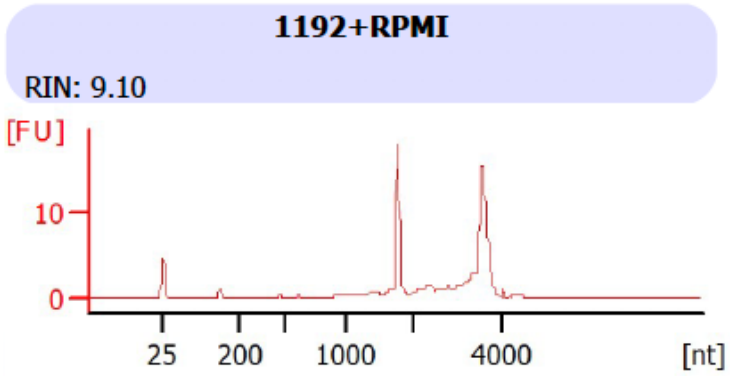
A cheaper way to isolate RNA

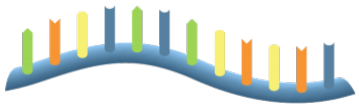


- Is not recommended as residual phenol will inhibit the library build and you cannot get rid of it.

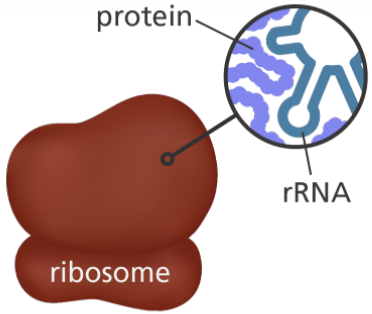
Check purity with RNA bioanalyzer



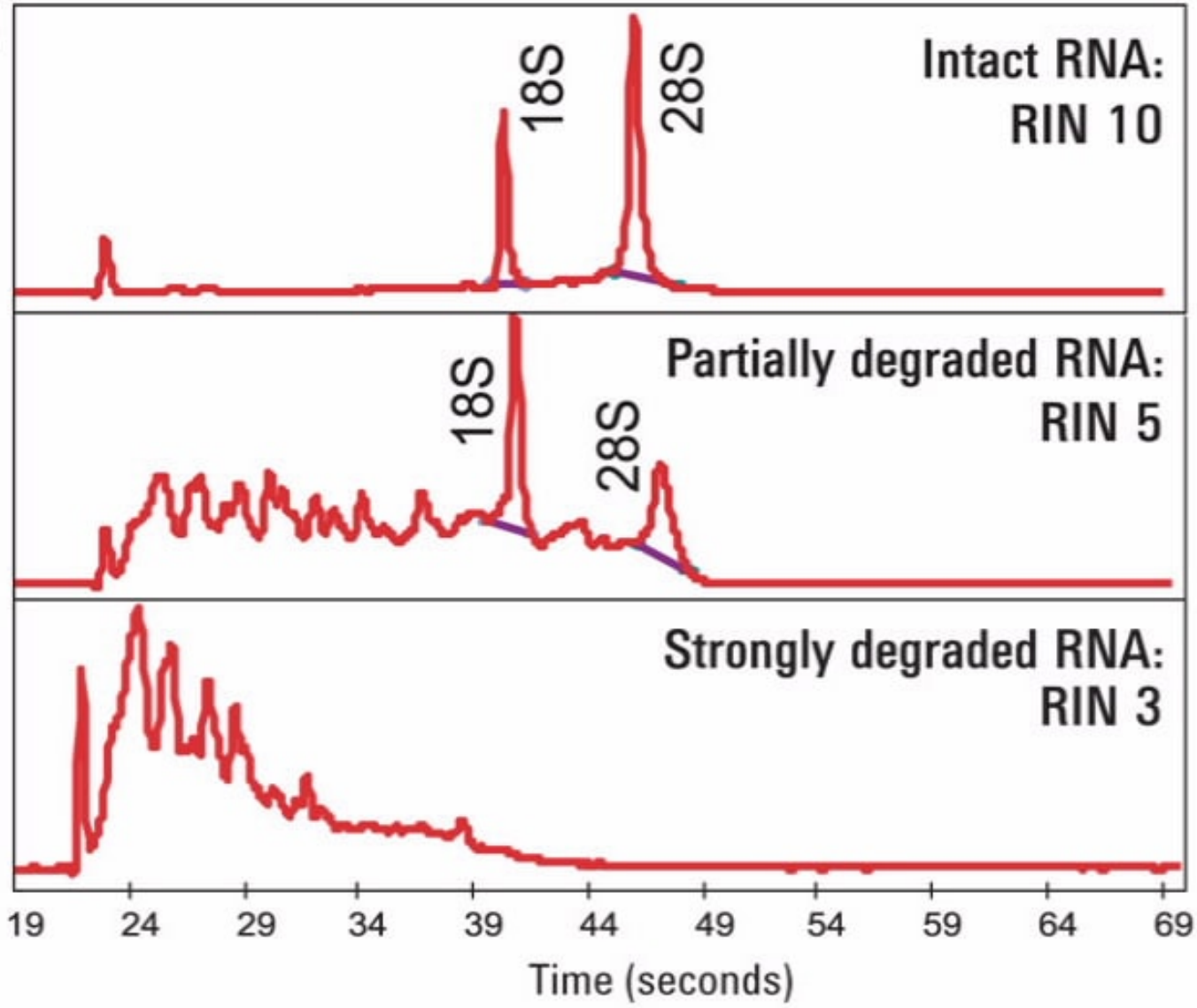




messenger RNA
(mRNA)



ribosomal RNA
(rRNA)


18S
28S
Intact RNA:
RIN 10
18S
28S
Partially degraded RNA:
RIN 5
Strongly degraded RNA:
RIN 3
19
24
29
34
39
44
49
54
59
64
69
Time (seconds)

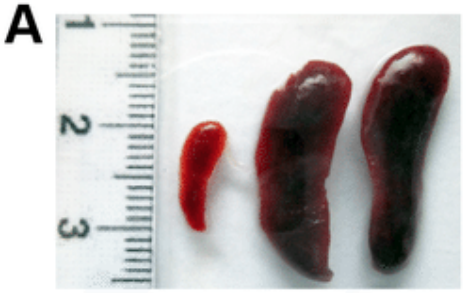
RNA Bioanalyzer requires Qubit

- Need to take a Qubit reading not NanoDrop

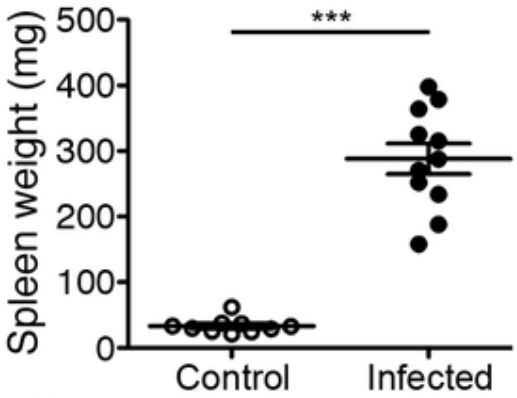
Experimental workflow before it gets sequenced

1

Samples of interest

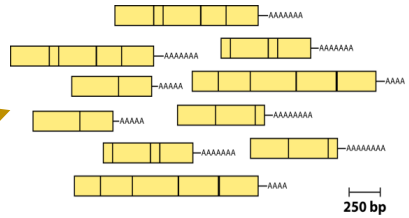


PMID: 27548618



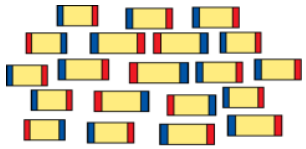
2

Isolate RNAs



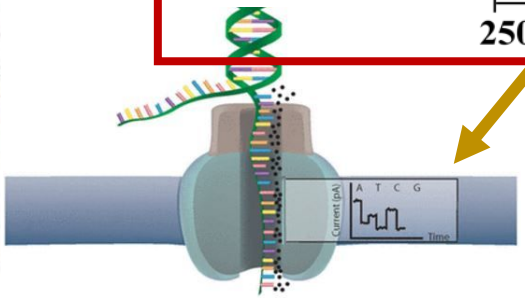
3

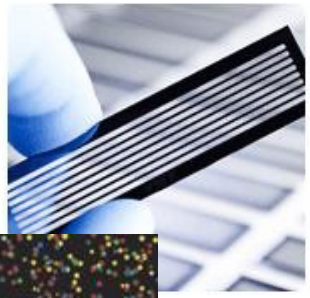
Library build



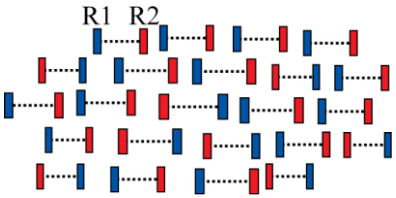
4

Illumina sequencing versus ONP





5

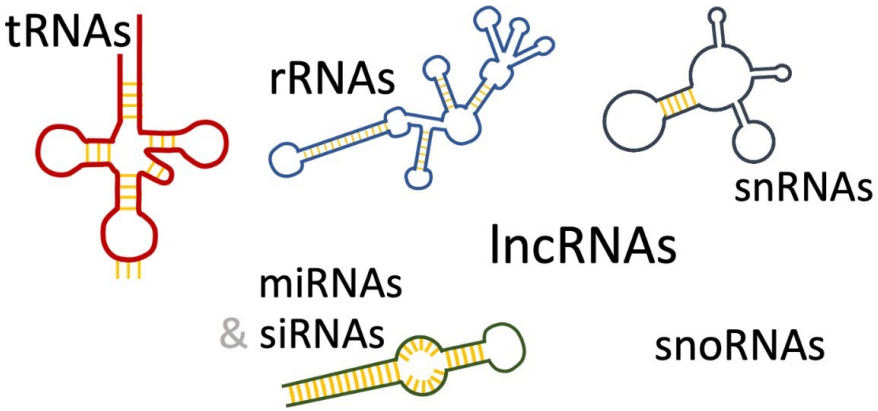


Reads (R1 and R2) generated

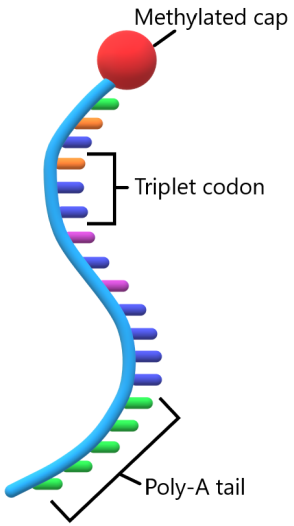
RNA composition

RNA comes in many different flavors

- Ribosomal-related RNAs:
  - rRNA, tRNA, snoRNA (up to 90% of RNAs)
- Protein-coding RNAs:
  - mRNA
- Regulatory RNAs:
  - microRNAs, lncRNAs



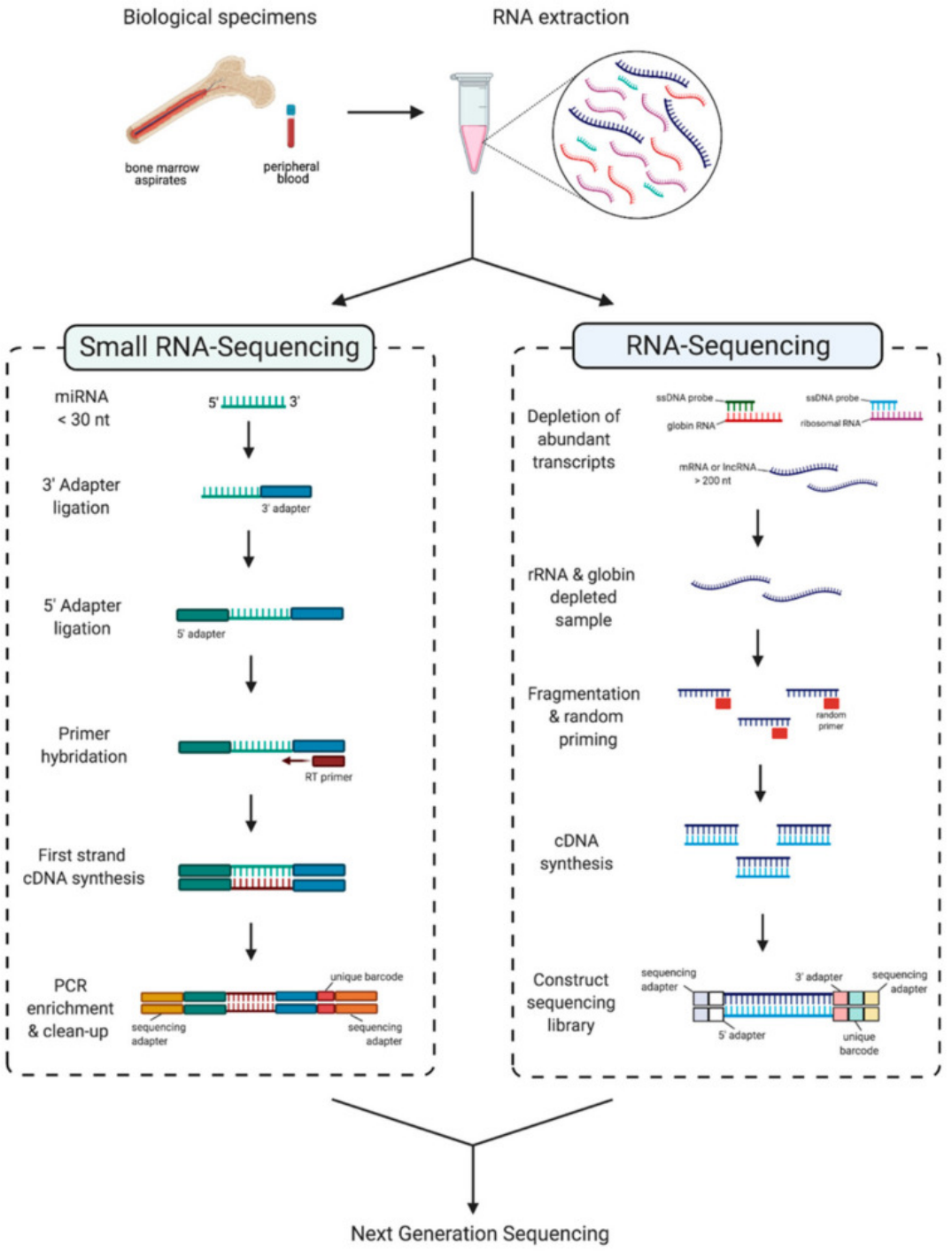
messenger RNA



- Kukurba, Mongomery (2015) Cold Spring Harb. Protoc.
  Gaida SM et al. (2015) Nat Comm 6:7045

“Comprehensive” transcriptome analysis

Two different protocols/kits!

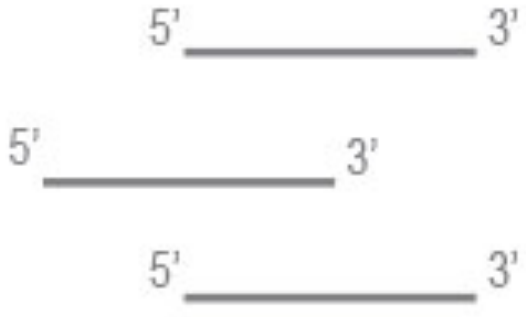


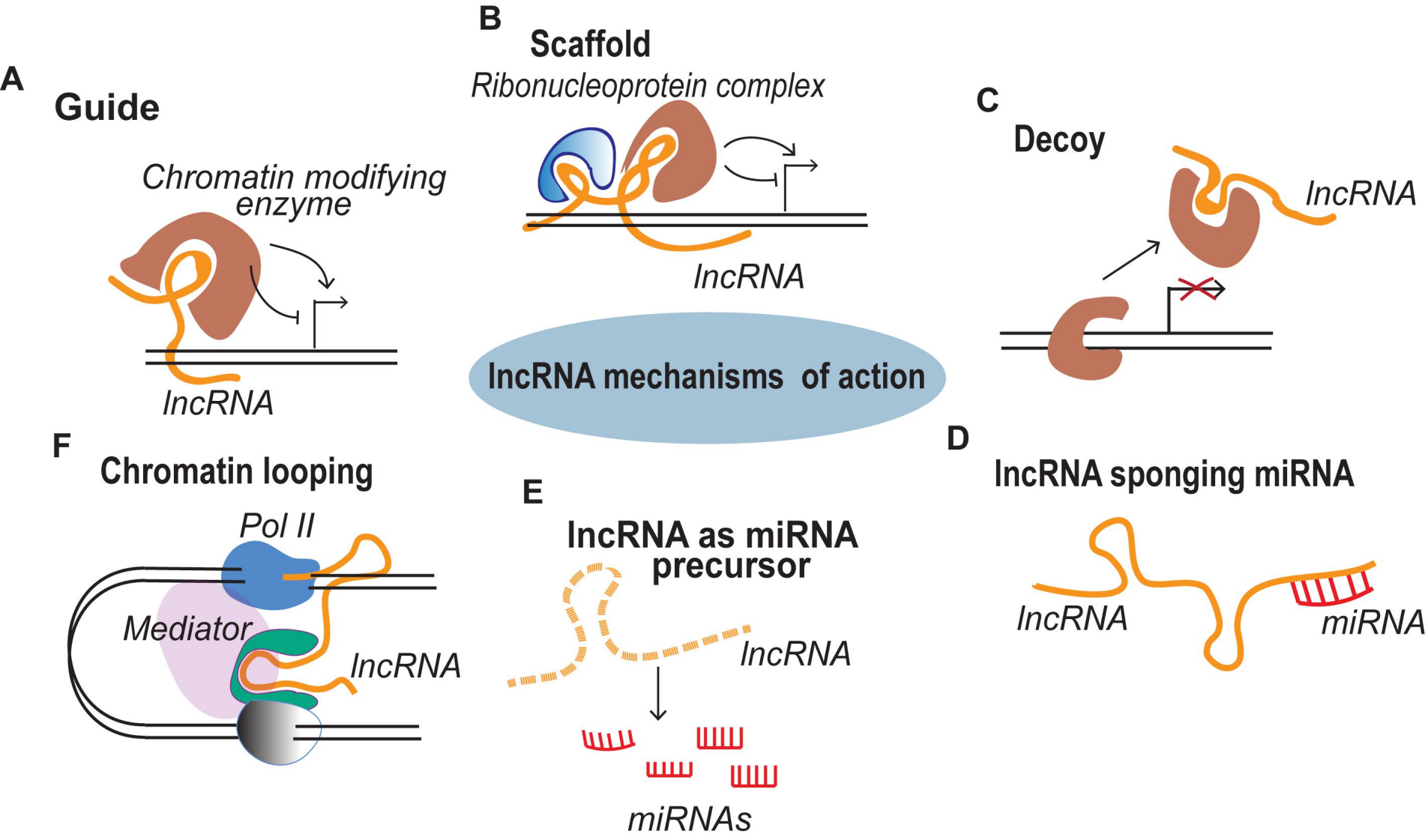
Before building the libraries need to perform target enrichment

- If rRNA is not removed, the majority of the final sequencing reads would be from rRNA and not mRNA
- Therefore, it necessary to enrich for mRNA
- Two common strategies:
  1. Poly A+ selection = mRNA only
  2. rRNA depletion = mRNA + *other species*

Poly-A versus rRNA depletion?

- For differential gene expression, most enrich using Poly(A)+
- However:
  - If you are aiming to obtain information about long non-coding RNA's perform ribosomal RNA depletion
  - Bacterial mRNAs are also not poly-adenylated




A
Guide
Chromatin modifying enzyme
lncRNA
B
Scaffold
Ribonucleoprotein complex
lncRNA
C
Decoy
lncRNA
lncRNA mechanisms of action
F
Chromatin looping
Pol II
Mediator
lncRNA
E
lncRNA as miRNA precursor
lncRNA
miRNAs
D
lncRNA sponging miRNA
lncRNA
miRNA

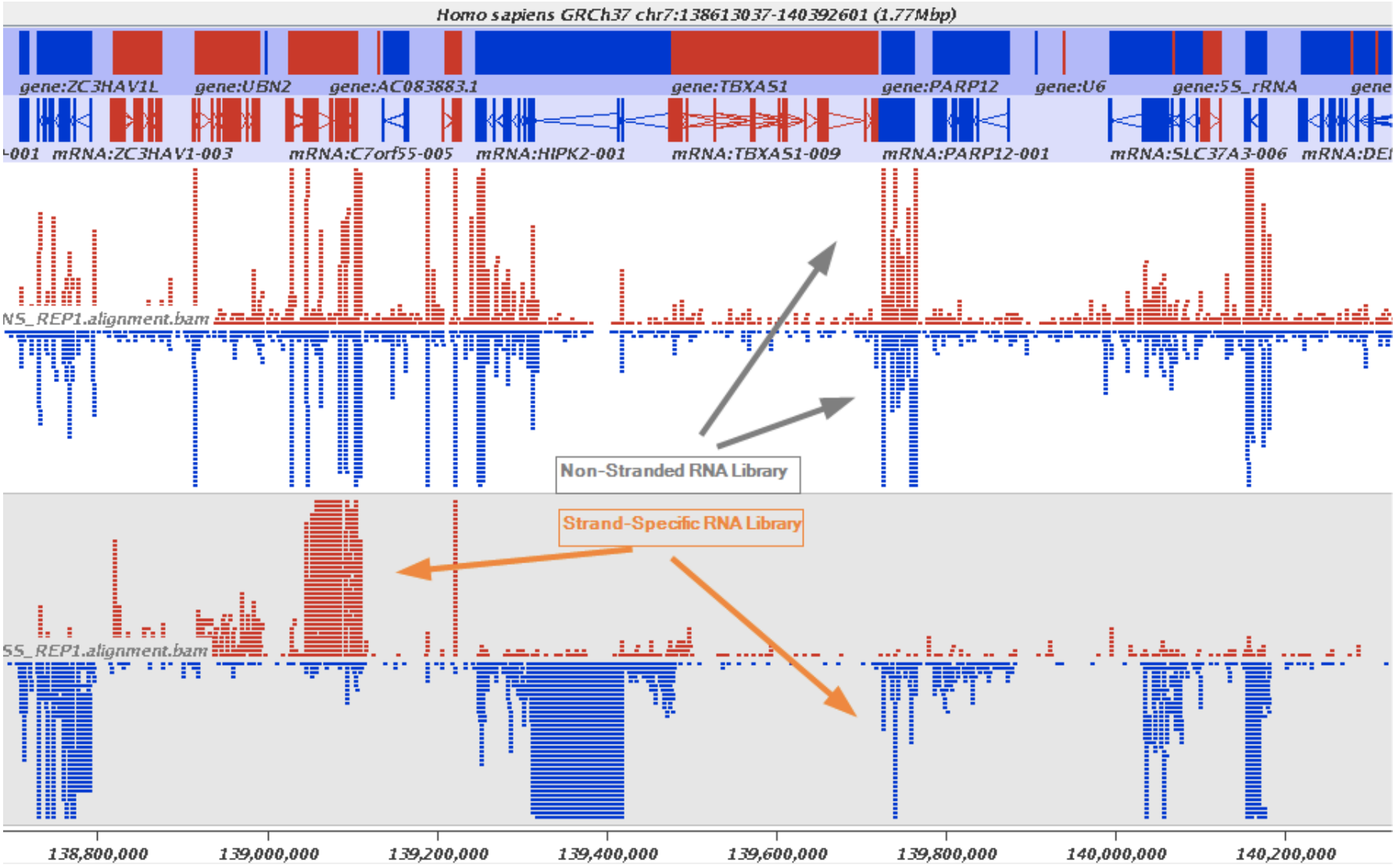
Strandedness

Another consideration is whether to generate strand-preserving libraries

Libraries can be stranded or unstranded

The implication of **stranded** libraries is that you could distinguish whether the reads are derived from forward or reverse-encoded transcripts

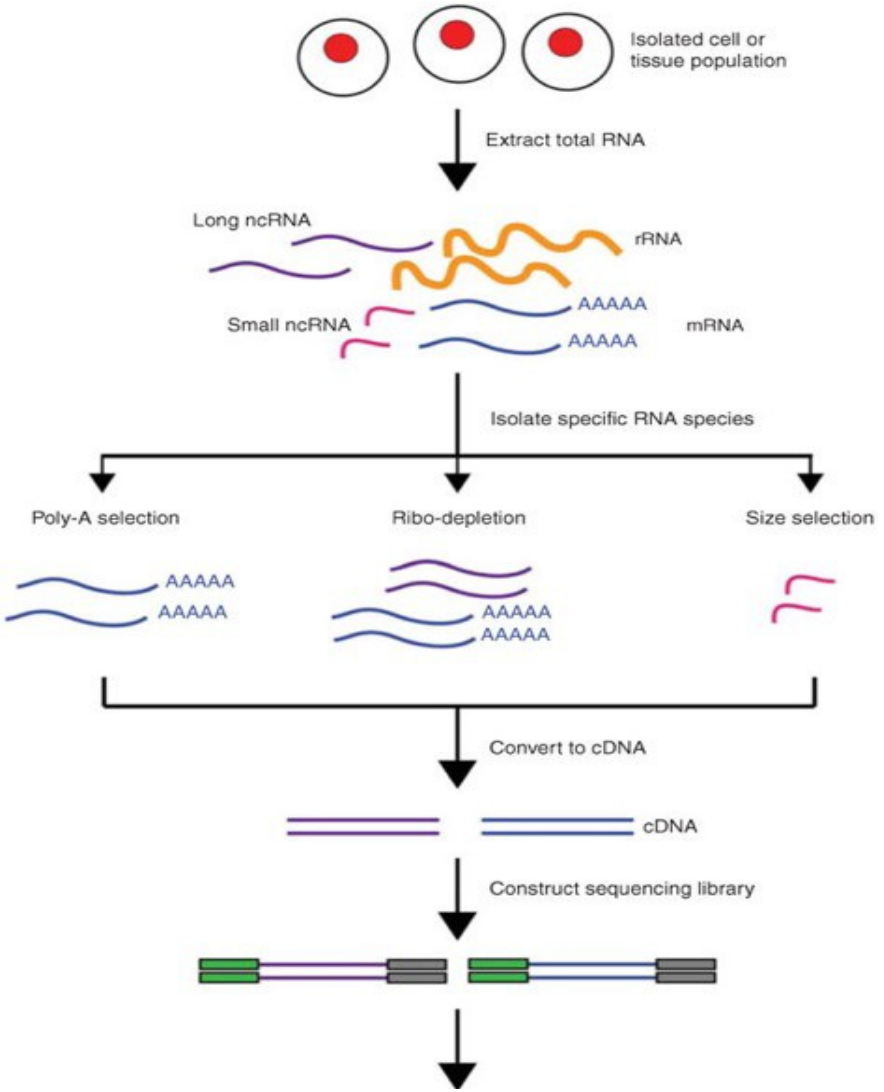
Strandedness



Red = + strand

Blue = - strand

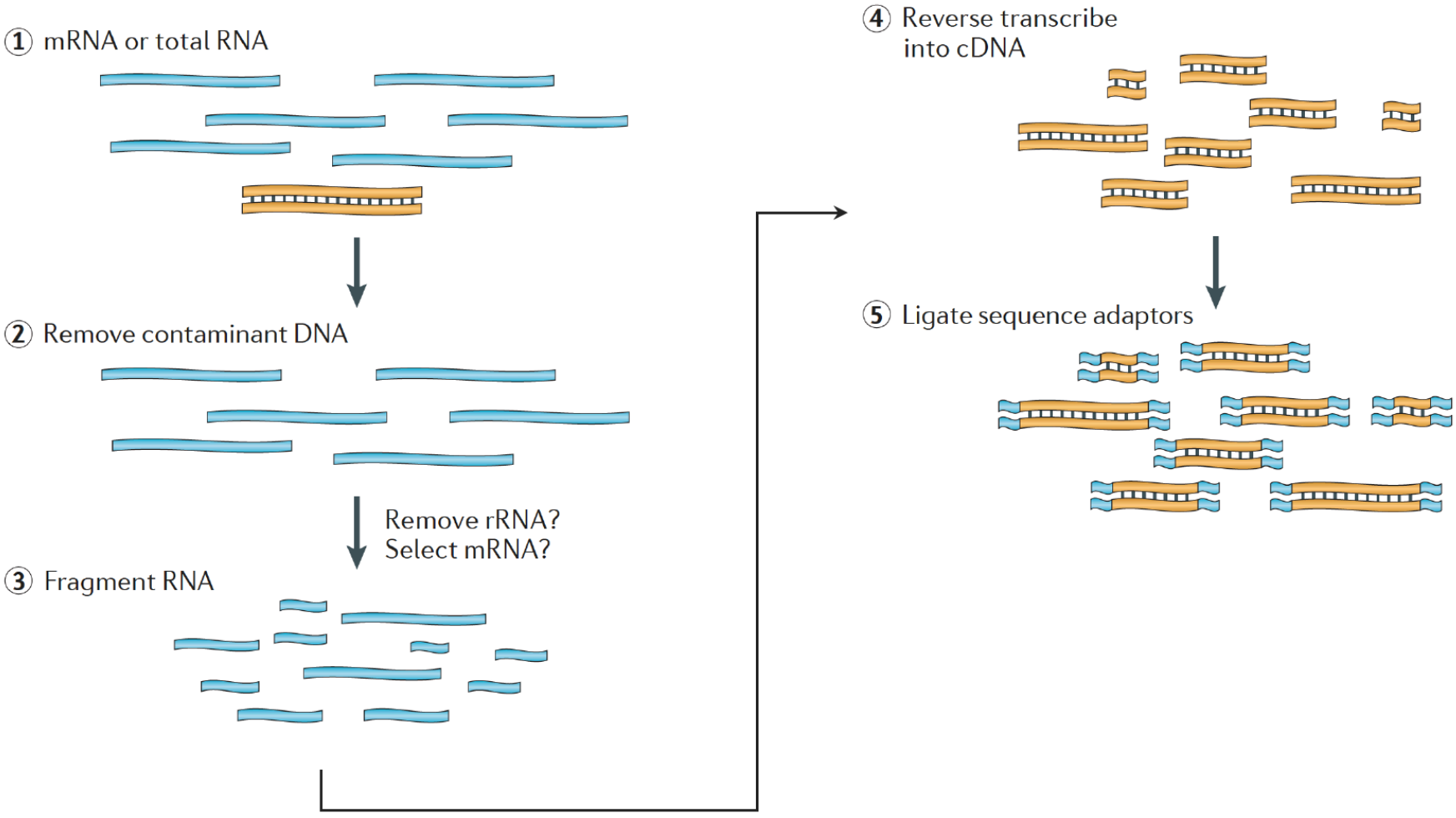
RNA-seq Library build steps



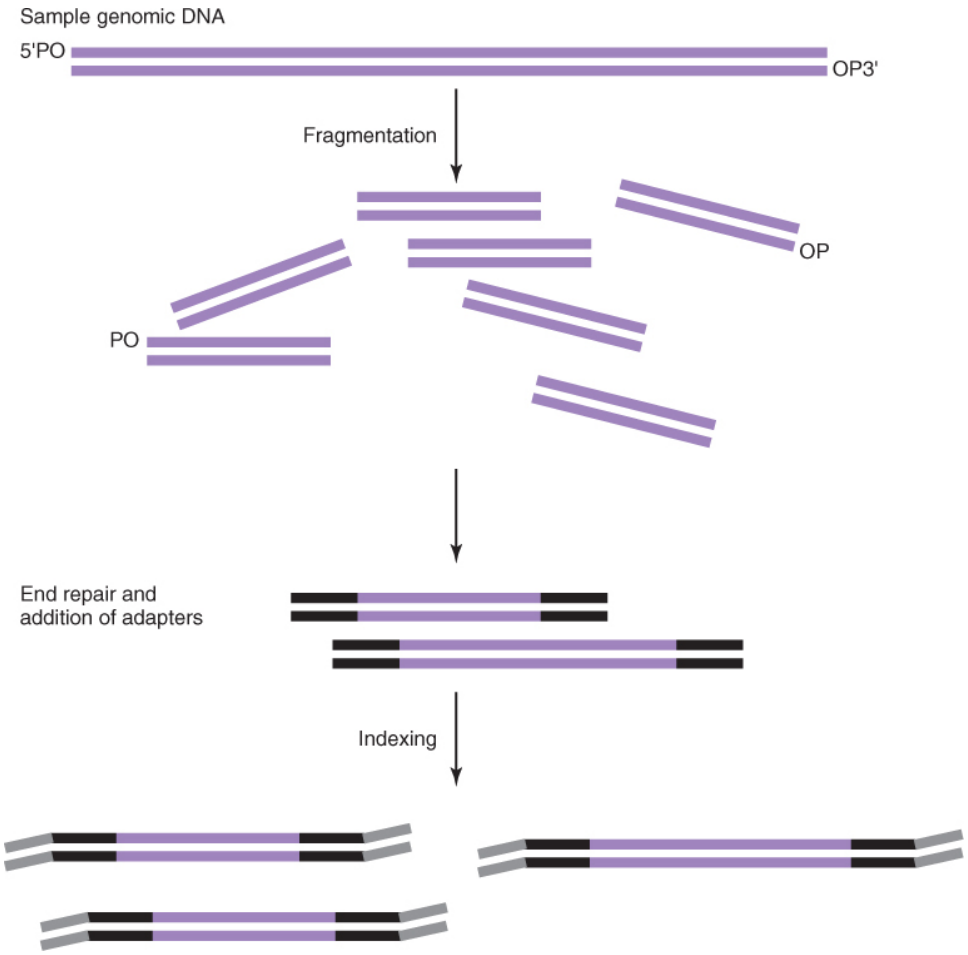
- ☑ Isolate RNA
- ☑ RNA species diversity
- ☑ target enrichment
- ☐ cDNA conversion / PCR amplification

- Kukurba, Mongomery (2015) Cold Spring Harb. Protoc.
  Gaida SM et al. (2015) Nat Comm 6:7045

Library Prep steps



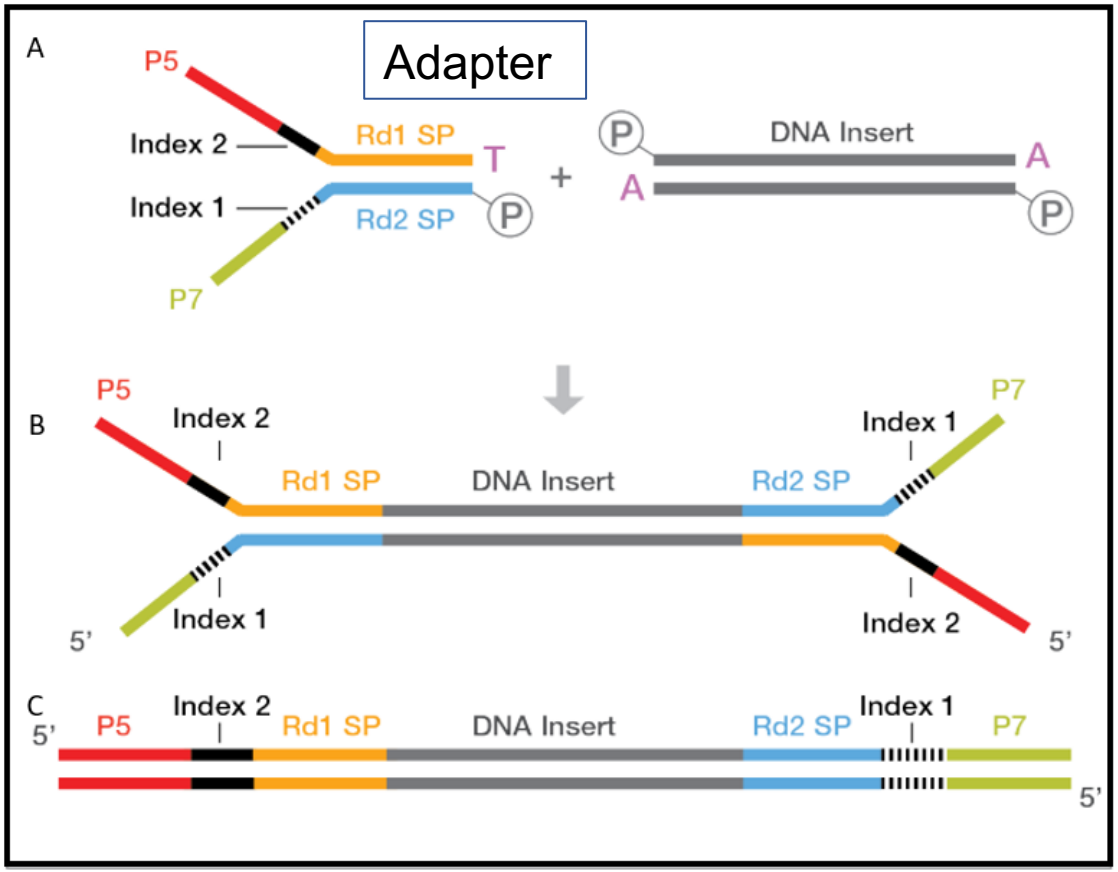
Adapters vs Index

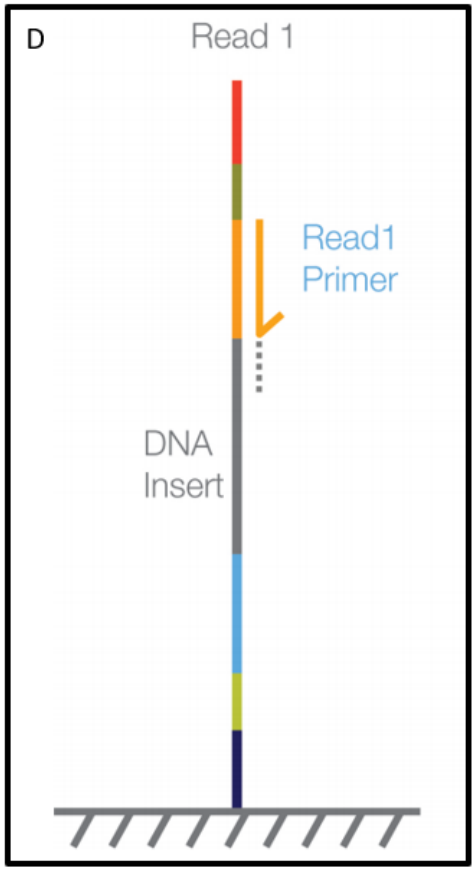


- Adaptors will serve as primer binding sites for amplification and sequencing, and for immobilization of fragments by hybridization
- Indices can also be used as a “barcode”/sample ID to combine many samples into 1 seq. run

Architecture of Standard Illumina NGS library

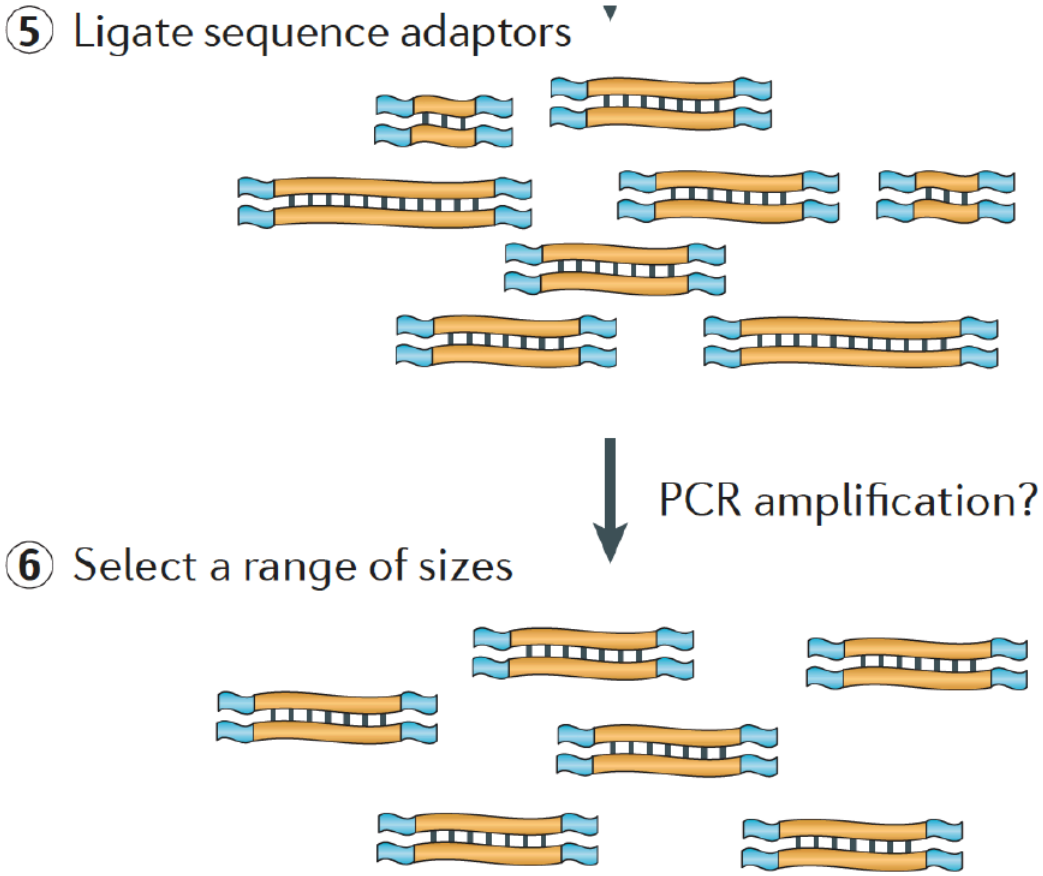
P5 and p7 sequences are required to bind the flow cell





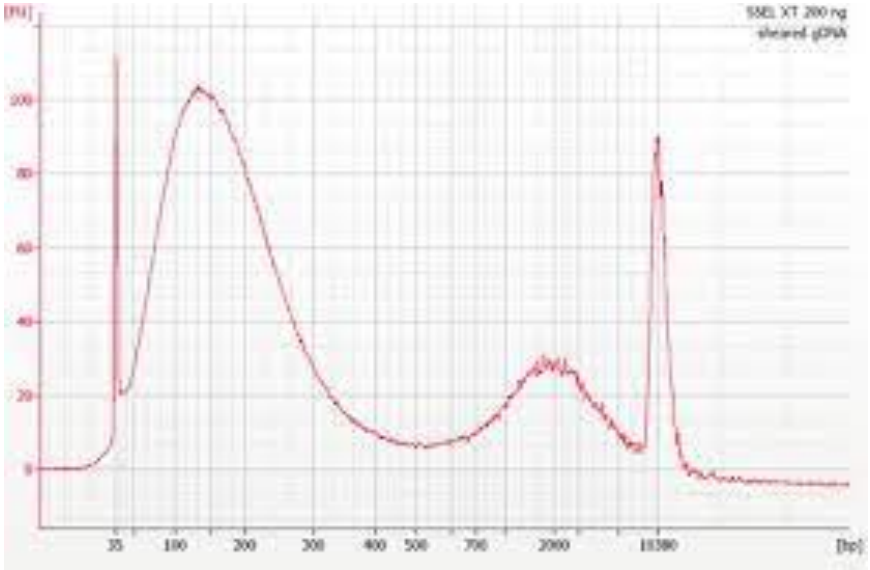
Unique index

Library Prep steps continued



Common mistakes during library build

- Addition of adapters and unique barcodes
- Number of PCR cycles



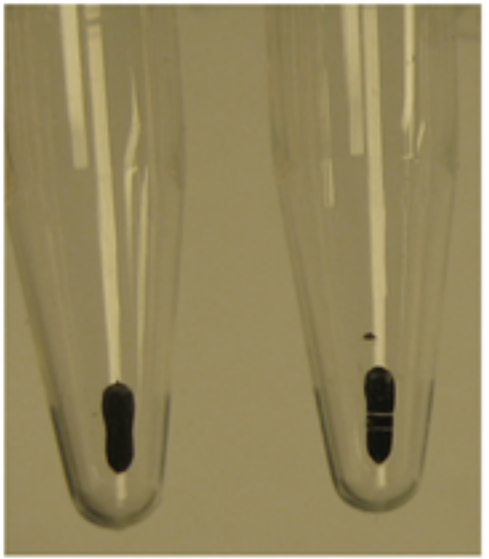
Common mistakes during library build

- Addition of adapters and unique barcodes
- Number of PCR cycles
- Ampure XP beads usage

Shiny Wet Pellet

Matt Dry Pellet

Cracked Overdried Pellet



Common mistakes during library build

- Addition of adapters and unique barcodes
- Number of PCR cycles
- Ampure XP beads usage
- RNA fragmentation

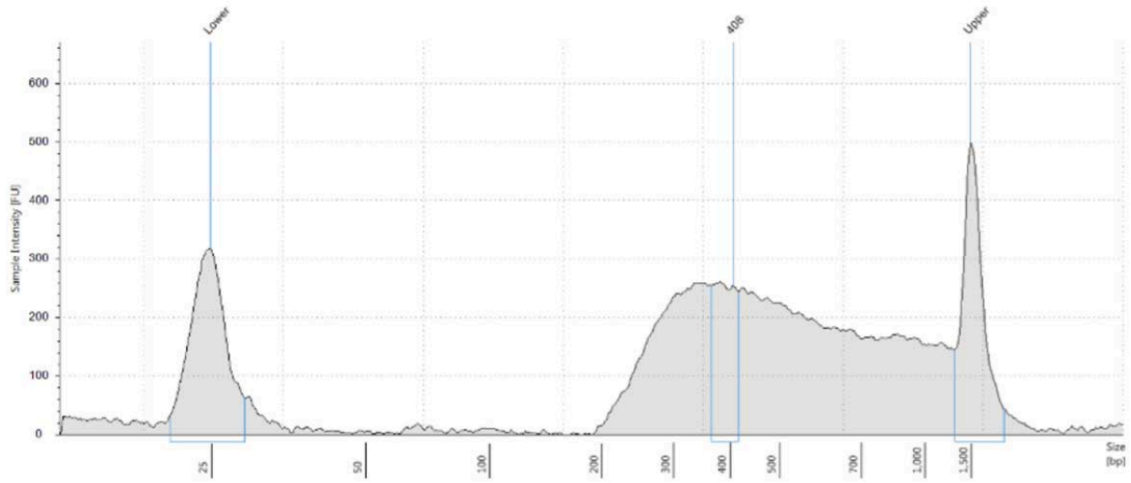
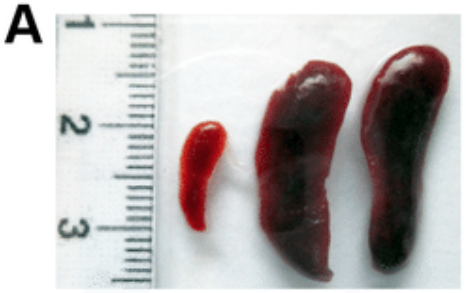


Figure 4. An electropherogram showing a sample with incomplete fragmentation due to suboptimal mixing. There are large fragments present which will be unable to undergo bridge amplification during cluster generation on the flow cell.

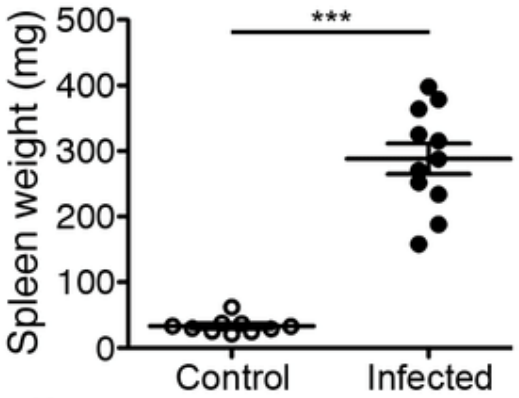
Experimental workflow before it gets sequenced

1

Samples of interest

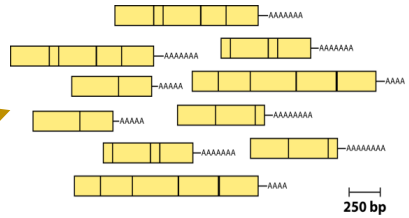


PMID: 27548618



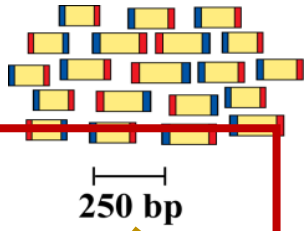
2

Isolate RNAs

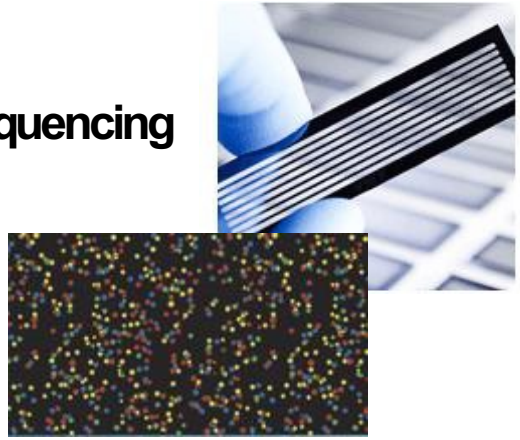


3

Library build

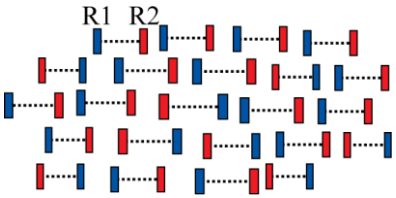


Illumina sequencing



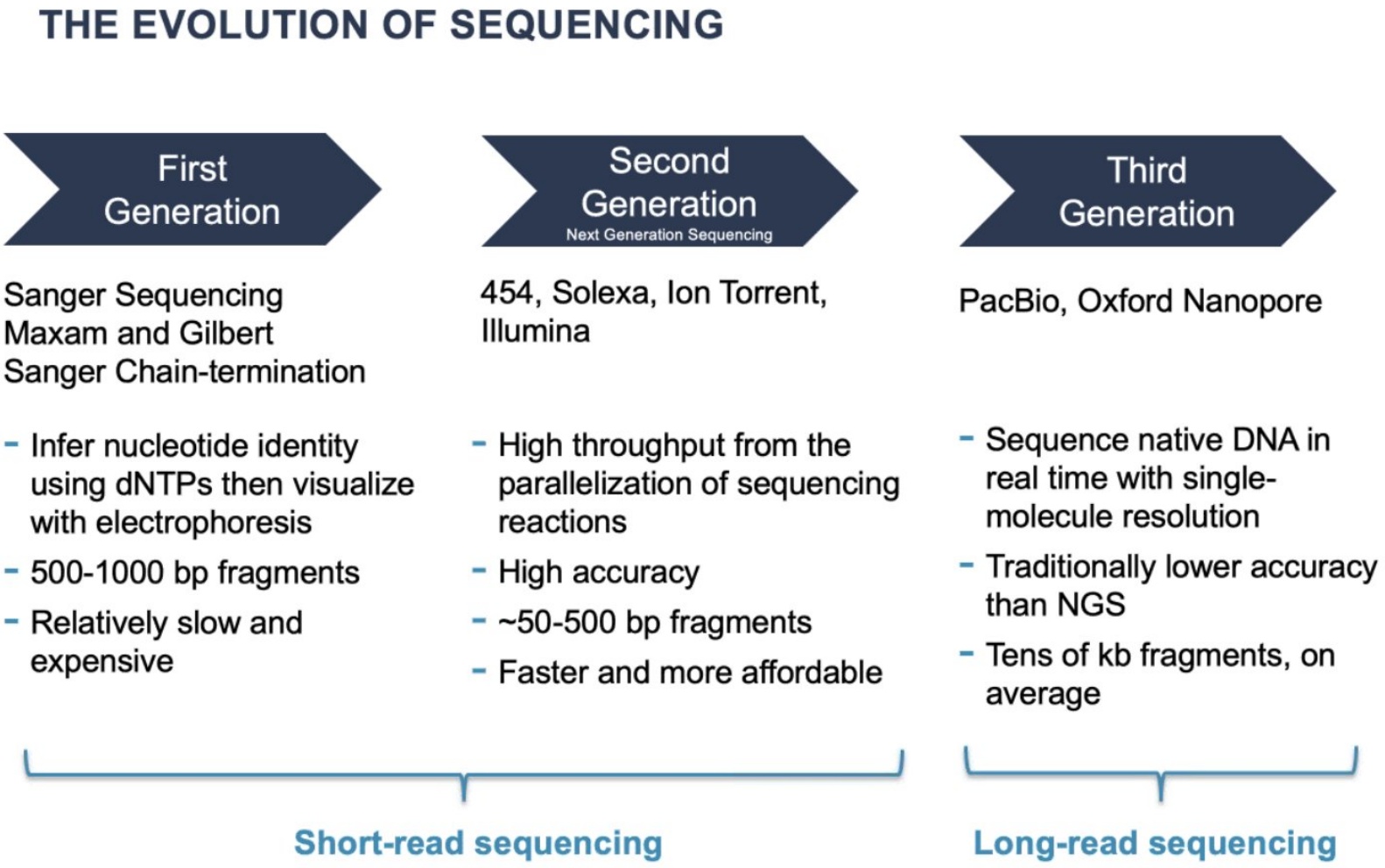
4

5



Reads (R1 and R2) generated

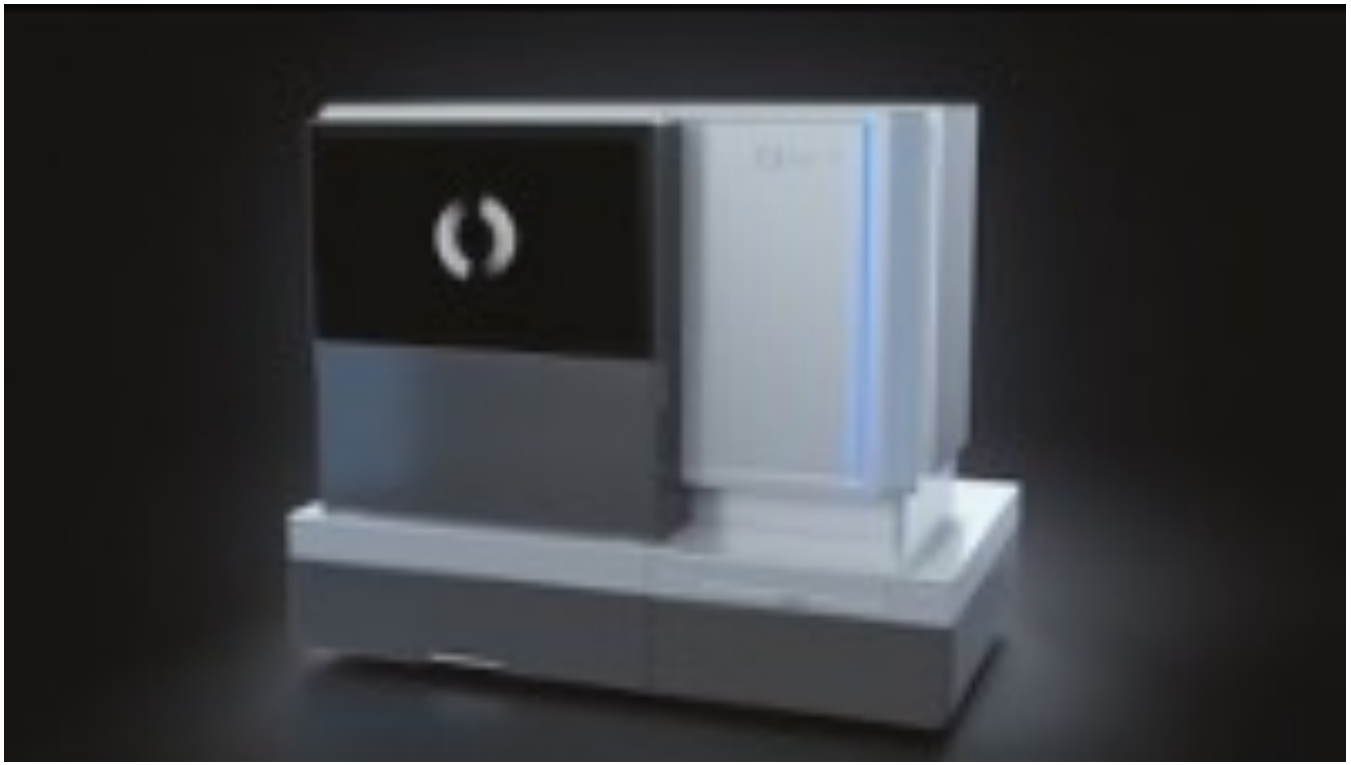
Two main approaches in NGS: short-read vs long-read



Second Generation Sequencing

- Typical characteristics:
  - Reads are short (100-300bp)
  - Quality of bases decreases as the length of the read increases
  - Includes HiSeq, NextSeq, NovaSeq platforms
  - Run Time varies but can be up to > 48hrs
  - Maximum Output is ~25 to 400 million reads per lane

At UVM, we have the G4 sequencer



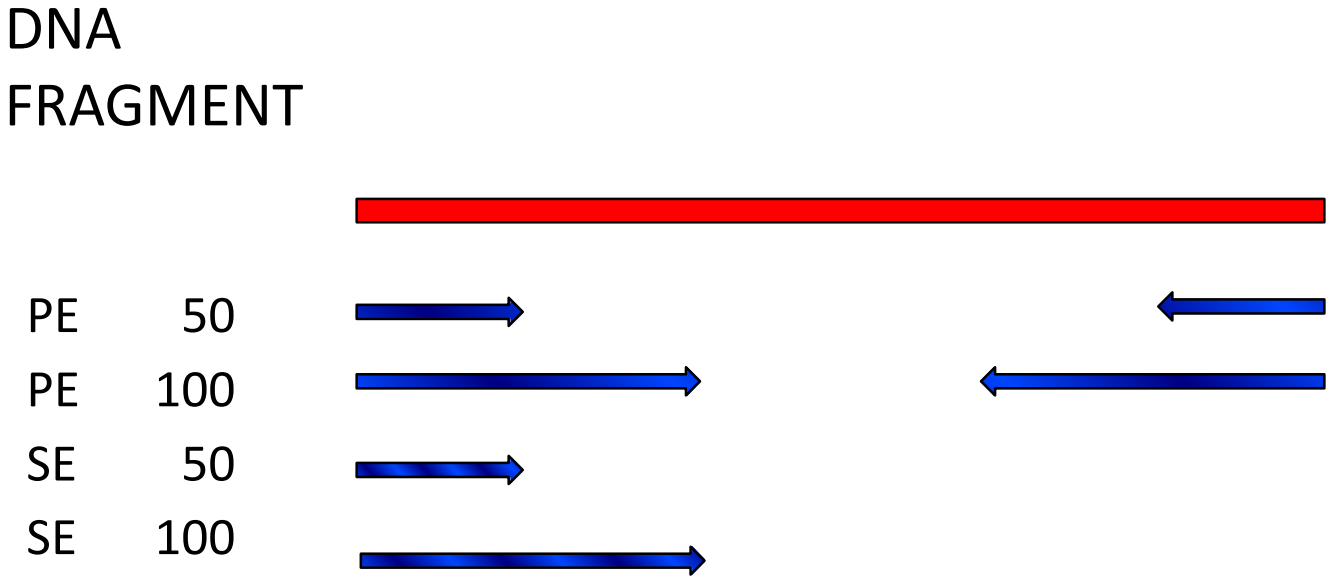
https://youtu.be/tBB0tAgFhiU?si=7Qtm0PTtGQuxfkk7

https://singulargenomics.com/g4/

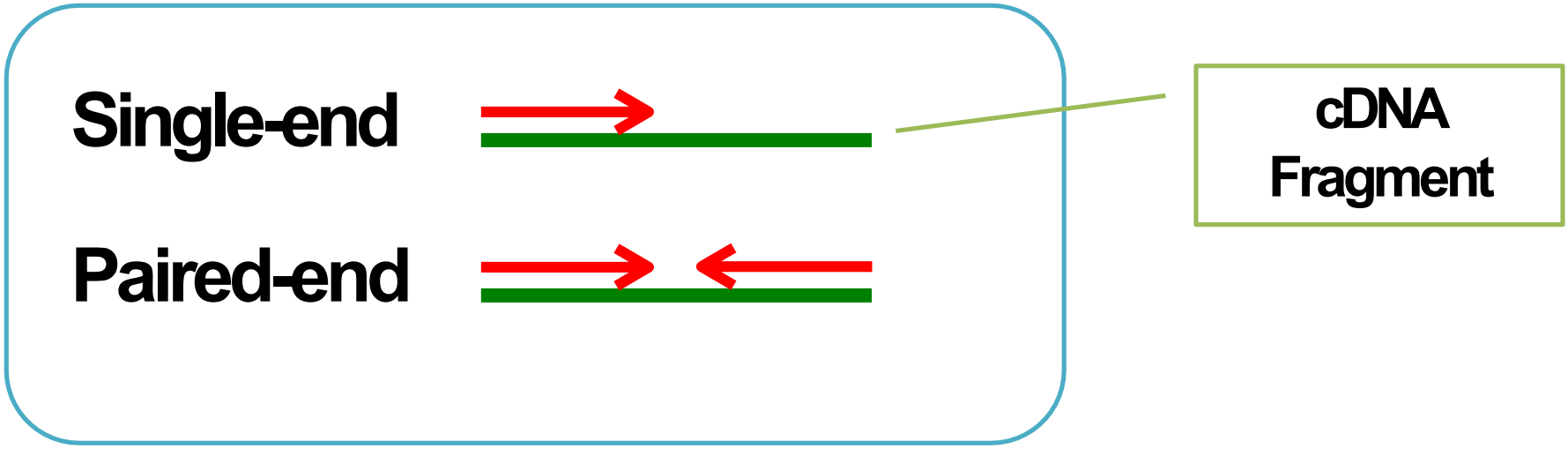
Sequencing Options

Illumina Sequencing options:

- Length of sequence (up to 300 bases)
- Paired-end (PE) or single-end (SE)



Single-end vs paired-end



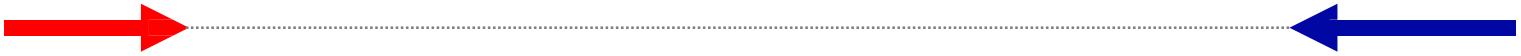
What you get from VIGR:

Single-end: ONE fastq file per sample

Paired-end: TWO fastq files per sample

What is the Advantage of Longer and PE Reads?

- Reads mapping to junctions
  - With longer reads we will have more reads spanning exons
  - Isoforms or distinguishing paralogs



- Paired end reads

  Knowing both ends of a fragment and an approximation of fragment size helps to determine the transcript from which it was derived.

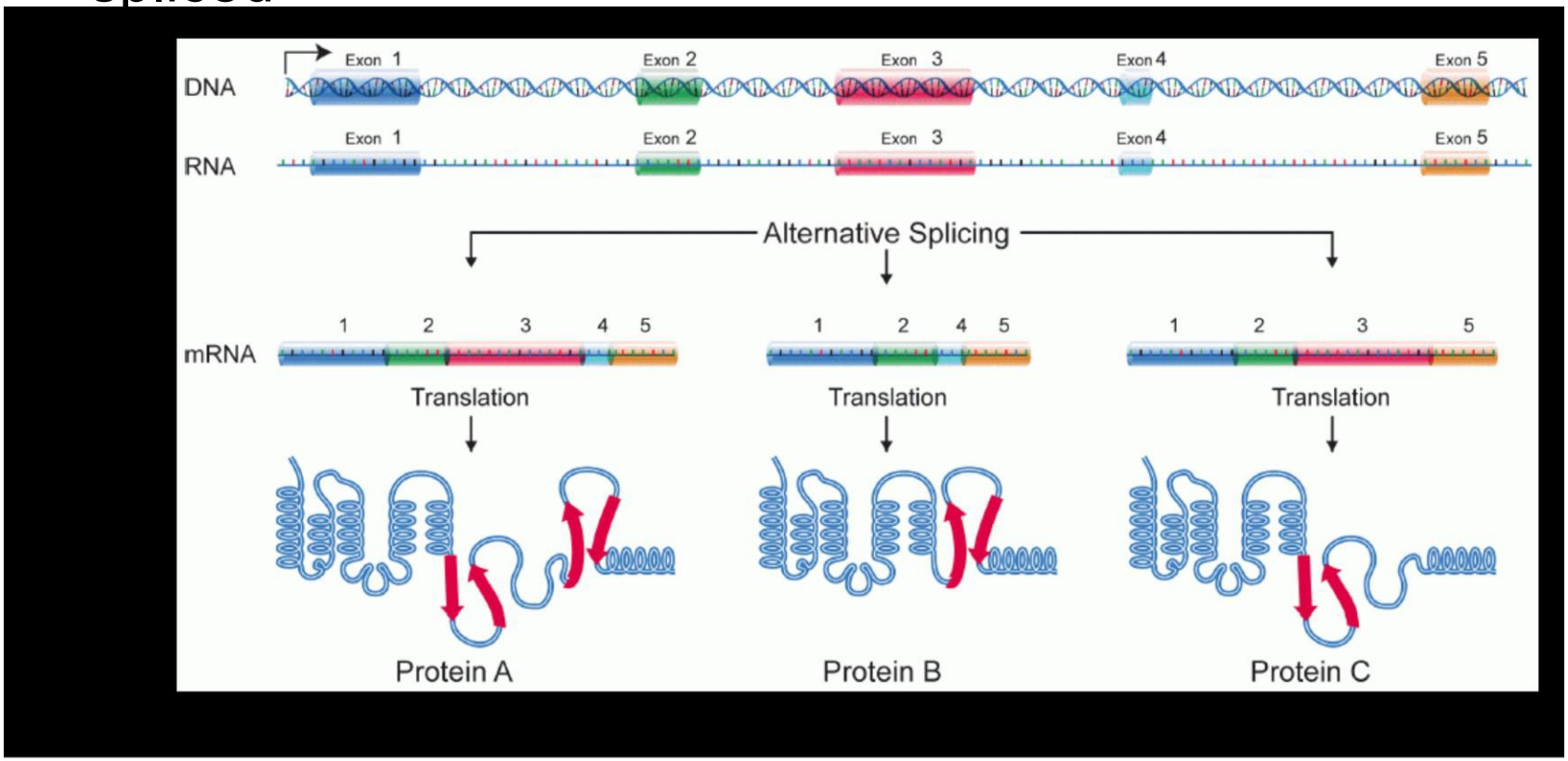
When does SE vs PE matter?

If your considering studying:

- long non-coding RNAs
- De novo transcriptome assembly
- Alternative splicing
  - Alternative RNA splicing is a process during gene expression that allows a single gene to produce many transcripts

Alternative Splicing

- Alternative splicing increases the biodiversity of proteins that can be encoded by the genome.
- In humans, ~95% of multi-exonic genes are alternatively spliced



Alternatively spliced exons yield three different protein isoforms

Other NGS terms

What is a read?

A read is a string of bases represented by their one letter codes. Here is an example of a read that is 50 bases long. `TTAACCTTGGTTTTGAACTTGAACACTTAGGGGATTGAAGATTCAACAACCCTAAAGCTTGGGGTAAAAC`

Other NGS terms

- **Read:** A raw sequence that comes from a sequencing machine.
- **Sequencing depth**: total number of sequences, reads, or bp generated in a single experiment

	Replicates per group		
	3	5	10
Fold change			
2	87%	98%	100%
Sequencing depth (millions of reads)			
3	19%	29%	52%
10	33%	51%	80%
15	38%	57%	85%

PMID: 26813401

In Summary, to quantify Differential Gene Expression

- Read length: 50 to 100 bp
- Paired vs single end: Single end (cheaper)
- Number of reads: > 15 million per sample
- Replicates: 3 biological replicates *minimum*

A well-planned experiment goes a long way!

Summary RNA-Seq Experiment Planning