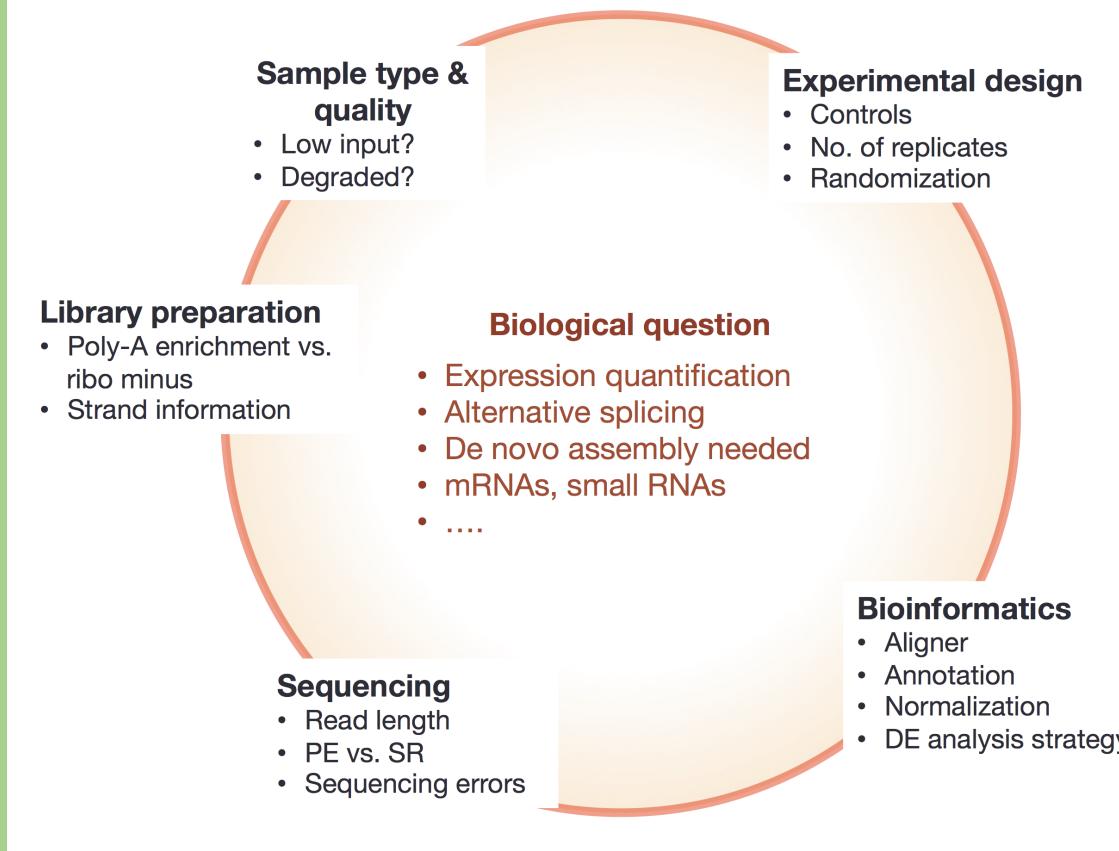


Everything's connected...



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

Princess Rodriguez, PhD

MMG 232
Spring 2023

Learning Objectives

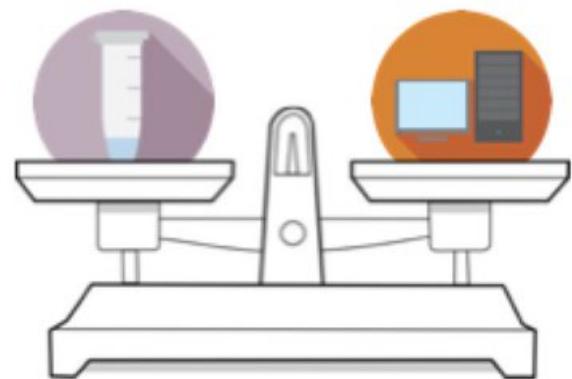
- Understand how an RNA-Seq experiment is planned and considerations that should be taken by the experimenter

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



Caveat

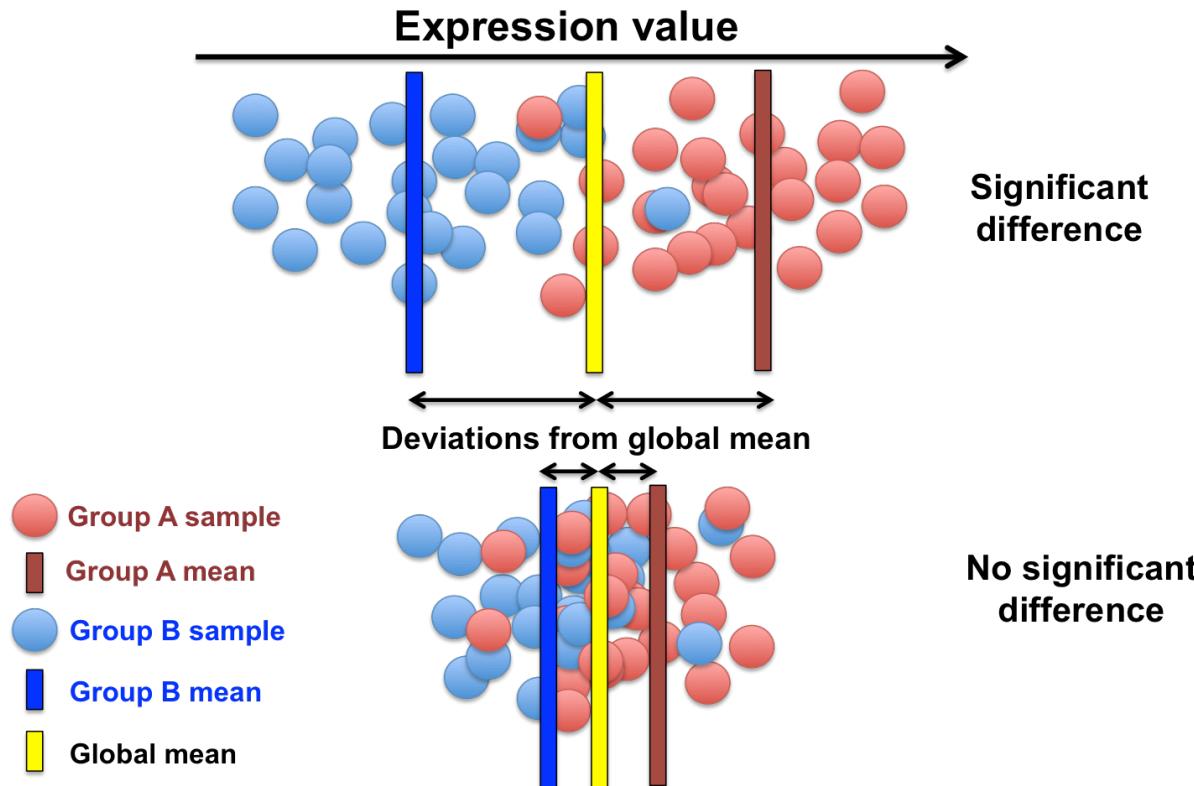
- Transcriptome is the collection of all transcript readouts present in a cell.
- RNA-seq can be used to explore and/or quantify the transcriptome of an organism, which can be utilized for many different *types* of experiments:

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:

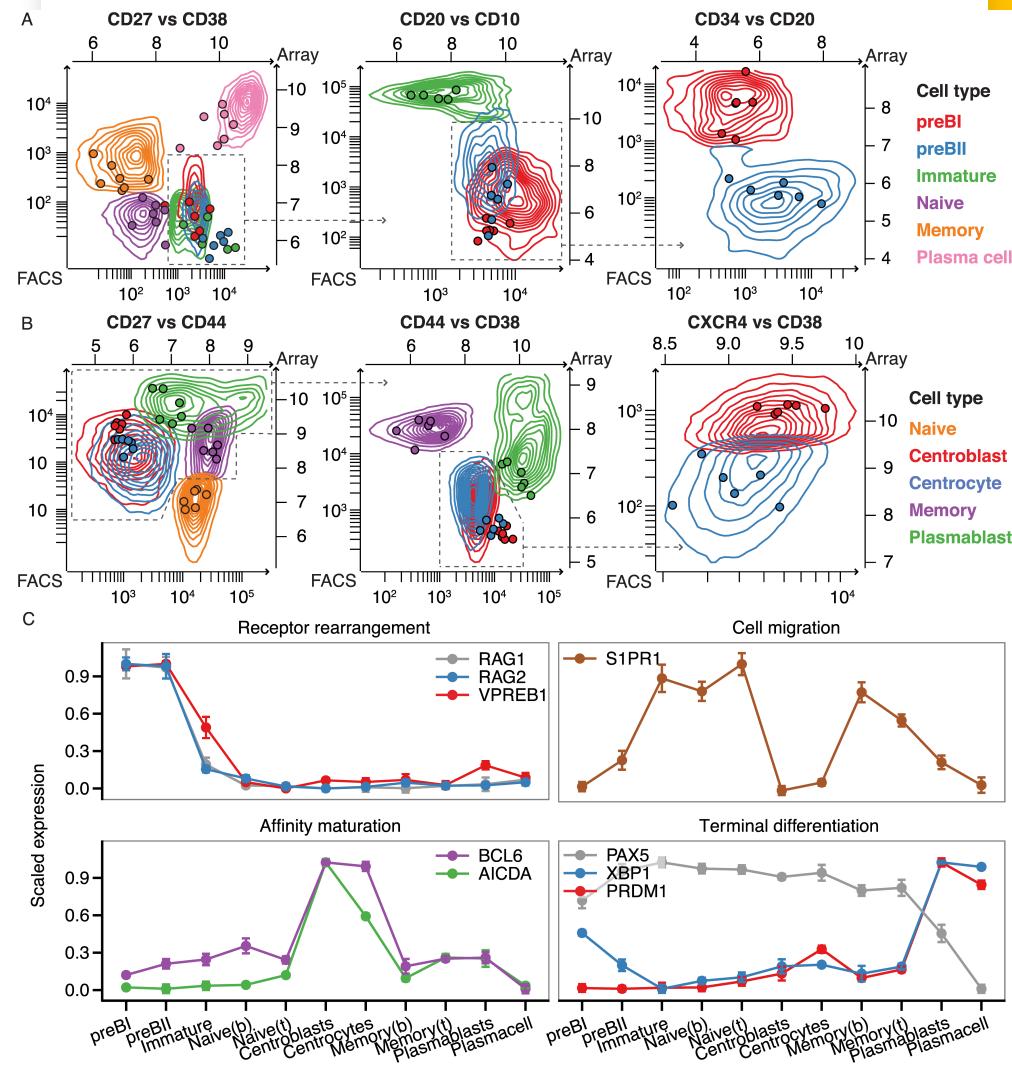
What genes are differentially expressed between conditions?



Basic types of questions answered:

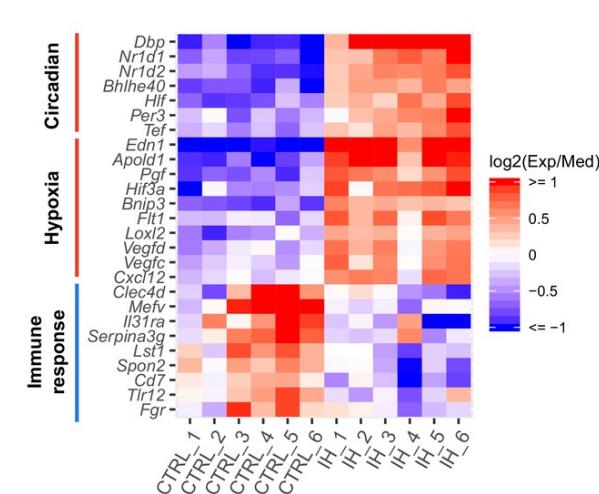
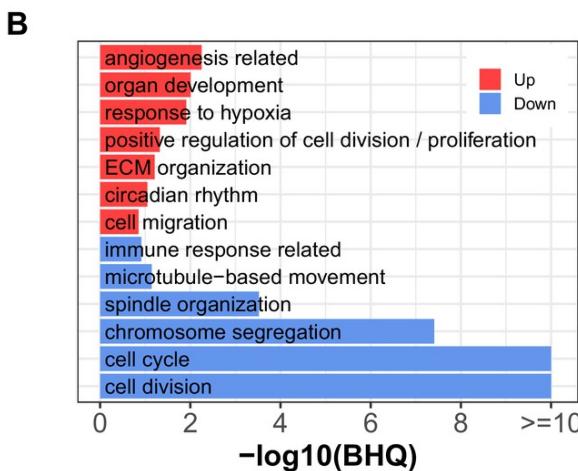
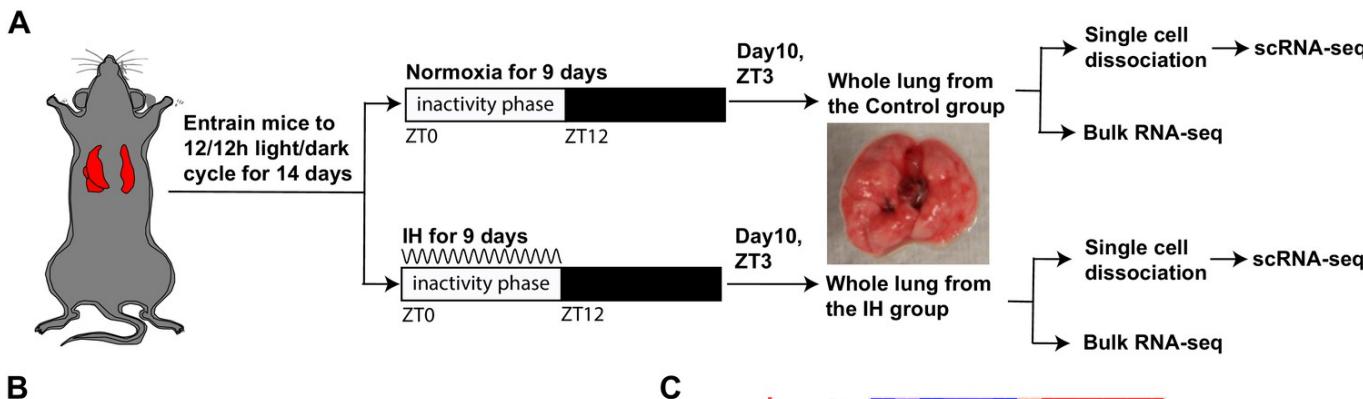
Are there any trends in gene expression over time or across conditions?

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



Basic types of questions answered:

What processes or pathways are important for my condition of interest?



Biological samples/Library preparation

Step 1

Sequence reads

Step 2

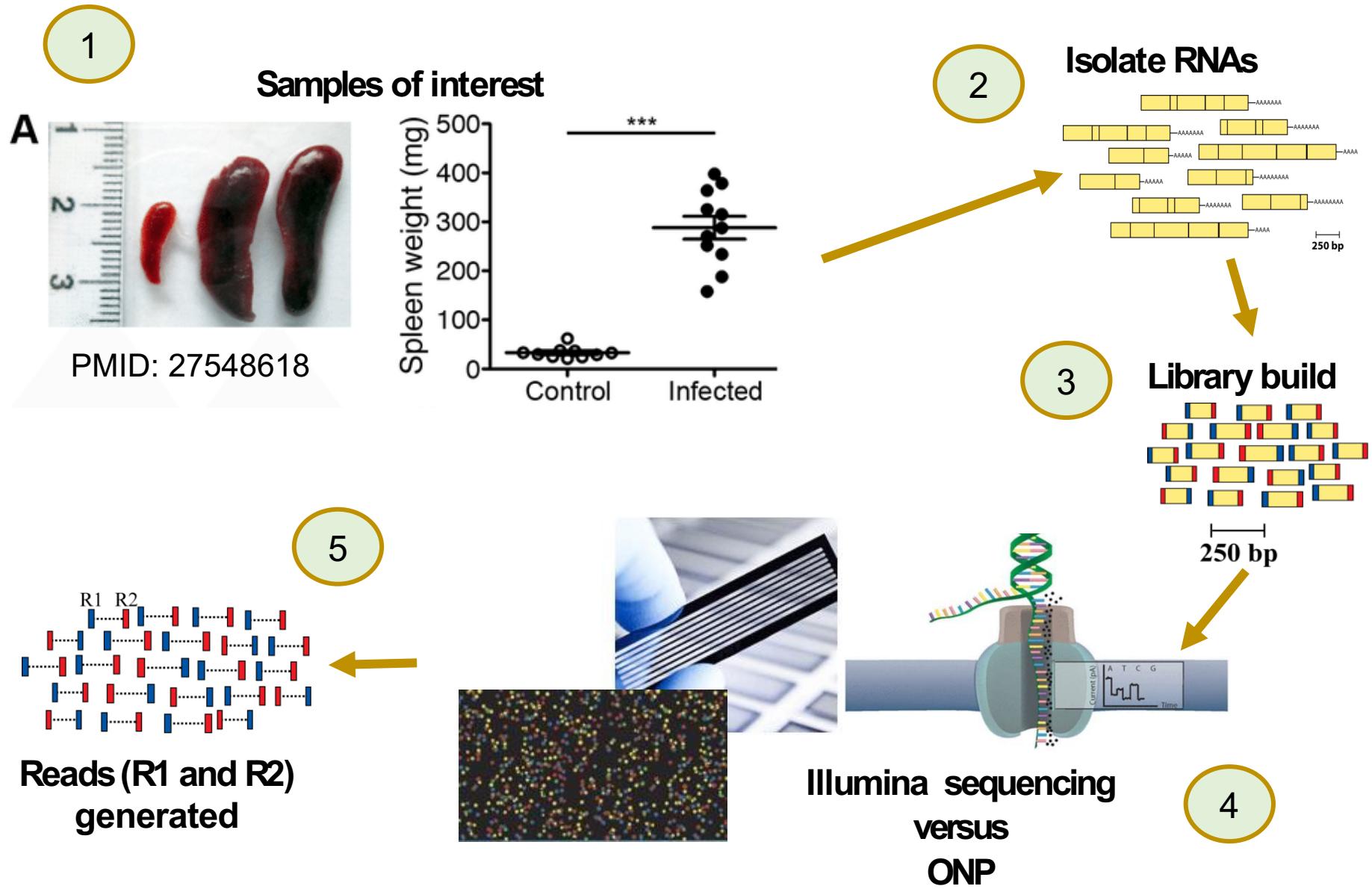
Mapping/
Quantification

DGE with R

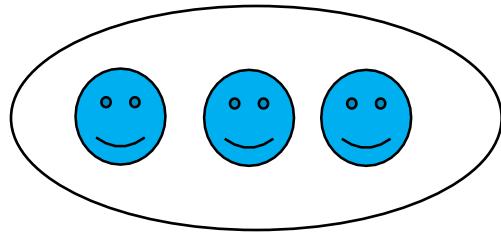
Step 3: Data Analysis

Functional
Analysis with R

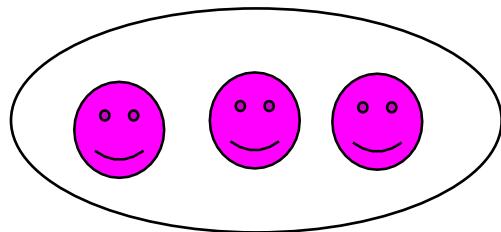
Experimental workflow



Biological Replicates



Condition 1

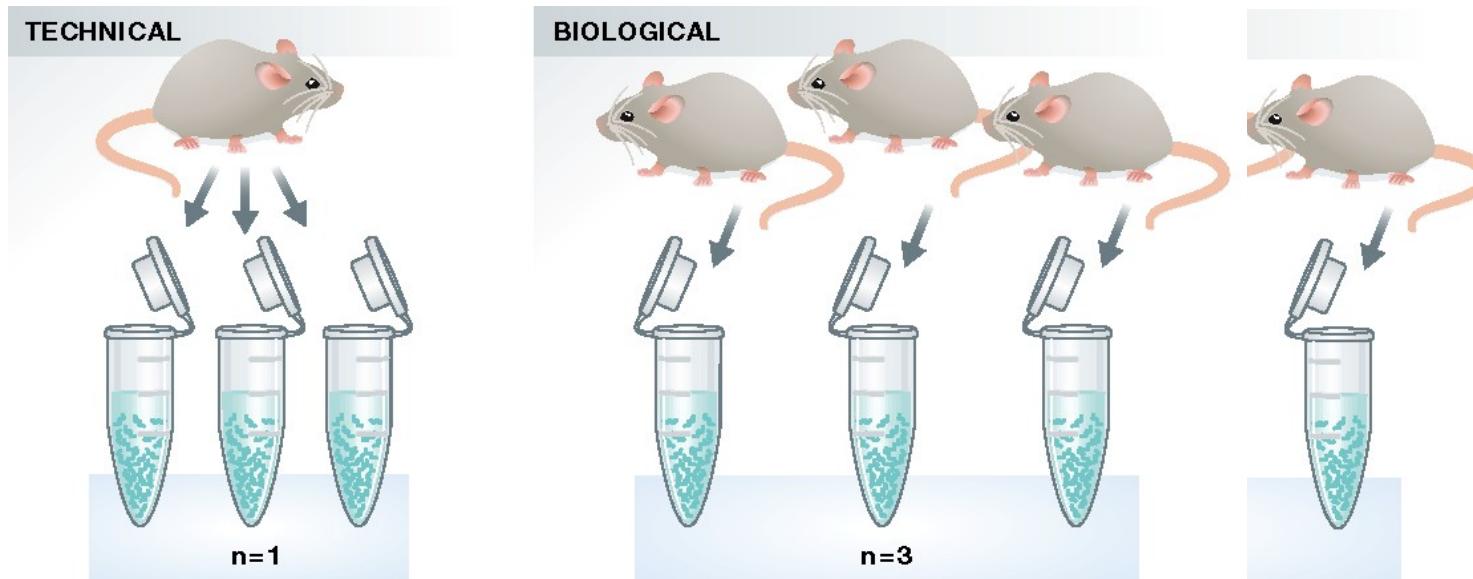


Condition 2

- ❖ 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



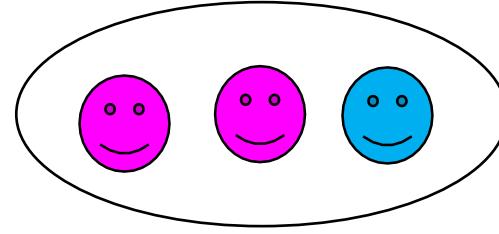
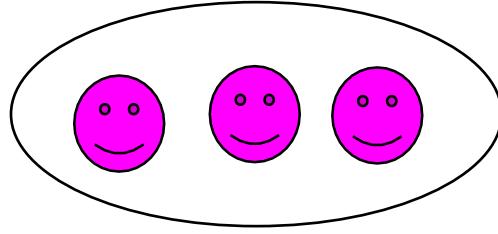
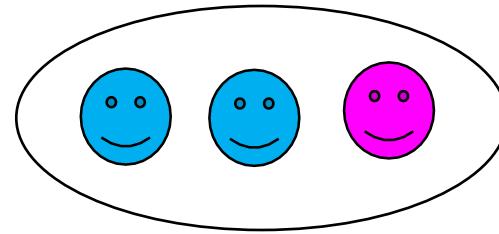
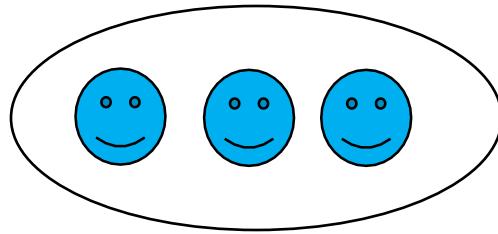
Probability of detecting DEGs

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



PMID: 26813401

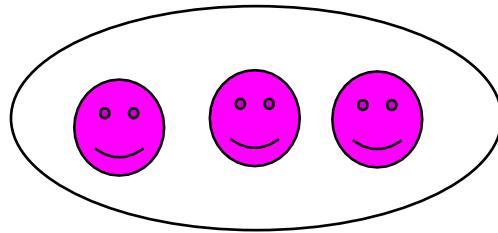
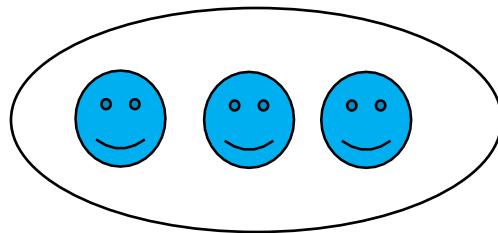
Grouping of Replicates



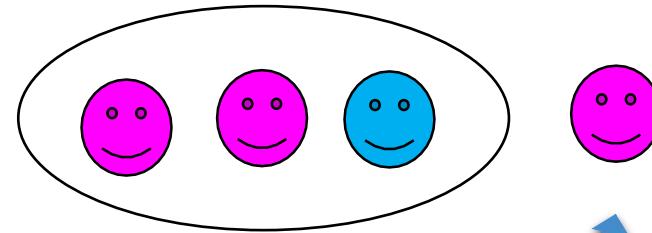
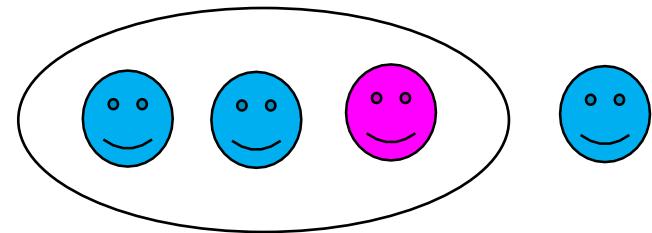
What you want

What you get

Grouping of Replicates



What you want

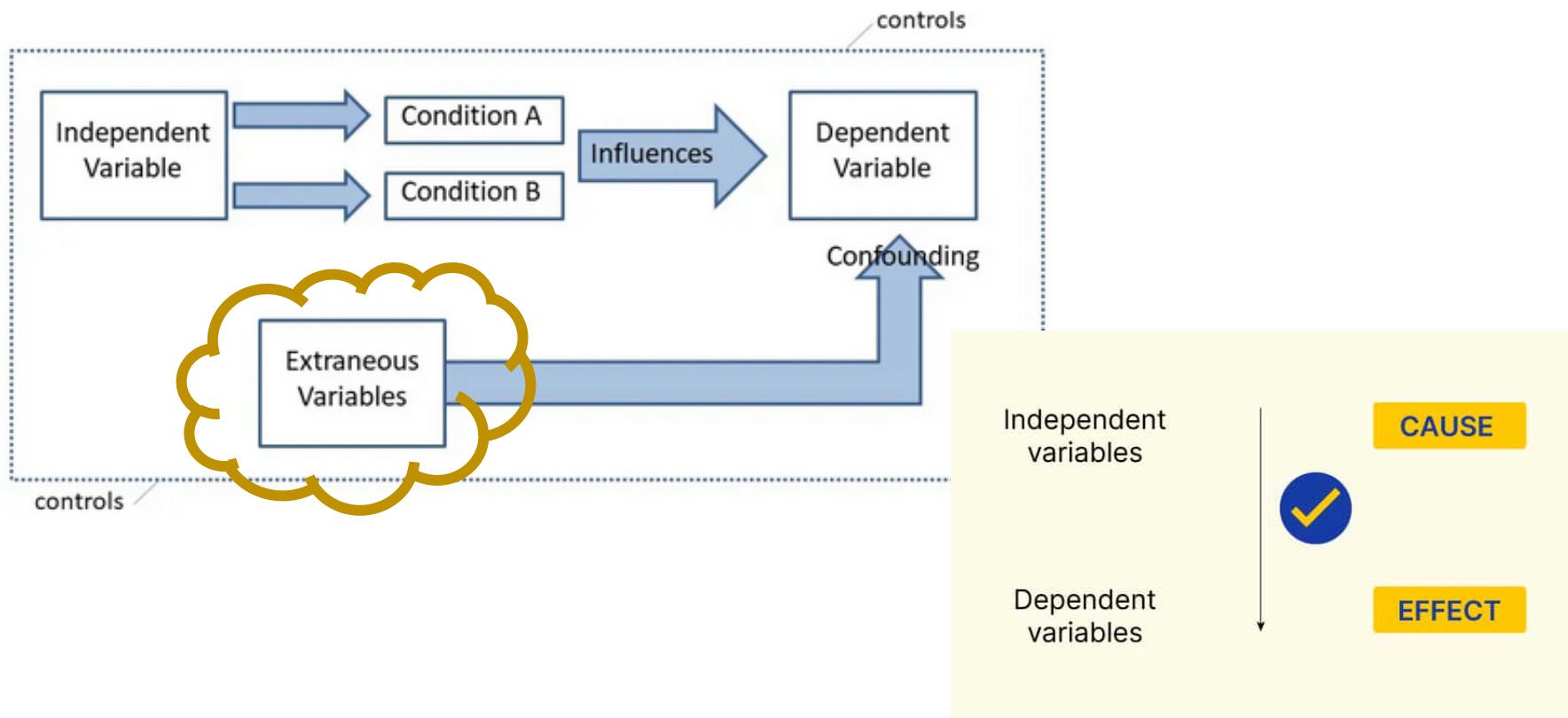


What you get

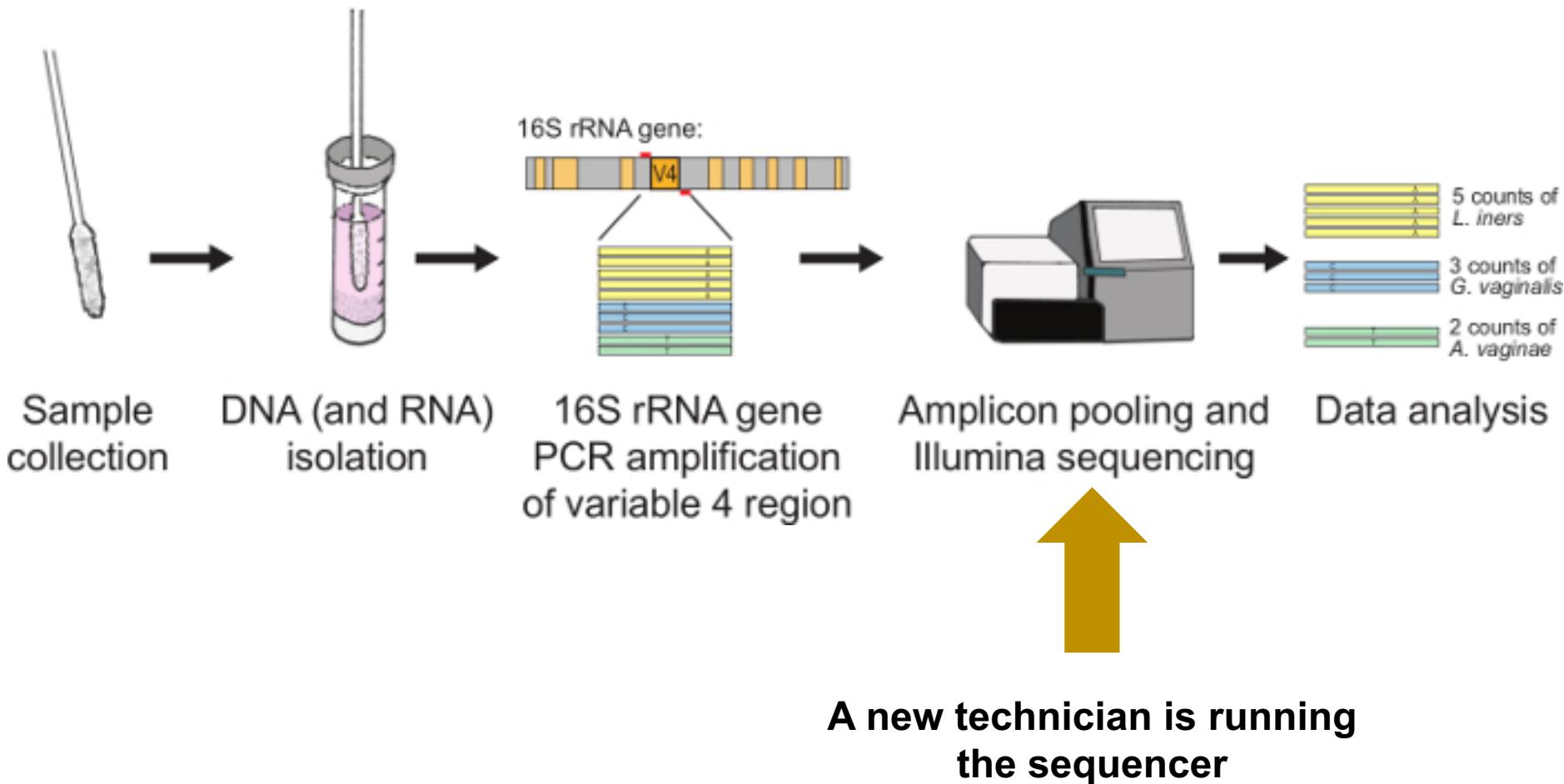
That spare comes in handy
Highly recommend especially
with mice!

What causes this? Confounding variables

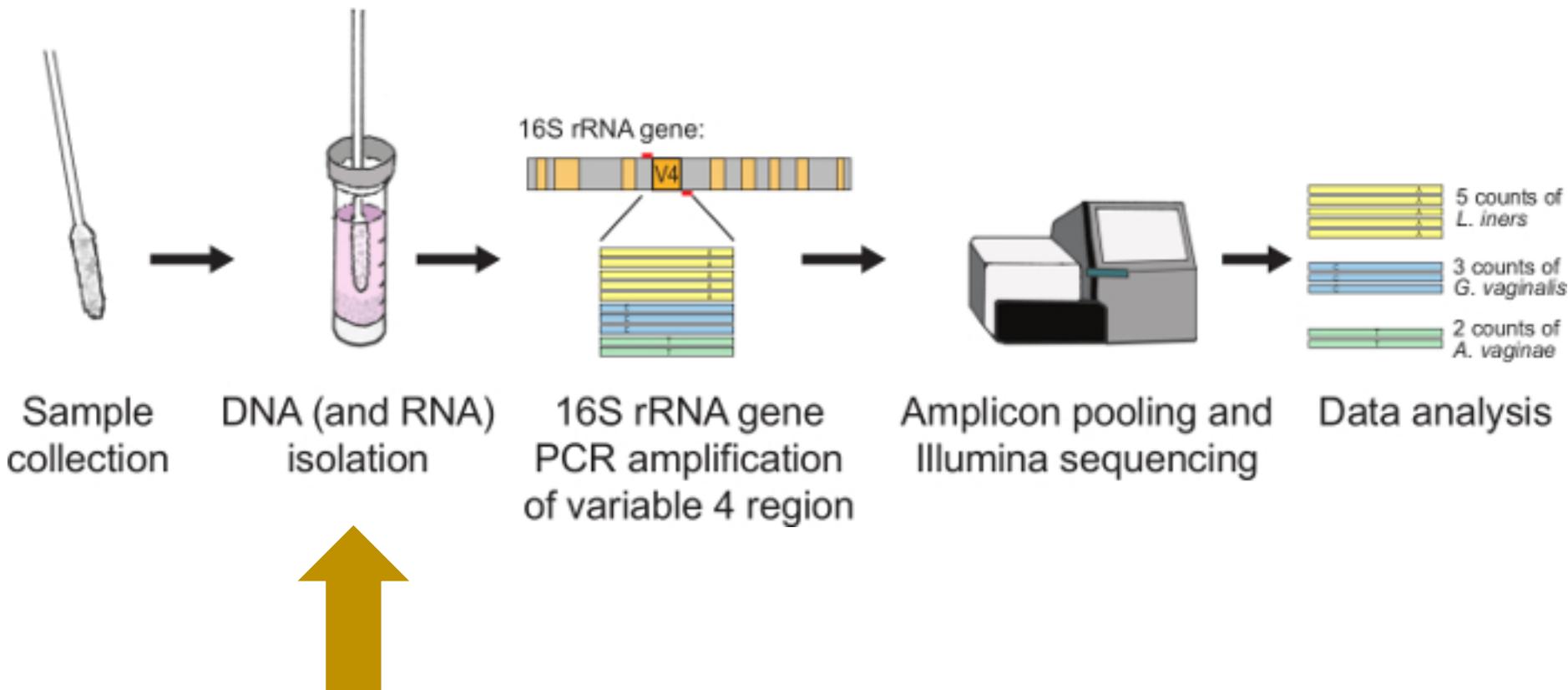
A variable that influences or *confounds* the relationship between an independent and dependent variable



Examples of confounding variables

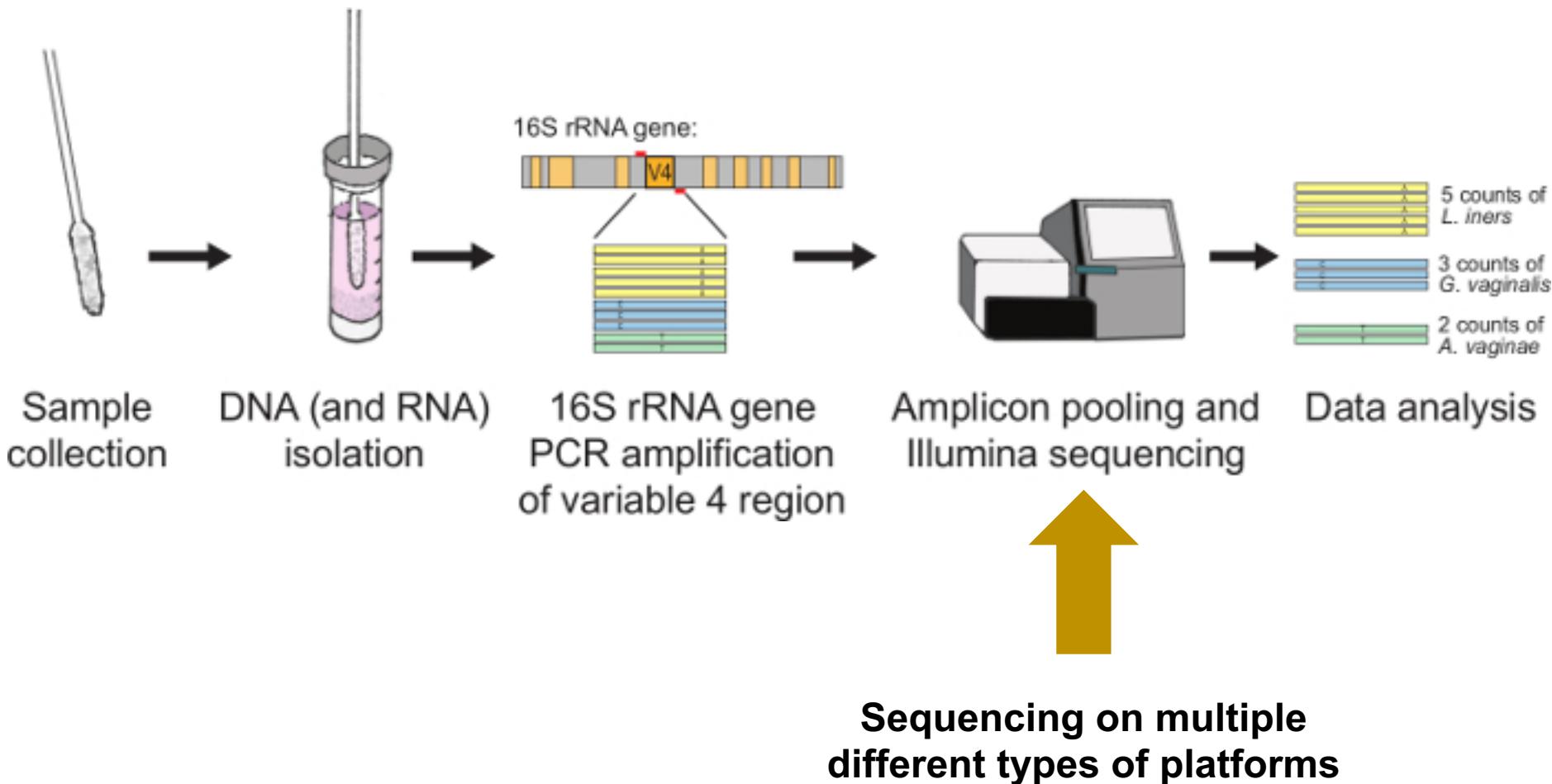


Examples of confounding variables

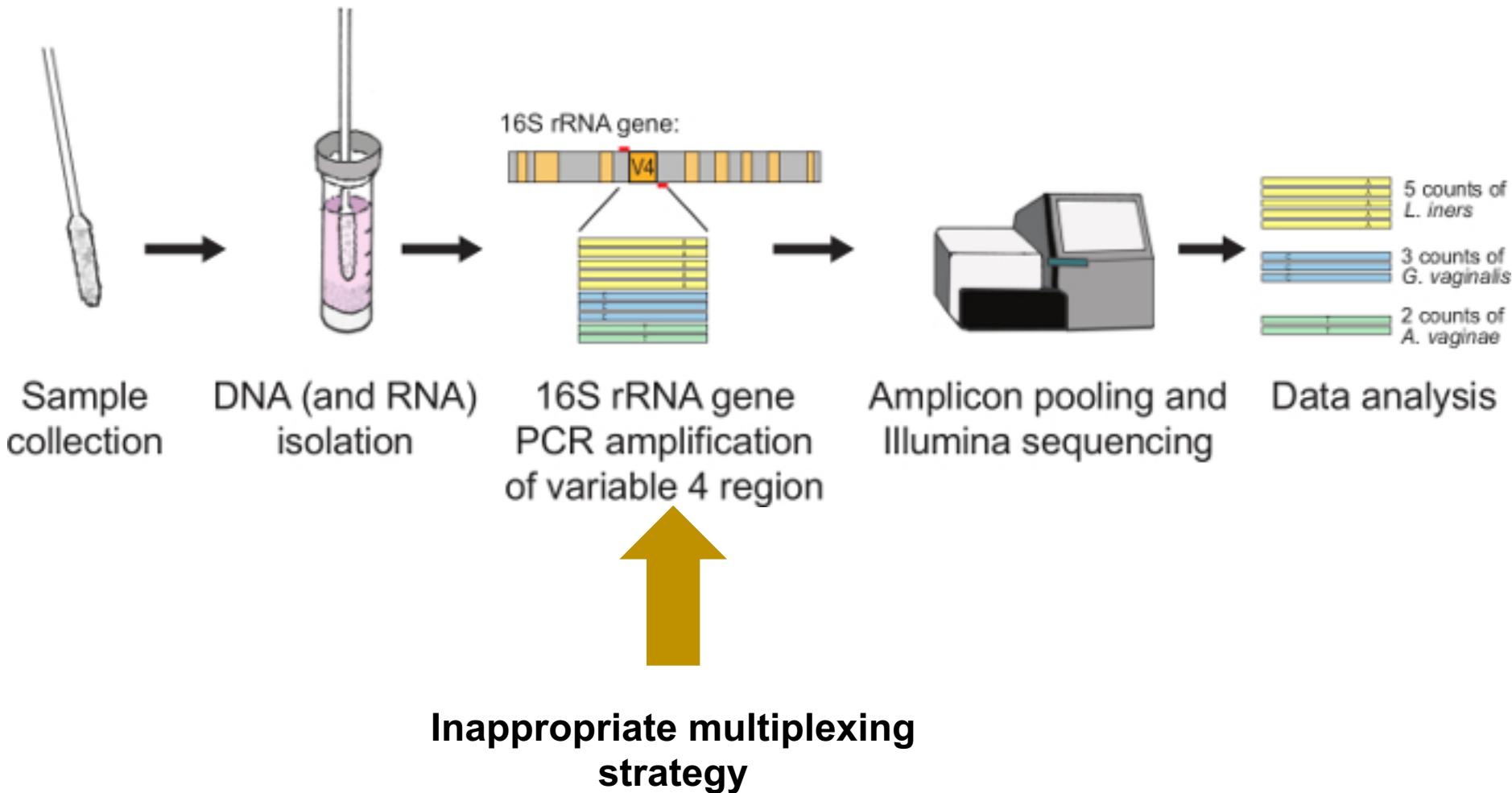


Extracting DNA/RNA with two different kits!

Examples of confounding variables



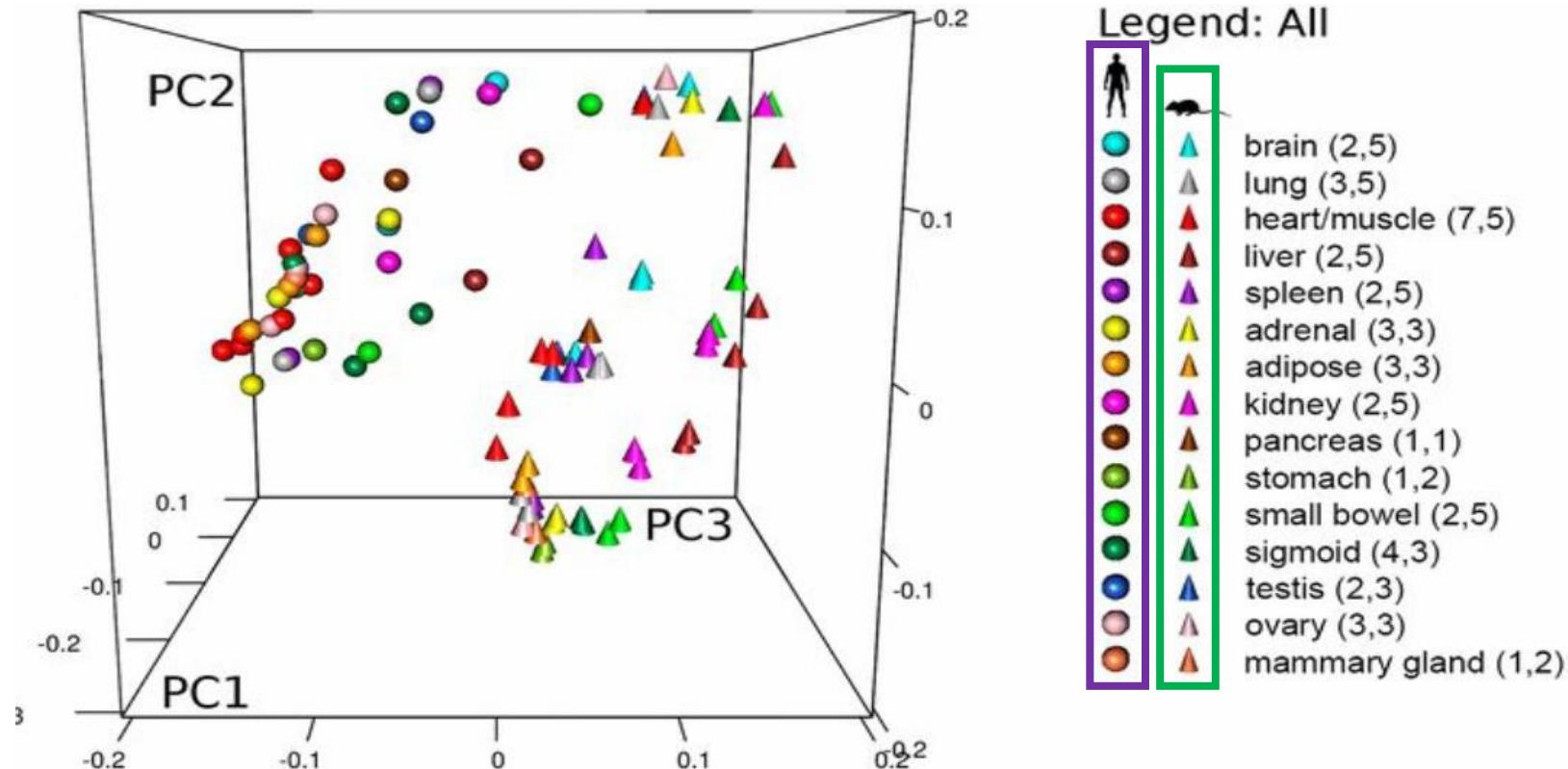
Examples of confounding variables



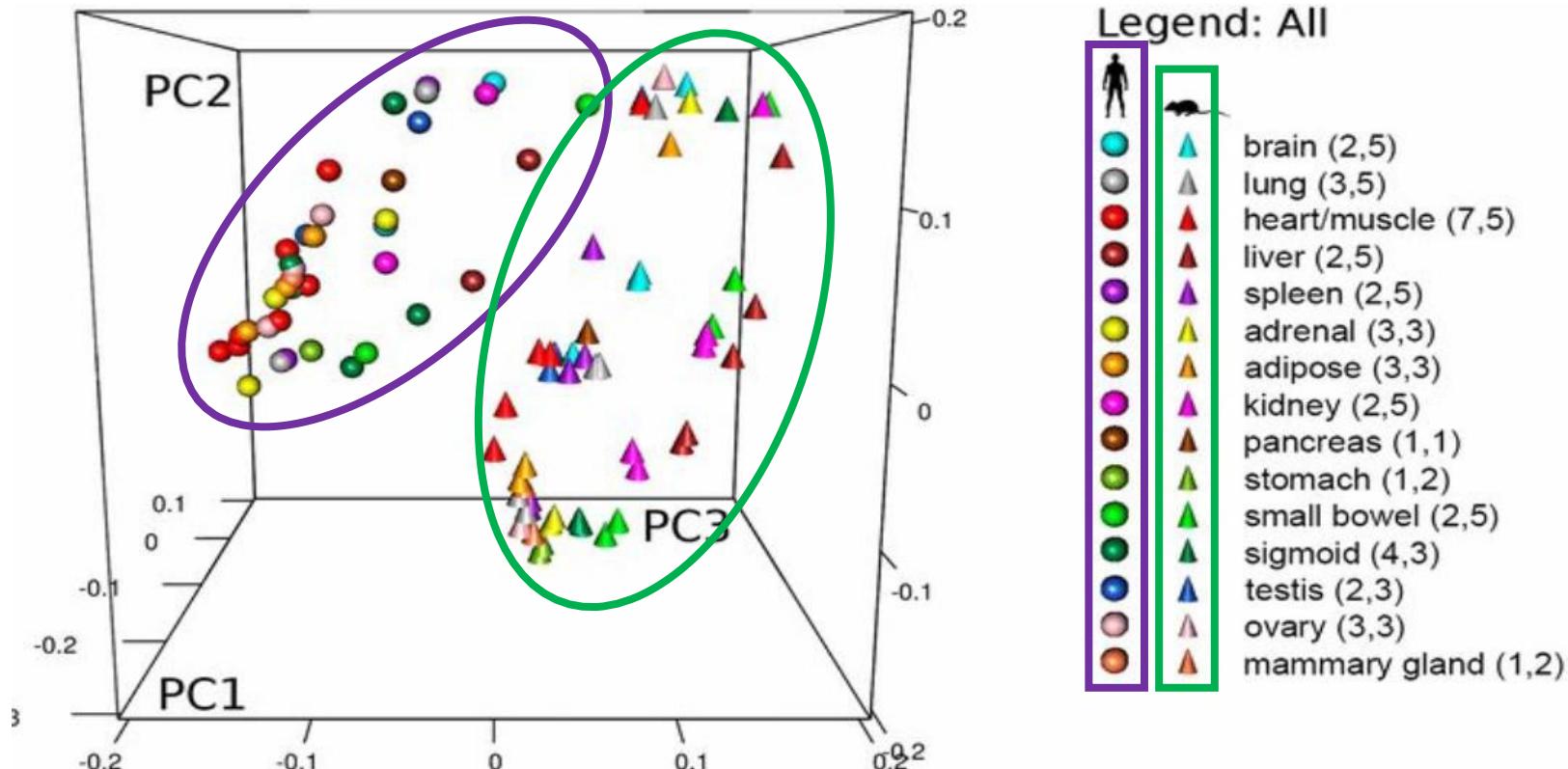
What this means?

- It's 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 goes a long way**

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



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



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.

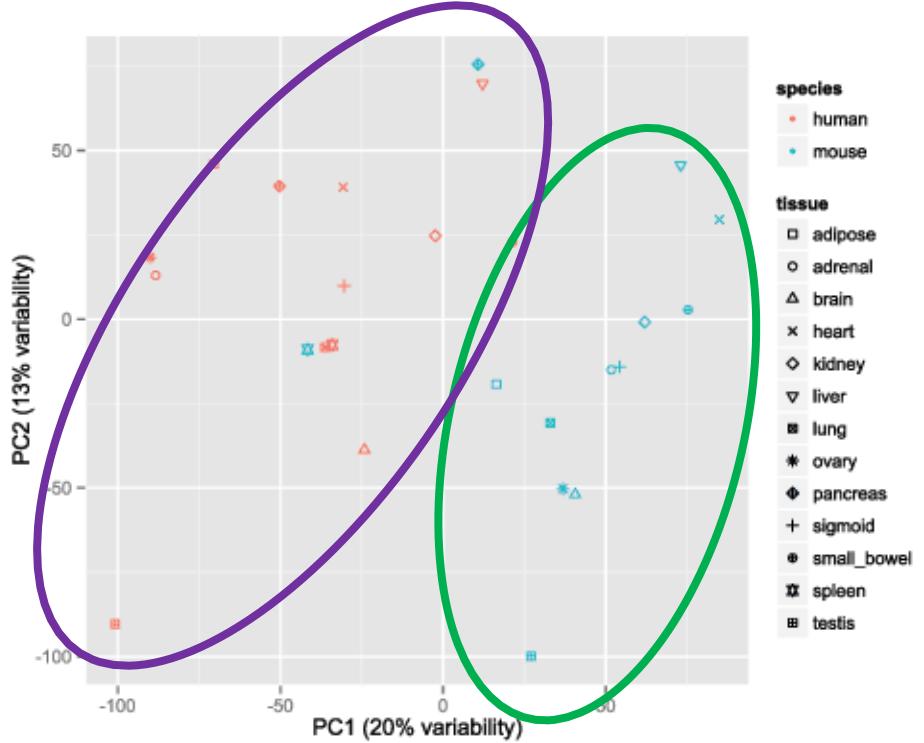
Sequence study design (sequencer ID, run ID, lane number):

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

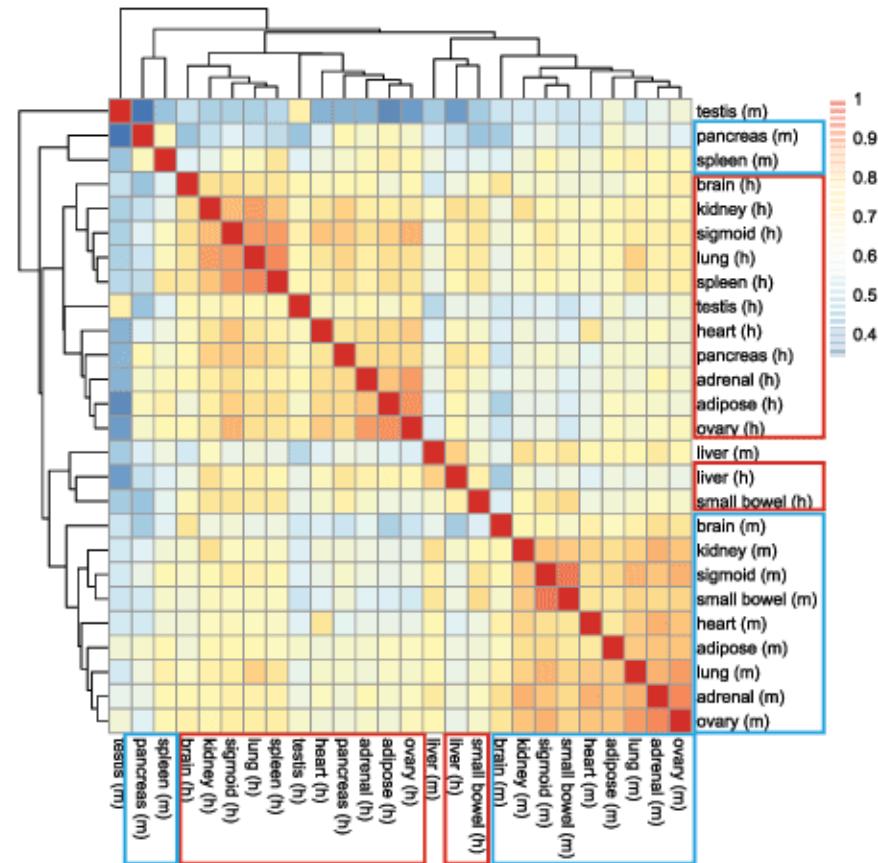
Before accounting for batch effect

a



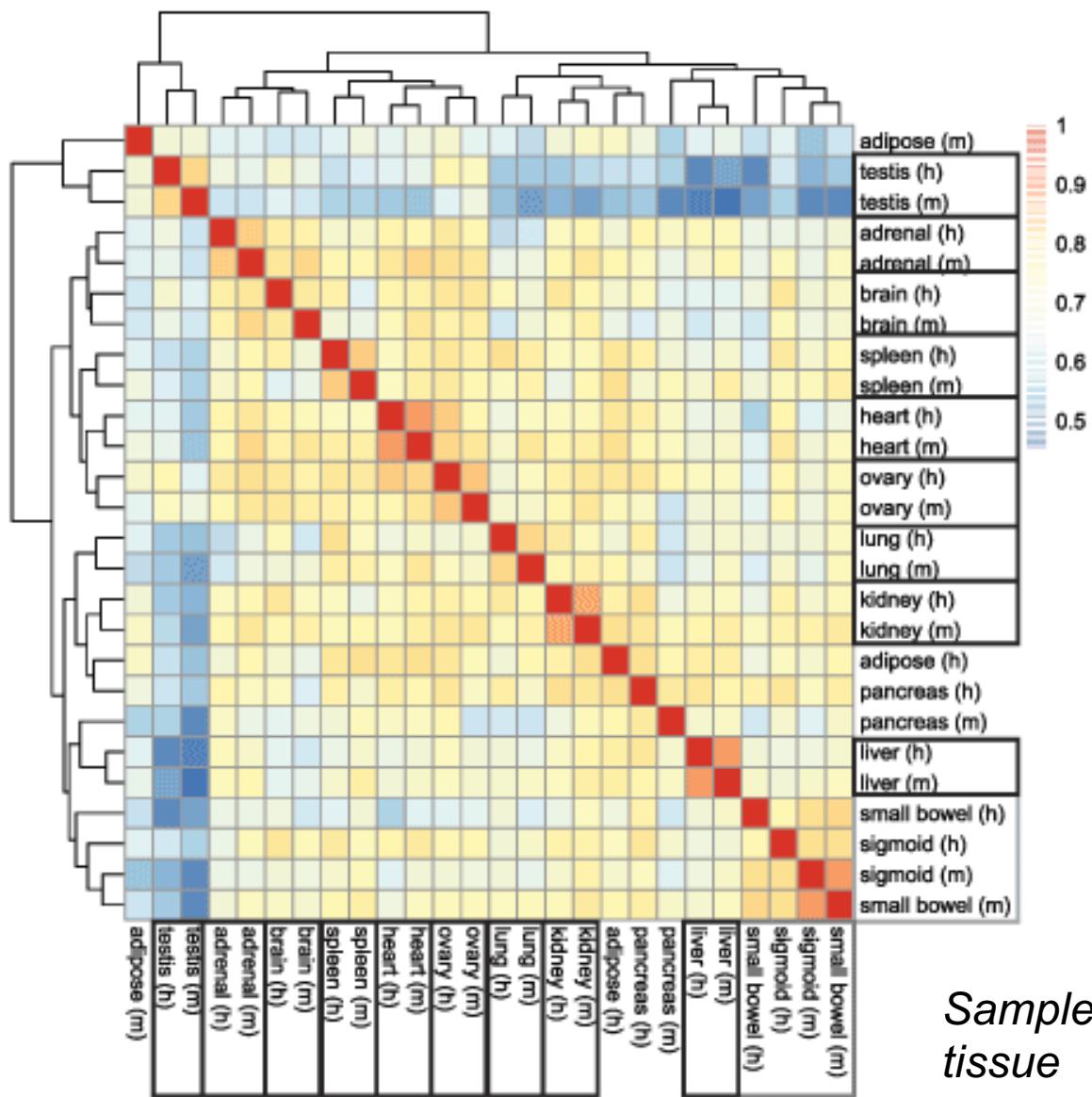
Samples grouped by animal

b



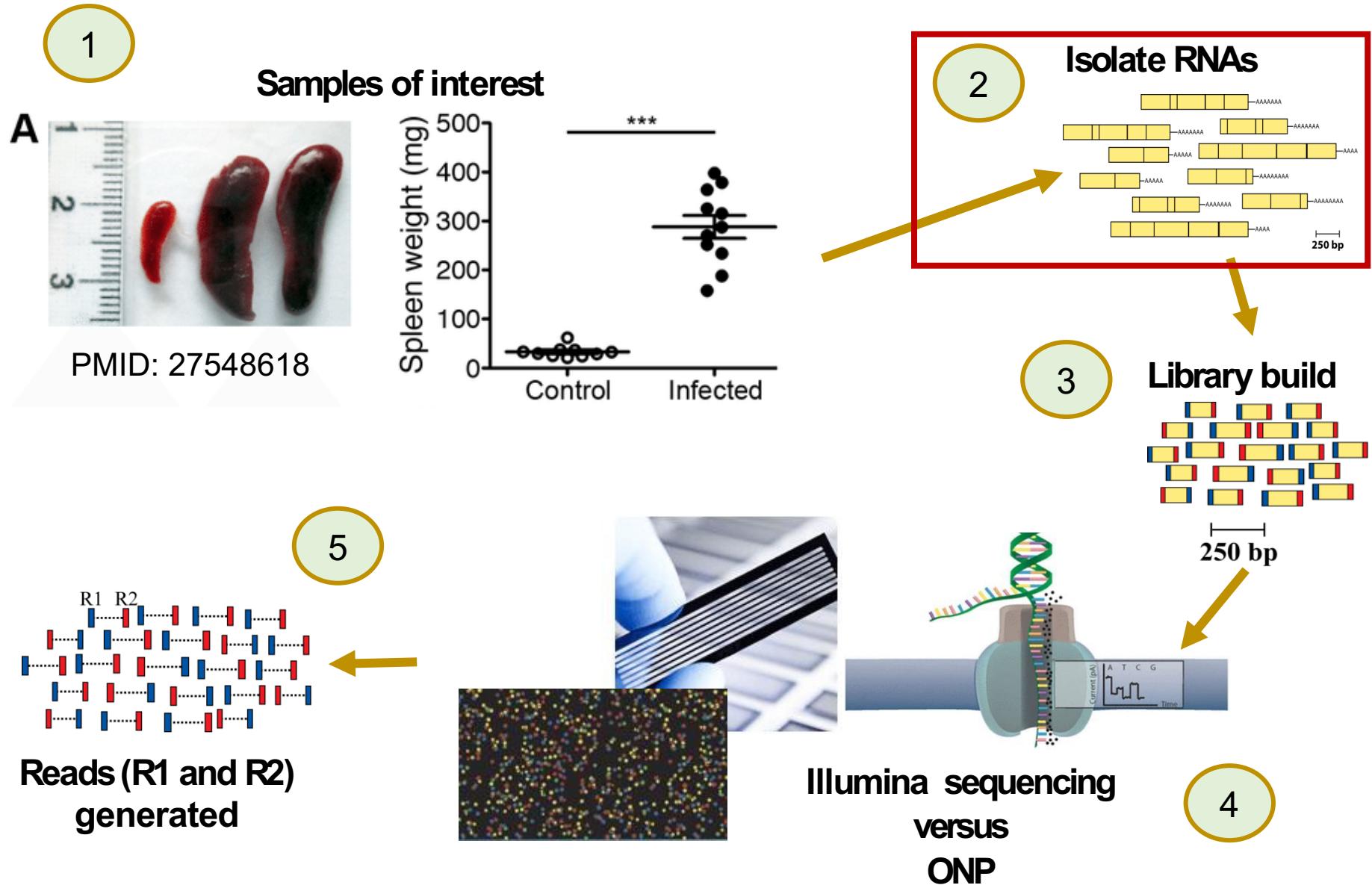
b

After accounting for batch effect



Samples now grouped by tissue

Experimental workflow



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



RNeasy Plus Mini Kit (50)

Cat. No. / ID: 74134

For 50 minipreps: RNeasy Mini Spin Columns, gDNA Eliminator Spin Columns, Collection Tubes, RNase-Free Water and Buffers

\$435.00

[Log in to see your account pricing.](#)

1. Column type / Plate type

Micro

Mini

96 well

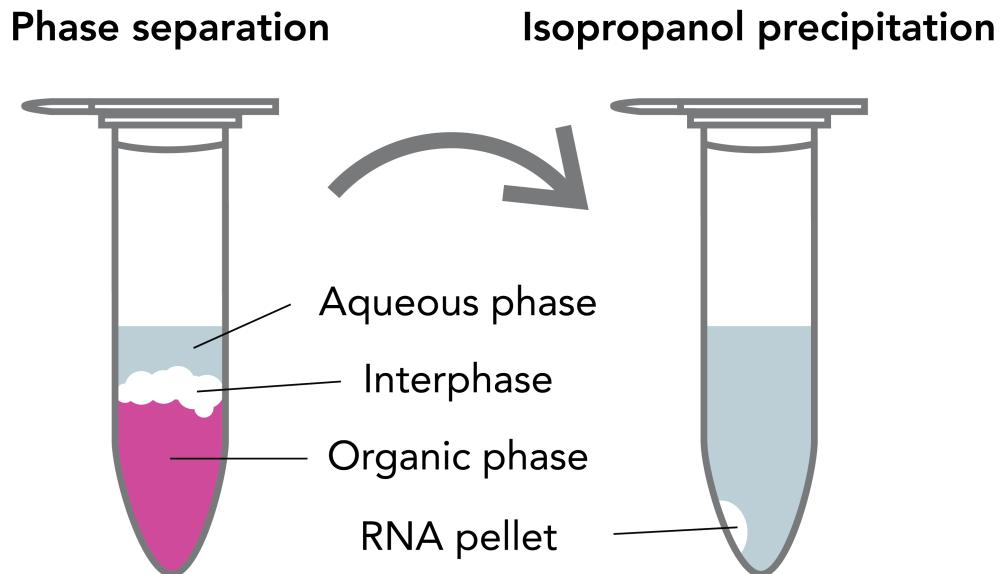
2. Preparations

50

250

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

The cheaper way to isolate RNA



- The UVM VIGR core **does not** recommend as residual phenol will inhibit the library build and you cannot get rid of it. They will not sequence your samples if you use this mode of isolation!

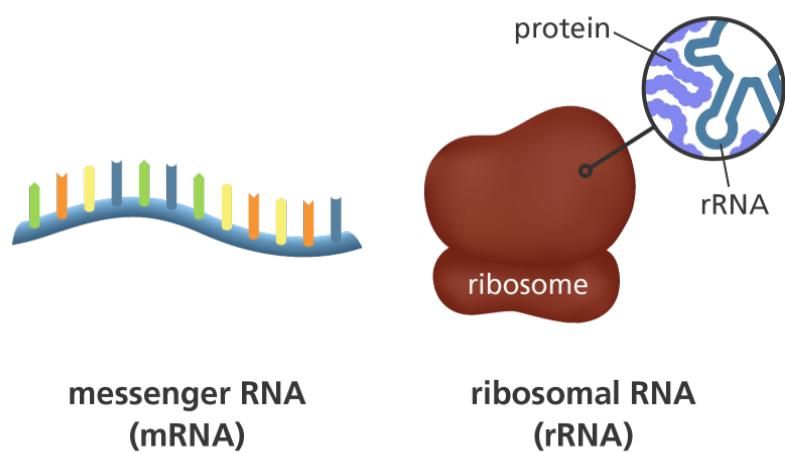
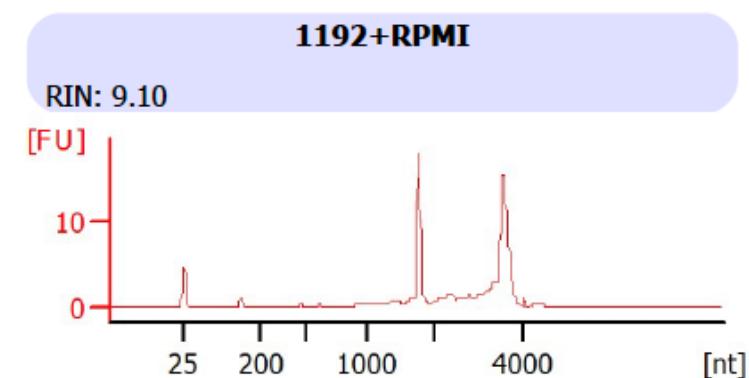
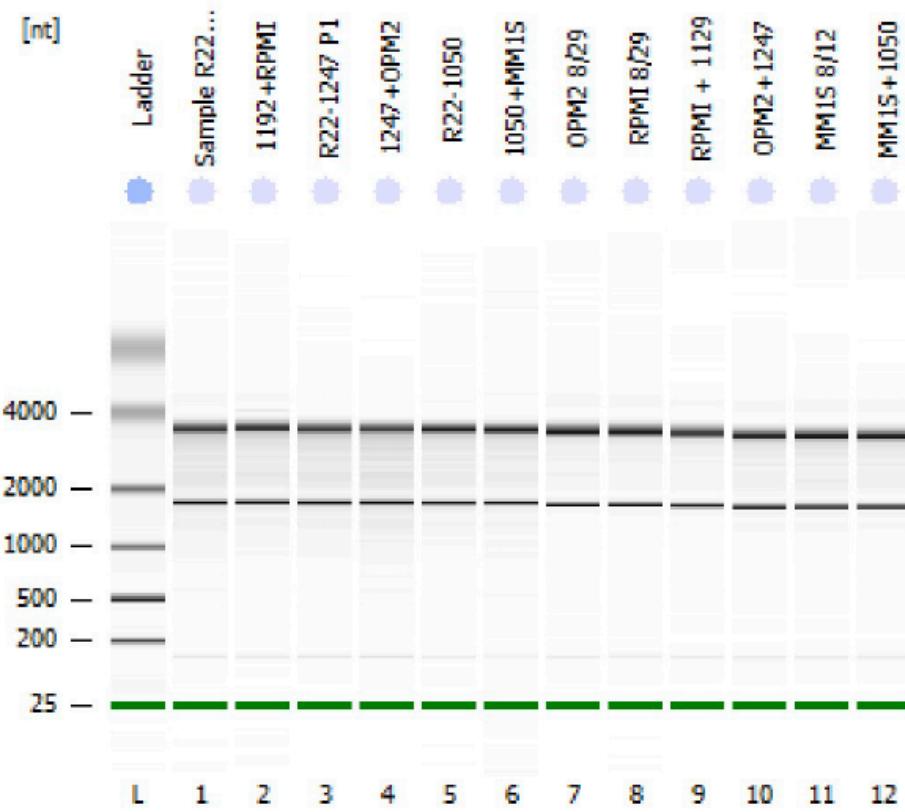
Other recommendations

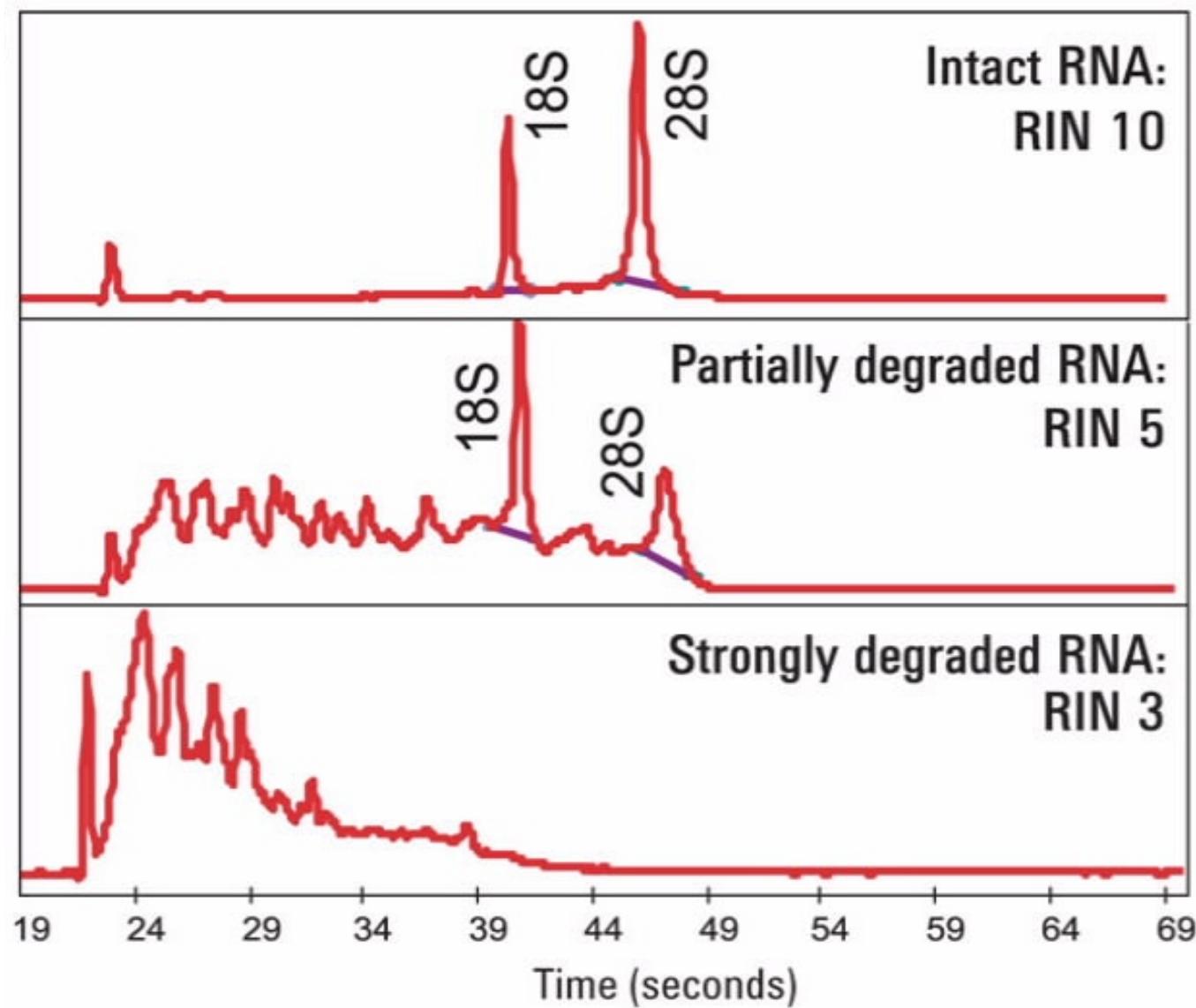
- Maintain RNA integrity with **Ribolock** (ThermoFisher: EO0381)
- Once isolated, RNA is stored in the -80C



Check purity with RNA bioanalyzer

Electrophoresis File Run Summary





RNA Bioanalyzer requires Qubit

- Need to take a Qubit reading not NanoDrop



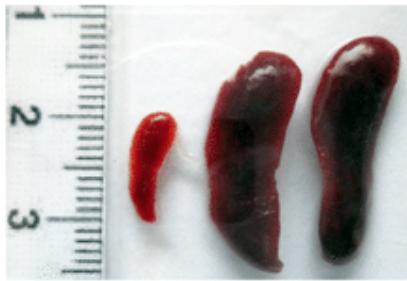
NO!

Experimental workflow before it gets sequenced

1

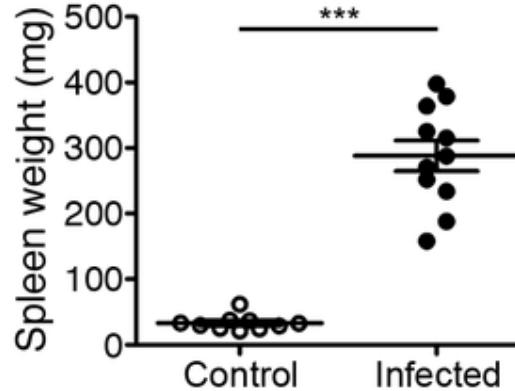
Samples of interest

A



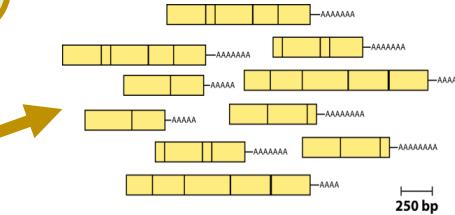
PMID: 27548618

Samples of interest



2

Isolate RNAs

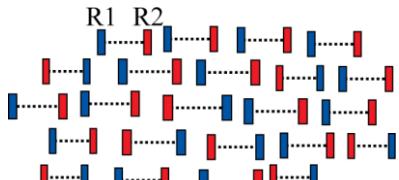


3

Library build

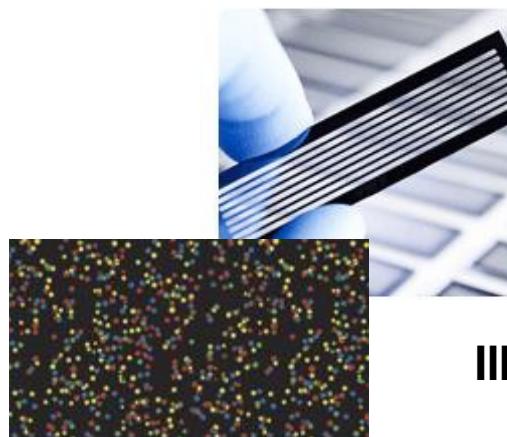


250 bp



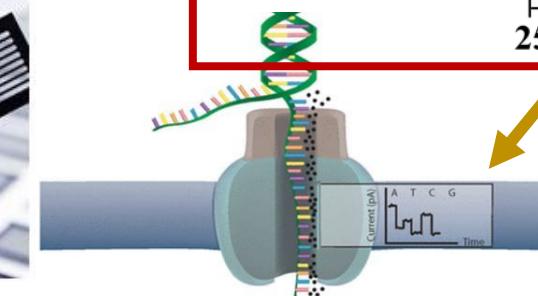
Reads (R1 and R2)
generated

5



Illumina sequencing
versus
ONP

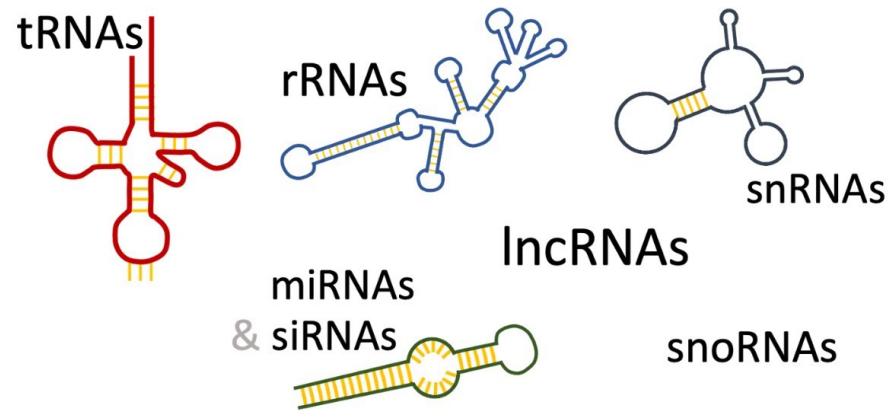
4



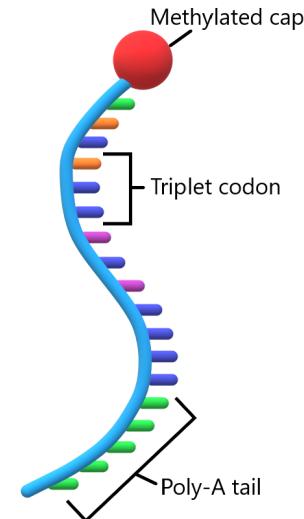
RNA composition

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

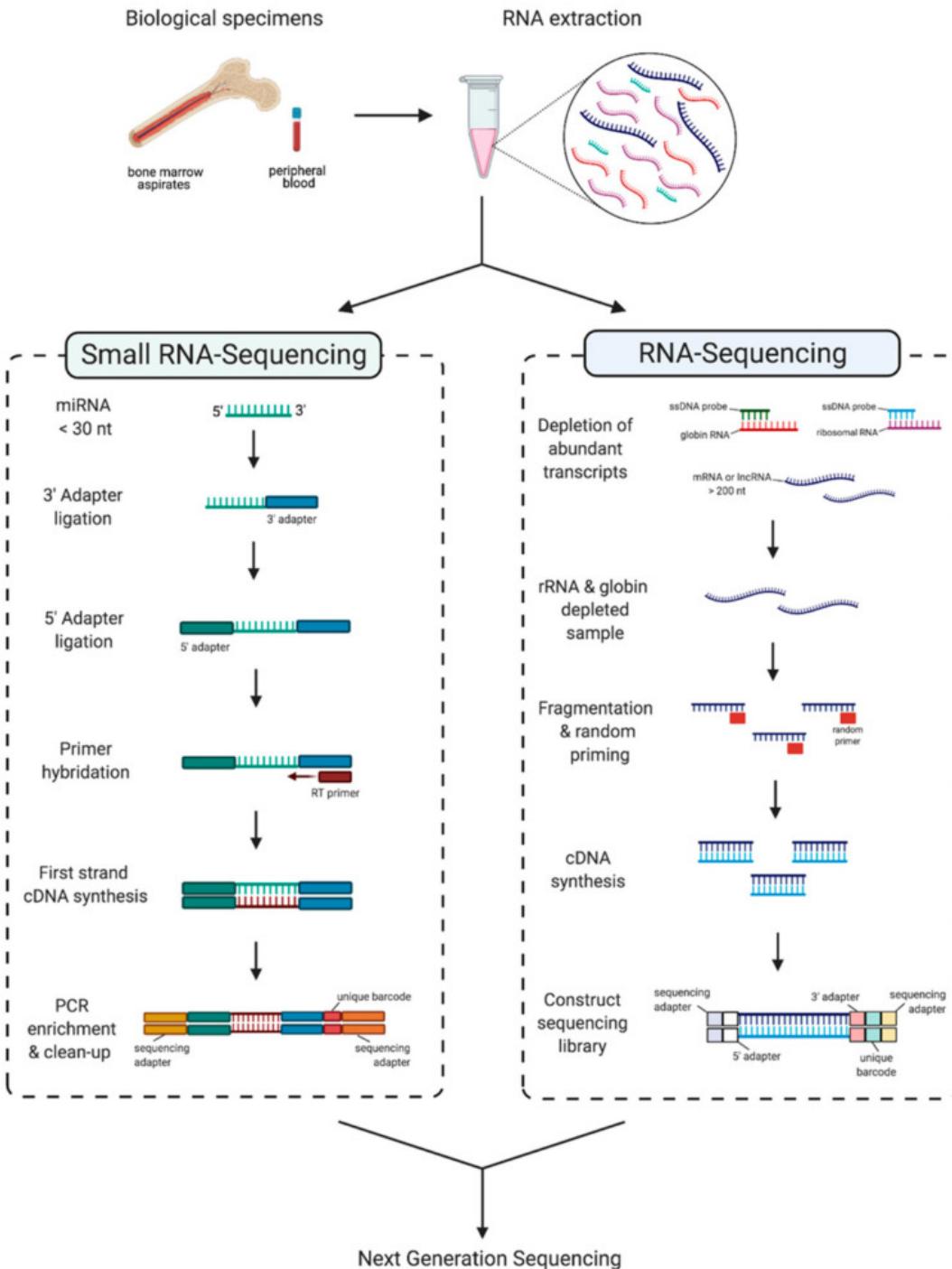


messenger RNA



“Comprehensive” transcriptome analysis

Two different protocols/kits!

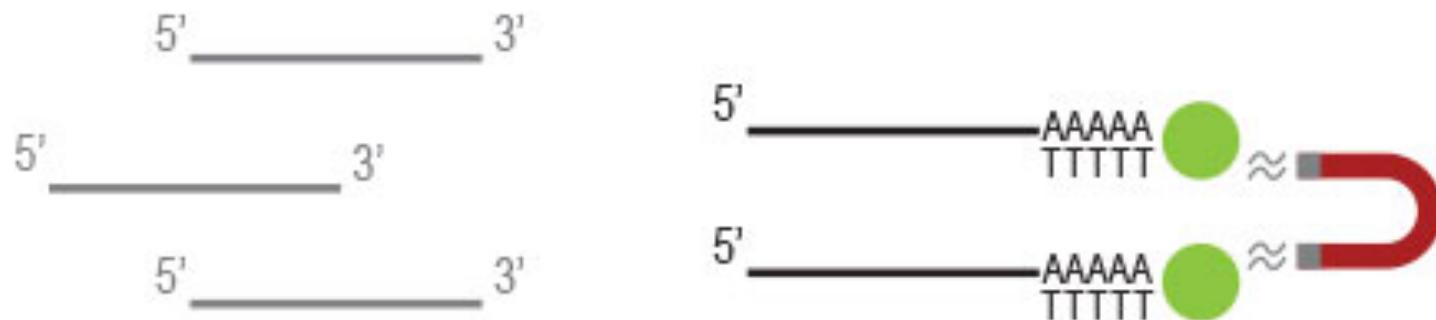


Before building the libraries need to perform target enrichment

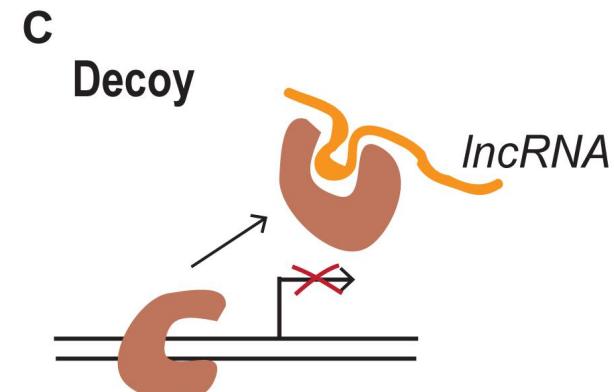
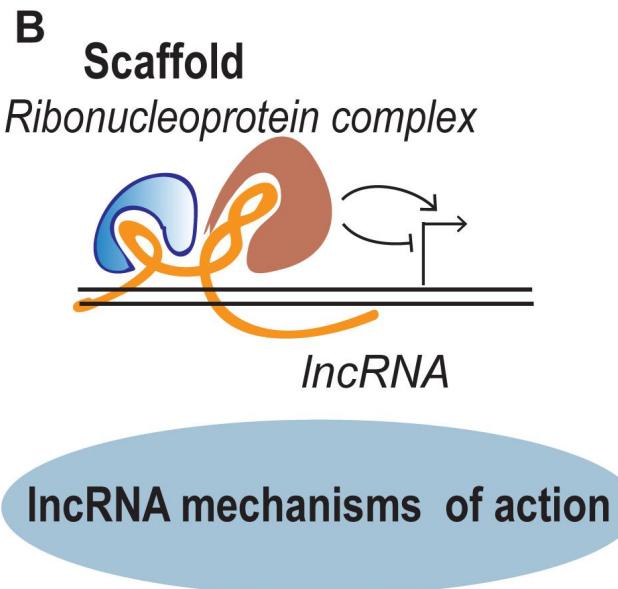
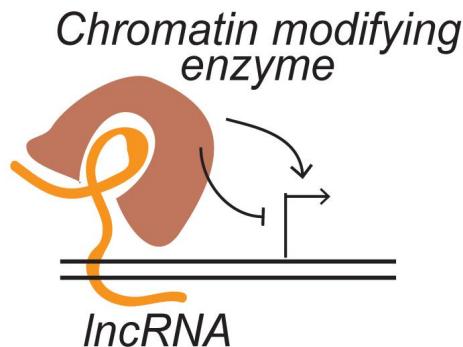
- If rRNA is not removed, the majority of the final sequencing reads would be from rRNA and not mRNA
- Therefore, it necessary to enrich for mRNA
- Two common strategies:
 1. Poly A+ selection = mRNA only
 2. rRNA depletion = mRNA + *other species*

Poly-A versus rRNA depletion?

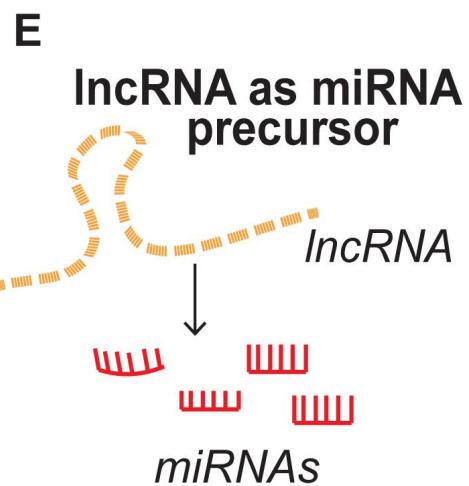
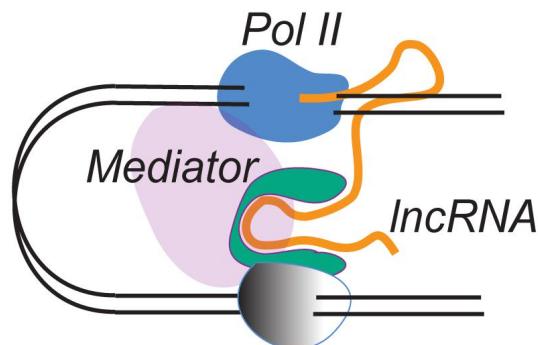
- For differential gene expression, most enrich using Poly(A)+
- However:
 - If you are aiming to obtain information about long non-coding RNA's perform ribosomal RNA depletion
 - Bacterial mRNAs are also not poly-adenylated



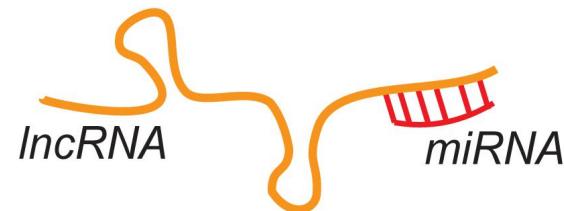
A Guide



F Chromatin looping



D lncRNA sponging miRNA



Strandedness



Another consideration is whether to generate strand-preserving libraries

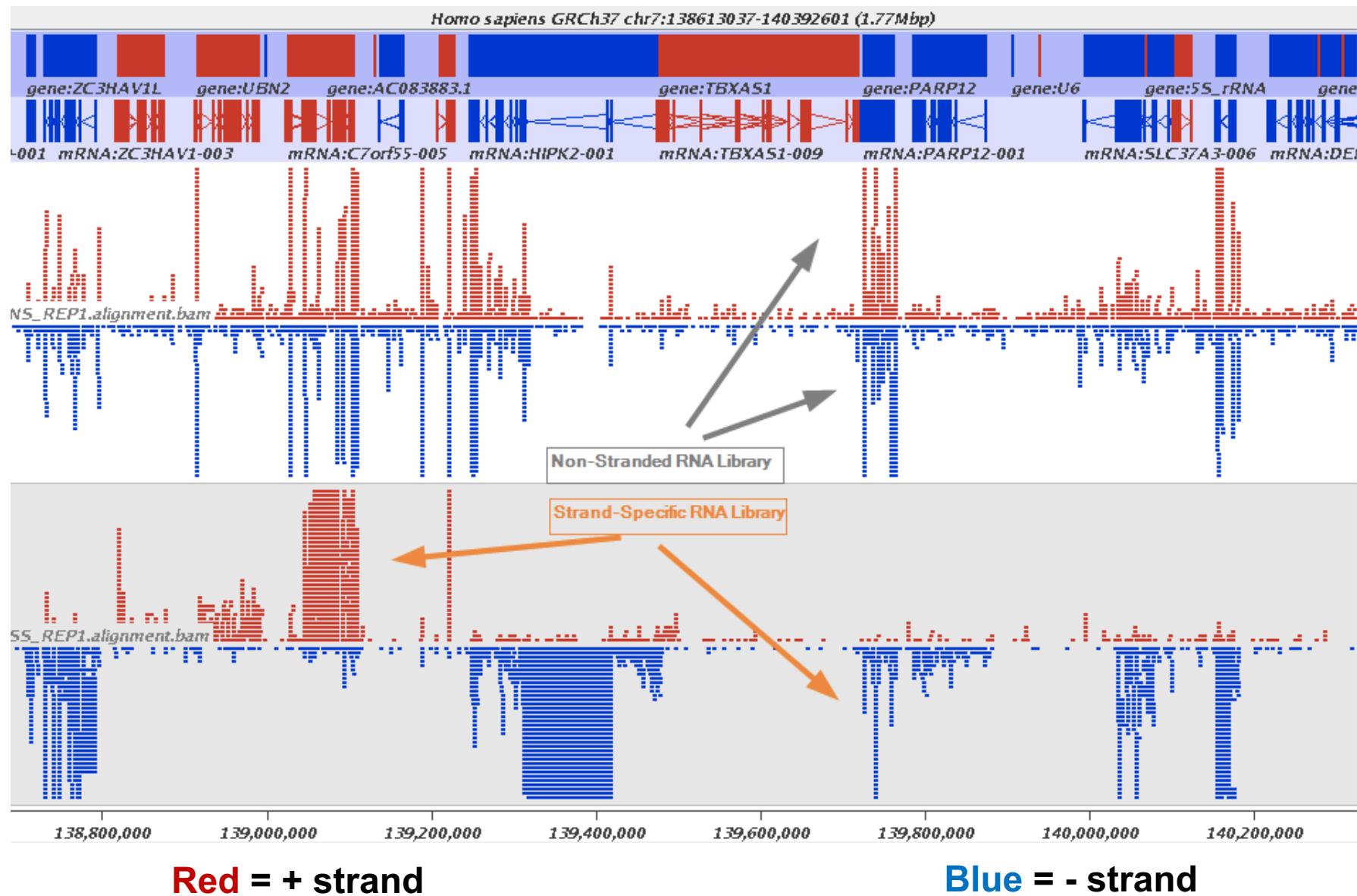


Libraries can be stranded or unstranded

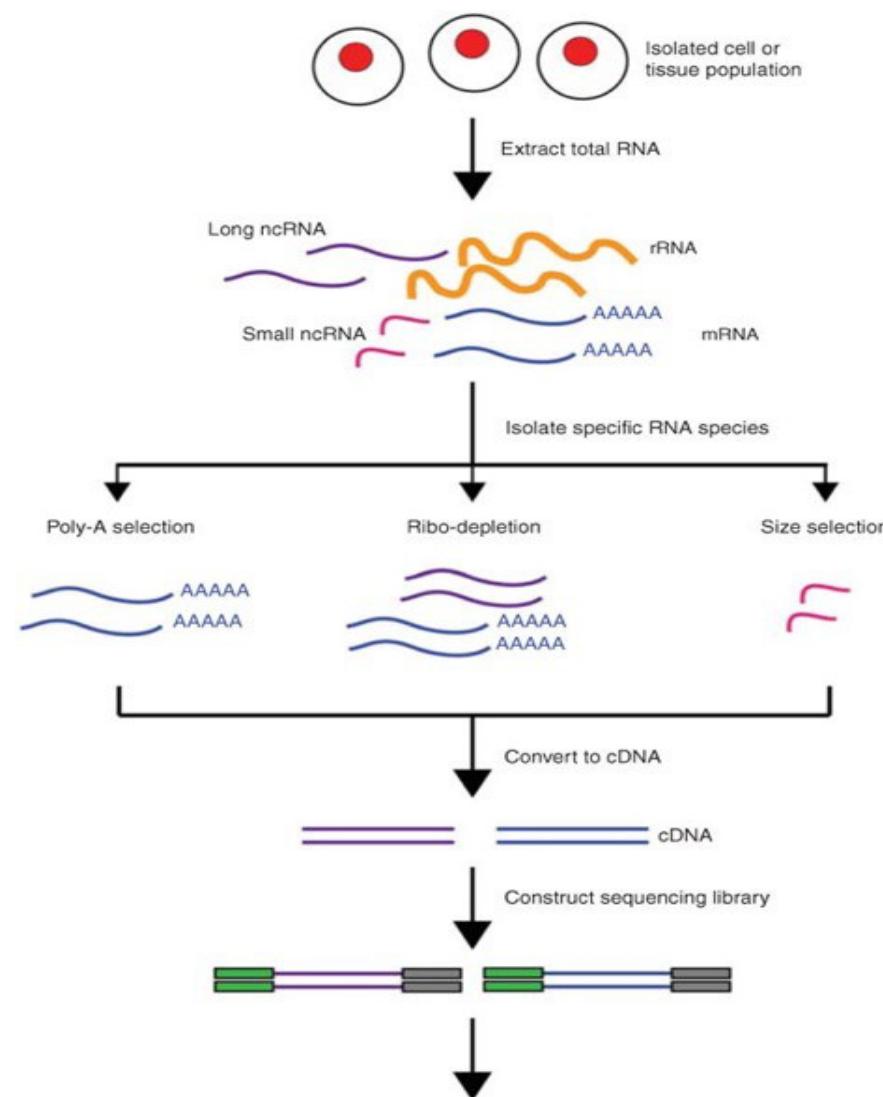


The implication of **stranded** libraries is that you could distinguish whether the reads are derived from forward or reverse-encoded transcripts

Strandedness



RNA-seq Library build steps



Isolate RNA



RNA species diversity



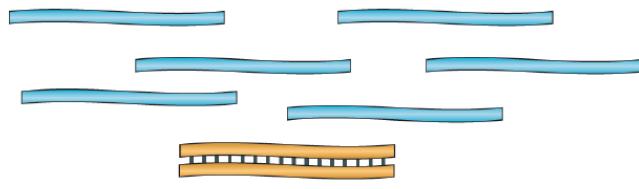
target enrichment



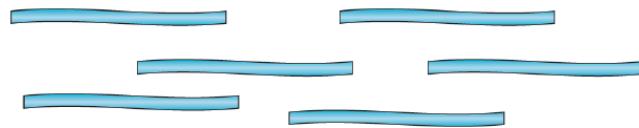
cDNA conversion /
PCR amplification

Library Prep steps

① mRNA or total RNA

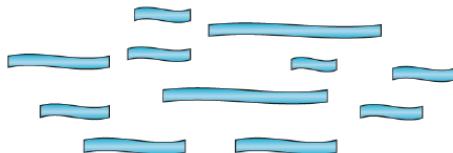


② Remove contaminant DNA

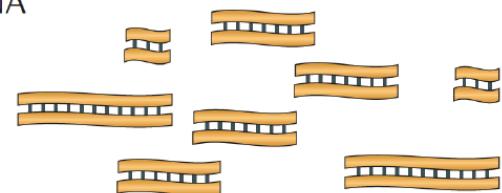


Remove rRNA?
Select mRNA?

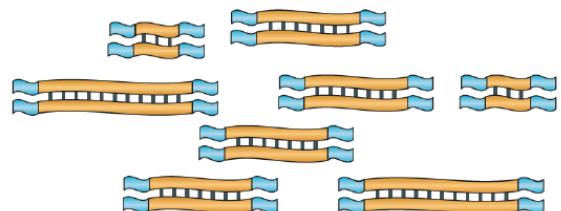
③ Fragment RNA



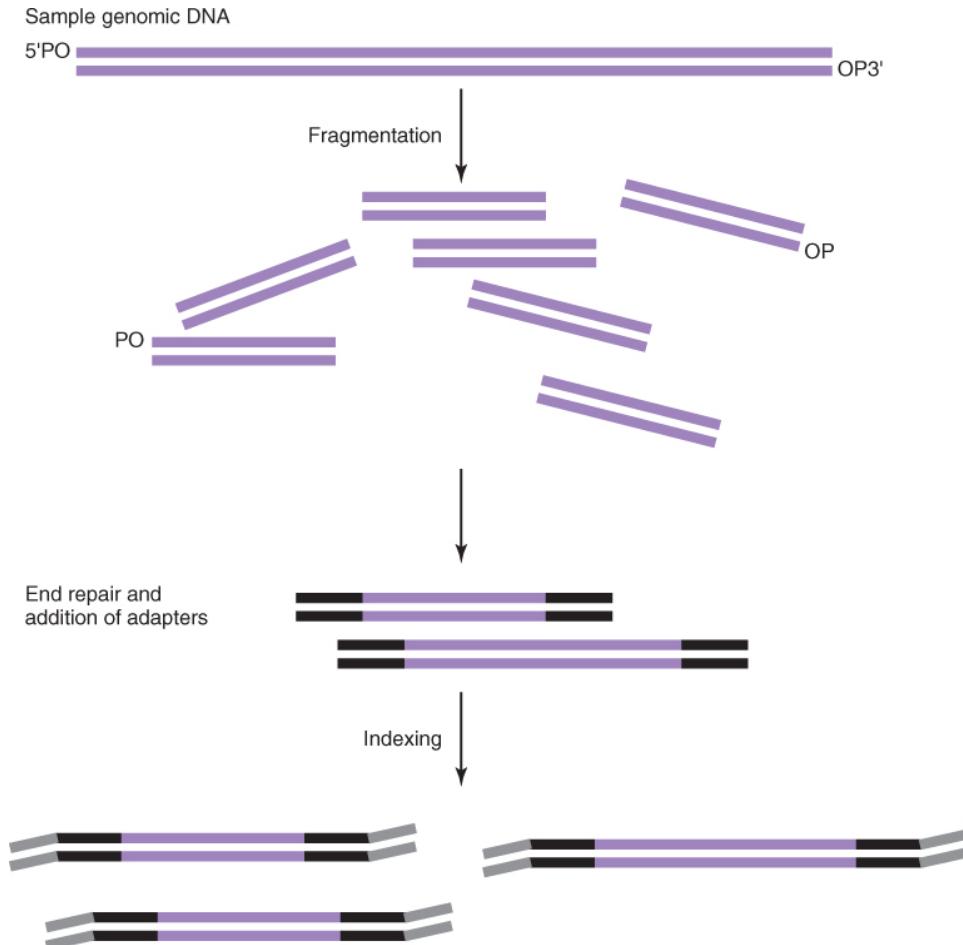
④ Reverse transcribe
into cDNA



⑤ Ligate sequence adaptors



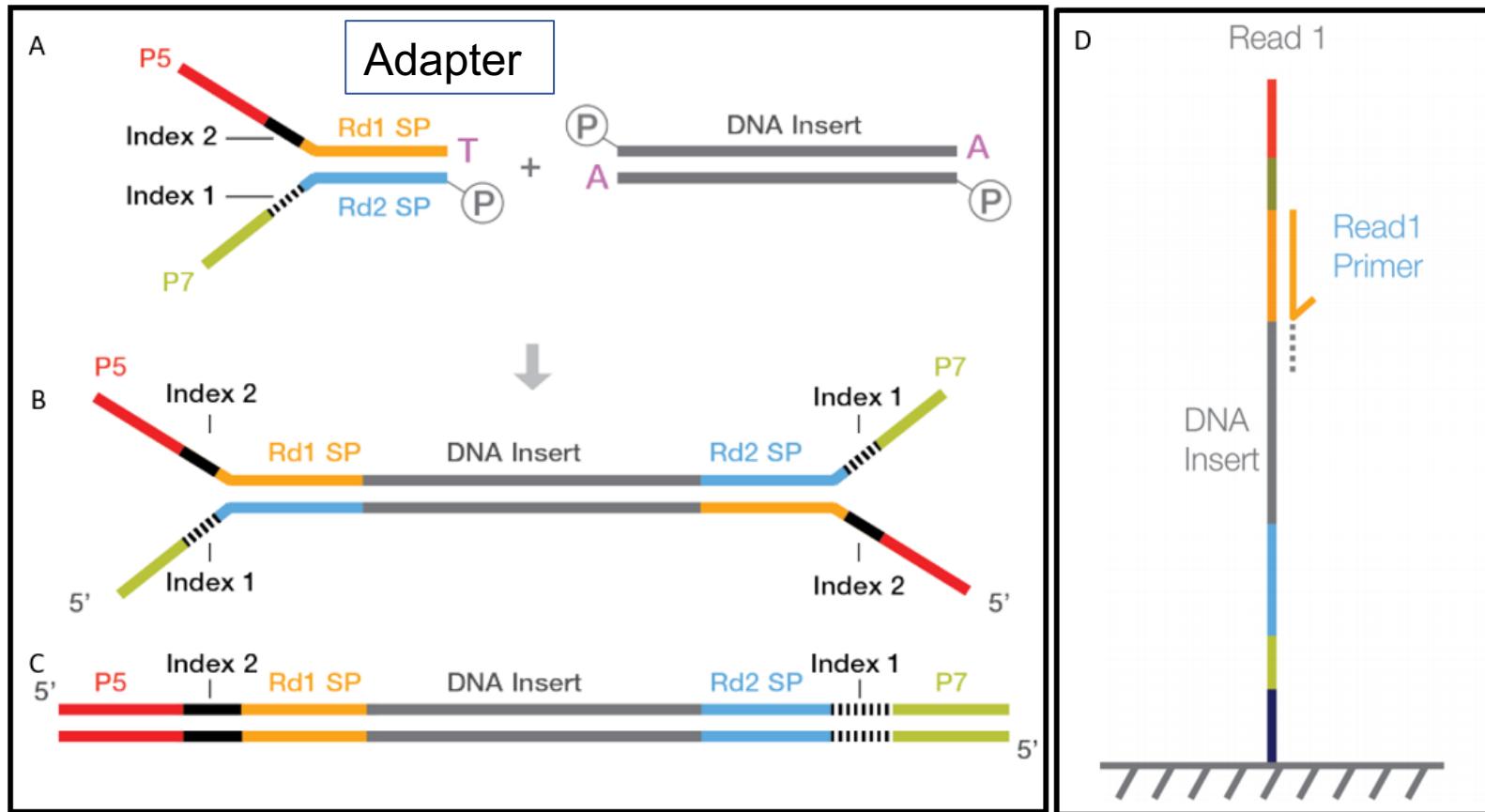
Adapters vs Index



- 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

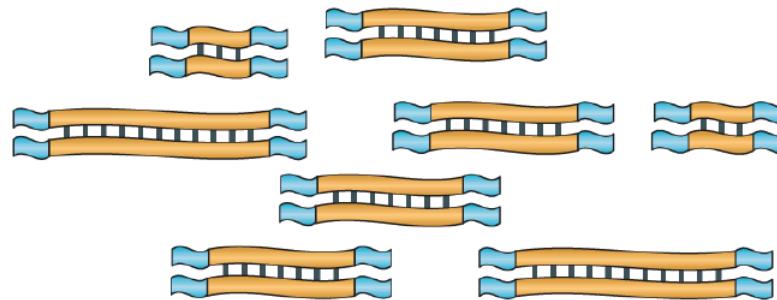
P5 and p7 sequences are required to bind the flow cell



Unique index

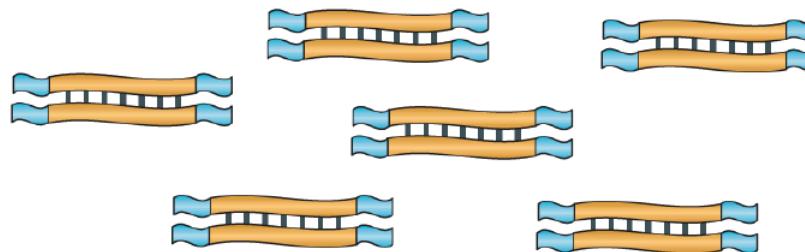
Library Prep steps continued

⑤ Ligate sequence adaptors



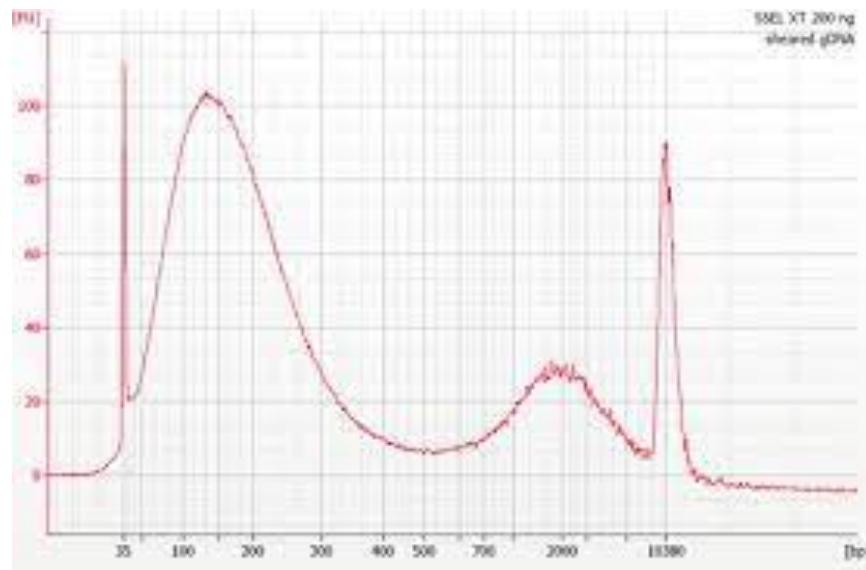
PCR amplification?

⑥ Select a range of sizes



Common mistakes during library build

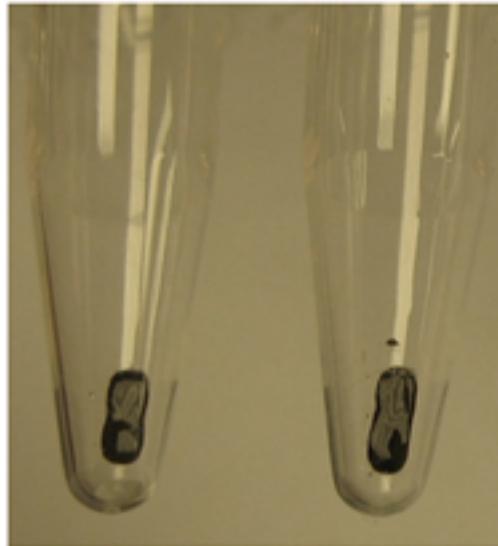
- Addition of adapters and unique barcodes
- Number of PCR cycles



Common mistakes during library build

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

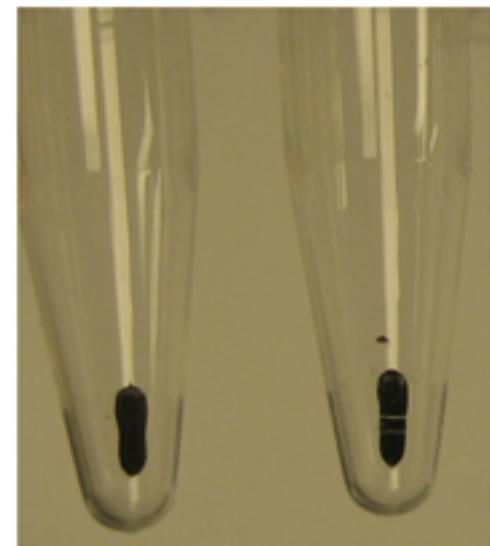
Shiny Wet Pellet



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

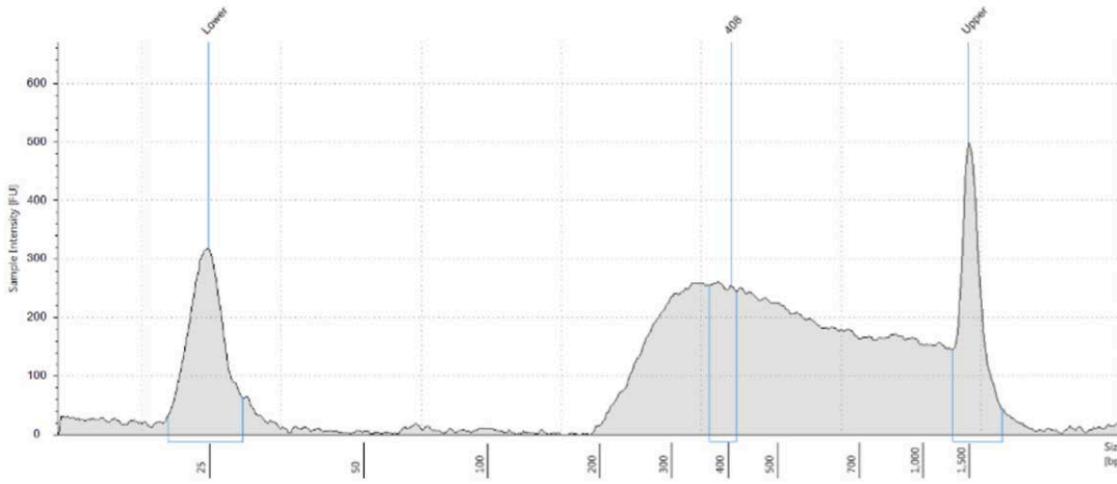


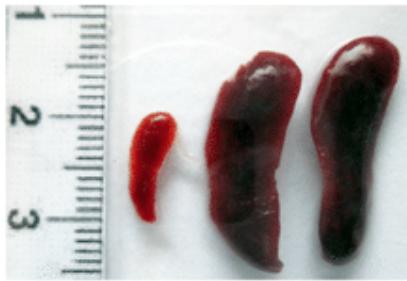
Figure 4. An electropherogram showing a sample with incomplete fragmentation due to suboptimal mixing. There are large fragments present which will be unable to undergo bridge amplification during cluster generation on the flow cell.

Experimental workflow before it gets sequenced

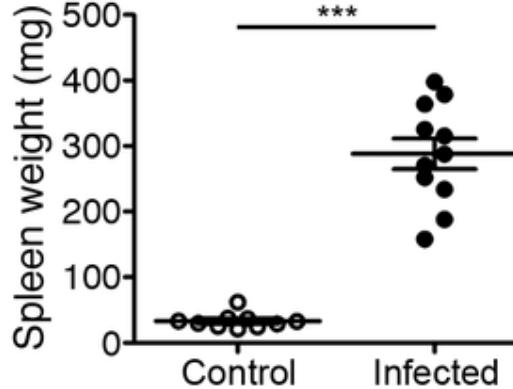
1

Samples of interest

A

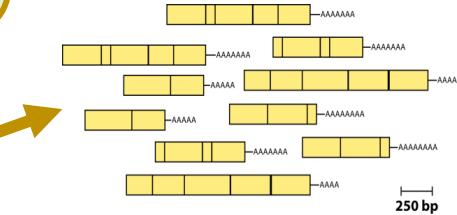


PMID: 27548618



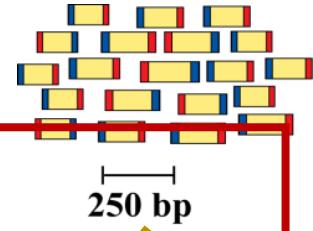
2

Isolate RNAs

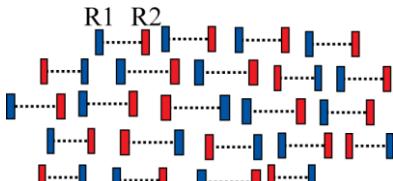


3

Library build

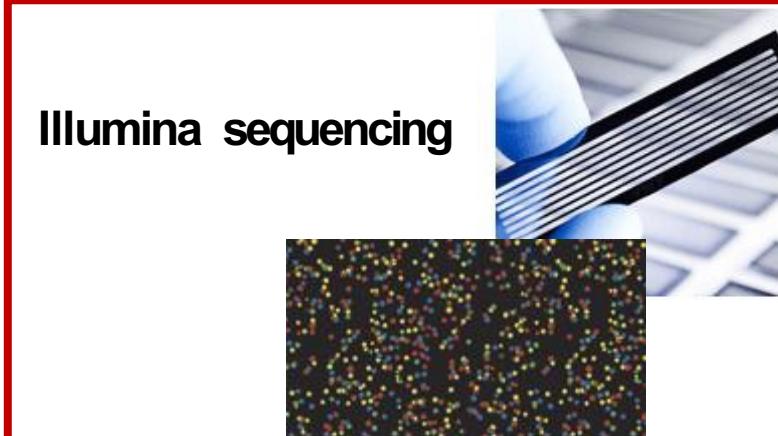


5



Reads (R1 and R2)
generated

Illumina sequencing



4

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

THE EVOLUTION OF SEQUENCING

First Generation

Sanger Sequencing
Maxam and Gilbert
Sanger Chain-termination

- Infer nucleotide identity using dNTPs then visualize with electrophoresis
- 500-1000 bp fragments
- Relatively slow and expensive

Second Generation Next Generation Sequencing

454, Solexa, Ion Torrent,
Illumina

- High throughput from the parallelization of sequencing reactions
- High accuracy
- ~50-500 bp fragments
- Faster and more affordable

Third Generation

PacBio, Oxford Nanopore

- Sequence native DNA in real time with single-molecule resolution
- Traditionally lower accuracy than NGS
- Tens of kb fragments, on average

Short-read sequencing

Long-read sequencing

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

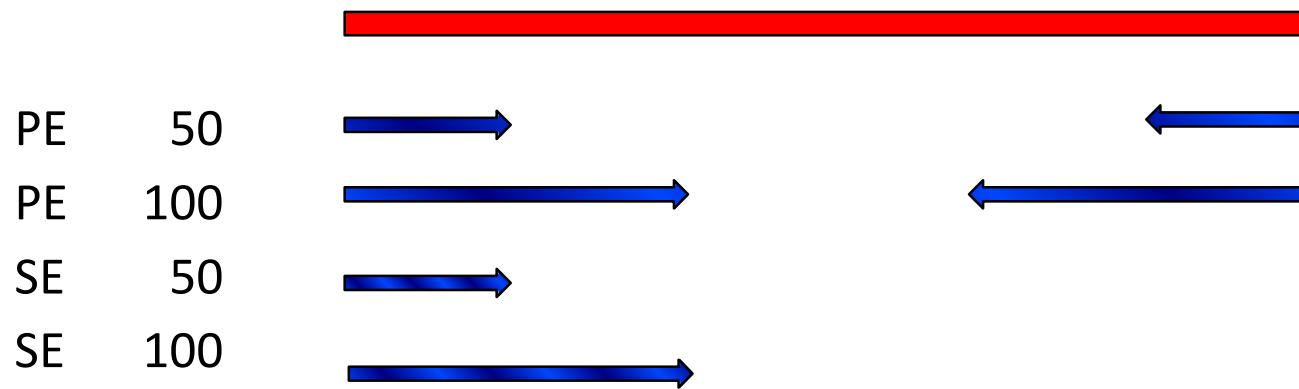


Sequencing Options

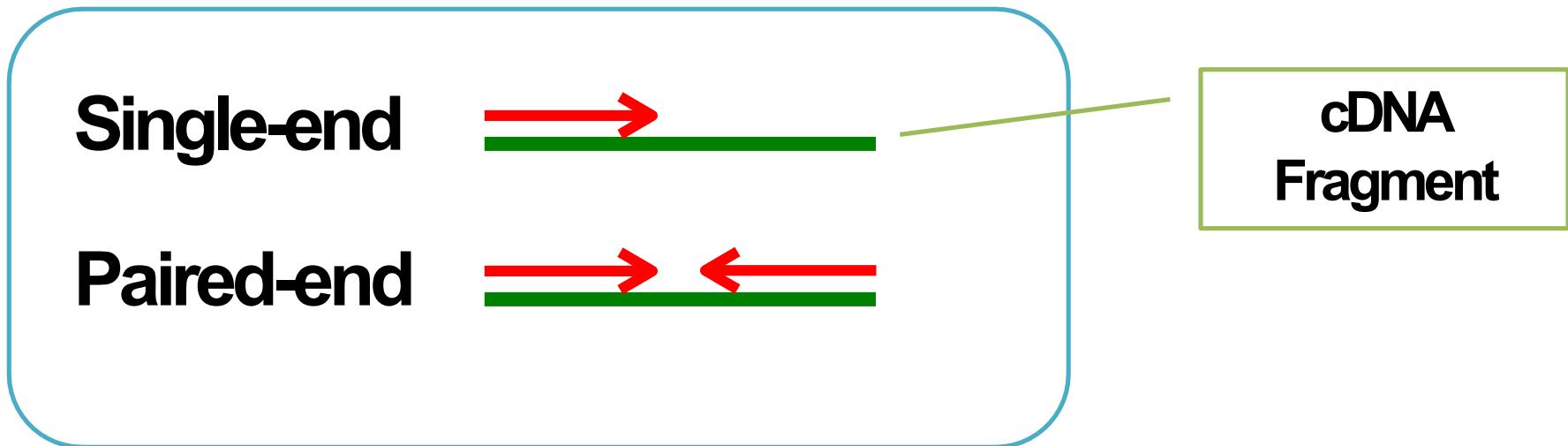
Illumina Sequencing options:

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

DNA
FRAGMENT



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.

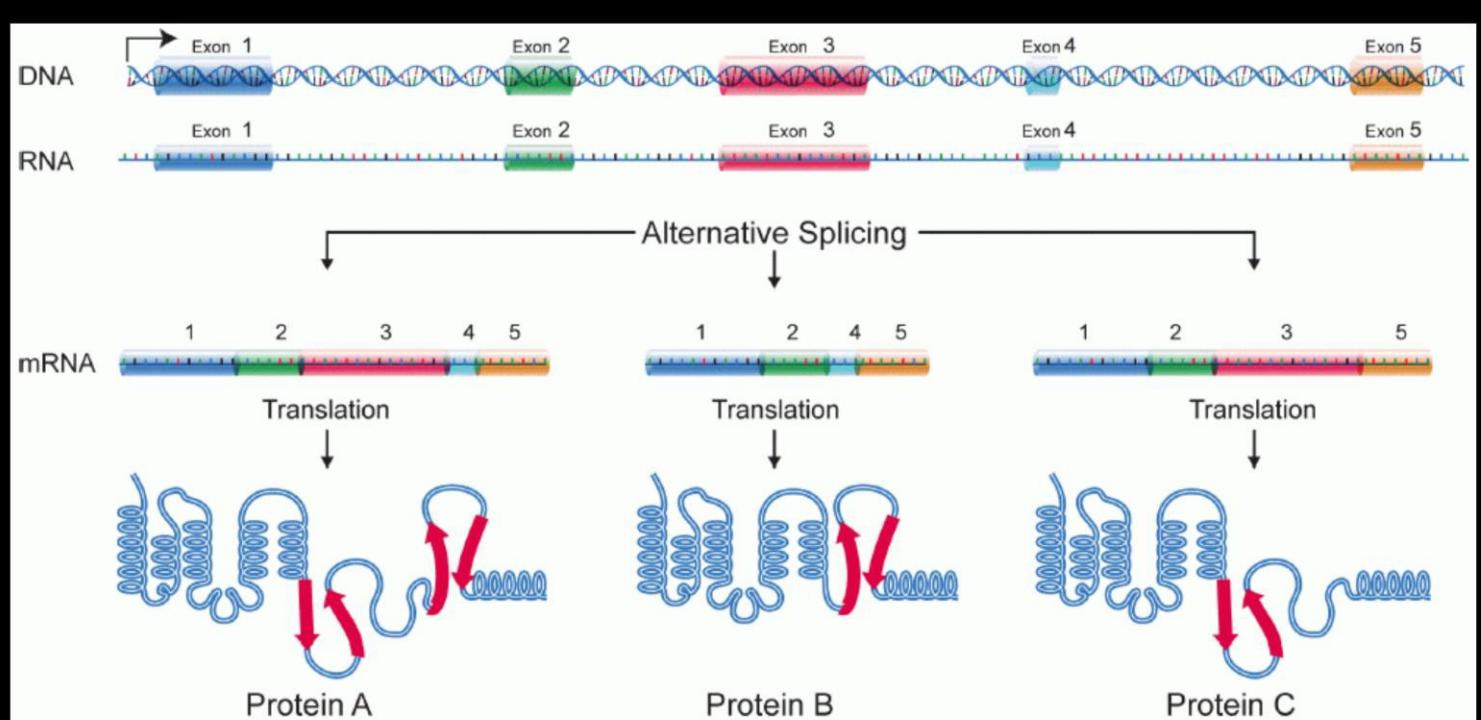
When does SE vs PE matter?

If you're 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. TTAACCTTGGTTTGAACTTGAACACTTAGGGGATTGAAGATTCAACAAACCCCTAAAGCTTGGGTAAAAC

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%

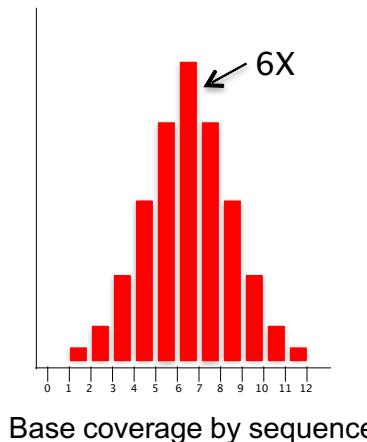
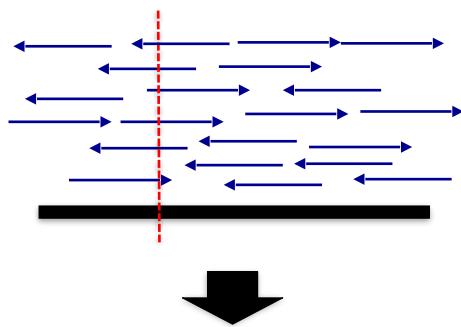
PMID: 26813401

Other NGS terms

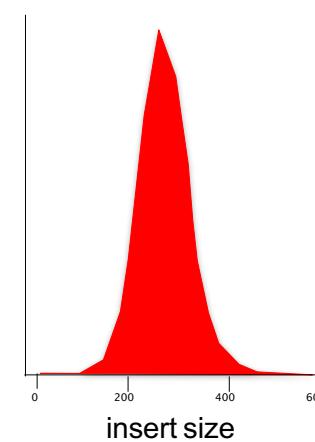
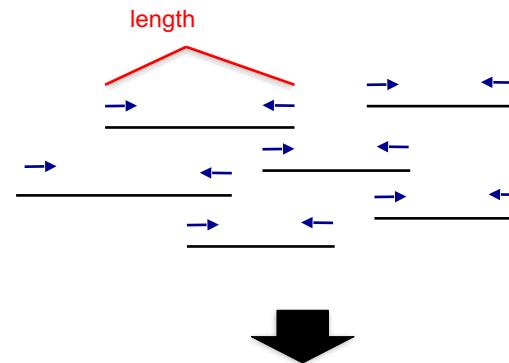
- **Read:** A raw sequence that comes from a sequencing machine.
- **Sequencing depth:** total number of sequences, reads, or bp generated in a single experiment
- **Coverage:** Total number of bases generated per size of the genome sequences
 - What does it mean if someone says they want 15X coverage?

NGS Sequencing Terminology

Sequence Coverage



Insert Size



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!

What you want to do versus what you can afford

Service	Cost
RNA bioanalyzer	~\$60 for 12 samples
Qubit quant	~\$4 per sample
DNA bioanalyzer	~\$82 for 11 samples
Library build (if core does it)	~\$2800 for 12 samples
Library kit (if you do it)	~\$1600 for 12 samples
Sequencing	~\$2000 per lane
Total ~\$4900 OR \$3700	

Time breakdown

Service	Time
RNA isolation	Time to collect samples (days to months) + two days
Library build (if core does it)	Can take up to a month
Library build (if you do it)	One week
Sequencing (HiSeq)	Can take up to a month

Summary RNA-Seq Experiment Planning

