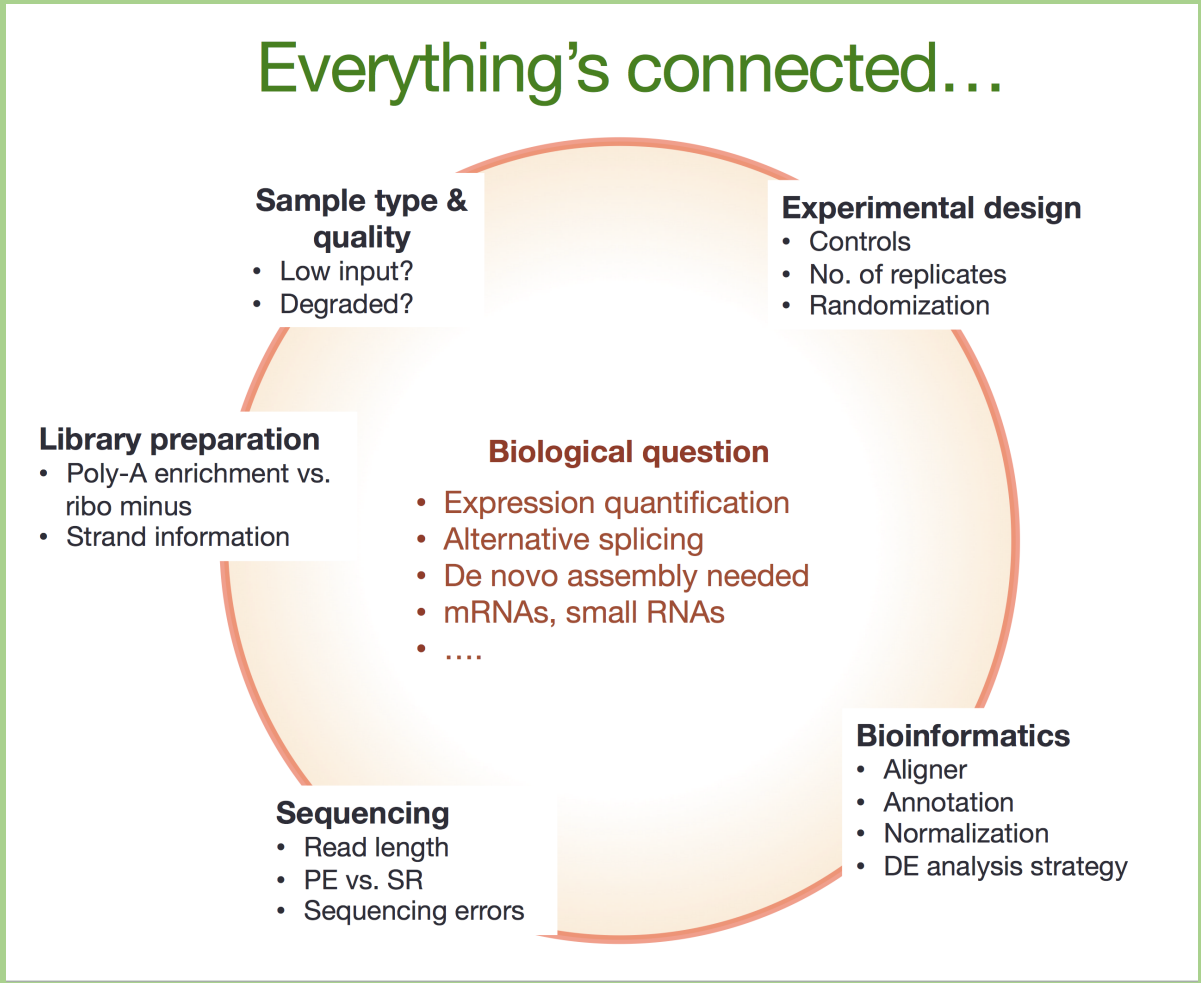


Lecture 1: Setting up an RNA-Seq experiment at UVM

Princess Rodriguez, PhD

MMG 232
Spring 2023

Learning Objectives

- Understand how an RNA-Seq experiment is planned, and considerations taken by the experimenter

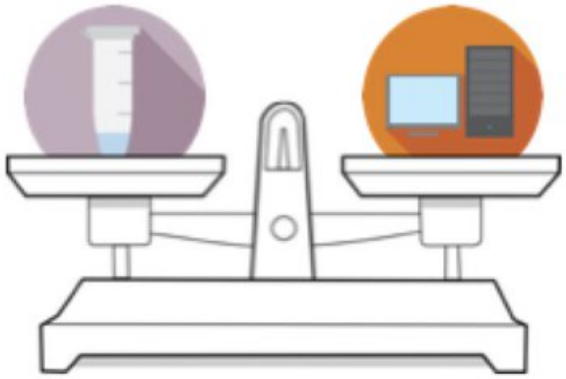
Why?

"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 types of experiments:

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

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
  - The recommendations for this type of experiment would differ slightly
- Meta-transcriptomics

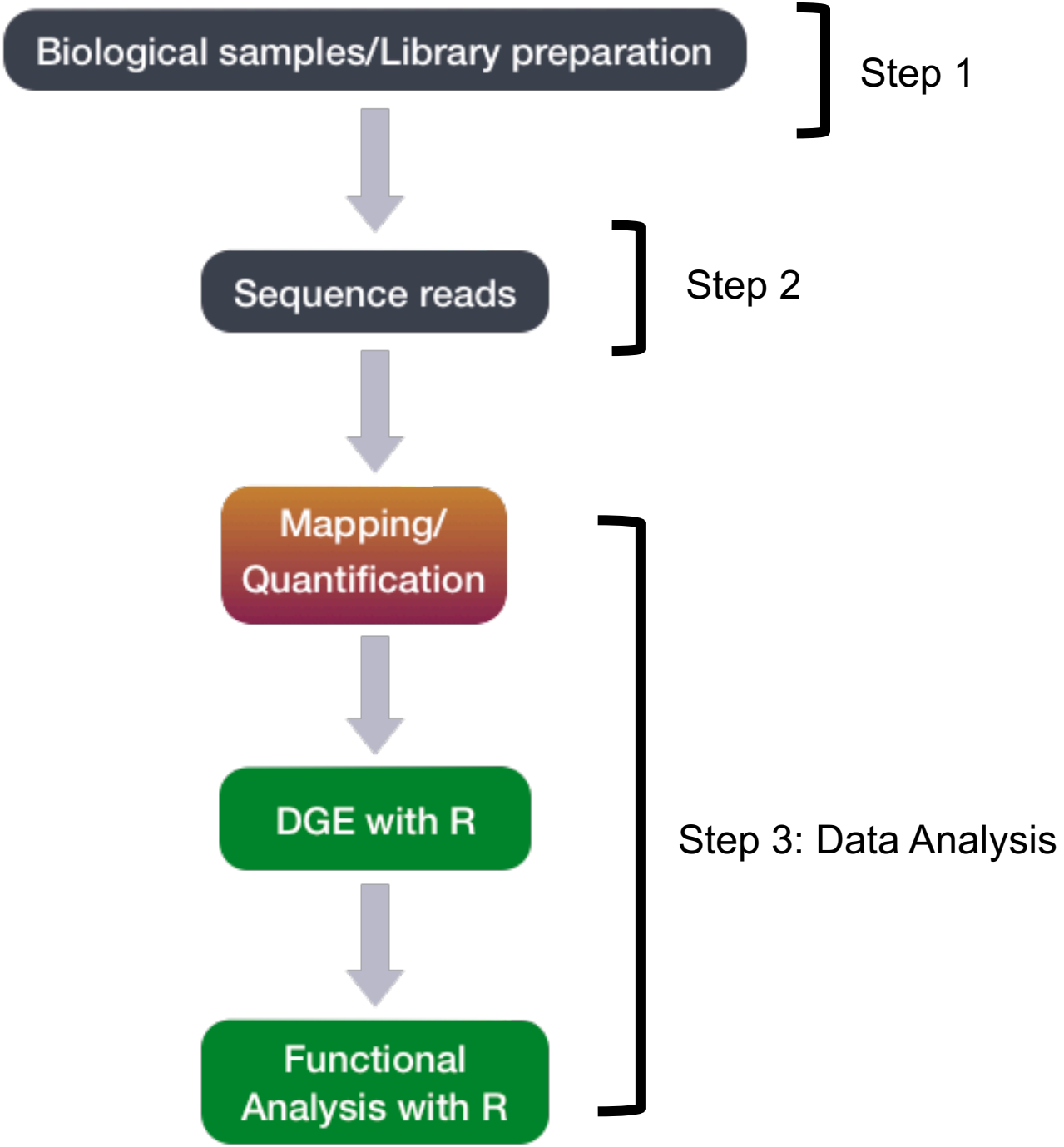
Types of questions answered:

- What genes are differentially expressed between conditions?

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

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

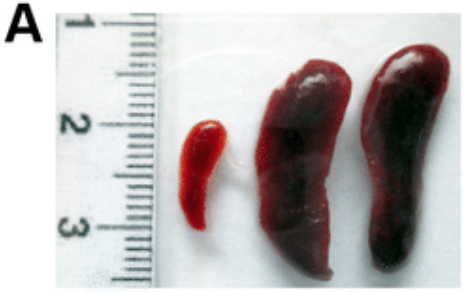
- What processes or pathways are important for my condition of interest?


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

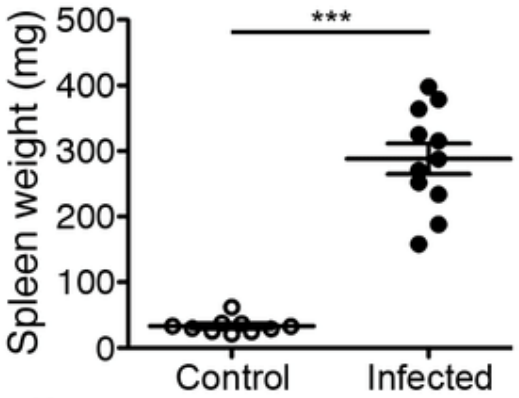
Experimental workflow before it gets sequenced

1

Samples of interest

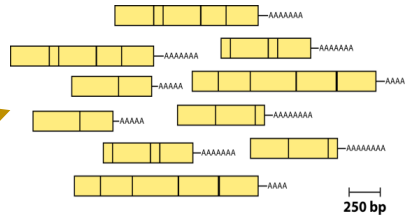


PMID: 27548618



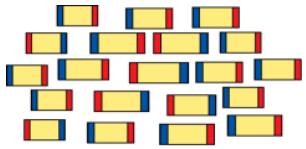
2

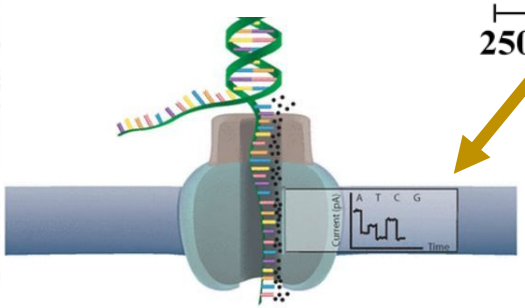
Isolate RNAs

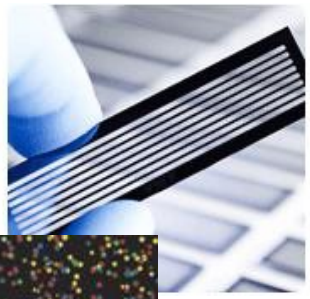


3

Library build



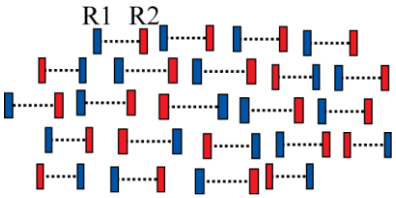




Illumina sequencing versus ONP

4

5

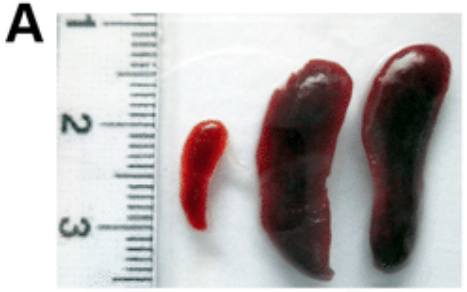


Reads (R1 and R2) generated

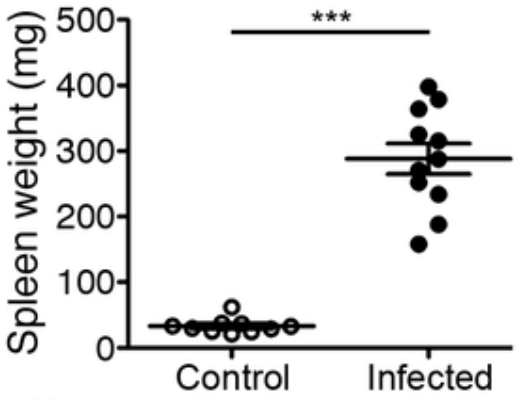
Experimental workflow before it gets sequenced

1

Samples of interest

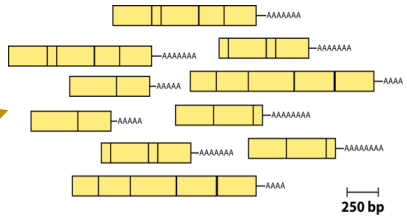


PMID: 27548618



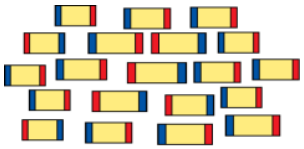
2

Isolate RNAs

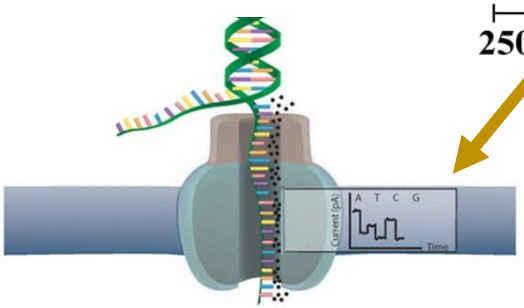


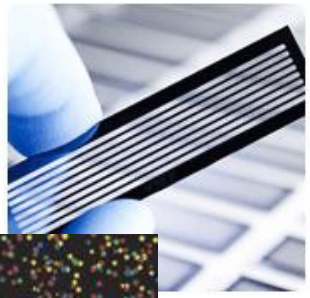
3

Library build



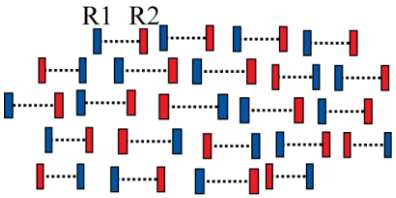
4





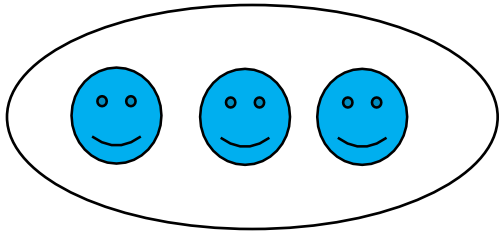
Illumina sequencing versus ONP

5

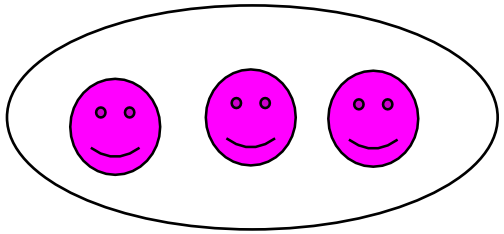


Reads (R1 and R2) generated

Biological Replicates



Condition 1

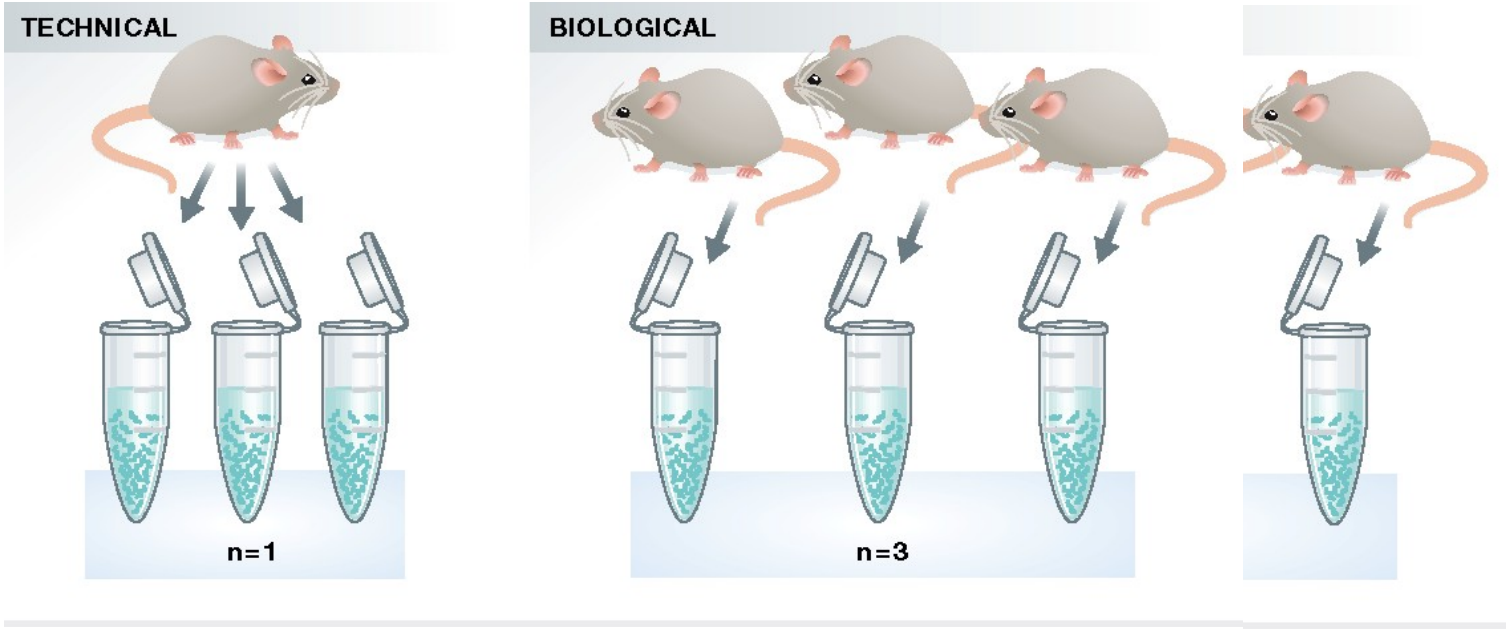


Condition 2

- To detect Differentially Expressed Genes (DEGs) between groups we should have several samples, which are also known as 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** (More is always better!)

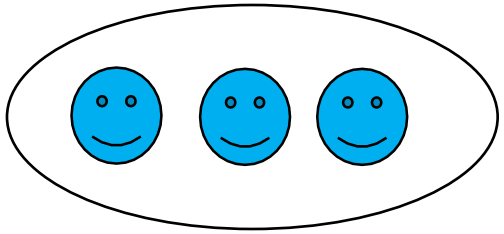


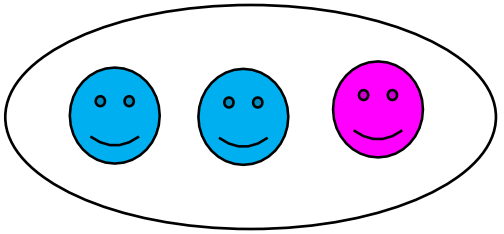
Probability of detecting DEGs

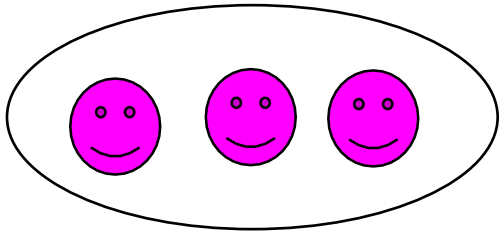
	Replicates per group		
	3	5	10
Fold change			
2	87%	98%	100%

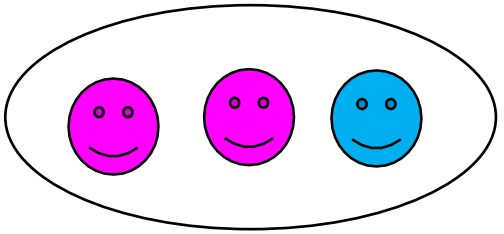
PMID: 26813401

Grouping of Replicates





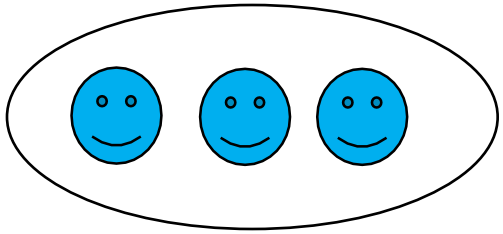


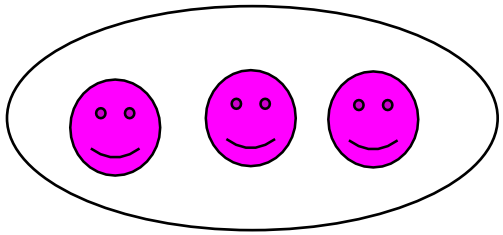


What you want

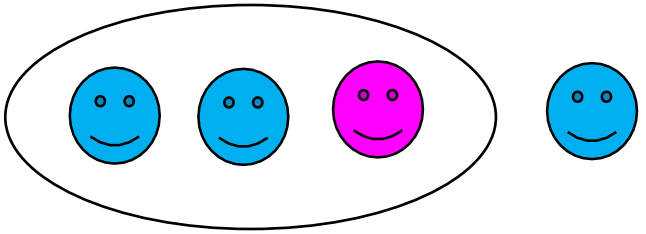
What you get

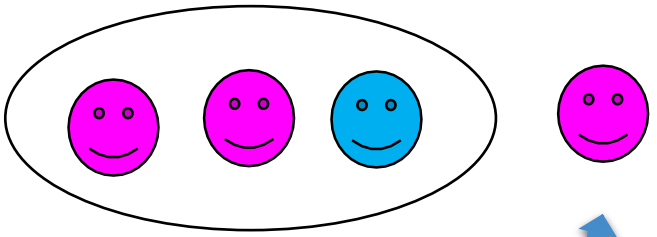
Grouping of Replicates





What you want



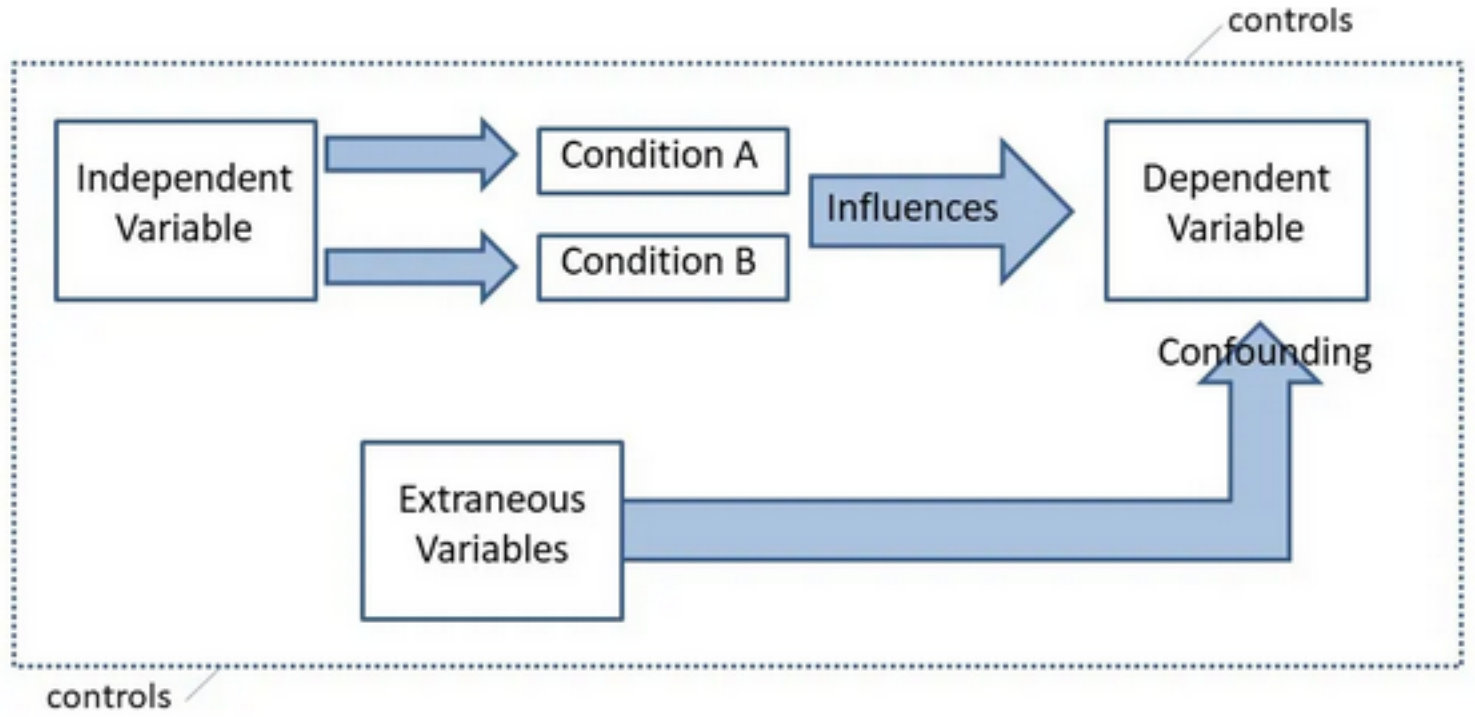


What you get

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



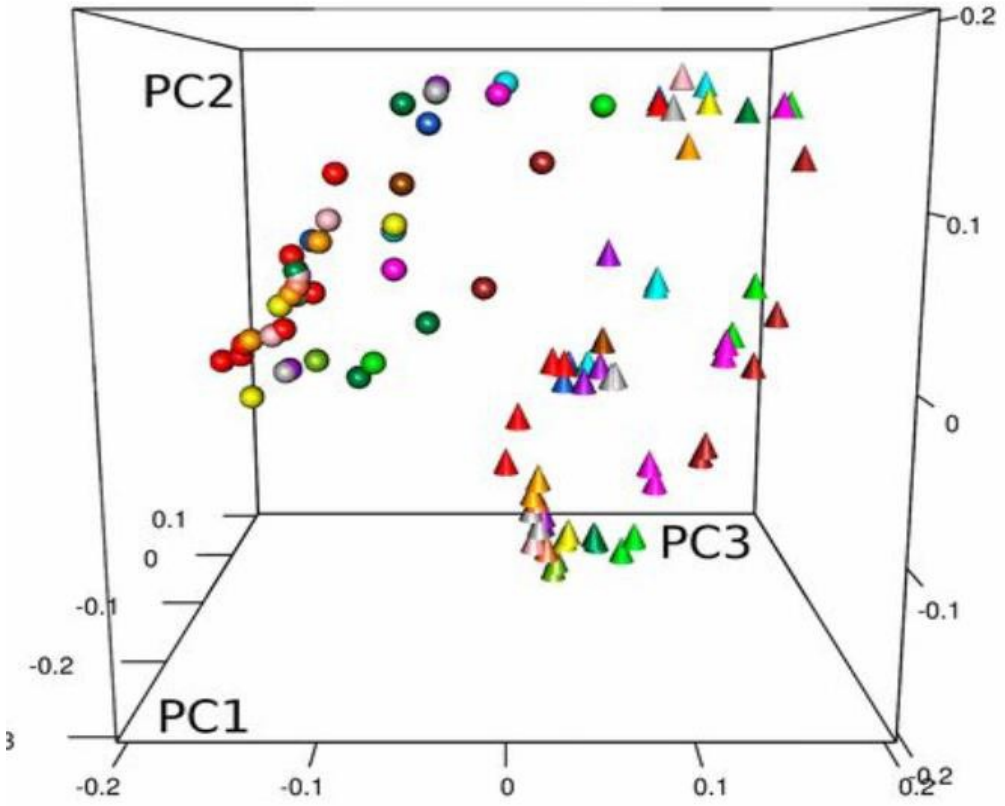
What causes this? Confounding variables

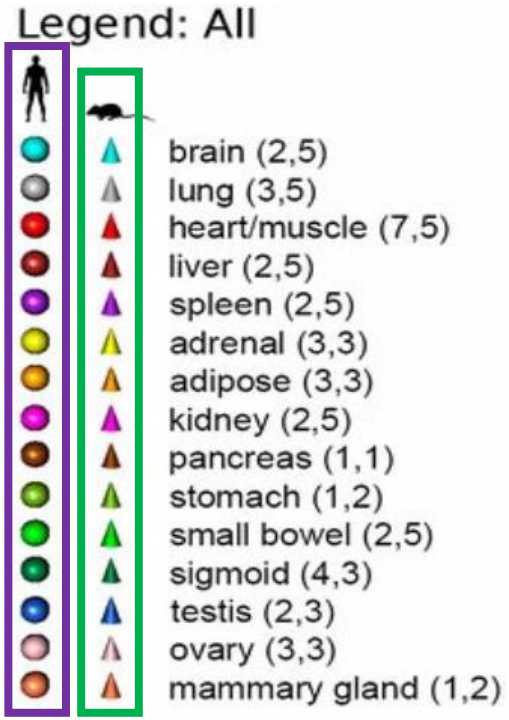
- Sometimes, its impossible for bioinformaticians to partition biological variation from technical variation, when these two sources of variation are confounded.
- No amount of statistical sophistication can separate confounded factors after data have been collected.
- *…sometimes, these confounding variables are not in your control!*
- A well-planned experiment with an additional sample, does end up saving you time and money down the road

Examples common in RNA-Seq

- New technician is processing your samples and/or running the sequencer
- Extracting RNA with different kits
- Inappropriate multiplexing strategy
- Sequencing on multiple different types of platforms

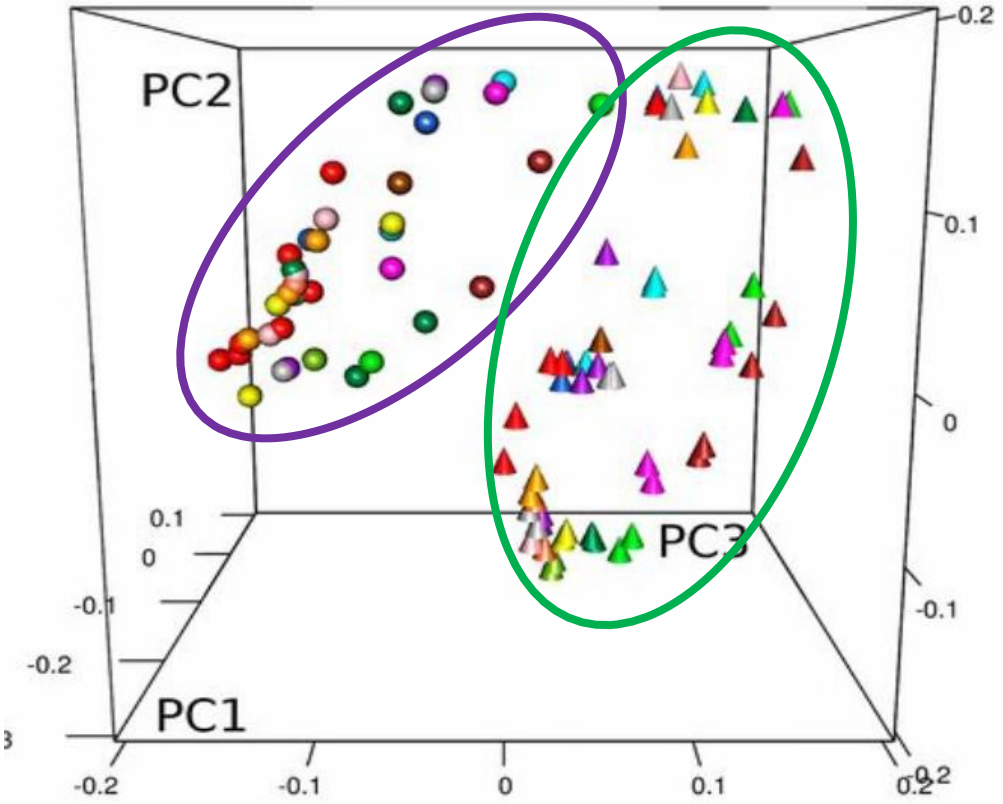
ENCODE reported that gene expression was likely to follow a species-specific rather tissue-specific pattern

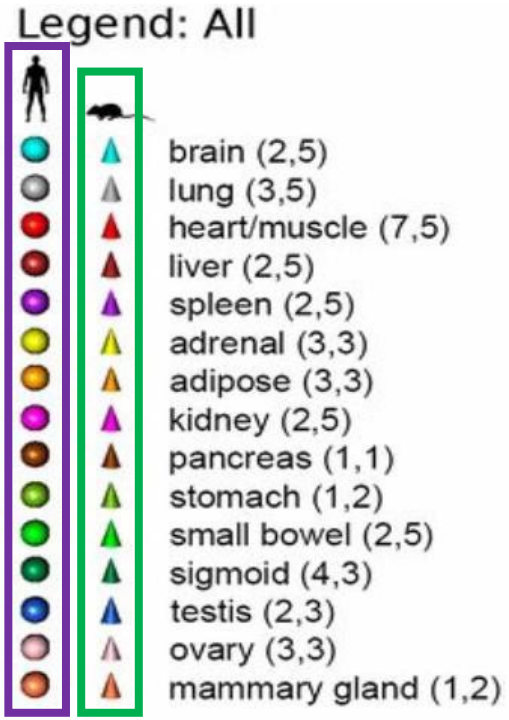




https://doi.org/10.1073/pnas.1413624111

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



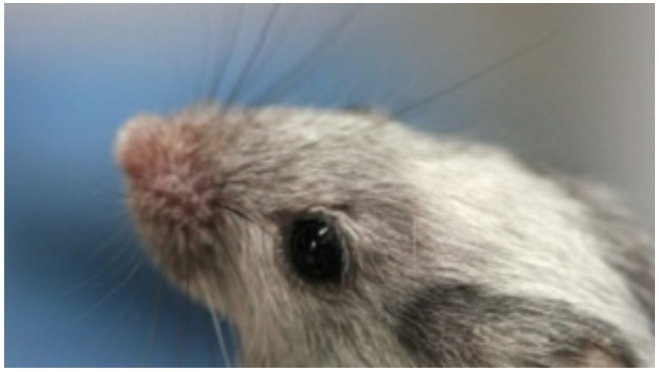


https://doi.org/10.1073/pnas.1413624111

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.

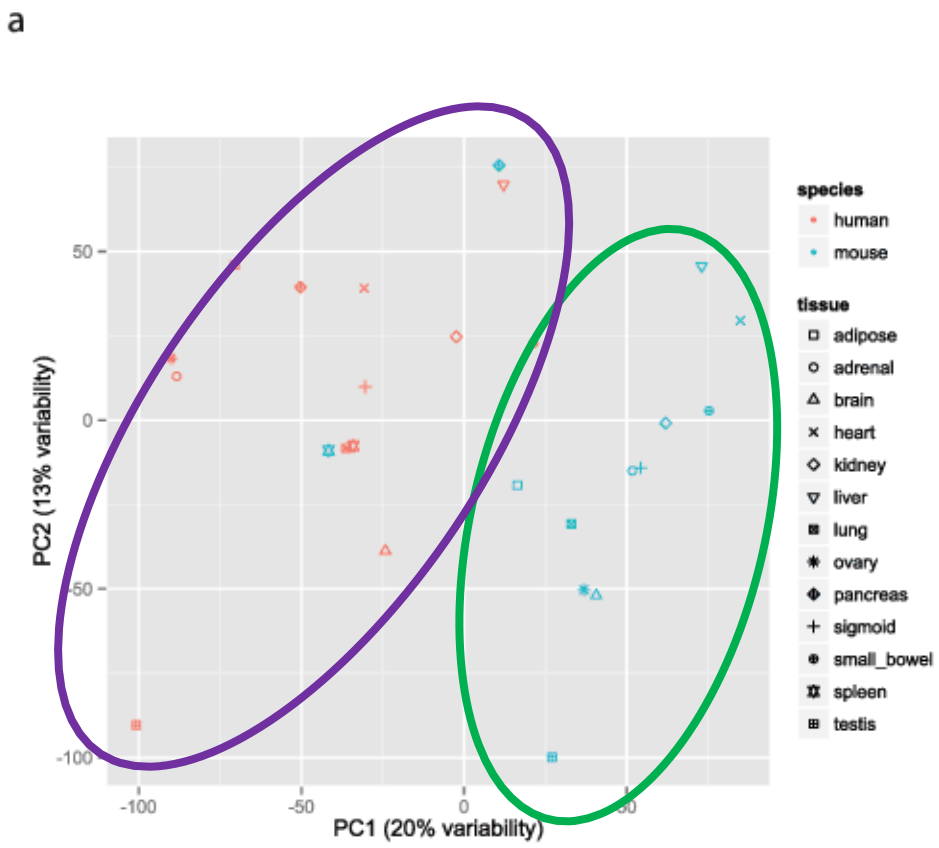
But earlier this month, Yoav Gilad of the University of Chicago called these results into question on Twitter. With a dozen or so 140-character dispatches (including three heat maps), Gilad suggested the results published in *PNAS* were an anomaly—a result of how the tissue samples were sequenced in 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.

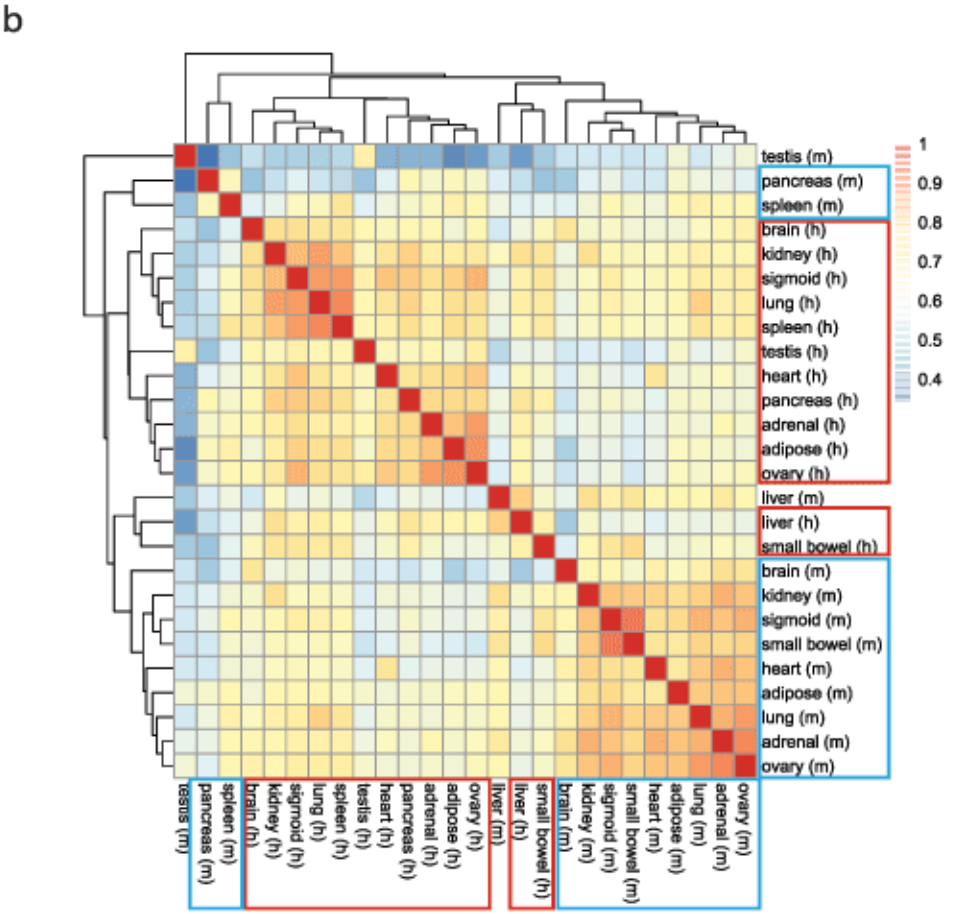
Sequence study design (sequencer ID, run ID, lane number):

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

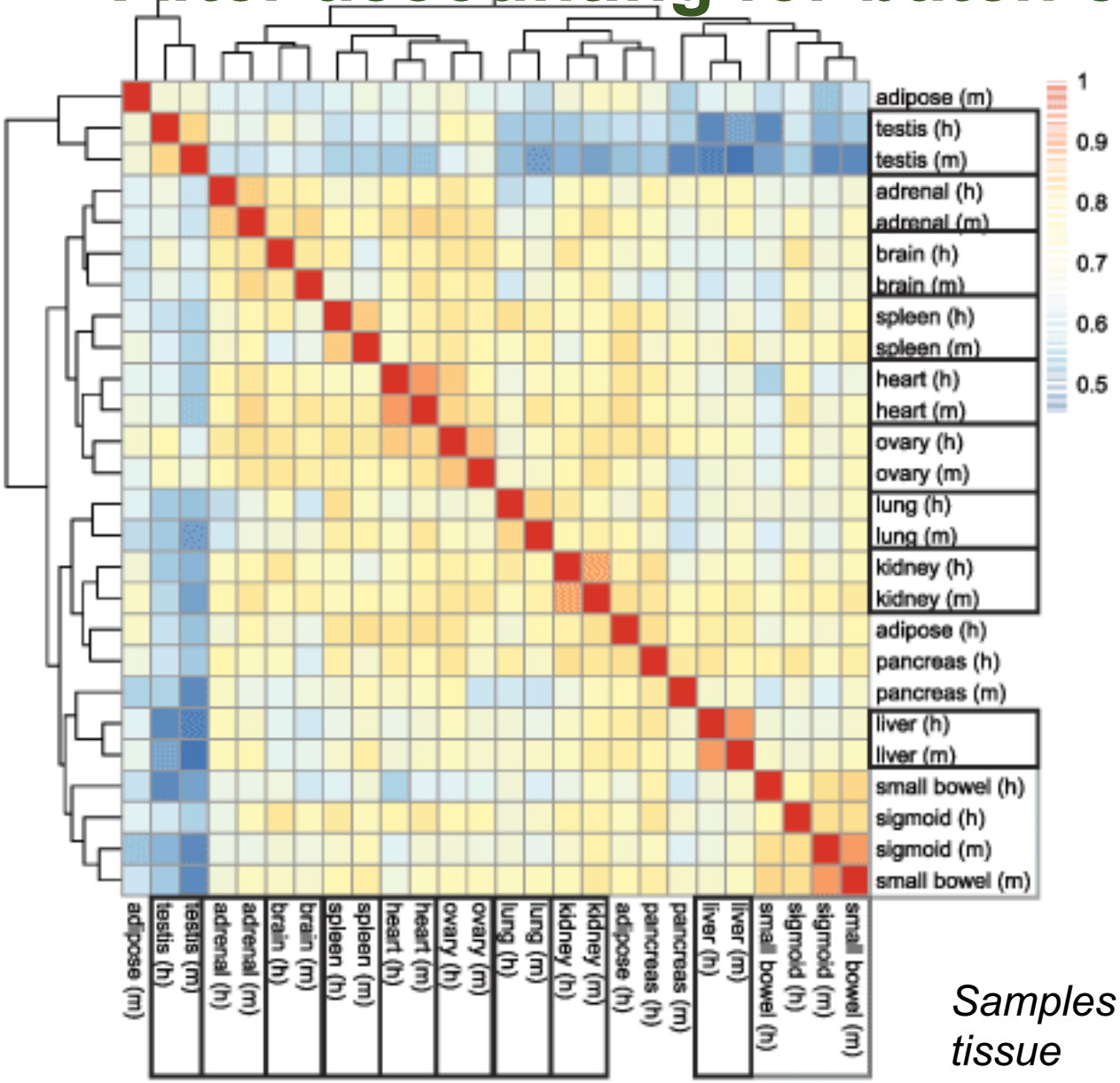




Samples grouped by animal

b

After accounting for batch effect

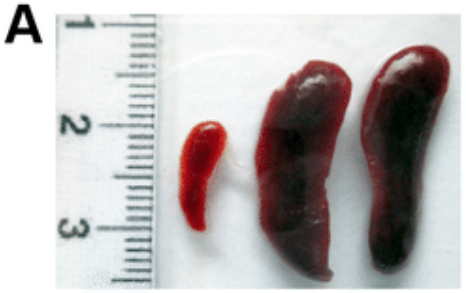


Samples now grouped by tissue

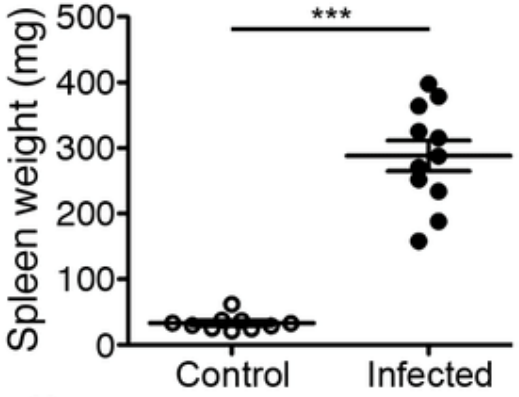
Experimental workflow before it gets sequenced

1

Samples of interest

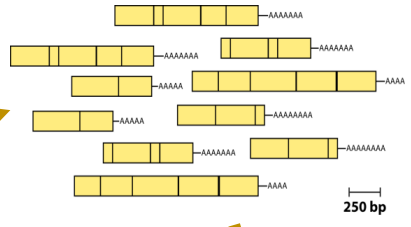


PMID: 27548618



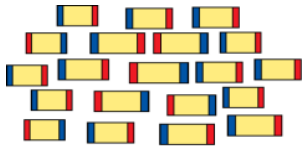
2

Isolate RNAs

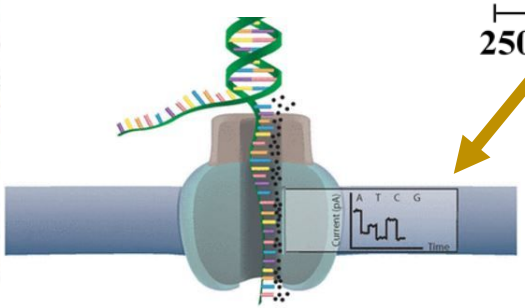


3

Library build

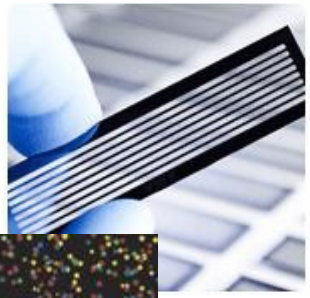


250 bp

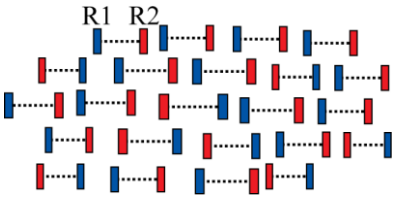


Illumina sequencing versus ONP

4



5



Reads (R1 and R2) generated

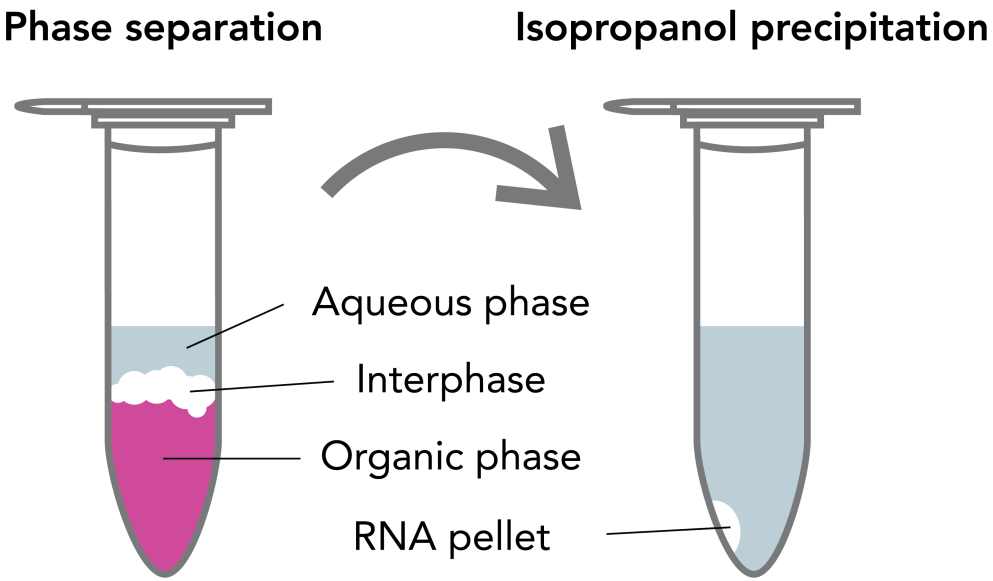
Working with RNA

- 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

The cheaper way to isolate RNA

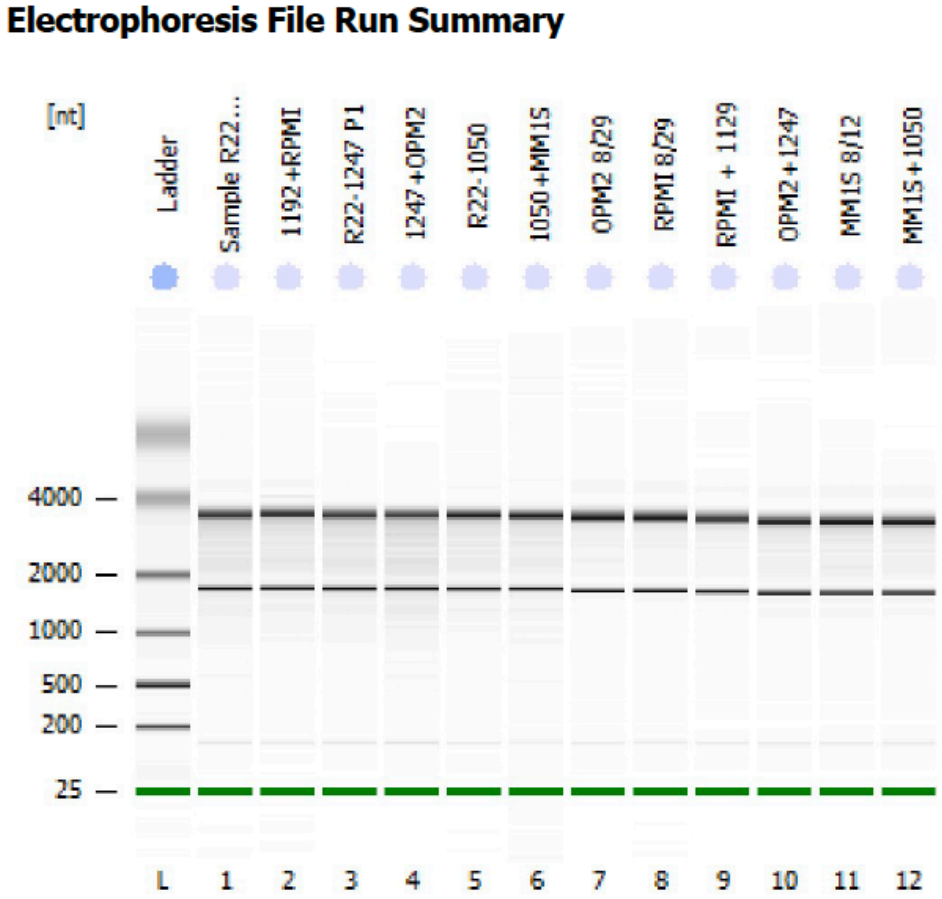


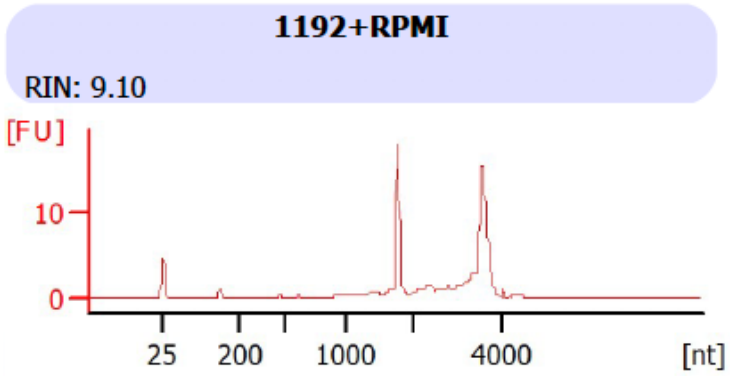
- The UVM VIGR core ***does not*** recommend as residual phenol will inhibit the library build and you cannot get rid of it. They will not sequence your samples if you use this mode of isolation!

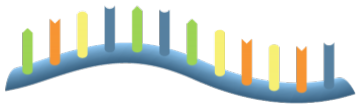
Other recommendations

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

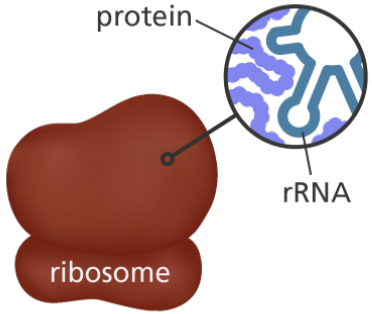
Check purity with RNA bioanalyzer



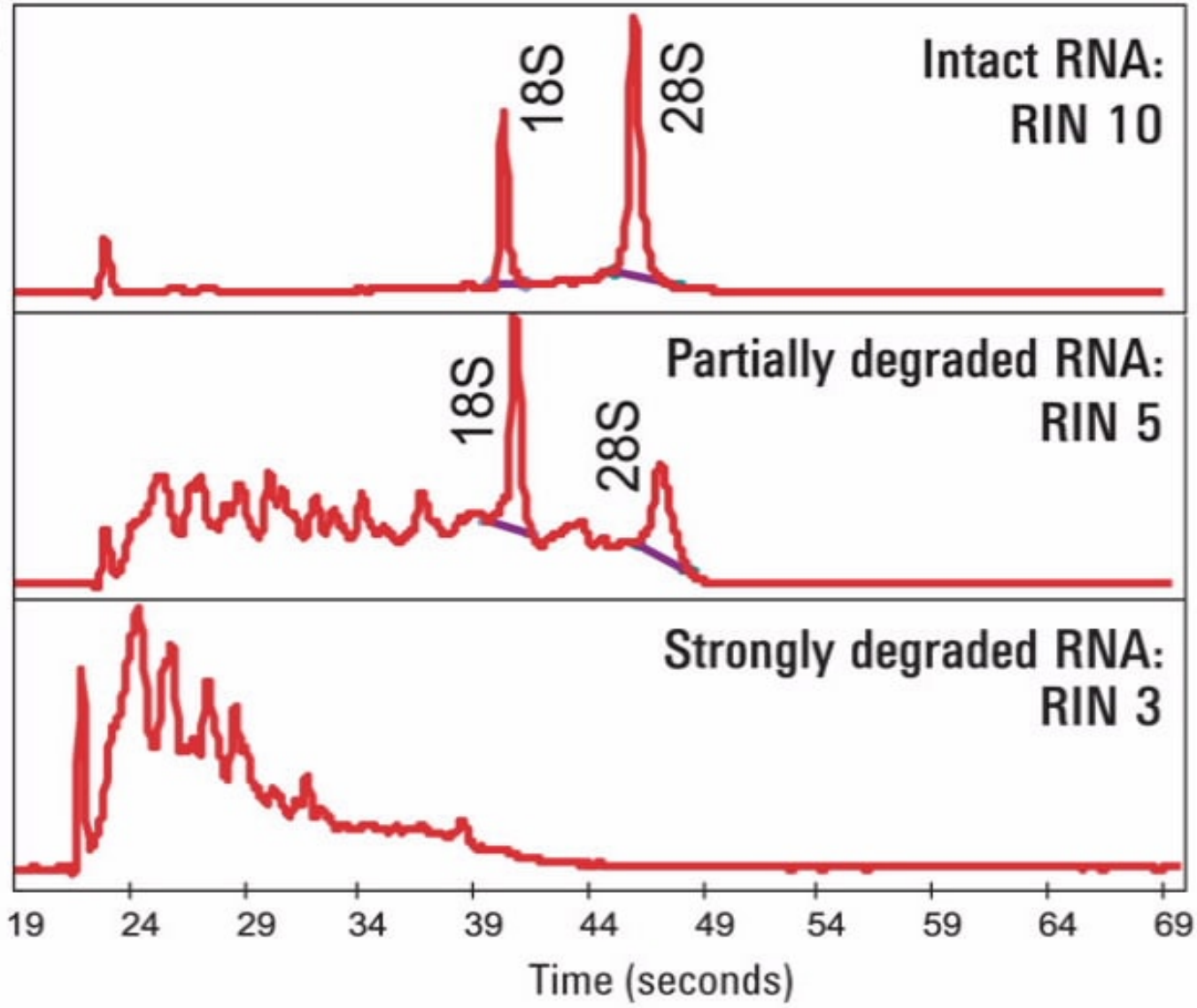




messenger RNA
(mRNA)



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)

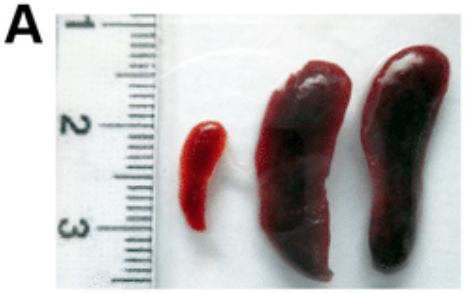
RNA Bioanalyzer requires Qubit

- Need to take a Qubit reading not NanoDrop

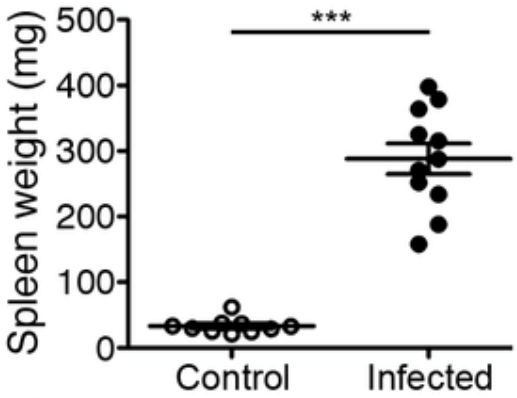
Experimental workflow before it gets sequenced

1

Samples of interest

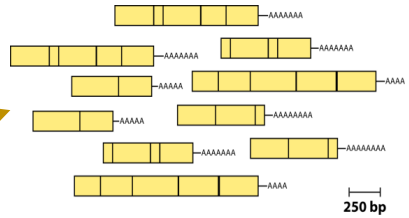


PMID: 27548618



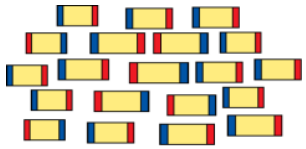
2

Isolate RNAs



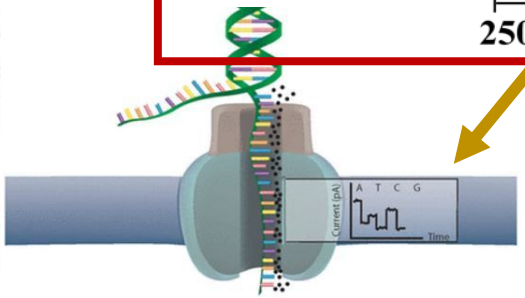
3

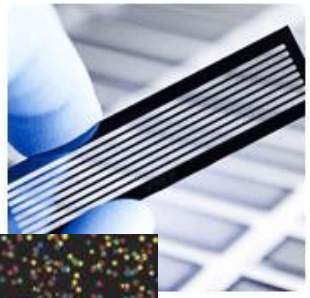
Library build



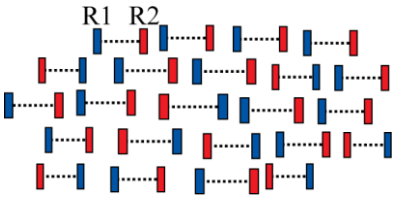
4

Illumina sequencing versus ONP





5

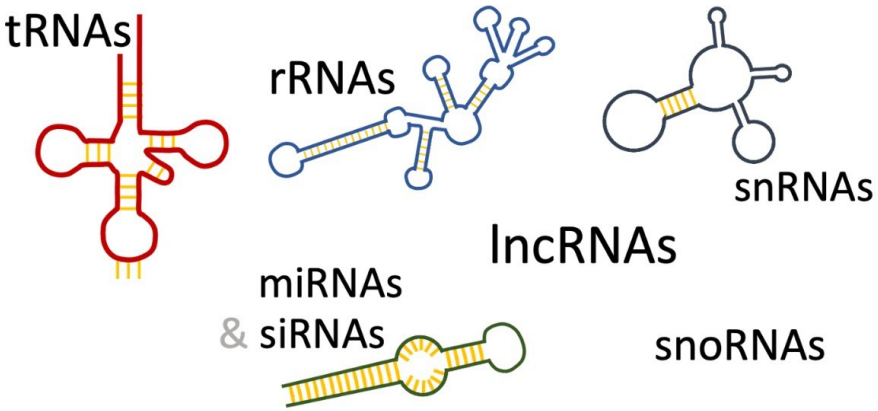


Reads (R1 and R2) generated

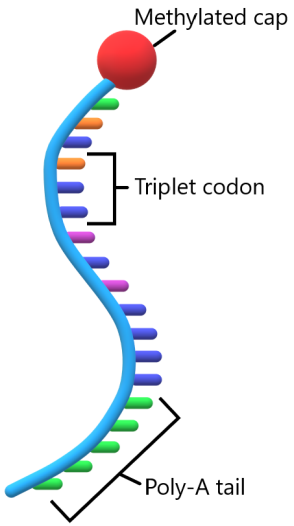
RNA composition

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

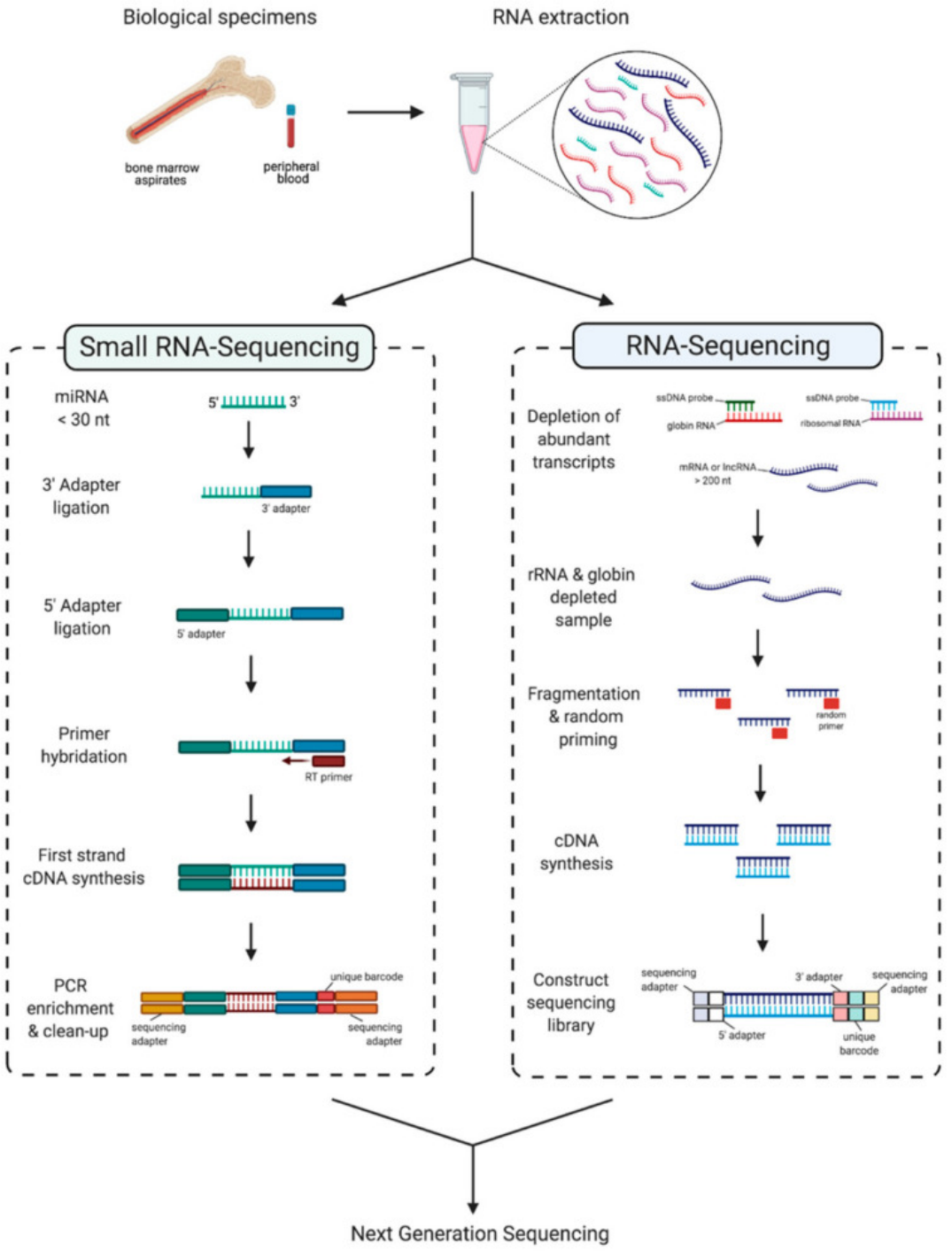


messenger RNA



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

Comprehensive transcriptome analysis

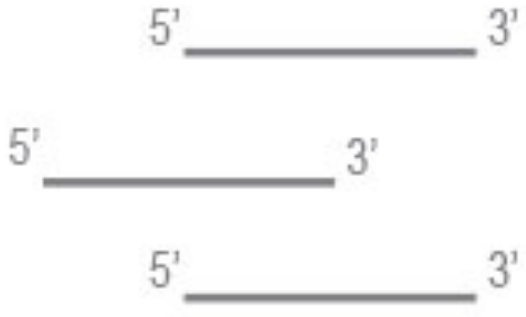


Before building the libraries need to perform target enrichment

- 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

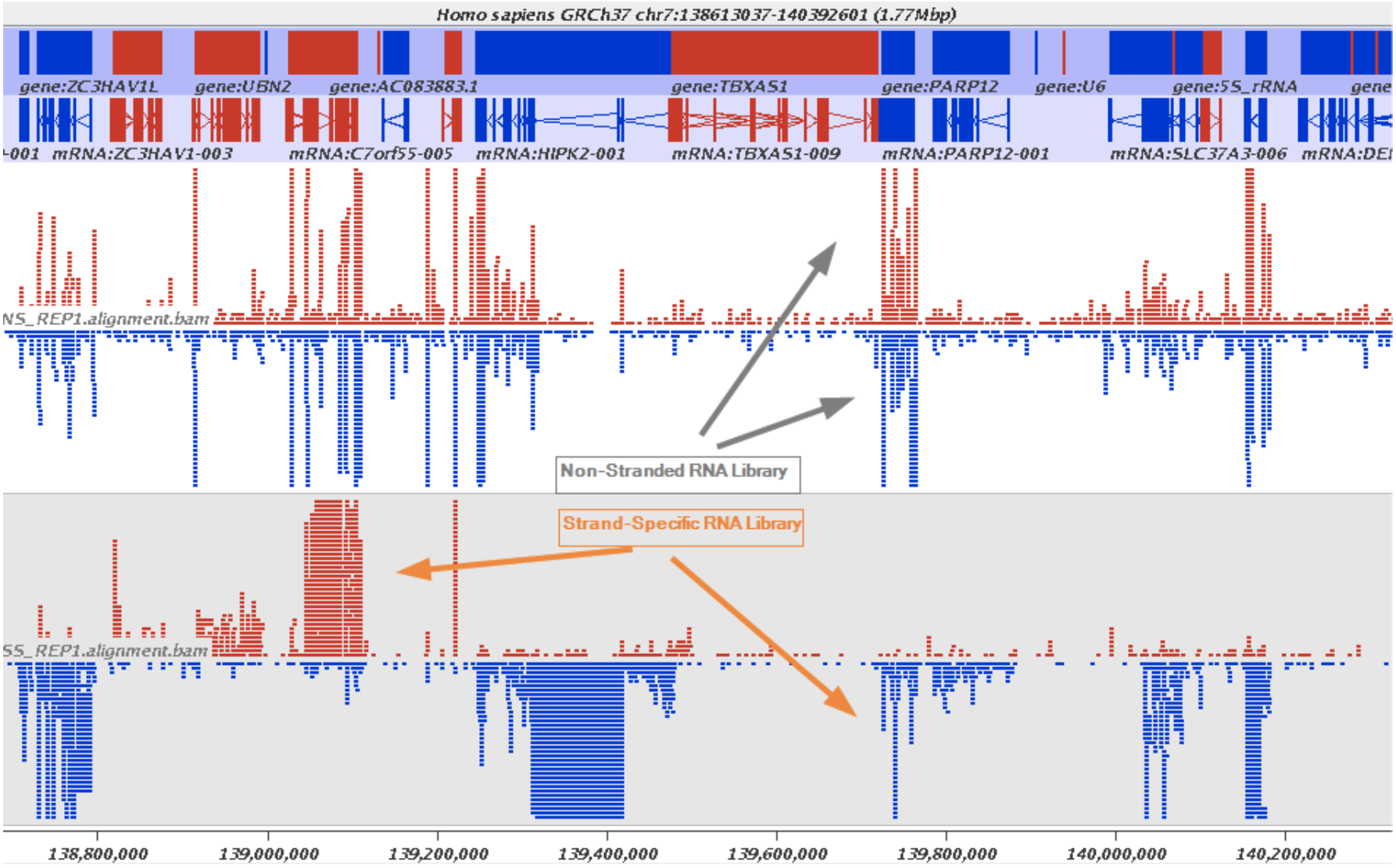


Strandedness

- 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

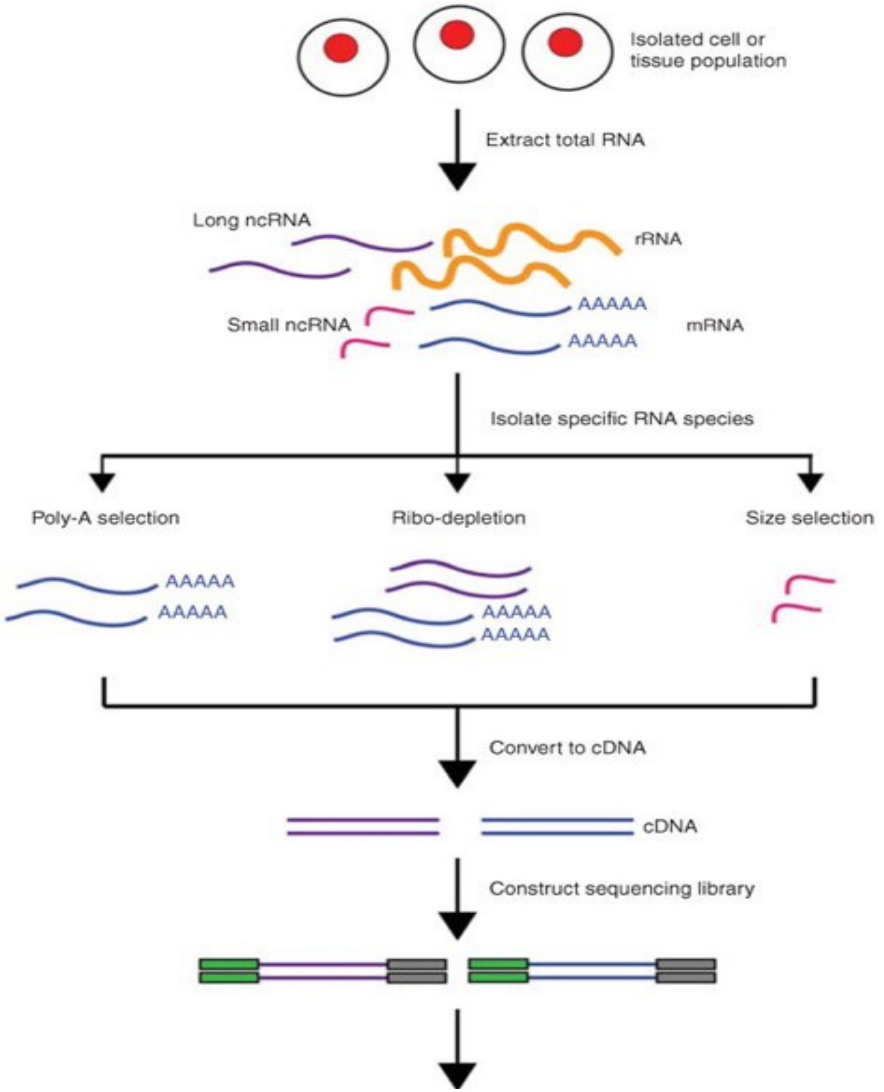
Strandedness



Red = + strand

Blue = - strand

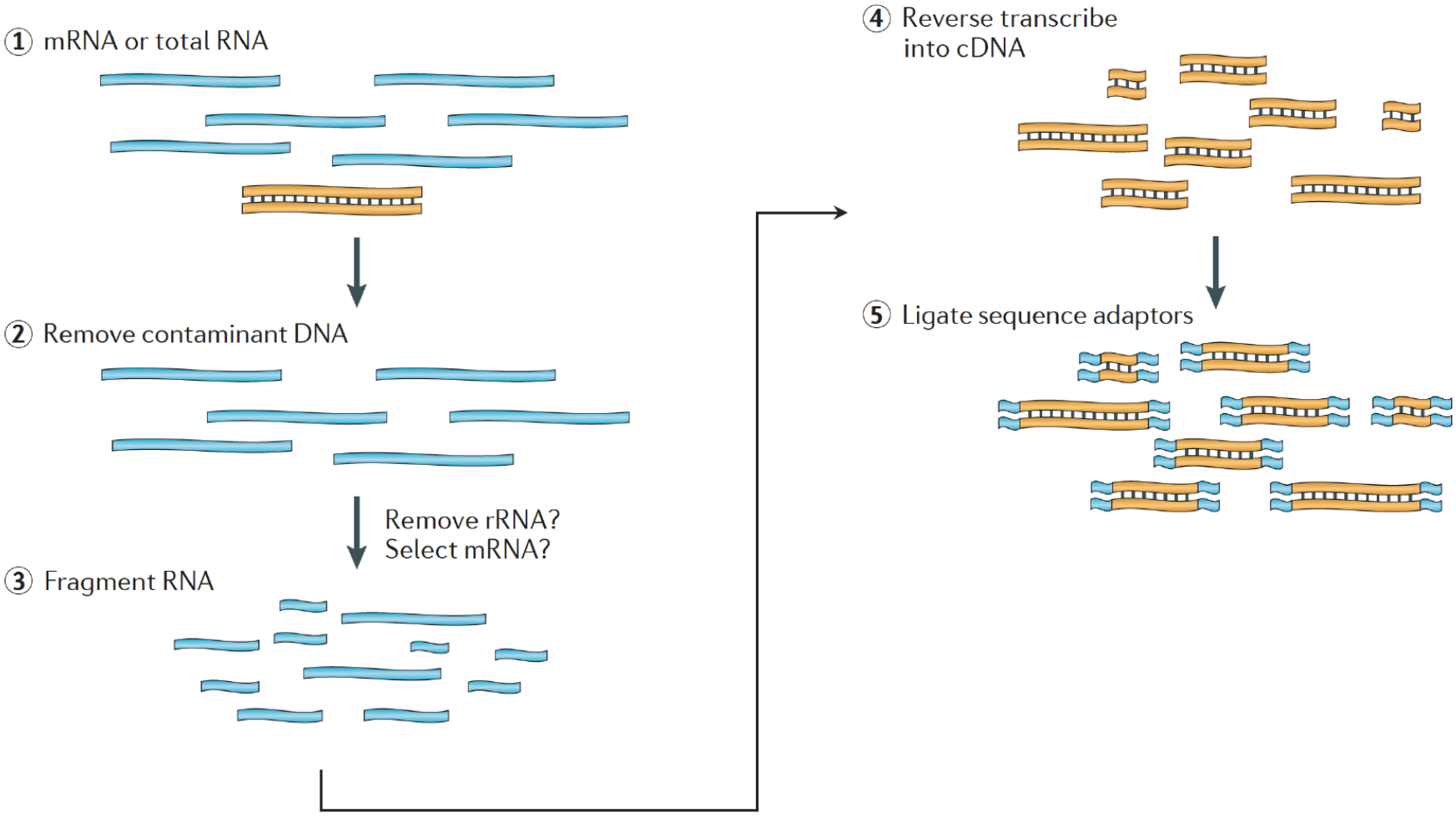
RNA-seq Library build steps



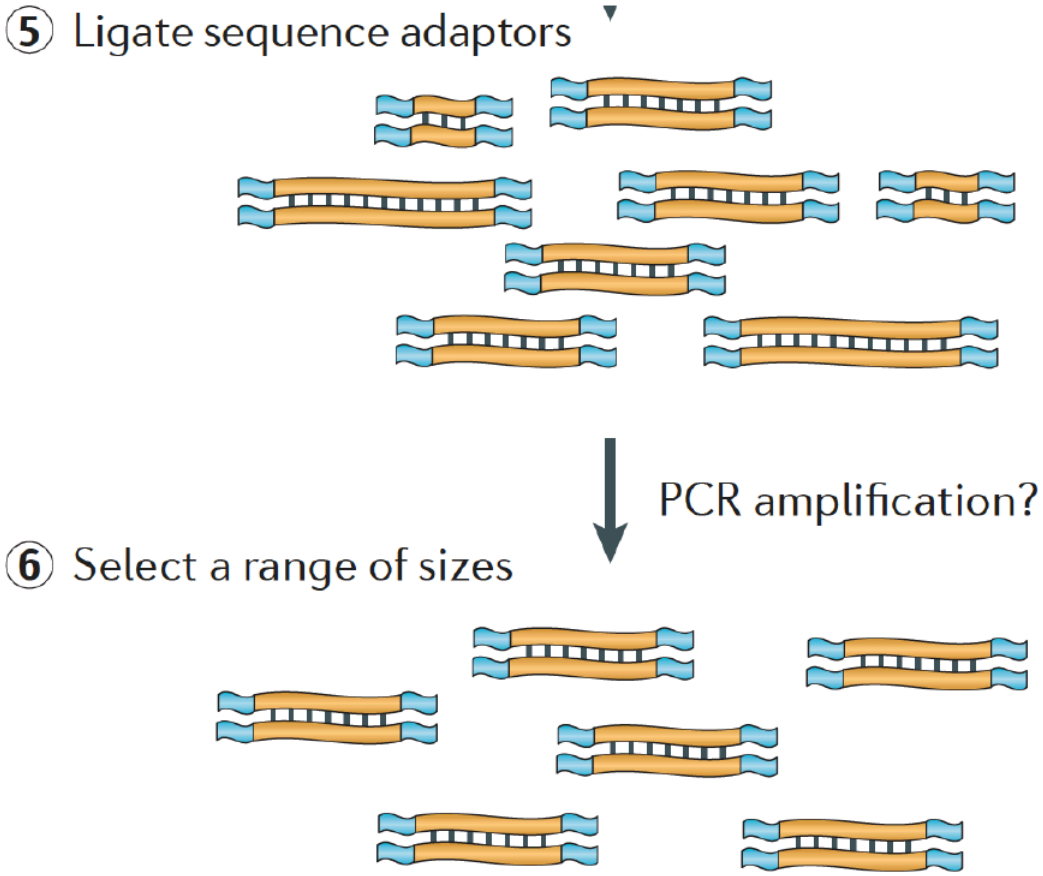
- ☑ Isolate RNA
- ☑ RNA species diversity
- ☑ target enrichment
- ☐ cDNA conversion / PCR amplification

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

Library Prep steps

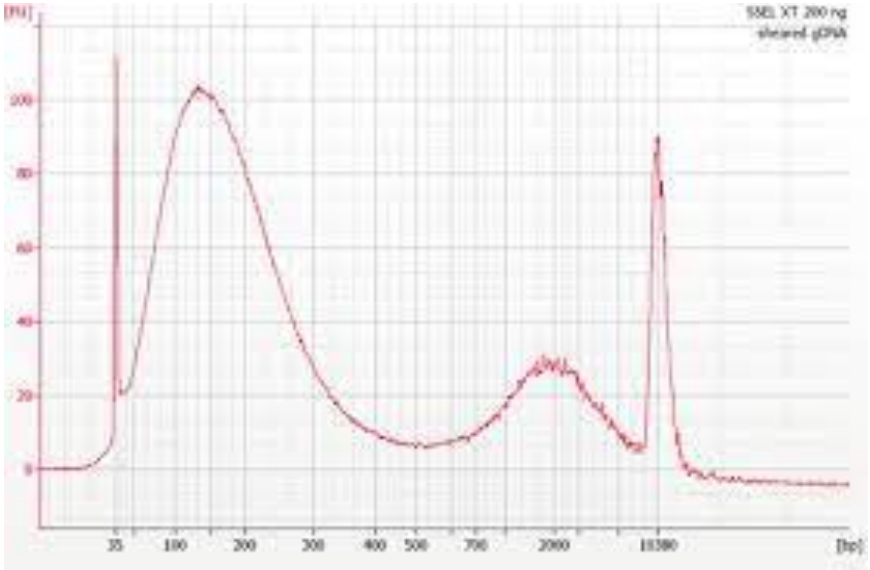


Library Prep steps continued



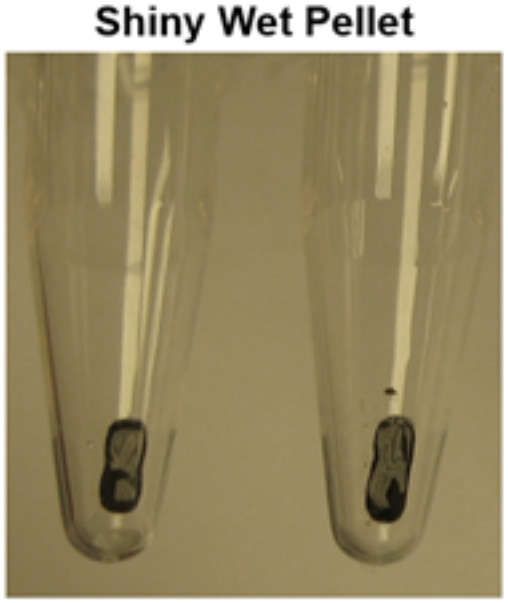
Common mistakes during library build

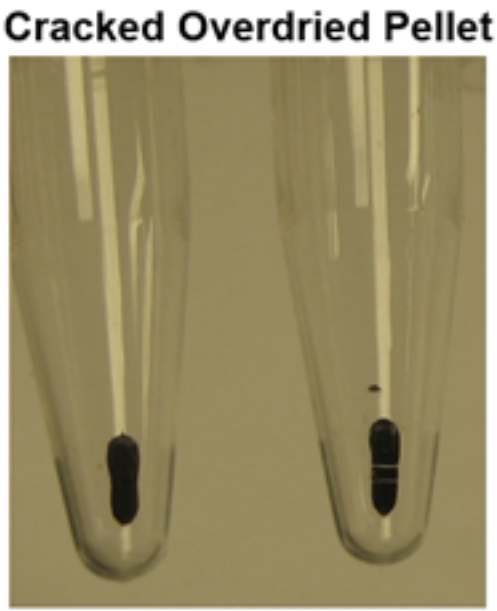
- Addition of adapters and unique barcodes
- Number of PCR cycles



Common mistakes during library build

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





Common mistakes during library build

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

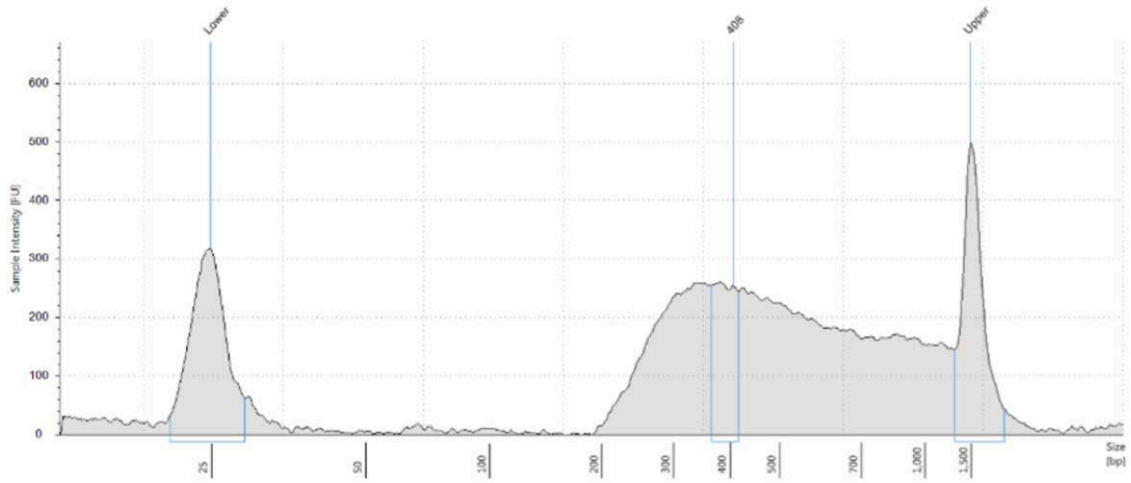
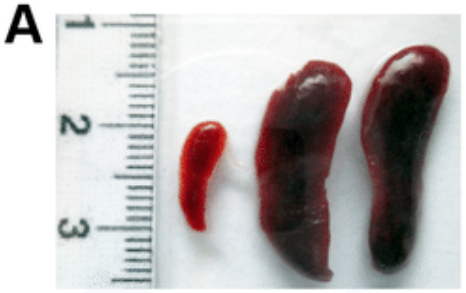


Figure 4. An electropherogram showing a sample with incomplete fragmentation due to suboptimal mixing. There are large fragments present which will be unable to undergo bridge amplification during cluster generation on the flow cell.

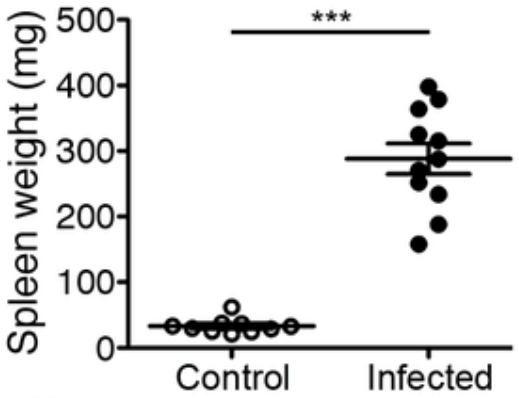
Experimental workflow before it gets sequenced

1

Samples of interest

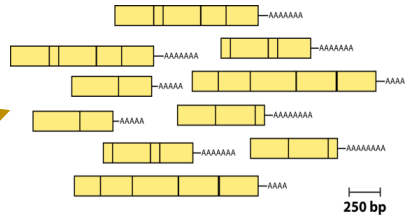


PMID: 27548618



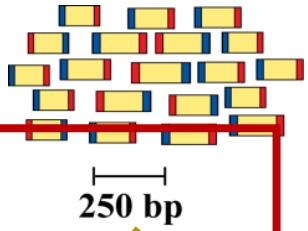
2

Isolate RNAs

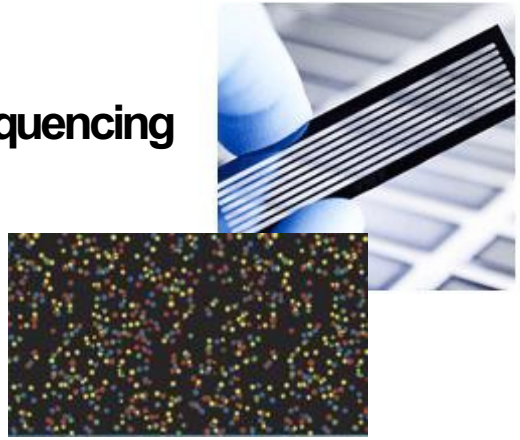


3

Library build

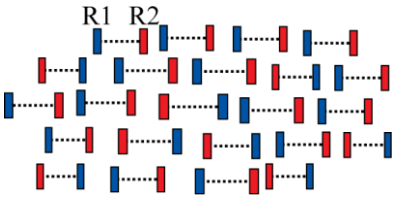


Illumina sequencing



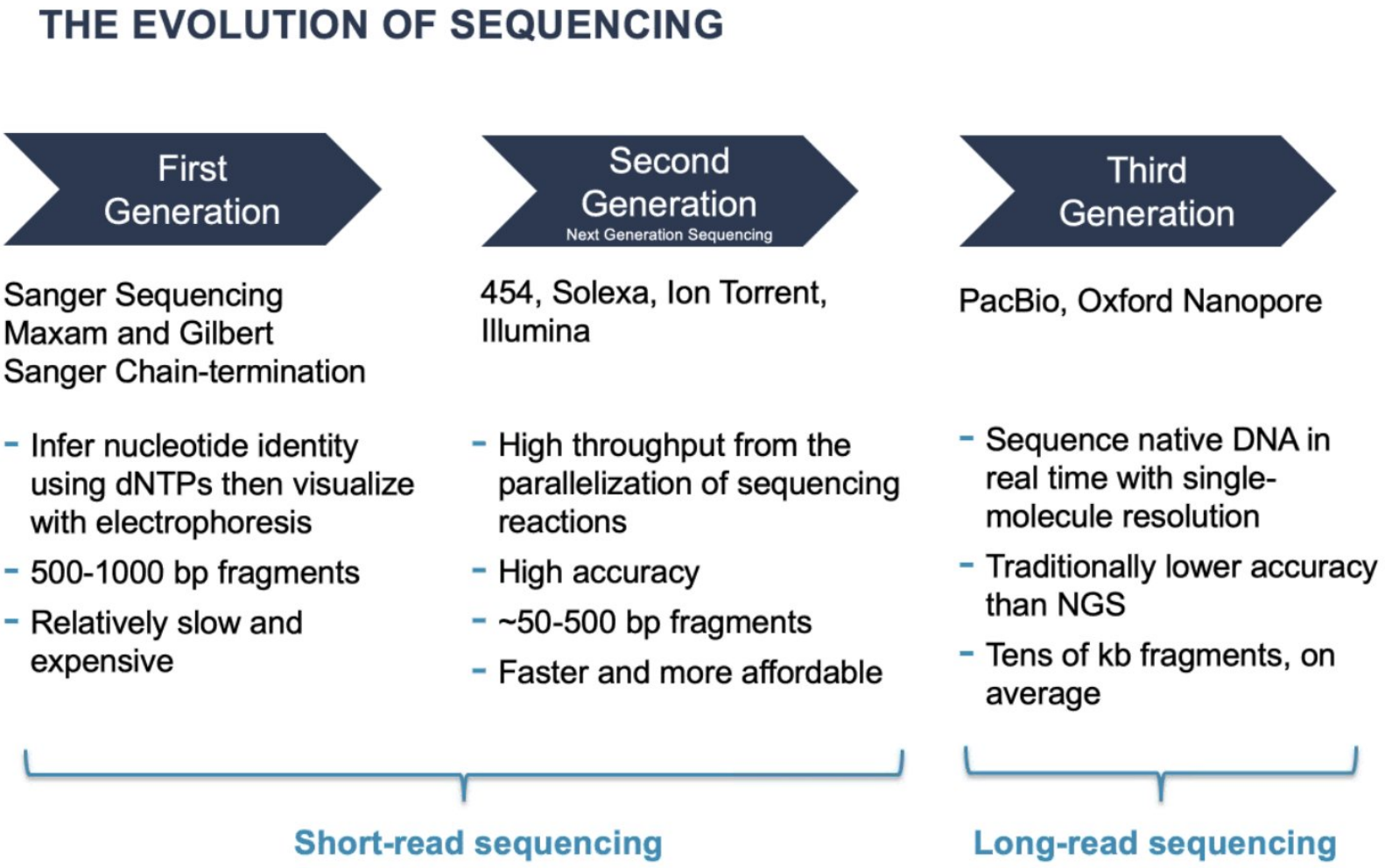
4

5



Reads (R1 and R2) generated

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



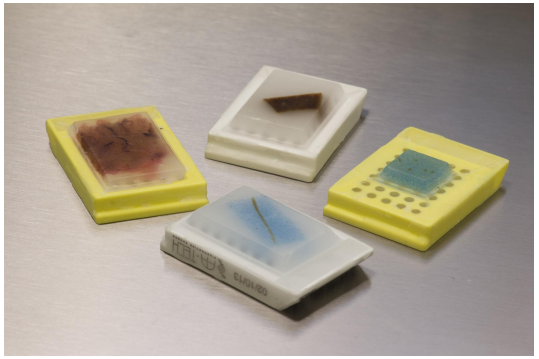
Second Generation Sequencing

- 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

Illumina/Solexa pros/cons

Pros

- Most library protocols are compatible with the Illumina system
- Highest-throughput of all platforms
- Can sequence fragmented DNA
  - FFPE samples



Cons

- Clustering across flow cell must be tightly controlled
- Overloading can result in poor sequence quality
- Low complexity samples must be mixed with PhiX to generate library diversity

Third Generation Sequencing

- Typical characteristics:
  - No amplification step – single molecule
  - Long (1,000bp+) sequence reads
  - Real-time monitoring of nucleotide incorporation
  - Reads the nucleotide sequence at the **single molecule** level
  - Often associated with nanopore technology

Oxford Nanopore Technology (ONT) sequencing

- Sequencers can be pocket-sized, USB connectable, cost $150 to $1,000
- Generates incredibly long reads: read length are typically between 10,000bp & as long as 1,000,000 bp!

Modifications to RNA

- …are lost during the reverse transcription step whereby RNA is converted back to cDNA in order to be sequenced
- This strips all edited bases and epigenetic information from the molecules
- TGS has emerged as a promising alternative to genome-wide map of RNA modifications.

ONP pros/cons

Pros

- Fast run times – takes a few hours
- Cheap
- No need for fluorescent nucleotides since it uses semi-conductor technology

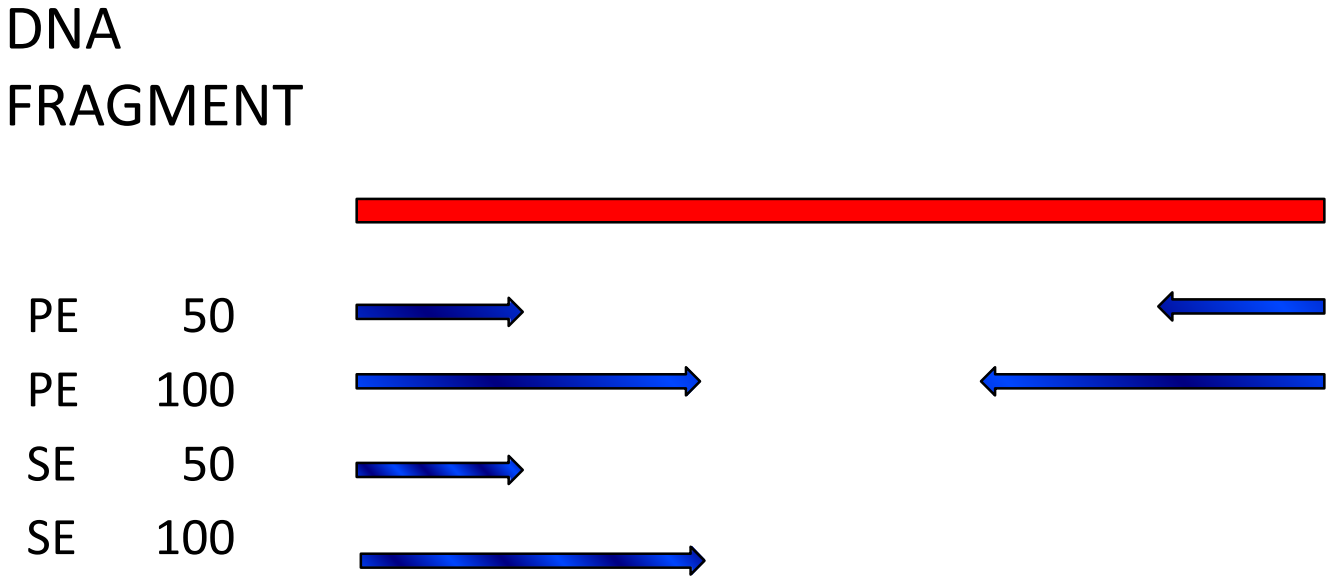
Cons

- High error/ lower per read accuracy
- No clinical applications yet, but high potential

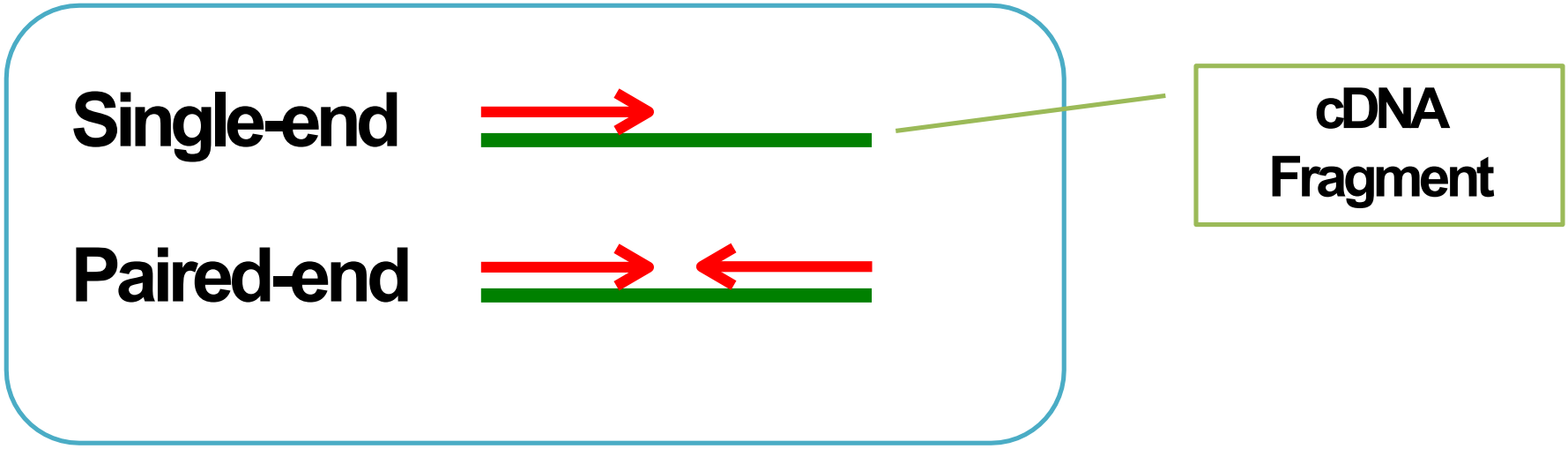
Sequencing Options

Illumina Sequencing options:

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



Single-end vs paired-end



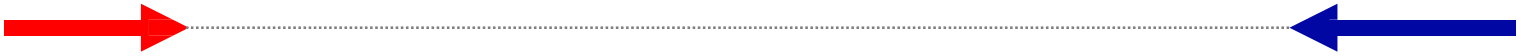
What you get from VIGR:

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.

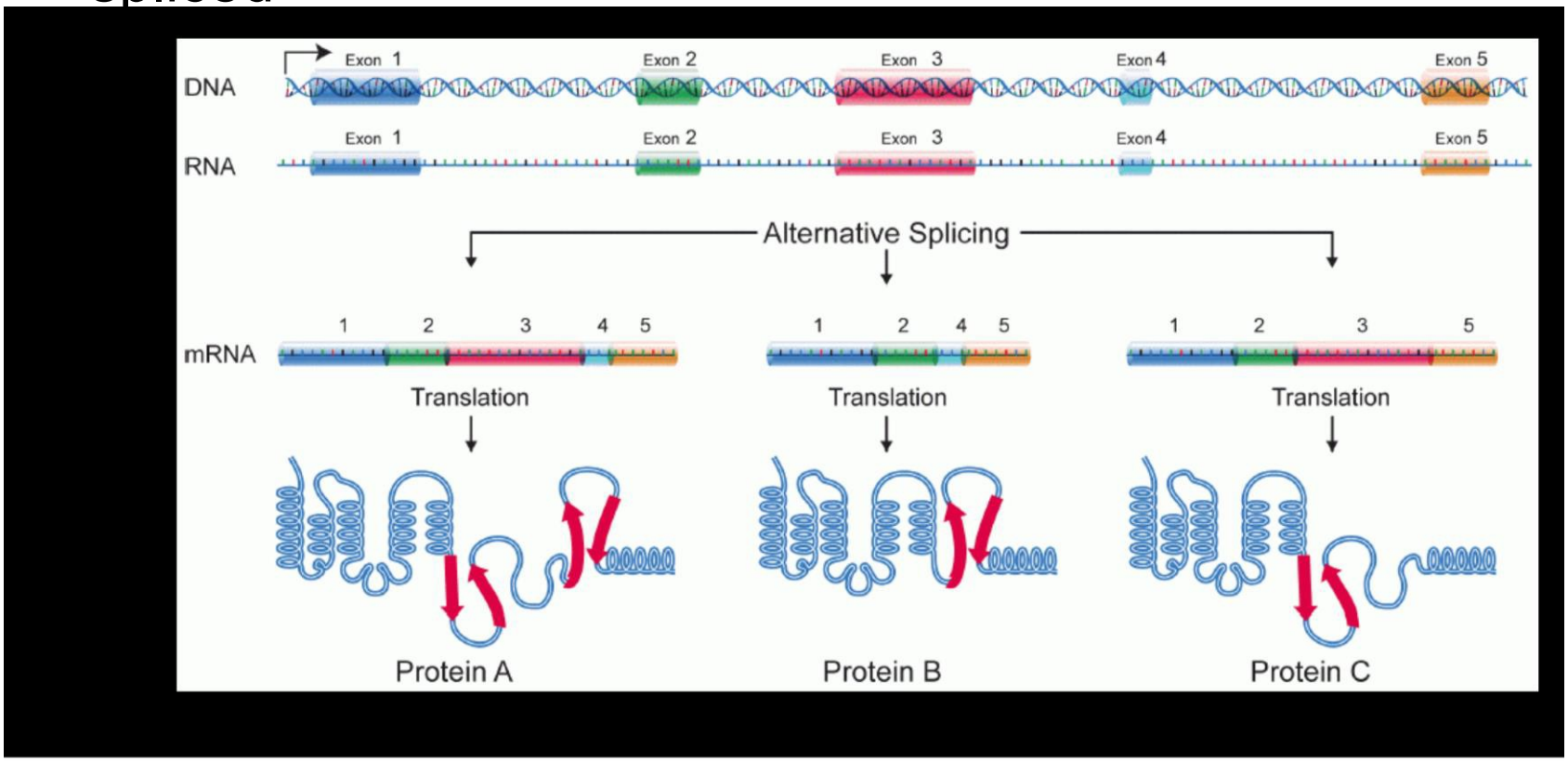
When does SE vs PE matter?

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%

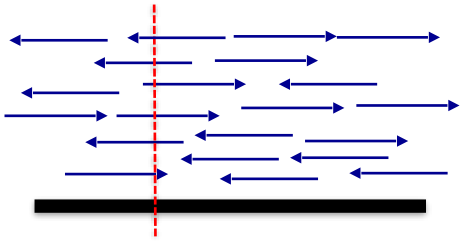
PMID: 26813401

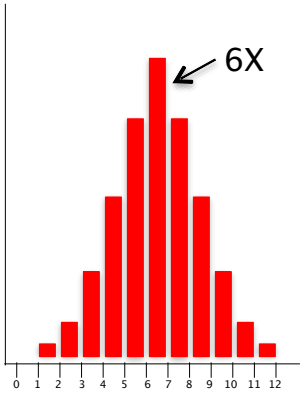
Other NGS terms

- **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

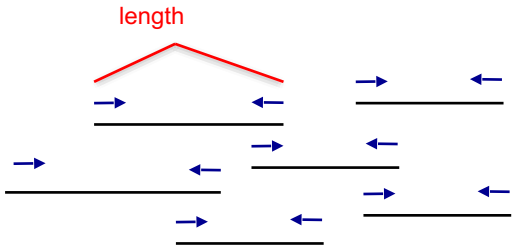
Sequence Coverage

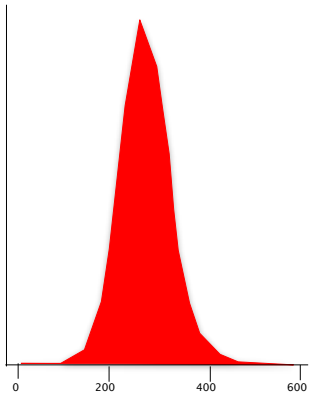




Base coverage by sequence

Insert Size





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*

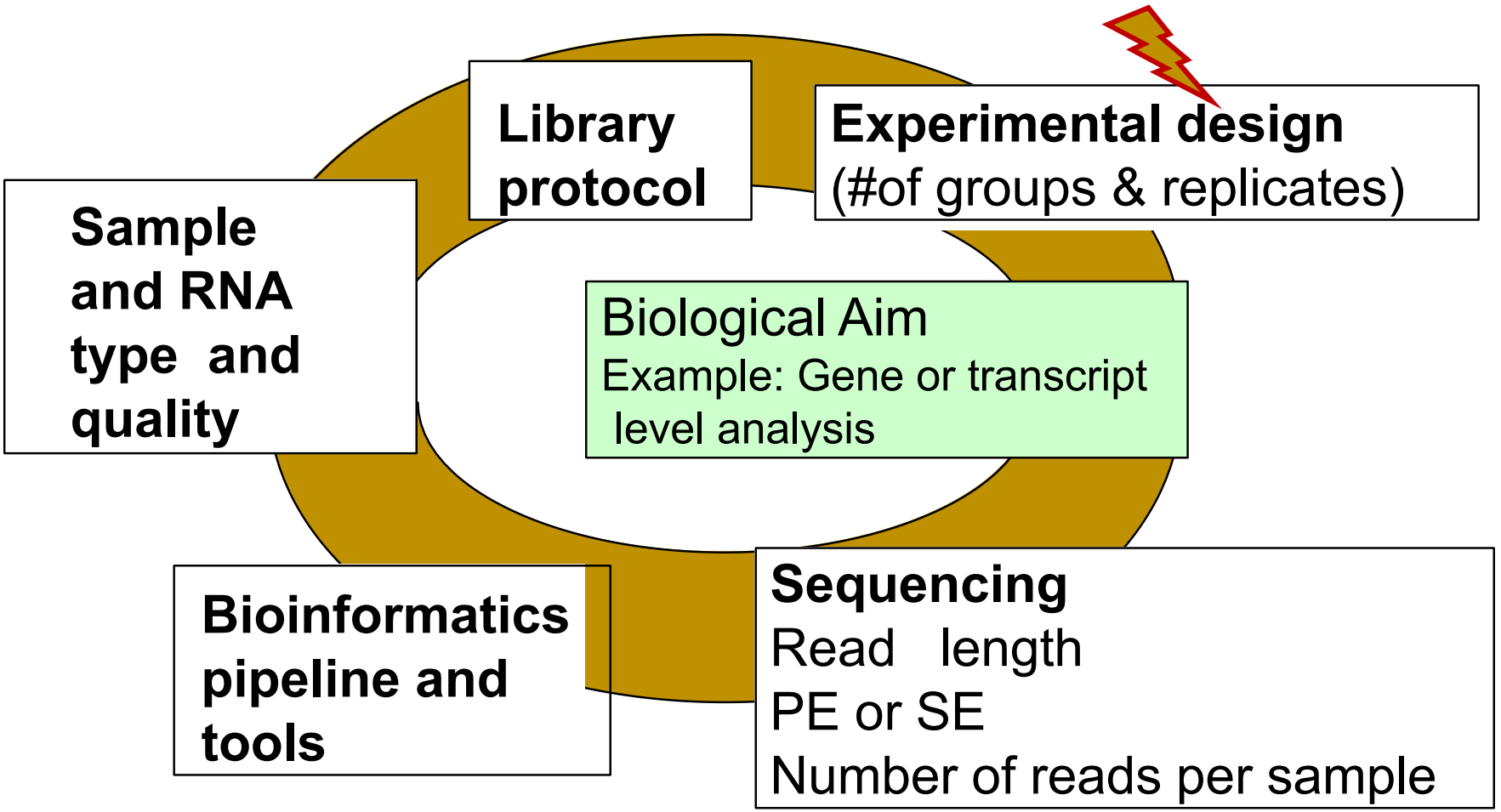
What you want to do versus what you can afford

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



Important considerations:

1. Number and type of **replicates**
2. Avoiding **confounding** variables
3. Avoiding (as much as possible) **batch effects**