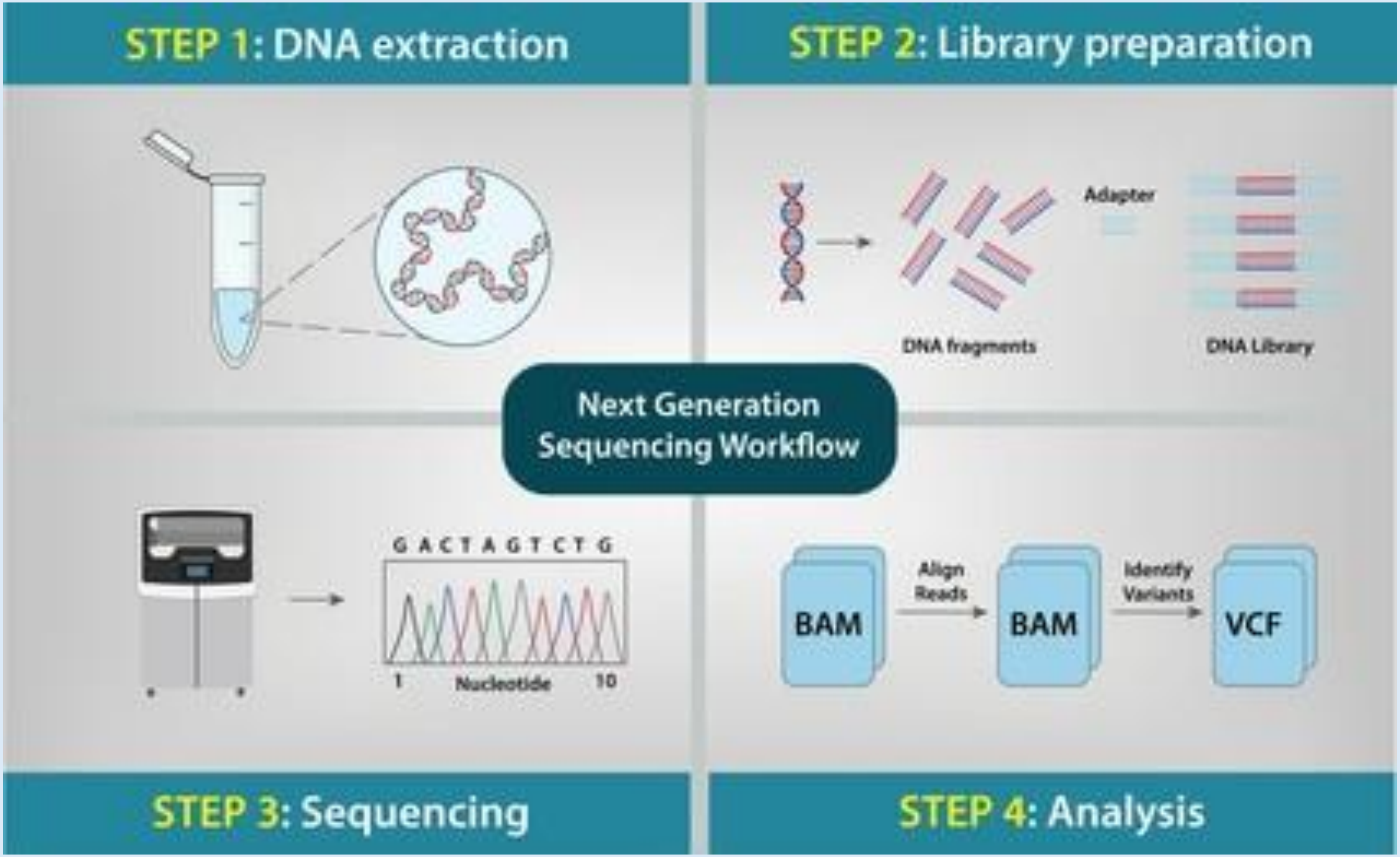


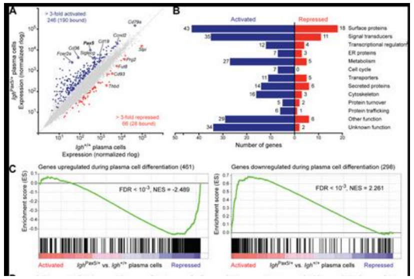
Final Project Discussion

Princess Rodriguez, PhD

MMG 3320
Spring 2026

Final Project Prompt

- You will analyze an NGS dataset of your choosing from "start" to "finish".
- You will begin by identifying your dataset.
- You will then download the data.
- You will then process it.
- You will then visualize it.
- You will then interpret and deliver your findings.



- Along the way you will perform QUALITY CONTROL

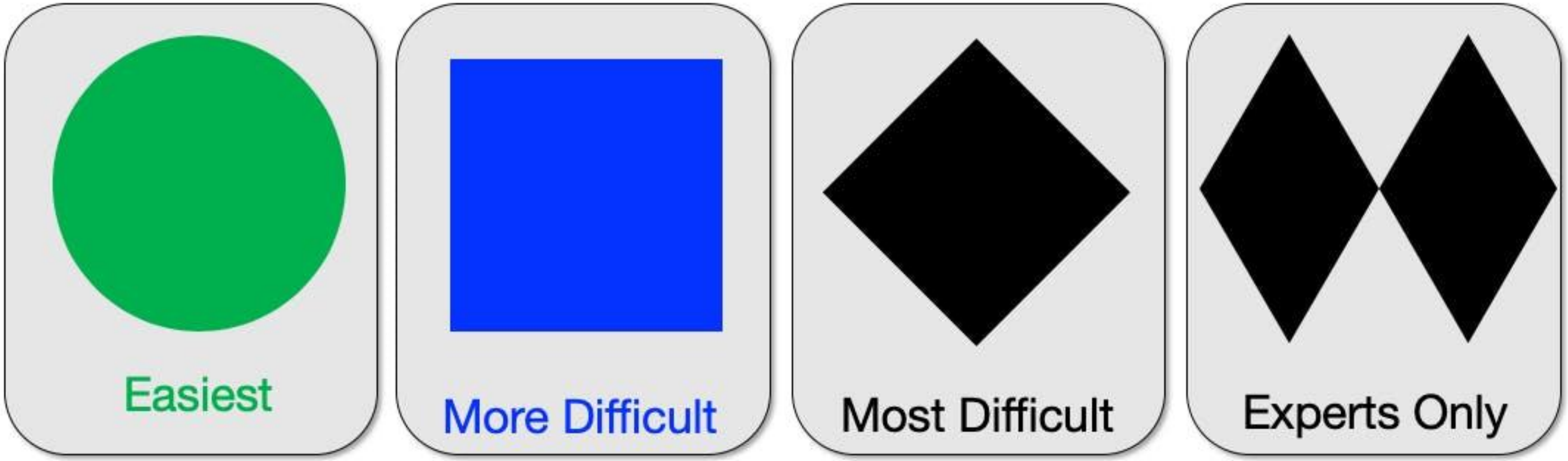
Final Project Delivery

- **All students** will deliver an **oral presentation** to communicate their results and interpretations.

- Everyone will submit a folder with their compiled analysis… more details to come!

- Everyone must be present for the final weeks' presentations.

Ski Trails



You will be asked to select a trail and a corresponding challenge.

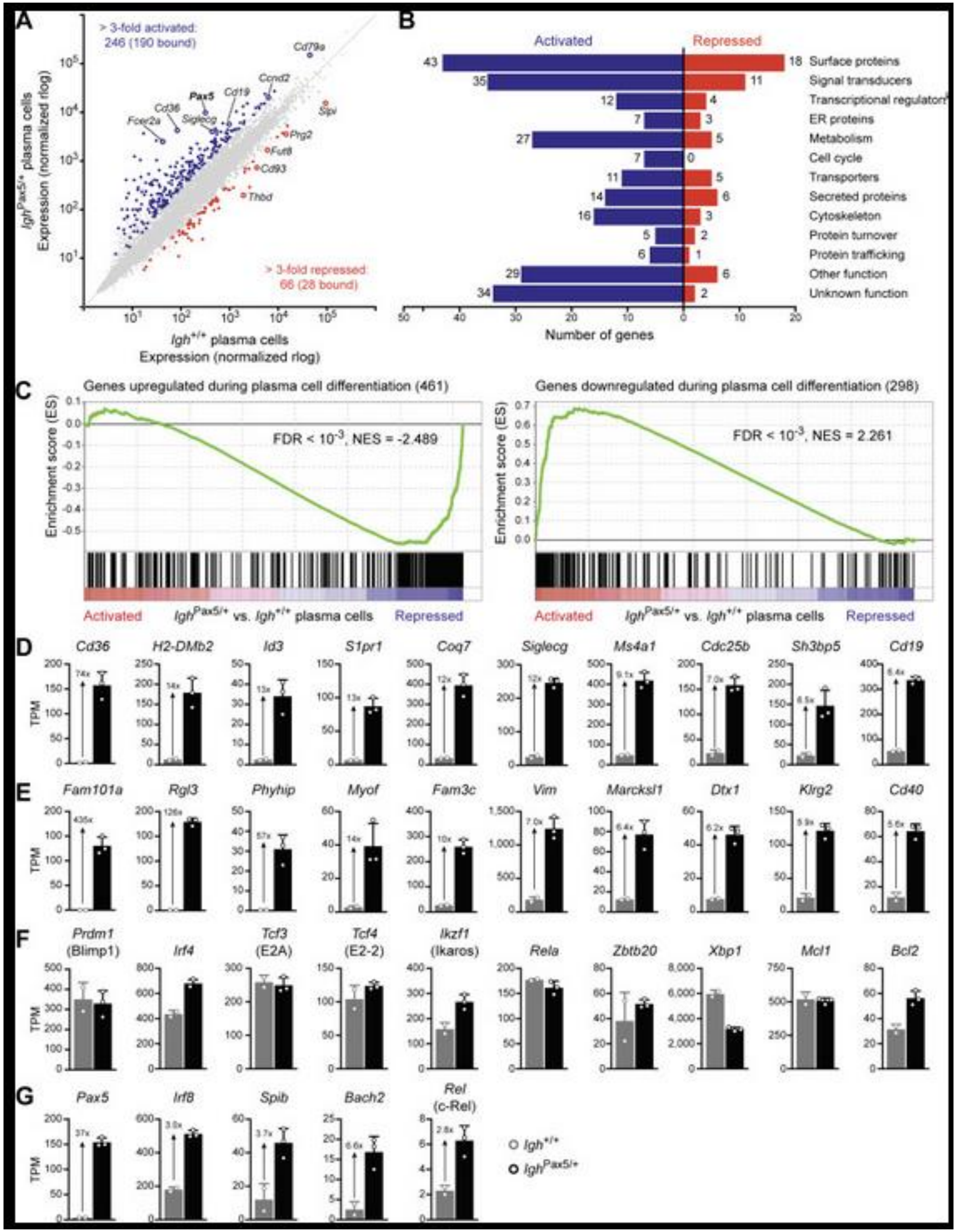
All challenge prompts below are *specific* to RNA-Seq. If you select a different kind of NGS dataset to analyze, I will generate a challenge prompt specific for your data type and trail.

Green Mountain Trail

Replicate a figure in a primary research article and then change one parameter at the visualization stage

Green Mountain Trail

Challenge 1: Adjusting the Threshold for Differential Gene Expression (DEG)

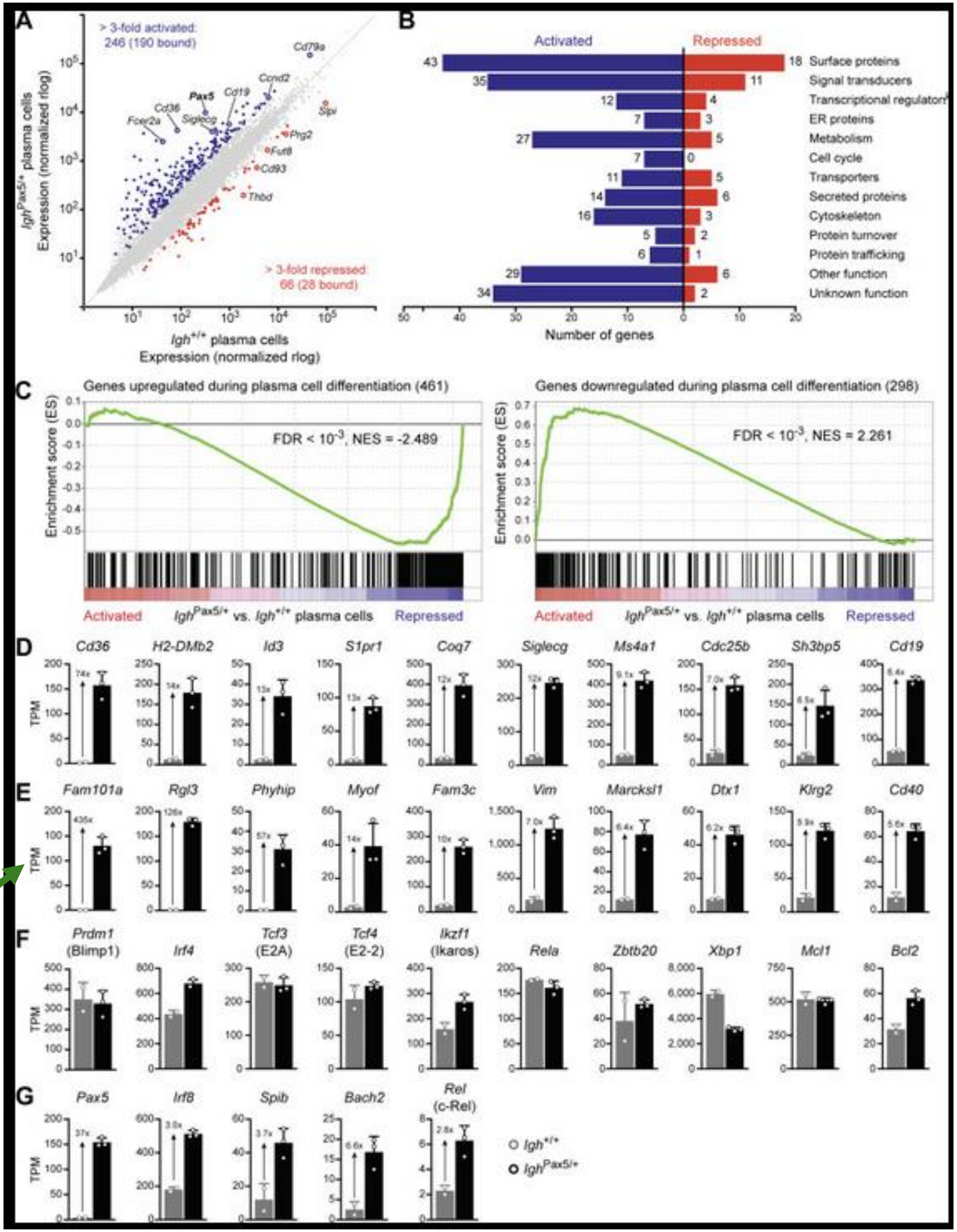


Green Mountain Trail

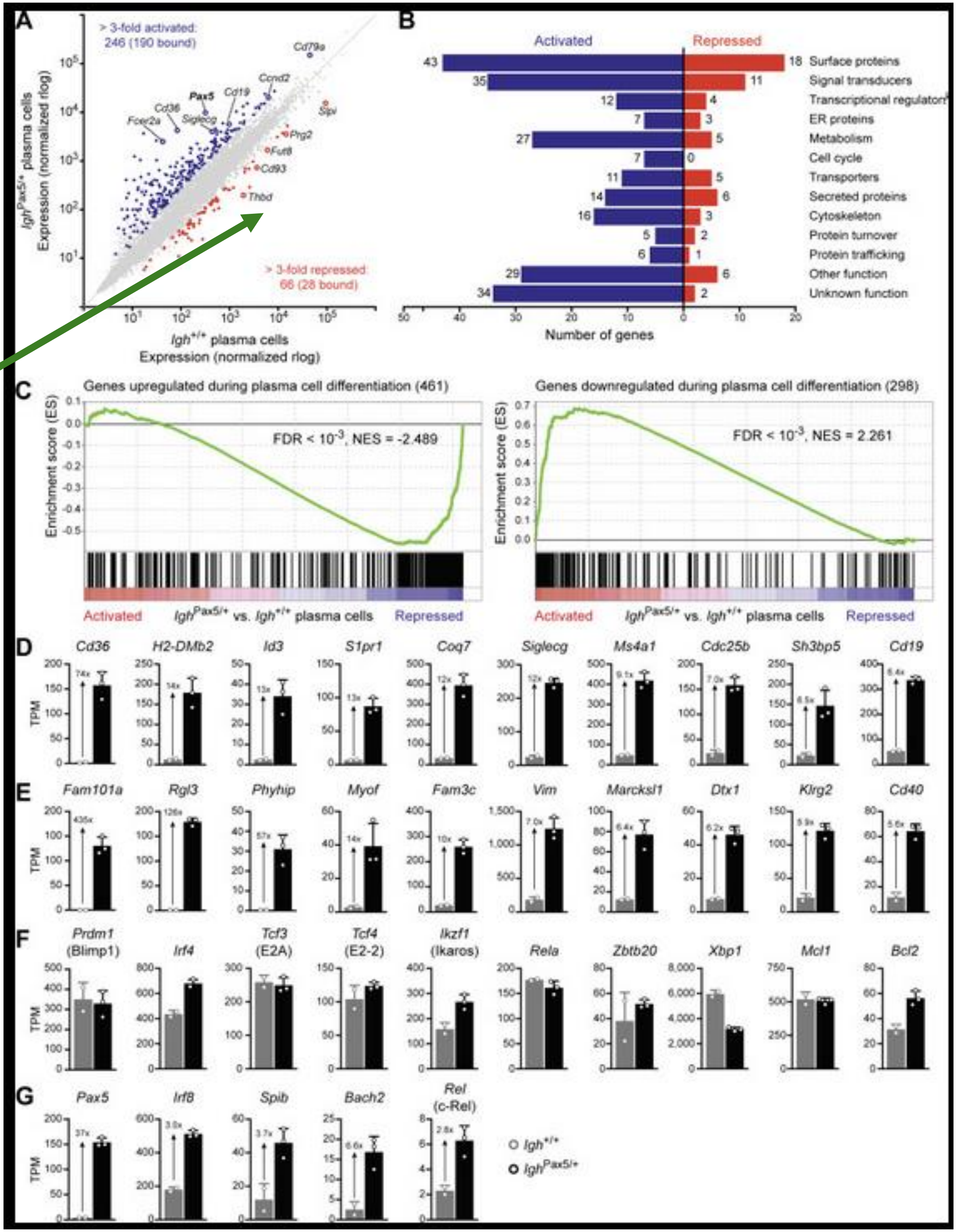
Challenge 2: Experimenting with Normalization Techniques

Compare visualizations of gene expression data normalized using two methods (e.g., **TPM, CPM, vs. DESeq2's variance-stabilizing transformation**). Assess how normalization affects downstream analyses like bar or box plots.

Do you know TPM means?



Green Mountain Trail



How would you label this MA plot?

Challenge 3: Changing Color Schemes/Labeling for Data Interpretation

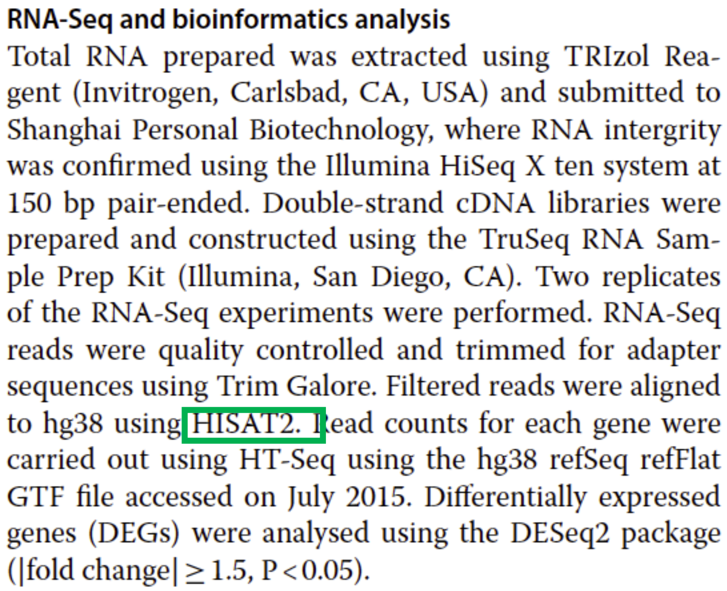
Adjust the color scale of a heatmap (e.g., changing from a red-green to a blue-yellow color scheme) and evaluate how the choice of visualization colors influences the clarity of expression trends and ease of data interpretation.

Blue Sky Trail

Compare and Contrast bioinformatic tools during the preprocessing stage and describe its impact on the data interpretation

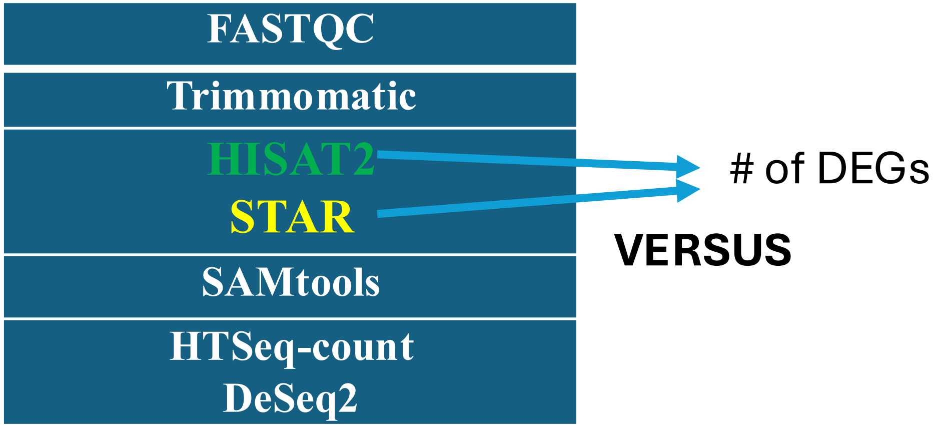
The bioinformatic pipeline we will learn in class is:

MMG3320	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 (|fold change| ≥ 1.5, P < 0.05).

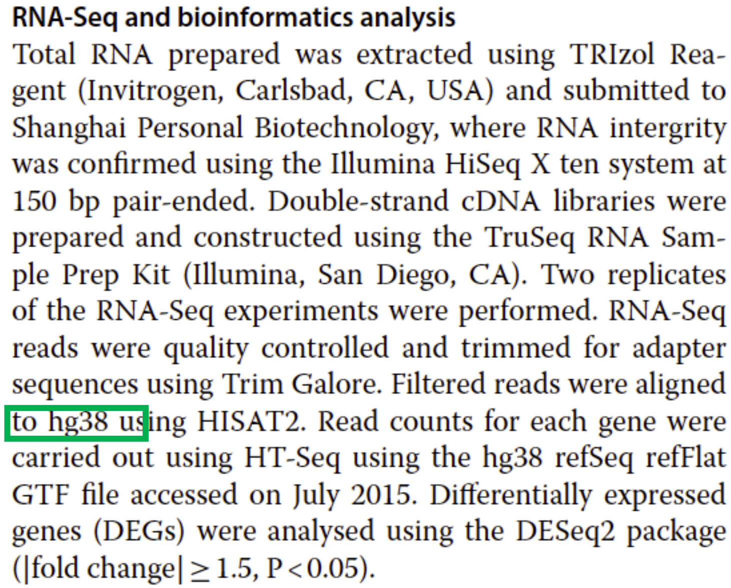


Information on the data processing can be found …?

GEO
Gene Expression Omnibus

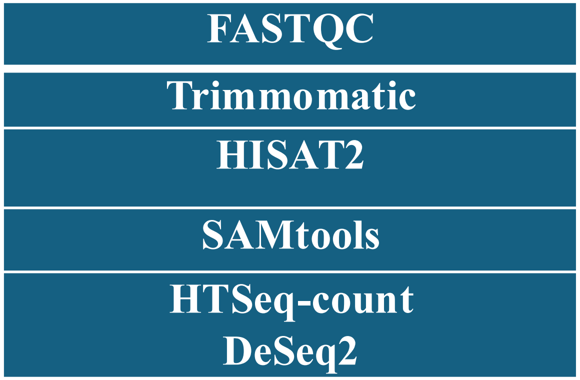
PubMed®

Challenge 1: Testing Different Alignment Tools
Align the RNA-Seq reads to the reference genome using two different aligners (e.g., HISAT2 vs. STAR). Compare metrics such as alignment rate, number of uniquely mapped reads, and runtime, and discuss how the choice of aligner might affect downstream analysis.



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

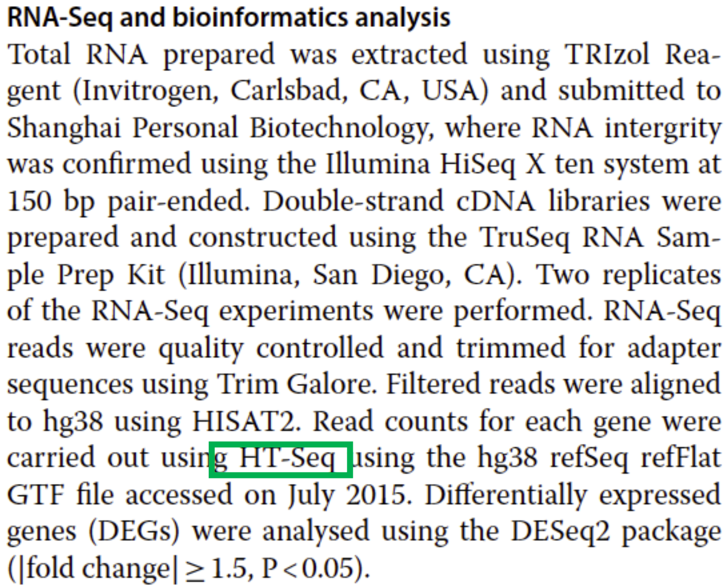
HISAT2

SAMtools

HTSeq-count
DeSeq2

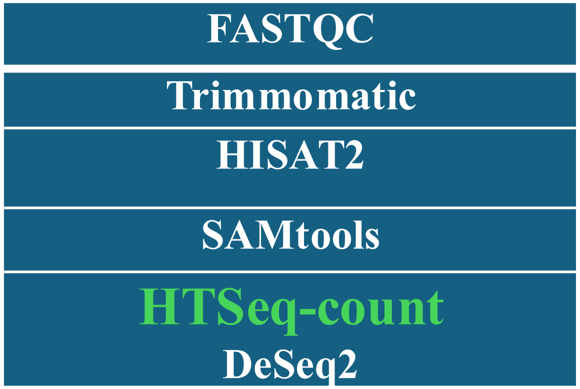
Challenge 2: Evaluating Reference Genome Versions

Map the RNA-Seq reads to two different versions of the reference genome (e.g., GRCh37 vs. GRCh38). Compare the alignment statistics and any differences in gene annotations. Discuss how the choice of reference genome might influence downstream results and biological interpretations.



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 (|fold change| ≥ 1.5, P < 0.05).



FASTQC

Trimmomatic

HISAT2

SAMtools

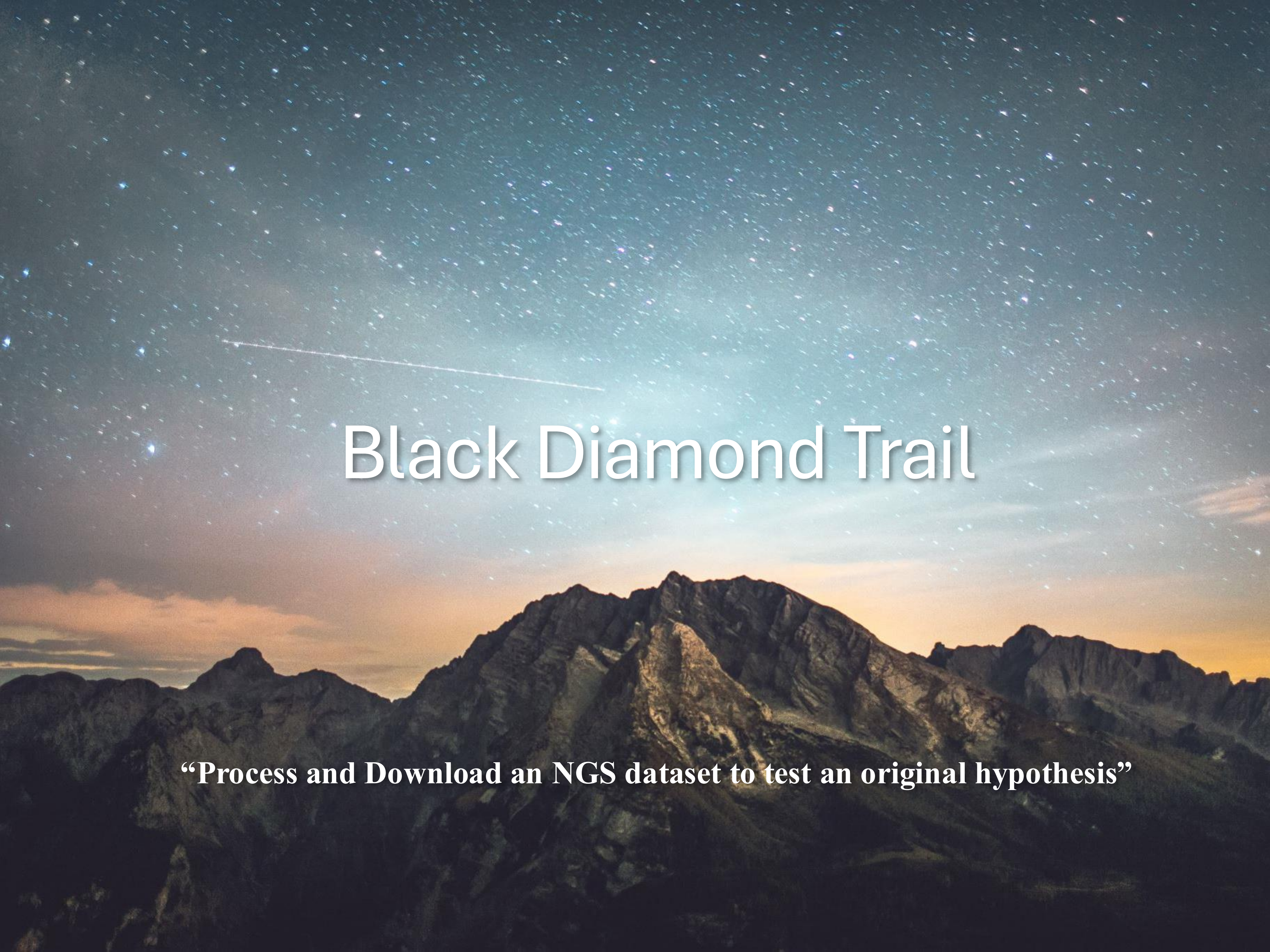
HTSeq-count

DeSeq2

Challenge 3: Comparing Count Generation Tools

Generate counts files using two different tools (e.g., HTSeq-count vs. featureCounts). Compare the total number of assigned reads, unassigned reads, and computational efficiency. Discuss how differences in counting strategies might influence downstream analyses such as differential expression.

The blue trail highlights the importance of tool selection during the preprocessing stage and its impact on the interpretation of RNA-Seq data.



Black Diamond Trail

“Process and Download an NGS dataset to test an original hypothesis”

Your overall approach will be different

You are going in with a hypothesis and using the dataset to test this hypothesis.

“Compared to macrophages, I hypothesize that Dendritic cells activated with LPS will express an upregulation of glycolytic genes as opposed to genes required for oxidative phosphorylation.”

“Compared to macrophages, I hypothesize that Dendritic cells activated with LPS will express an upregulation of glycolytic genes as opposed to genes required for oxidative phosphorylation.”

	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~~

Challenge 1: Creating Time-Series or Condition-Specific Plots
If your data includes multiple time points or conditions, create a figure (e.g., line plots or heatmaps) to visualize expression changes for key genes across these conditions. Highlight patterns or trends and discuss how they support or refute your biological hypothesis.

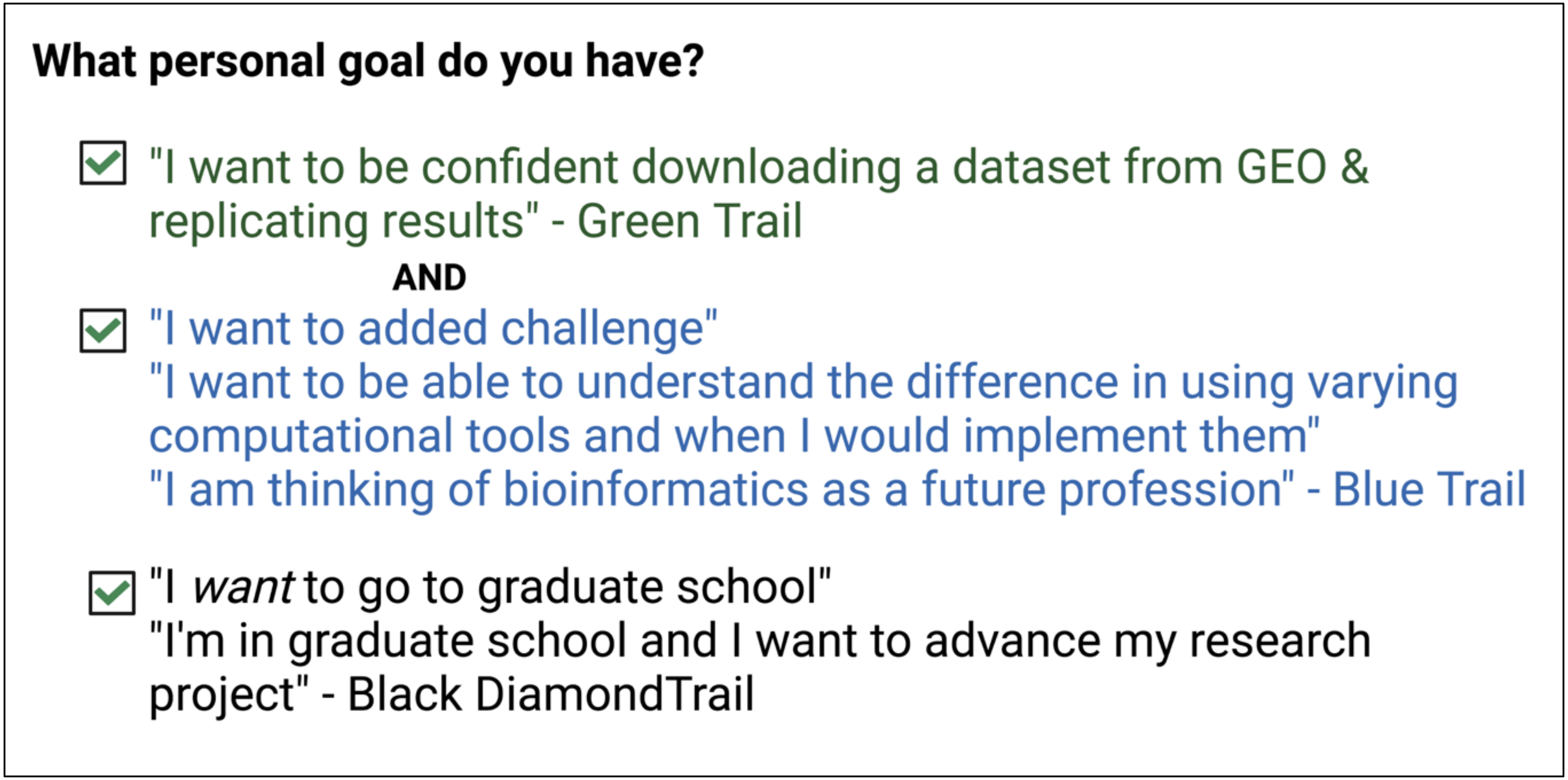
Challenge 2: Comparing Pathway Expression Across Groups
Use pathway analysis to identify key pathways enriched in a subset of your data. Create a customized plots (e.g. bar plots, dot plots, network graphs) to compare pathway activity between experimental groups not compared in the published work. Discuss how the visualization highlights the differences in pathway regulation.

Challenge 3: Annotating Single-Gene Expression Differences
Select a gene of interest from your dataset and create a violin plot or boxplot comparing its expression across conditions or groups. Customize the figure to include statistical annotations (e.g., p-values or fold changes) and explain why this gene is biologically significant.

For all black trail challenges you will be required to design a multi-panel figure that integrates multiple layers of analysis (e.g., a heatmap for expression patterns, a volcano plot for DEG results, and a GO enrichment bar chart). Explain how the combination of figures tells a cohesive story and enhances the overall interpretation of the data.

The black trail encourages students to think critically about data visualization while developing skills to create professional, publication-quality figures that clearly convey their <u>original</u> findings.

How do I select a trail?



What personal goal do you have?

☑ "I want to be confident downloading a dataset from GEO & replicating results" - Green Trail

AND

☑ "I want to added challenge"
"I want to be able to understand the difference in using varying computational tools and when I would implement them"
"I am thinking of bioinformatics as a future profession" - Blue Trail

☑ "I *want* to go to graduate school"
"I'm in graduate school and I want to advance my research project" - Black DiamondTrail

General

a. This is an individual assignment unless granted permission.

b. Each student will be allocated 15 minutes to present their findings and answer questions from the audience during the last week of class. All students are required to attend these sessions. The audience will be able to ask you questions during the presentation.

Selecting an NGS dataset

Acceptable	**Unacceptable**
RNA-Seq	Single-cell RNA-Seq
ChIP-Seq	Microarray
ATAC-Seq	Spatial Transcriptomics
Permission required: *Research-specific dataset Metagenomics WGS/WES	

Beware

	Selecting a dataset	**Download dataset**	**Index Genome**	**Alignment**
Estimated time to complete	1-2 weeks	24 hours	1hr – 3 days	***3-7 days +***
Comment		Per 5GB = 1.5 hrs = one sample	Depends on how large the genome is	Dependent on the number of samples Dependent on alignment strategy
Homework Assignment	~100 points Select dataset, and justify why dataset and trail were selected	~100 points FASTQC + interpretation		~150 points Alignment stats + interpretation Decision to be made on how to proceed based on interpretation
Due dates (tentative)	Mid Feb	Late Feb/ Early March		Early to Mid March

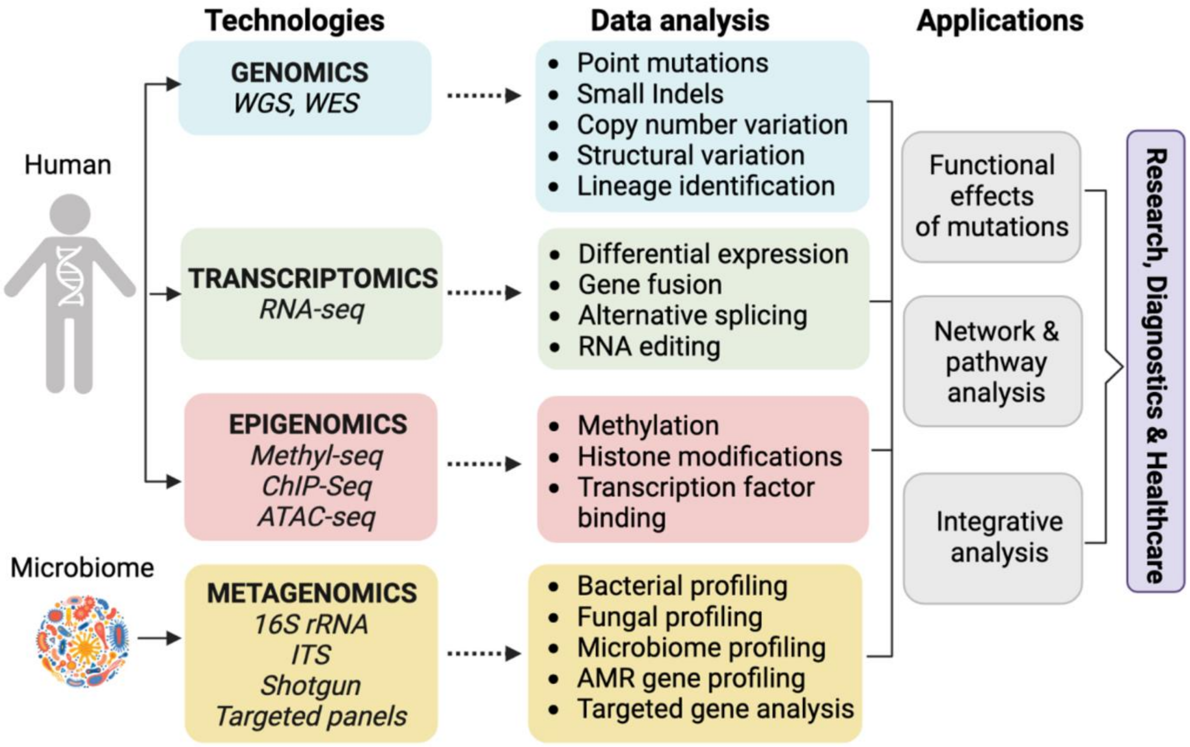
Important Disclosures

- While in-class, we will be going through the *basic steps* of data processing using a dataset that is publicly available.
- This project requires that you use what you learned in-class and apply it to a different NGS dataset.
- ***We both will not know the quality of the published dataset you selected until about March.*** Therefore, depending on what we find we may need to pivot and change the intention of the final project goals.
- I am most familiar with advising on a human or mouse system. However, other organisms are completely fine to select. You will be in charge of understanding if for example "*…there are pathway analysis tools available for Drosophila…" or where to find the GTF file for bacteria.*
- ***We will hit many unforeseen hiccups.*** This is completely normal in the realm of bioinformatics! Be prepared to troubleshoot.
- I do not have control over how fast or slow your data will process on the VACC. The alignment step is the most COMPUTATIONAL HEAVY STEP of the ENTIRE pipeline. Please do not leave this for the last minute as the VACC does have multiple users!

Lessons from Last Year

- If you select black trail but then see around April that your analysis is more aligned with green trail, this is 100% okay. But you must consult with me and tell me at least a week prior to your presentation that you will be ***changing trails***. There will be a major point deduction if your presentation and trail selected do not match!
- If you select the black trail, I expect an original hypothesis to be tested. Points will be deducted if this original hypothesis is not present or tested.
- I had multiple students throughout the years who opted to analyze a dataset “sitting in their lab.” Some of these students were wildly successful, others were not.

Common experimental designs for NGS



Technologies

Data analysis

Applications

Human

GENOMICS
WGS, WES

- Point mutations
- Small Indels
- Copy number variation
- Structural variation
- Lineage identification

TRANSCRIPTOMICS
RNA-seq

- Differential expression
- Gene fusion
- Alternative splicing
- RNA editing

EPIGENOMICS
Methyl-seq
ChIP-Seq
ATAC-seq

- Methylation
- Histone modifications
- Transcription factor binding

Microbiome

METAGENOMICS
16S rRNA
ITS
Shotgun
Targeted panels

- Bacterial profiling
- Fungal profiling
- Microbiome profiling
- AMR gene profiling
- Targeted gene analysis

Functional effects of mutations

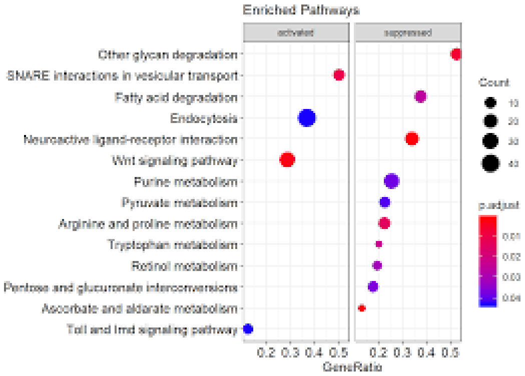
Network & pathway analysis

Integrative analysis

Research, Diagnostics & Healthcare

Bulk RNA-Seq: When to Use it?

- Measures average gene expression across a tissue or cell population



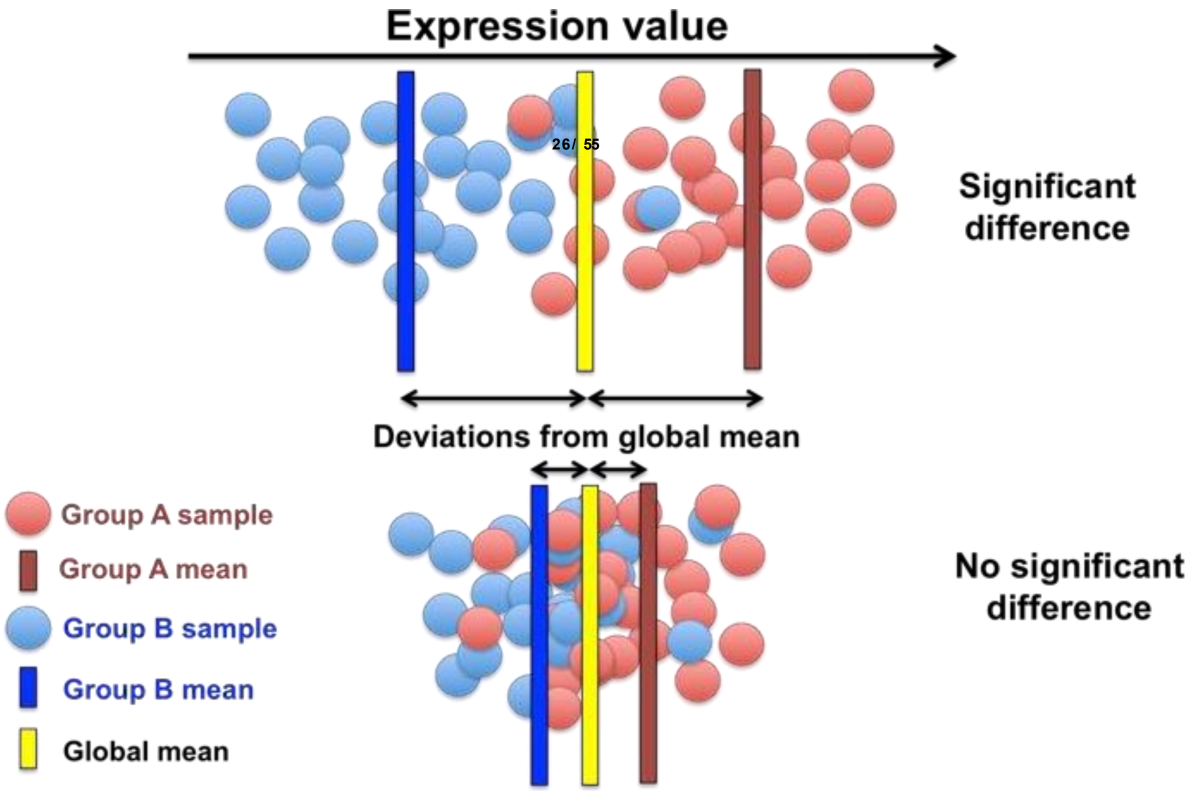
Best for:
❖ Comparing disease vs control
❖ Treatment response studies

- Pros: cost effective & robust
- Limitations: obscures cell-type composition

- Useful when the research question is about global expression changes or pathways, but not cell-type resolution.

Basic types of questions answered:

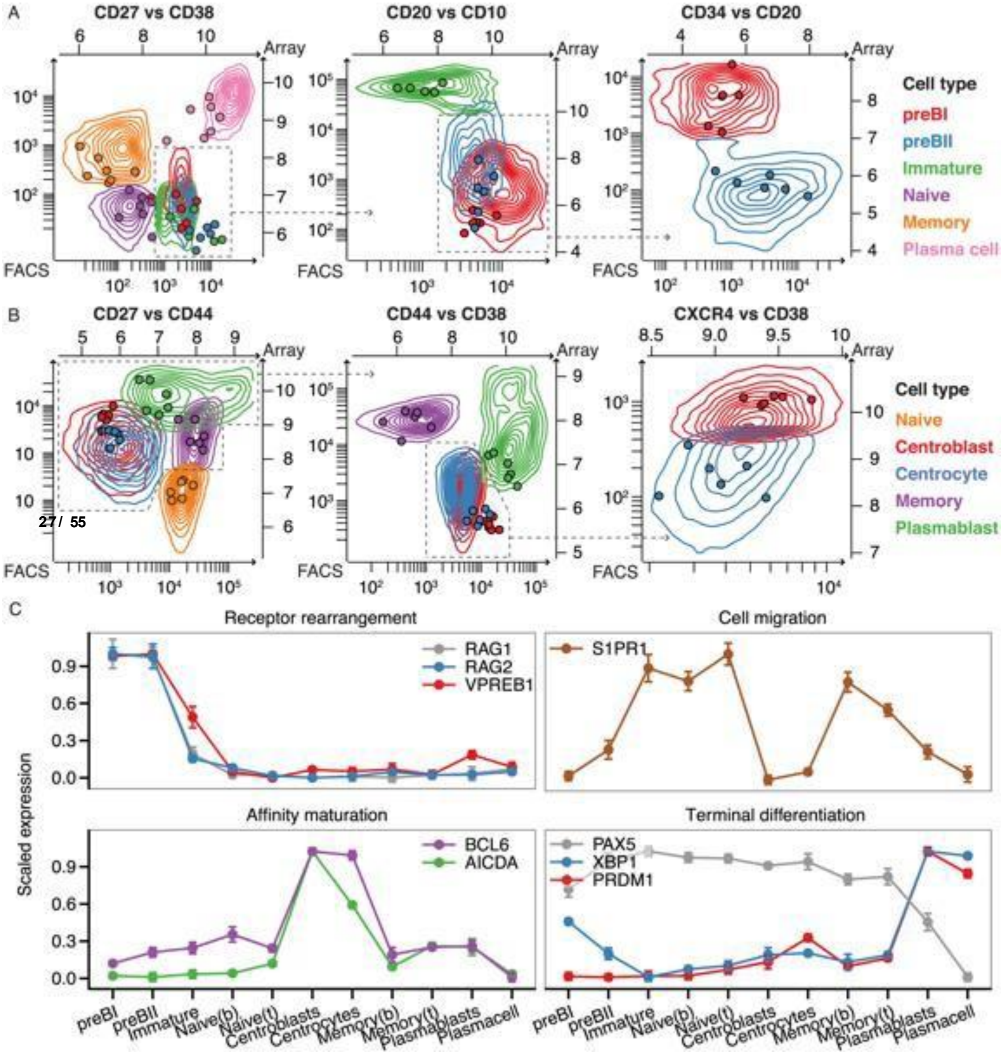
What genes are differentially expressed between conditions?



Other questions answered:

Are there any trends in gene expression across development?

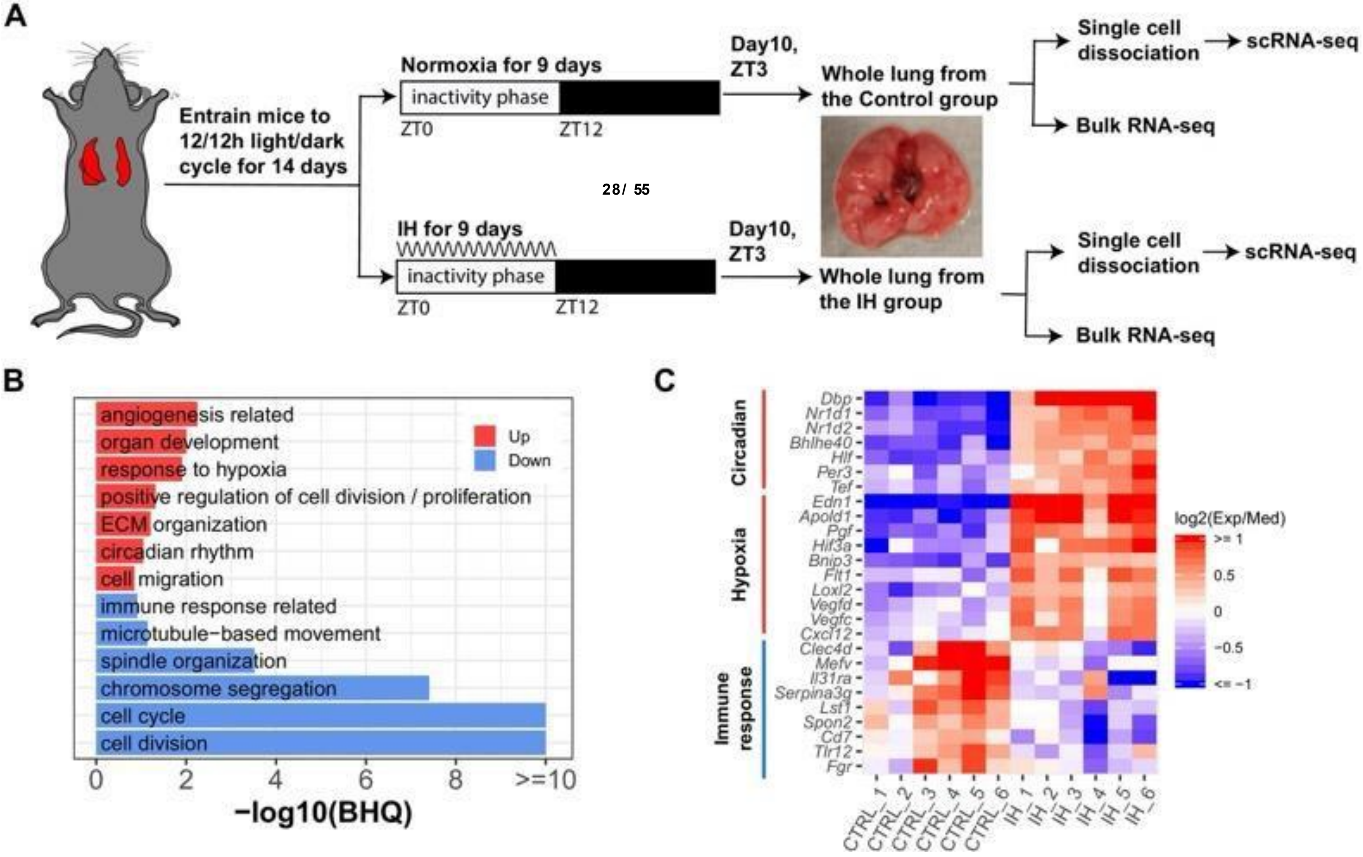
Which groups of genes change similarly over time or across conditions?



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

Basic types of questions answered:

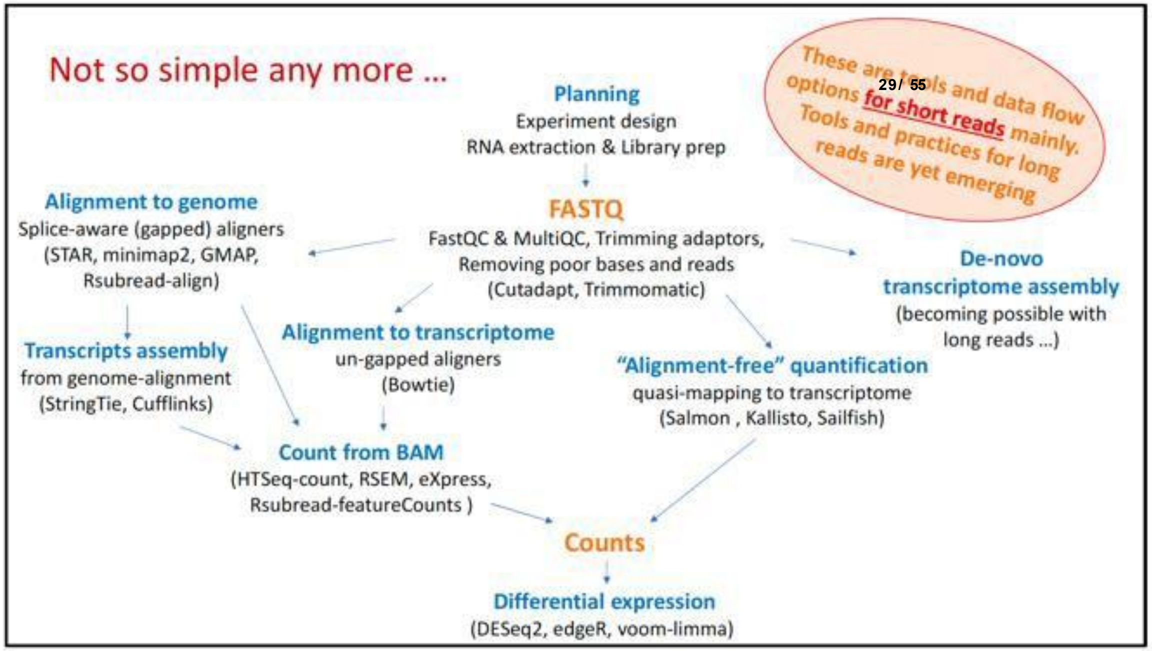
What processes or pathways are enriched in condition of interest?



https://elifesciences.org/articles/63003

Basic Principals

- Study Design
- Quality Assessment (UNIX)
- Trimming & Preprocessing (UNIX)
- Alignment (UNIX)
- Visualization of BAMs/counts (R)



Not so simple any more ...

These are tools and data flow options **for short reads** mainly. Tools and practices for long reads are yet emerging

Planning
Experiment design
RNA extraction & Library prep

FASTQ
FastQC & MultiQC, Trimming adaptors,
Removing poor bases and reads
(Cutadapt, Trimmomatic)

Alignment to genome
Splice-aware (gapped) aligners
(STAR, minimap2, GMAP,
Rsubread-align)

De-novo transcriptome assembly
(becoming possible with
long reads ...)

Transcripts assembly
from genome-alignment
(StringTie, Cufflinks)

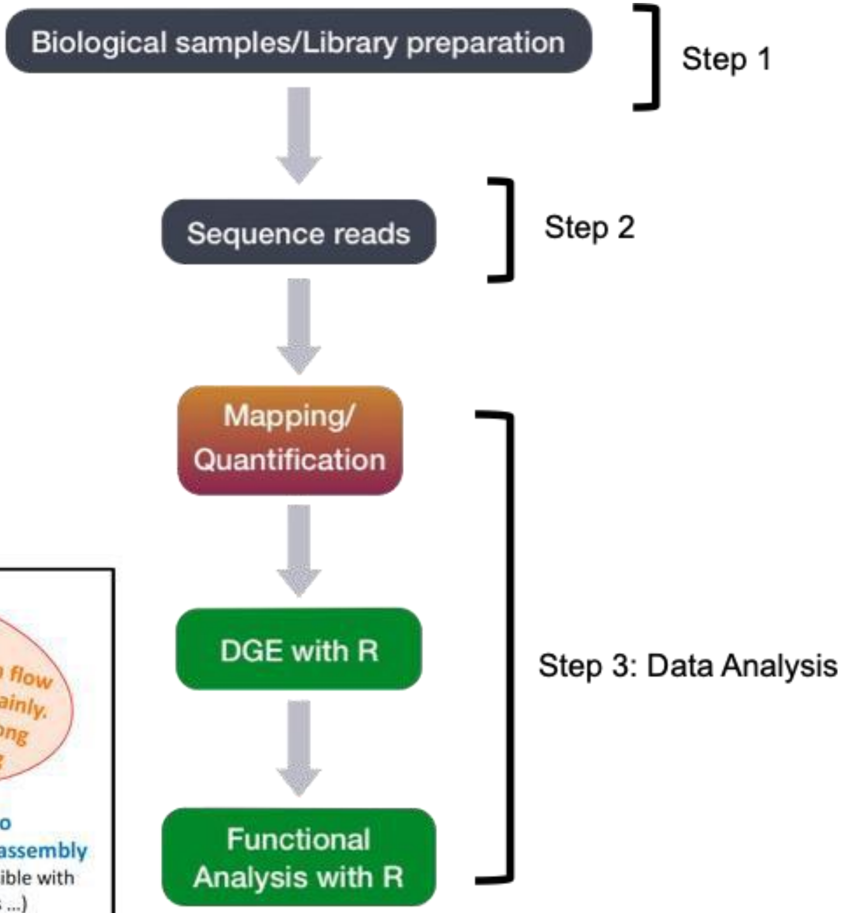
Alignment to transcriptome
un-gapped aligners
(Bowtie)

"Alignment-free" quantification
quasi-mapping to transcriptome
(Salmon , Kallisto, Sailfish)

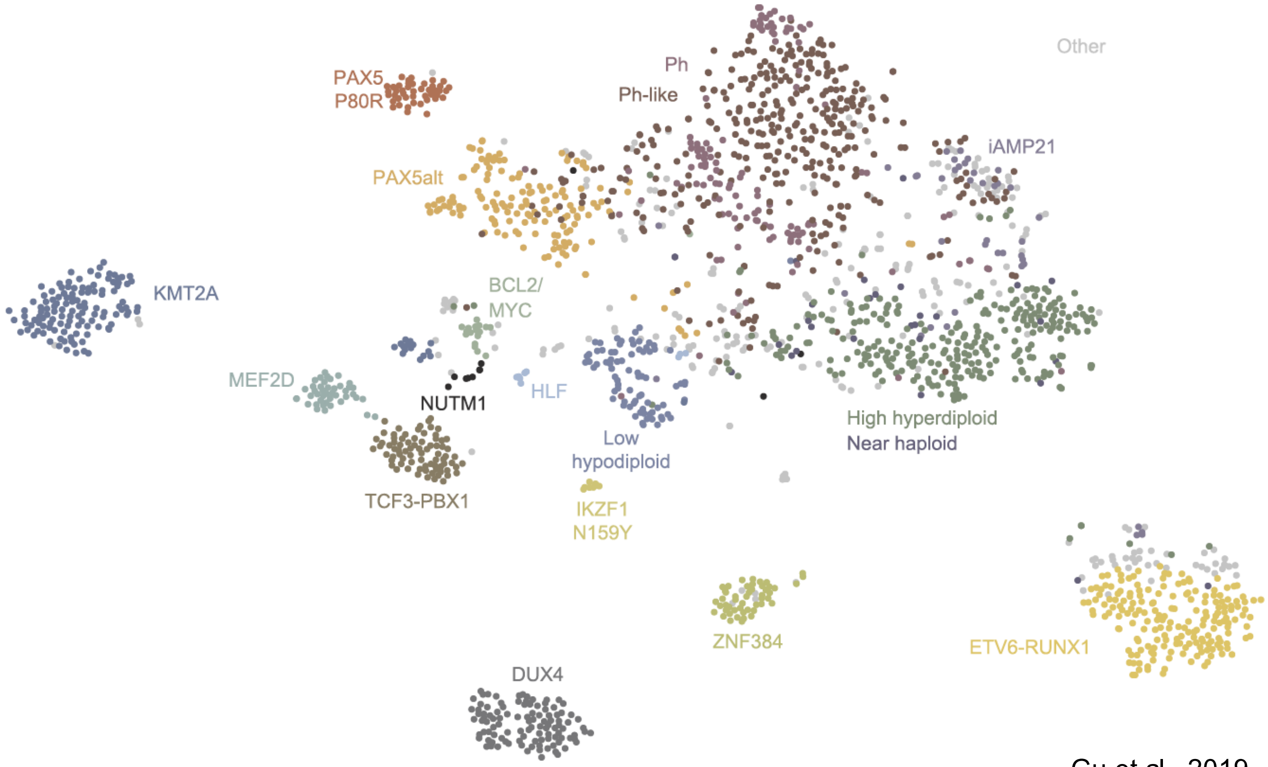
Count from BAM
(HTSeq-count, RSEM, eXpress,
Rsubread-featureCounts)

Counts

Differential expression
(DESeq2, edgeR, voom-limma)



RNA-Seq led to the identification of new subtypes in B-ALL



Gu et al., 2019

Motivation - Single cell level insights

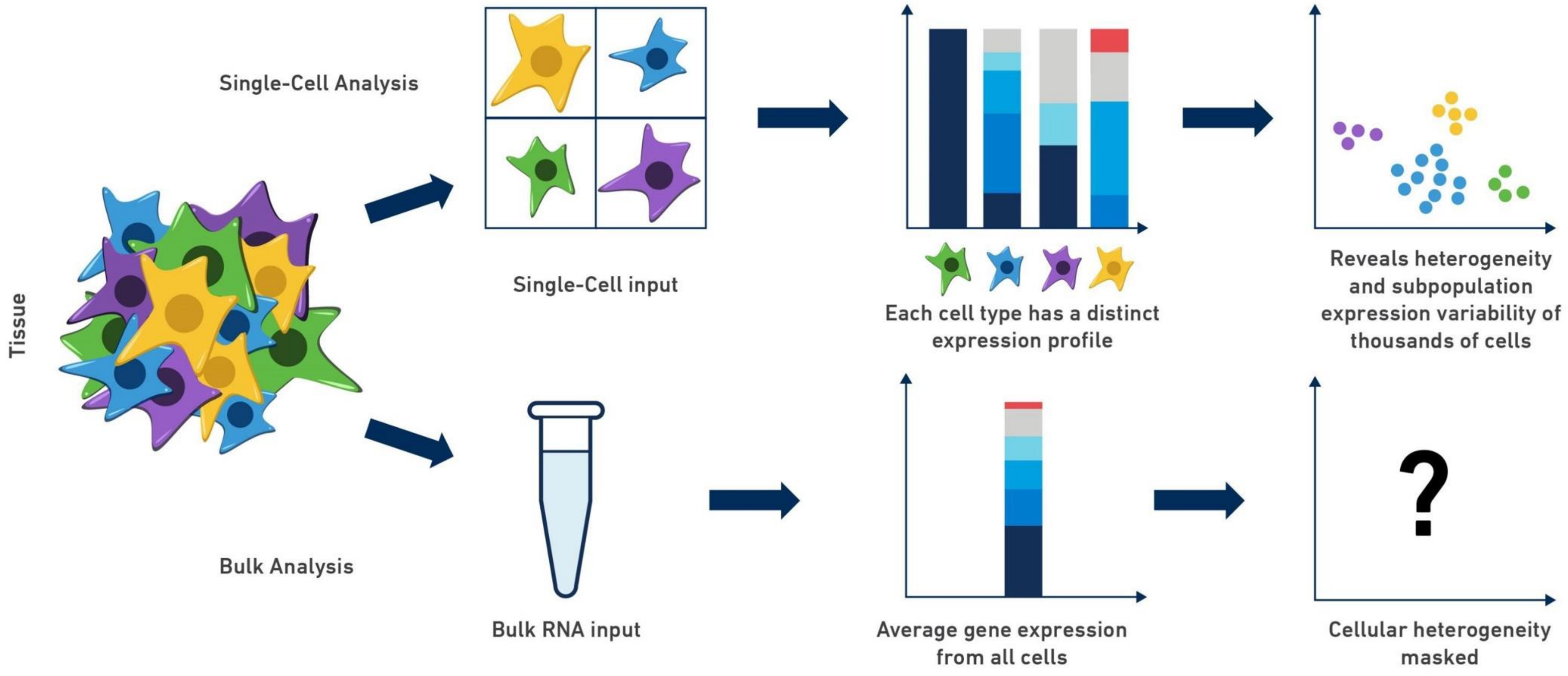
Bulk RNA-seq

Taste the **average** of all fruits

Single cell RNA-seq

Taste each fruit **individually**

Bulk RNA- vs. single cell RNA-seq


Tissue
Single-Cell Analysis
Single-Cell input
Each cell type has a distinct expression profile
Reveals heterogeneity and subpopulation expression variability of thousands of cells
Bulk Analysis
Bulk RNA input
Average gene expression from all cells
?
Cellular heterogeneity masked

Single-Cell RNA-Seq

- Measures expression **cell-by-cell**

- Best for:

❖ Identifying novel cell types/states
❖ Uncovering tumor heterogeneity
❖ Immune profiling
❖ Understanding treatment-resistant populations

- Pros: High-resolution
- Limitations: Expensive, technological challenges as the data is noisy

- Select scRNA-seq when you suspect heterogeneity matters

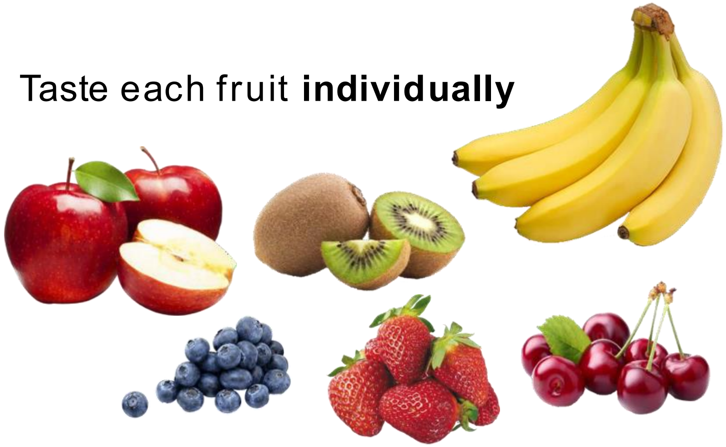
Single cell level SPATIAL insights

Bulk RNA-seq

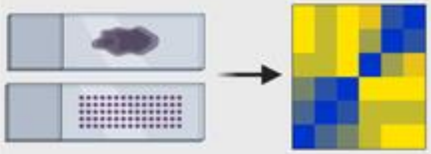
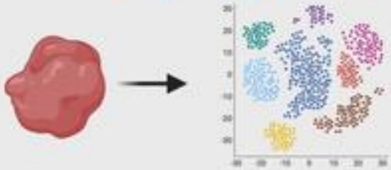
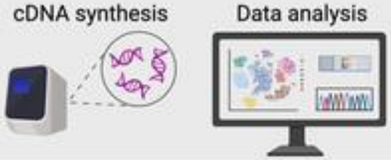
Taste the
average of
all fruits

Single cell RNA-seq

Taste each fruit **individually**



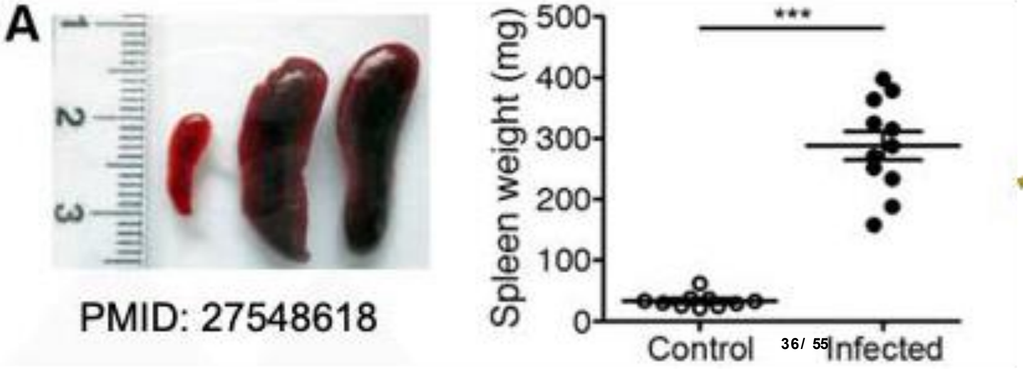
Spatial transcriptomics

	Bulk RNA sequencing 	Single-cell RNA Sequencing 	High Throughput Spatial Transcriptomics 
Resolution	Patient-level gene expression	Single-cell resolution	Spatial and single-cell resolution
Data Format	Aggregated gene expression from whole tissue	Gene expression at the individual level	Gene expression mapped to 2D tissue coordinates
Biological Insights	Identifies differentially expressed genes at the tissue level	Identifies cell subpopulations and heterogeneity	Reveals spatially variable genes, tissue structure, and cell-cell interactions
Advantages/ Disadvantages	✓ High sensitivity for overall gene expression ✕ No spatial or single-cell resolution	✓ High resolution for individual cell types ✕ Loses spatial context and cell-cell interactions	✓ Preserves tissue architecture and spatial relationships ✕ Lower sensitivity for minimally-expressed genes

Experimental workflow

1

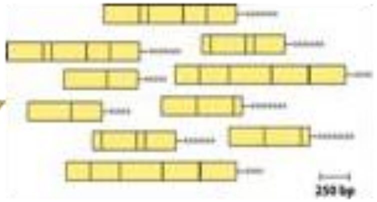
Samples of interest



PMID: 27548618

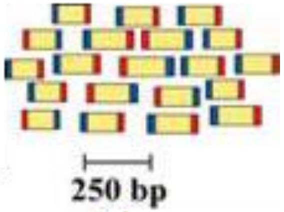
2

Isolate RNAs

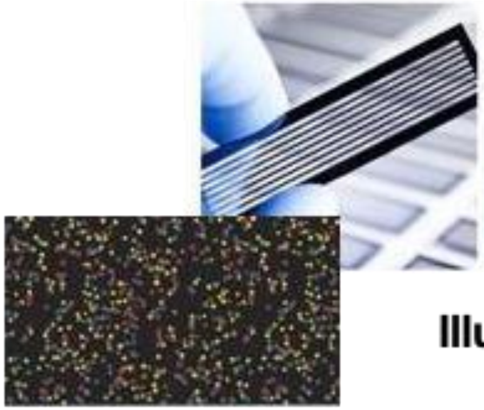


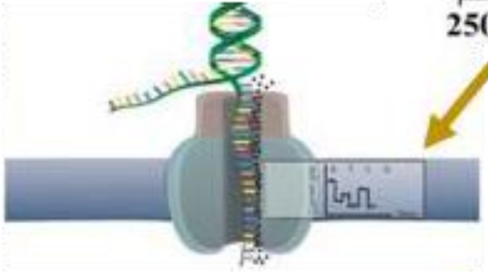
3

Library build



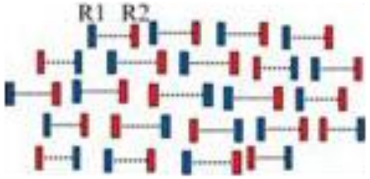
4





Illumina sequencing
versus
ONP

5



Reads (R1 and R2) generated

Biological Replicates

Experimental replicates can be performed as **technical replicates** or **biological replicates**.

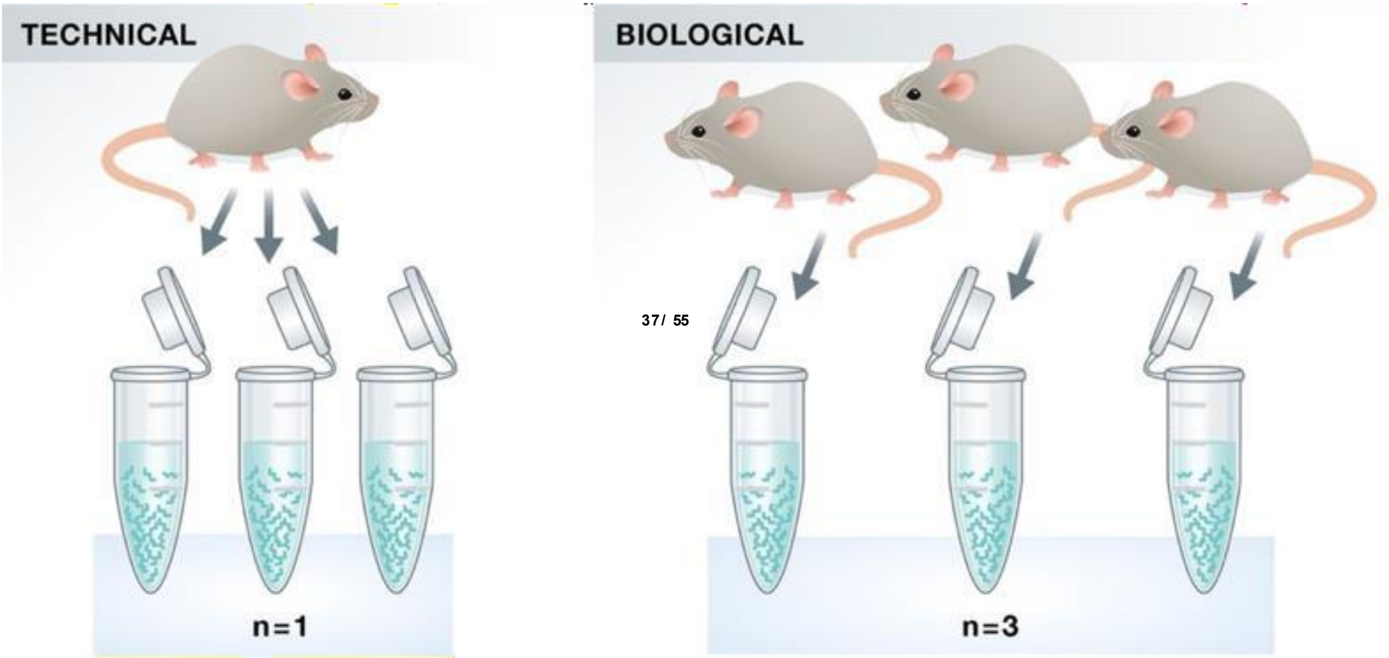
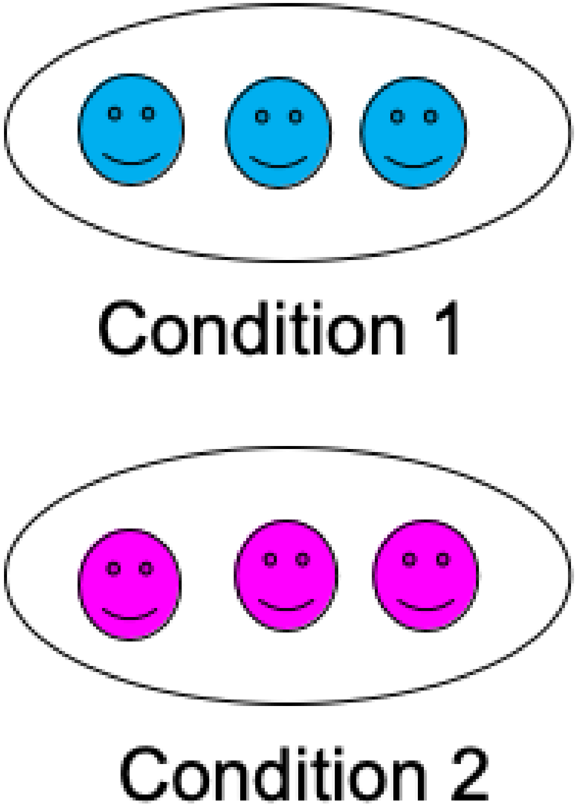


Figure 16: Biological Replicates

*Image credit: [Klaus B., EMBO J (2015) **34**: 2727-2730]*

- **Technical replicates:** use the same biological sample to repeat the technical or experimental steps in order to accurately measure technical variation and remove it during analysis.
- **Biological replicates** use different biological samples of the same condition to measure the biological variation between samples.

Biological Replicates



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

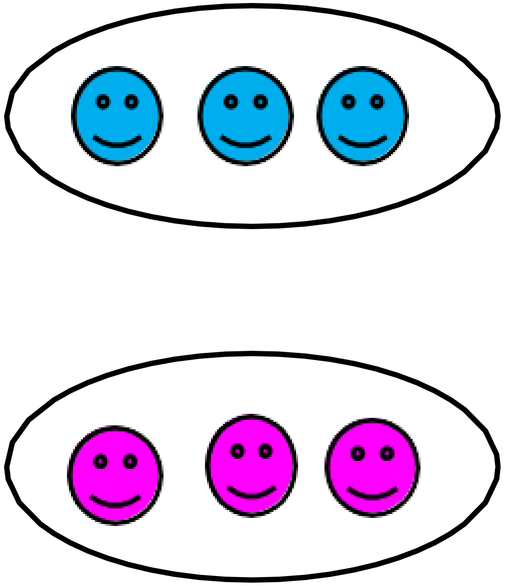
Probability of detecting DEGs

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

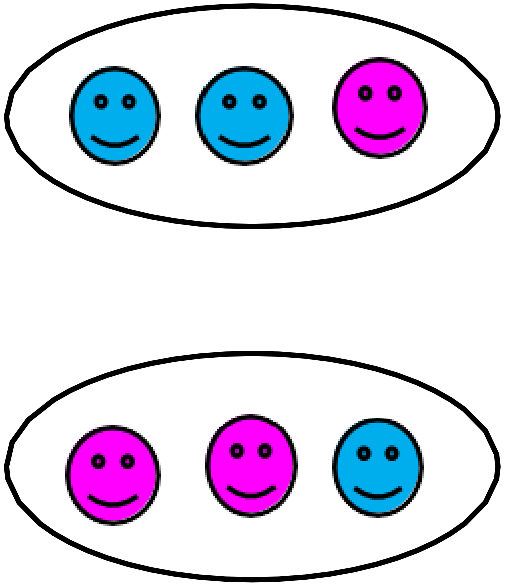
40/ 55

PMID: 26813401

Grouping of Replicates

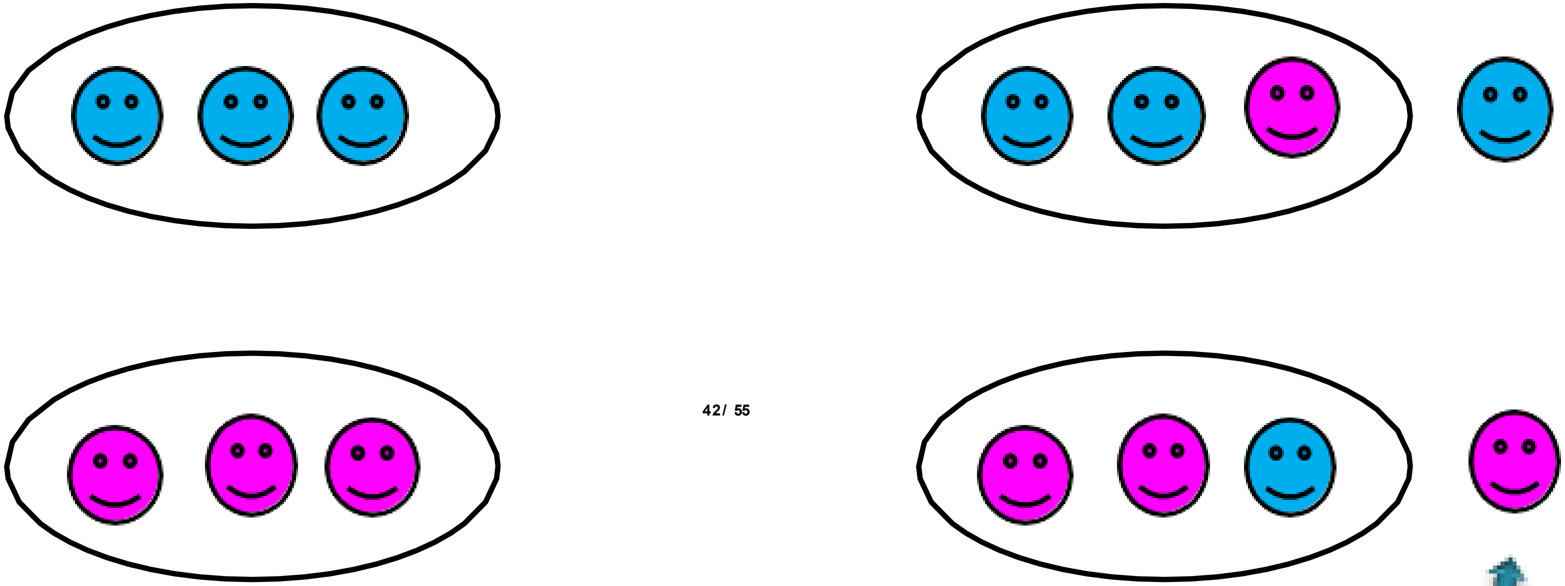


What you want



What you get

Grouping of Replicates



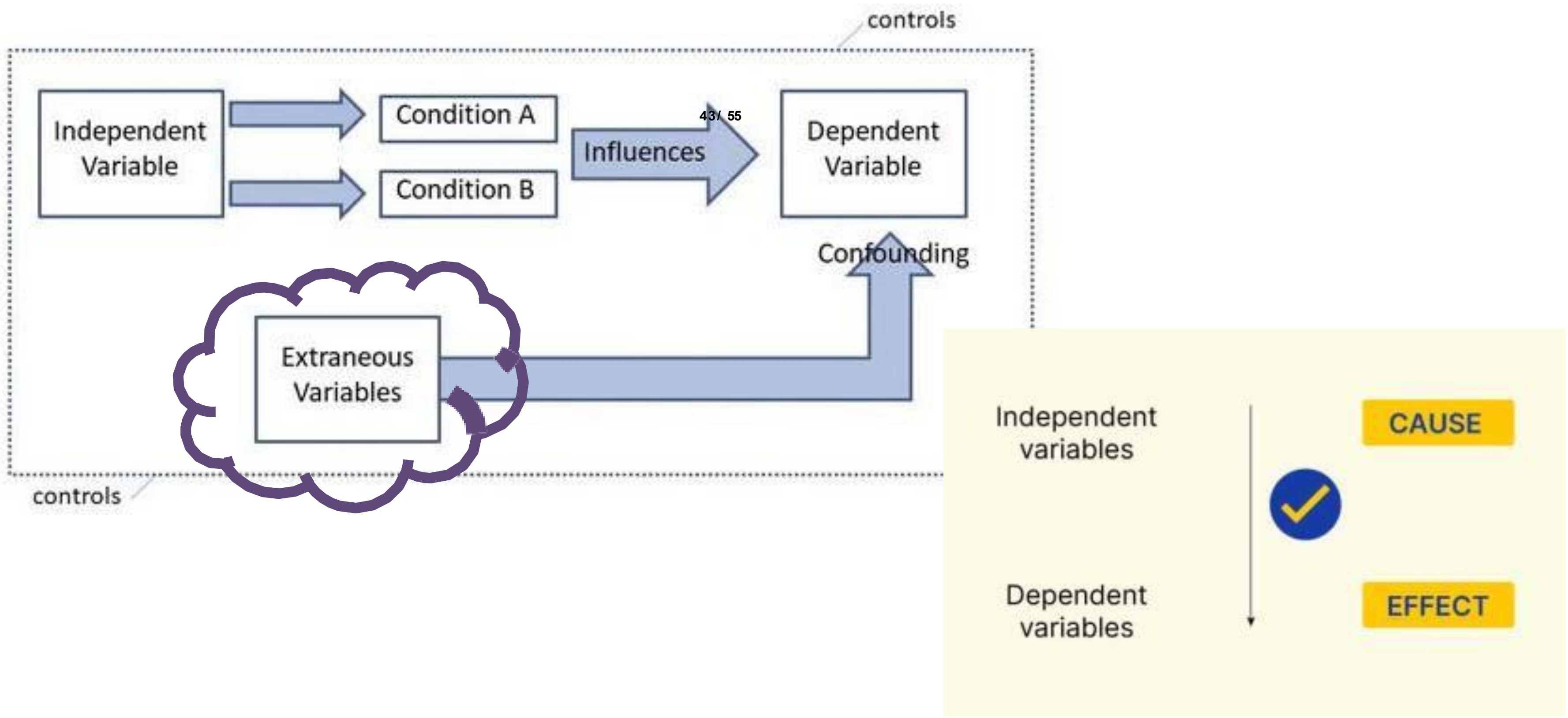
What you want

What you get

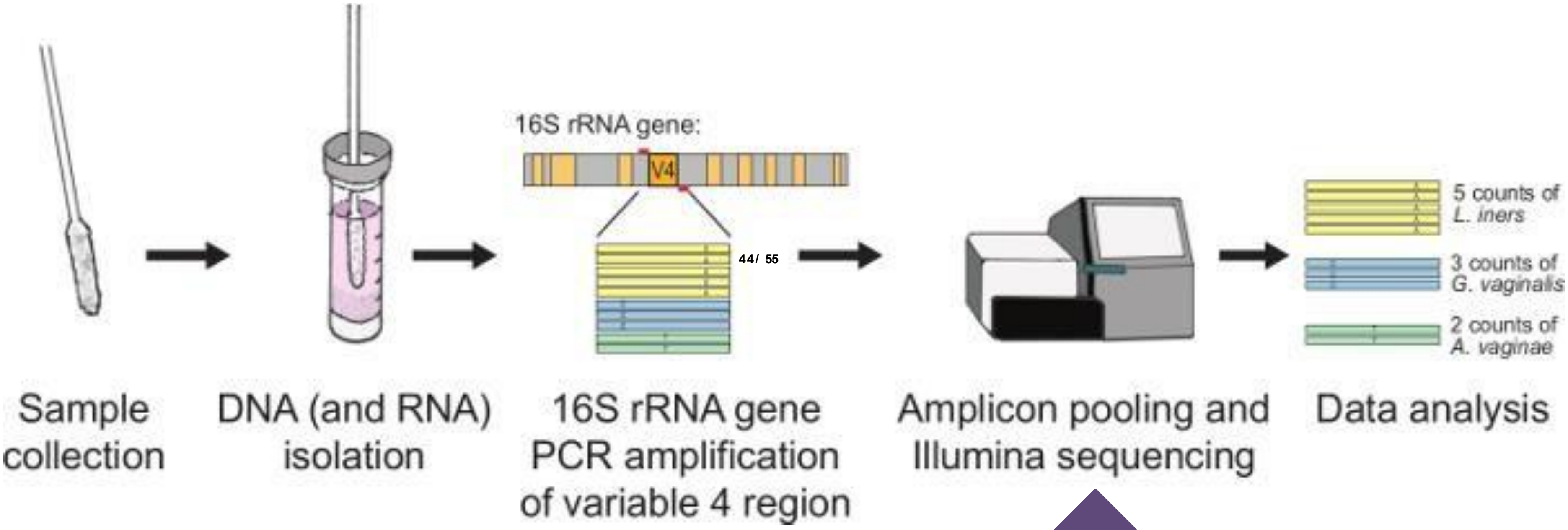
That spare comes in handy
Highly recommend especially
with mice!

What causes this?
Confounding variables

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



Examples of confounding variables



Sample collection

DNA (and RNA) isolation

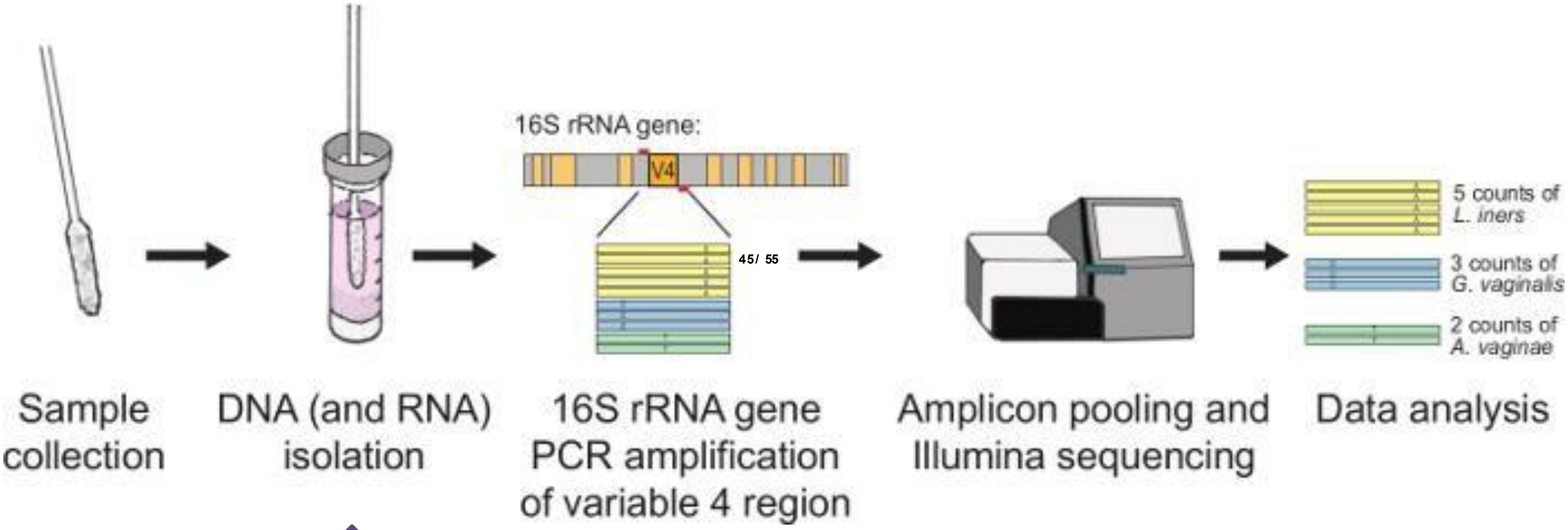
16S rRNA gene PCR amplification of variable 4 region

Amplicon pooling and Illumina sequencing

Data analysis

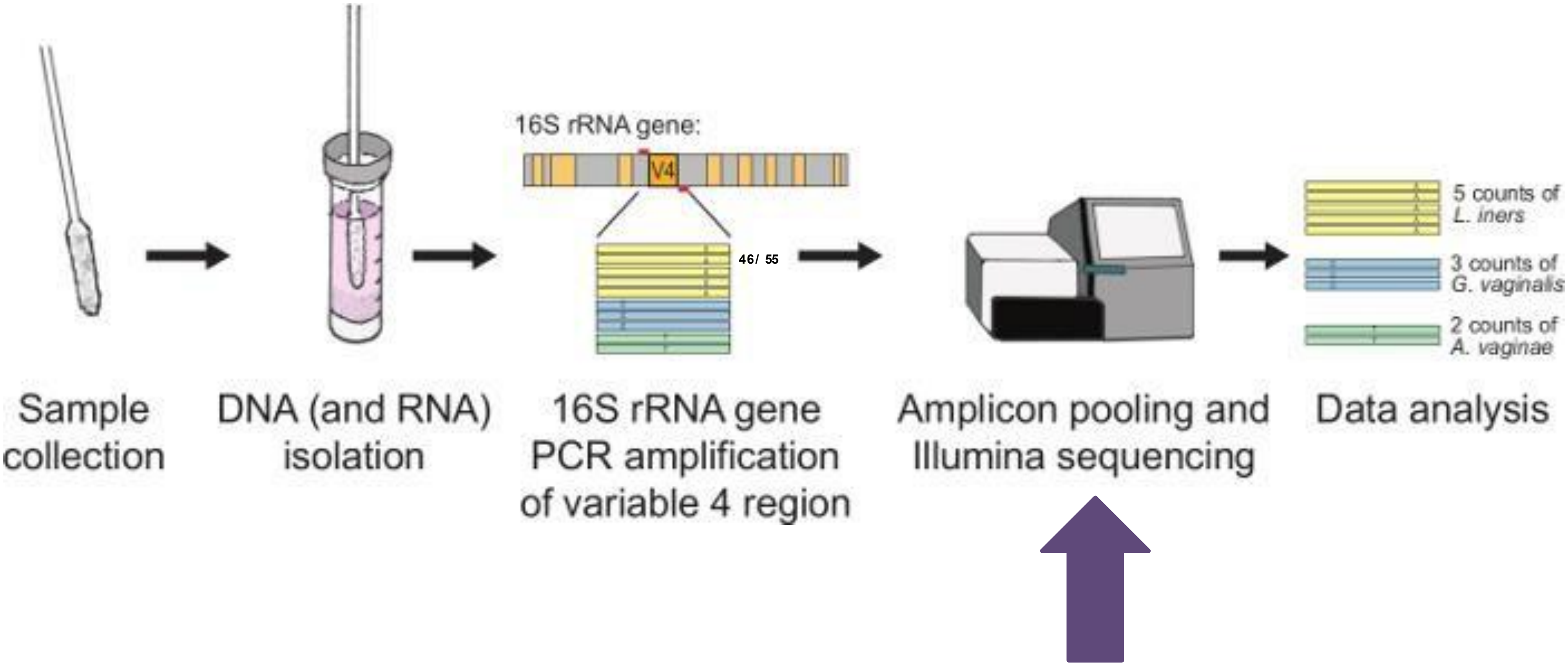
A new technician is running the sequencer

Examples of confounding variables



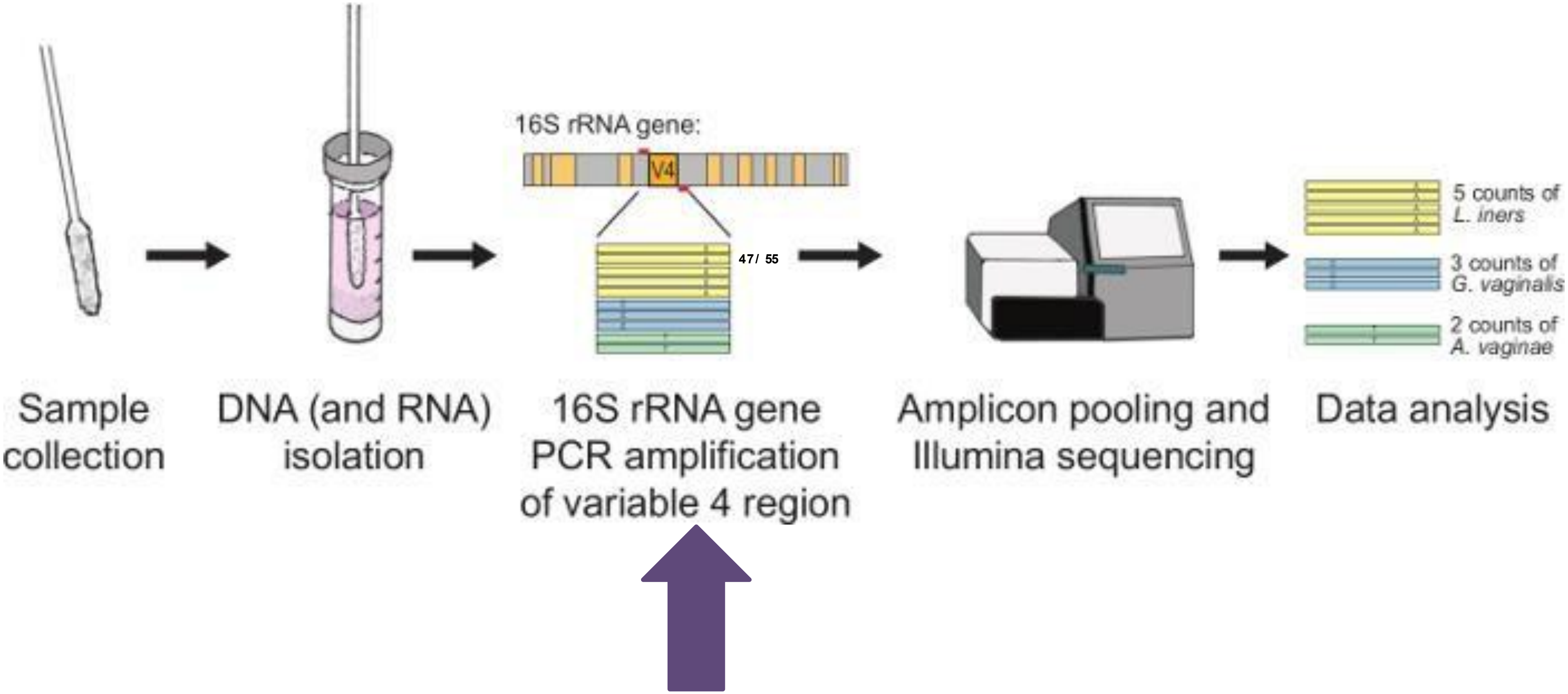
Extracting DNA/RNA with two different kits!

Examples of confounding variables



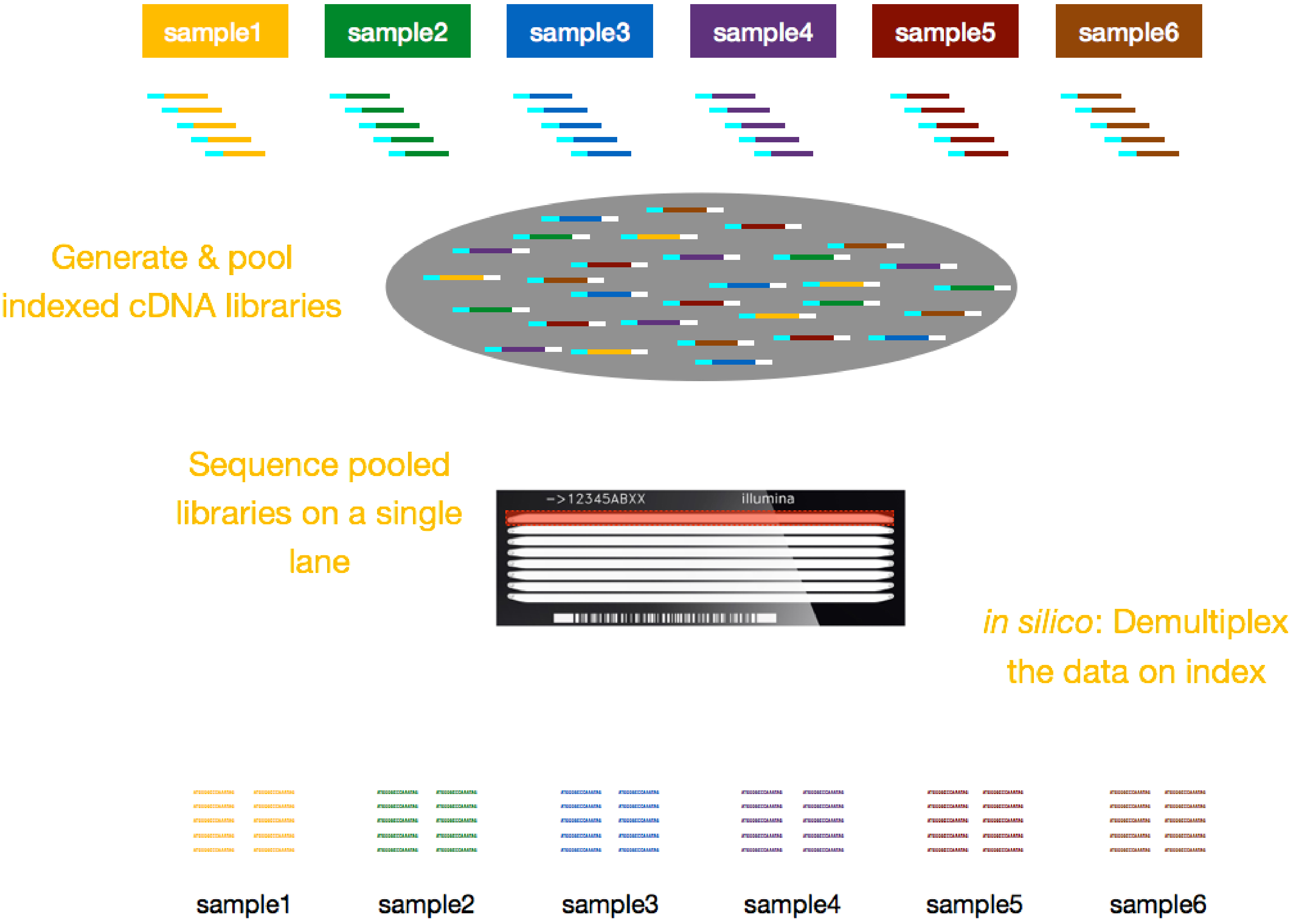
Sequencing on multiple different types of platforms

Examples of confounding variables



Inappropriate multiplexing strategy

Multiplexing

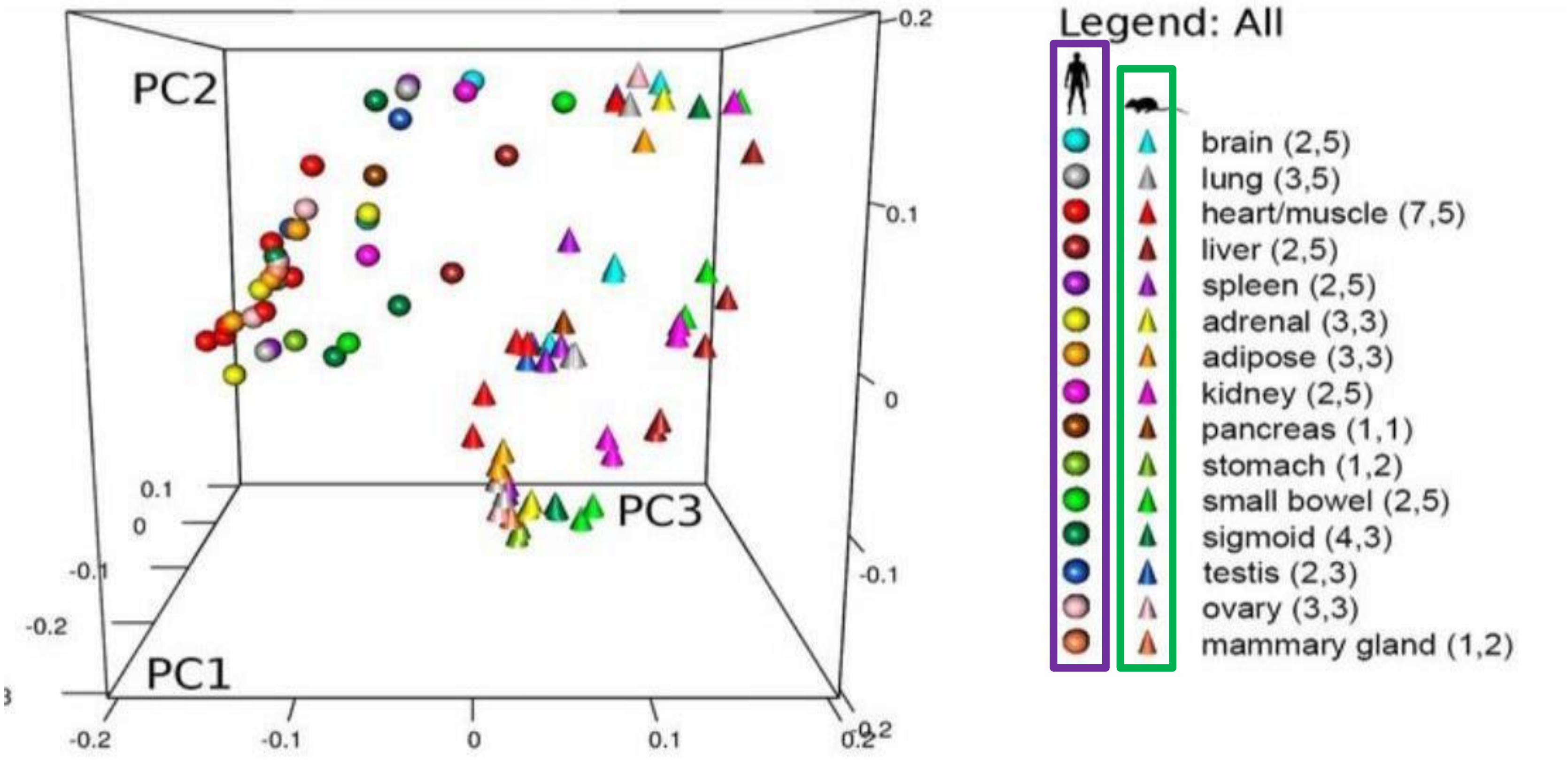


Generate & pool indexed cDNA libraries

Sequence pooled libraries on a single lane

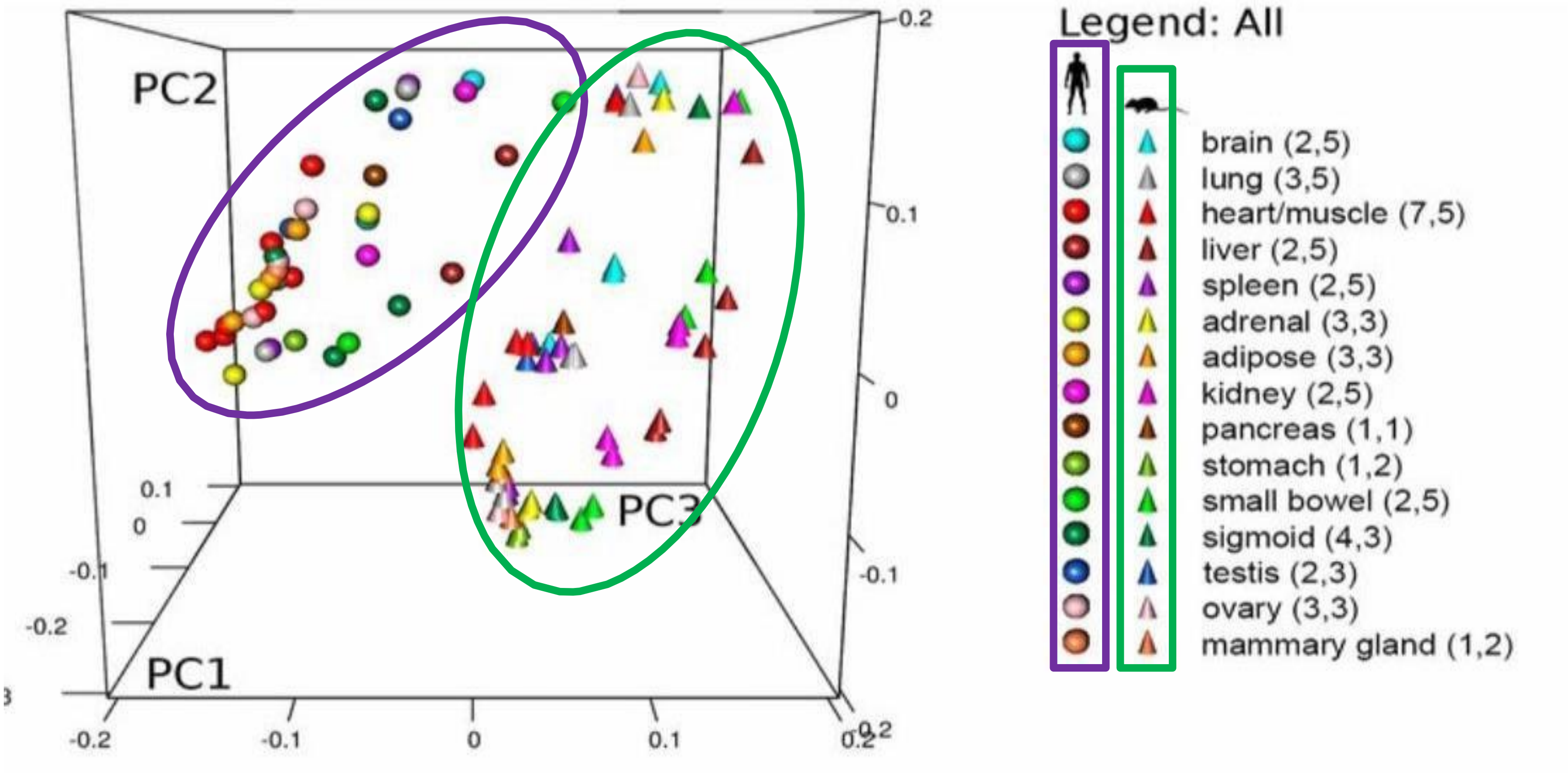
in silico: Demultiplex the data on index

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

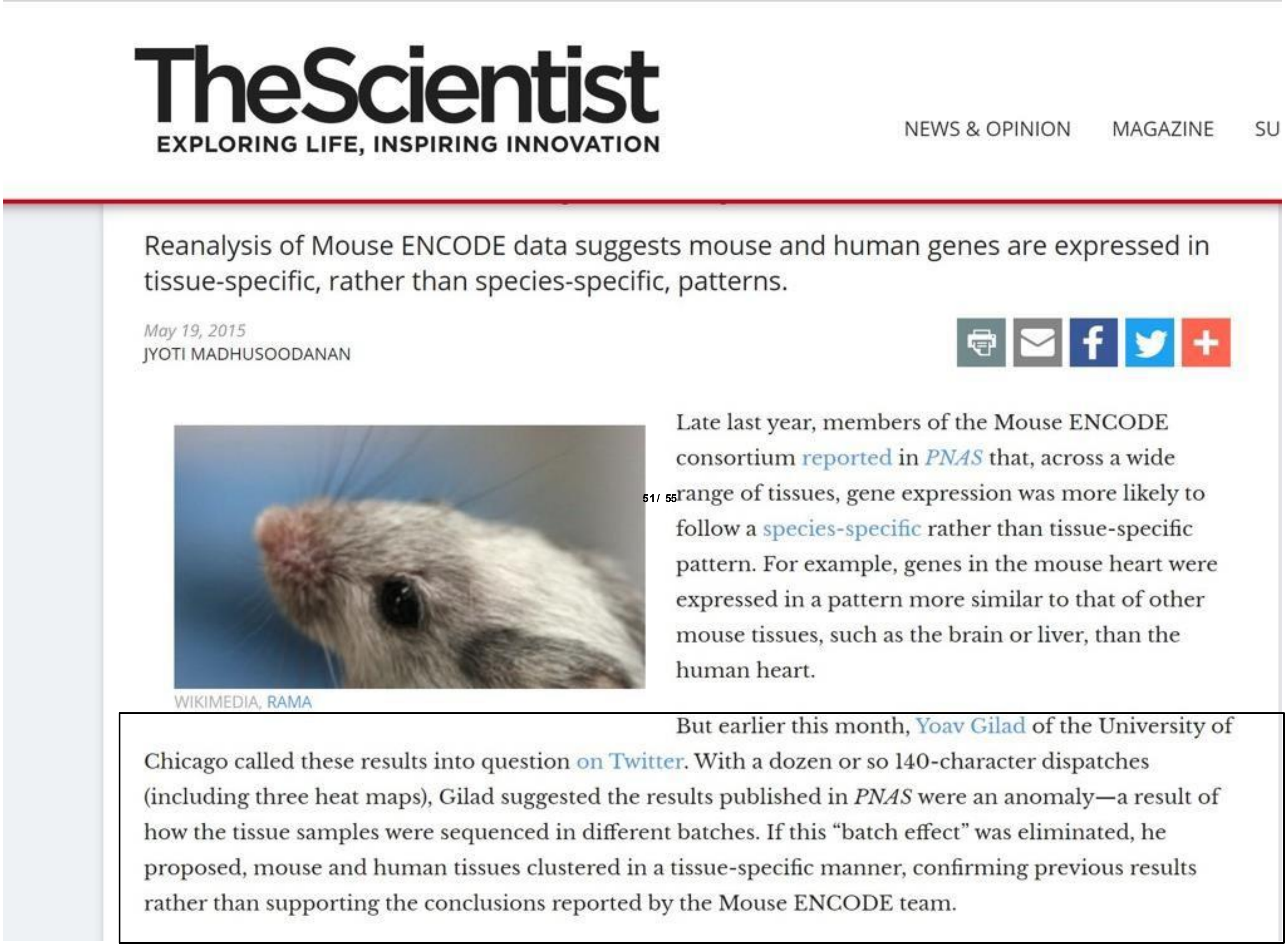


https://doi.org/10.1 073/p nas.141362 4111

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



https://doi.org/10.1 073/p nas.141362 4111



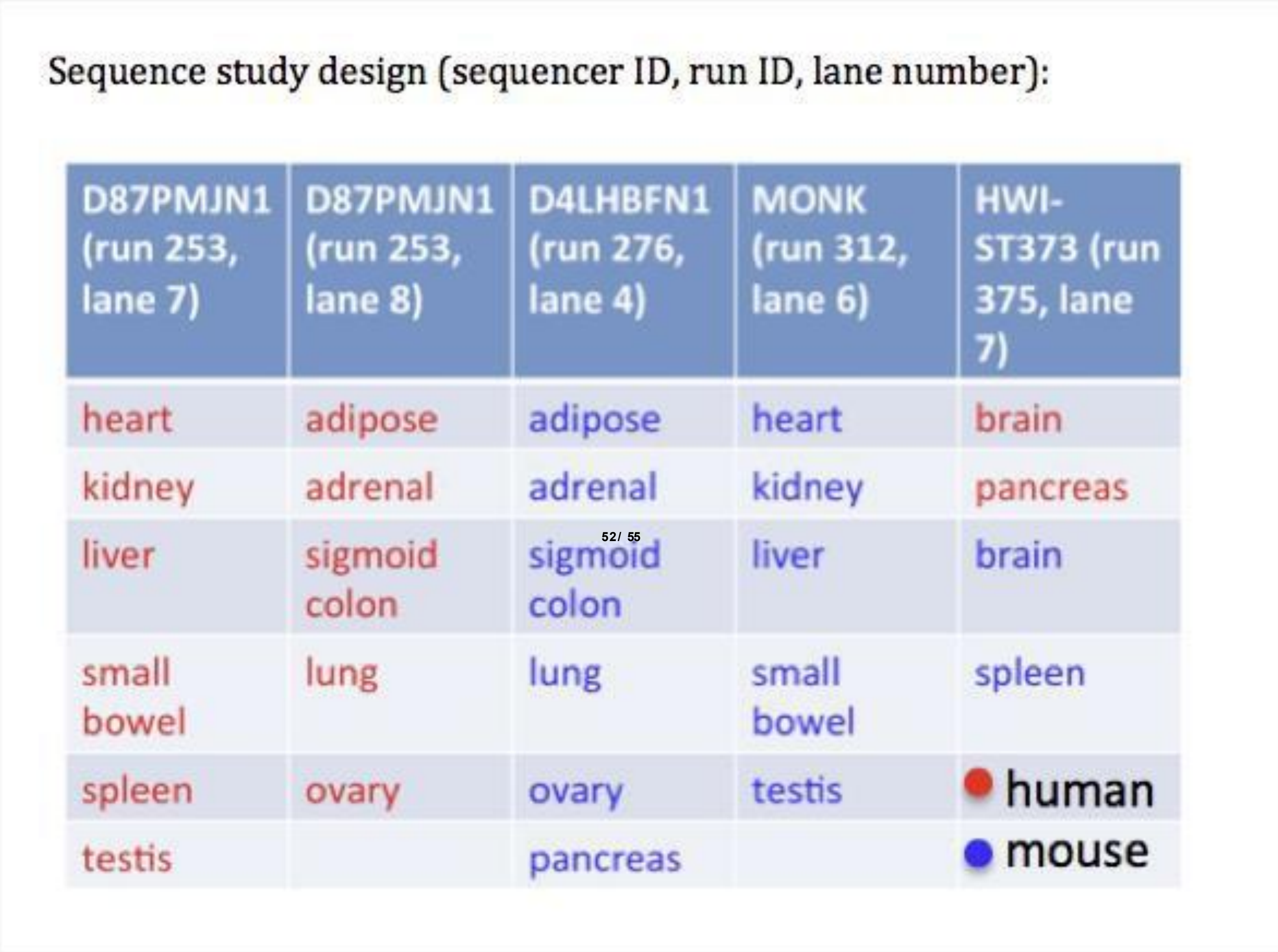
Reanalysis of Mouse ENCODE data suggests mouse and human genes are expressed in tissue-specific, rather than species-specific, patterns.

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.

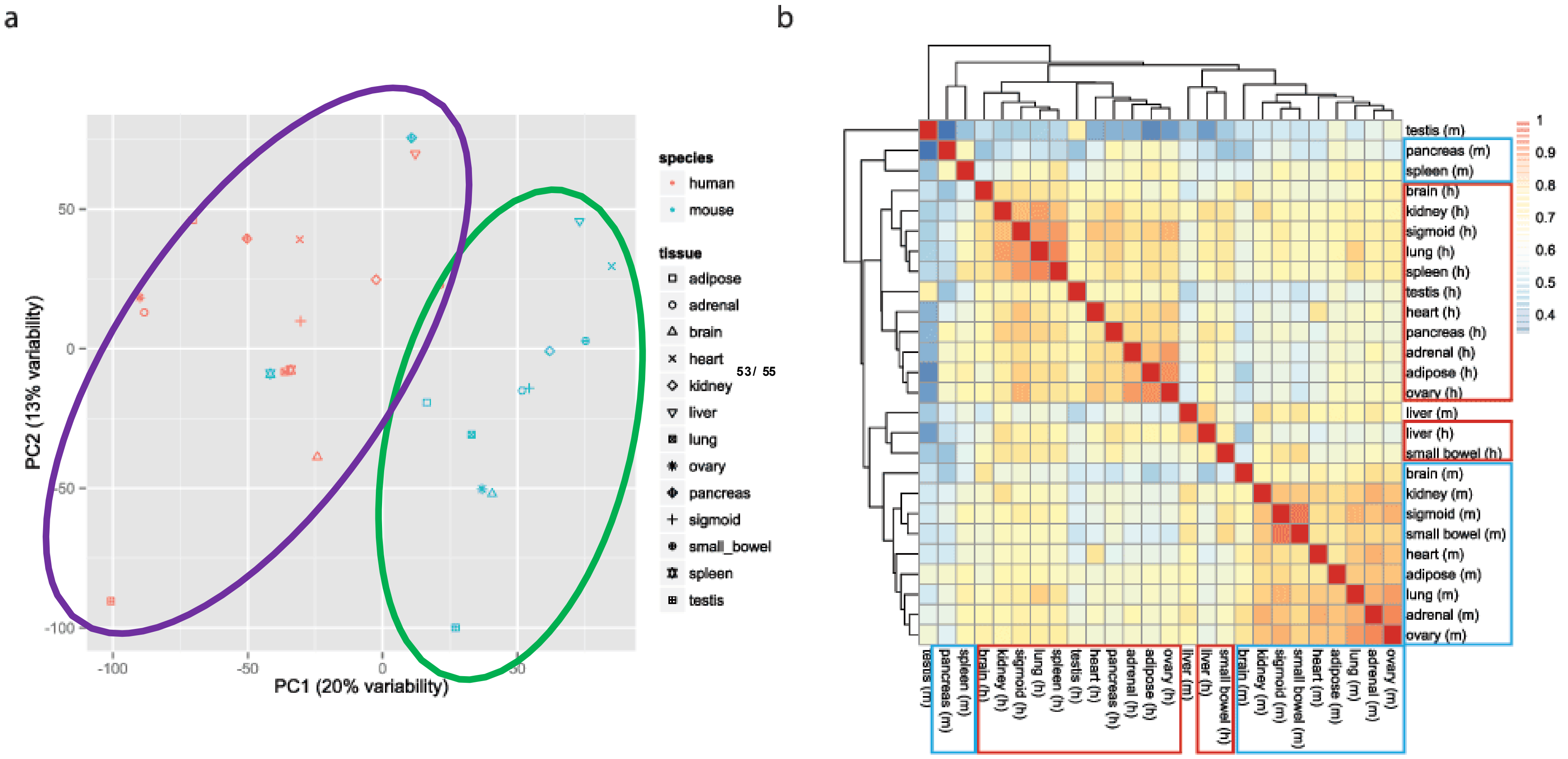


Sequence study design (sequencer ID, run ID, lane number):

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

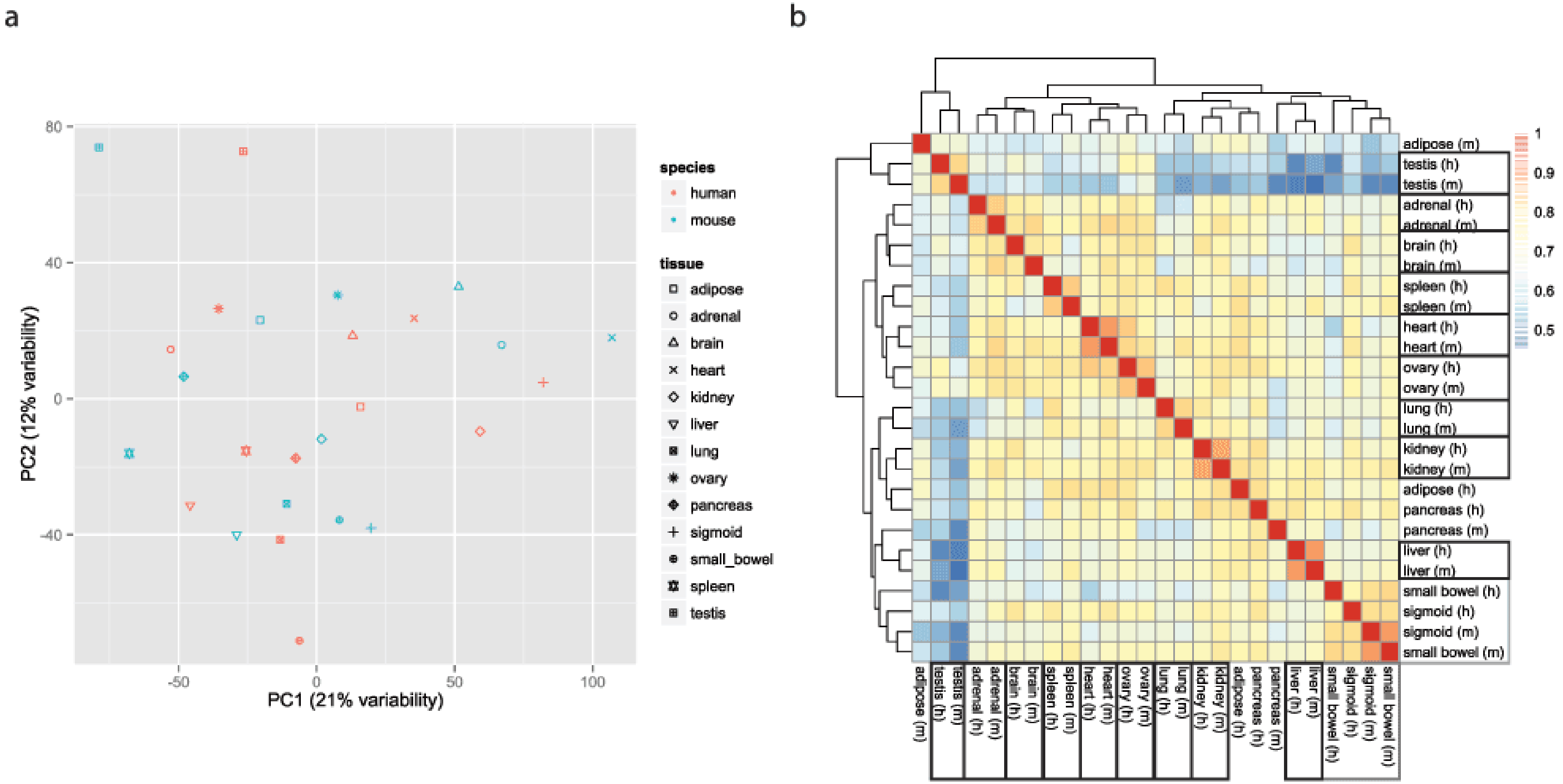
Sequencing lane (a batch effect) was almost completely confounded with species in the PNAS study. From @Y_Gilad

Before accounting for batch effect



Samples grouped by animal

After accounting for batch effect



Samples now grouped by tissue!

What does this all means?

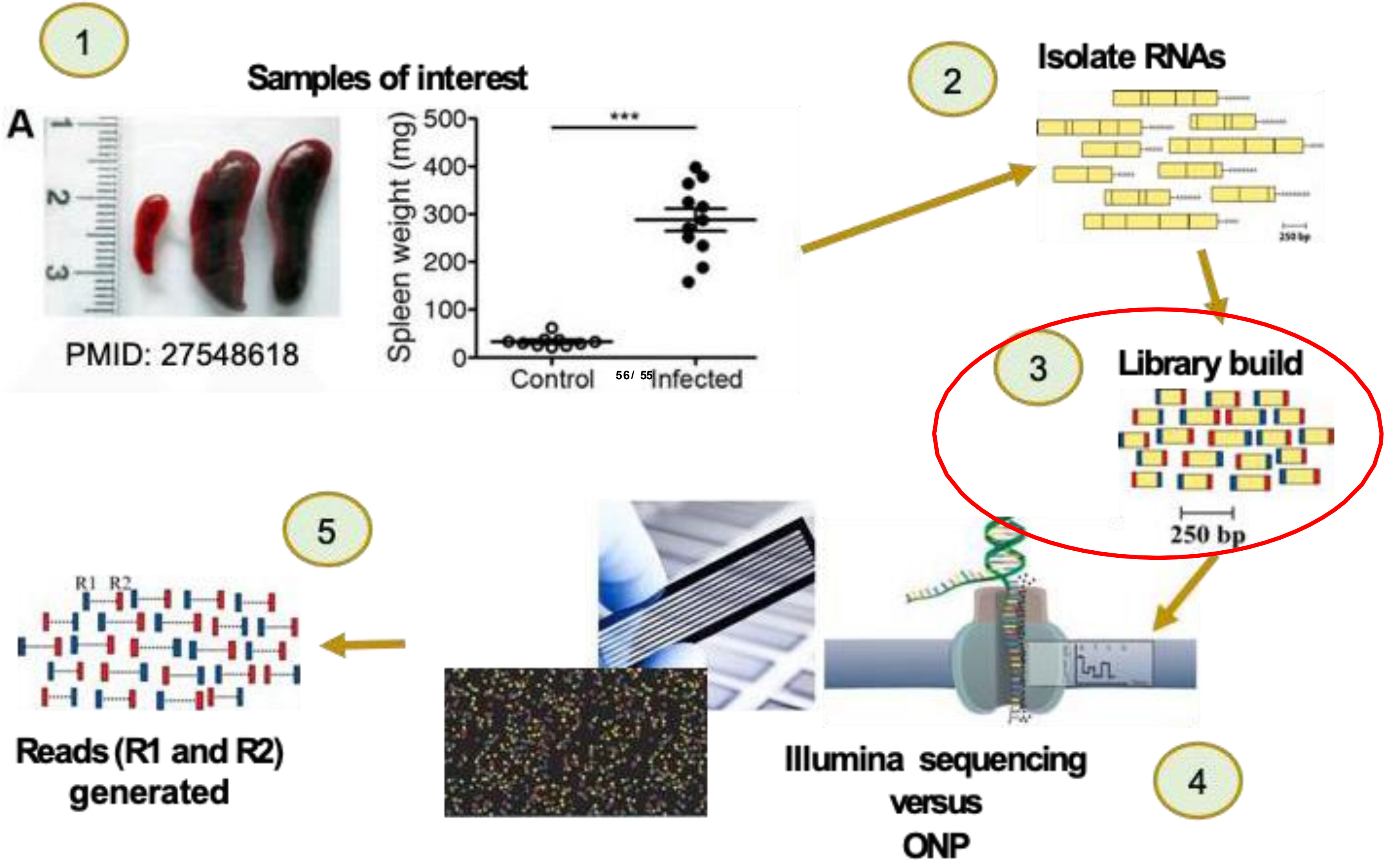
- Its sometimes impossible for bioinformaticians to partition biological variation from technical variation, when these two sources of variation ***are confounded***.

- 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. Its up to you to recognize this!

Experimental workflow



1 Samples of interest

PMID: 27548618

2 Isolate RNAs

3 Library build

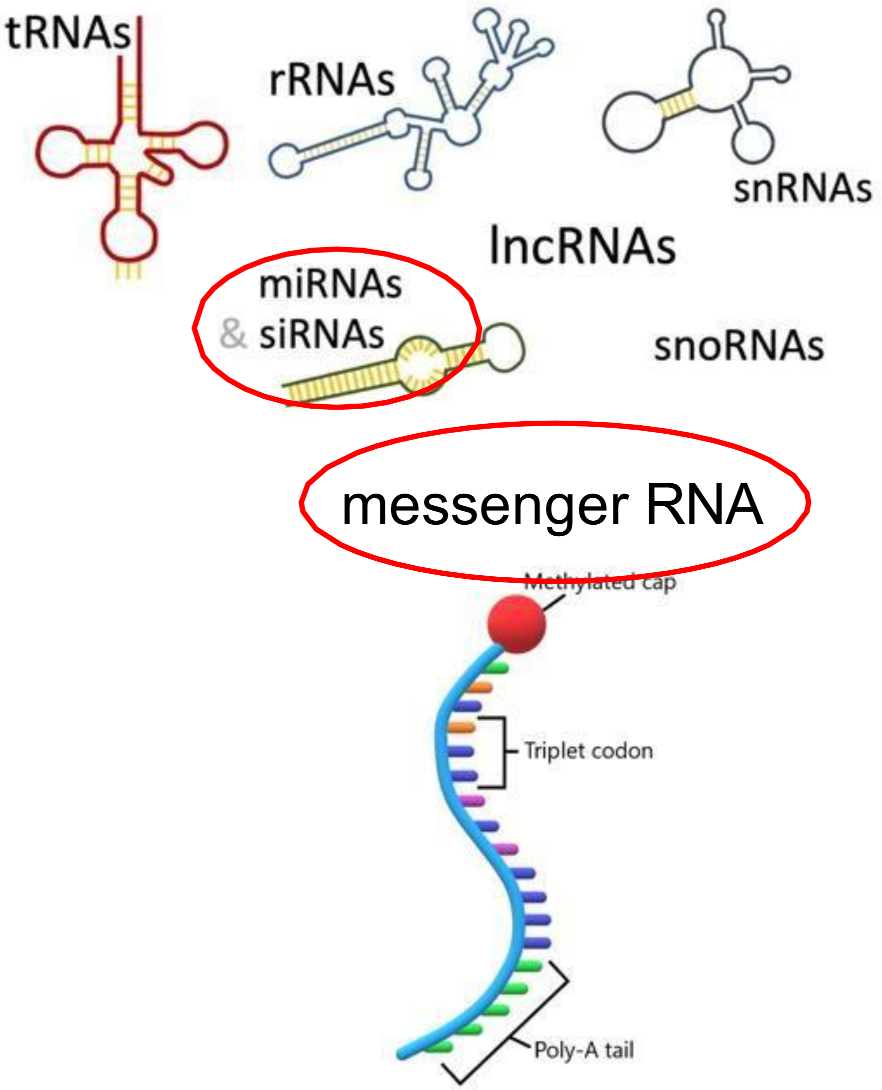
4 Illumina sequencing versus ONP

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

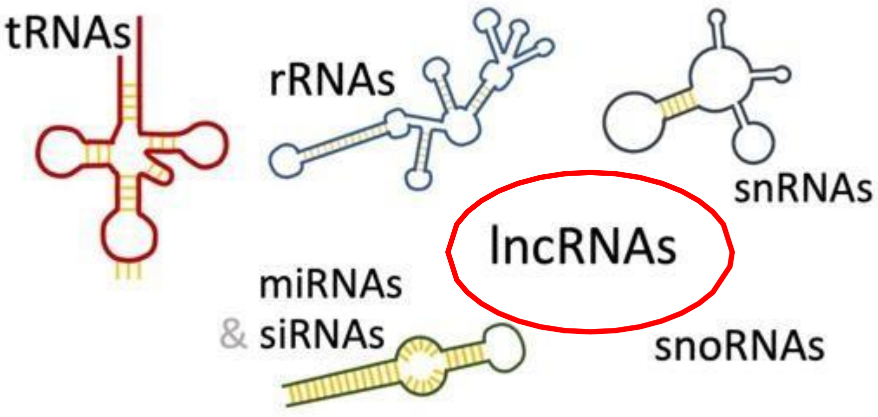


- Kukurba, Mongomery (2015) Cold Spring Harb. Protoc. Gaida SM et al. (2015) Nat Comm 6:7045

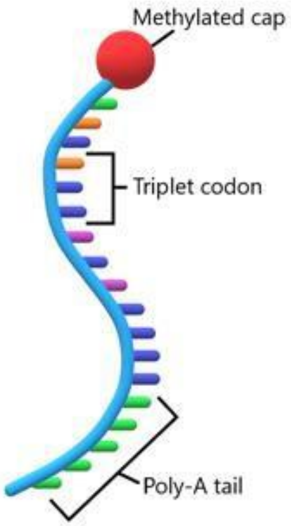
RNA composition

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

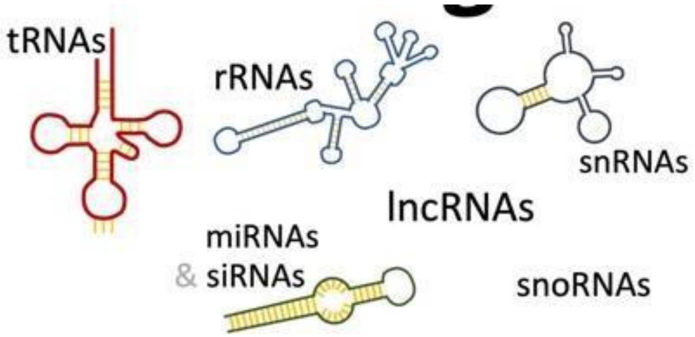


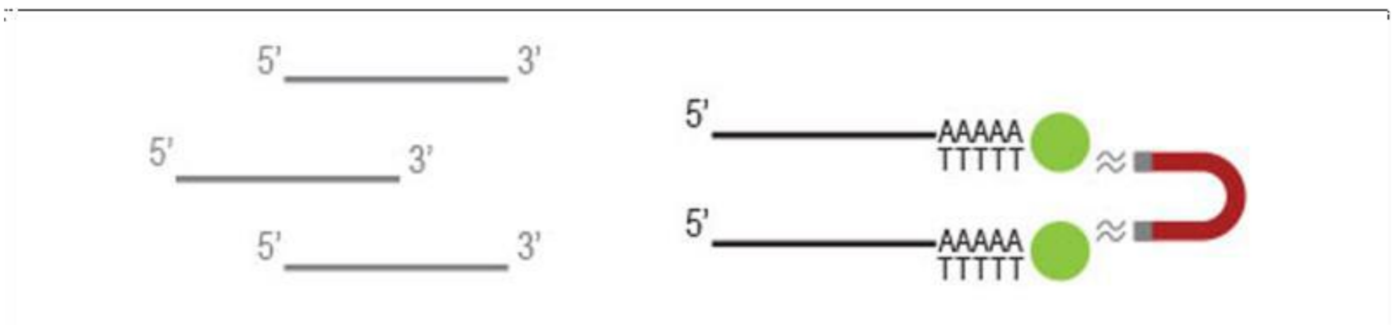
messenger RNA



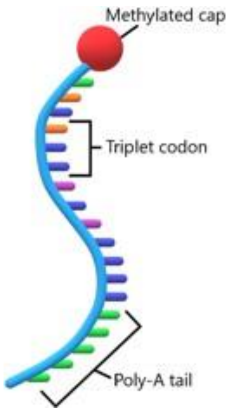
- Kukurba, Mongomery (2015) Cold Spring Harb. Protoc.
Gaida SM et al. (2015) Nat Comm 6:7045

The RNA sample undergoes either selection of the mRNA (polyA selection) or depletion of the rRNA. The resulting RNA is fragmented.



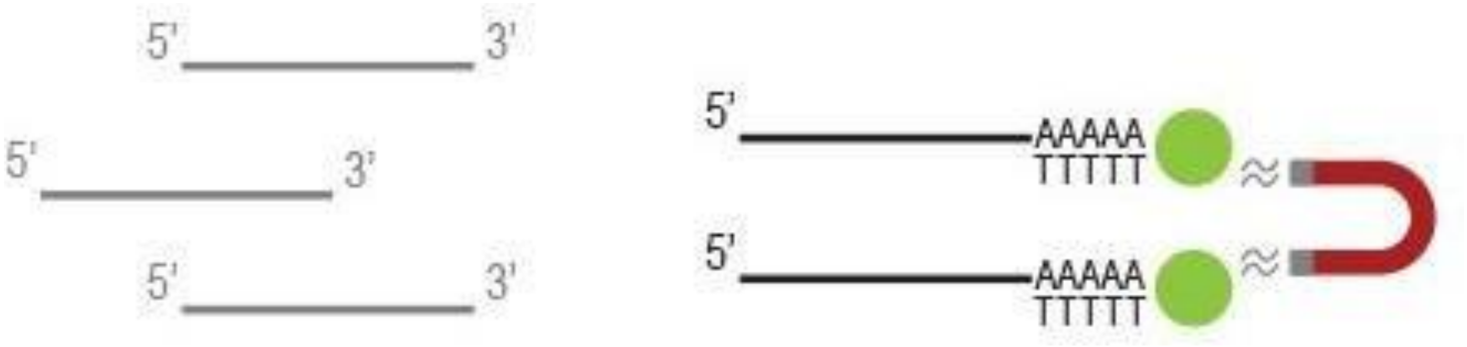


messenger RNA

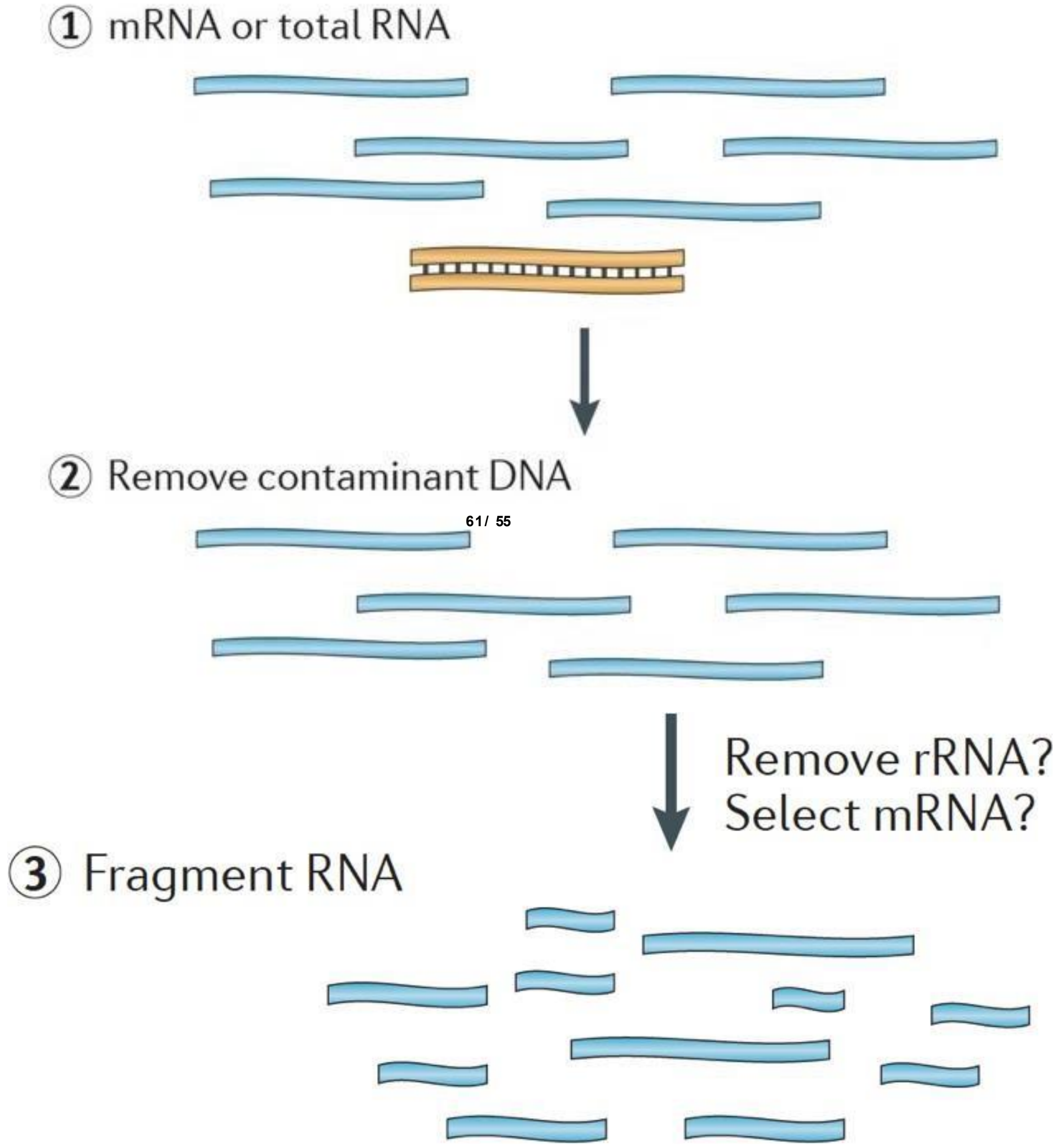


Poly-A versus rRNA depletion?

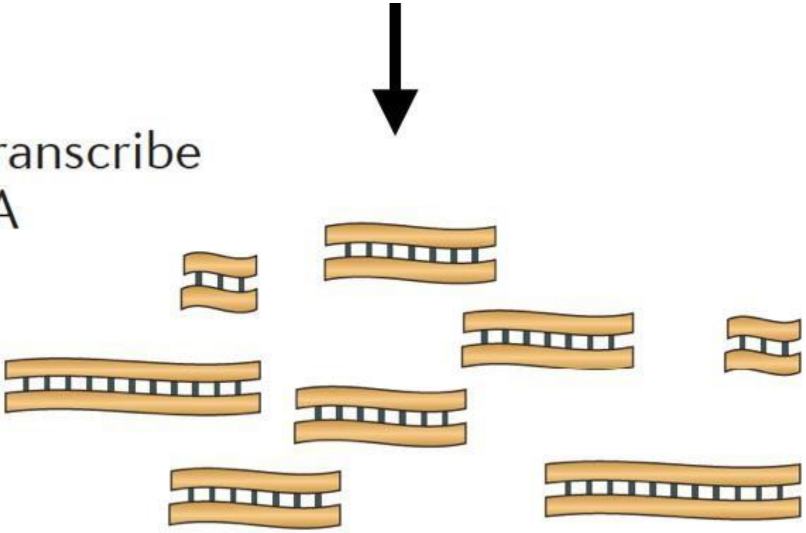
- If you are aiming to obtain information about long non-coding RNA's I recommend performing ribosomal RNA depletion
- Bacterial mRNAs are also not poly-adenylated



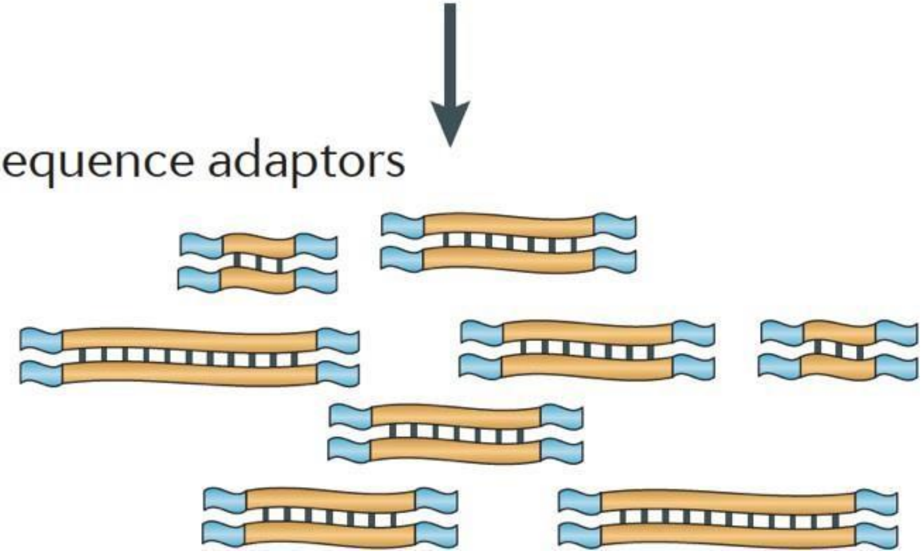
Illumina Library preparation



④ Reverse transcribe into cDNA



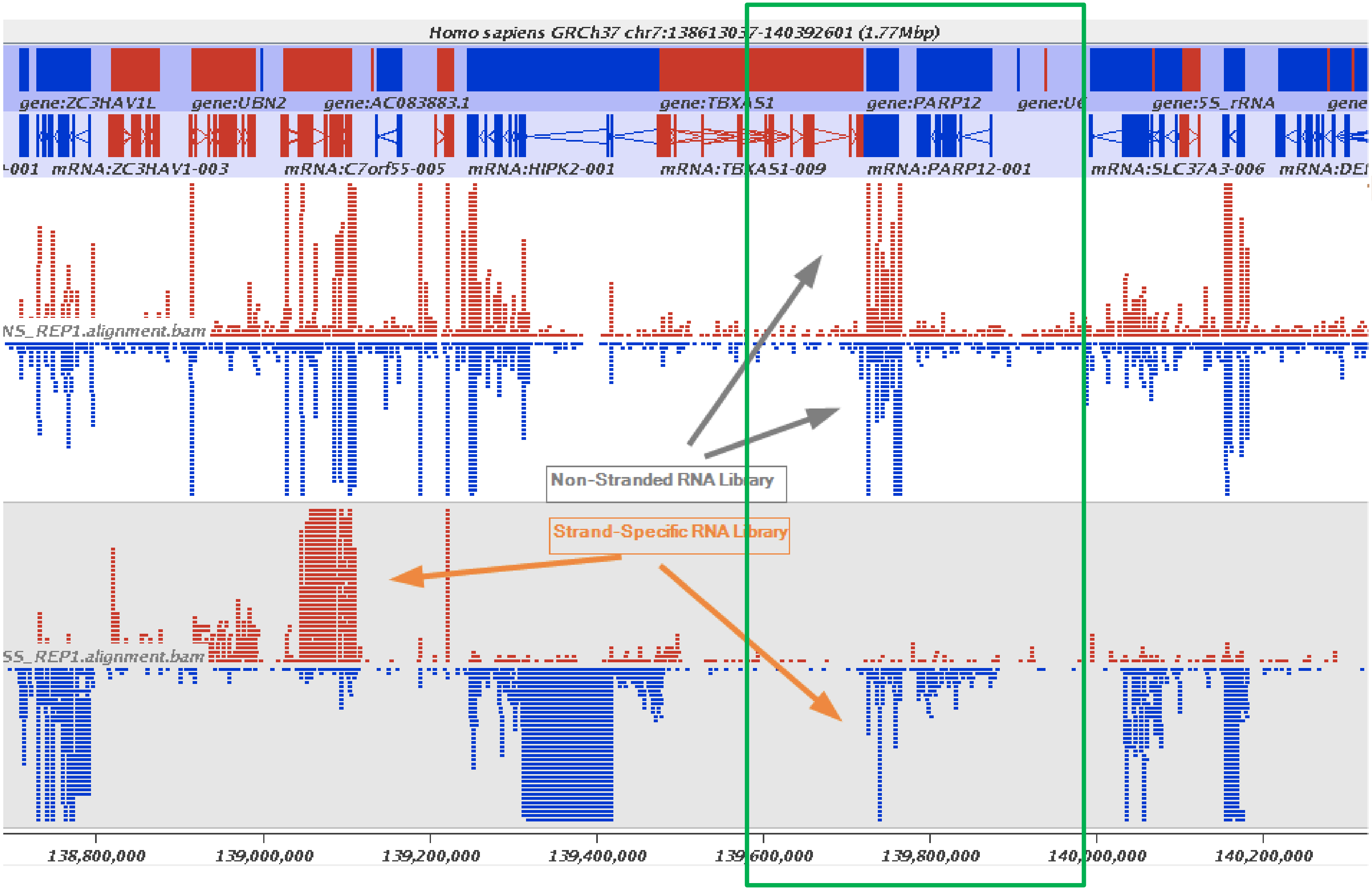
⑤ Ligate sequence adaptors



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



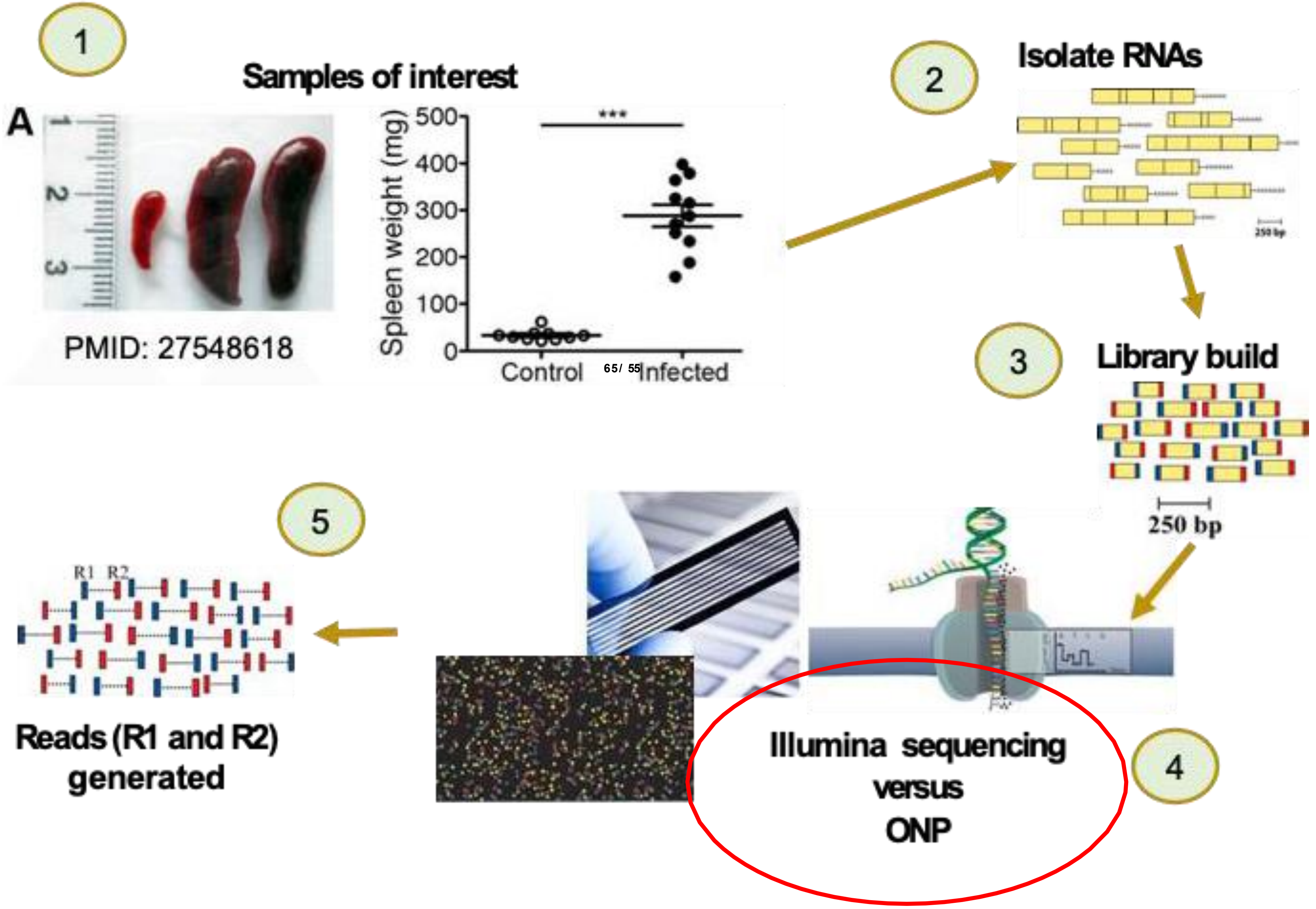
Non-Stranded RNA Library

Strand-Specific RNA Library

ly see

Red = + strand **Blue = - strand**

Experimental workflow



1

Samples of interest

PMID: 27548618

2

Isolate RNAs

3

Library build

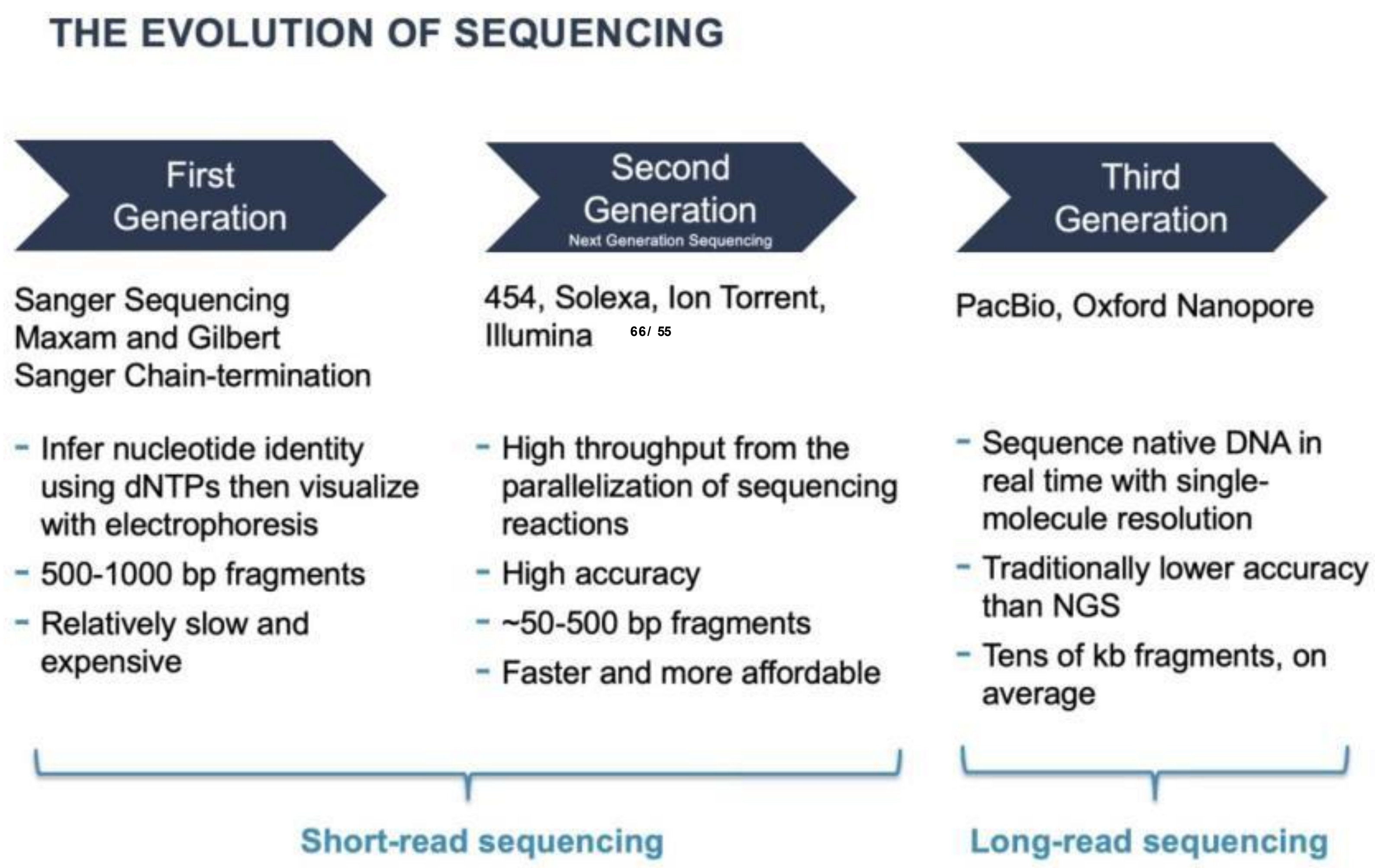
4

Illumina sequencing
versus
ONP

5

Reads (R1 and R2)
generated

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

The bioinformatic pipeline for these are different!

Single-end versus Paired-end

After preparation of the libraries, sequencing can be performed to generate the nucleotide sequences of the ends of the fragments, which are called **reads**. You will have the choice of sequencing a single end of the cDNA fragments (single-end reads) or both ends of the fragments (paired-end reads).

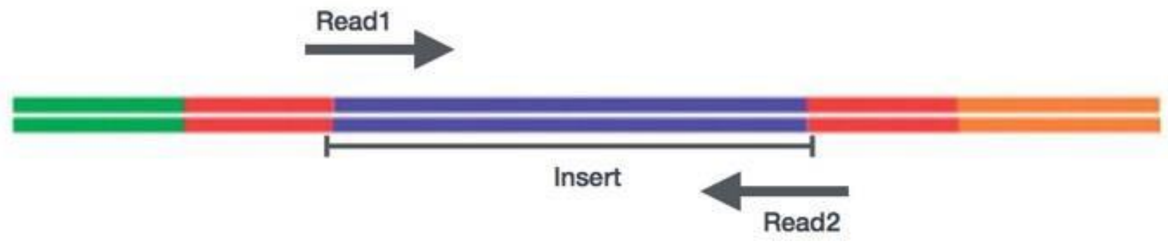
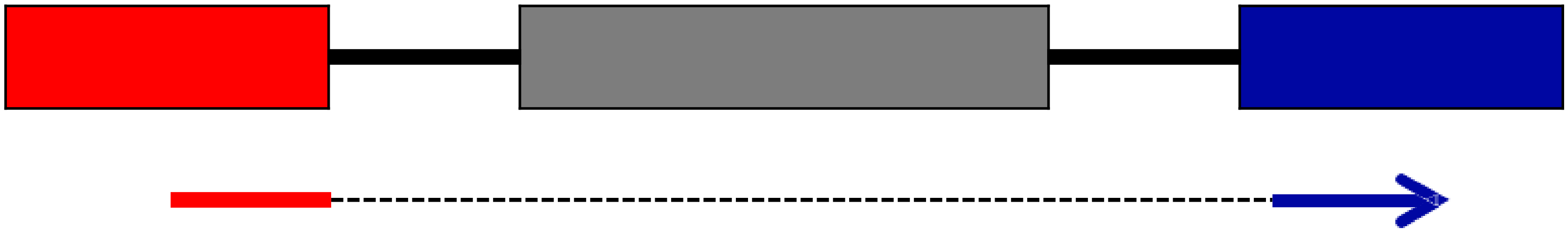


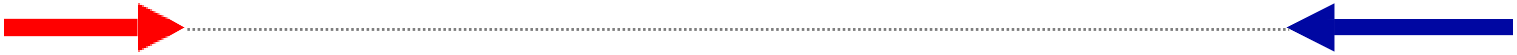
Figure 10: Paired End Reads

- SE => Only Read1 => one FASTQ file/sample
- PE => Read1 + Read2 => **two FASTQ files/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

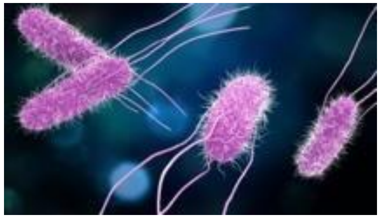
  Knowing both ends of a fragment and an approximation of fragment size helps to determine the transcript from which it was derived.

In Summary, to quantify Differential Gene Expression

- Technology: Illumina
- Read length: 50bp to 300 bp
- Paired vs single end: *doesn't matter but important to note*
- Number of reads: > 15 million per sample
- Replicates: 3 biological replicates *minimum*

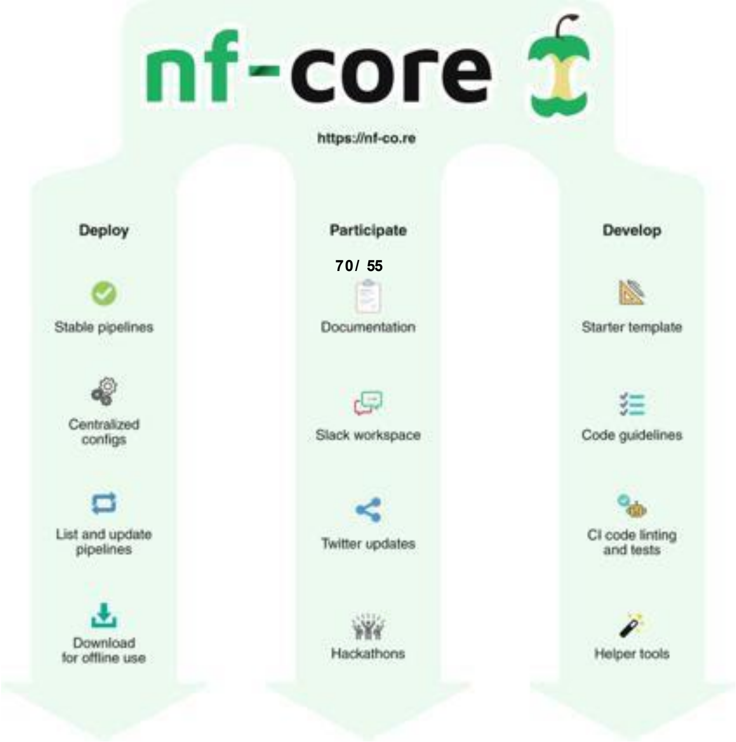
A well-planned experiment goes a long way!

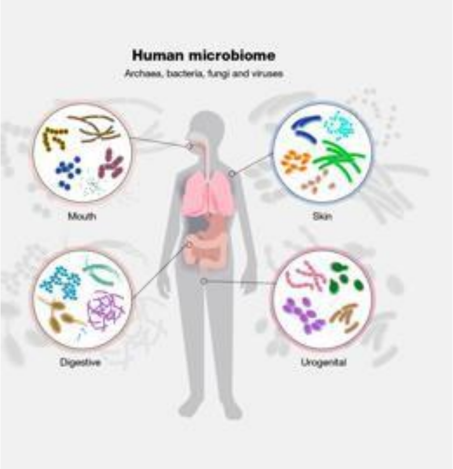
Final projects from the years have spanned the following topics:



Salmonella enterica

Applications of organoids as research models

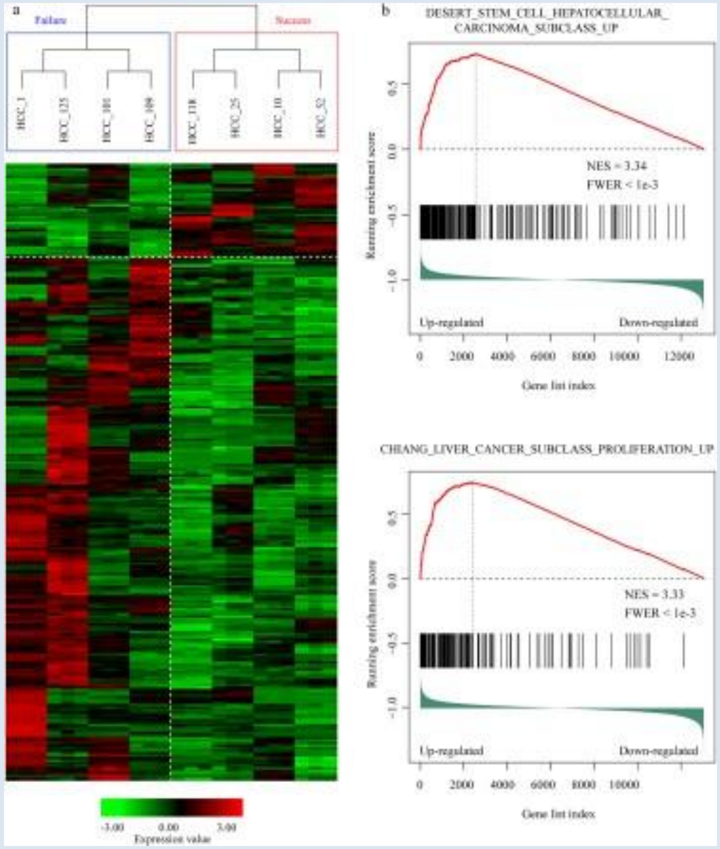




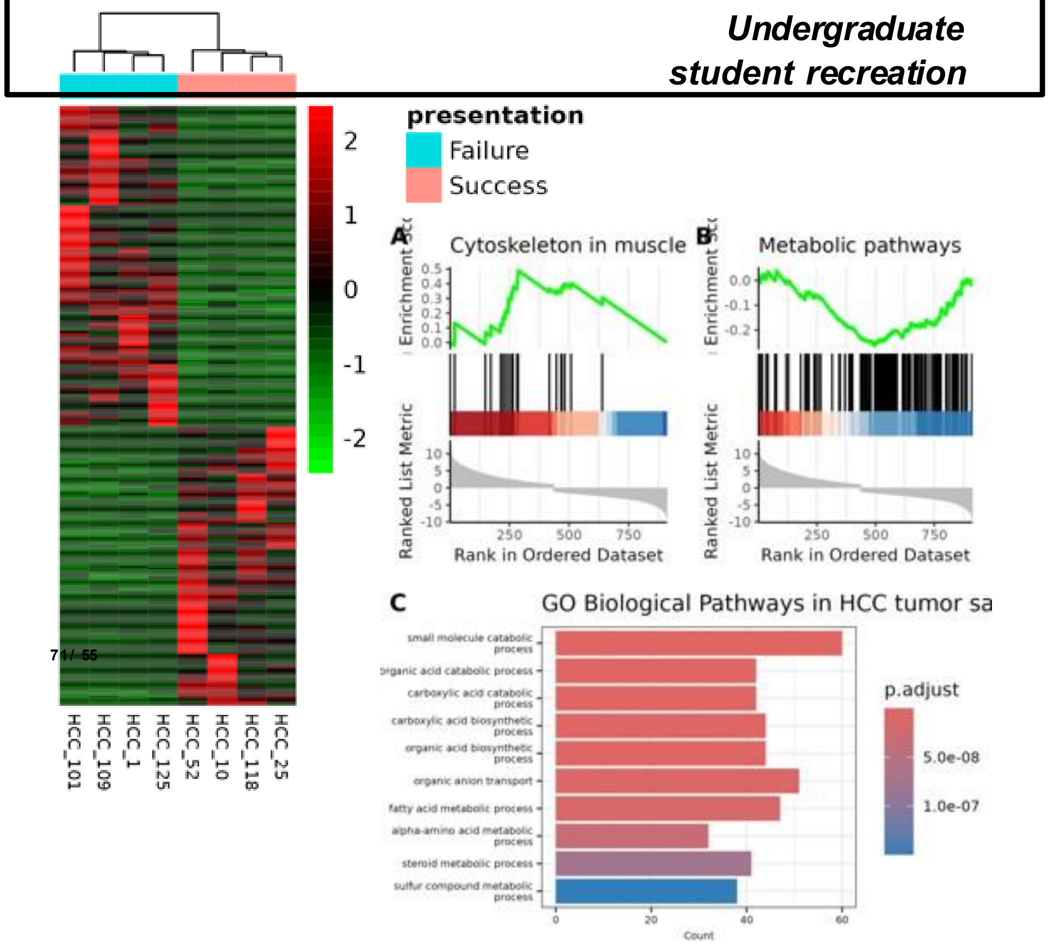
Dengue

Breast cancer

And more….!



Original Published Work



Undergraduate student recreation

Green Trail

UG credentials:

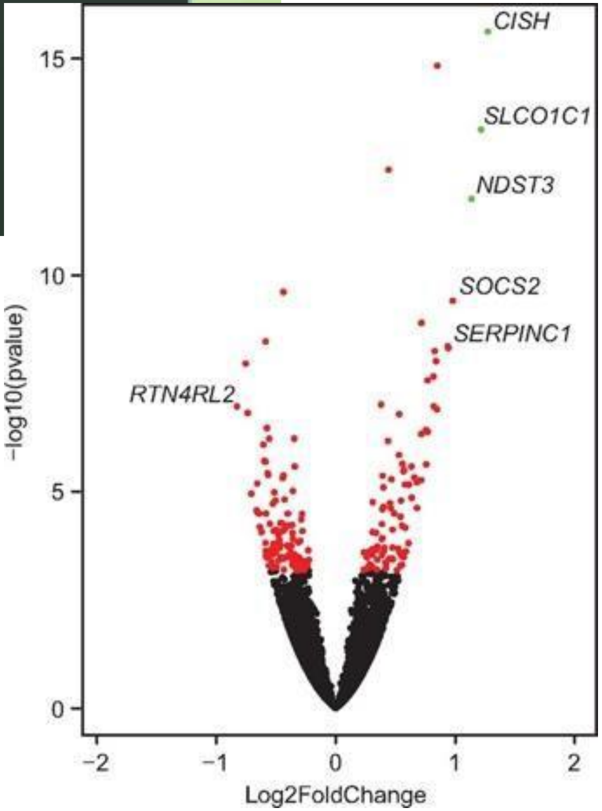
1-semester of intro bioinformatics

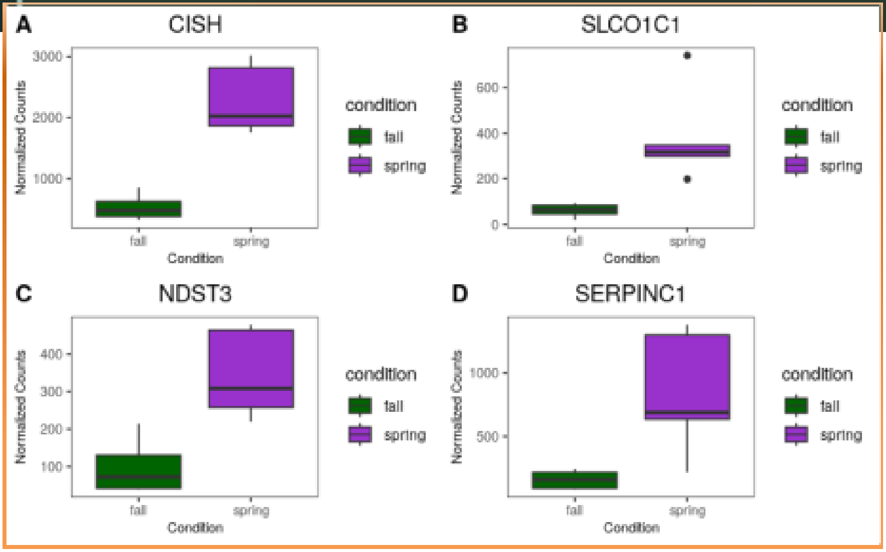
Research Question

Due to the strong relationship between the kidney and the heart, which differentially expressed genes in bear kidneys are related to cardiac pathways?

72/ 55

Black Trail
UG credentials:
1-semester of intro bioinformatics





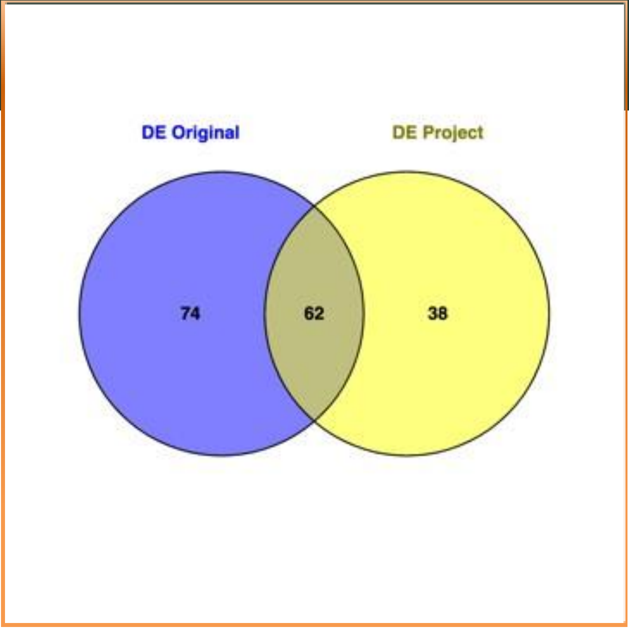
Black Trail

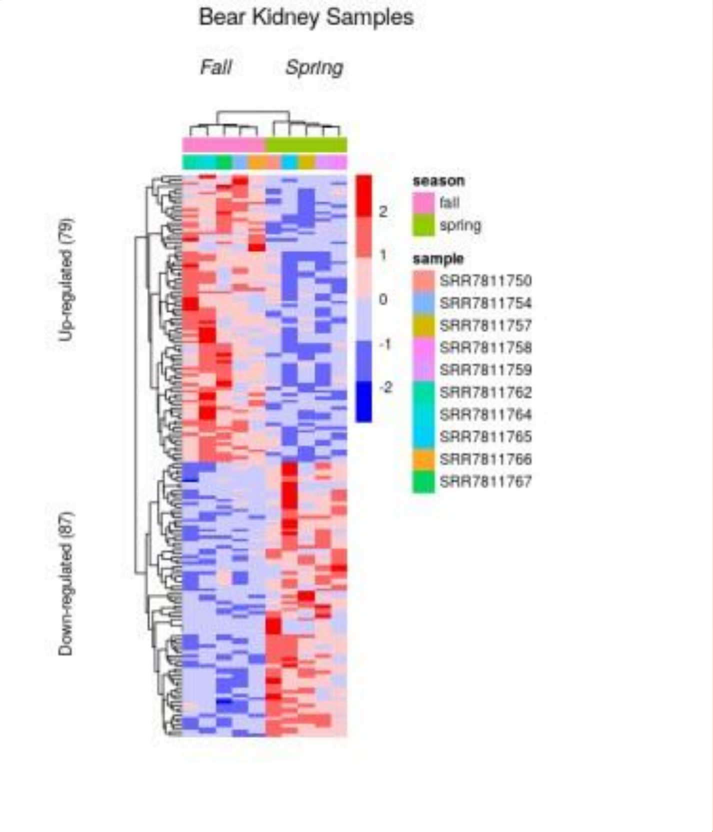
UG credentials:

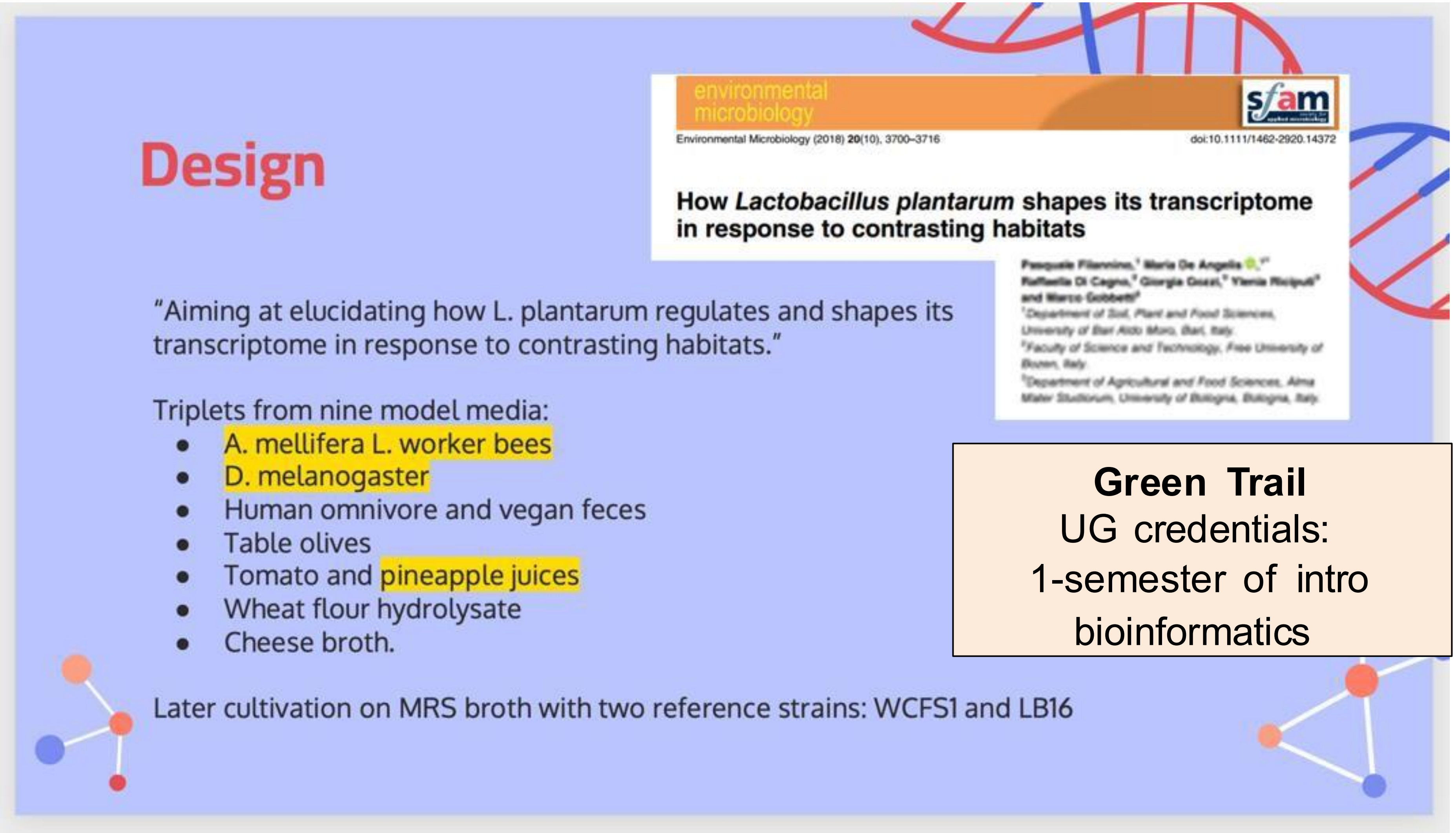
1-semester of intro bioinformatics

Research Question

Due to the strong relationship between the kidney and the heart, which differentially expressed genes in bear kidneys are related to cardiac pathways?







Design

environmental microbiology

sfam

Environmental Microbiology (2018) 20(10), 3700–3716 doi:10.1111/1462-2920.14372

How *Lactobacillus plantarum* shapes its transcriptome in response to contrasting habitats

Pasquale Filannino,[1] Maria De Angelis,[1*] Raffaella Di Cagno,[2] Giorgia Gozzi,[3] Ylenia Riciputi[3] and Marco Gobbetti[2]

[1]*Department of Soil, Plant and Food Sciences, University of Bari Aldo Moro, Bari, Italy.*
[2]*Faculty of Science and Technology, Free University of Bozen, Italy.*
[3]*Department of Agricultural and Food Sciences, Alma Mater Studiorum, University of Bologna, Bologna, Italy.*

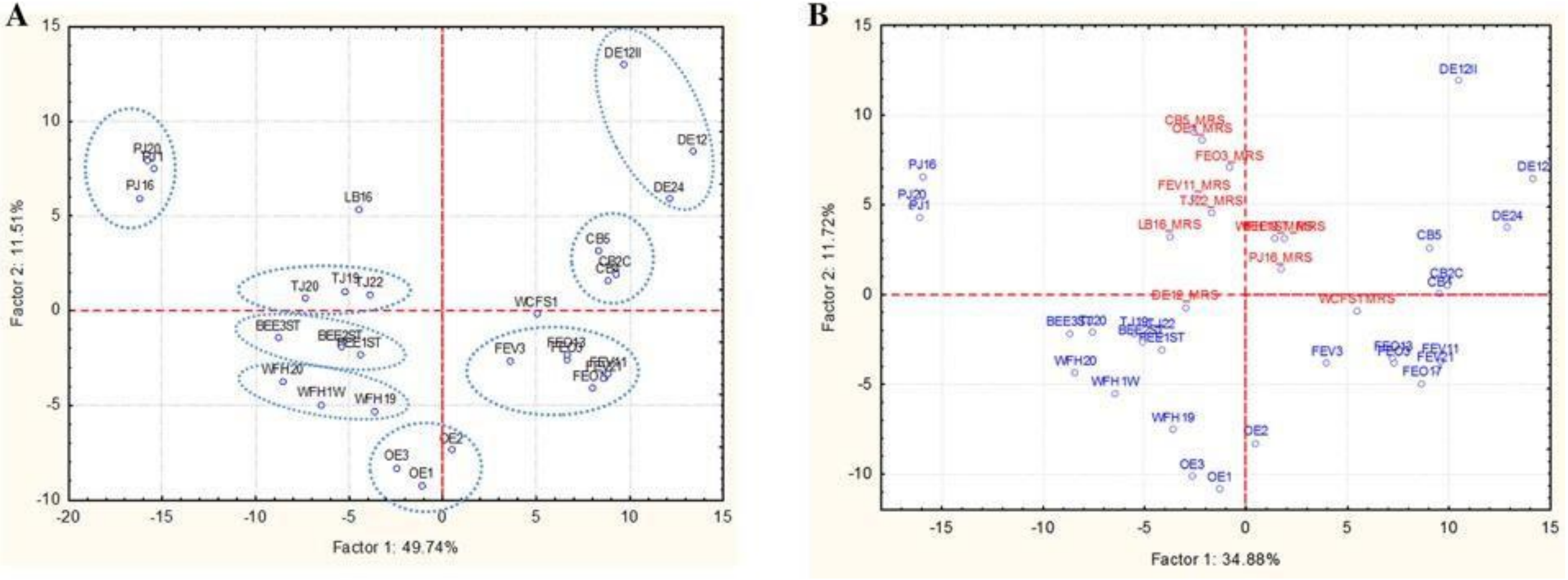
"Aiming at elucidating how L. plantarum regulates and shapes its transcriptome in response to contrasting habitats."

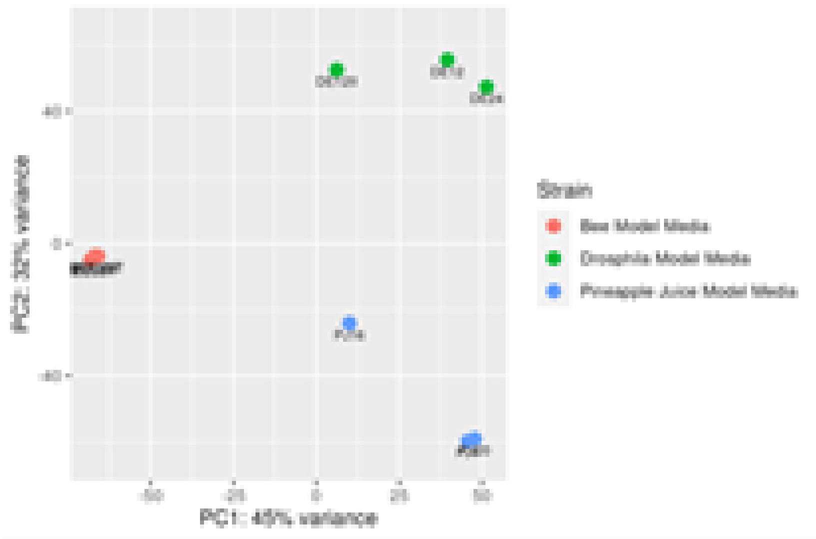
Triplets from nine model media:

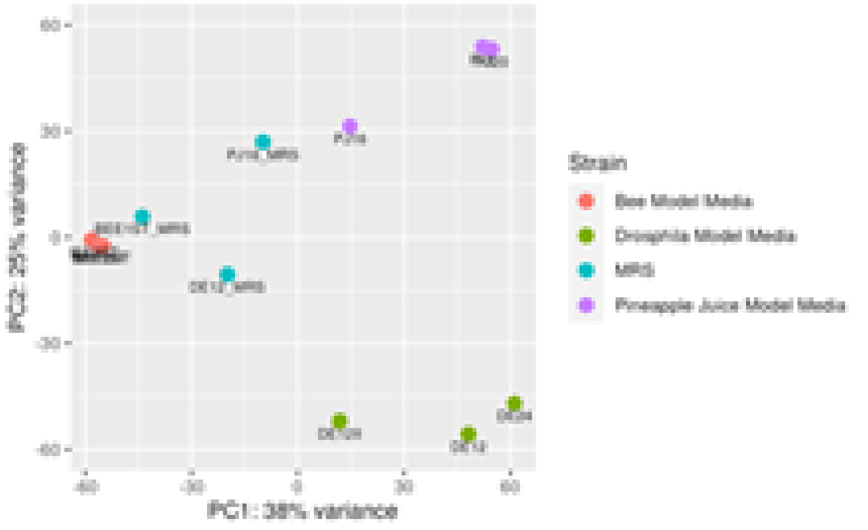
- A. mellifera L. worker bees
- D. melanogaster
- Human omnivore and vegan feces
- Table olives
- Tomato and pineapple juices
- Wheat flour hydrolysate
- Cheese broth.

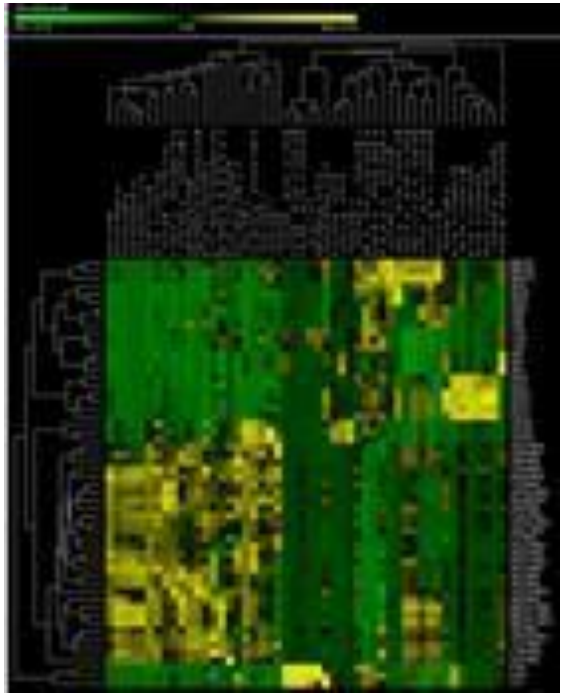
Later cultivation on MRS broth with two reference strains: WCFS1 and LB16

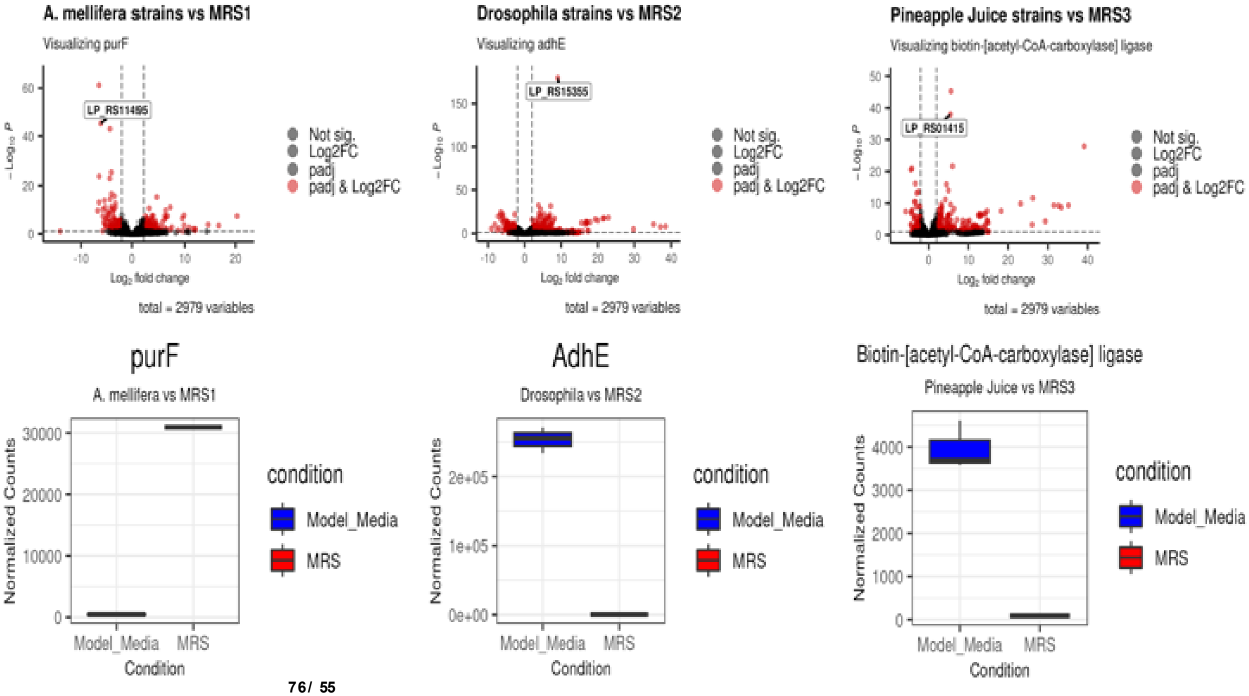
Green Trail
UG credentials:
1-semester of intro bioinformatics


A
Factor 1: 49.74%
Factor 2: 11.51%
PJ20
PJ1
PJ16
LB16
DE12II
DE12
DE24
CB5
CB2C
CB4
TJ20
TJ19
TJ22
WCFS1
BEE3ST
BEE2ST
BEE1ST
WFH20
WFH1W
WFH19
FEV3
FEO13
FEO3
FEV11
FEV21
FEO17
OE3
OE2
OE1
B
Factor 1: 34.88%
Factor 2: 11.72%
DE12II
PJ16
PJ20
PJ1
CB5_MRS
OE1_MRS
FEO3_MRS
FEV11_MRS
TJ22_MRS
LB16_MRS
WBEE1ST_MRS
PJ16_MRS
DE12_MRS
WCFS1MRS
DE12
DE24
CB5
CB2C
CB4
BEE3ST
TJ20
TJ19
TJ22
BEE2ST
BEE1ST
WFH20
WFH1W
WFH19
FEV3
FEO13
FEO3
FEV11
FEV21
FEO17
OE2
OE3
OE1

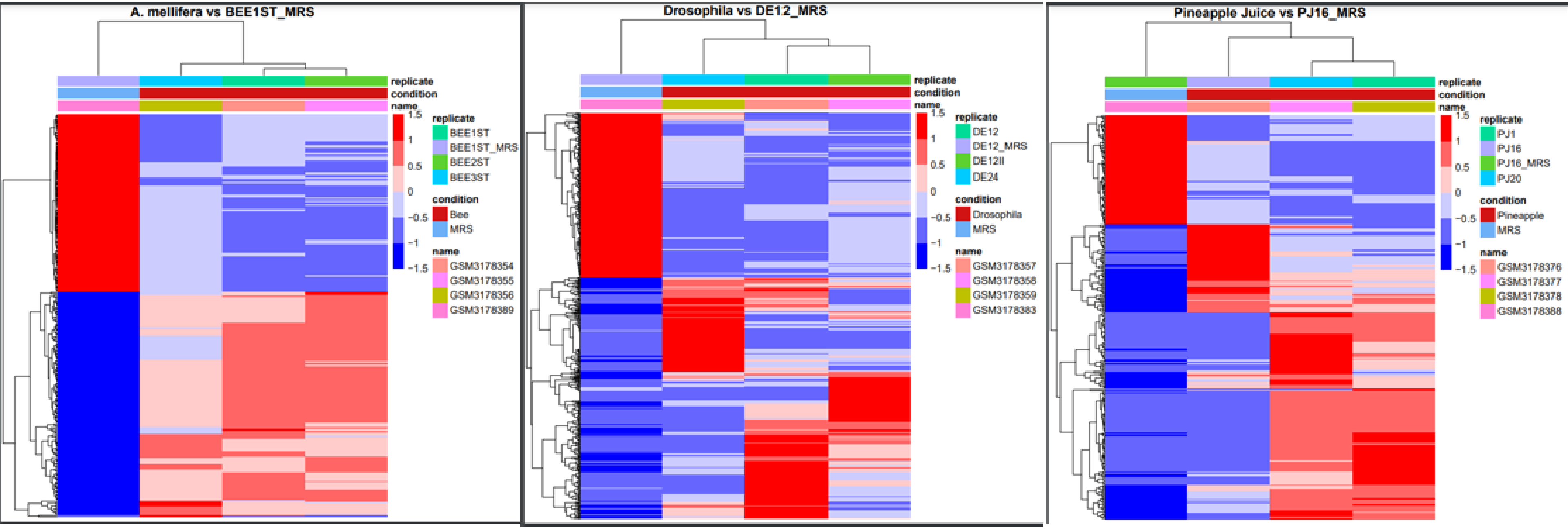

PC1: 45% variance
PC2: 32% variance
DE12II
DE12
DE24
PJ16
Strain
Bee Model Media
Drosphila Model Media
Pineapple Juice Model Media


PC1: 38% variance
PC2: 25% variance
PJ16_MRS
PJ16
BEE1ST_MRS
DE12_MRS
DE12II
DE12
DE24
Strain
Bee Model Media
Drosphila Model Media
MRS
Pineapple Juice Model Media





A. mellifera strains vs MRS1
Visualizing purF
LP_RS11495
−Log10 P
Log2 fold change
total = 2979 variables
Not sig.
Log2FC
padj
padj & Log2FC
Drosophila strains vs MRS2
Visualizing adhE
LP_RS15355
total = 2979 variables
Not sig.
Log2FC
padj
padj & Log2FC
Pineapple Juice strains vs MRS3
Visualizing biotin-[acetyl-CoA-carboxylase] ligase
LP_RS01415
total = 2979 variables
Not sig.
Log2FC
padj
padj & Log2FC
purF
A. mellifera vs MRS1
Normalized Counts
Model_Media
MRS
Condition
condition
Model_Media
MRS
AdhE
Drosophila vs MRS2
Normalized Counts
Model_Media
MRS
Condition
condition
Model_Media
MRS
Biotin-[acetyl-CoA-carboxylase] ligase
Pineapple Juice vs MRS3
Normalized Counts
Model_Media
MRS
Condition
condition
Model_Media
MRS



A. mellifera vs BEE1ST_MRS
replicate
condition
name
replicate
BEE1ST
BEE1ST_MRS
BEE2ST
BEE3ST
condition
Bee
MRS
name
GSM3178354
GSM3178355
GSM3178356
GSM3178389
Drosophila vs DE12_MRS
replicate
condition
name
replicate
DE12
DE12_MRS
DE12II
DE24
condition
Drosophila
MRS
name
GSM3178357
GSM3178358
GSM3178359
GSM3178383
Pineapple Juice vs PJ16_MRS
replicate
condition
name
replicate
PJ1
PJ16
PJ16_MRS
PJ20
condition
Pineapple
MRS
name
GSM3178376
GSM3178377
GSM3178378
GSM3178388

HW #5 (Due Feb 20th)

For this homework assignment, please identify the primary research article and samples you would like to perform this bioinformatic reanalysis on.

Keep in mind that each reanalysis will be performed with a specific, larger "goal" in mind.

These goals are specific to the trail selected and can be broadly summarized as: 1) to replicate the findings from the authors (**Green Mountain**), 2) alter the bioinformatic pipeline and understand how this impacts the final findings (**Blue Sky**), or 3) use the dataset to test an original hypothesis (**Black Diamond**).