Overview of RNA-Seq

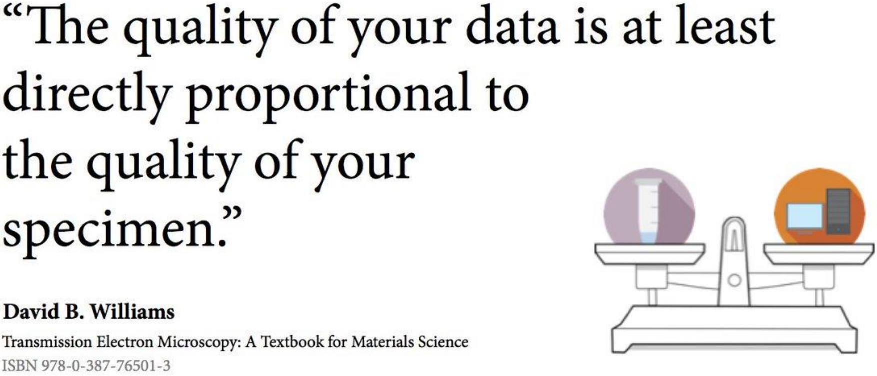
Dr. Princess Rodriguez

2025-01-29

Learning Objectives:

- Understand applications of RNA sequencing
- Introduce the overall differential expression workflow
- Understand experimental design concepts such as replicates and batch effects
- Understand different types of library preps, their requirements and uses.

Why?



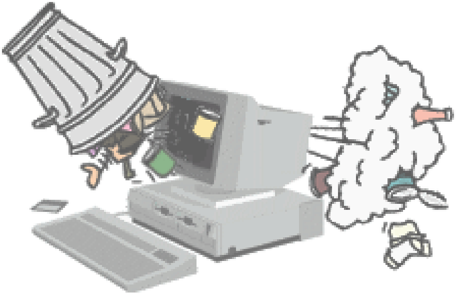
“The quality of your data is at least directly proportional to the quality of your specimen.”

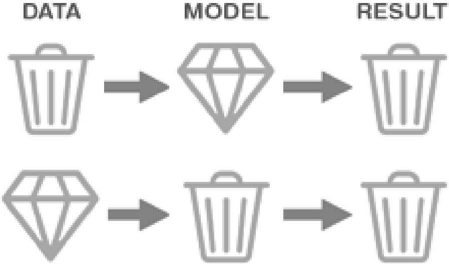
David B. Williams

Transmission Electron Microscopy: A Textbook for Materials Science

ISBN 978-0-387-76501-3

Garbage In, Garbage Out





Your input will define the quality of output you get!

Overview of RNA-seq

RNA-seq is an exciting experimental technique that is utilized to explore and/or quantify gene expression within or between conditions.

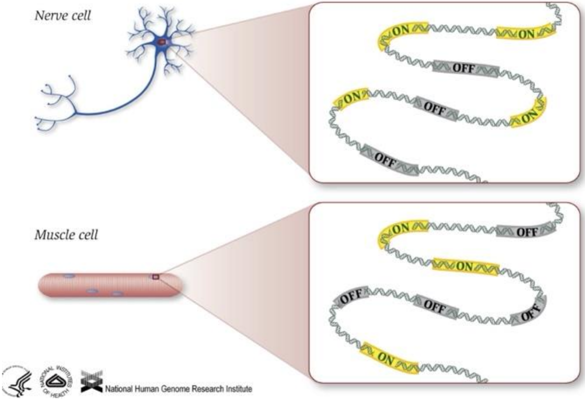


Figure 1: Gene Expression in Cells

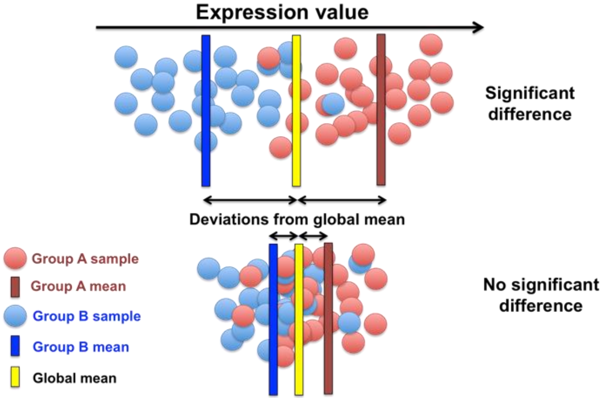
The Transcriptome

The transcriptome is defined as a collection of all the transcript readouts present in a cell. RNA-seq data can be used to explore and/or quantify the transcriptome of an organism, which can be utilized for the following types of experiments:

- **Differential Gene Expression**: *quantitative* evaluation and comparison of transcript levels between conditions
- **Transcriptome assembly**: building the profile of transcribed regions of the genome, a *qualitative* evaluation
- **Refinement of gene models**: building better gene models and verifying them using transcriptome assembly
- **Metatranscriptomics**: community transcriptome analysis

Basic types of questions answered:

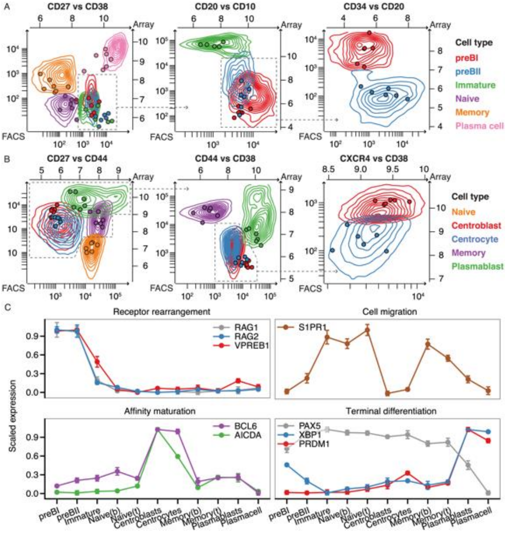
What genes are differentially expressed between conditions?



Other questions answered:

Are there any trends in gene expression across development?

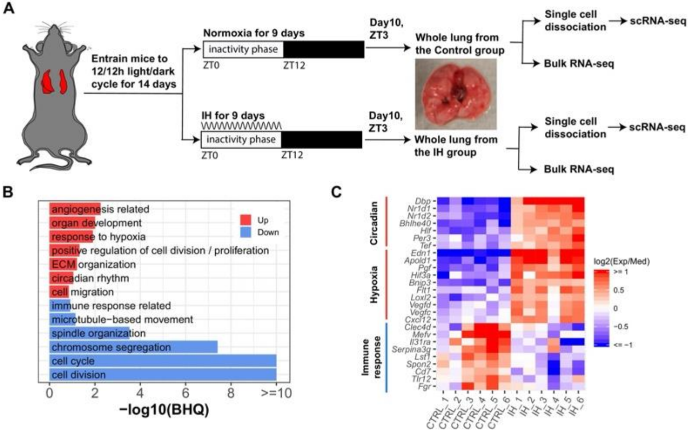
Which groups of genes change similarly over time or across conditions?



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

Basic types of questions answered:

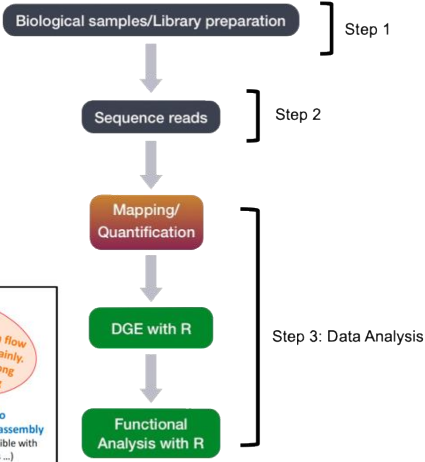
What processes or pathways are enriched in condition of interest?



https://elifesciences.org/articles/63003

Basic Principals

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



Biological samples/Library preparation

Step 1

Sequence reads

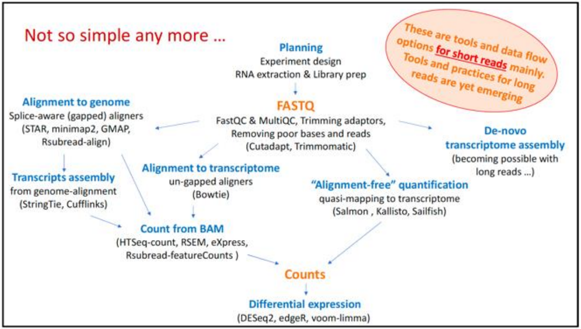
Step 2

Mapping/ Quantification

DGE with R

Functional Analysis with R

Step 3: Data Analysis



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)

Transcripts assembly
from genome-alignment (StringTie, Cufflinks)

Alignment to transcriptome
un-gapped aligners (Bowtie)

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

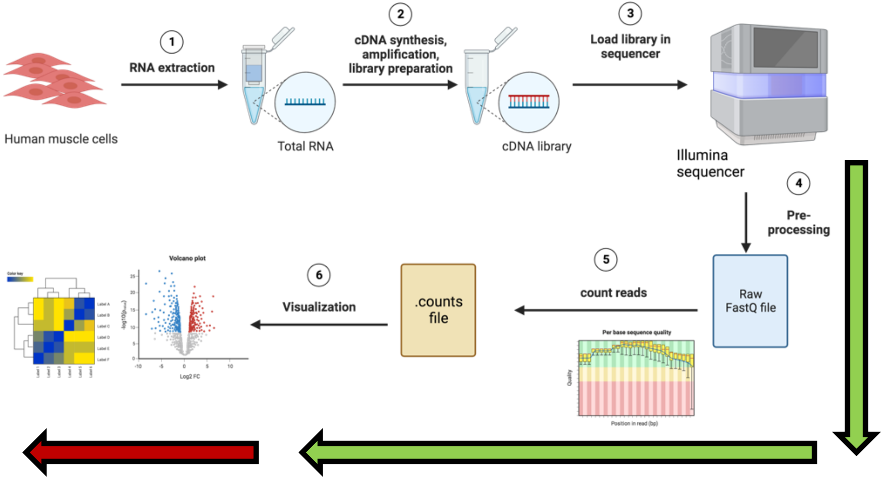
"Alignment-free" quantification
quasi-mapping to transcriptome (Salmon , Kallisto, Sailfish)

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

Counts

Differential expression
(DESeq2, edgeR, voom-limma)

RNA-Seq Bioinformatic Pipeline



Human muscle cells

1 RNA extraction

Total RNA

2 cDNA synthesis, amplification, library preparation

cDNA library

3 Load library in sequencer

Illumina sequencer

4 Pre-processing

Raw FastQ file

5 count reads

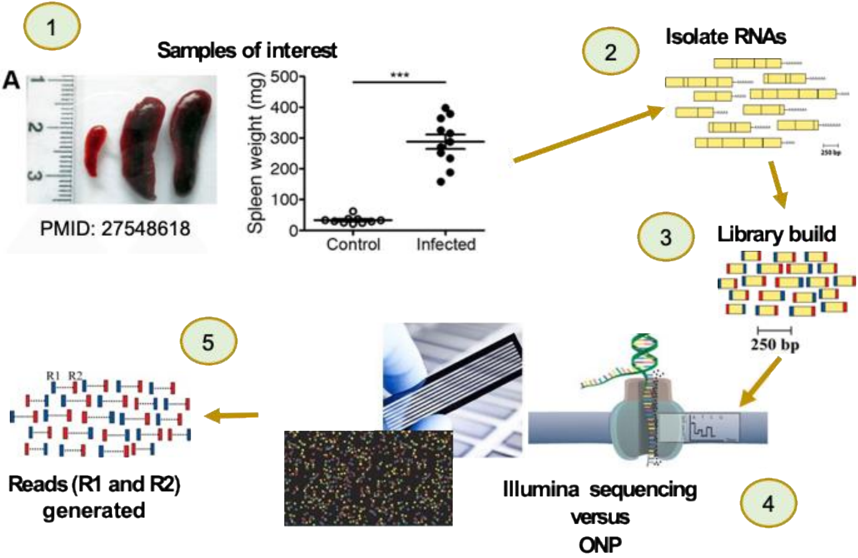
.counts file

6 Visualization

R/RStudio

UNIX/SHELL scripting

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**.

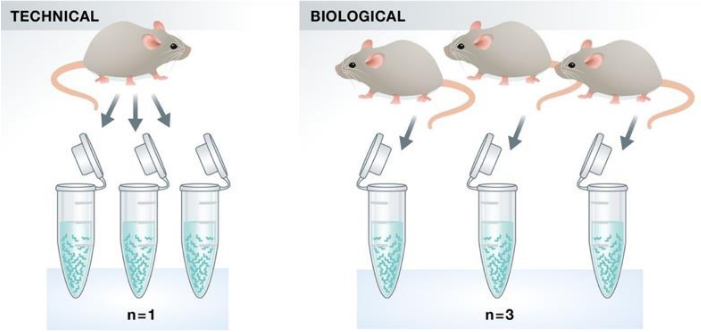
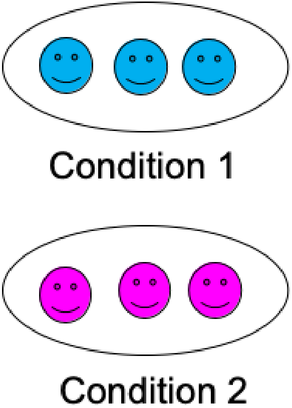


Figure 16: Biological Replicates

*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



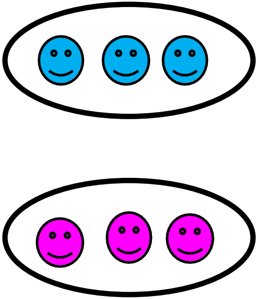
- To detect Differentially Expressed Genes (DEGs) between groups we should have several samples, which are also known as biological replicates

Probability of detecting DEGs

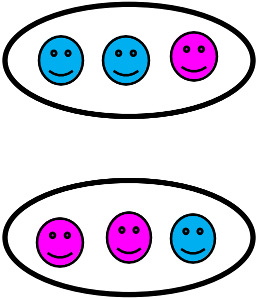
	Replicates per group		
	3	5	10
Fold change			
2	87%	98%	100%

PMID: 26813401

Grouping of Replicates

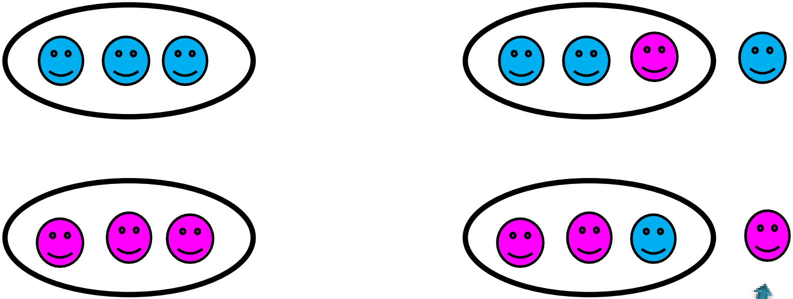


What you want



What you get

Grouping of Replicates



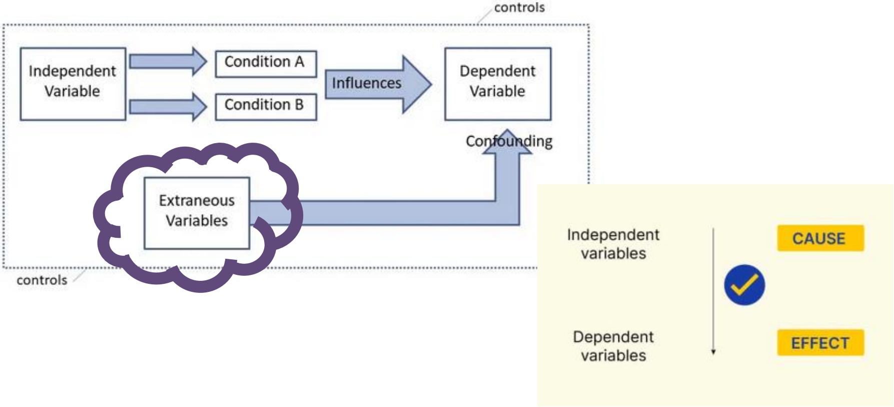
What you want

What you get

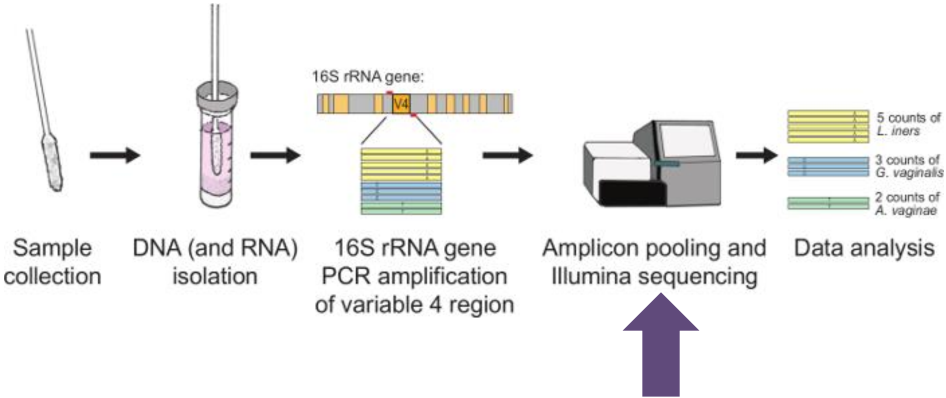
That spare comes in handy
Highly recommend especially
with mice!

What causes this?
Confounding variables

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

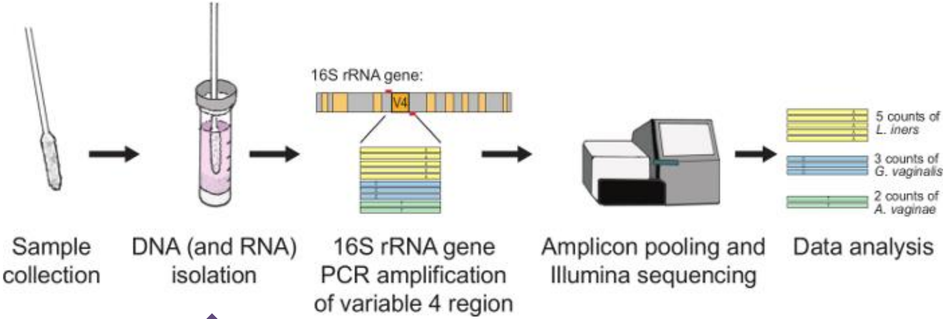


Examples of confounding variables



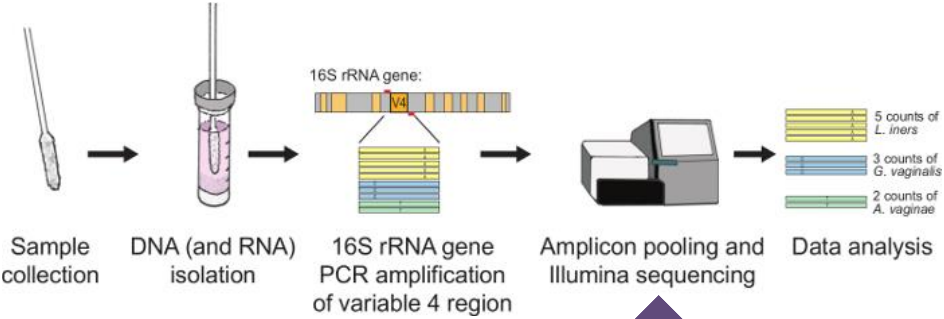
A new technician is running the sequencer

Examples of confounding variables



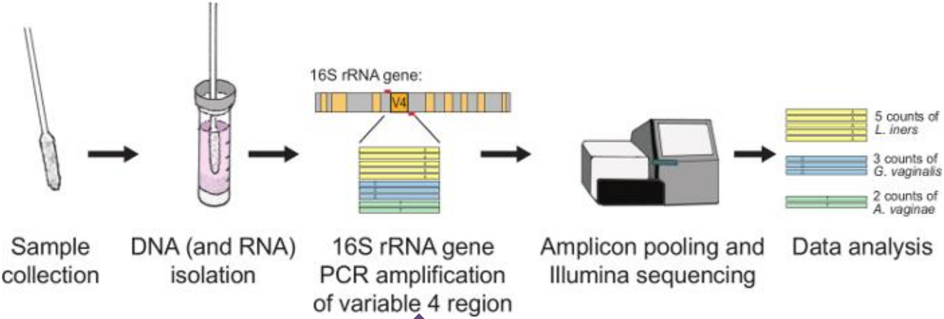
Extracting DNA/RNA with two different kits!

Examples of confounding variables



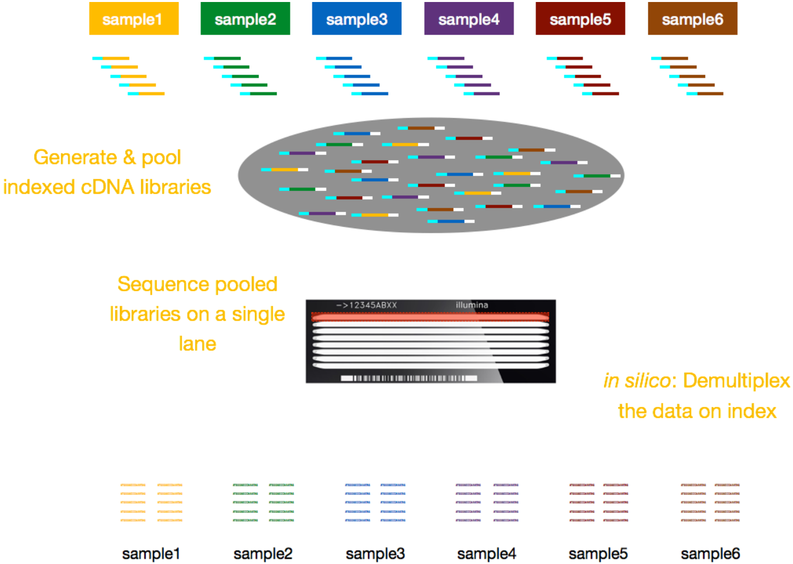
Sequencing on multiple different types of platforms

Examples of confounding variables



Inappropriate multiplexing strategy

Multiplexing

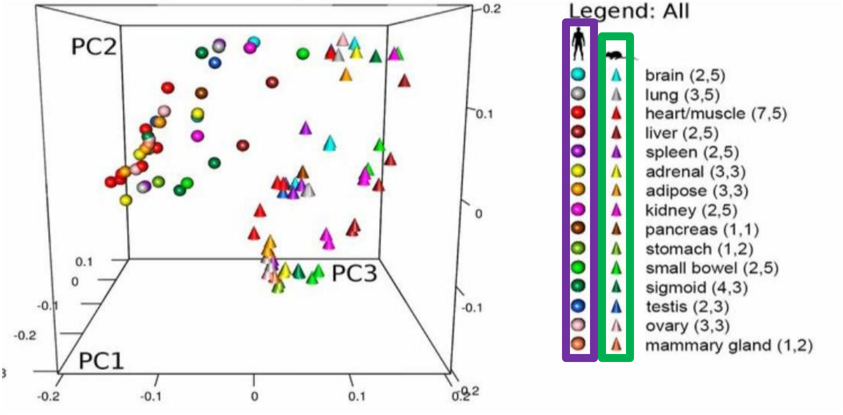


Generate & pool indexed cDNA libraries

Sequence pooled libraries on a single lane

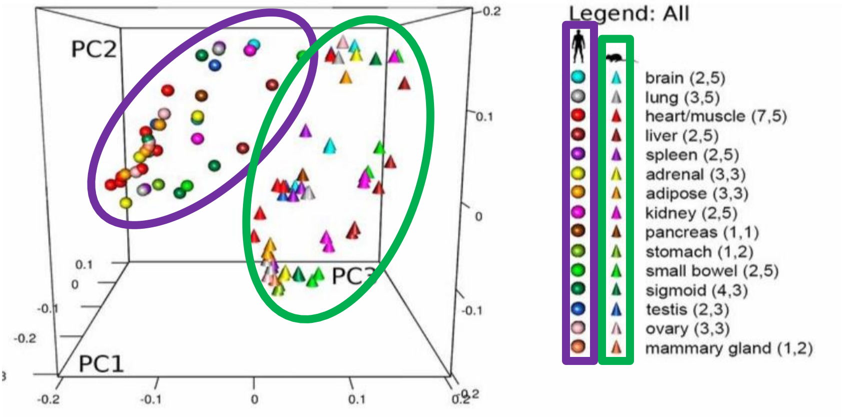
in silico: Demultiplex the data on index

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

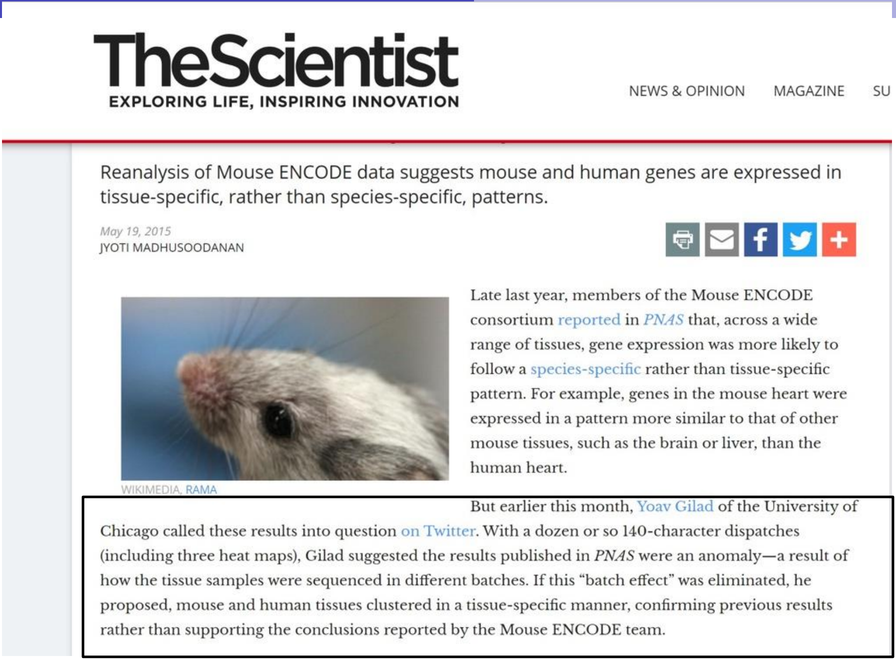


https://doi.org/10.1073/pnas.1413624111

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



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



TheScientist

EXPLORING LIFE, INSPIRING INNOVATION

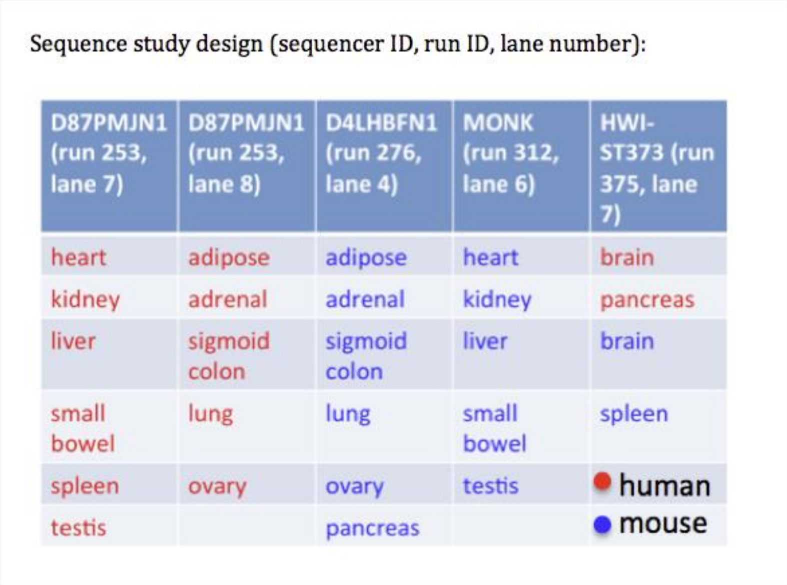
Reanalysis of Mouse ENCODE data suggests mouse and human genes are expressed in tissue-specific, rather than species-specific, patterns.

May 19, 2015
JYOTI MADHUSOODANAN

WIKIMEDIA, RAMA

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

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

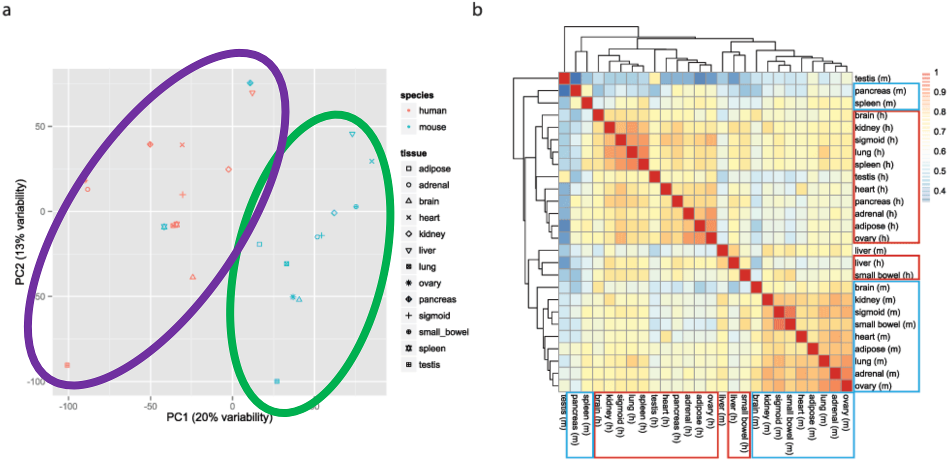


Sequence study design (sequencer ID, run ID, lane number):

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

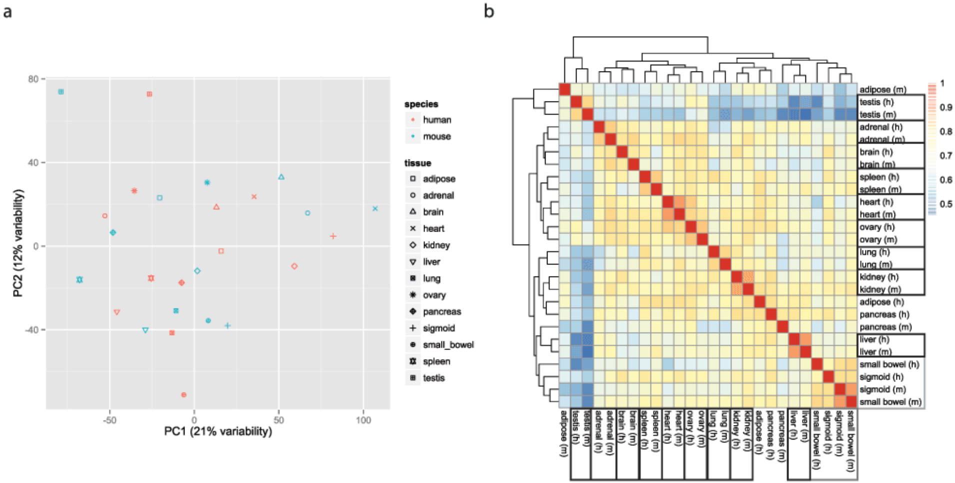
Sequencing lane (a batch effect) was almost completely confounded with species in the PNAS study. From @Y_Gilad

Before accounting for batch effect



Samples grouped by animal

After accounting for batch effect



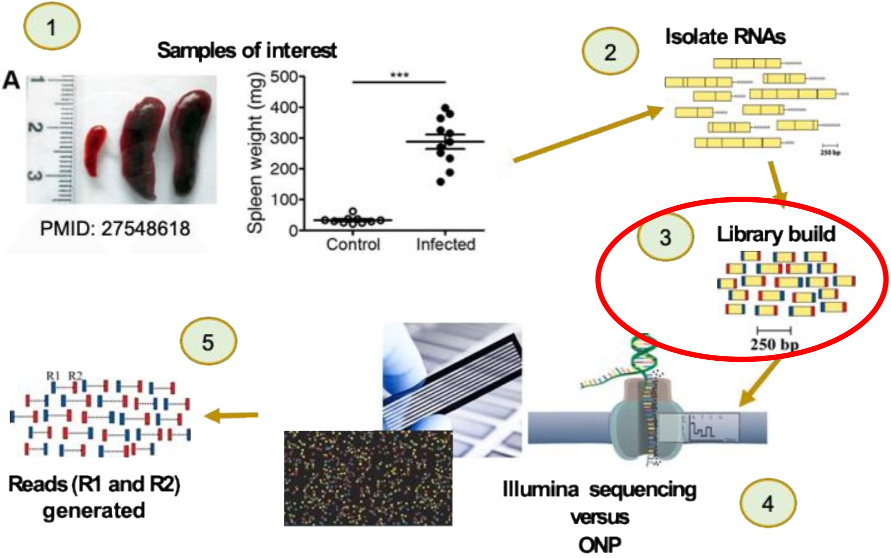
Samples now grouped by tissue!

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

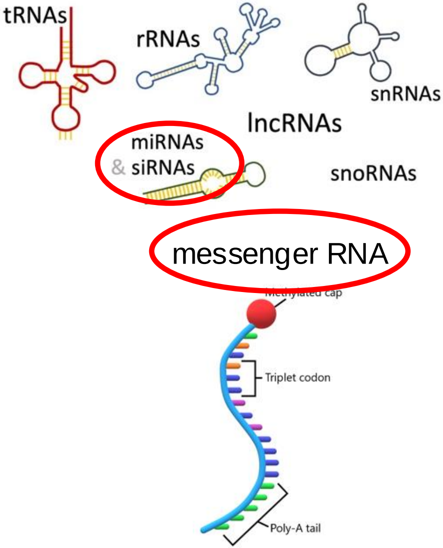
4 Illumina sequencing versus ONP

5 Reads (R1 and R2) generated

RNA composition

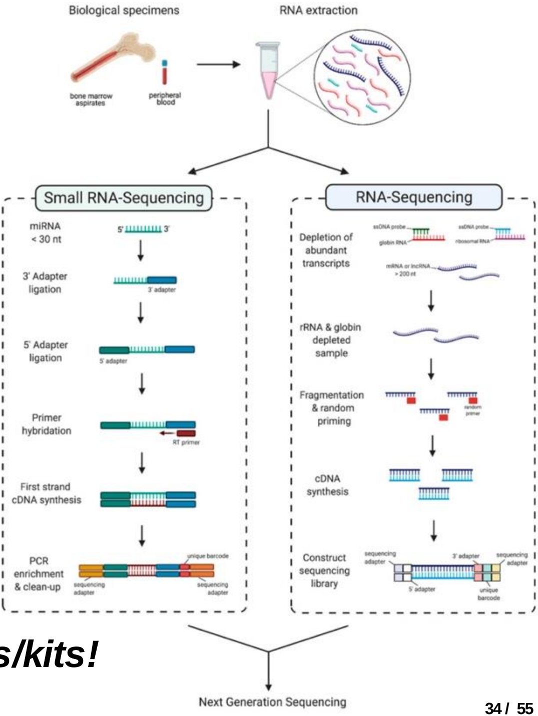
RNA comes in many different flavors

- 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

“Comprehensive” transcriptome analysis

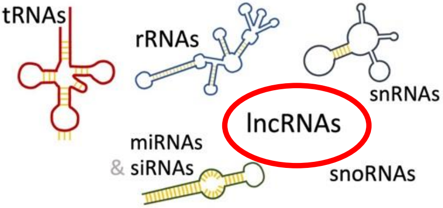


Two different protocols/kits!

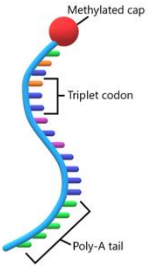
RNA composition

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

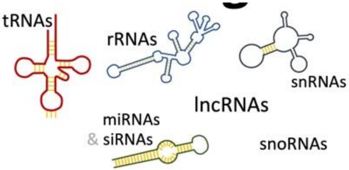


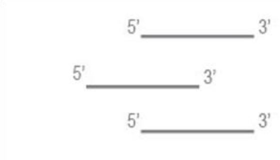
messenger RNA

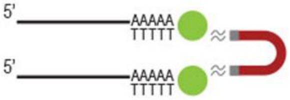


- Kukurba, Mongomery (2015) Cold Spring Harb. Protoc.
Gaida SM et al. (2015) Nat Comm 6:7045

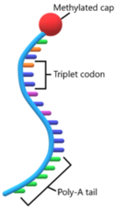
The RNA sample undergoes either selection of the mRNA (polyA selection) or depletion of the rRNA. The resulting RNA is fragmented.







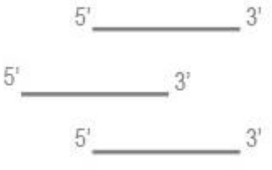
messenger RNA

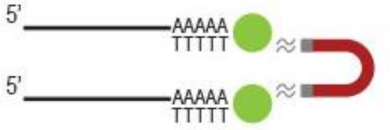


Poly-A versus rRNA depletion?

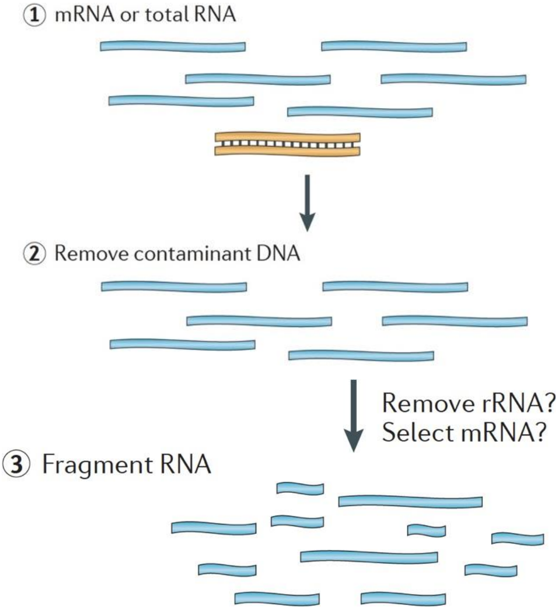
If you are aiming to obtain information about long non-coding RNA’s I recommend performing ribosomal RNA depletion

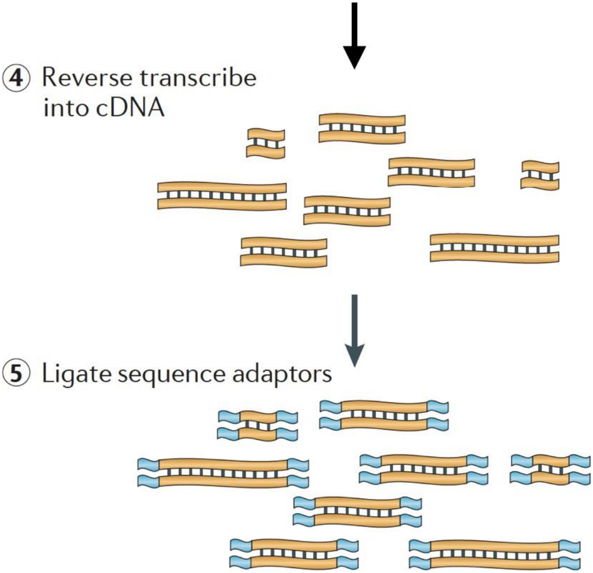
Bacterial mRNAs are also not poly-adenylated





Illumina Library preparation




④ Reverse transcribe into cDNA
⑤ Ligate sequence adaptors

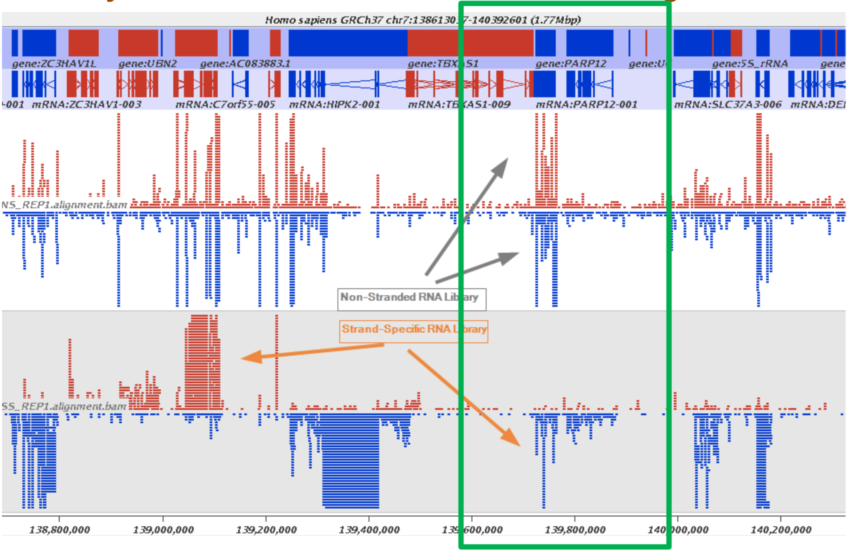
Strandedness

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

Simple and accurate analysis of overlapping genes: Clearly see that PARP12 is encoded on the negative strand

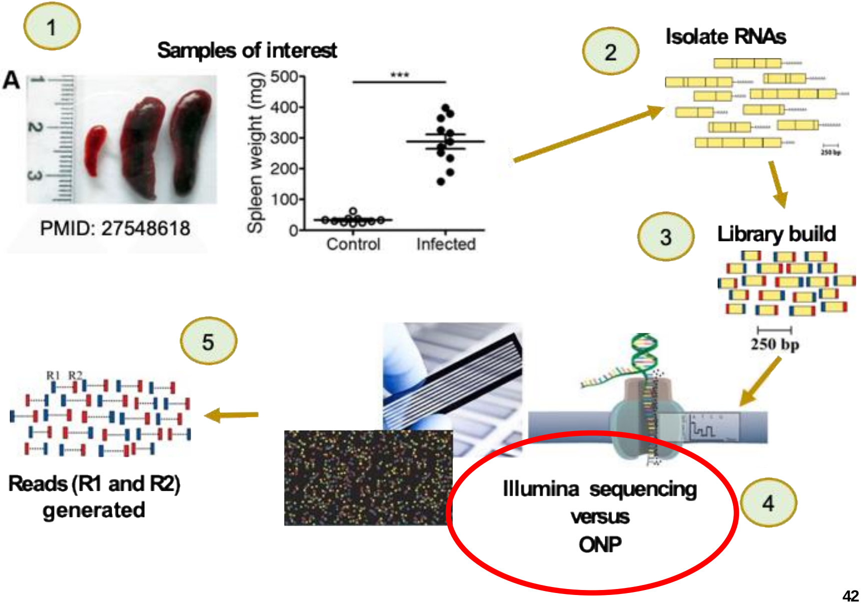


Non-Stranded RNA Library

Strand-Specific RNA Library

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

Experimental workflow



1 Samples of interest

PMID: 27548618

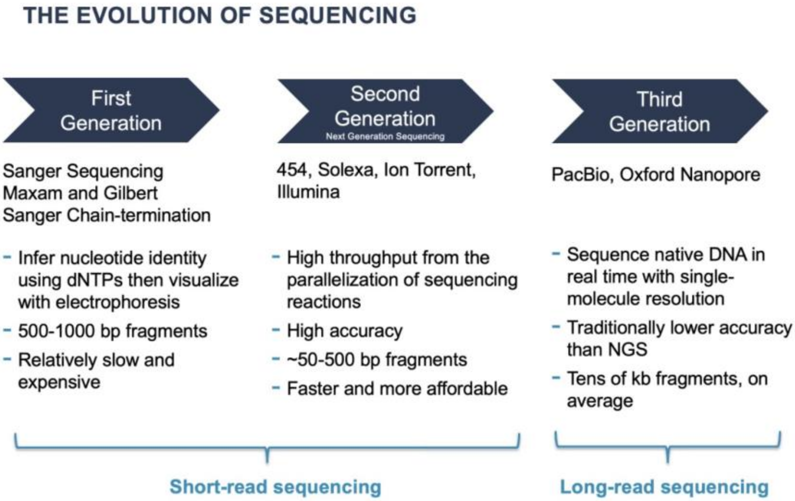
2 Isolate RNAs

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).

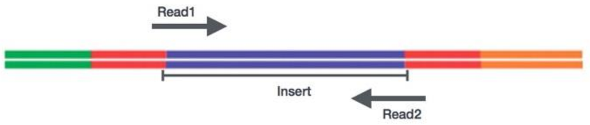
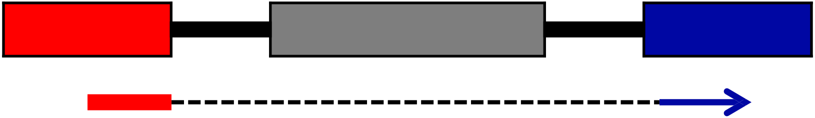


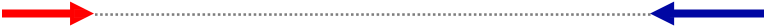
Figure 10: Paired End Reads

- 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.

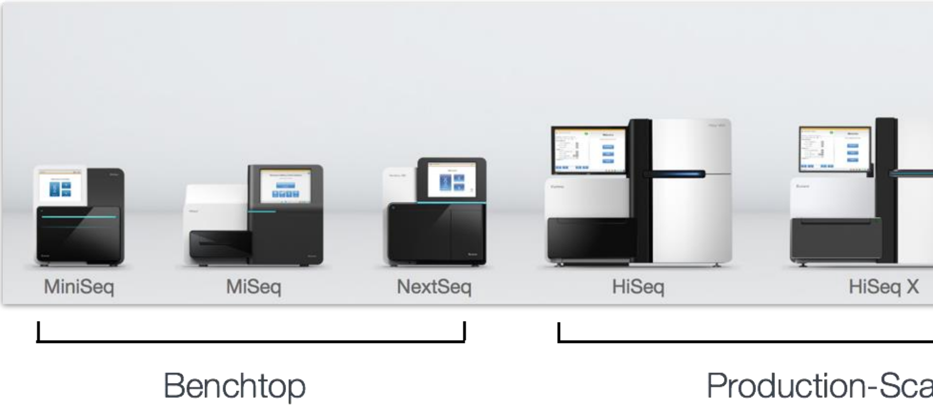
In Summary, to quantify Differential Gene Expression

- 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!

Different sequencing platforms

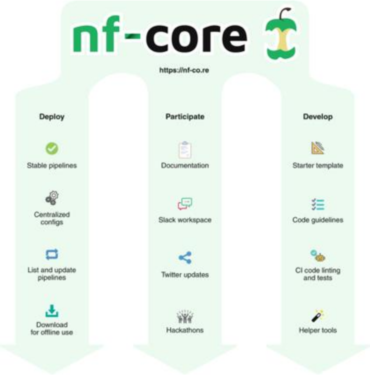
There are a variety of Illumina platforms to choose from to sequence the cDNA libraries.

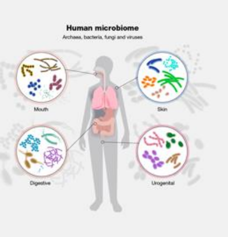


Final projects from the years have spanned the following topics:

Salmonella enterica

Applications of organoids as research models

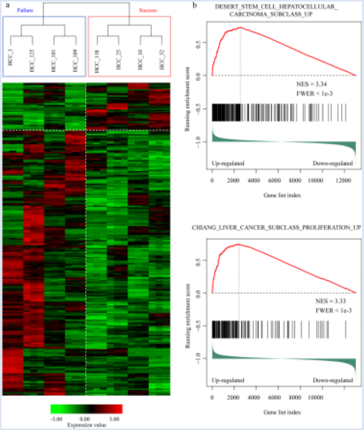




Dengue

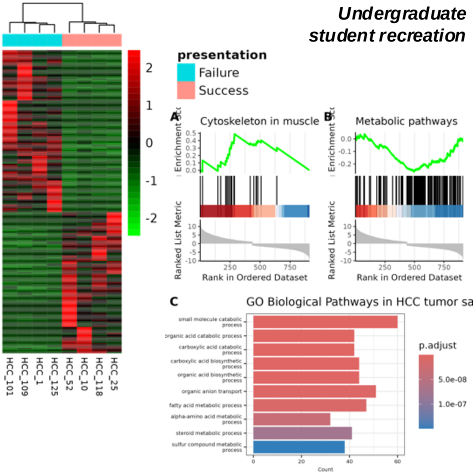
Breast cancer

And more….!



Original Published Work

Undergraduate student recreation

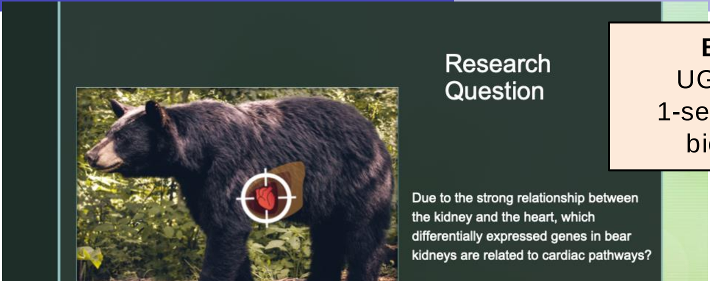


Green Trail

UG credentials:

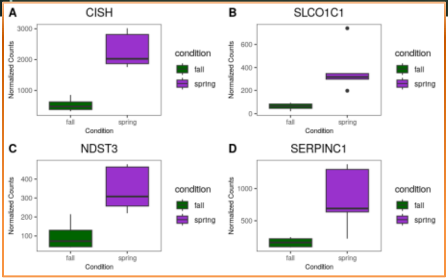
1-semester of intro bioinformatics

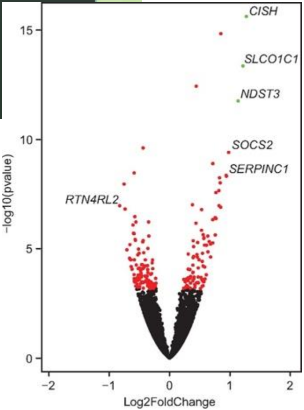
Black Trail
UG credentials:
1-semester of intro bioinformatics



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?

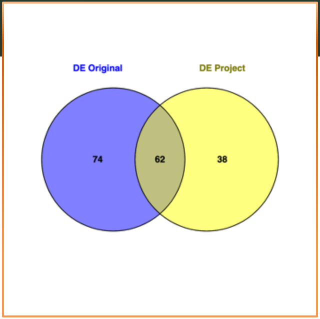


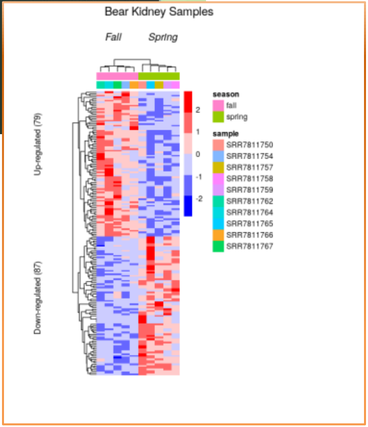


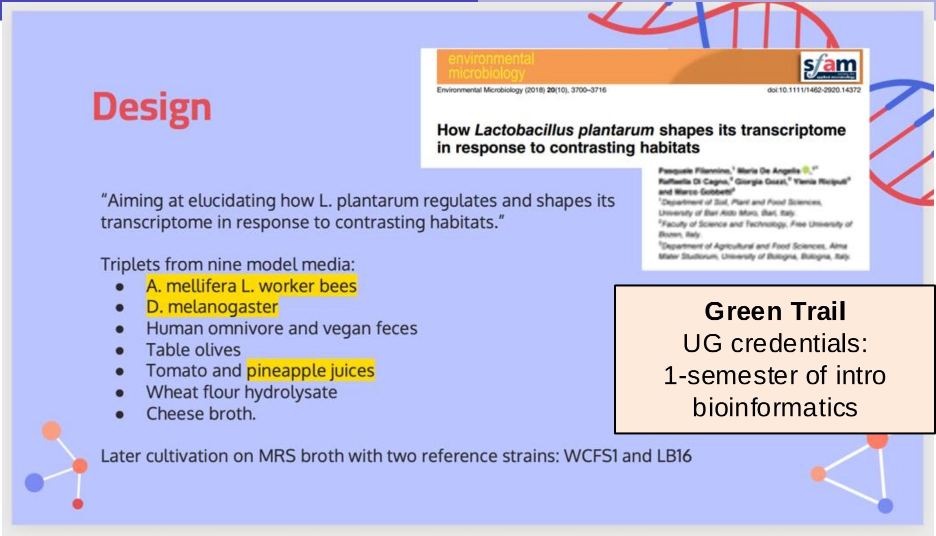
Black Trail
UG credentials:
1-semester of intro
bioinformatics

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 (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] Yiwia 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.*

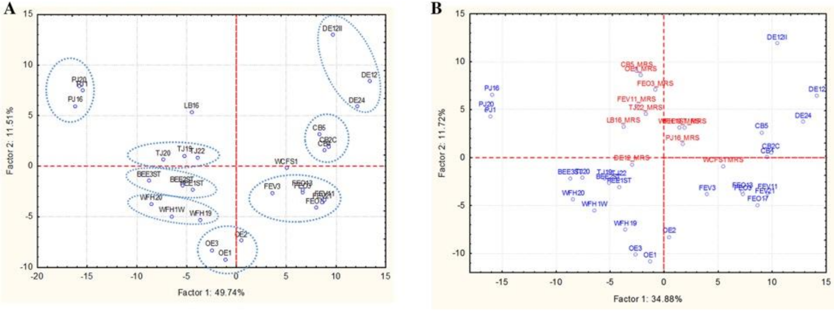
“Aiming at elucidating how L. plantarum regulates and shapes its transcriptome in response to contrasting habitats.”

Triplets from nine model media:

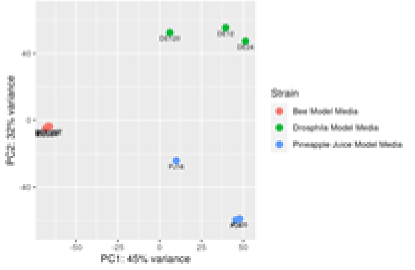
- A. mellifera L. worker bees
- D. melanogaster
- Human omnivore and vegan feces
- Table olives
- Tomato and pineapple juices
- Wheat flour hydrolysate
- Cheese broth.

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

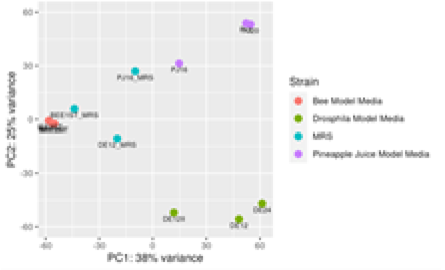
Green Trail
UG credentials:
1-semester of intro bioinformatics



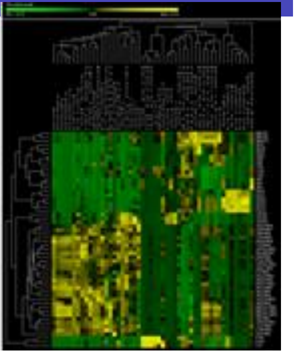
A
Factor 1: 49.74%
Factor 2: 11.51%
-20
-15
-10
-5
0
5
10
15
PJ20
PJ1
PJ16
LB16
DE12II
DE12
DE24
CB5
CB2C
CB1
TJ20
TJ19
TJ22
WCFS1
BEE3ST
BEE2ST
BEE1ST
FEV3
FEO13
FEO19
FEV11
FEO17
WFH20
WFH1W
WFH19
OE2
OE3
OE1
B
Factor 1: 34.88%
Factor 2: 11.72%
DE12I
DE12
DE24
PJ16
PJ20
PJ1
CB5_MRS
OE1_MRS
FEO3_MRS
FEV11_MRS
TJ22_MRS
LB16_MRS
WBEE1ST_MRS
PJ16_MRS
DE12_MRS
WCFS1MRS
CB5
CB2C
CB1
BEE3ST
TJ20
TJ19
TJ22
BEE2ST
BEE1ST
FEV3
FEO13
FEO9
FEV11
FEV21
FEO17
WFH20
WFH 1W
WFH 19
DE2
OE3
OE1

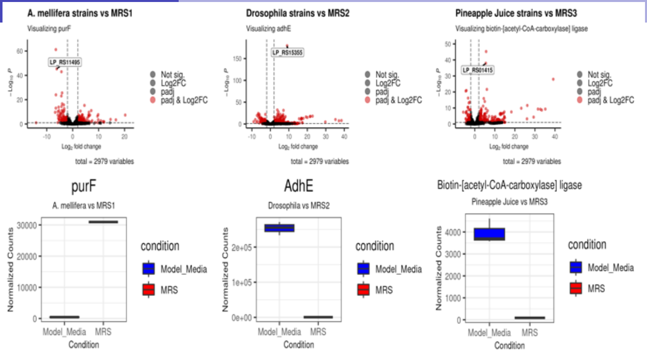


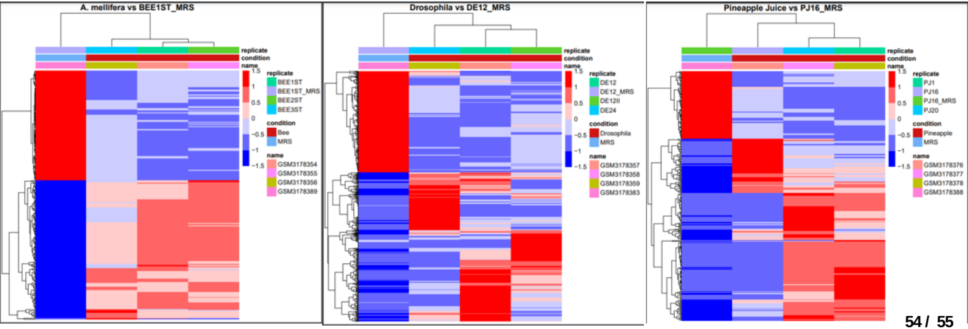
PC1: 45% variance
PC2: 32% variance
40
0
-40
-50
-25
0
25
50
Strain
● Bee Model Media
● Drosphila Model Media
● Pineapple Juice Model Media
PJ16



PC1: 38% variance
PC2: 25% variance
60
30
0
-30
-60
-60
-30
0
30
60
Strain
● Bee Model Media
● Drosphila Model Media
● MRS
● Pineapple Juice Model Media
PJ16_MRS
PJ16
DE12_MRS




A. mellifera strains vs MRS1
Visualizing purF
LP_RS11495
$-\log_{10} P$
Log₂ fold change
total = 2979 variables
Drosophila strains vs MRS2
Visualizing adhE
LP_RS15355
Pineapple Juice strains vs MRS3
Visualizing biotin-[acetyl-CoA-carboxylase] ligase
LP_RS01415
● Not sig.
● Log2FC
● padj
● padj & Log2FC
purF
A. mellifera vs MRS1
AdhE
Drosophila vs MRS2
Biotin-[acetyl-CoA-carboxylase] ligase
Pineapple Juice vs MRS3
Normalized Counts
Condition
Model_Media
MRS
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
DE12
DE12_MRS
DE12II
DE24
condition
Drosophila
MRS
name
GSM3178357
GSM3178358
GSM3178359
GSM3178383
Pineapple Juice vs PJ16_MRS
replicate
PJ1
PJ16
PJ16_MRS
PJ20
condition
Pineapple
MRS
name
GSM3178376
GSM3178377
GSM3178378
GSM3178388
1.5
1
0.5
0
-0.5
-1
-1.5

Citation

This lesson has been developed by members of the teaching team at the [Harvard Chan Bioinformatics Core (HBC).](#) These are open access materials distributed under the terms of the [Creative Commons Attribution license](#) (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Authors: Mary Piper, Meeta Mistry, Radhika Khetani
Other sources - https://umich-brcf-bioinf.github.io/rnaseq_demystified_workshop/site/Module3a_Design_Prep_Seq#2_Experimental_Design_and_Practicalities