



# Introduction to RNA-seq for differential gene expression (DGE) analysis

---

**Owen M. Wilkins, PhD**

Bioinformatics scientist

Center for Quantitative Biology, Geisel School of Medicine at Dartmouth

Email: [DataAnalyticsCore@groups.dartmouth.edu](mailto:DataAnalyticsCore@groups.dartmouth.edu)

Website: (<https://sites.dartmouth.edu/cqb/projects-and-cores/data-analytics-core/>)

---

07/18/22



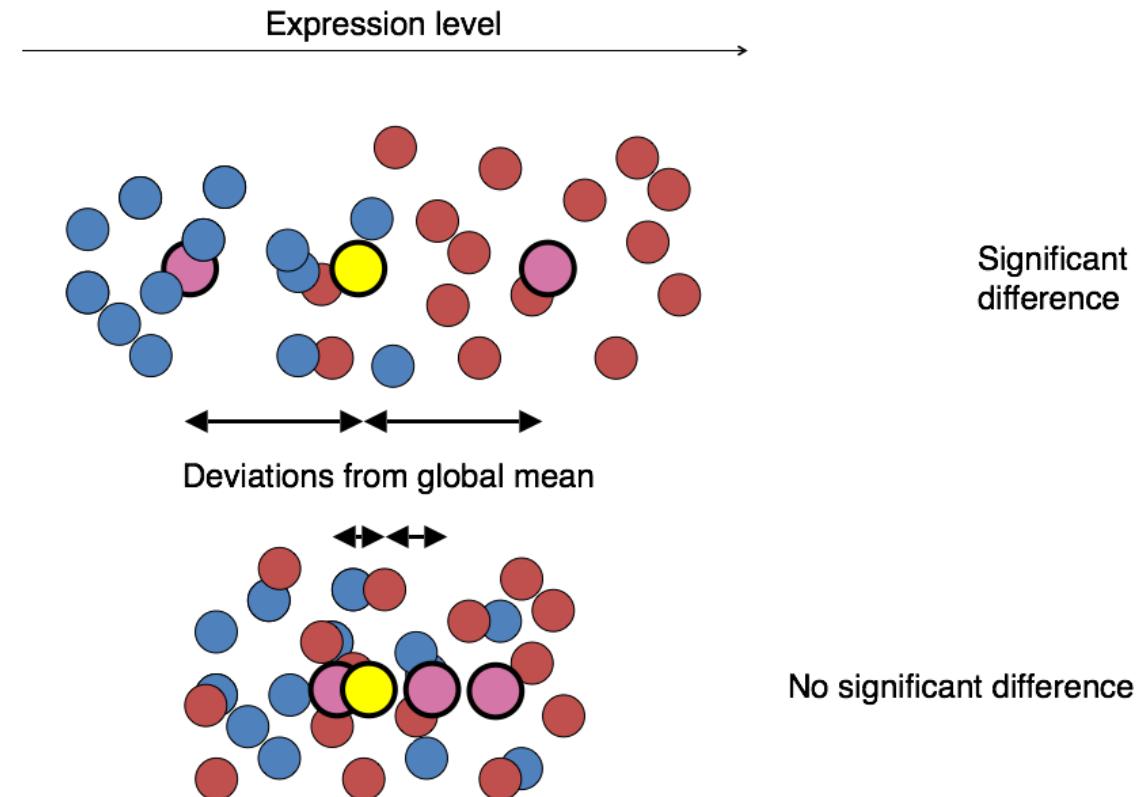
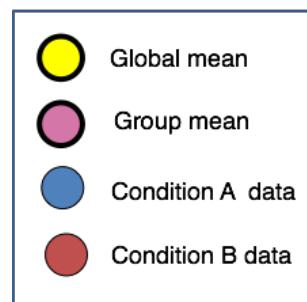
**Dartmouth**  
GEISEL SCHOOL OF MEDICINE

# Differential expression analysis

- Aim: Test quantitative expression changes between two or more experimental groups
- Perform separate statistical test for read counts of each gene

Sample metadata:

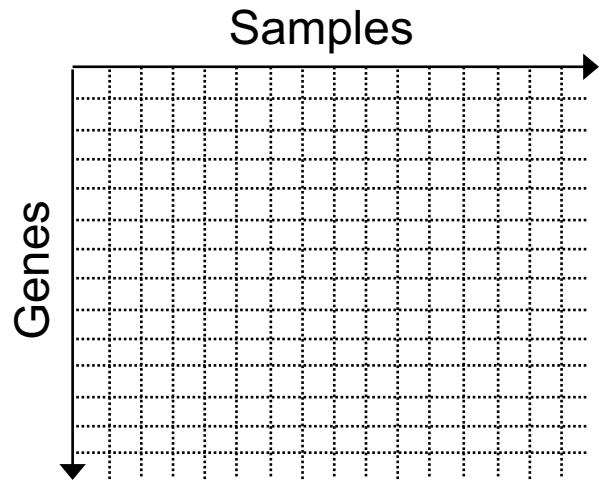
	Sex	Age	Tx-group
Sample_1	F	33	vehicle
Sample_1	F	21	vehicle
Sample_1	M	22	tx
Sample_1	F	29	tx
....	....	....	....
Sample_X	M	35	tx



[https://hbctraining.github.io/DGE\\_workshop/lessons/04\\_DGE\\_DESeq2\\_analysis.html](https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html)

# Differential expression analysis

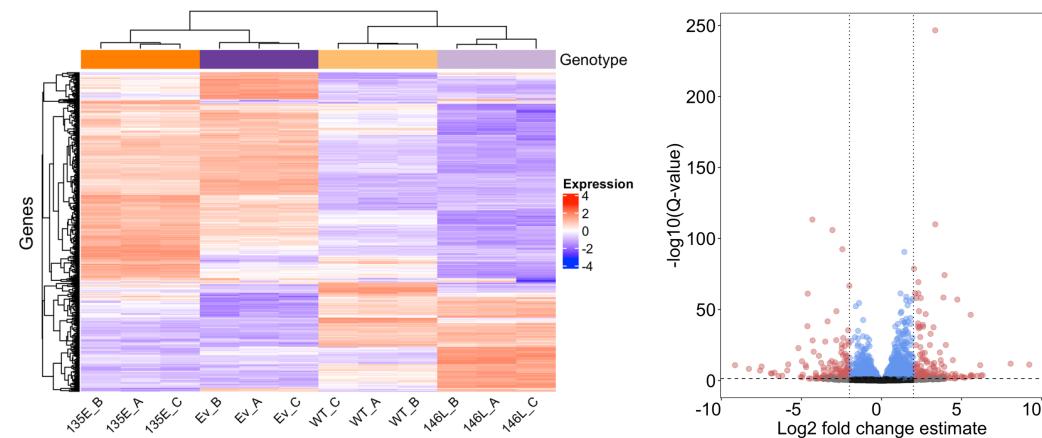
Gene expression matrix  
(read counts)



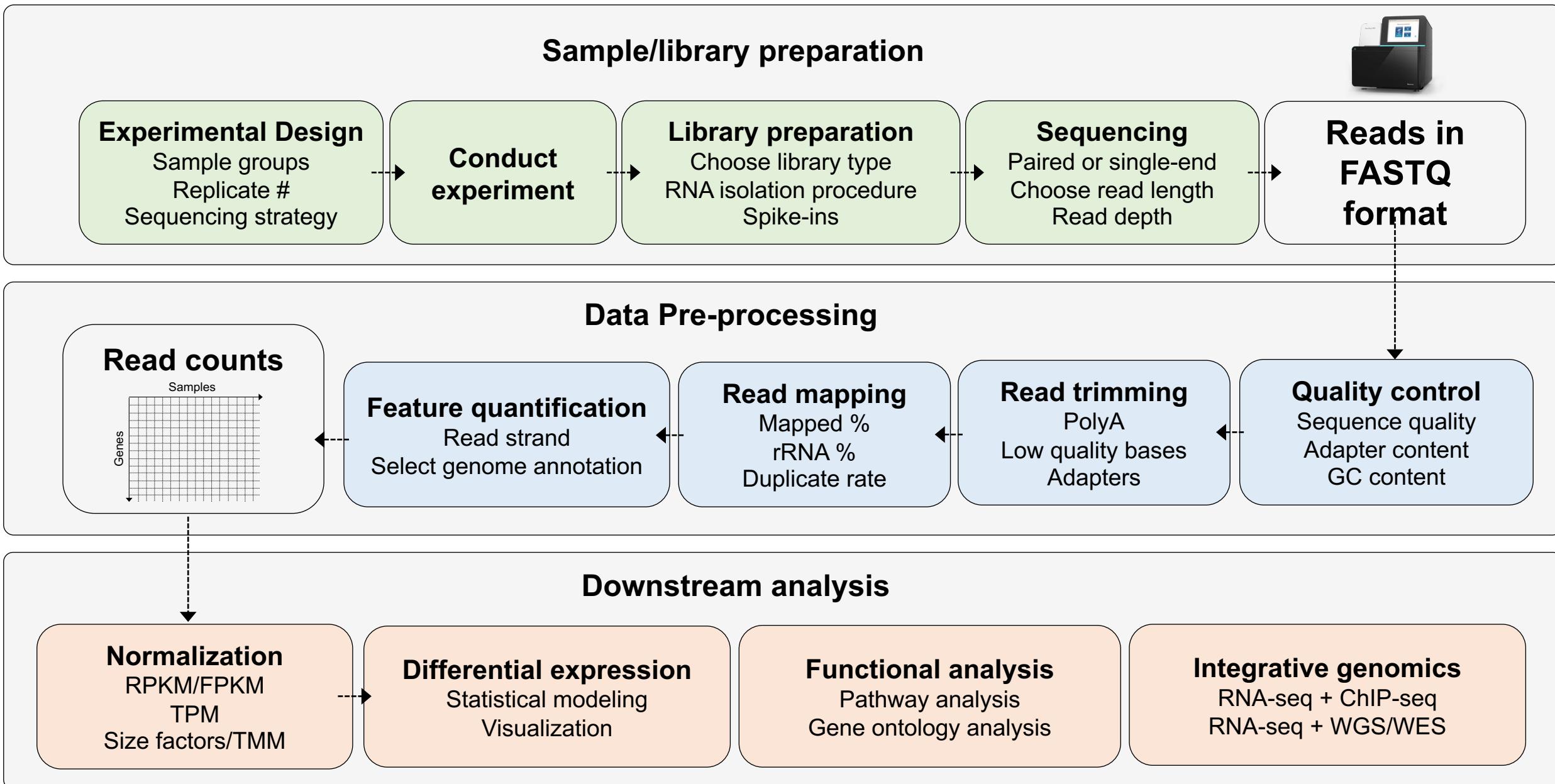
- Normalization
- Exploratory analysis
- Statistical modeling
- Results prioritization
- Visualization

Differential expression results

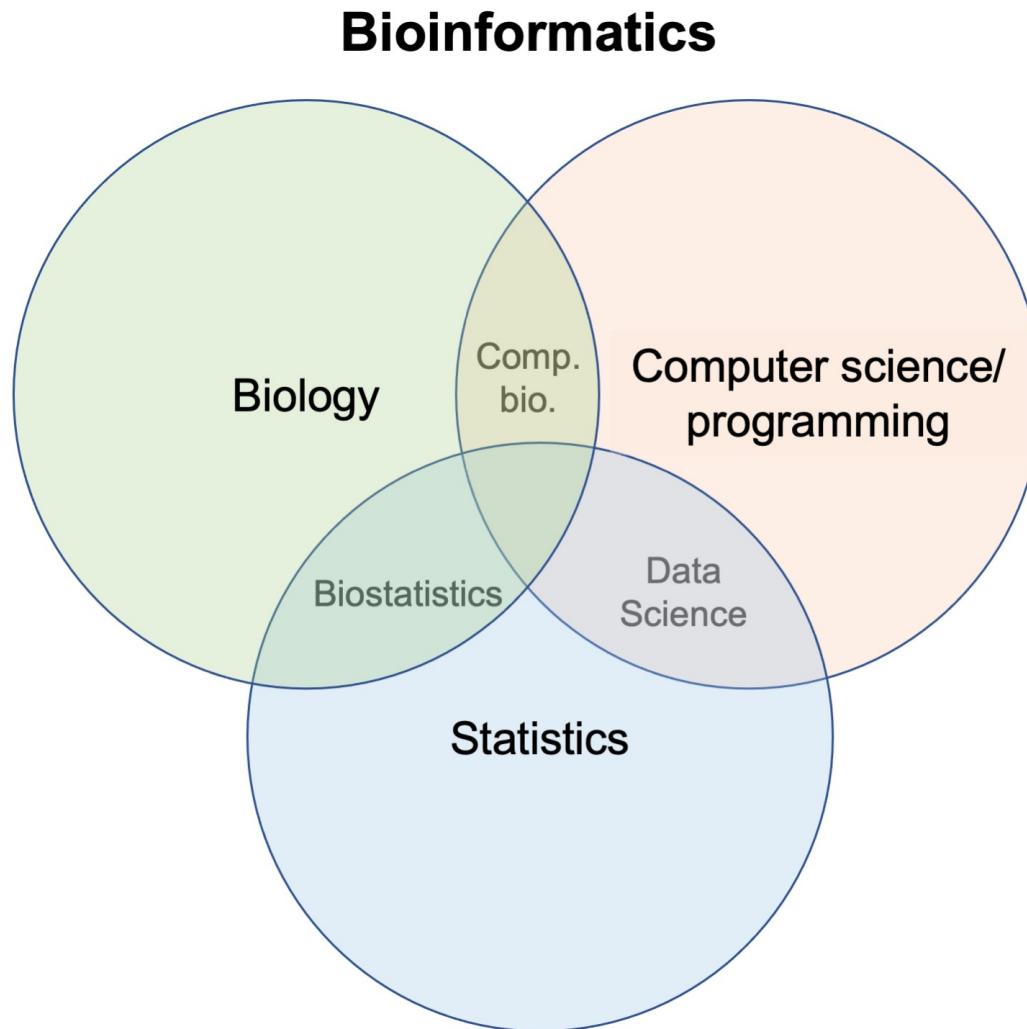
ID	Ifc	IfcSE	stat	pvalue	padj	Gene
ENSG032219	2.89	0.12	-25.4	1.6E-251	1.7E-249	ARID4A
ENSG012951	-3.44	0.15	-24.5	1.2E-132	1.4E-122	KLK10
ENSG016754	-3.14	0.12	-24.1	2.9E-123	1.4E-118	KLK5
ENSG080097	-2.78	0.11	-20.7	5.7E-115	2.4E-109	LGALS1
ENSG006277	5.92	0.28	19.8	1.9E-80	3.5E-92	UCHL1
.....	.....	.....	.....	.....	.....	.....



# Overview of a typical RNA-seq experiment



# Technical skills for RNA-seq analysis



## Data pre-processing:

- Requires understanding of library preparation and sequencing technology
- More dependent on programming & computational skill set
- Often performed using established pipelines

## Down-stream analysis

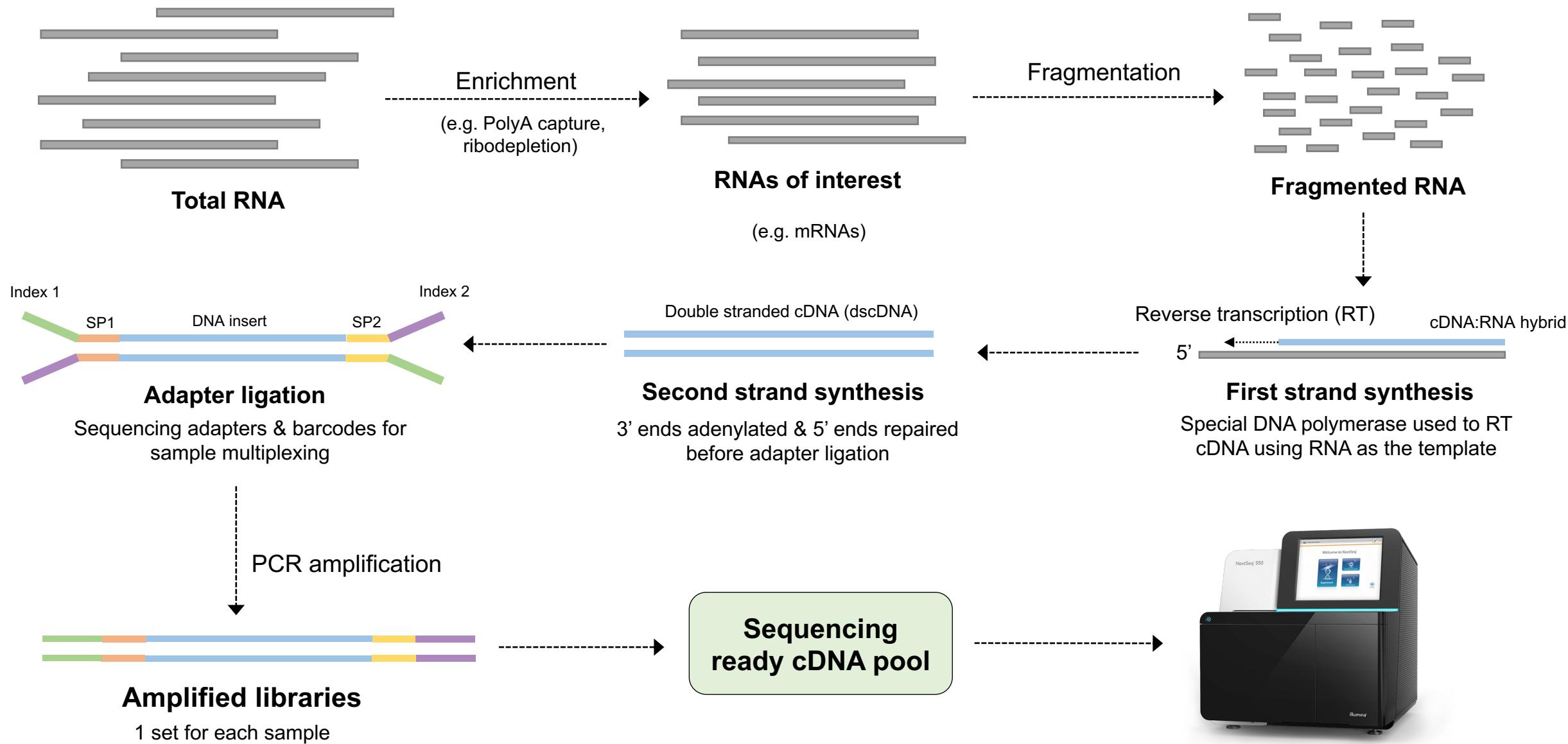
- Requires basic knowledge of statistical concepts
- More dependent on statistical programming & data visualization skill set

# Talk outline:



- **Library preparation**
  - Basic protocol
  - Types of libraries
- **Data pre-processing**
  - Quality control
  - Read trimming
  - Read mapping
  - Read counting/expression quantification
- **Sample preparation & experimental design**
  - Replicates
  - Sequencing depth
  - Sequencing configuration (read type & length)

# Library preparation for RNA-seq

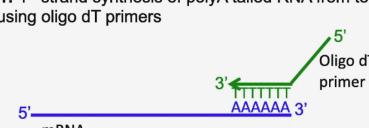


# Applications of different library types

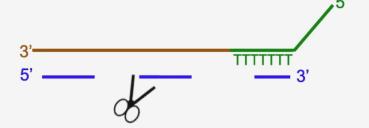
## 3' End

### 3' method (LEXO)

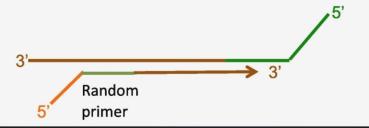
**Step 1:** 1<sup>st</sup> strand synthesis of polyA tailed RNA from total RNA using oligo dT primers



**Step 2:** Degradation of the RNA template



**Step 3:** 2<sup>nd</sup> strand synthesis with random primers containing 5' Illumina-compatible linker sequences



**Step 4:** Amplification using random primers that add barcodes and cluster generation sequences



**Step 5:** Sequencing

Adapted from Fukua et al, 2019. *Genom. Biol.*

Differential Expression

Lower cost/High Throughput

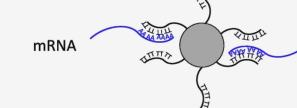
Low Input and Low-Quality Samples

FFPE

## Full-length - PolyA

### Traditional method (KAPA)

**Step 1:** Capture polyA tailed RNA from total RNA using magnetic oligo dT beads



**Step 2:** mRNA fragmentation



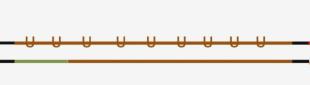
**Step 3:** 1<sup>st</sup> strand synthesis with random primers



**Step 4:** 2<sup>nd</sup> strand synthesis with dUDP



**Step 5:** A-tailing and barcoded adapter ligation



**Step 6:** Amplification (dUTP strand is not amplified)



**Step 7:** Sequencing

Full Length mRNA  
Differential Expression  
Splice Variants  
SNV Detection  
Low Input with Amplification

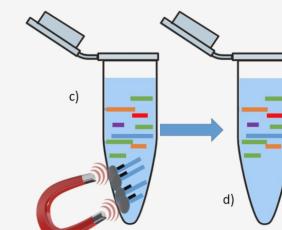
## Ribodepletion



a) Ribo Zero Probe

Ribo Zero Probe

b) Ribo Zero Probe



Adapted from Illumina.com

Full length mRNA + lincRNAs

Differential Expression

Splice Variants

SNVs

FFPE

Throughput

Cost / Data Richness

Image Credit:  
Lexogen Inc

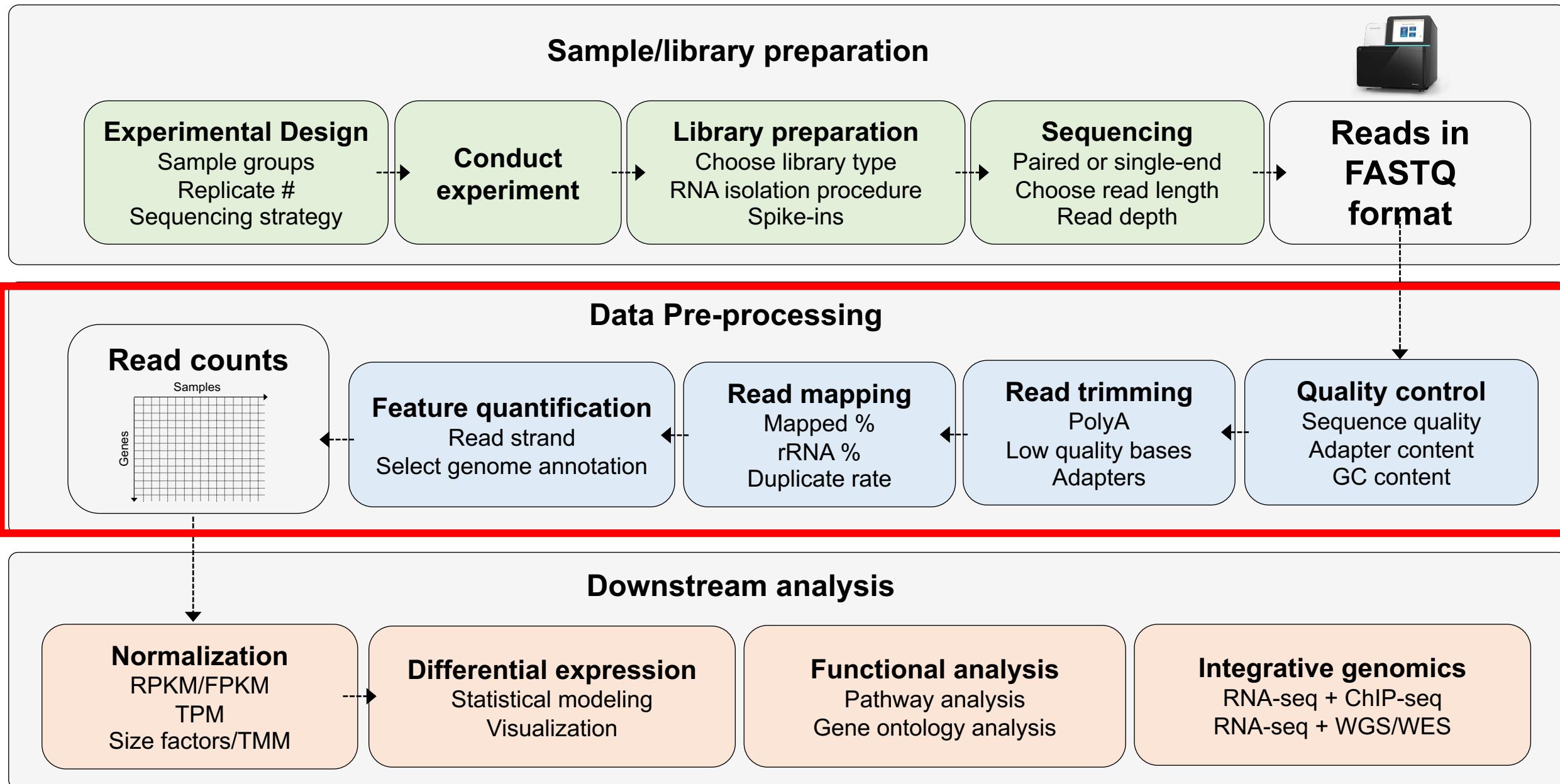
# Sample preparation



- **Be consistent with sample prep**
  - Practice protocol 1<sup>st</sup>, don't get better over course of collecting more samples
- **Minimize batches**
  - 1 batch is ideal, otherwise, smallest number possible
  - MUST randomly distribute samples from experimental conditions across batches
  - Treat each batch EXACTLY the same
- **Collect replicates**
  - Statistics cannot be done on one sample! (statistics is study of populations)
  - The more you collect, the more power you have to discover DEGs
  - Make each replicate as similar as possible e.g. same passage number of cell line
- **Work with your genomics core (they do this a lot)**
- **Pilot experiments can be valuable**

**Spend time to creating a high-quality dataset upfront,  
rather than trying to improve & rescue it later**

# Overview of a typical RNA-seq experiment



# FASTQ files

- 4 lines per record
  - Single file or 1 per direction for paired-end reads (R1.fastq vs. R2.fastq)

## Single FASTQ entry:

- Row 1: Header line (starts with @)
  - Row 2: Base calls
  - Row 3: ‘+’ or line 1
  - Row 4: Base qualities

In bash:

```
> zcat $sample.fastq.gz | head -n 12
```

```
[d41294d@discovery7 ATACseq_6-4-19]$ zcat 648-PKA-Cre-C3Tag_S4_R2_001.fastq.gz | head -n12
@NB501031:325:HK52YBGXB:1:11101:9675:1055 2:N:0:ACCACTGT
CTGGGCAGCTGGGGGGTGGGGTGGGAAAGACACAGGAGCCACACAGAGAGAGACATCCATCCTGTCTCCATTGCT
+
AAAAAEEEEAEFFFFF/E/EE6EEEEEEAEFFFFF//EEEEEAEEEAEFF/E/EEEA<EEEEEEEEE/EAEF
@NB501031:325:HK52YBGXB:1:11101:7613:1055 2:N:0:ACCACTGT
GTGTTAGGGACTTCTCAAGGAAGTTCTATAGATAGAGGCCAGTACTCTTGAGTGACAGGGGGAACATCTGTGTAAA
+
6AA6AEAAE/EEEEEEEEE/AEAEFFFFF6EEEAE/EEEE//EE/AEE/E/<EEEEEE/EAAE<EEAE
@NB501031:325:HK52YBGXB:1:11101:16664:1055 2:N:0:ACCACTGT
ATGCTGGAGTTCTGTGGCACCCACCTCTCAAACACTCATTCATCCATTGAATGGGACATAGGGACAATAGTGAT
+
AAAAAEEAEAEFFFFF/A<AEAEFFFFF/E//EEEEAEFFFFF/EE/EEEEEEA6/EA/EE/</EEE/E
```

Phred Score	Probability of incorrect call	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

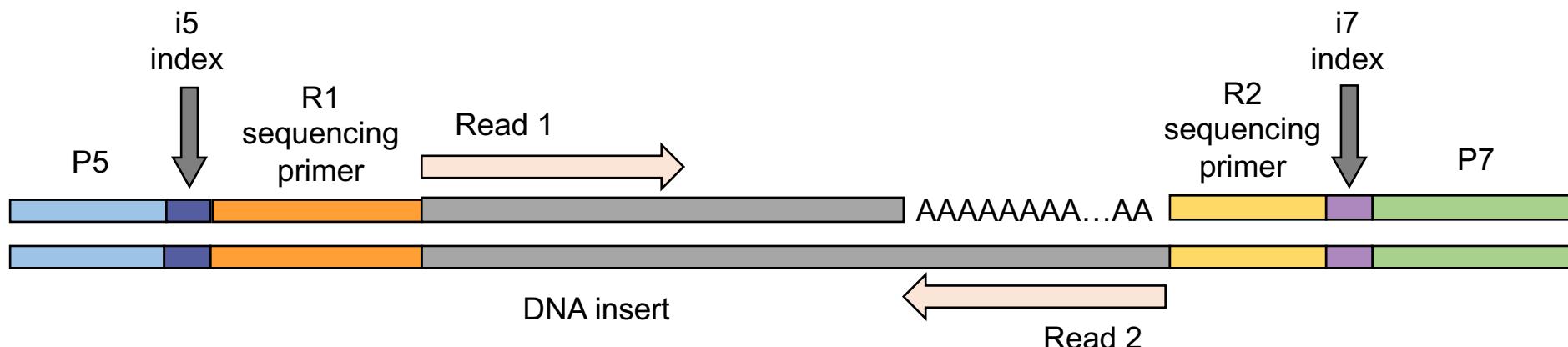
# Read trimming (pre-processing)

- Process of ‘cleaning’ NGS reads to improve **speed & quality** of downstream analysis
- NGS reads contain sequences not wanted for downstream steps:
- Trimming algorithms search reads for specified sequences and remove matches based on user specified behavior

Commonly trimmed sequences:

- Sequencing adapters
- PolyA tails
- Low quality bases
- ‘N’ base calls (no call)

Typical Illumina RNA-seq library



## Legend:

DNA insert

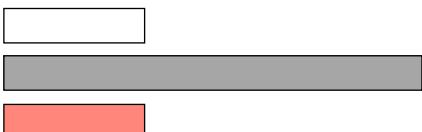


Adapter

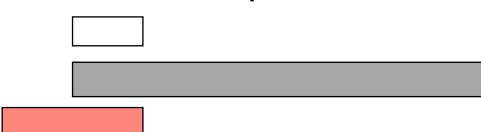


Trimmed

Full adapter at start of read 1

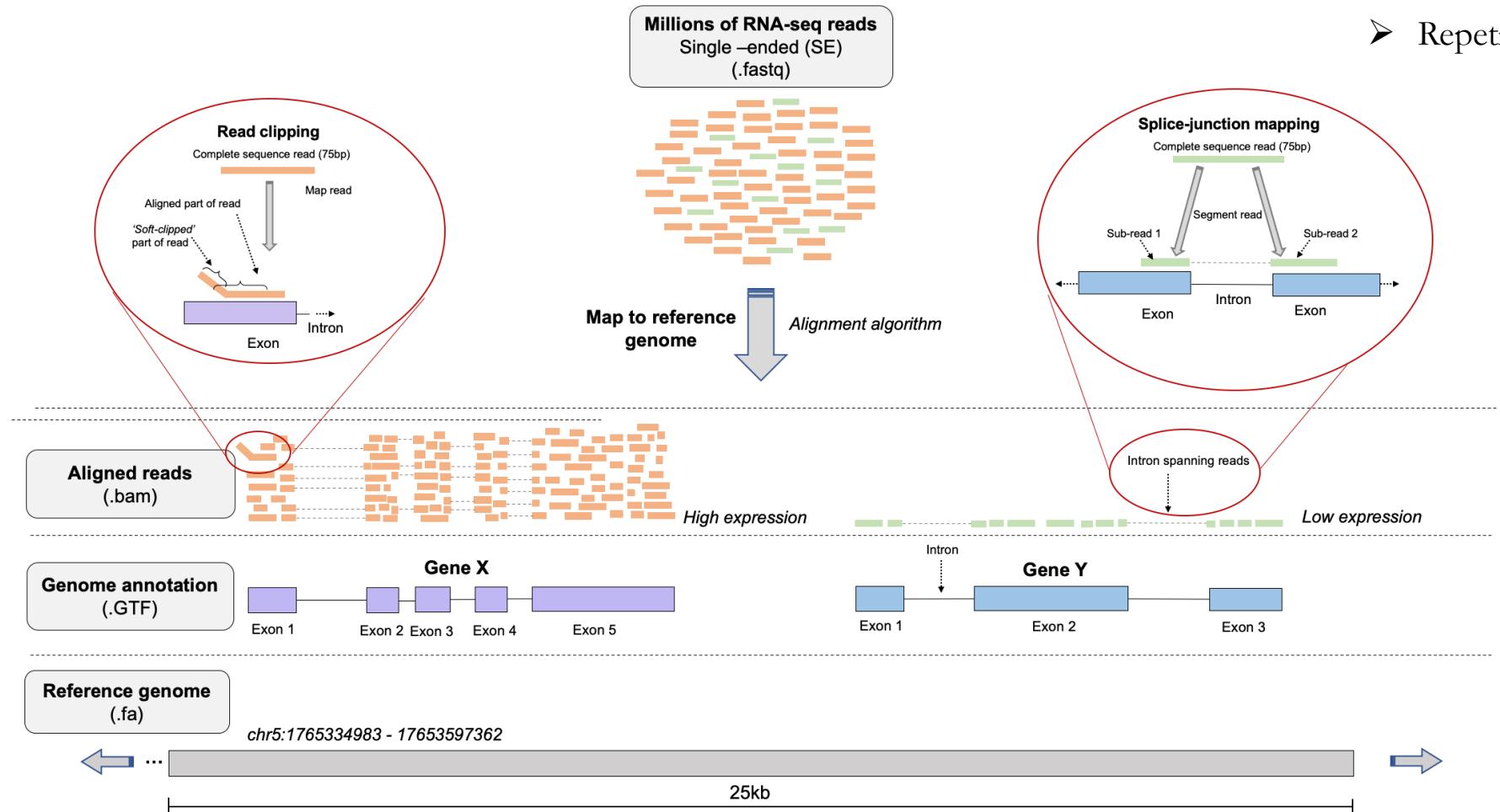


Partial adapter at end of read 2



# Read mapping to a reference genome

- Reads can be aligned to a reference genome for well studied organisms. e.g. hg38, mm10
- Where is the most likely origin for a read in a reference genome?

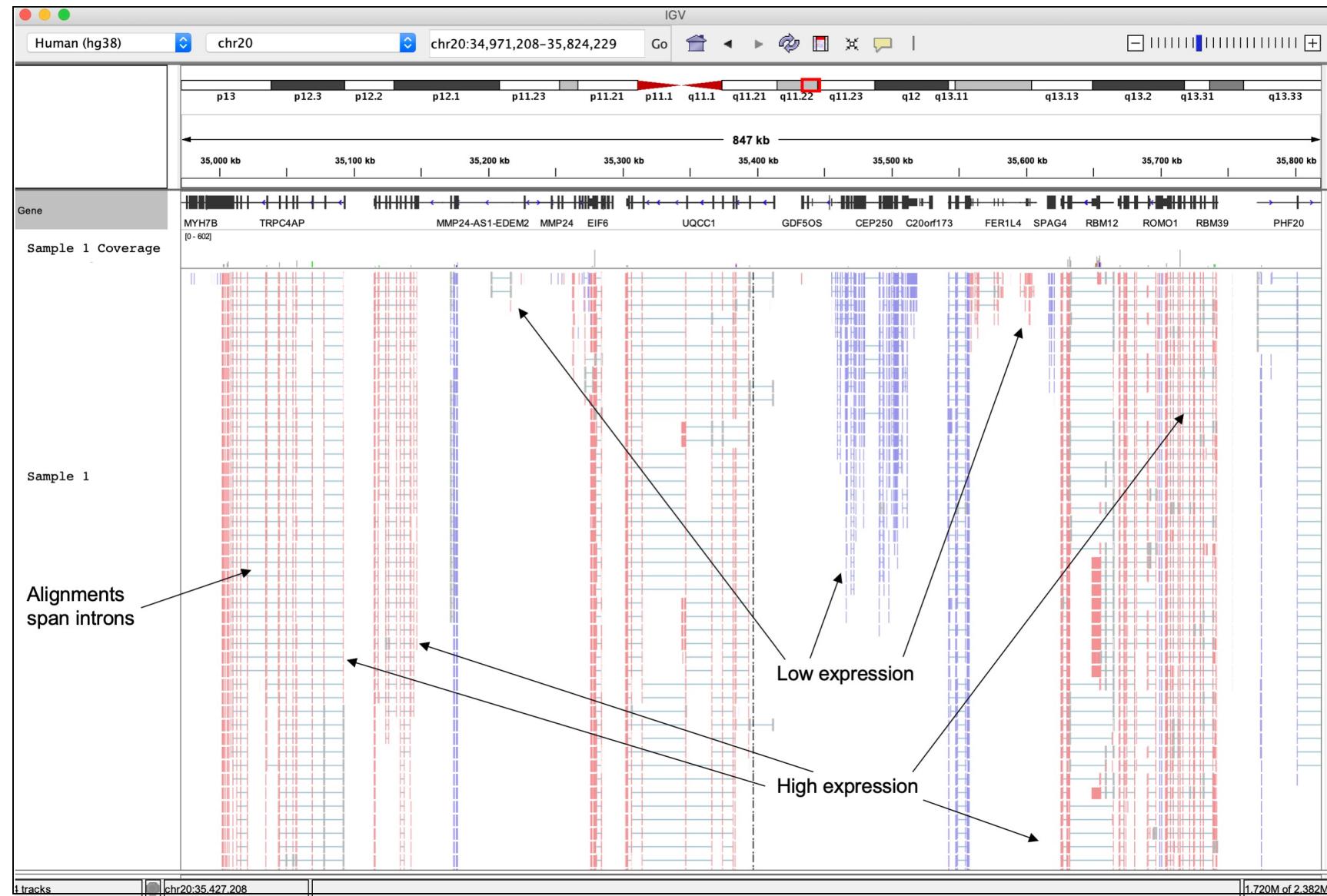


## Challenges:

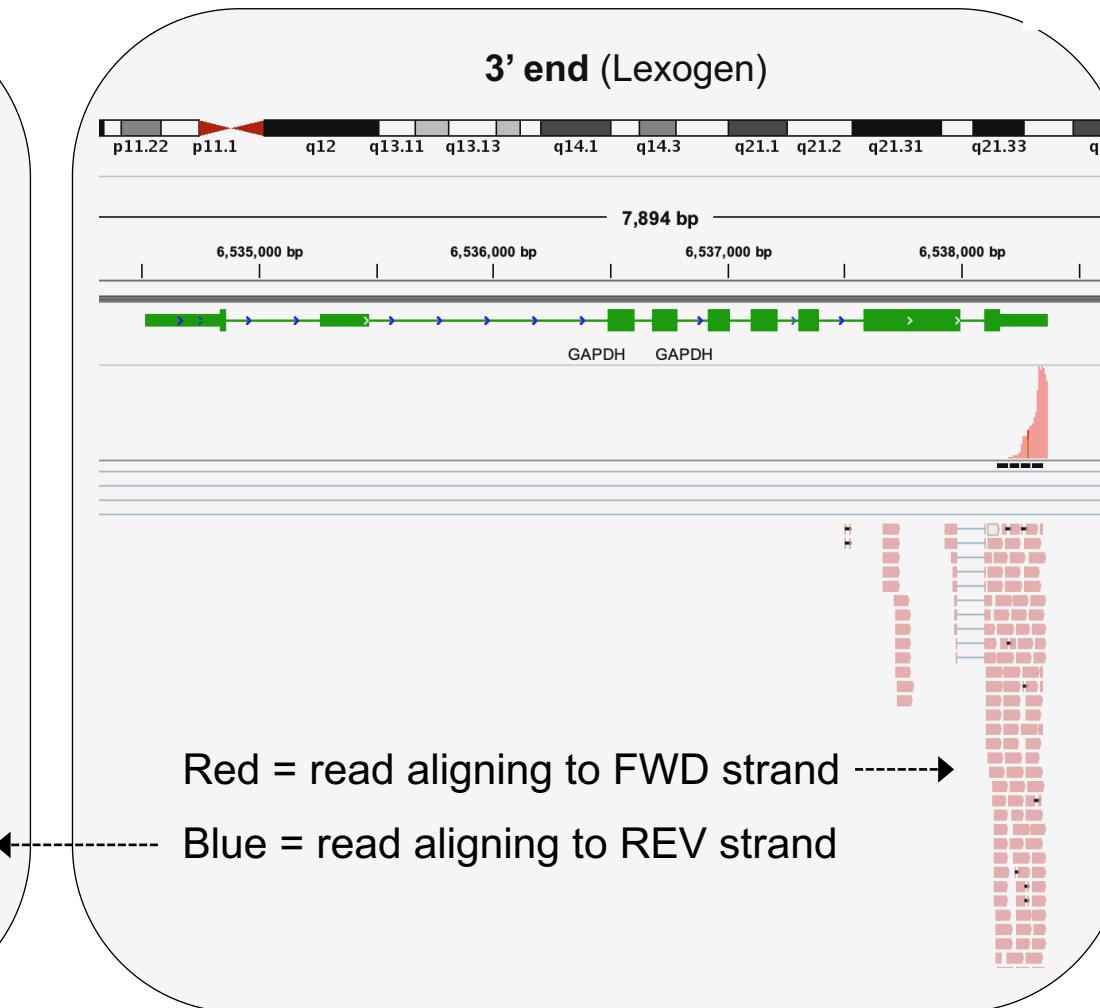
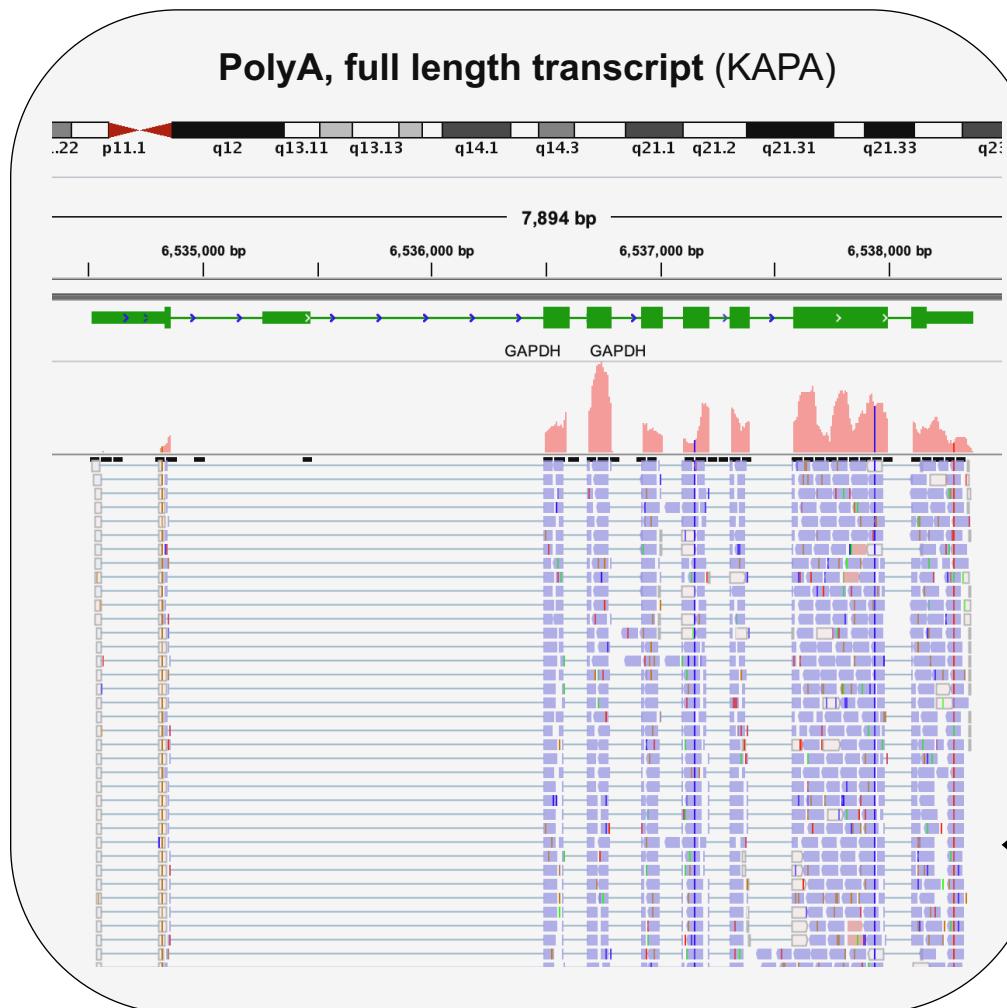
- Enormous search space
- Computationally intensive
- Genome vs transcriptome
- Genetic variation
- Repetitive sequences

# Aligned reads in browser

- Real data example using Integrative Genomic Viewer (IGV)
  - This is just 1 sample



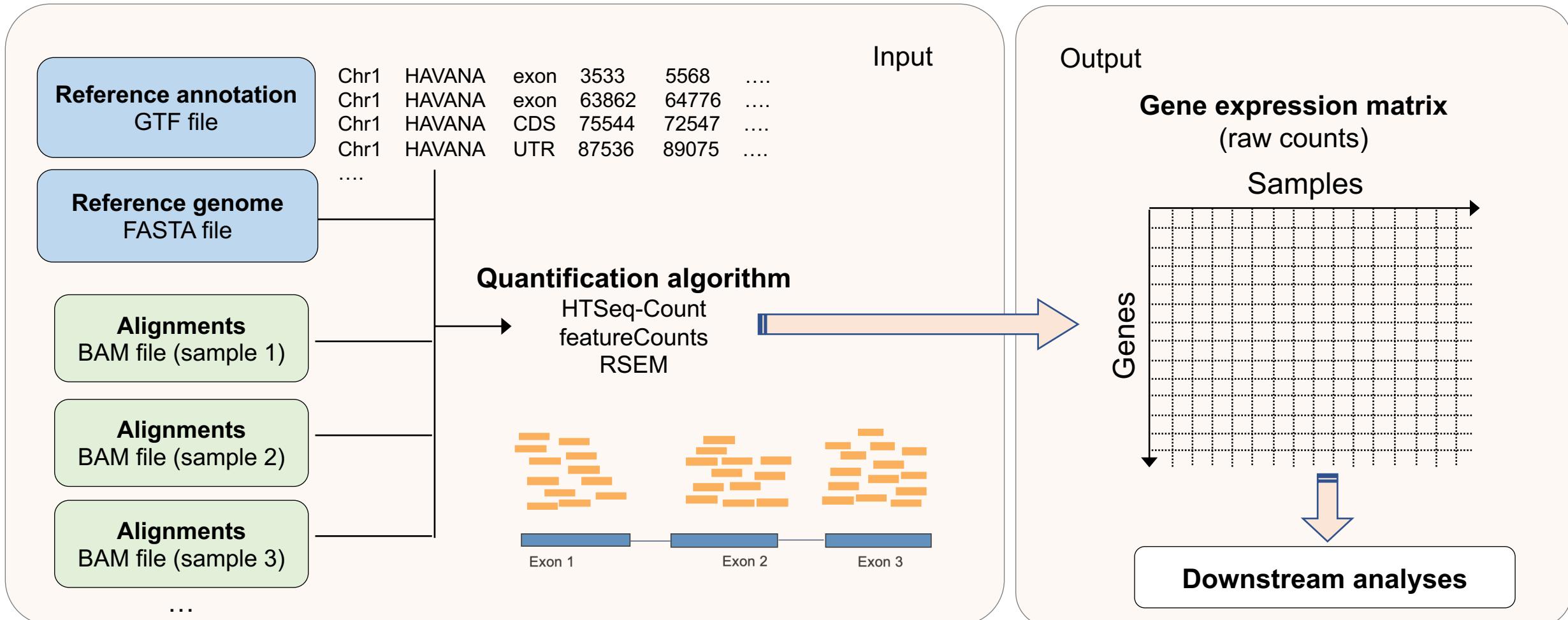
# Data from different library types looks different



Quality control, analysis pipelines, and possible hypotheses therefore inherently differ

# Quantification (read counting)

- Count reads at each genomic position for desired feature of interest (usually exons for RNA-seq)



# Replicates



- Arguably more important than read depth or length for DE

**How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?**

---

NICHOLAS J. SCHURCH,<sup>1,6</sup> PIETÀ SCHOFIELD,<sup>1,2,6</sup> MAREK GIERLIŃSKI,<sup>1,2,6</sup> CHRISTIAN COLE,<sup>1,6</sup> ALEXANDER SHERSTNEV,<sup>1,6</sup> VIJENDER SINGH,<sup>2</sup> NICOLA WROBEL,<sup>3</sup> KARIM GHARBI,<sup>3</sup> GORDON G. SIMPSON,<sup>4</sup> TOM OWEN-HUGHES,<sup>2</sup> MARK BLAXTER,<sup>3</sup> and GEOFFREY J. BARTON<sup>1,2,5</sup>

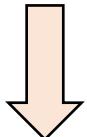
<sup>1</sup>Division of Computational Biology, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom

<sup>2</sup>Division of Gene Regulation and Expression, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom

<sup>3</sup>Edinburgh Genomics, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

<sup>4</sup>Division of Plant Sciences, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom

<sup>5</sup>Division of Biological Chemistry and Drug Discovery, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom



*"With 3 biological replicates, 9 of the 11 tools evaluated found only 20%–40% of the significantly differentially expressed (SDE) genes"*

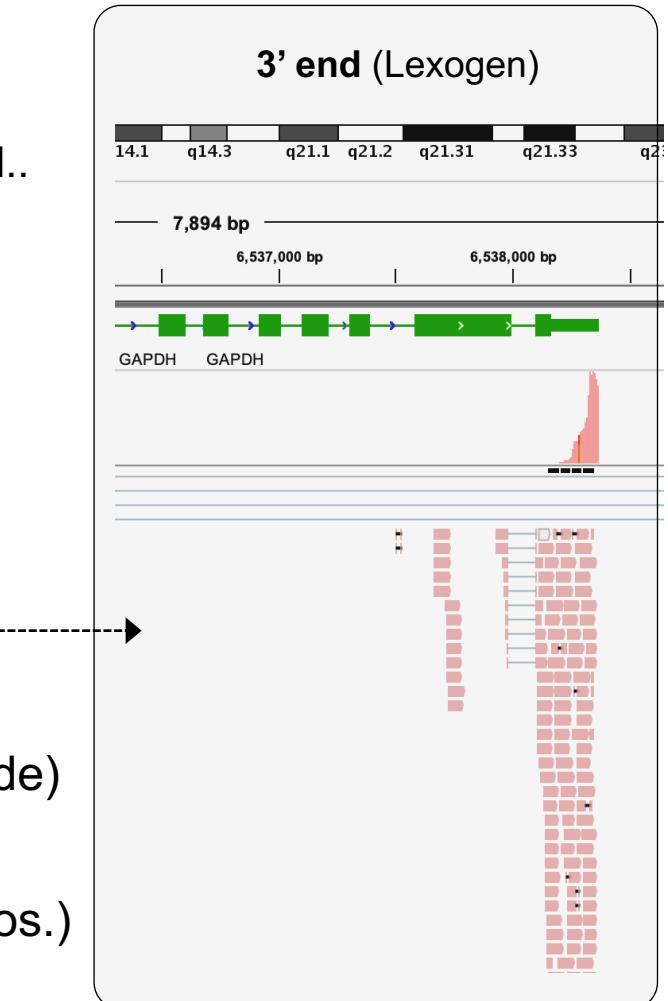
- Schurch *et al*, RNA, 2016

- Suggested minimum no. of replicates should = 6
- More heterogeneity = more replicates needed (e.g. human tissues vs. cultured cell lines)

# Sequencing depth (or coverage?)



- ‘Coverage’ doesn’t have much meaning for transcriptome data
  - We are less concerned with how many times we cover a specific locus w/ a read..
  - More meaningful for WES, WGS, Genome Assembly
- Generally total of 10-30 million reads for DGE of eukaryotic genomes
- Some species require many fewer than this
- Technology also affects required read number (3'-end data needs fewer)
- Checking saturation can help you assess if you’ve sequenced enough (next slide)
- Try to avoid generating libraries of differing complexity (vastly different reads nos.)

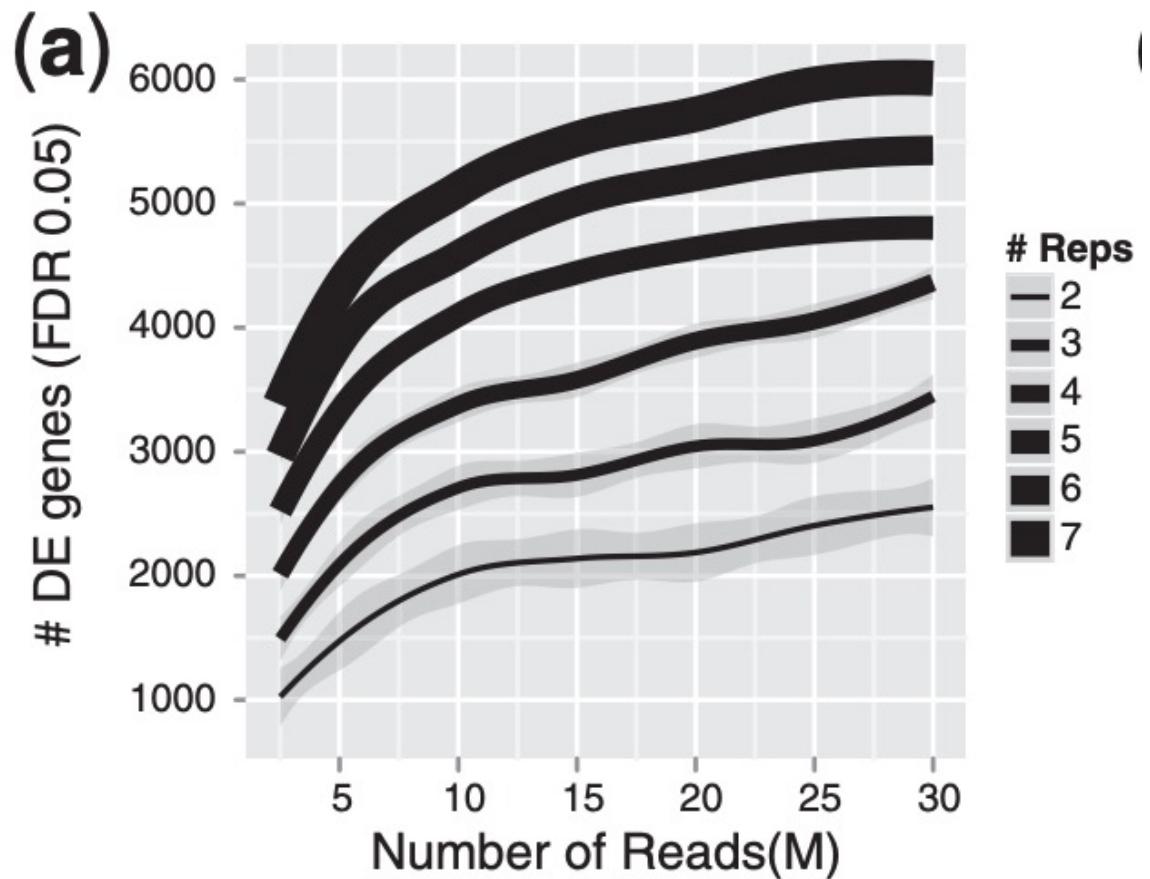


# Depth vs. Replicates



## Which is more important?

- No. of DEGs increases w/ replicate no.
- Diminishing returns after 10-15M reads (for this dataset)
- Additional replicates are more valuable than sequencing really deeply (for DE analysis)



Liu *et al*, 2014, *Bioinformatics*

# Replicates & Statistical power

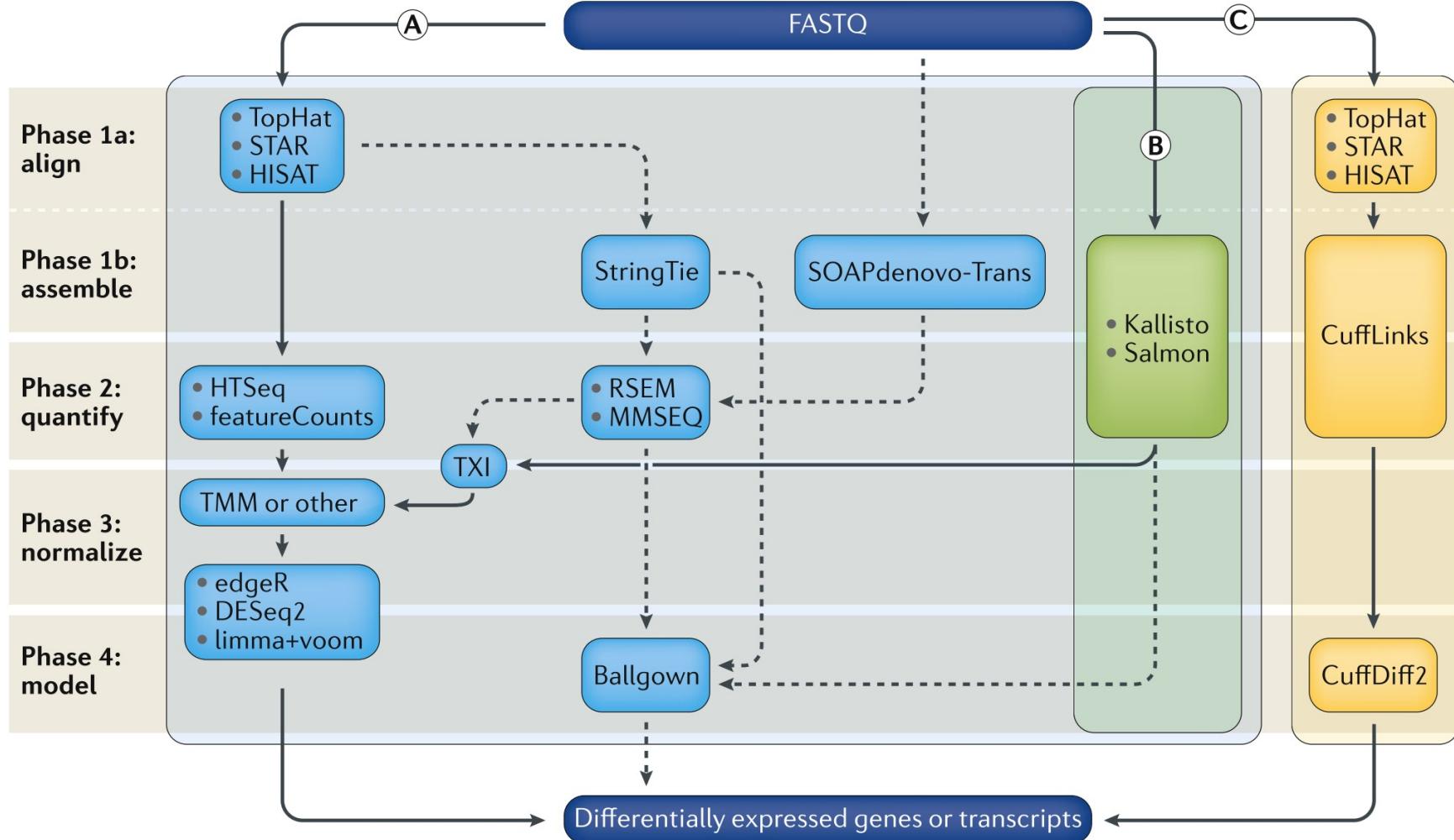


- More replicates improve power to detect smaller fold changes (effect size)
- More replicates improve power at the same sequencing depth

**Table 1** Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

Effect size (fold change)	Replicates per group		
	3	5	10
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

# Differential expression workflow(s)



- Several tools available for each step
- Each have various strengths, weaknesses, applications

# Differential analysis methods



**Table I:** Software packages for detecting differential expression

Method	Version	Reference	Normalization <sup>a</sup>	Read count distribution assumption	Differential expression test
edgeR	3.0.8	[4]	TMM/Upper quartile/RLE (DESeq-like)/None (all scaling factors are set to be one)	Negative binomial distribution	Exact test
DESeq	1.10.1	[5]	DESeq sizeFactors	Negative binomial distribution	Exact test
baySeq	1.12.0	[6]	Scaling factors ( <u>quantile</u> /TMM/total)	Negative binomial distribution	Assesses the posterior probabilities of models for differentially and non-differentially expressed genes via empirical Bayesian methods and then compares these posterior likelihoods
NOIseq	1.1.4	[7]	<u>RPKM</u> /TMM/Upper quartile	Nonparametric method	Contrasts fold changes and absolute differences within a condition to determine the null distribution and then compares the observed differences to this null
SAMseq (samr)	2.0	[8]	SAMseq specialized method based on the mean read count over the null features of the data set	Nonparametric method	Wilcoxon rank statistic and a resampling strategy
Limma	3.14.4	[9]	TMM	voom transformation of counts	Empirical Bayes method
Cuffdiff 2 (Cufflinks)	2.0.2-beta	[10]	<u>Geometric</u> (DESeq-like)/quartile/classic-fpkm	Beta negative binomial distribution	t-test
EBSeq	1.1.7	[11]	DESeq median normalization	Negative binomial distribution	Evaluates the posterior probability of differentially and non-differentially expressed entities (genes or isoforms) via empirical Bayesian methods

Seyednasrollah *et al*, *Brief. In Bioinfo.* 2013

- Non-parametric methods can perform better for large sample sizes
- DESeq2 & Limma perform well across many circumstances

# Differential analysis methods



## scientific reports

OPEN

### Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis

Luis A. Corchete<sup>1,2,3,4,5</sup>, Elizabeta A. Rojas<sup>1,2,3</sup>, Diego Alonso-López<sup>2</sup>, Javier De Las Rivas<sup>2,3</sup>, Norma C. Gutiérrez<sup>1,2,3,5</sup> & Francisco J. Burguillo<sup>4</sup>

Soneson and Delorenzi *BMC Bioinformatics* 2013, **14**:91  
<http://www.biomedcentral.com/1471-2105/14/91>



RESEARCH ARTICLE

Open Access

### A comparison of methods for differential expression analysis of RNA-seq data

Charlotte Soneson<sup>1\*</sup> and Mauro Delorenzi<sup>1,2</sup>

➤ See 'Useful links' folder on GitHub



Rapaport et al. *Genome Biology* 2013, **14**:R95  
<http://genomebiology.com/2013/14/9/R95>



METHOD

Open Access

### Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data

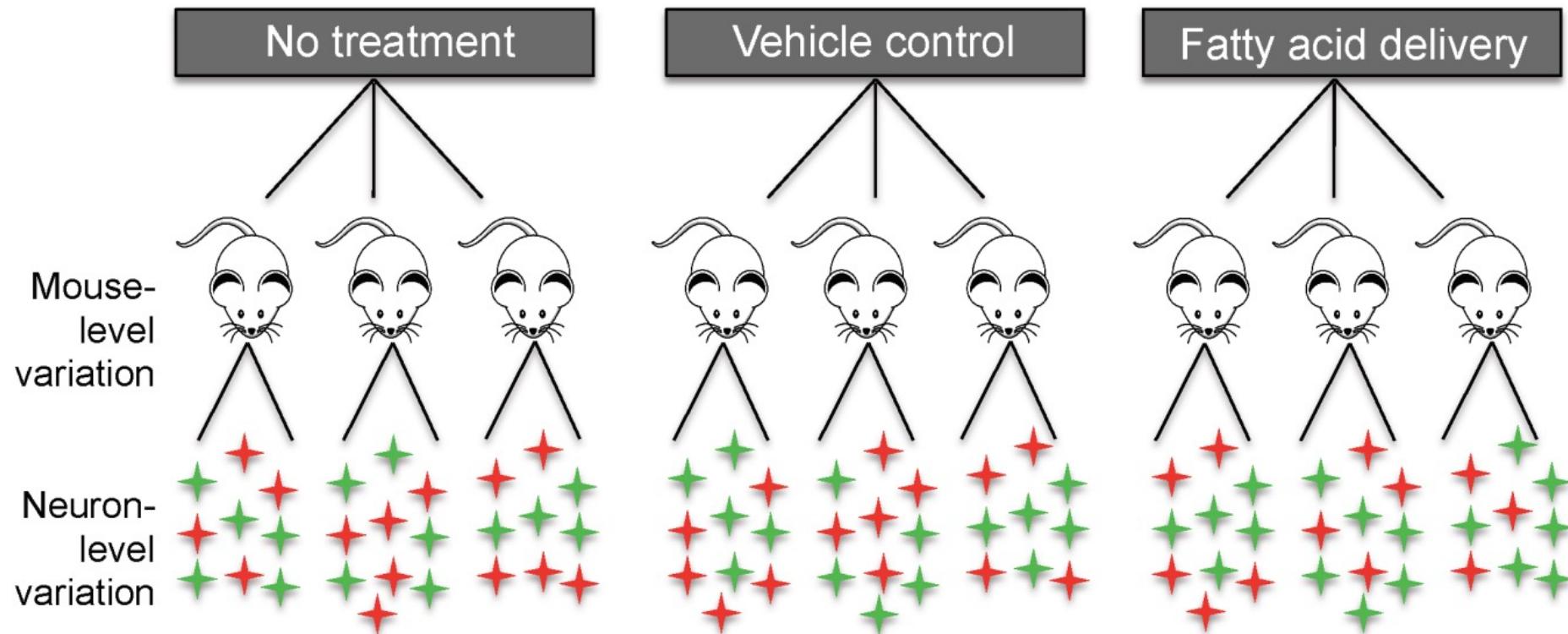
Franck Rapaport<sup>1</sup>, Raya Khanin<sup>1</sup>, Yupu Liang<sup>1</sup>, Mono Pirun<sup>1</sup>, Azra Krek<sup>1</sup>, Paul Zumbo<sup>2,3</sup>, Christopher E Mason<sup>2,3</sup>, Nicholas D Soccia<sup>1</sup> and Doron Betel<sup>3,4\*</sup>

### Comparison of software packages for detecting differential expression in RNA-seq studies

Fatemeh Seyednasrollah, Asta Laiho and Laura L. Elo

Submitted: 20th August 2013; Received (in revised form): 9th October 2013

# Hierarchical study designs



Moen *et al*, 2016, *Plos One*

- Within-sample correlation must be accounted for
- Applies to time-course experiments also
- Limma can handle hierarchical study designs



# Take home messages

- Work with your genomics core to design an experiment that will address your hypothesis
- You should always know and understand the type of library used to generate your data
- If you don't perform data pre-processing, you should understand the basics steps involved
- Replicates are critical for differential expression analysis
- Various DE analysis tools are available. Experimental design can affect which one is most suitable

# Questions?