# BIO634: Next generation Sequencing 2

## Transcriptome and Biological Interpretation

8th-9th May, 2023

Deepak Kumar Tanwar

URPP Evolution in Action
Embedded bioinformatician

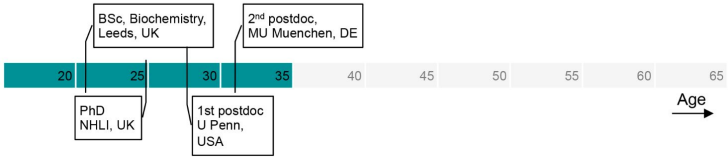

@d_k_tanwar


deepak.tanwar@evolution.uzh.ch

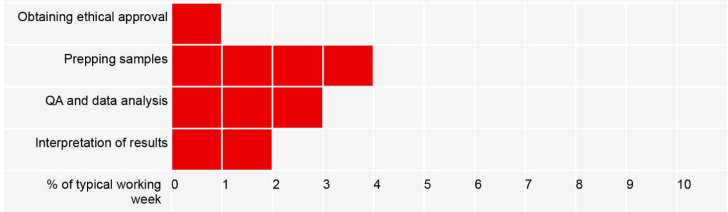
## **Leon** (bioinformatics user)

Leon is on his second postdoctoral fellowship, working on quorum sensing in bacteria. “I’m using a combination of transcriptomics, proteomics and metabolomics to understand these pathogenic changes better” he explains. “I end up with big spreadsheets of protein or gene IDs and I’m trying to piece together which signaling pathways are involved in flipping to the pathogenic state”. He has been on an introductory Unix course but is much more comfortable with GUIs than with the command line. “I just have a visual brain”, he says.

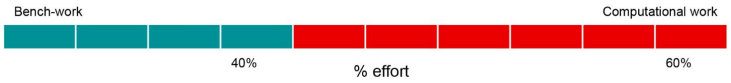
### Career timeline



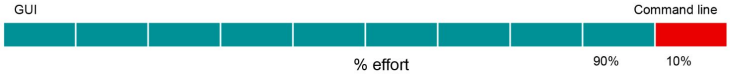

### Typical activities




### Distribution of time between bench-work and computational work




### Preference for using GUI vs command line




### Drivers

- Understanding what makes a usually harmless bacterium pathogenic in the lungs of people with cystic fibrosis

### Goals

- QA of -omics data
- Statistical analysis of data
- Data integration and pathway analysis

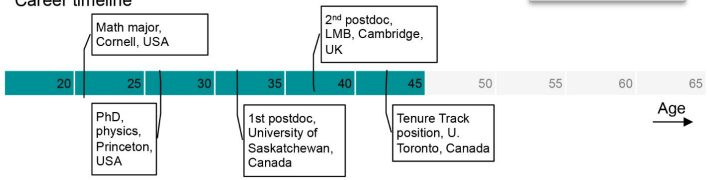
### Pain points

- Lack of access to departmental compute farm
- Sporadic to non-existent access to bioinformatics support

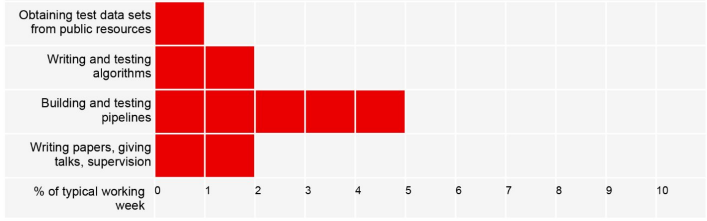
## **Martha** (bioinformatics scientist)

Martha is a senior bioinformatician in an international structural genomics consortium. Her biggest project is on predicting the functions of proteins whose structures have just been solved; she’s building a structure-to-function prediction pipeline for the project. This is funded partly by the NIH and partly through industrial funding. She also has a fascination for predicting structure and usually has a student or two working on structural prediction projects.

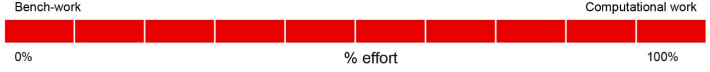
### Career timeline



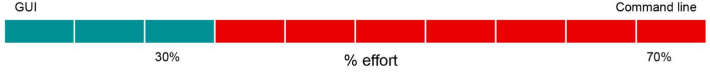

### Typical activities




### Distribution of time between bench work and computational work




### Preference using for GUI vs command line




### Drivers

- Understanding the relationship between sequence, structure and function
- Application to target discovery and validation

### Goals

- Create a structure-to-function pipeline for molecular biologists
- Predict structures de novo from models of similar, solved structures

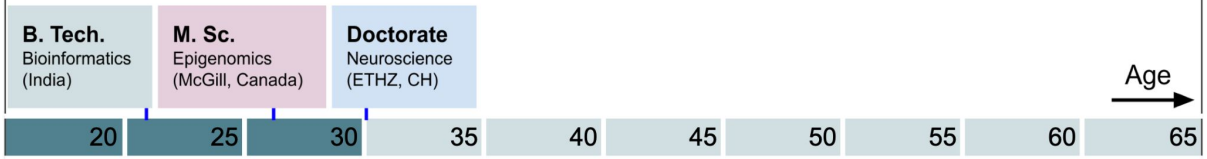
### Pain points

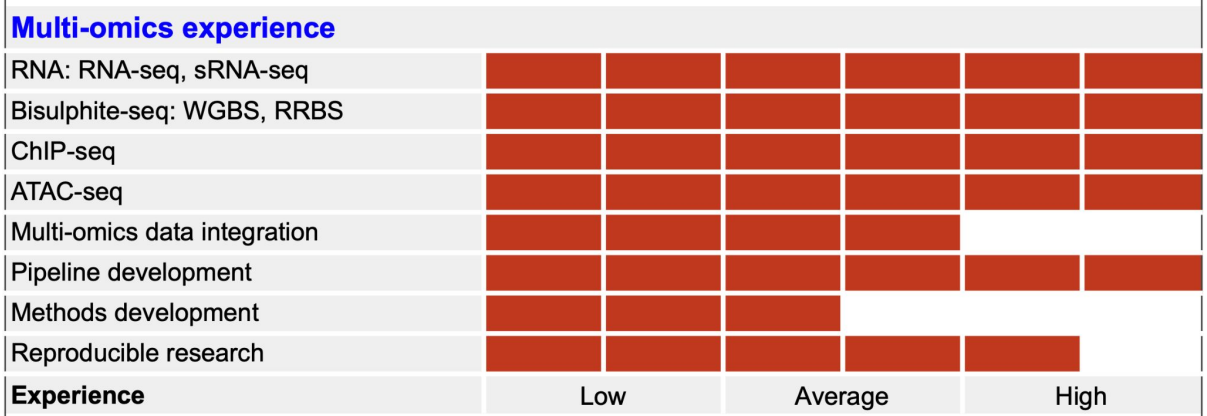
- Sometimes the guys in the lab expect her to fix their computers for them
- Finding students and more senior staff with adequate math

**Deepak Tanwar** (embedded bioinformatics scientist)

Deepak is a bioinformatician at the URPP Evolution in Action. Deepak has a strong background in multi-omics research and has expertise in reproducible data analysis, benchmarking, and methods development.

Education timeline



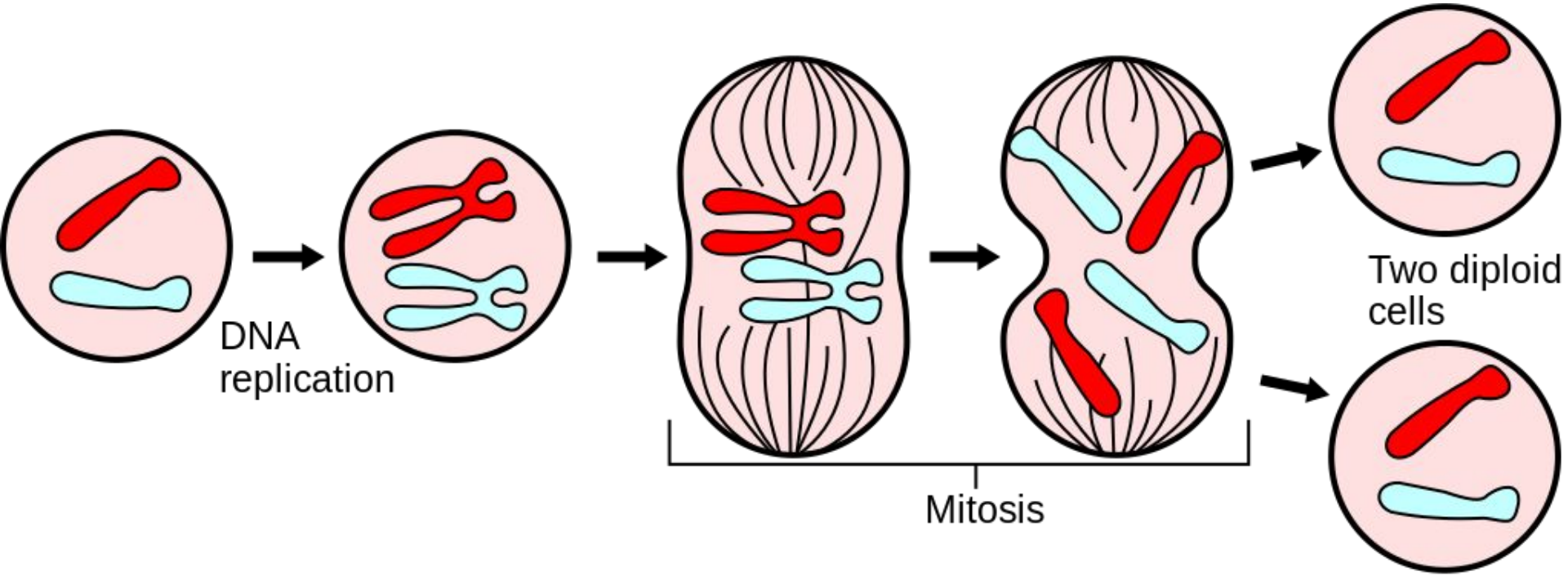




Distribution of time between bench work and computational work

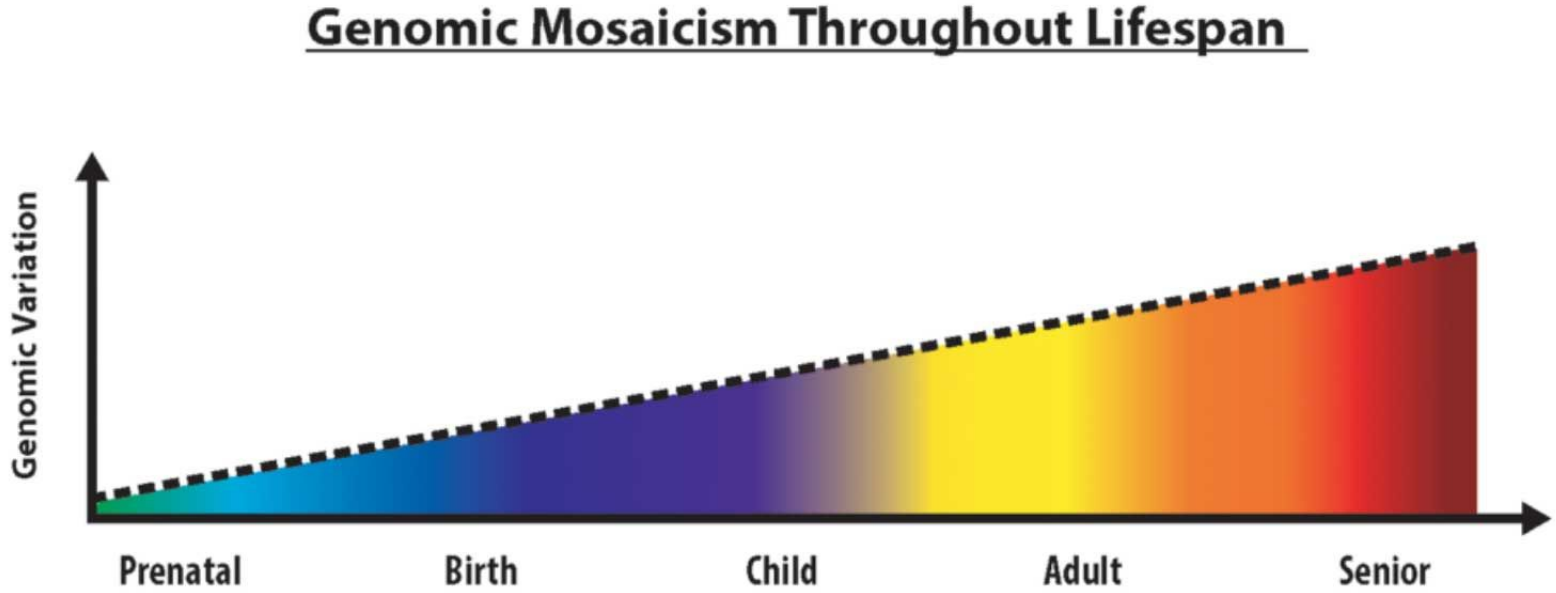
```
cd ~/Desktop

git clone https://github.com/urppeia/bio634.git

cd bio634

./login.sh          |          ./win_login.sh
```

# Introduction

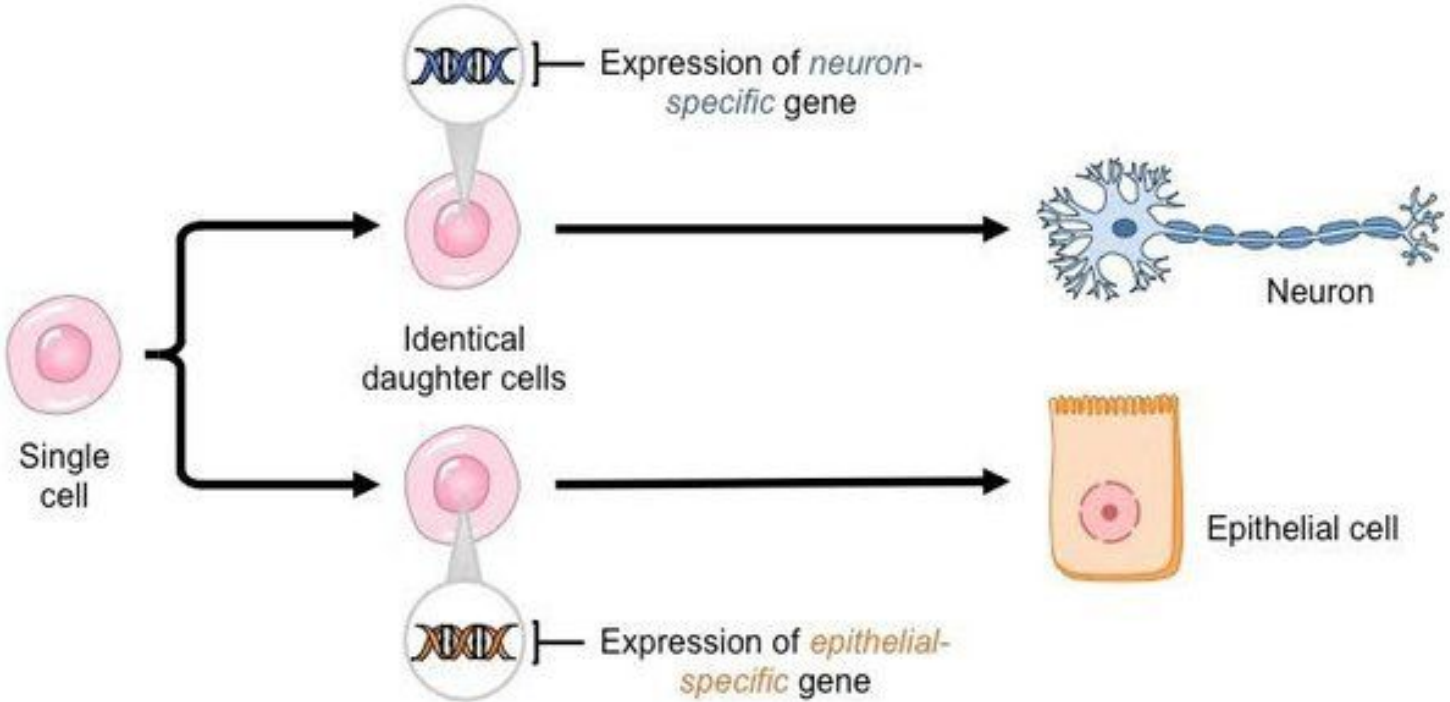
# All cells have same DNA

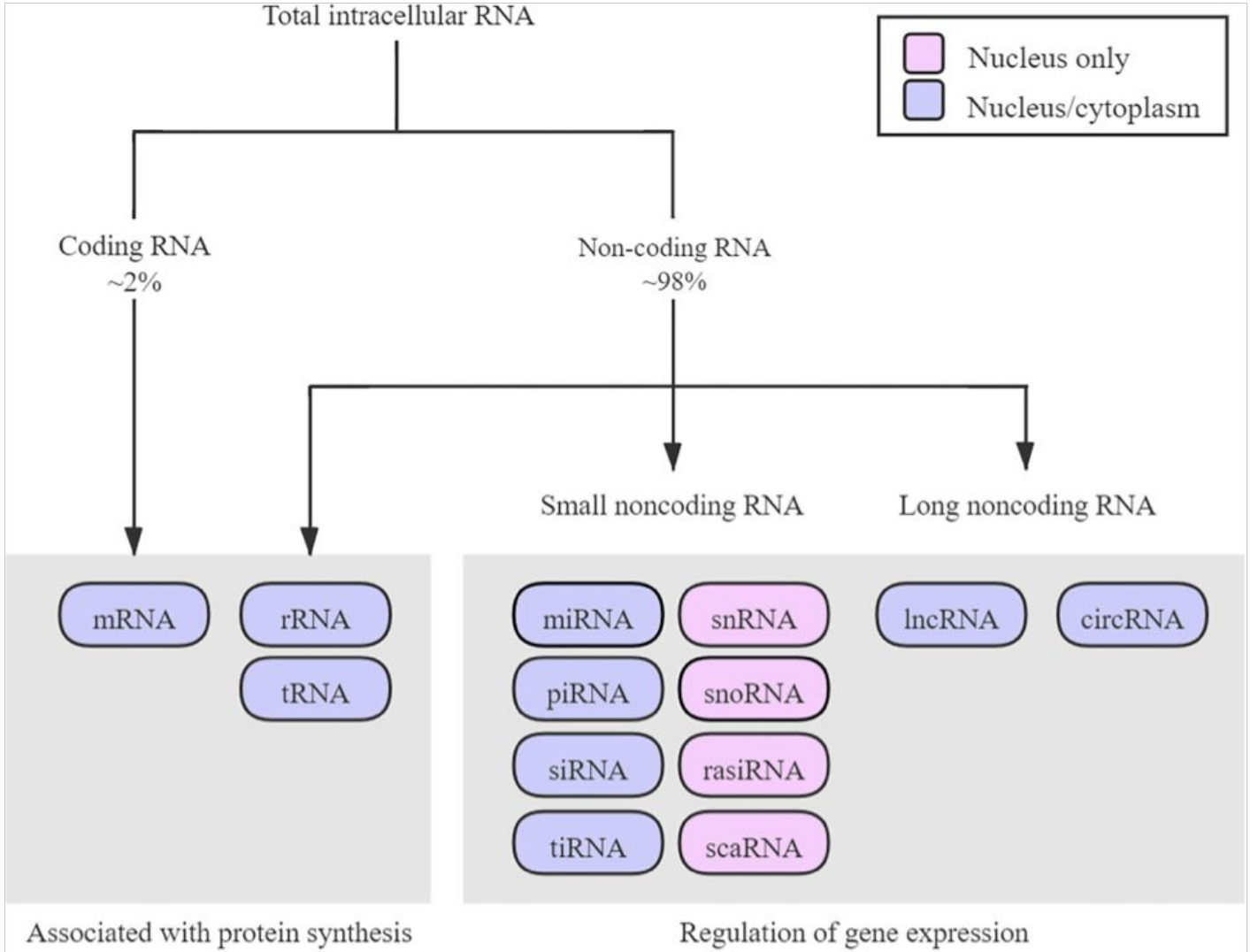


# Not all our cells have the same DNA

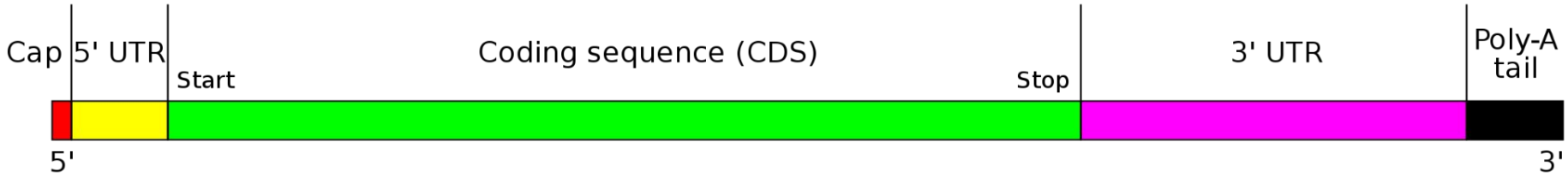


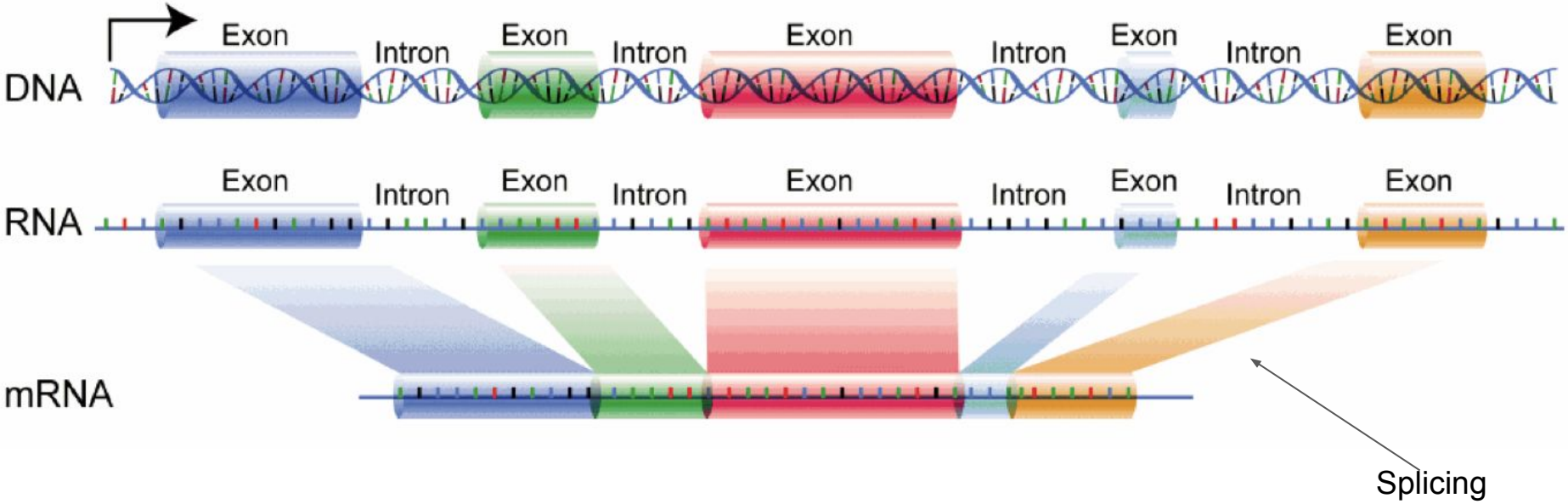
# Specific RNA expression make specific cells




Total intracellular RNA
Nucleus only
Nucleus/cytoplasm
Coding RNA
~2%
Non-coding RNA
~98%
Small noncoding RNA
Long noncoding RNA
mRNA
rRNA
tRNA
miRNA
snRNA
lncRNA
circRNA
piRNA
snoRNA
siRNA
rasiRNA
tiRNA
scaRNA
Associated with protein synthesis
Regulation of gene expression

**The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)**



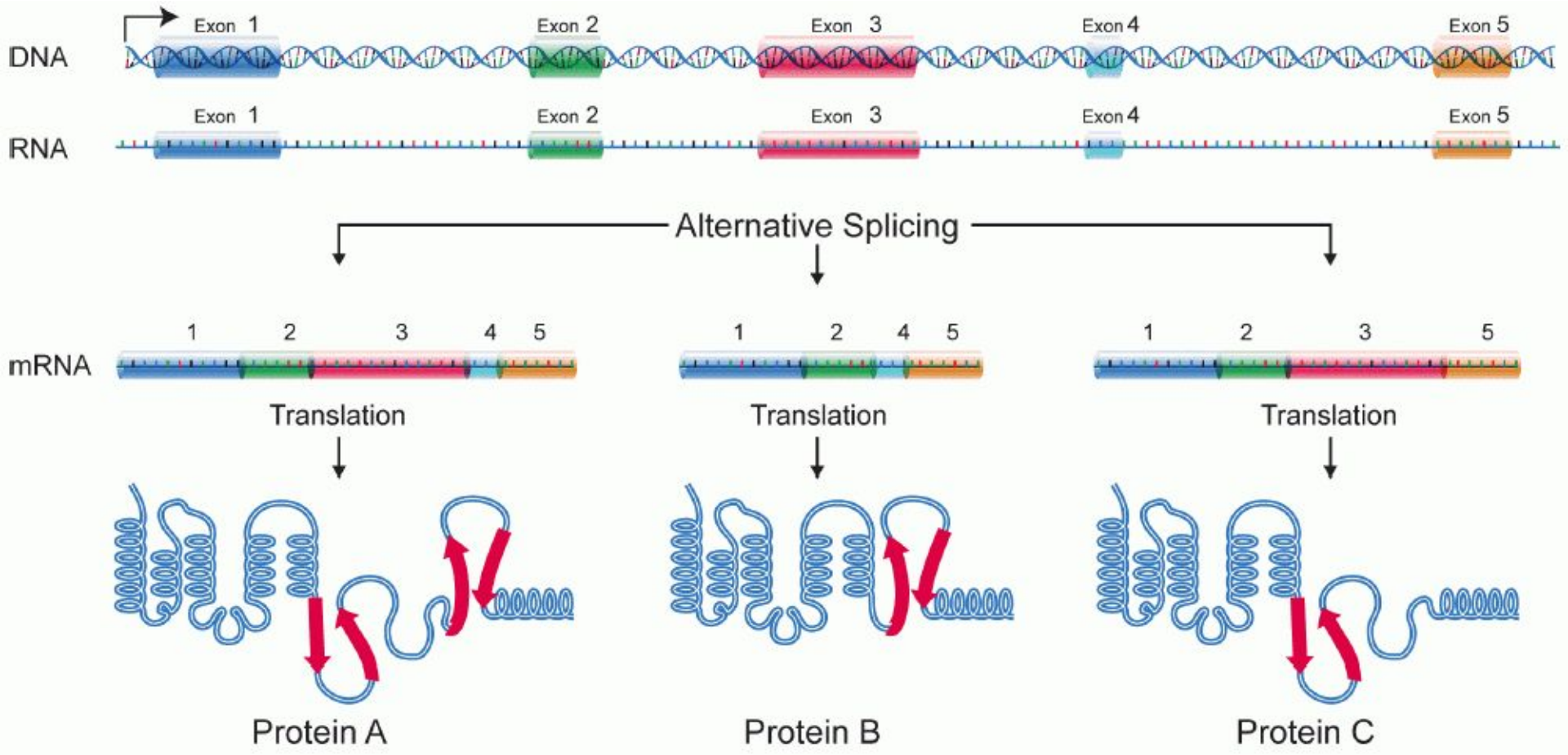

DNA
Exon
Intron
Exon
Intron
Exon
Intron
Exon
Intron
Exon
RNA
Exon
Intron
Exon
Intron
Exon
Intron
Exon
Intron
Exon
mRNA
Splicing

Cell Genomics

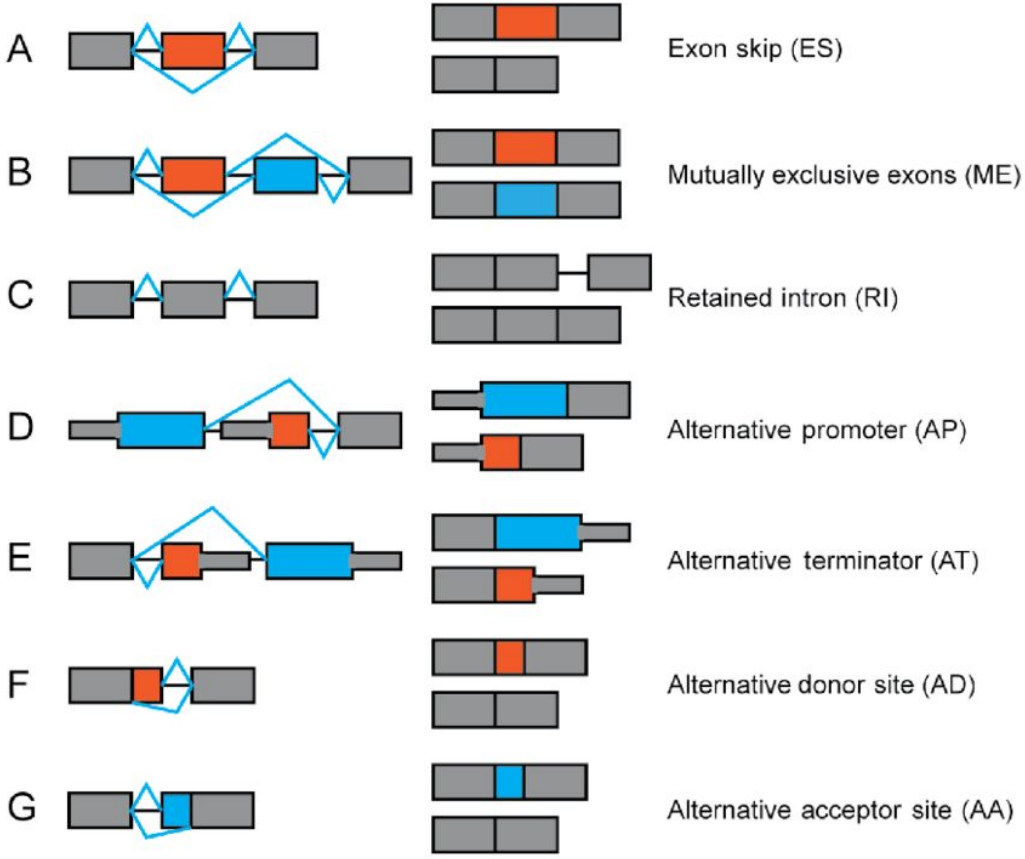
Perspective

# Not all exons are protein coding: Addressing a common misconception

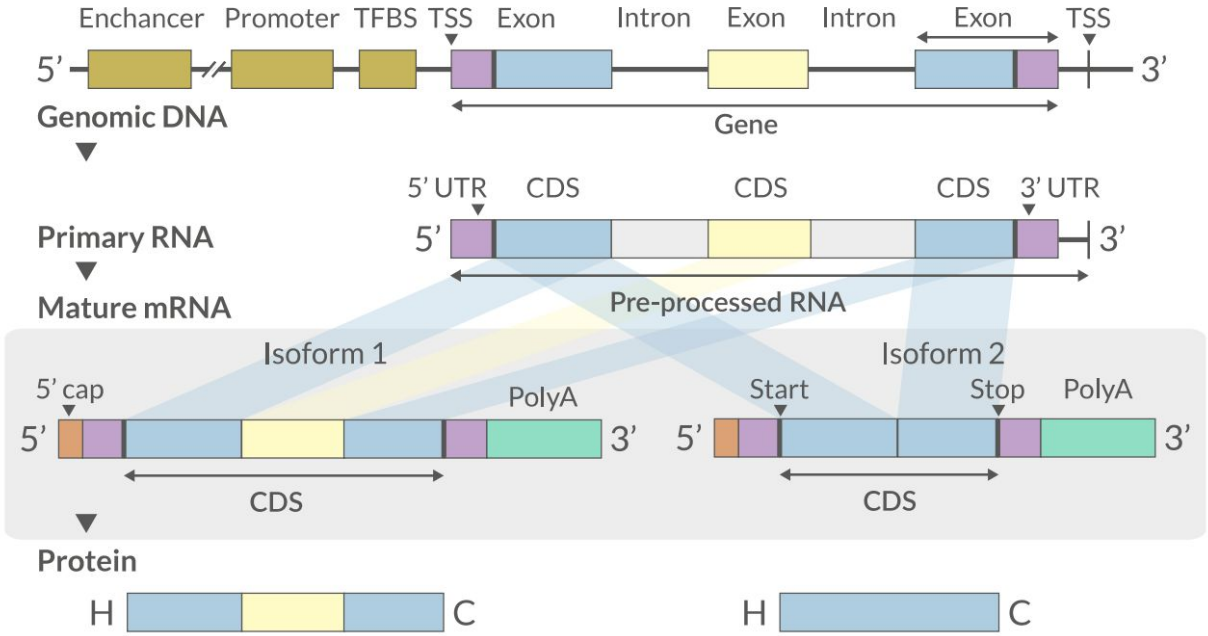
Julie L. Aspden,[1,2,3] Edward W.J. Wallace,[4] and Nicola Whiffin[5,6,*]


DNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
RNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Alternative Splicing
mRNA
1 2 3 4 5
1 2 4 5
1 2 3 5
Translation
Translation
Translation
Protein A
Protein B
Protein C

# One gene, many different mRNAs



# Summary so far




- The transcriptome is spatially and temporally dynamic
- Data comes from functional units (coding regions)
- Only a tiny fraction of the genome

# Why should we sequence RNA?

# Experiment design

# Designing the right experiment

Clear objectives

Amenable to statistical analysis

Reproducible

# How much should one sequence?

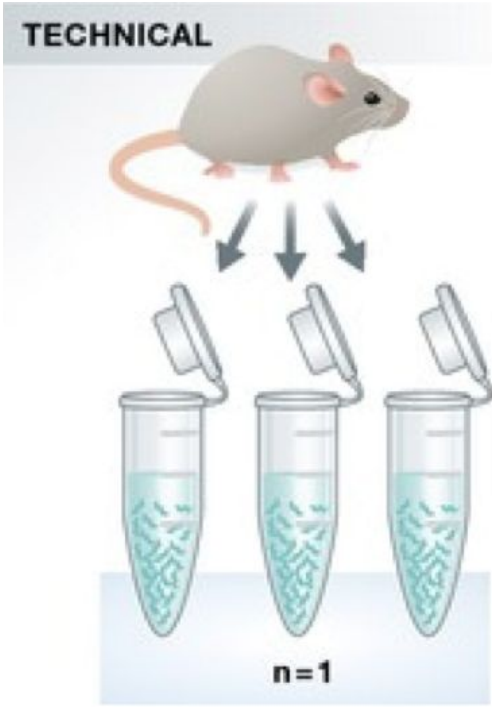
The coverage is defined as:

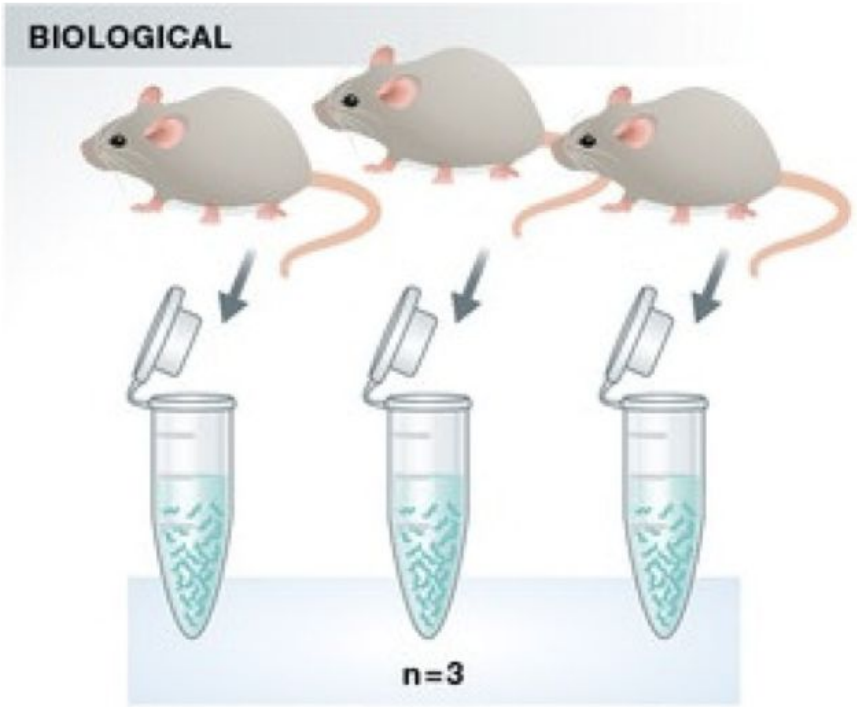
$$\frac{Read\ Length \times Number\ of\ Reads}{Length\ of\ Target\ Sequence}$$

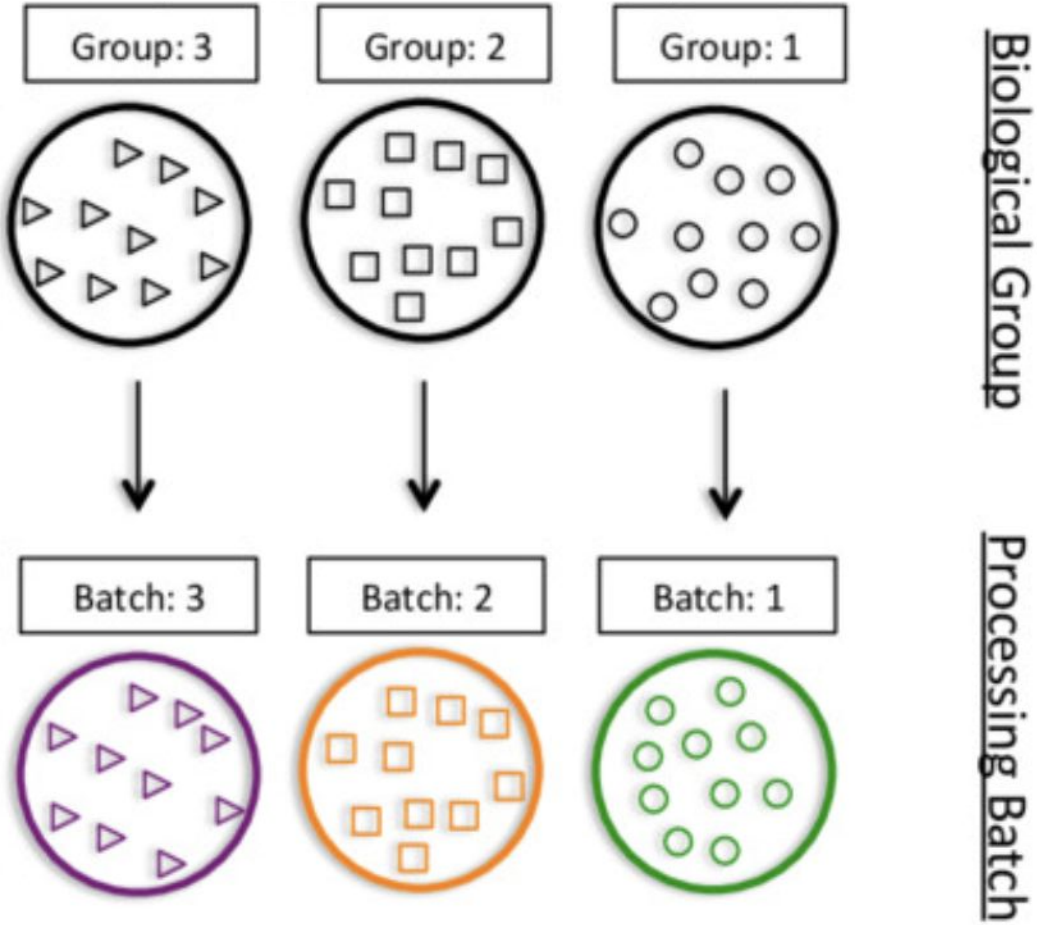
The amount of sequencing needed for a given sample is determined by the goals of the experiment and the nature of the RNA sample.

- For a general view of differential expression: 5–25 million reads per sample
- For alternative splicing and lowly expressed genes: 30–60 million reads per sample.
- In-depth view of the transcriptome/assemble new transcripts: 100–200 million reads
- Targeted RNA expression requires fewer reads.
- miRNA-Seq or Small RNA Analysis require even fewer reads.

# Importance of replicates

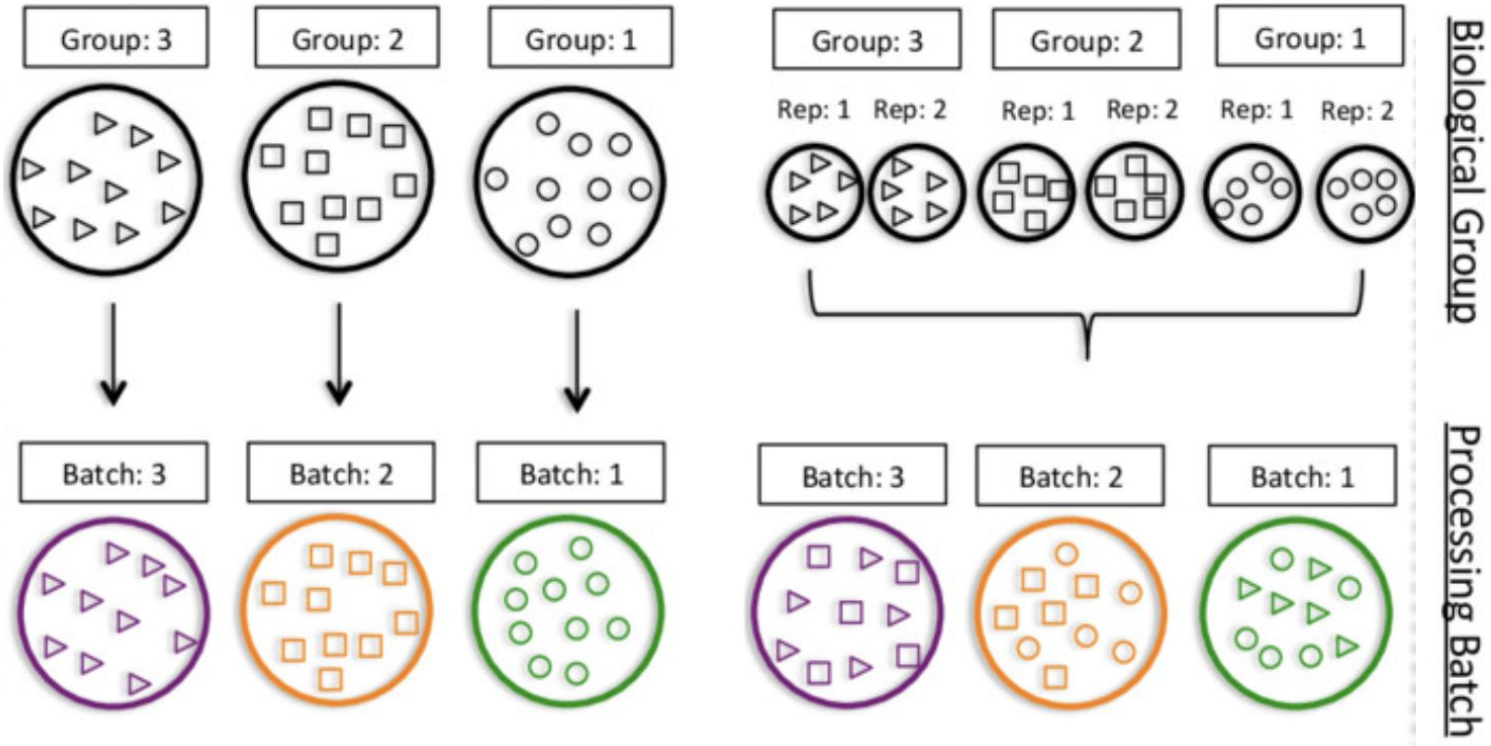





**What changes would you make here to make the experimental design more optimal?**

# Batch effects?



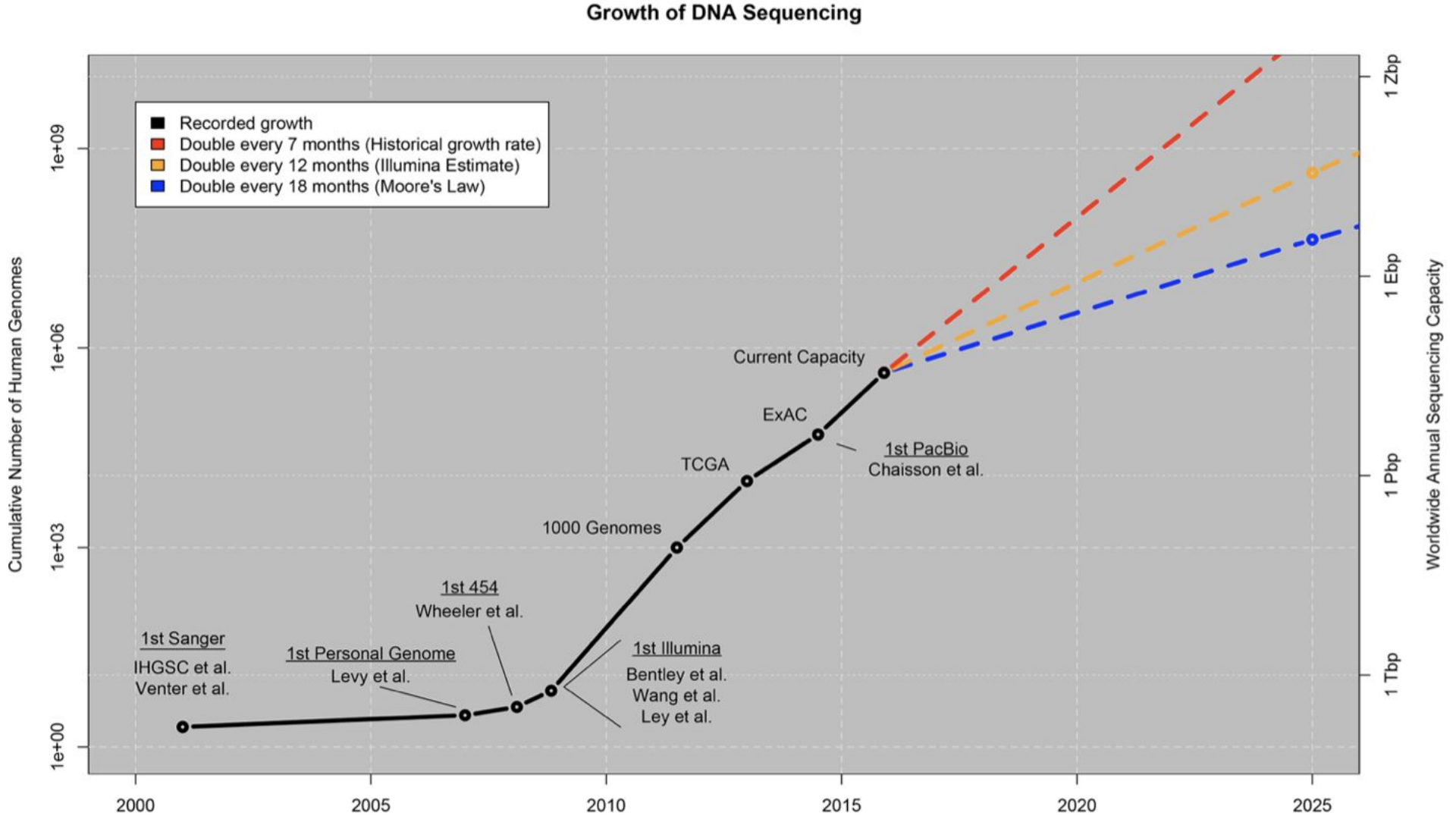

**Write down everything.**

# The Importance of Experimental Design

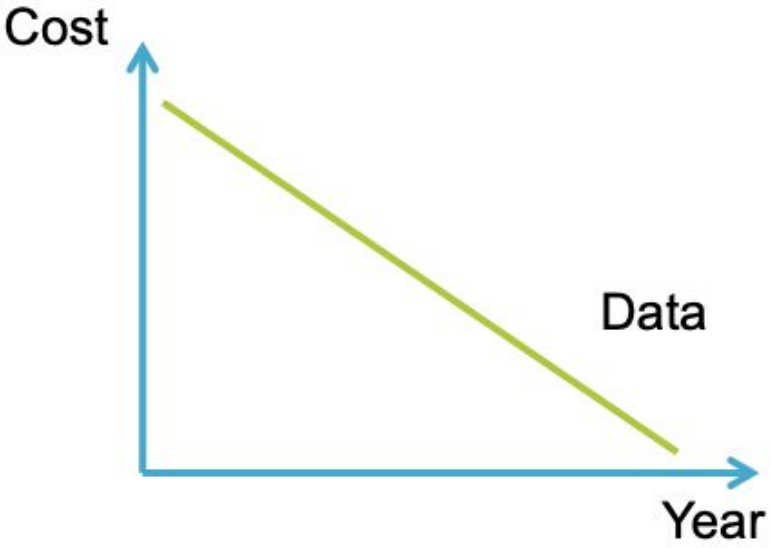

Let's see if the subject responds to magnetic stimuli… ADMINISTER THE MAGNET!

Interesting…there seems to be a significant decrease in heart rate. The fish must sense the magnetic field.

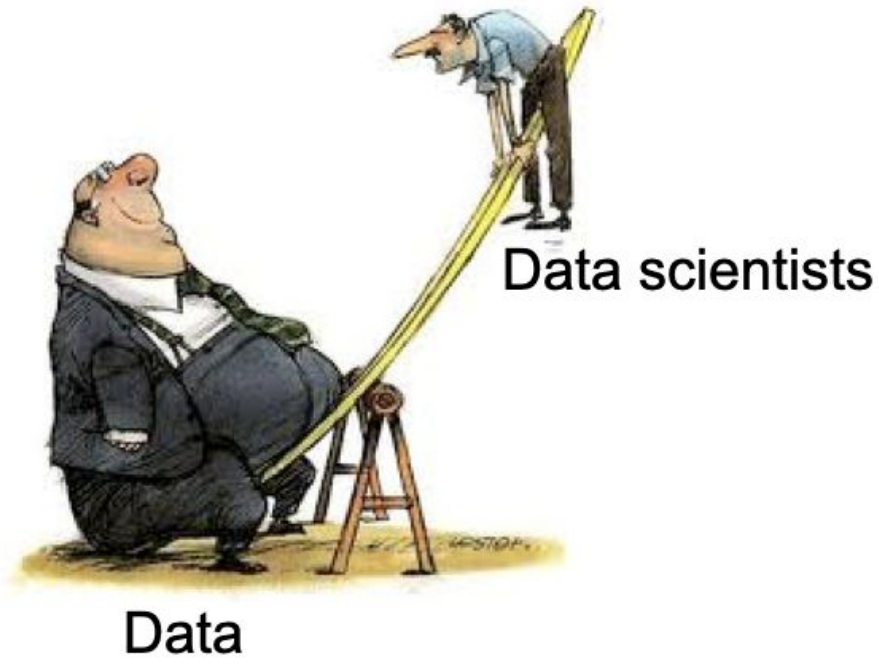
# High-throughput sequencing


Growth of DNA Sequencing
Recorded growth
Double every 7 months (Historical growth rate)
Double every 12 months (Illumina Estimate)
Double every 18 months (Moore's Law)
Cumulative Number of Human Genomes
1e+00
1e+03
1e+06
1e+09
Worldwide Annual Sequencing Capacity
1 Tbp
1 Pbp
1 Ebp
1 Zbp
1st Sanger
IHGSC et al.
Venter et al.
1st Personal Genome
Levy et al.
1st 454
Wheeler et al.
1st Illumina
Bentley et al.
Wang et al.
Ley et al.
1000 Genomes
TCGA
ExAC
1st PacBio
Chaisson et al.
Current Capacity
2000
2005
2010
2015
2020
2025

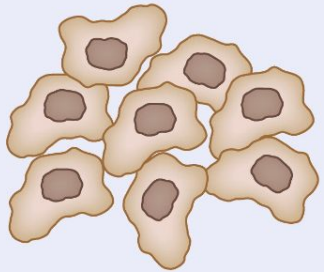
# Growing burden on Bioinformaticians



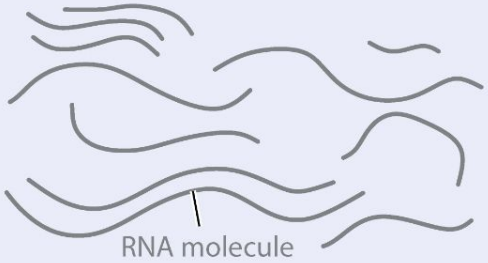

"Per base"



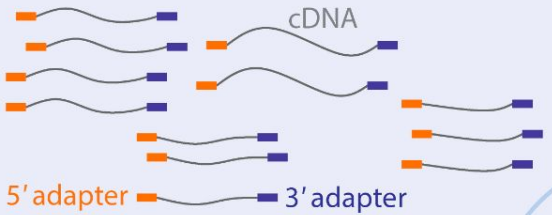
## Sample of interest



## Total RNA extraction and target enrichment



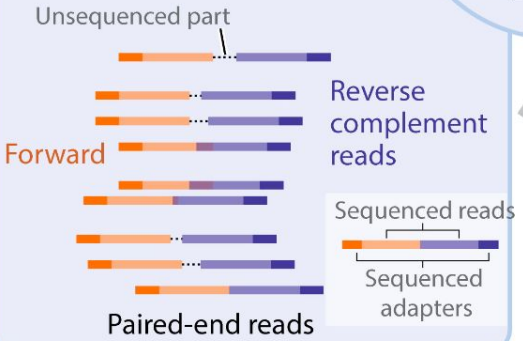

## Fragment, reverse transcribe, ligate adapters, amplify



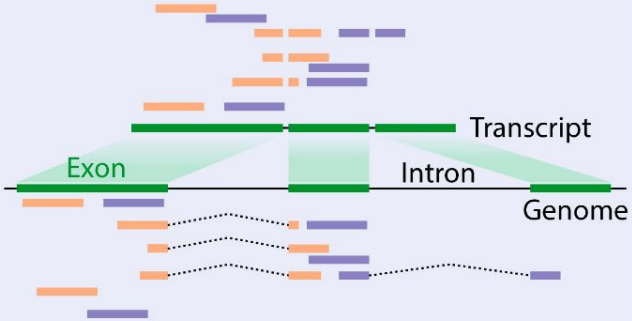

cDNA library

## Sequencing machine

## Sequencing



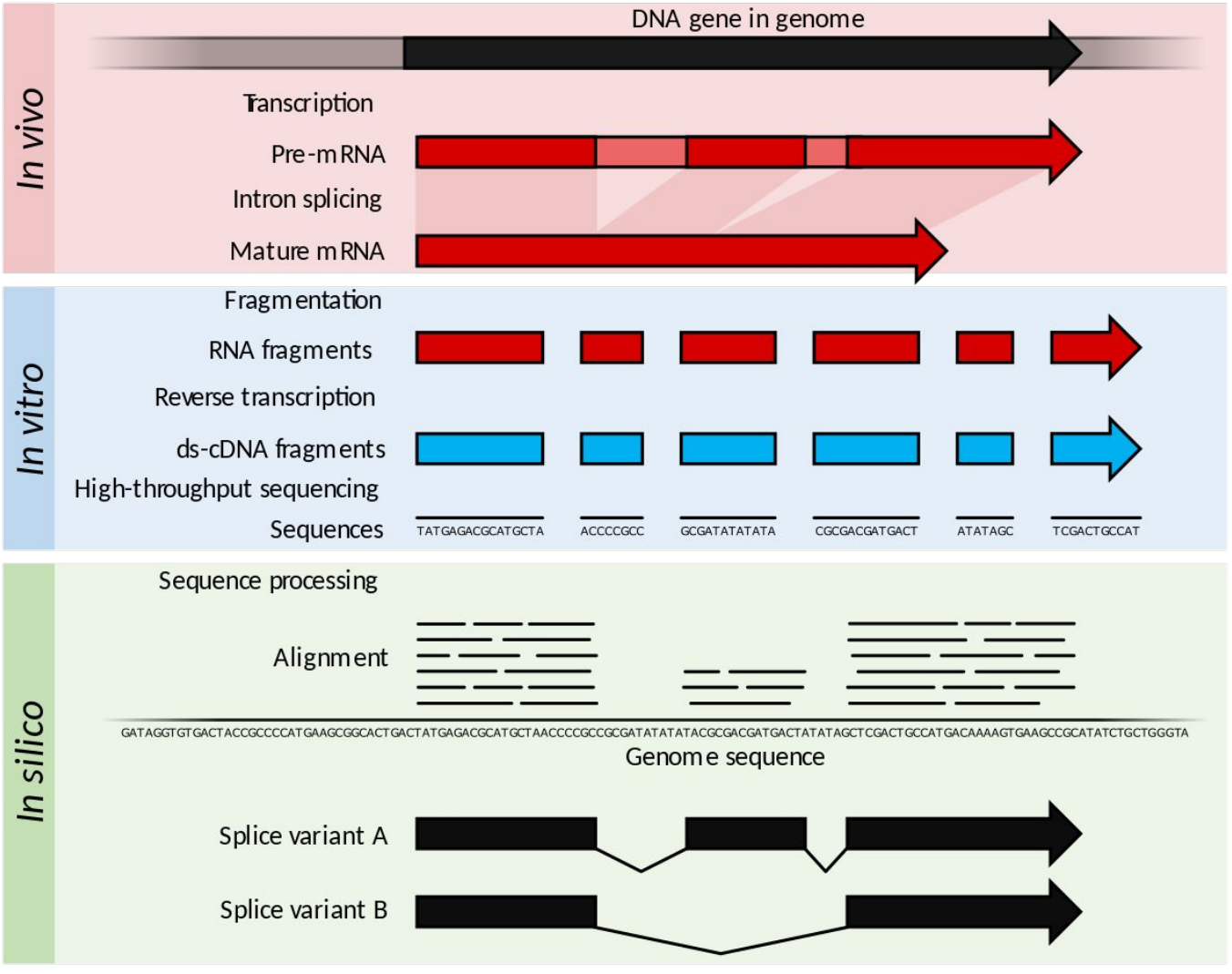

## Transcriptome/genome mapping

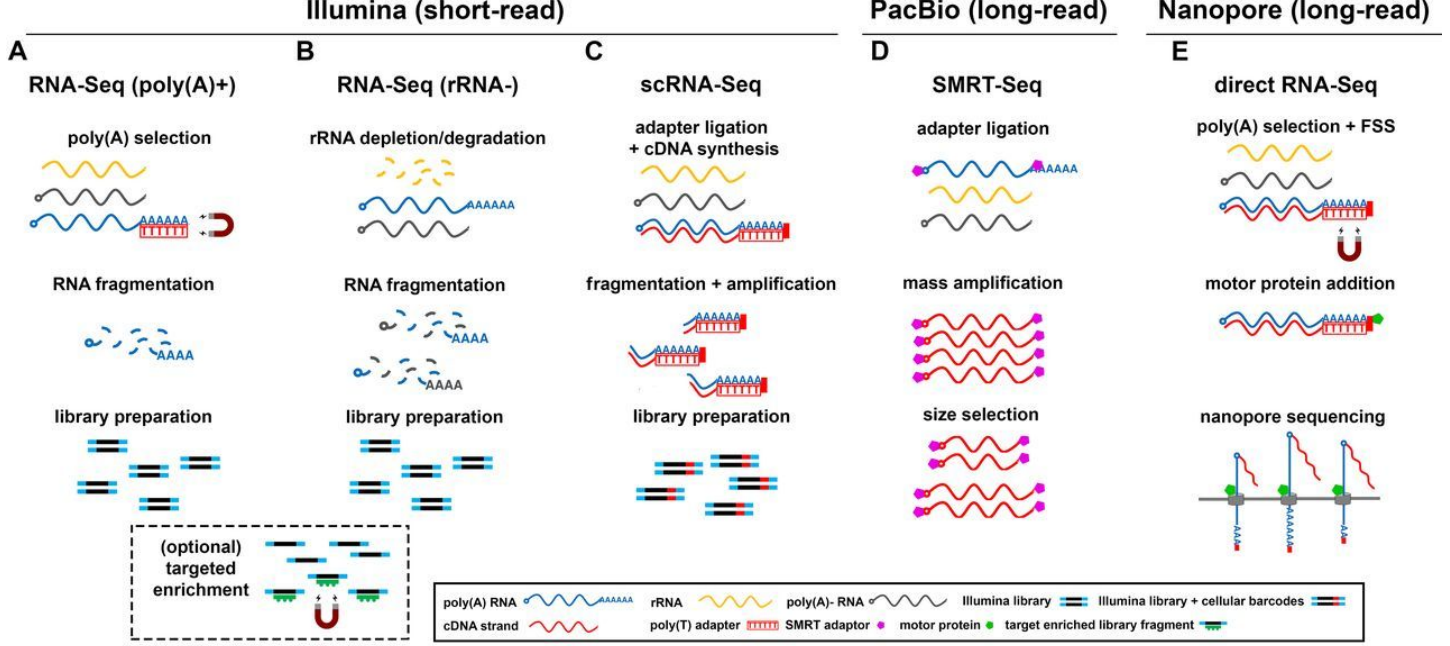


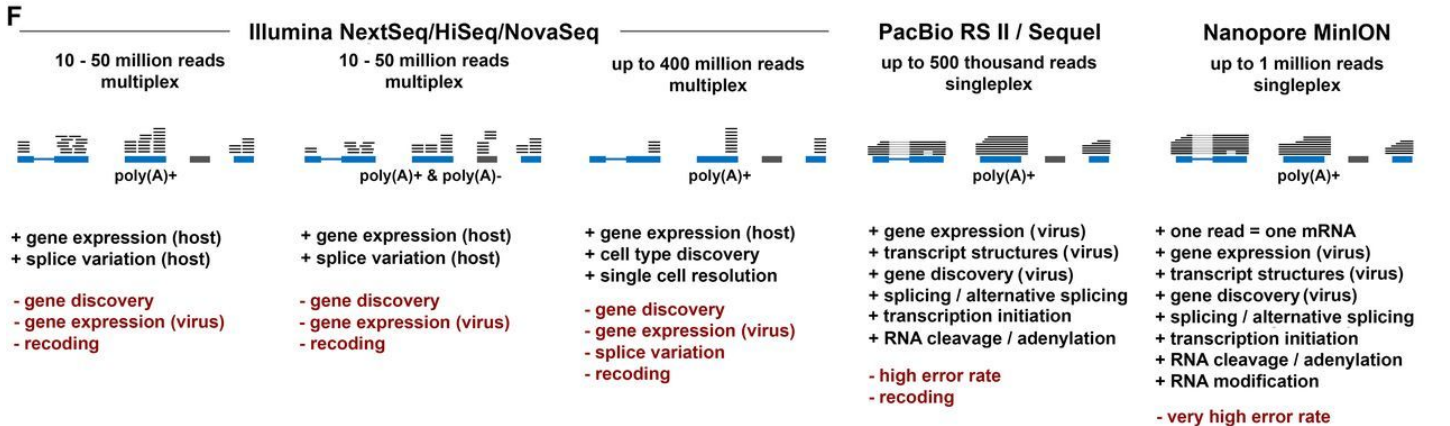

## Data analysis

- Differential expression
- Variant calling
- Annotation
- Novel transcript discovery
- RNA editing

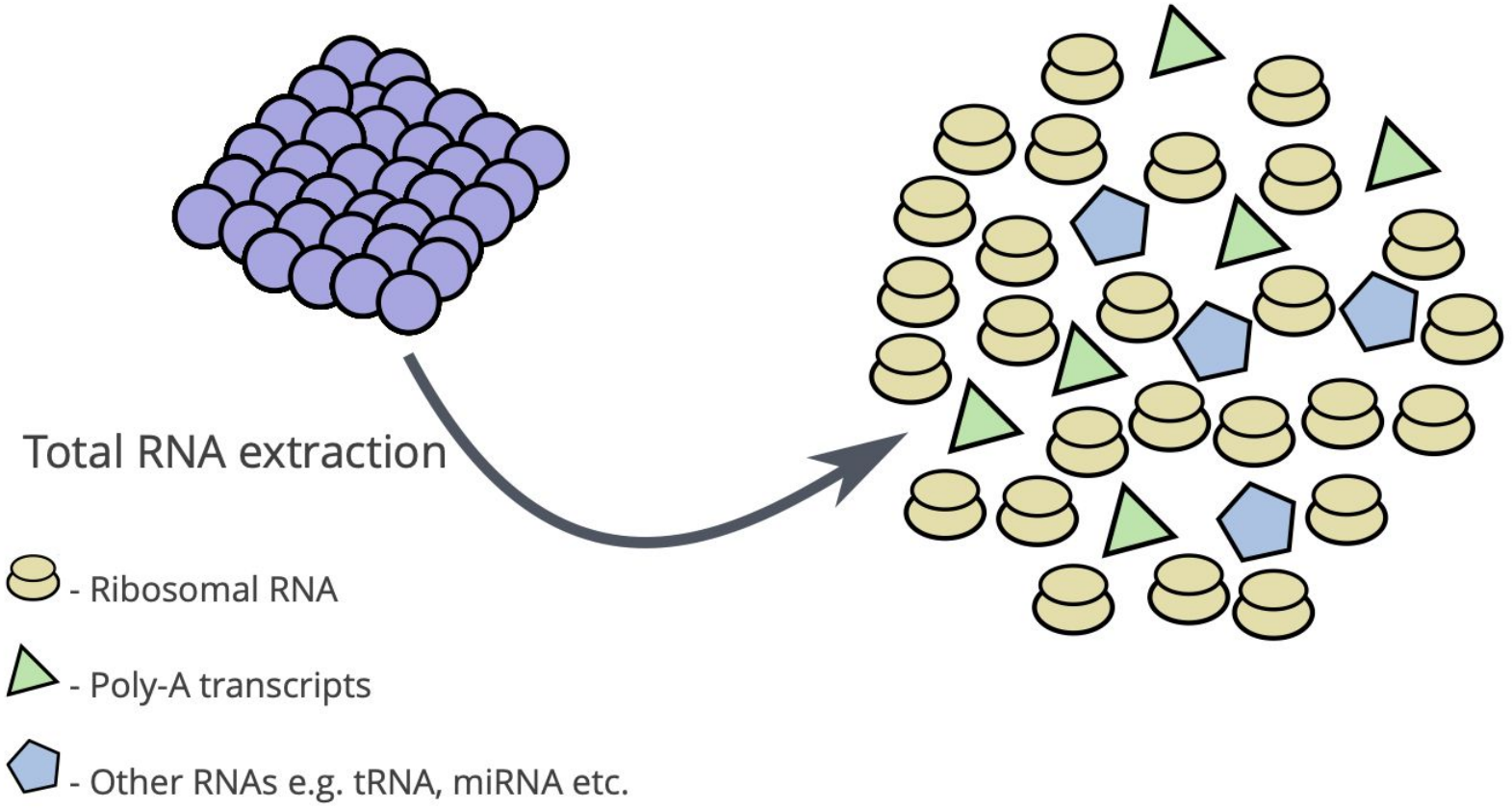
Van den Berge K, et al. 2019.
*Annu. Rev. Biomed. Data Sci.* 2:139–73


In vivo
DNA gene in genome
Transcription
Pre-mRNA
Intron splicing
Mature mRNA
In vitro
Fragmentation
RNA fragments
Reverse transcription
ds-cDNA fragments
High-throughput sequencing
Sequences
TATGAGACGCATGCTA
ACCCCGCC
GCGATATATATA
CGCGACGATGACT
ATATAGC
TCGACTGCCAT
In silico
Sequence processing
Alignment
GATAGGTGTGACTACCGCCCCATGAAGCGGCACTGACTATGAGACGCATGCTAACCCCGCCGCGATATATATACGCGACGATGACTATATAGCTCGACTGCCATGACAAAAGTGAAGCCGCATATCTGCTGGGTA
Genome sequence
Splice variant A
Splice variant B


Illumina (short-read)
PacBio (long-read)
Nanopore (long-read)
A
RNA-Seq (poly(A)+)
poly(A) selection
RNA fragmentation
library preparation
(optional) targeted enrichment
B
RNA-Seq (rRNA-)
rRNA depletion/degradation
RNA fragmentation
library preparation
C
scRNA-Seq
adapter ligation + cDNA synthesis
fragmentation + amplification
library preparation
D
SMRT-Seq
adapter ligation
mass amplification
size selection
E
direct RNA-Seq
poly(A) selection + FSS
motor protein addition
nanopore sequencing
poly(A) RNA
cDNA strand
rRNA
poly(T) adapter
poly(A)- RNA
SMRT adaptor
motor protein
Illumina library
Illumina library + cellular barcodes
target enriched library fragment

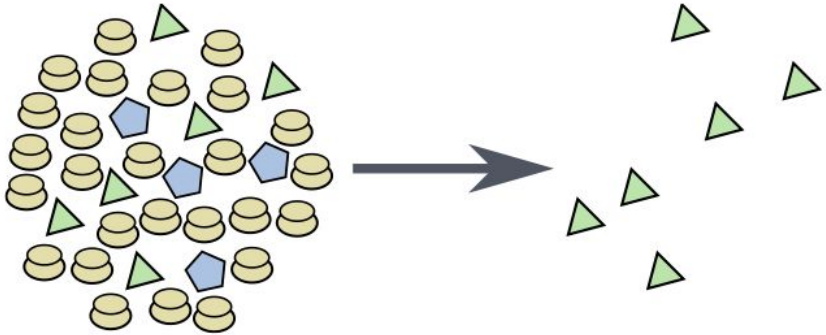

F
Illumina NextSeq/HiSeq/NovaSeq
10 - 50 million reads multiplex
poly(A)+
+ gene expression (host)
+ splice variation (host)
- gene discovery
- gene expression (virus)
- recoding
10 - 50 million reads multiplex
poly(A)+ & poly(A)-
+ gene expression (host)
+ splice variation (host)
- gene discovery
- gene expression (virus)
- recoding
up to 400 million reads multiplex
poly(A)+
+ gene expression (host)
+ cell type discovery
+ single cell resolution
- gene discovery
- gene expression (virus)
- splice variation
- recoding
PacBio RS II / Sequel
up to 500 thousand reads singleplex
poly(A)+
+ gene expression (virus)
+ transcript structures (virus)
+ gene discovery (virus)
+ splicing / alternative splicing
+ transcription initiation
+ RNA cleavage / adenylation
- high error rate
- recoding
Nanopore MinION
up to 1 million reads singleplex
poly(A)+
+ one read = one mRNA
+ gene expression (virus)
+ transcript structures (virus)
+ gene discovery (virus)
+ splicing / alternative splicing
+ transcription initiation
+ RNA cleavage / adenylation
+ RNA modification
- very high error rate

# Library preparation



# Library preparation

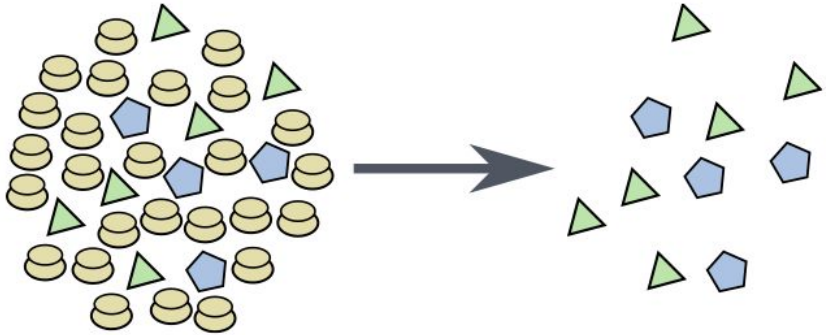
## Poly-A Selection



Poly-A transcripts e.g.:

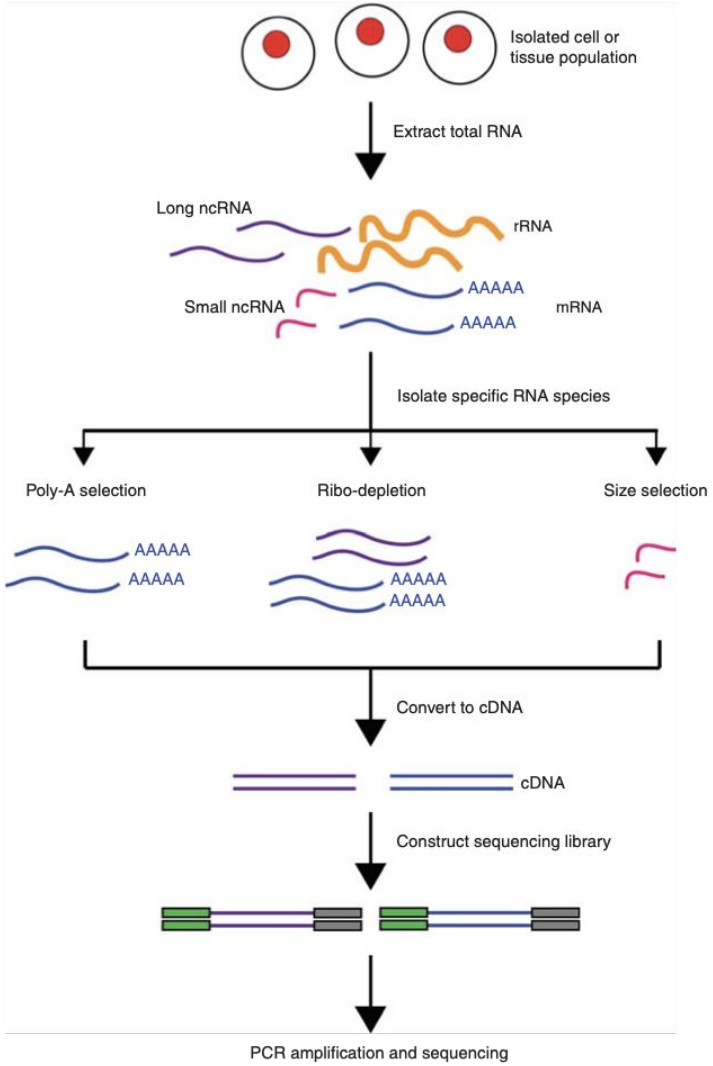
- mRNAs
- immature miRNAs
- snoRNA

## Ribominus selection

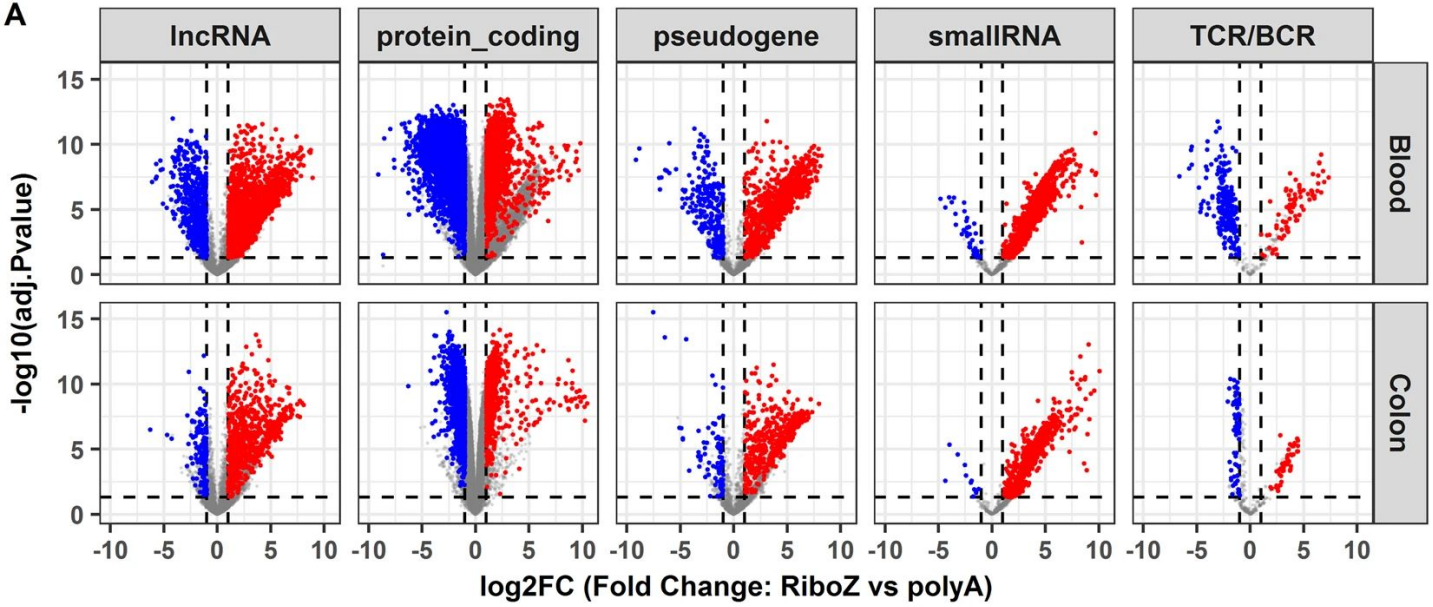


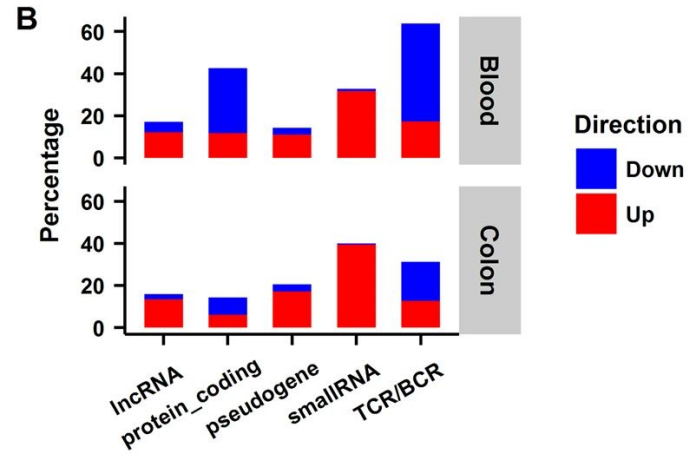
Poly-A transcripts + Other mRNAs e.g.:

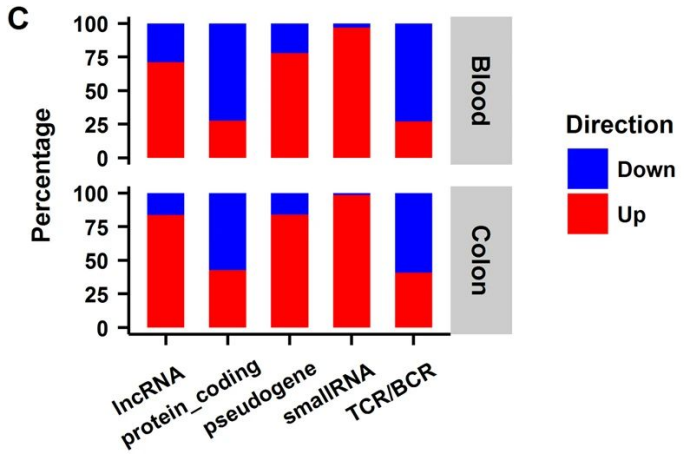
- tRNAs
- mature miRNAs
- piRNAs



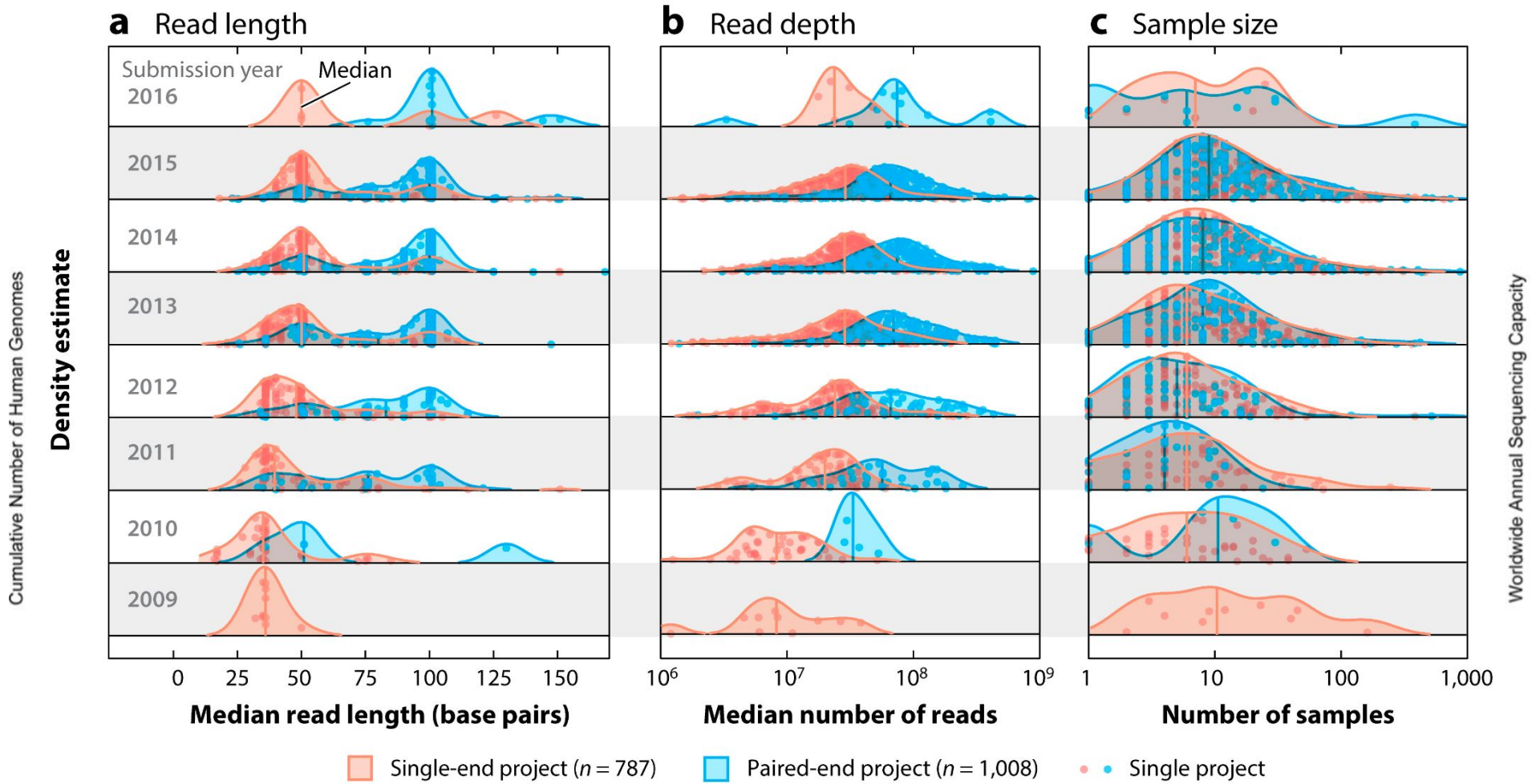

# When would you do a size selection?



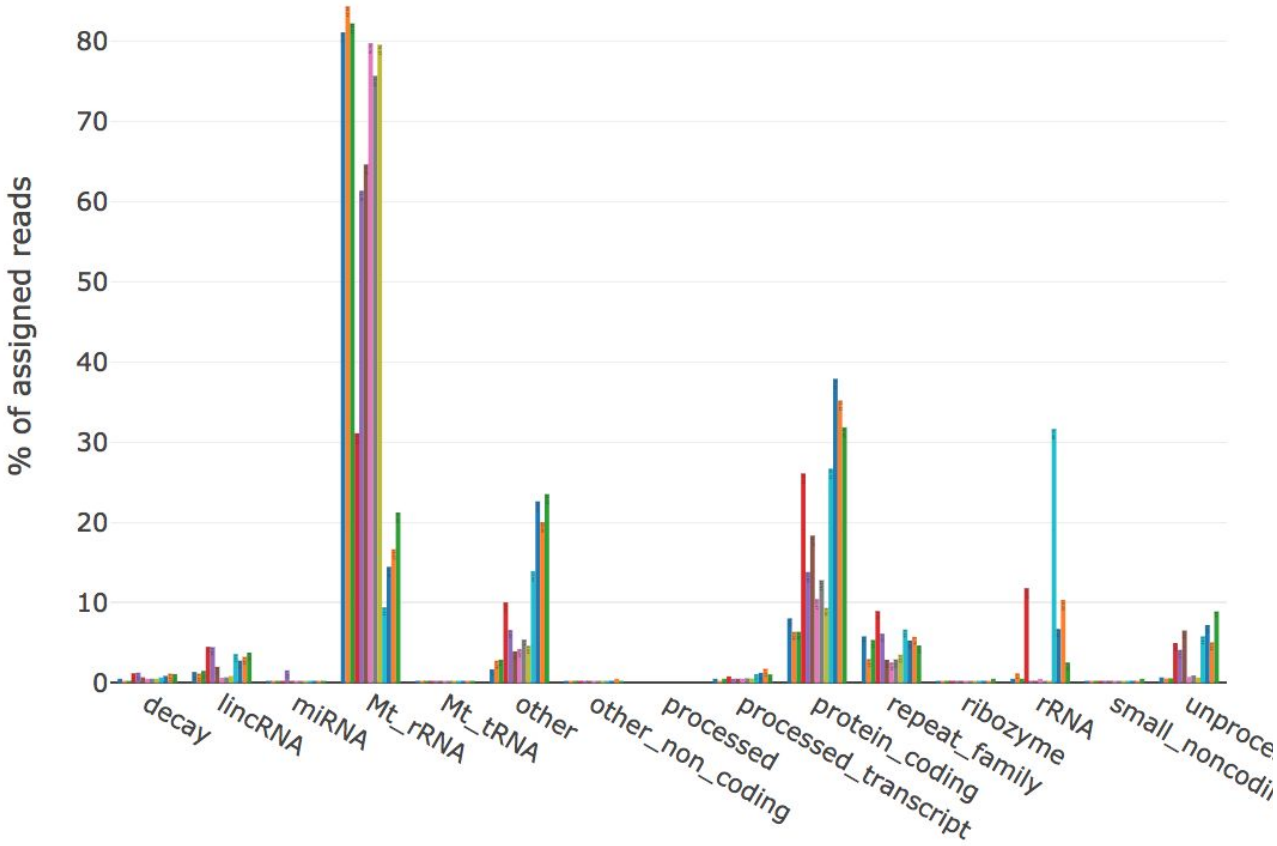

"polyA+ selection and rRNA depletion both selectively omit a distinct set of RNAs, so different fractions of the transcriptome are sequenced; thus, generating incompatible datasets."



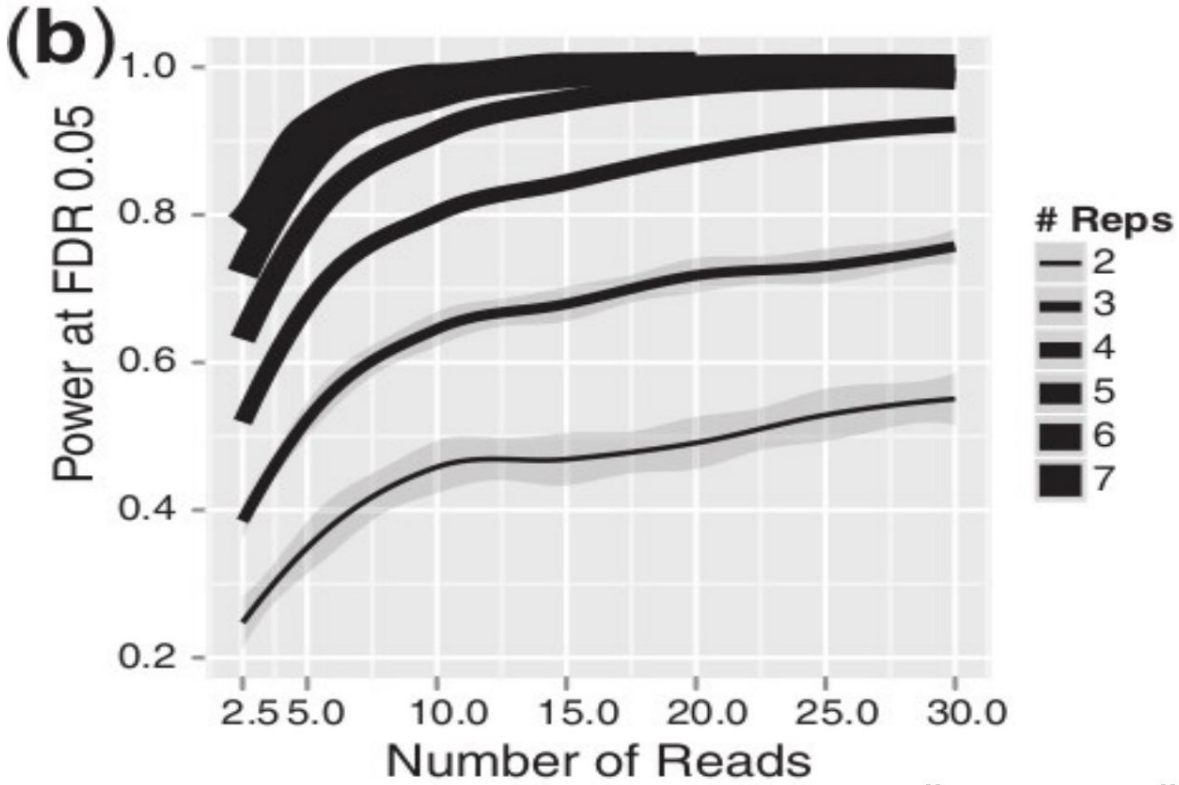

Van den Berge K, et al. 2019.
*Annu. Rev. Biomed. Data Sci.* 2:139–73

# Find problems in datasets generated, based on previous figure.

# What is the problem for RNA-seq data for this graph?



# Should one have more replicates or more sequencing depth?




Generally more replicates should be preferred if they can be obtained.

From Liu et al. 2014. RNA-seq differential expression studies: more sequence or more replication?

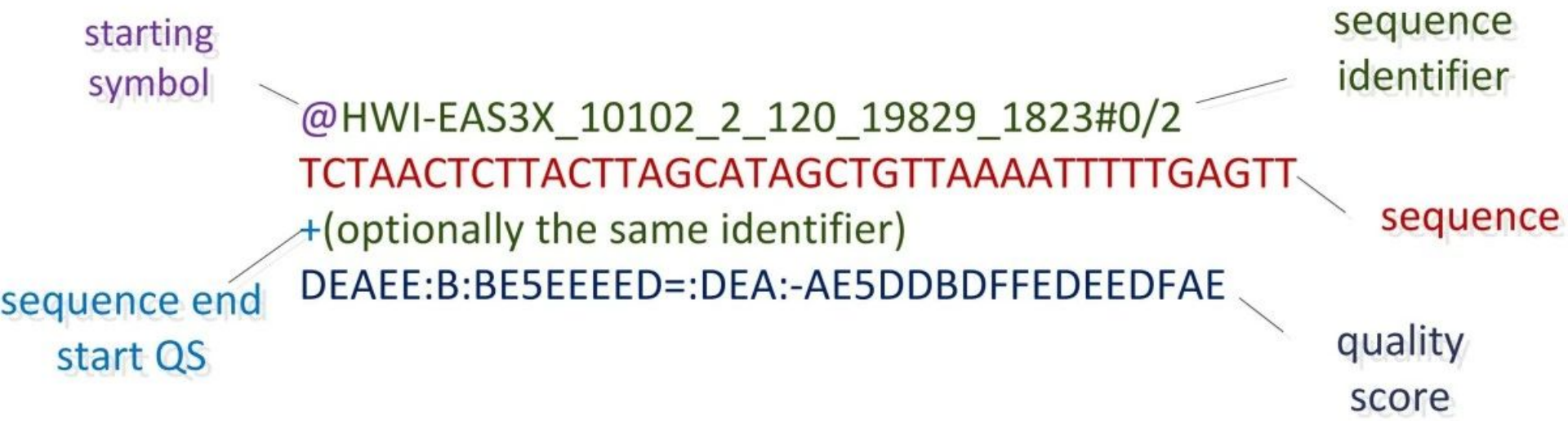
# Raw data formats

# FASTA format

> chr1

NNNNNNNNNNNNNNNNATGCATGCATGCATGCATGCATGCATGCATGCATG
CATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATG
CATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATG
CATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATG
CATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATG
CATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATG
CATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATG
CATGCATGC

# FASTQ format



# Data for the course

https://bioinfo.evolution.uzh.ch/project.zip

# Bash commands fresh up

# Command to make a directory

# Command to make a directory

**mkdir**

# Command to change directory

# Command to change directory

**cd**

# Command to copy

# Command to copy

**cp**

# Command to rename

# Command to rename

**mv**

# Command to move

# Command to move

**mv**

# Command to go to home directory

# Command to go to home directory

**cd**

**cd ~**

**cd $HOME**

# Command to show your username

# Command to show your username

**whoami**

# Command to list files

# Command to list files

**ls**

# Command to show content of a file

# Command to show content of a file

**cat file**

**less file**

**more file**

# Command to make an empty file

Command to make an empty file

**touch filename**

# Command to get help for a tool/ command

Command to get help for a tool/ command

**help command**

**command -h**

**command --help**

# Command to softlink

# Command to softlink

**ln**

# Command to check if a command exists

# Command to check if a command exists

**which**

# Command to do 2nd task only after 1st task is done

Command to do 2nd task only after 1st task is done

**task1 && task2**

# Quality control

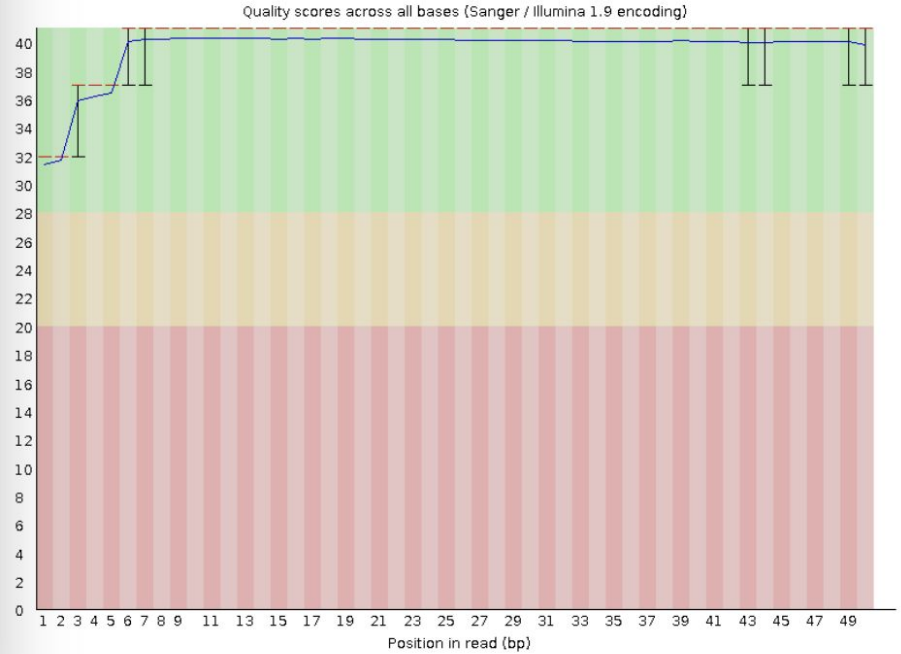
# Basic statistics

FastQC

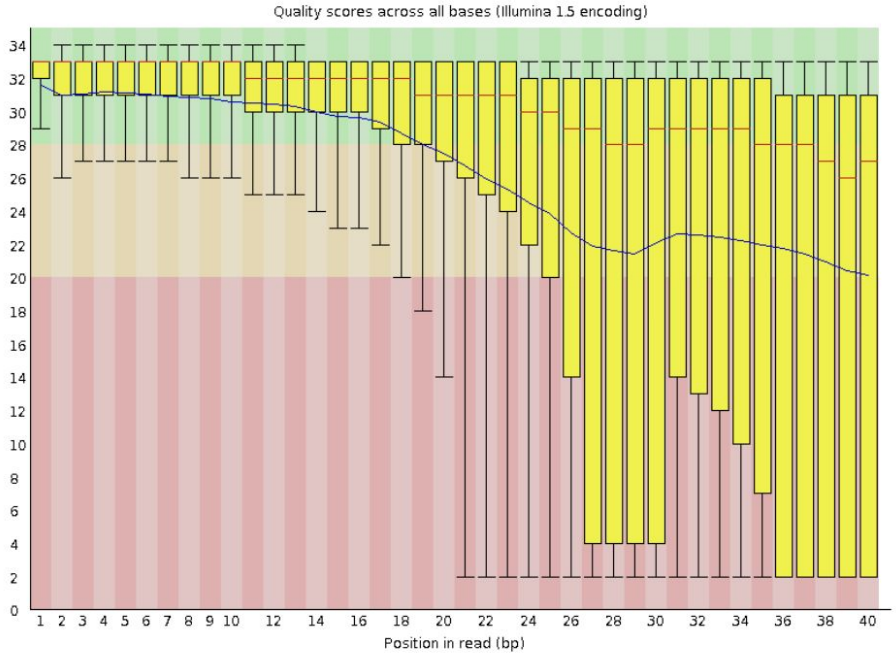
| Measure | Value |
|---|---|
| Filename | Reference_sample |
| File type | Conventional base calls |
| Encoding | Sanger/ Illumina 1.9 |
| Total sequences | 143,077,301 |
| Sequences flagged as poor quality | 0 |
| Sequence length | 100 |
| %GC | 47 |
| Duplicates | 37,625,112 |
| Sequencing type | Single end |

# Per base sequence quality

## Good Data




## Bad Data



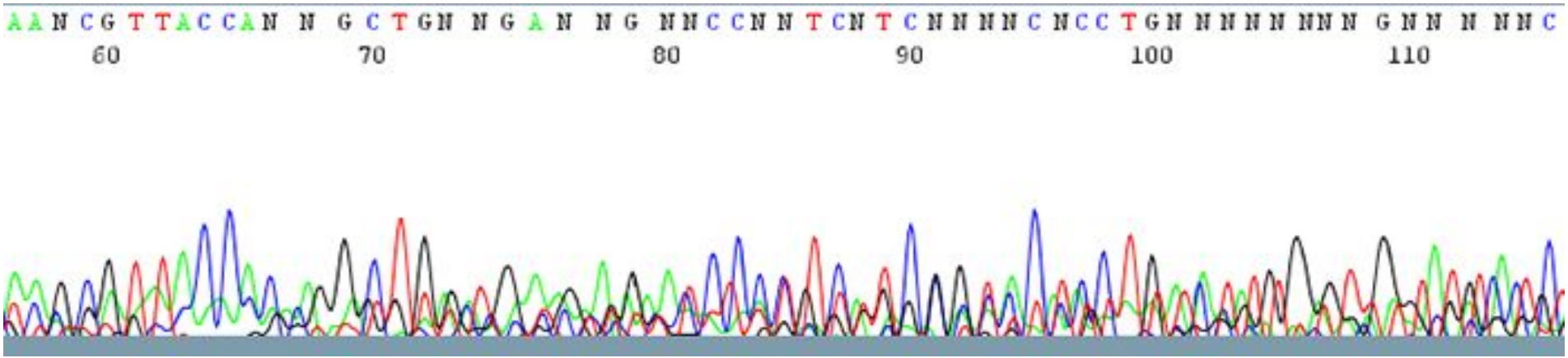
```
+SEQ_ID
!''*(((((***+))%%%++)(%%%%).1**
```

A quality value Q is an integer representation of the probability *p* that the corresponding base call is incorrect.

$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

| Phred Quality Score | Probability of incorrect base call | Base call accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1000 | 99.9% |
| 40 | 1 in 10000 | 99.99% |
| 50 | 1 in 100000 | 99.999% |

# Base cannot be determined

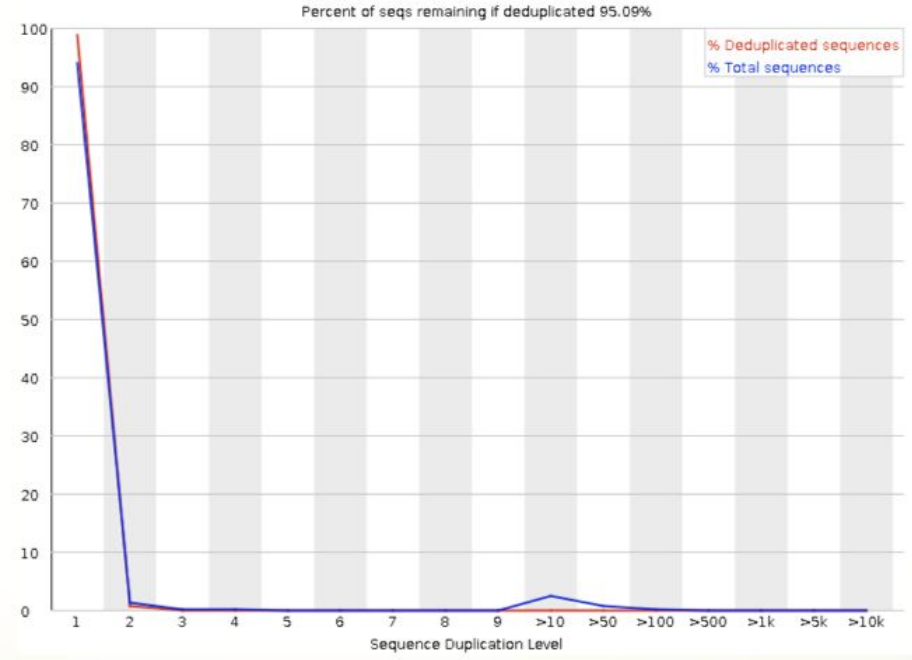


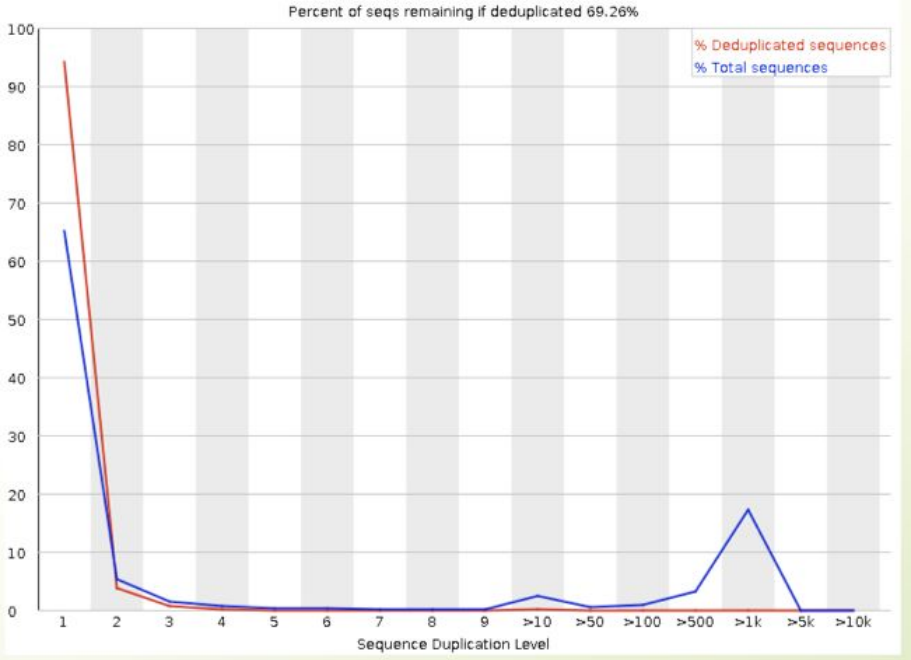
# Removing bad data

**NNNNNN**TGGCATATACGCGTGGCATATA**ADAPTER**

- Remove N
- Remove Adapter
- Remove bases with bad quality
- Remove reads get shorter than 30 bp

# Duplication problem



# Performing quality check with fastqc

1. Check if fastqc is installed
2. Make a directory: analysis in the project directory
3. Go to analysis directory and make a directory: 01_qc_raw
4. Go to 01_qc_raw and make directories: log and output

# Shell script vs Makefile

# Shell script to run analysis

```
#!/usr/bin/bash    or #!/bin/bash

for i in ../../raw/*.gz

        do

        fastqc $i -o ./output/ --noextract -f fastq -t 2

done
```

**What changes can you make to this script?**

# Shell script to run analysis

```
#!/usr/bin/bash    or #!/bin/bash

for i in ../../raw/*.gz

        do

        fastqc -v >./log/${i}.log && fastqc $i -o ./output/ --noextract -f fastq -t 2 2>>
./log/${i}.log

done
```

# Shell script to run analysis

```
#!/usr/bin/bash    or #!/bin/bash

for i in ../../raw/*.gz

        do

        fastqc -v >./log/${i}.stdout && fastqc $i -o ./output/ --noextract -f fastq -t 2 2>
./log/${i}.stderr

done
```

# Power of Makefile

```
# This makefile will run the FastQC software to check the quality of FastQ files

SHELL:=/bin/bash
source_dir=../../raw
target_dir=./output
files := $(wildcard $(source_dir)/*.fq.gz)
targets := $(patsubst $(source_dir)/%.fq.gz, $(target_dir)/%_fastqc.zip, $(files))

all: $(targets)
$(target_dir)/%_fastqc.zip: $(source_dir)/%.fq.gz
	fastqc -v > ./log/$(basename $(notdir $@)).stdout && fastqc $< -o ./output/
--noextract -f fastq -t 2 2>./log/$(basename $(notdir $@)).stderr
```

# Power of Makefile

```
# This makefile will run the FastQC software to check the quality of FastQ files

SHELL:=/bin/bash
source_dir=../data/raw
target_dir=./output
files := $(wildcard $(source_dir)/*.fq.gz)
targets := $(patsubst $(source_dir)/%.fq.gz, $(target_dir)/%_fastqc.zip, $(files))

all: $(targets)
$(target_dir)/%_fastqc.zip: $(source_dir)/%.fq.gz
	fastqc -v > ./log/$(basename $(notdir $@)).stdout && fastqc $< -o ./output/
--noextract -f fastq -t 2 2>./log/$(basename $(notdir $@)).stderr
```

# Makefile for simple task

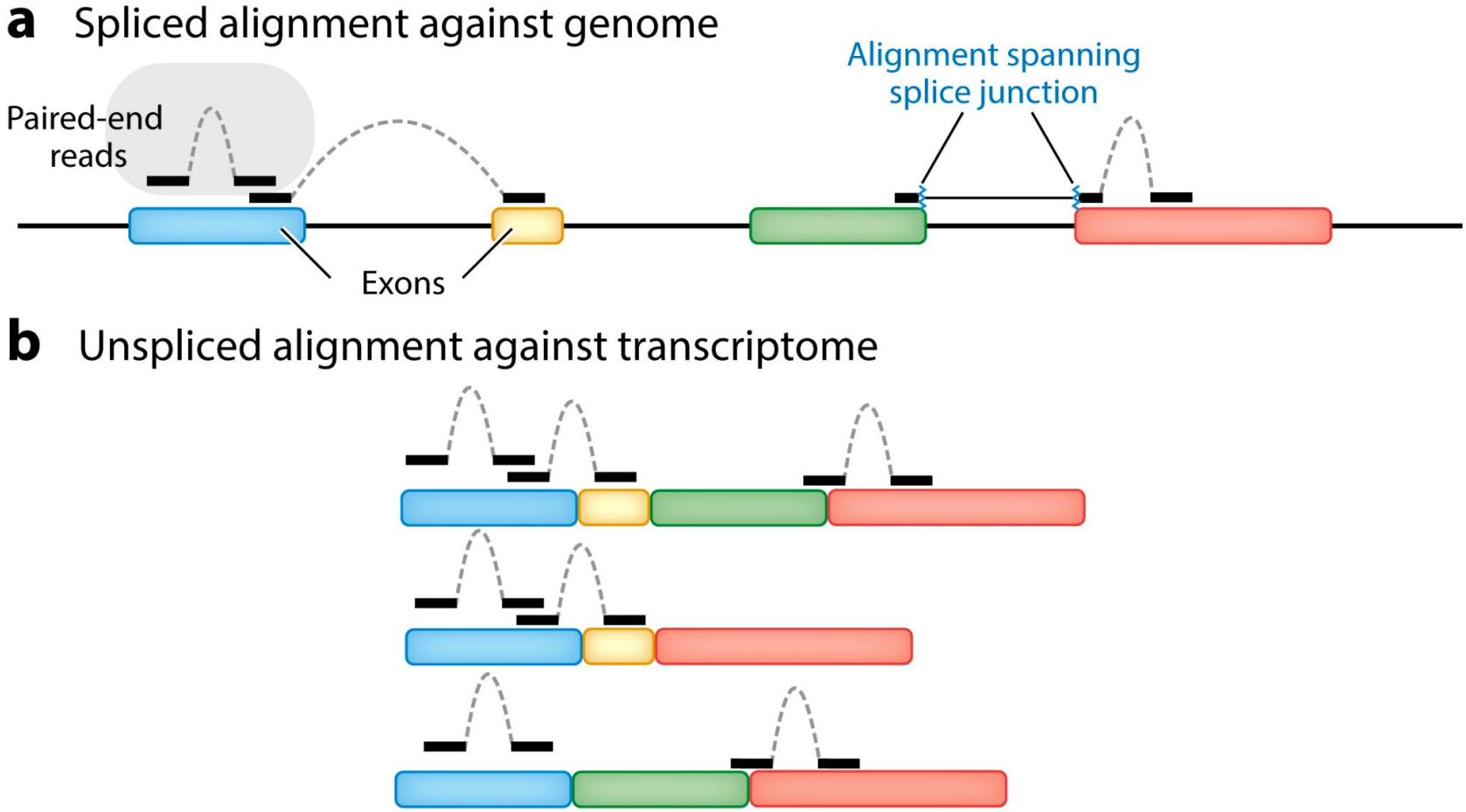
```
# Makefile to count unique lines


all: task1

task1:
	cat file.txt | uniq -c
```

# Removing bad data TrimGalore

**NNNNNN**TGGCATATACGCGTGGCATATA**ADAPTER**

- Remove N
- Remove Adapter
- Remove bases with bad quality (phred scores less than 30)
- Remove reads get shorter than 30 bp

# Alignment and read counts




Van den Berge K, et al. 2019.
*Annu. Rev. Biomed. Data Sci.* 2:139–73

# Alignment

https://youtu.be/4WRANhDiSHM

https://tinyheero.github.io/2015/09/02/pseudoalignments-kallisto.html

**Watch the video and read the link**

# Why not use BLAST?

# Why not use BLAST?

| Aligner | Human reference runtime (hrs) | Max mem used (GB) | Number of AMD 64 bit core processors |
|---|---|---|---|
| Bowtie2 | 0.62 | 9 | 17 |
| BWA | 0.66 | 9 | 17 |
| BLAST | 9.4 | 12 | 17 |

# Questions

**Question 1**

What is the application of BWT in biology?

**Question 2**

Which tools are built on BWT for biology?

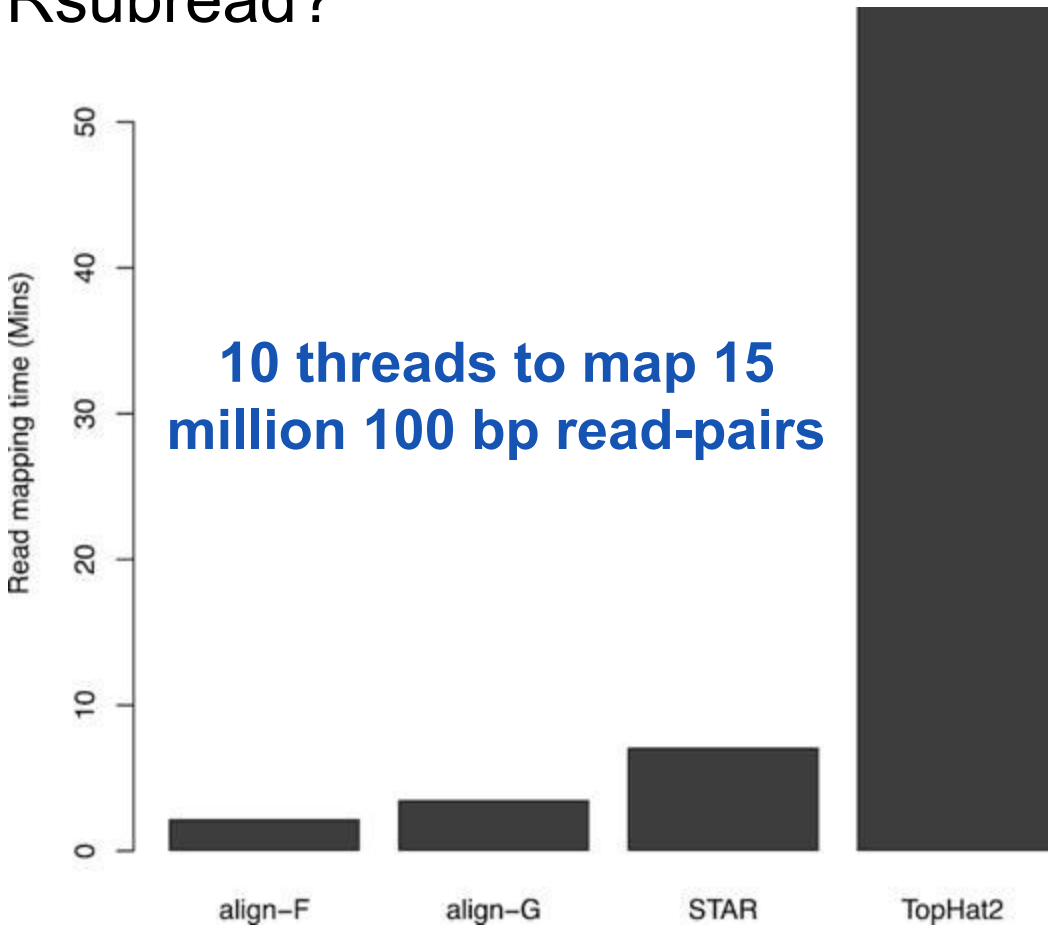
**Question 3**

Difference between alignment and Pseudoalignment?

**Question 4**

Which tools can perform Pseudoalignment and for which *seq data?

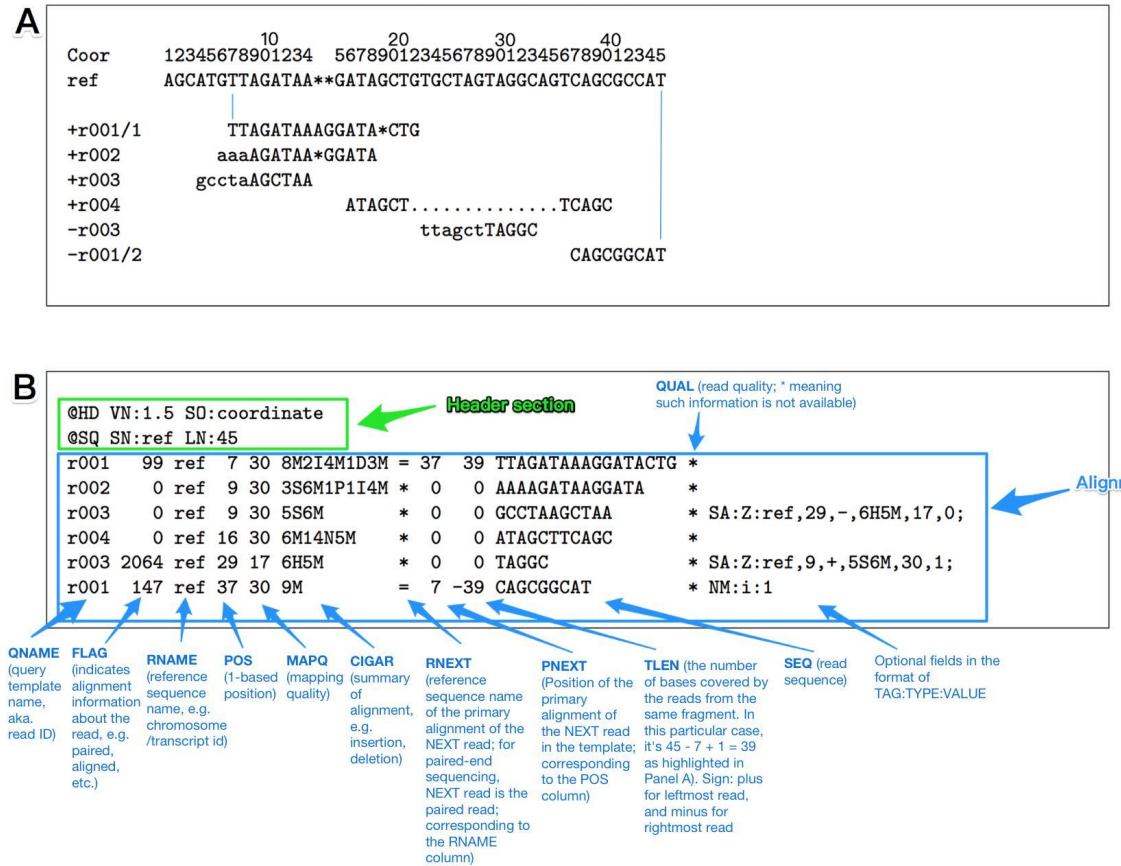
# Why to use Rsubread?



# Files needed for alignment

1. Trimmed FASTQ files
2. A Reference genome
3. Gene definition file
4. Tools to run alignment

# SAM format


A
10 20 30 40
Coor 12345678901234 5678901234567890123456789012345
ref AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1 TTAGATAAAGGATA*CTG
+r002 aaaAGATAA*GGATA
+r003 gcctaAGCTAA
+r004 ATAGCT.....TCAGC
-r003 ttagctTAGGC
-r001/2 CAGCGGCAT
B
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
Header section
QUAL (read quality; * meaning such information is not available)
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
Alignment section
QNAME (query template name, aka. read ID)
FLAG (indicates alignment information about the read, e.g. paired, aligned, etc.)
RNAME (reference sequence name, e.g. chromosome /transcript id)
POS (1-based position)
MAPQ (mapping quality)
CIGAR (summary of alignment, e.g. insertion, deletion)
RNEXT (reference sequence name of the primary alignment of the NEXT read; for paired-end sequencing, NEXT read is the paired read; corresponding to the RNAME column)
PNEXT (Position of the primary alignment of the NEXT read in the template; corresponding to the POS column)
TLEN (the number of bases covered by the reads from the same fragment. In this particular case, it's 45 - 7 + 1 = 39 as highlighted in Panel A). Sign: plus for leftmost read, and minus for rightmost read
SEQ (read sequence)
Optional fields in the format of TAG:TYPE:VALUE

# SAM Flags

| Bit | | Description |
|---|---|---|
| 1 | 0x1 | template having multiple segments in sequencing |
| 2 | 0x2 | each segment properly aligned according to the aligner |
| 4 | 0x4 | segment unmapped |
| 8 | 0x8 | next segment in the template unmapped |
| 16 | 0x10 | SEQ being reverse complemented |
| 32 | 0x20 | SEQ of the next segment in the template being reverse complemented |
| 64 | 0x40 | the first segment in the template |
| 128 | 0x80 | the last segment in the template |
| 256 | 0x100 | secondary alignment |
| 512 | 0x200 | not passing quality controls |
| 1024 | 0x400 | PCR or optical duplicate |
| 2048 | 0x800 | supplementary alignment |

**Example, flag 83 =** 64+16+2+1 means it's first read (0x40) of pair-end reads (0x1) and it's mapped on minus strand (0x10) and both reads mapped (0x2).

**https://broadinstitute.github.io/picard/explain-flags.html**

# Questions

1. Count the number of aligned reads in the BAM file
2. Store the unaligned reads from bam file in a new file
3. Display only header of the BAM file
4. Display first few lines of a BAM file with header
5. Display first few lines of a BAM file without header

Duplicated reads: which one is technical/ biological?

**Reference** ATCGGACTACTAATCCTCGCGCGATATAC

**Read 1**

ATCGGACT

ATCGGACT

**Read 2**

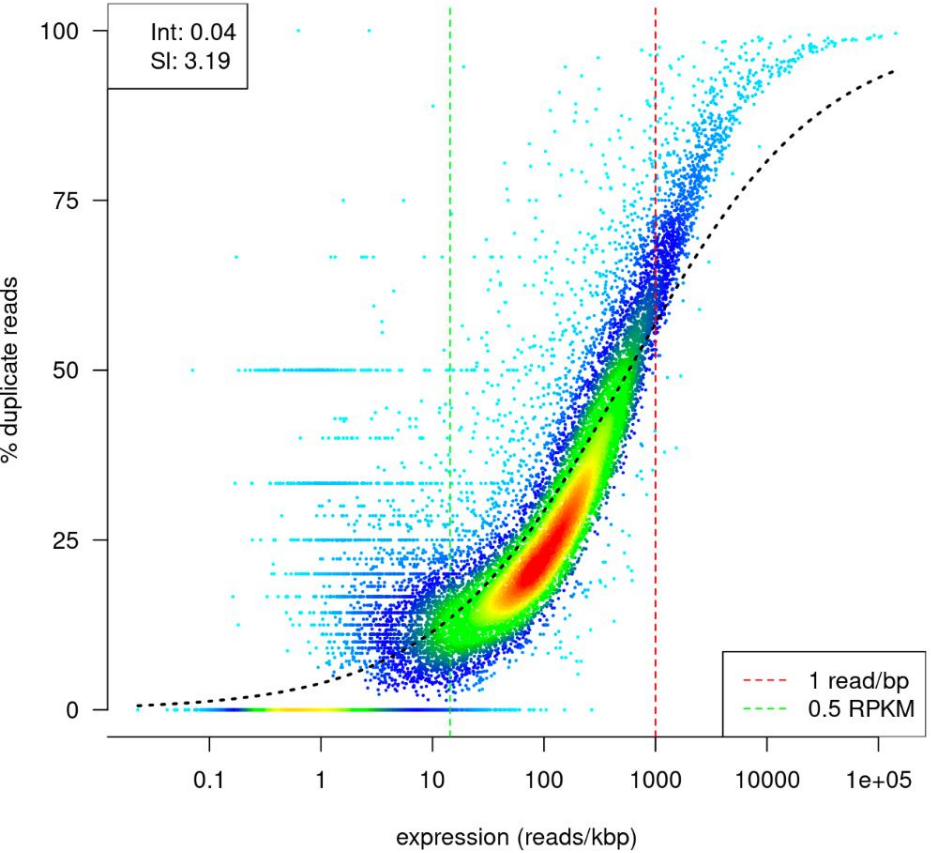
ACTACTAATCCTCGCG

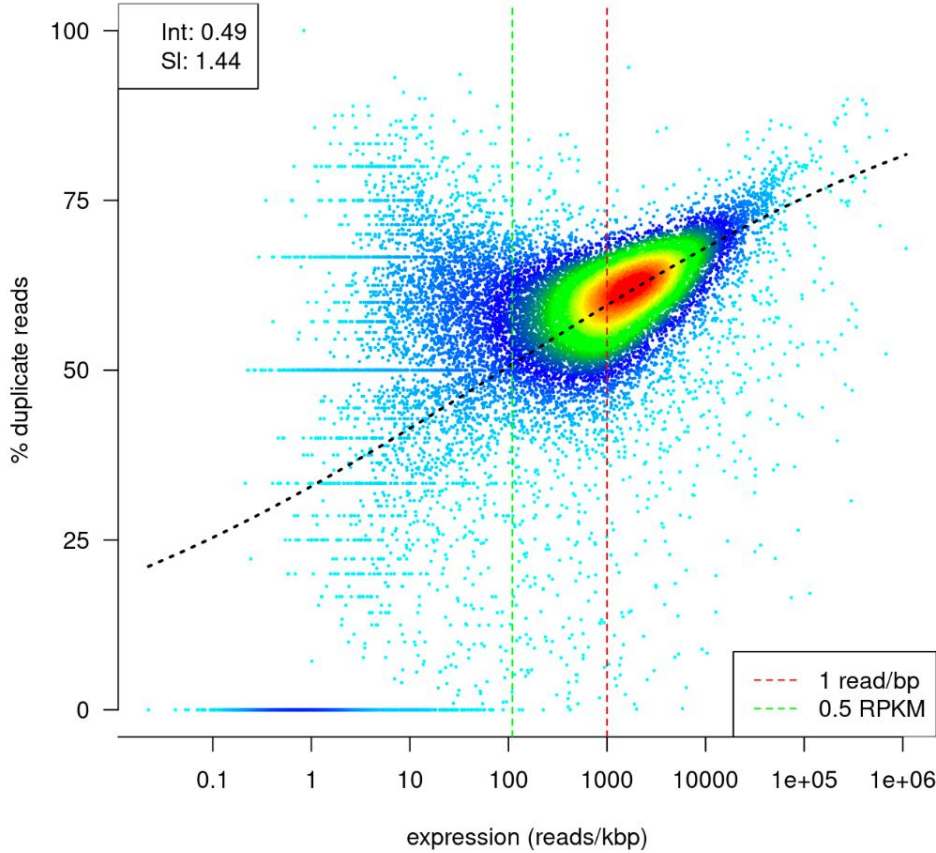
ACTACTAATCCTCGCG

# Reads duplication problem in RNA-seq

1. Should we remove duplicates, why or why not?
2. Which library preparation method would be best to distinguish between technical and biological duplicated reads?
3. How many PCR cycles should one use for optimal technical duplicates?
4. Is paired-end sequencing better than single-end sequencing for duplication problem?

# What does this plot tell you?



# Make a flowchart of RNA-seq pipeline so far