R for Bioinformatics

Analyses and challenges of RNA-seq data

CECAD COLOGNE

R Ladies Cologne

About me

Cascais, Portugal

TÉCNICO LISBOA

IMM Lisboa Instituto de Medicina Molecular

CECAD COLOGNE

CECAD Research Center, Cologne

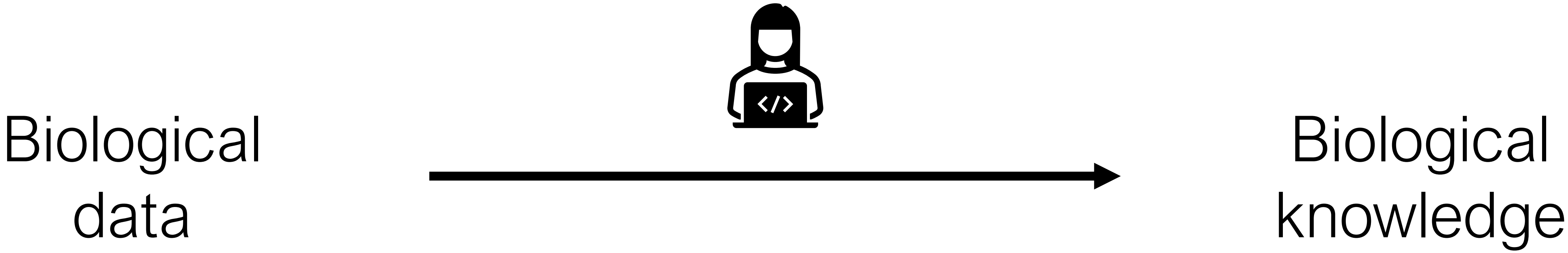
About today

- Introduction to Bioinformatics
- High-throughput RNA-sequencing
- RNA-sequencing data analysis workflow
    - Particularities of the data
    - Common challenges
    - Common analyses

What is Bioinformatics?

“Computer-aided biology”

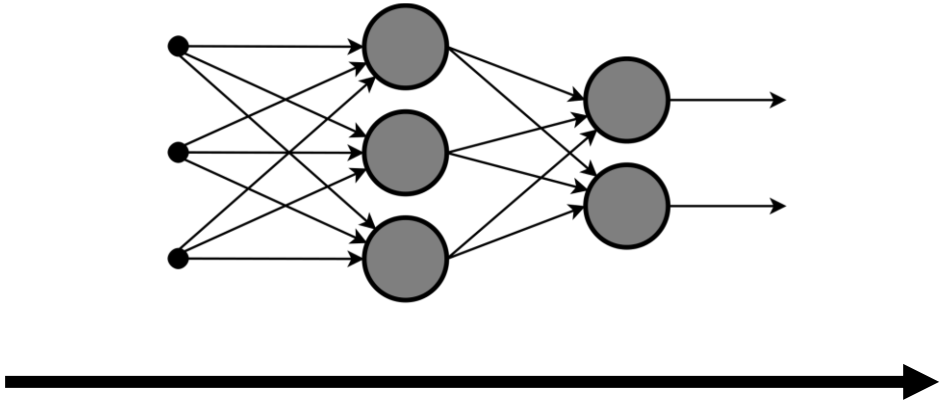
Prof. Dr. Barry Grant



Biological data

Biological knowledge

What is Bioinformatics?



Bacterial vs viral pneumonia diagnosis

Kermany *et al.* Cell (2018)

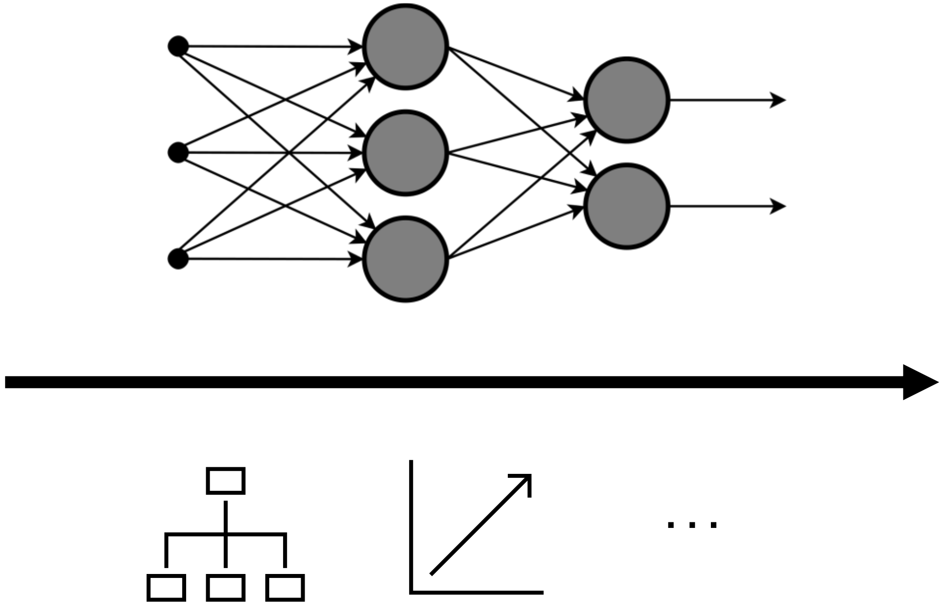
What is Bioinformatics?

High-throughput sequencing data

Medical imaging

Biomolecule structure

…



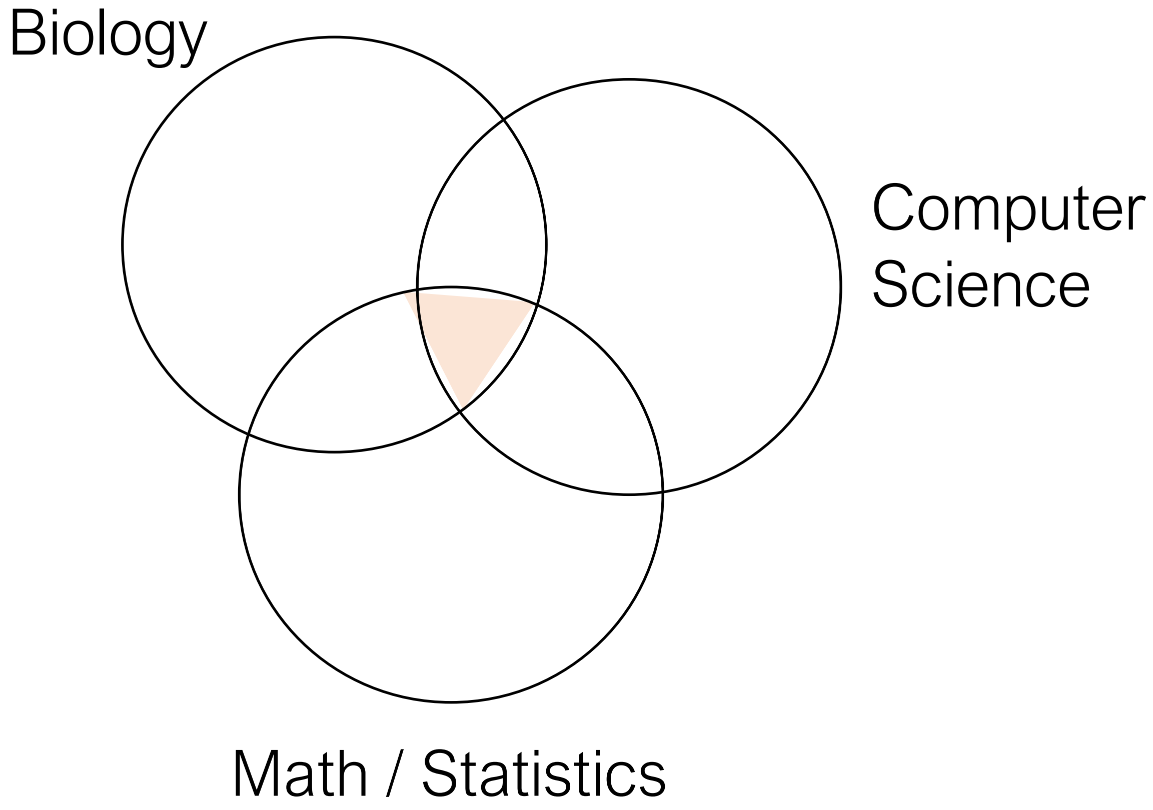
Fundamental biology

Medical diagnosis

Mechanisms of disease

What does a bioinformatician do?

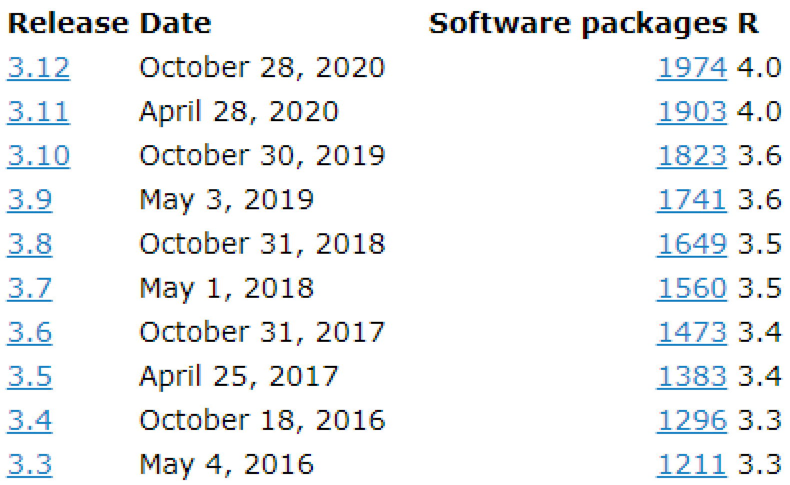
- Develop own software
- Apply previously developed software to particular questions and data
  - Understand the data
  - Apply analysis pipelines
  - Refine analysis pipelines after first findings and data exploration



Bioinformatics using R

Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS



Release	Date	Software packages	R
3.12	October 28, 2020	1974	4.0
3.11	April 28, 2020	1903	4.0
3.10	October 30, 2019	1823	3.6
3.9	May 3, 2019	1741	3.6
3.8	October 31, 2018	1649	3.5
3.7	May 1, 2018	1560	3.5
3.6	October 31, 2017	1473	3.4
3.5	April 25, 2017	1383	3.4
3.4	October 18, 2016	1296	3.3
3.3	May 4, 2016	1211	3.3

- Open source
- Open development
- Documented
- Reviewed
- Common platform for community

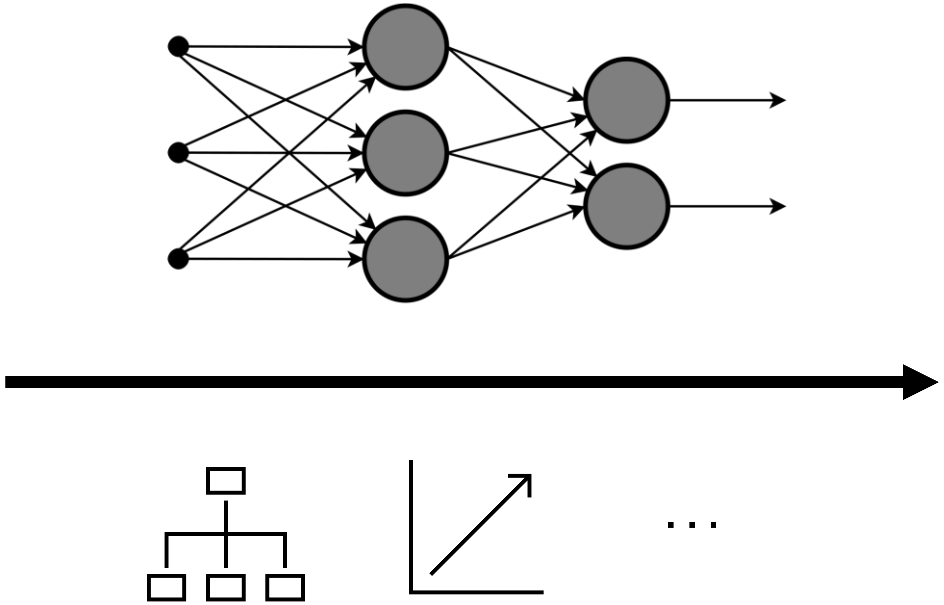
What is Bioinformatics?

High-throughput sequencing data

Medical imaging

Biomolecule structure

…

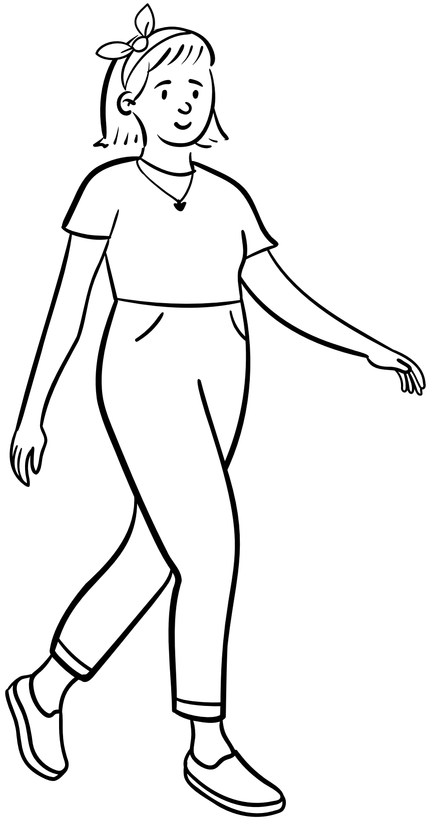


Fundamental biology

Medical diagnosis

Mechanisms of disease

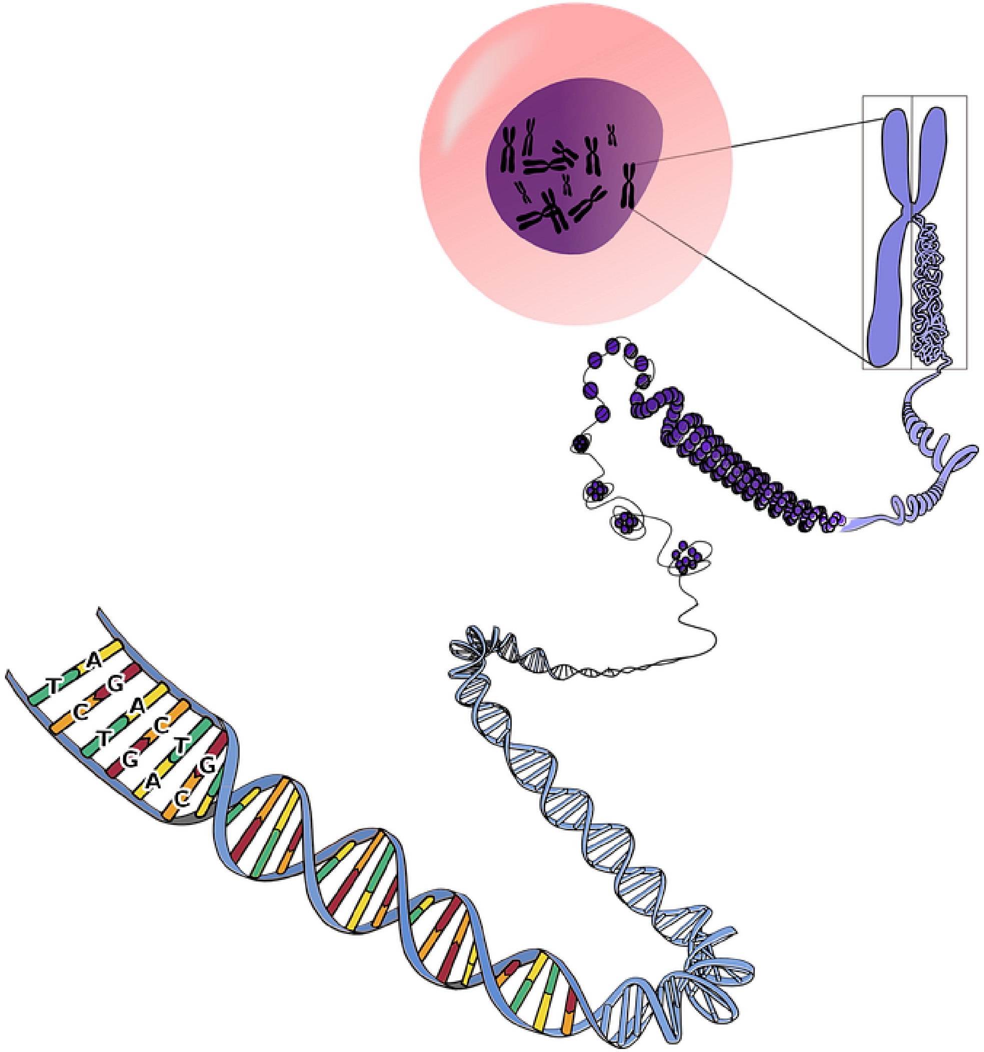
Genetic information



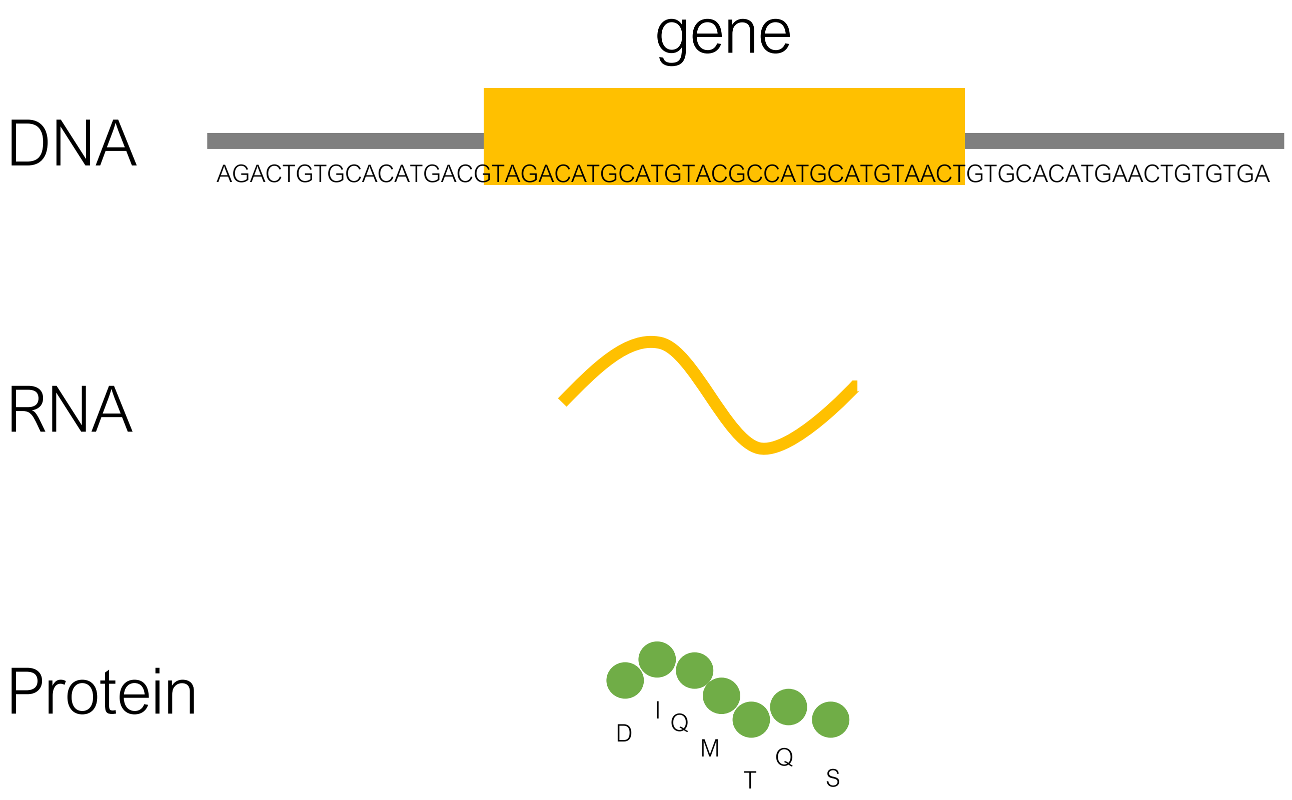
1. What is DNA?

2. How does it encode genetic information?

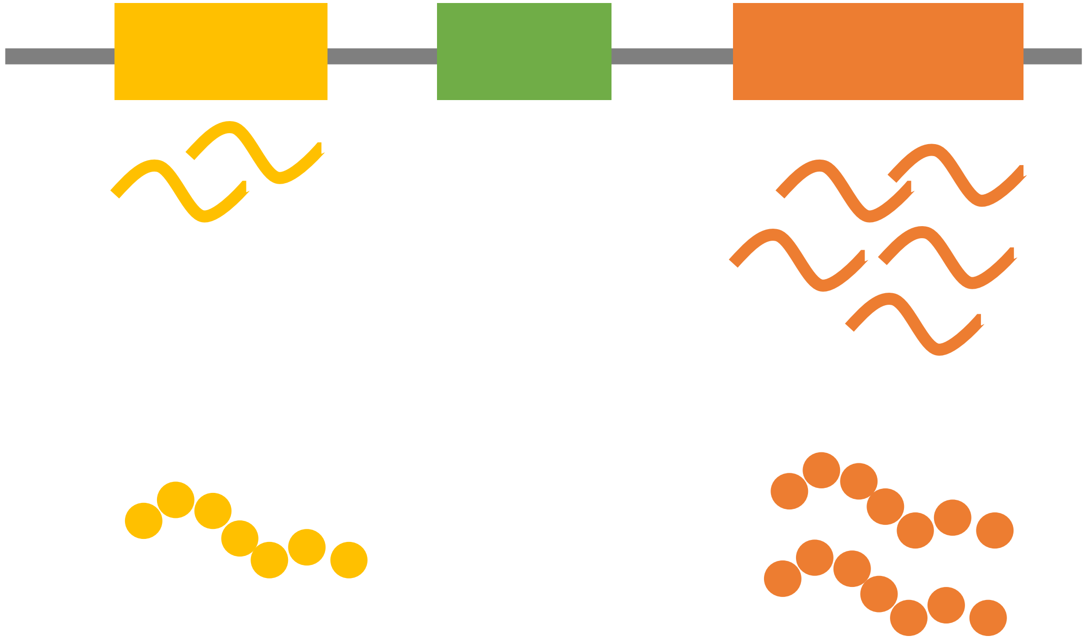
AGACTG



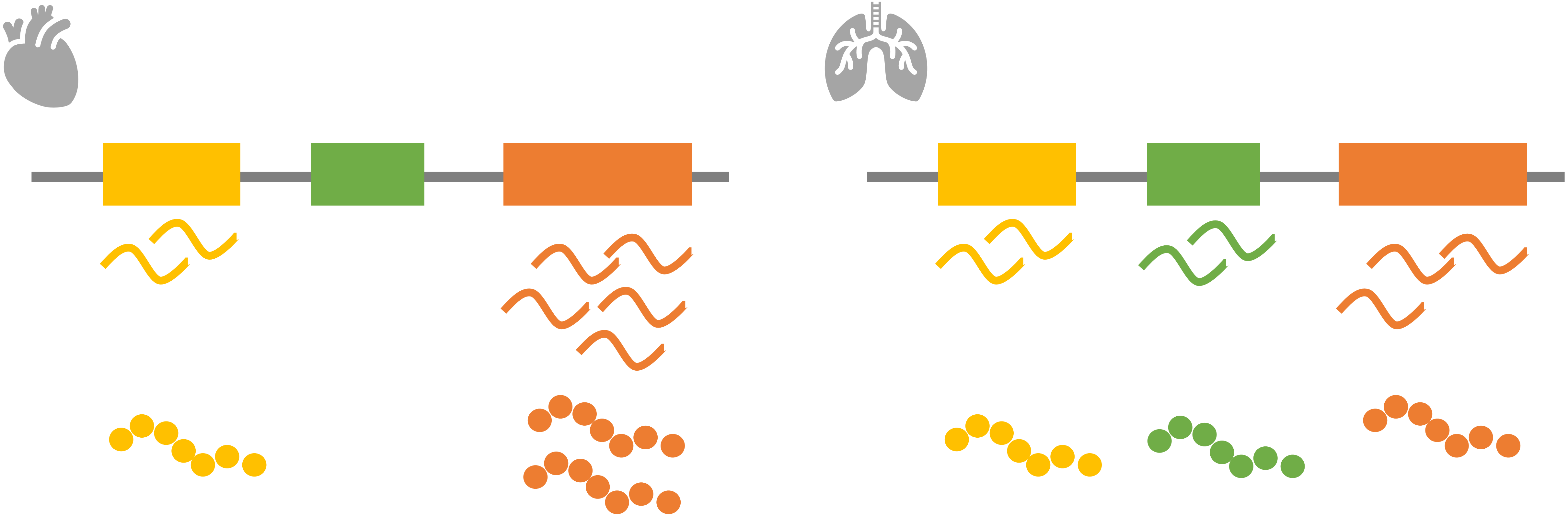
Genetic information


gene
DNA
AGACTGTGCACATGACGTAGACATGCATGTACGCCATGCATGTAACTGTGCACATGAACTGTGTGA
RNA
Protein
D
I
Q
M
T
Q
S

Gene expression



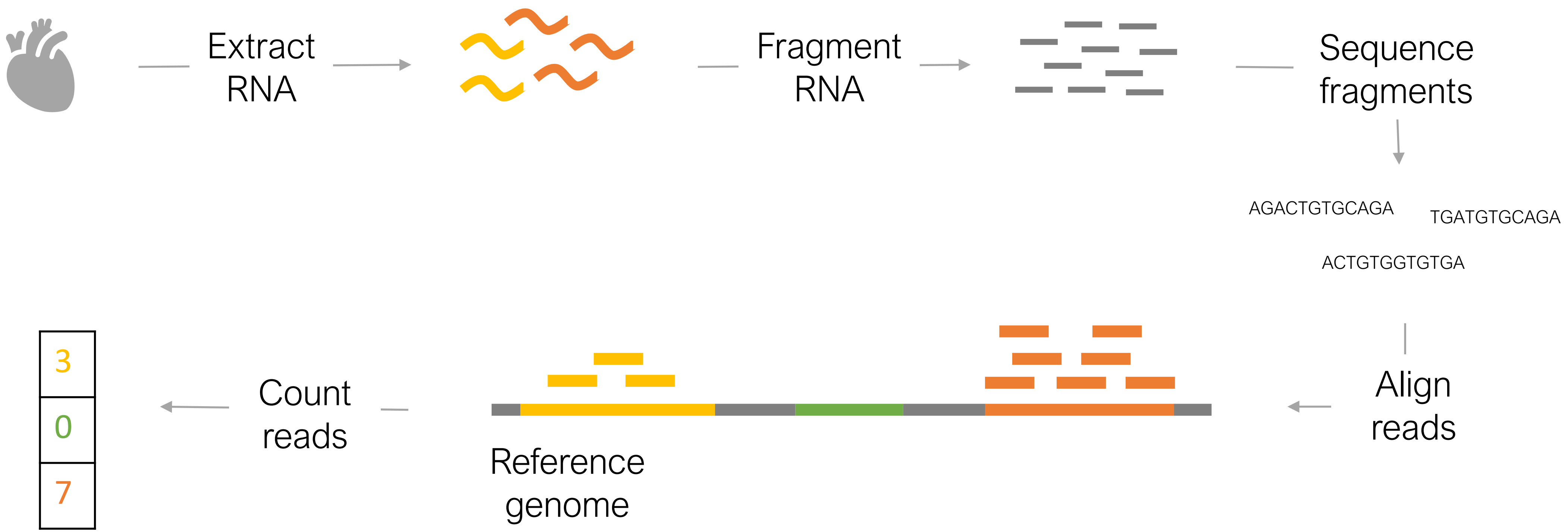
Gene expression



Gene expression tells us about cellular activity.

RNA-sequencing data

Humans have 20.000+ genes encoding for proteins -> need for scalable method



Extract RNA

Fragment RNA

Sequence fragments

AGACTGTGCAGA

TGATGTGCAGA

ACTGTGGTGTGA

Align reads

Reference genome

Count reads

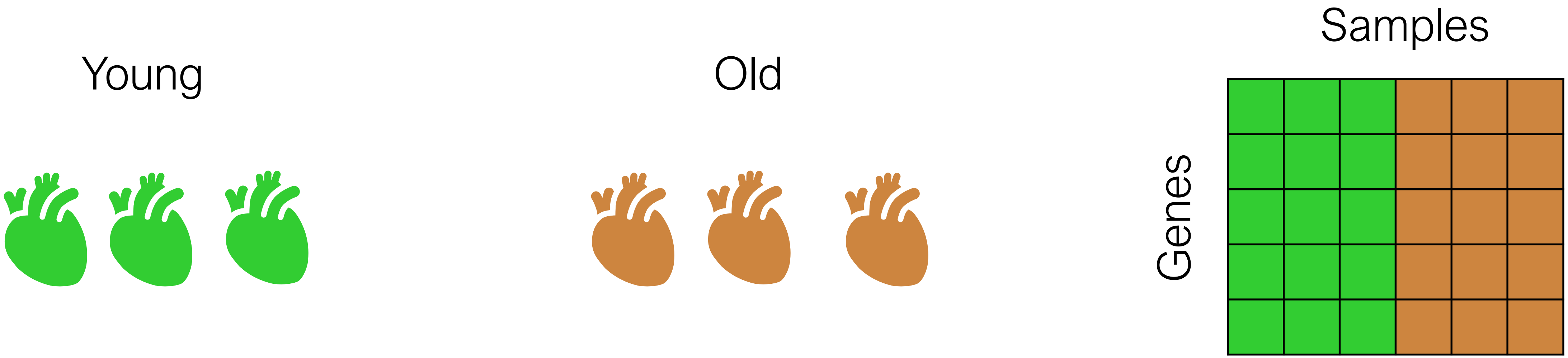
3

0

7

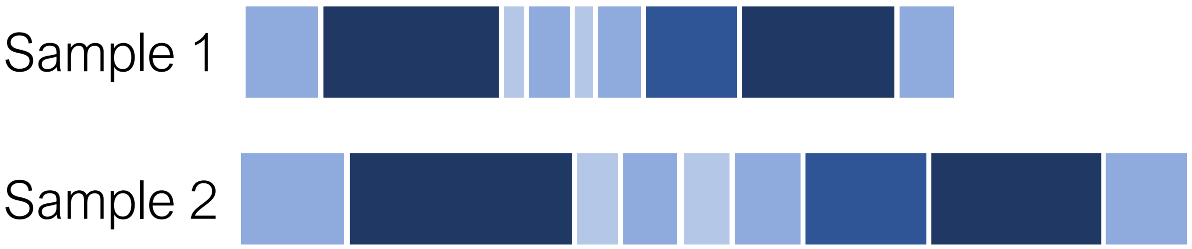
RNA-sequencing for group comparison

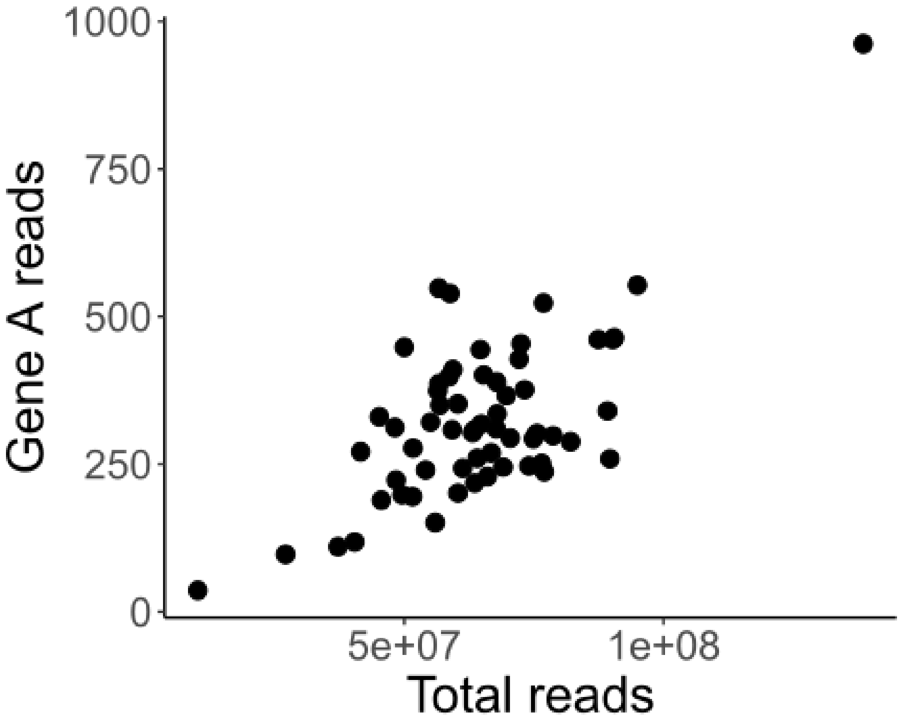
What distinguishes old hearts from young hearts?



Effect of total number of reads

Number of reads assigned to a given gene depends on the total number of reads.

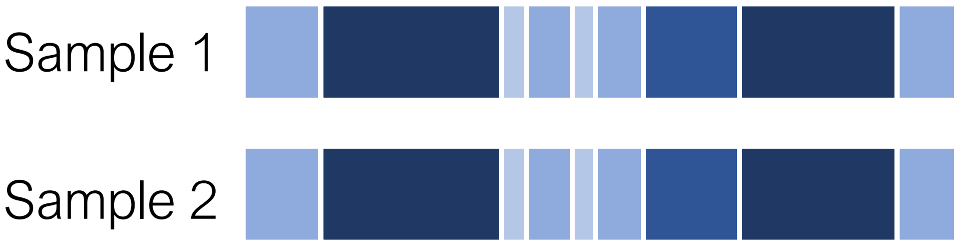


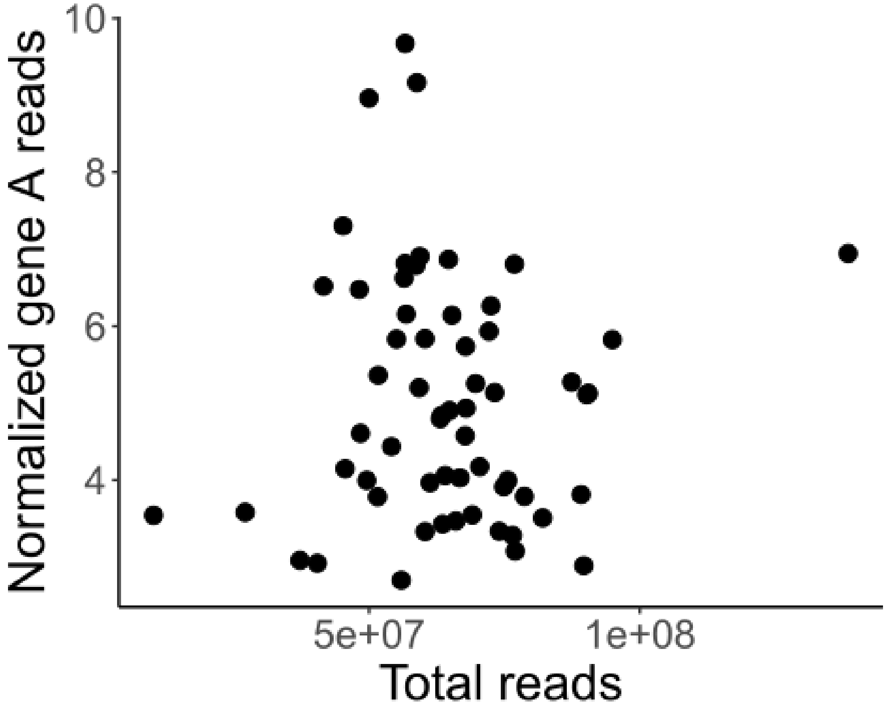


Simple solution: scale by total number of reads

Effect of total number of reads

Number of reads assigned to a given gene depends on the total number of reads.



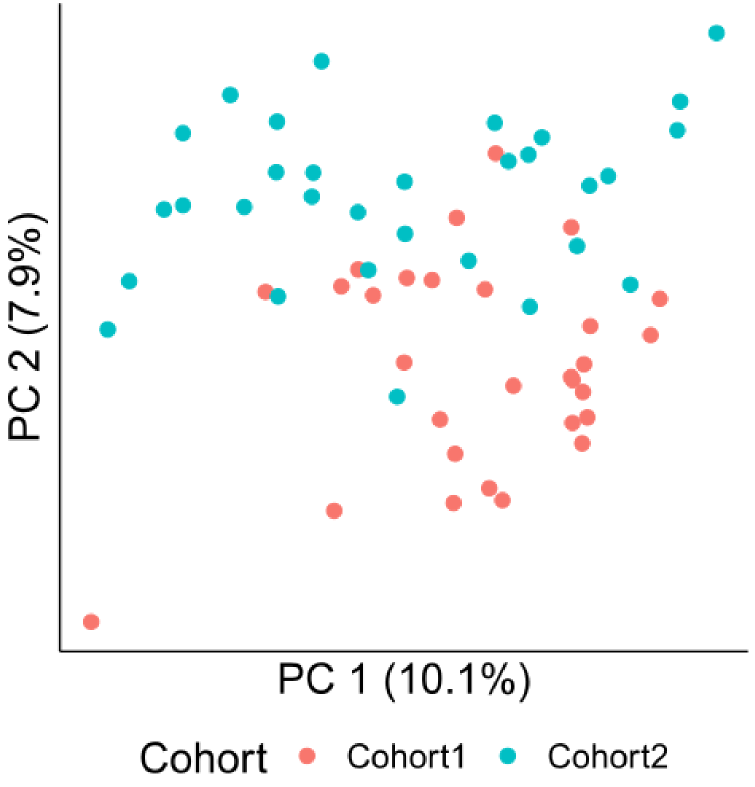


Simple solution: scale by total number of reads

Effect of confounding variables

Sources of variation in the data beyond the biologically interesting signal.

	Old	Young
Cohort 1	8	21
Cohort 2	22	9

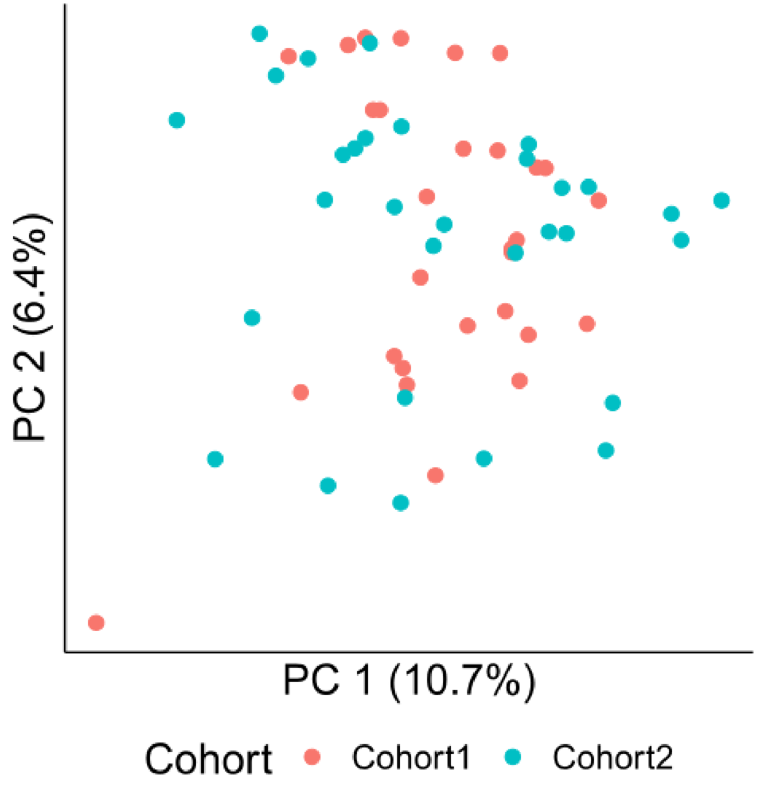


Simple solution: regress out cohort

Effect of confounding variables

Sources of variation in the data beyond the biologically interesting signal.

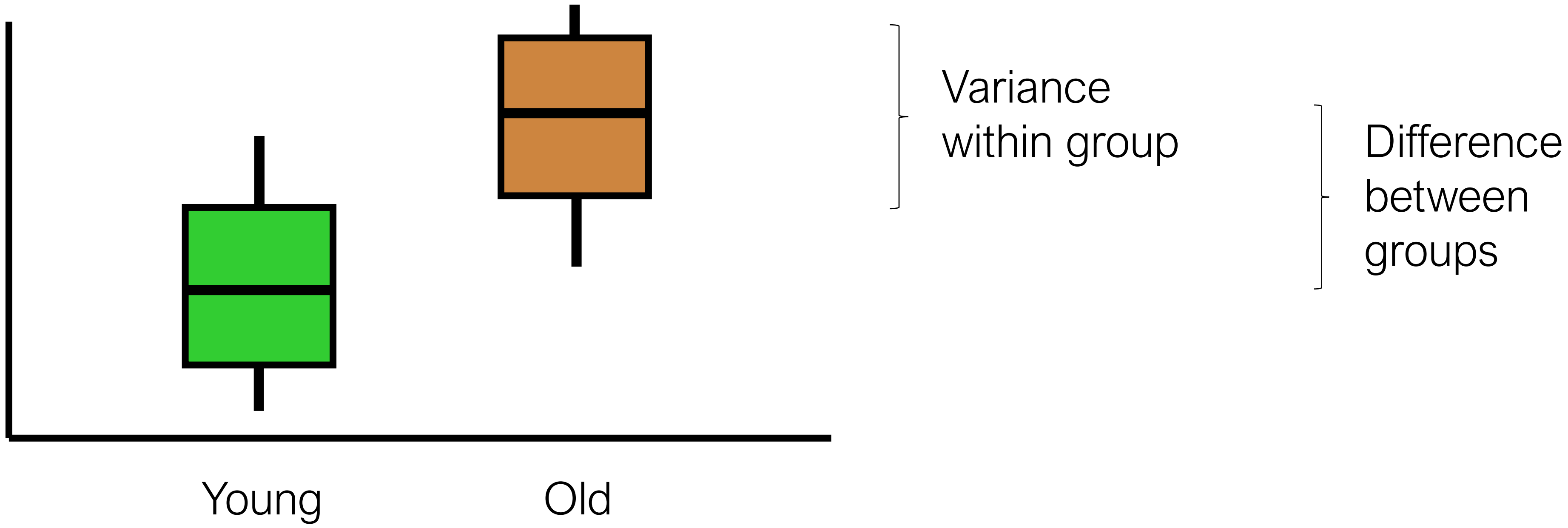
	Old	Young
Cohort 1	8	21
Cohort 2	22	9



Simple solution: regress out cohort

Differential expression analysis

Statistical quantification of expression differences between sample groups.



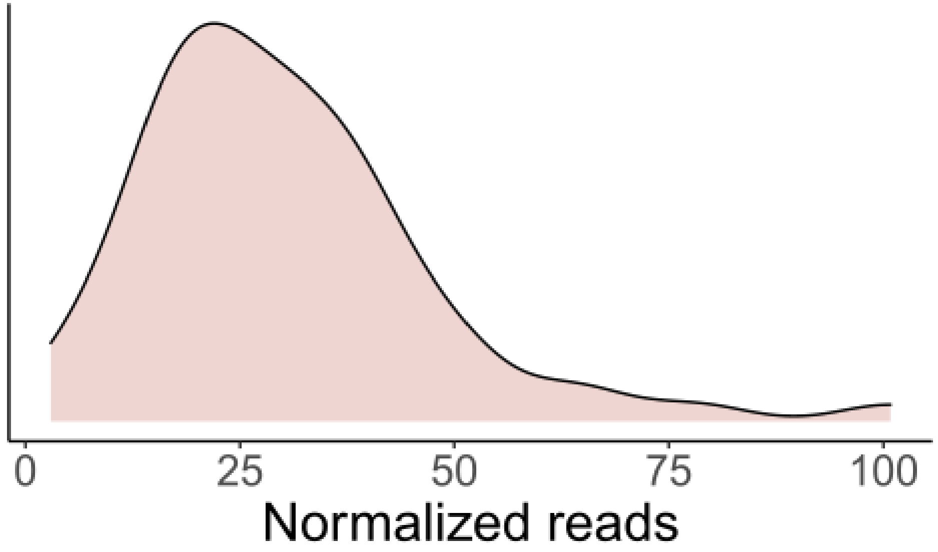
Variance within group

Difference between groups

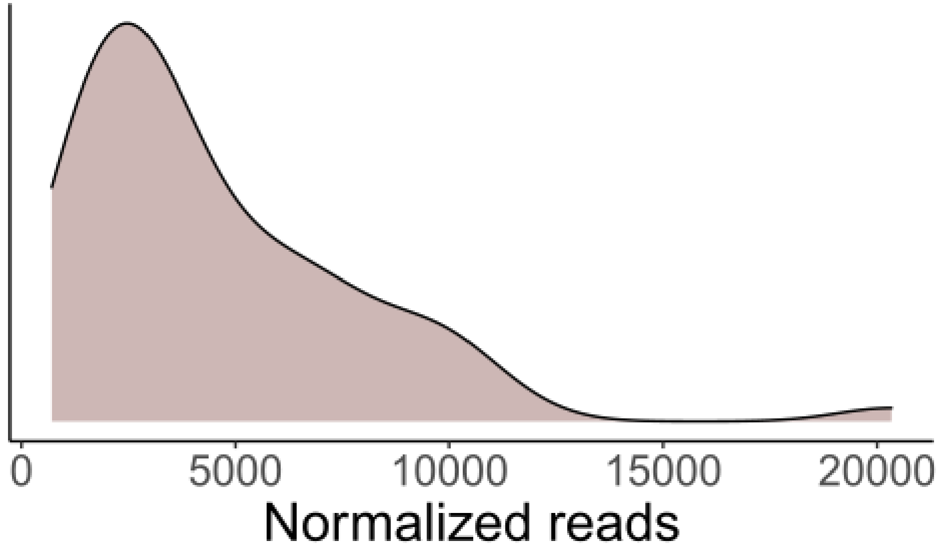
Model for data distribution

Read count data follows a Negative Binomial distribution.

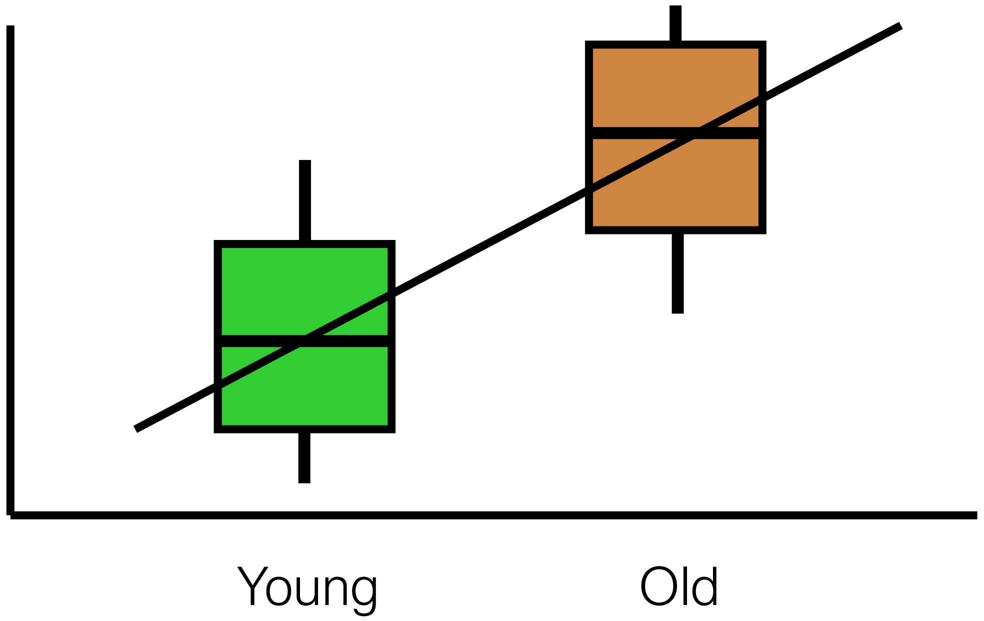
Lowly expressed gene



Highly expressed gene



Differential expression analysis



Read counts ~ Confounding Effects + Biological Effect

Take home messages

- Bioinformatics as “computer-aided biology”
- Gene expression as a read out of cellular activity
- Features of RNA-sequencing data

Q & A

?