Introduction to Bulk-RNAseq Analysis

Center for Health Data Science

Health Data Science Sandbox

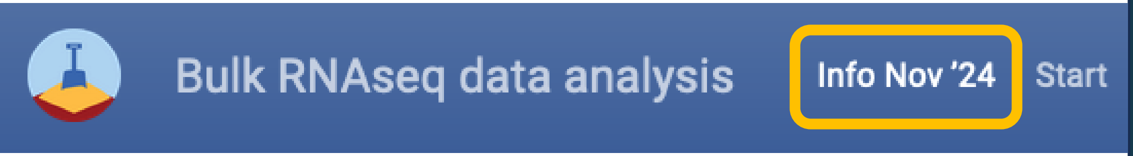
MATERIALS

Companion Website:

https://hds-sandbox.github.io/bulk_RNAseq_course/develop/

In the top menu, go to:

Info Nov ‘24



HeaDS

OVERVIEW

1. Who are we?
2. About this course
3. Experimental planning
4. Workshop data

TEACHERS

HeaDS:

SUND Data Lab

- Thilde Terkelsen
- Henrike Zschach
- Helene Wegener

HDS Sandbox

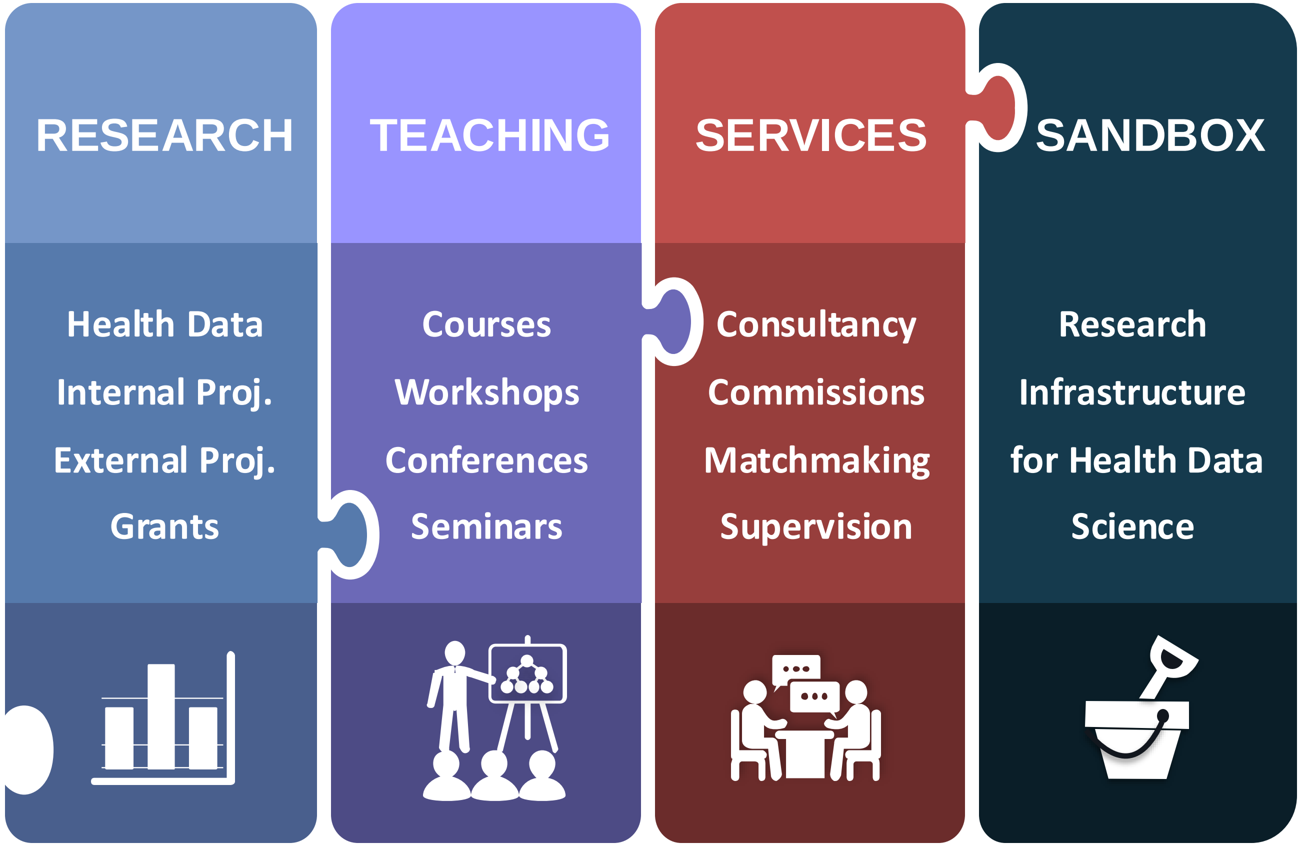
- Jennifer Bartell
- Alba Refoyo Martinez

CPR (reNEW):

- Adrija Kalvisa

CENTER FOR HEALTH DATA SCIENCE

- **Hub for health data science research**
- **Conduct Health Data Science research**
- Develop National Health **Data Science Sandbox** for Training and Research
- Host the **SUND Data Lab**



RESEARCH

Health Data
Internal Proj.
External Proj.
Grants

TEACHING

Courses
Workshops
Conferences
Seminars

SERVICES

Consultancy
Commissions
Matchmaking
Supervision

SANDBOX

Research
Infrastructure
for Health Data
Science

NATIONAL HEALTH DATA SCIENCE SANDBOX

- Gain specialized data skills with Sandbox training
- Self study - Apps on HPC (UCloud & GenomeDK)
- Many resources at hds-sandbox.github.io

Latest in-person workshops (sign up this spring!):

- **HPC-Launch (1 day)**
  Research Data Management &Computing for HDS
- **HPC-Pipes (2 days)**
  Workflow languages & environment management to build omics pipelines
- **Sandbox Genomics course (3.5 days)**
  GWAS from preprocessing to polygenic scores

Genomics
- NGS analysis
- Population genomics

HPC Lab
- Pipelines & workflow lang
- Research data mgmt

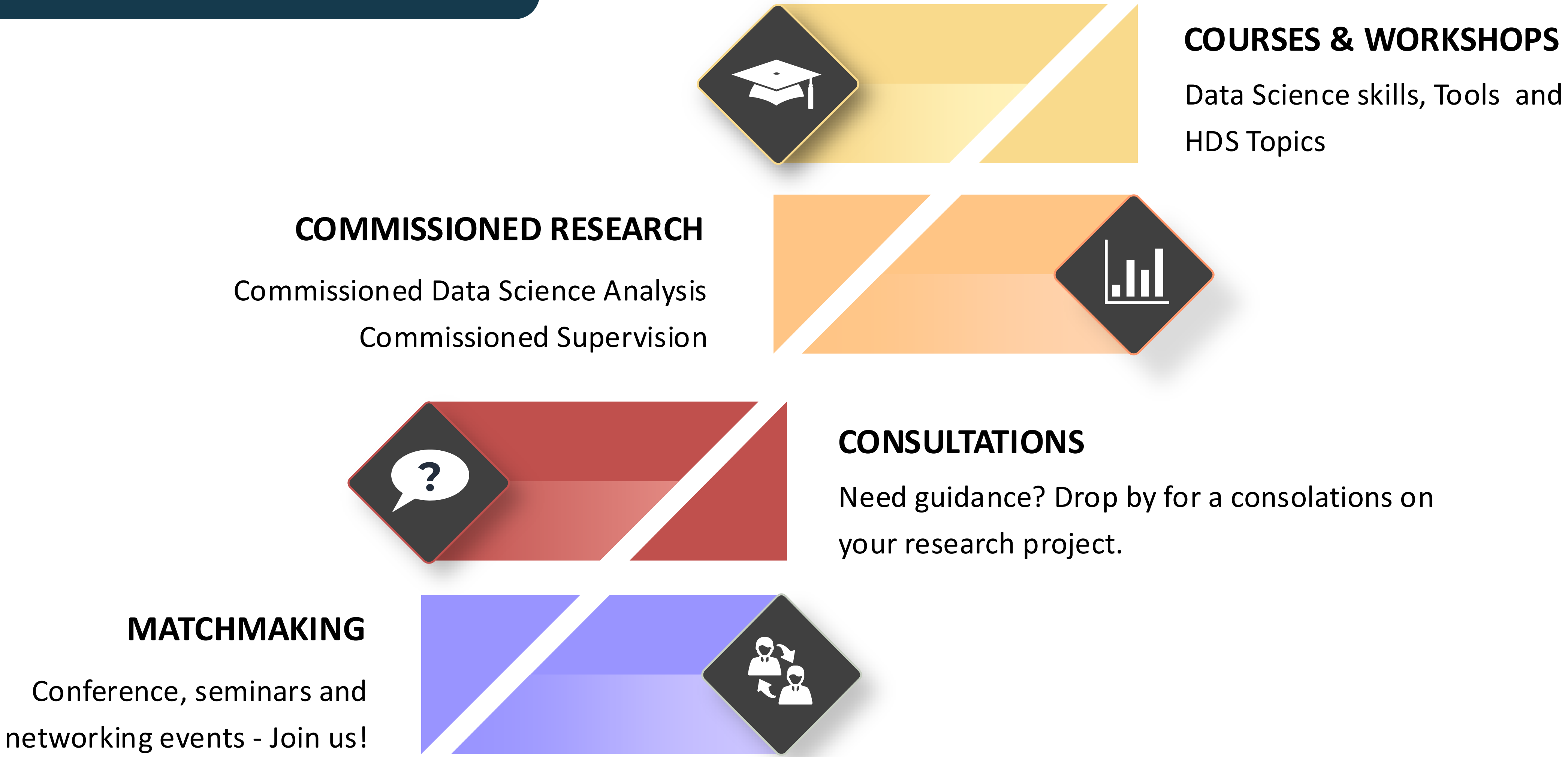
Transcriptomics
- Bulk RNA-Seq
- Single cell RNA-Seq

Health Records
- Biostatistics
- Predictive models

Proteomics
- Clinical proteomics
- CollabFold

HeaDS

SUND DATA LAB



COURSES & WORKSHOPS

Data Science skills, Tools and HDS Topics

COMMISSIONED RESEARCH

Commissioned Data Science Analysis

Commissioned Supervision

CONSULTATIONS

Need guidance? Drop by for a consolations on your research project.

MATCHMAKING

Conference, seminars and networking events - Join us!

HeaDS

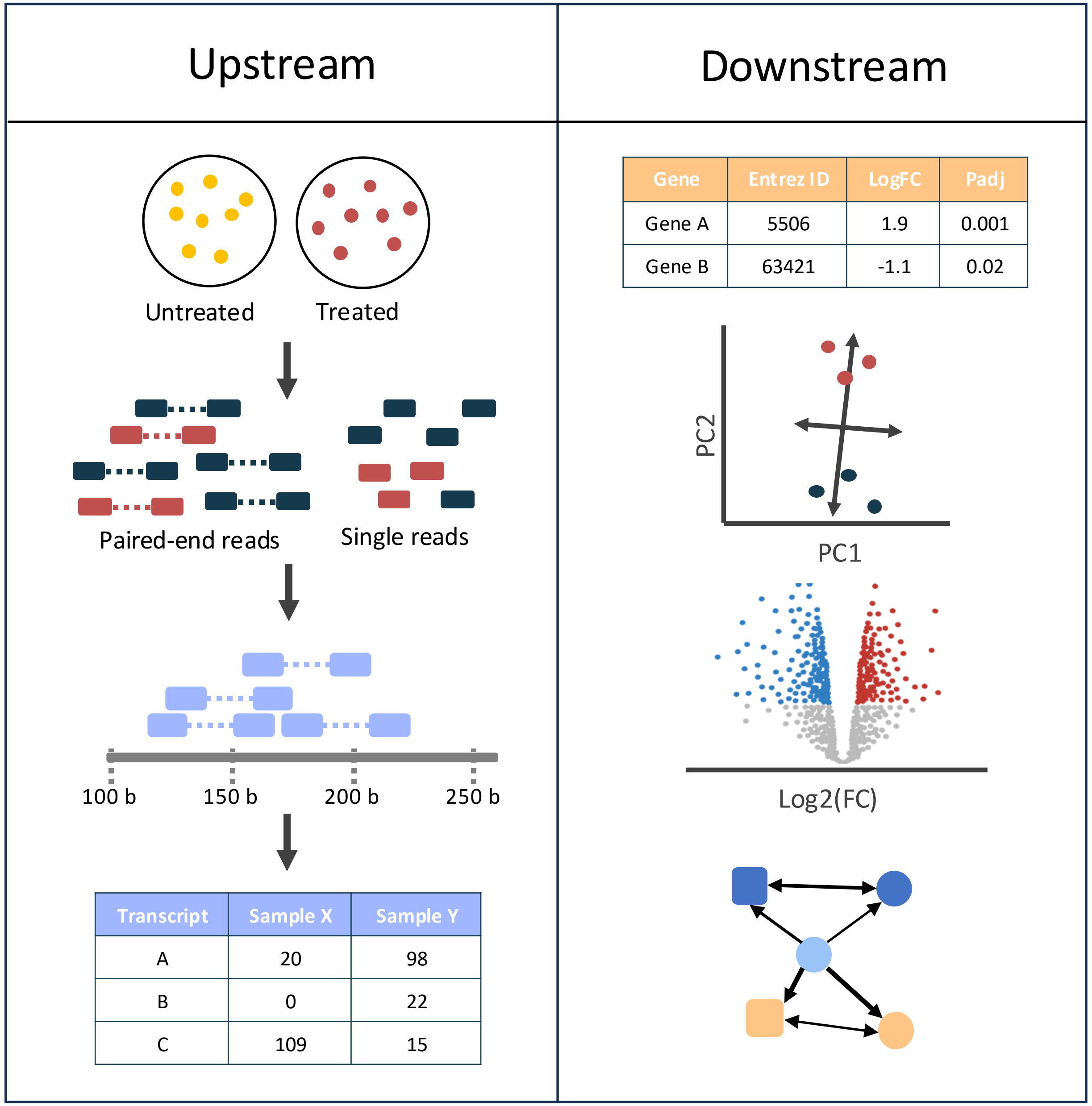
About this course

Motivation:

Teach researchers how to analyse bulk RNAseq data.

We go through:

- Experimental Design
- Pre-processing of reads
- Quality Checks
- Data Normalizing
- Exploratory Data Analysis (EDA)
- Differential Expression Analysis (DEA)
- Functional Analysis



Upstream

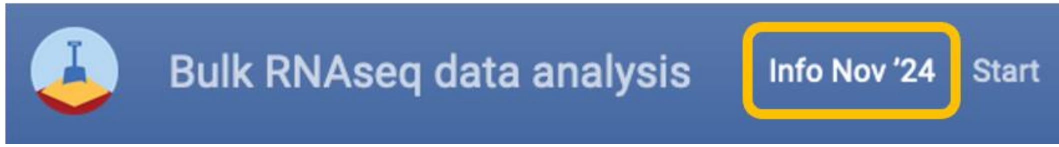
Downstream

Transcript	Sample X	Sample Y
A	20	98
B	0	22
C	109	15

Gene	Entrez ID	LogFC	Padj
Gene A	5506	1.9	0.001
Gene B	63421	-1.1	0.02

Program

Also available at :



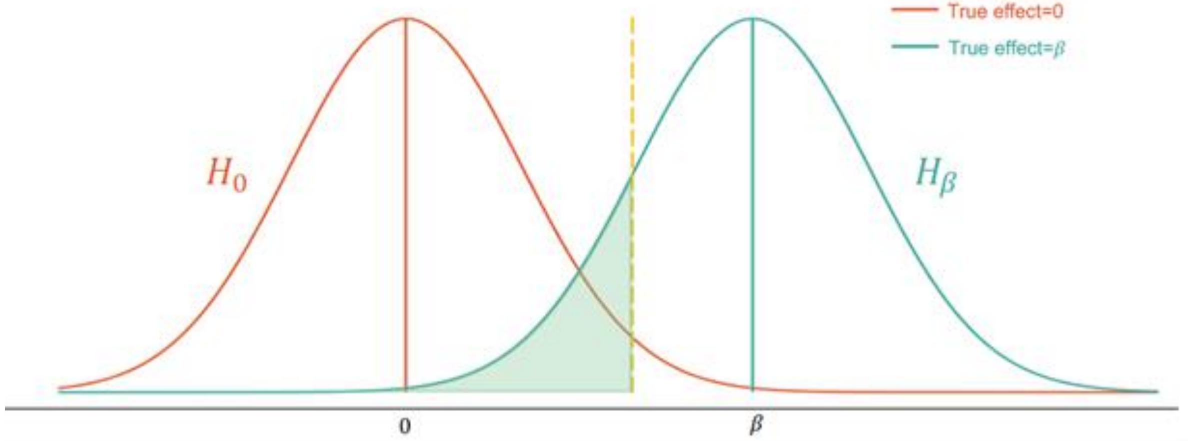
Day 1		Day2		Day3	
Time	**Subject**	**Time**	**Subject**	**Time**	**Subject**
9:00	Intro to Course	9:00	UCloud setup	9:00	UCloud setup
09:15	Experimental Design	9:30	RNAseq count matrix and normalization	9:45	DEA visualization
09:45	Preprocessing and library prep				
10:15	Break	10:15	Break	10:30	Break
10:30	Trimming, QC & Alignment	10:30	Exercise: Count Matrix	10:45	Gene annotation and databases
11:30	Feature counts & MultiQC	11:30	Exploratory data analysis	11:15	Exercise: Gene annotation
12:00	Lunch Break	12:00	Lunch	12:00	Lunch
13:00	Feature Counts & Pseudoaligners	13:00	Exercise: Exploratory data analysis	13:00	Exercise: Gene annotation
13:45	Intro to HPC and Ucloud			13:30	Functional analysis
14:30	Break	14:30	Break	14:15	Break
15:00	Nextflow pipelines and nf-core	14:45	Differential Expression Analysis	14:30	Functional analysis
16:00	Looking at pipeline result	15:15	Exercise: Differential Expression Analysis	15:15	Workflow summary
				15:30	Bring your own data
16:30	Q&A	16:30	Q&A	16:30	Wrap-up & Course evaluation

Experimental planning

Special considerations to account for before an RNAseq experiment.

Ignoring these will greatly **affect the quality** of your analysis.

1. Proper experiment **controls**
2. Number and type of **replicates**
3. Issues related to **confounding**
4. Addressing **batch effects**

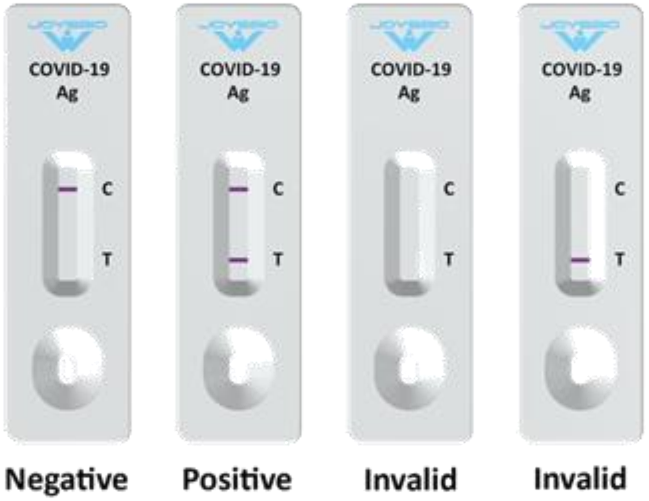


Experimental planning

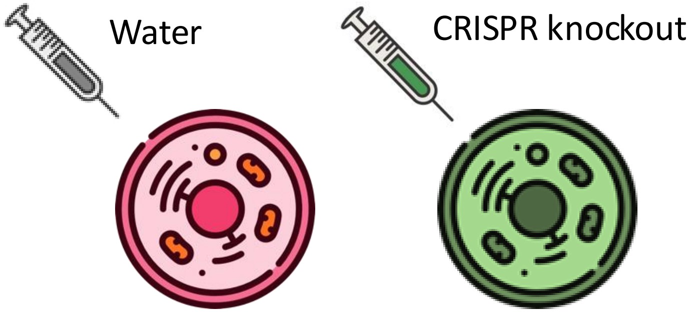
1. Proper experiment **controls**

- Minimize the effect of irrelevant variables
- Control deviations that might influence outcomes
- Types of controls:
  - **Positive:** A treatment with a known result
  - **Negative:** No response is expected

Covid tests: Positive control



Gene knockout: Negative control



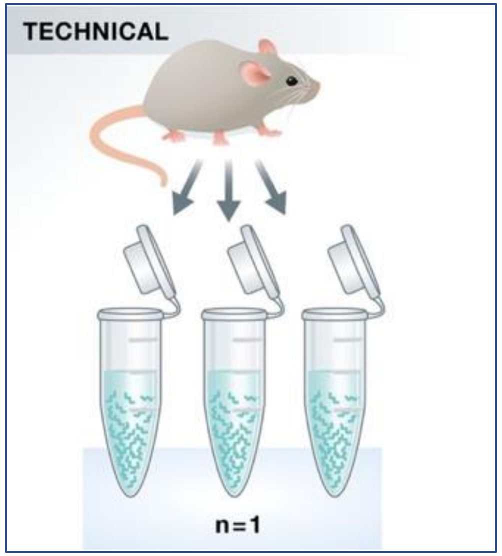
Experimental planning

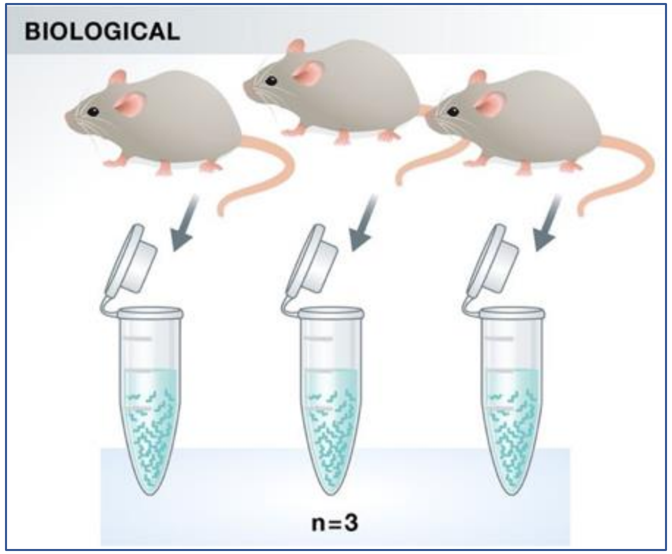
2. Number and type of **replicates**

- Needed to account for variation between samples

More biological replicates equals to:

- Better estimates of biological variation
- Improved model accuracy
- Power to detect differentially expressed genes



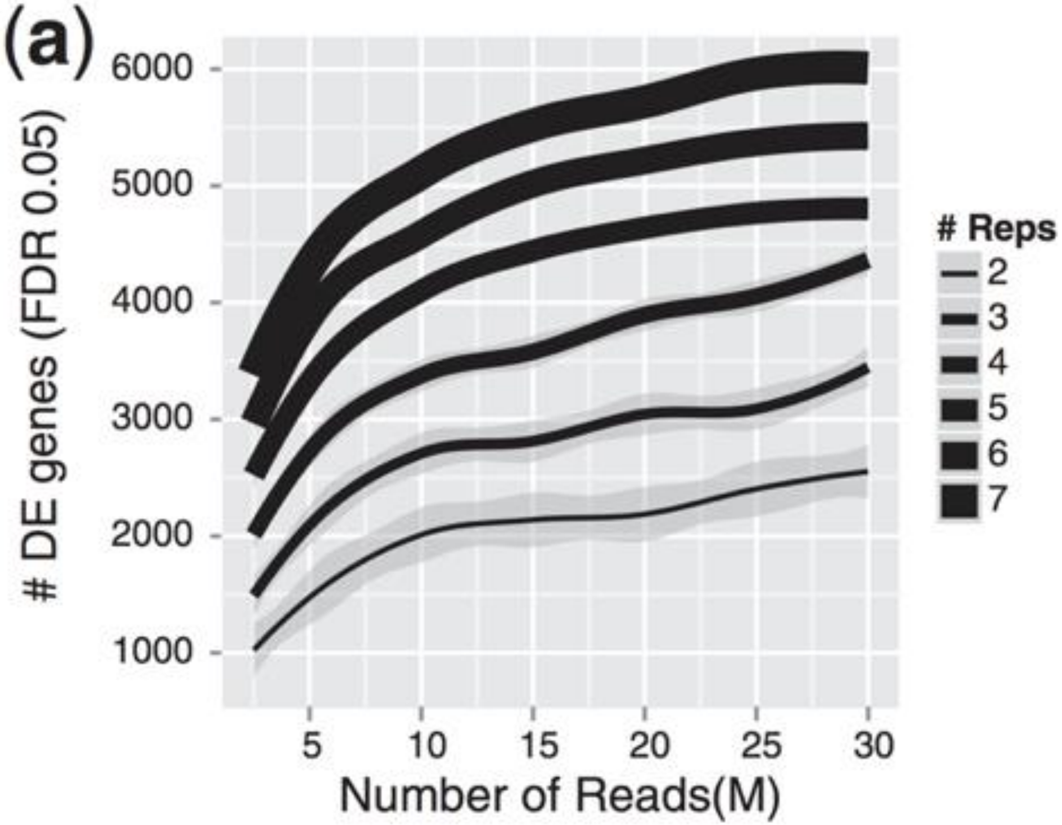


Experimental planning

2. Number and type of **replicates**

- Biological replicates **>** sequencing depth
  - Depends on your experiment

- More replicates **=** more differentially expressed genes **with greater confidence**

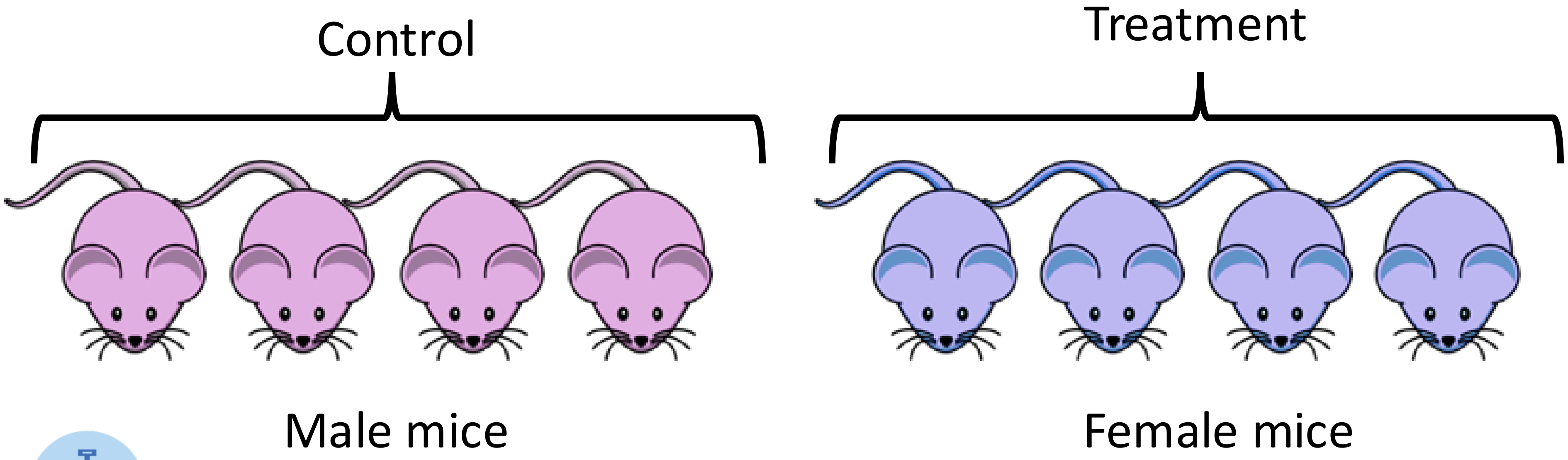


Liu, Y., et al., Bioinformatics (2014) **30**(3): 301–304

Experimental planning

3. Issues related to **confounding**

Avoid situations where we **cannot distinguish the separate effects of two different sources of variation**.



Control

Treatment

Male mice

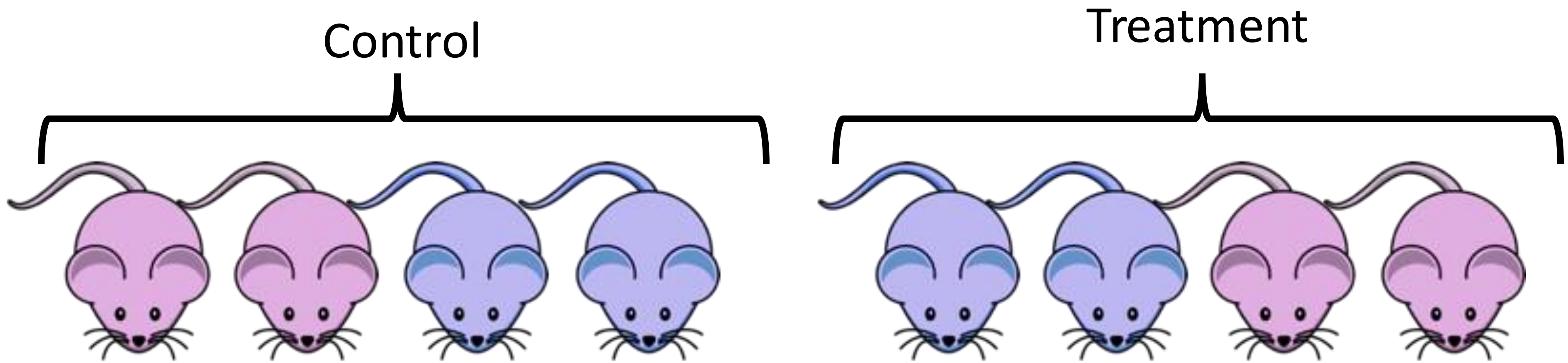
Female mice

Experimental planning

3. Issues related to **confounding.**

Avoid this by **randomizing** your design**:**

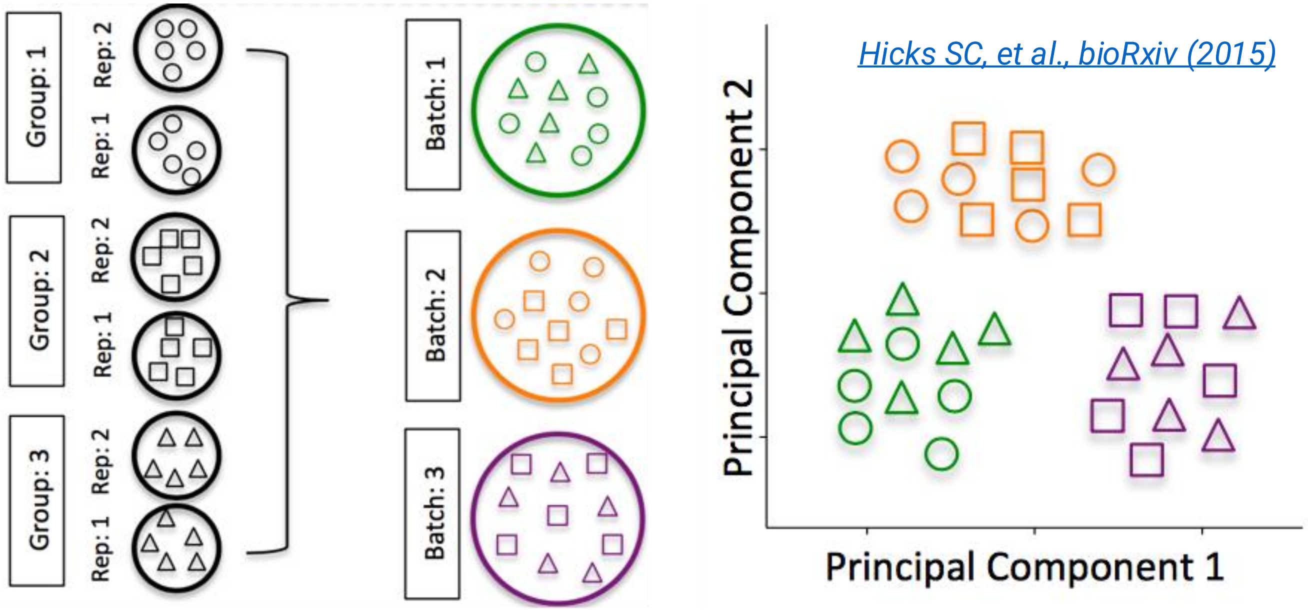
- Sample collection
- Treatment / Exposure



Experimental planning

4. **Batch effects**

Unwanted variance as consequence of technical issues such as sample collection, storage, experimental protocol, etc.



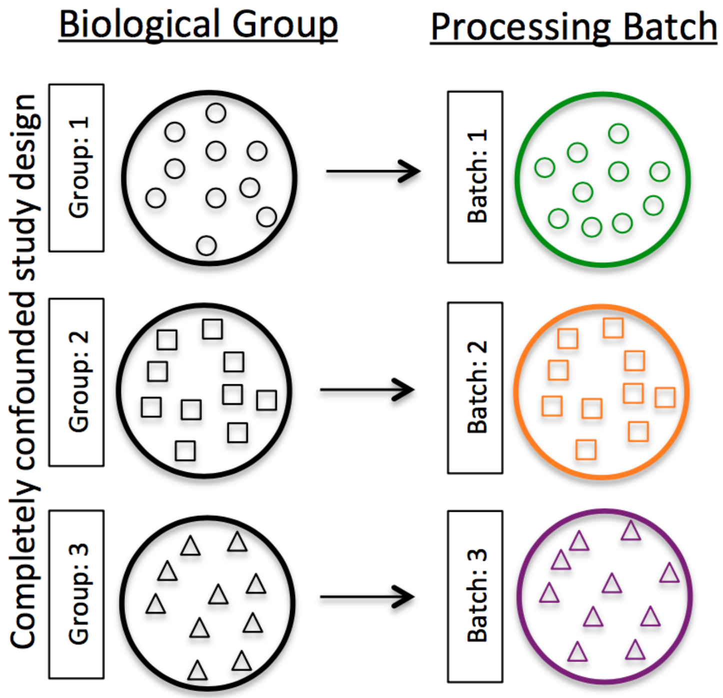
Hicks SC, et al., bioRxiv (2015)

Experimental planning

4. Addressing **batch effects:**

- Experimental **setup**
- **Technical equipment / reactants**
- **Processing** batch (days / people / place)

Batches can be corrected statistically **only** if they are **not confounded**!



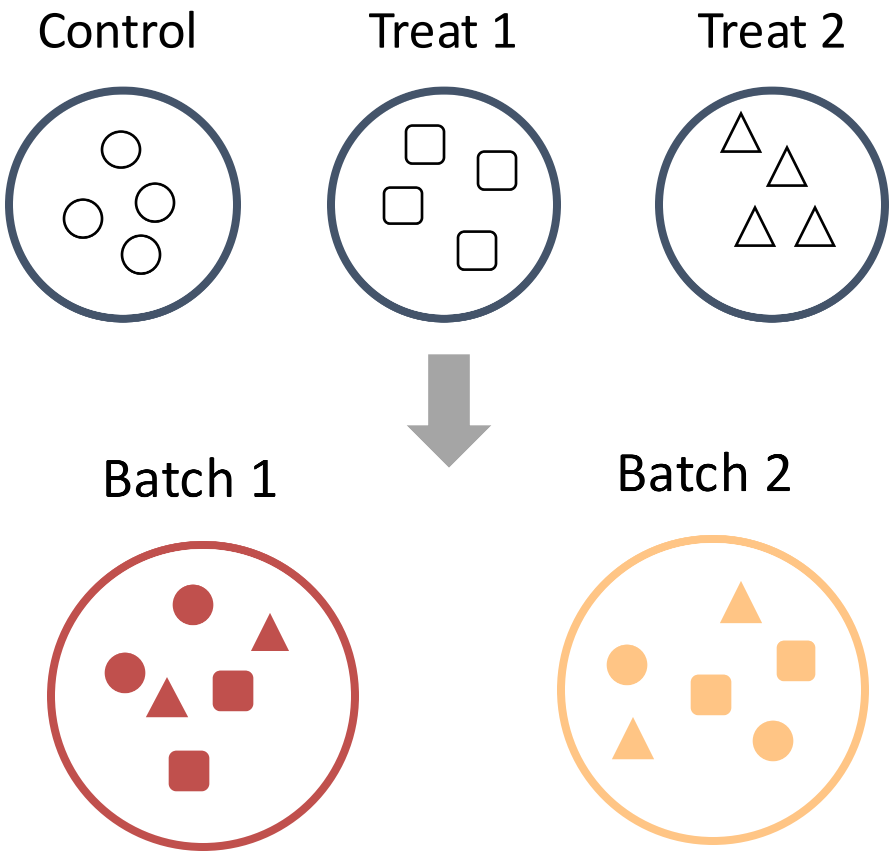
Hicks SC, et al., bioRxiv (2015)

Experimental planning

4. Avoid confounding:

- Unbiased data collection
- Groups are **balanced**
- Samples are **randomized**
- Batch information is recorded

(who, when, where, plate, machine)



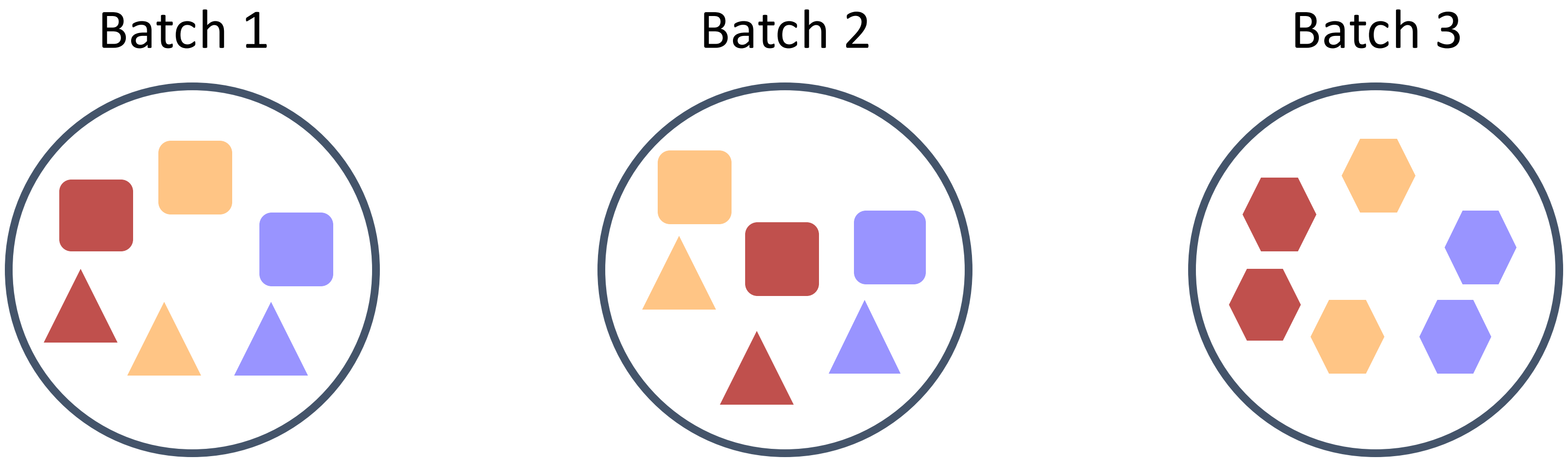
Quiz

- What is a technical replicate?
- What is a biological replicate?
- What is partial and complete confounding?

Quiz

Is this experimental design good? Why or why not?

■ Treatment 1 ▲ Treatment 2 ⬢ Control Color = Cell line



Batch 1

Batch 2

Batch 3

Quiz

Is this experimental design good? Why or why not?

■ Control sample ▲ Cancer sample Color = Patient

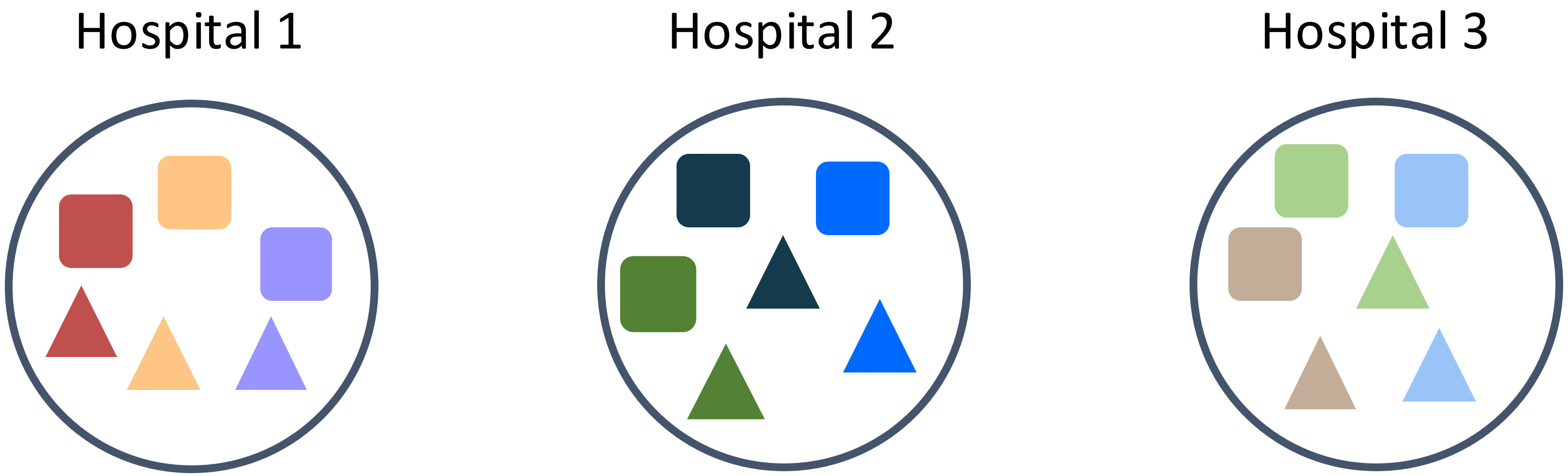


Table discussion

You are a doctor who has discovered the secret to eternal life! Once you inject your mice with the secret substance Vampirium they live 50% longer than controls (but unfortunately they also start to crave drinking blood).

Before you go into the phase 1 human trial, you think you should just check how the transcriptome is affected by Vampirium injection.

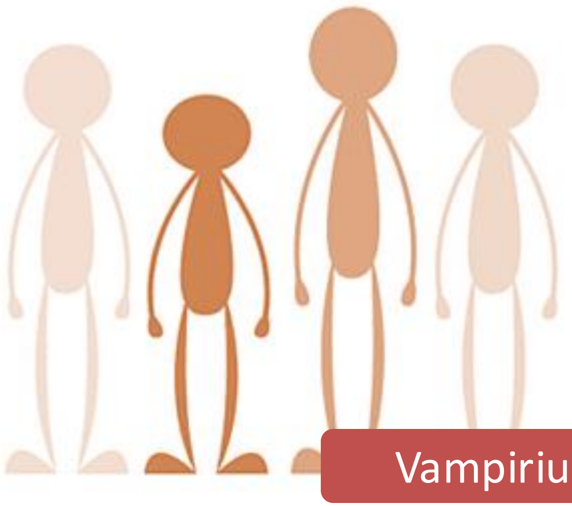
Discuss what should be the set-up for your experimental design? Consider:

- Cell lines vs. live mice
- Cases, Controls & Replicates (technical, biological)
- How to minimize confounding and batch effects

Workshop data

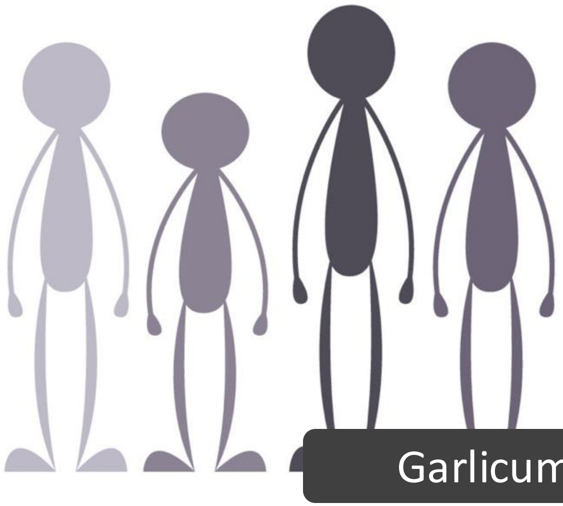
RQ: What happens to the gene expression profiles of subjects after injection with vampirium? Does the drug (Garlicum) work to combat the ‘unlucky side effect’ of vampirism?

Biting and drinking blood



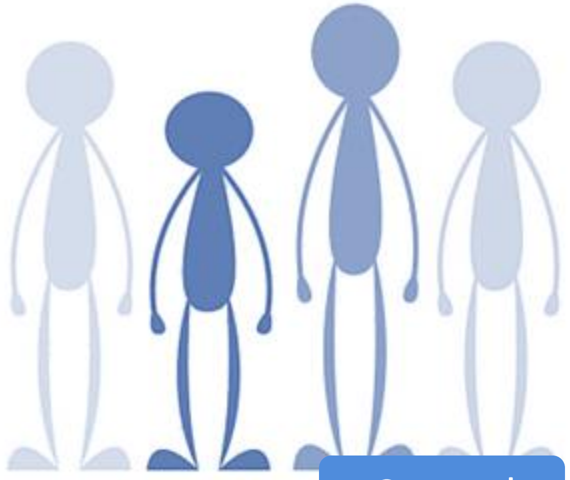
Vampirium

Drug to combat disease



Garlicum

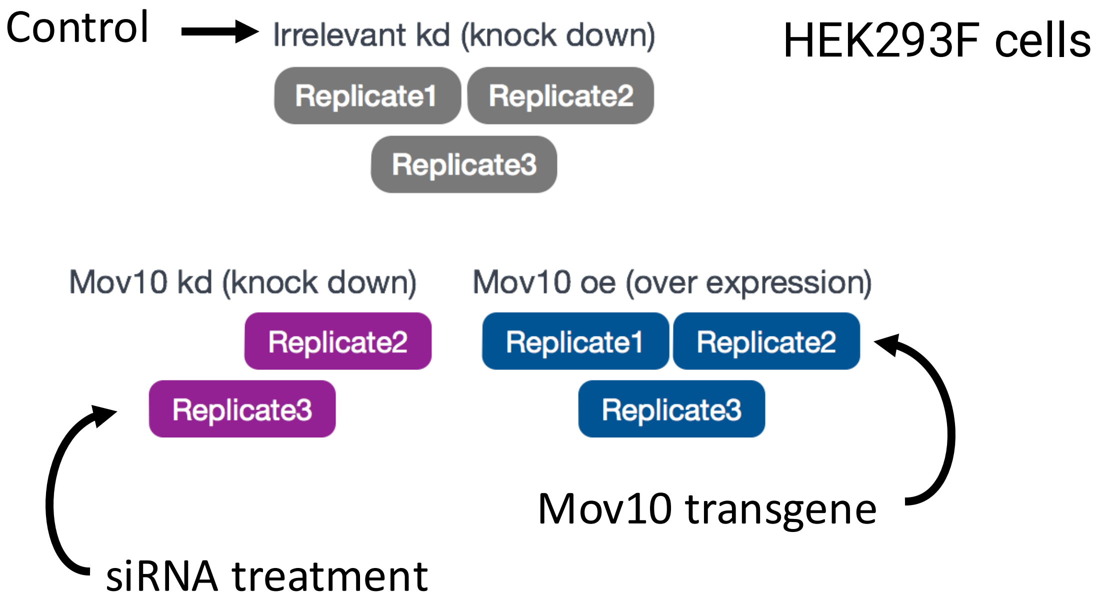
Normal samples



Control

Workshop data

- Starting from **sequencing reads** (Kenny PJ et al, Cell Rep 2014)
- **RQ:** Fragile X syndrome association between Mov10 and FMRP gene



Control → Irrelevant kd (knock down)

Replicate1 Replicate2

Replicate3

HEK293F cells

Mov10 kd (knock down)

Replicate2

Replicate3

siRNA treatment

Mov10 oe (over expression)

Replicate1 Replicate2

Replicate3

Mov10 transgene