DATALAB
SCIENCE FOR HEALTH

Introduction to experimental design in “omics” experiments

Bioing. Elmer A. Fernández (PhD)
DataLab
CONICET

elmer.fernandez@unc.edu.ar,
elmerfernandez@fpmlab.org.ar

CONICET

*

FUNDACIÓN PARA EL PROGRESO DE LA MEDICINA

Where are we?

We have an hypothesis that we need to verify…So, we need to define an experiment…

What does it mean?

What do we need to know?

What are the possible scenarios ?

Experimental scenarios

We need to generate the data

Design an experiment with a suitable experimental framework that meets all relevant statistical requirements.

We will use already existing data

Review existing research and identify samples that closely align with my requirements.

Extract detailed information on the samples and the data generation methodologies used.

Experimental scenarios

We will use already existing data

Review existing research and identify samples that closely align with my requirements.

Extract detailed information on the samples and the data generation methodologies used.

We need to generate the data

Design an experiment with a suitable experimental framework that meets all relevant statistical requirements.

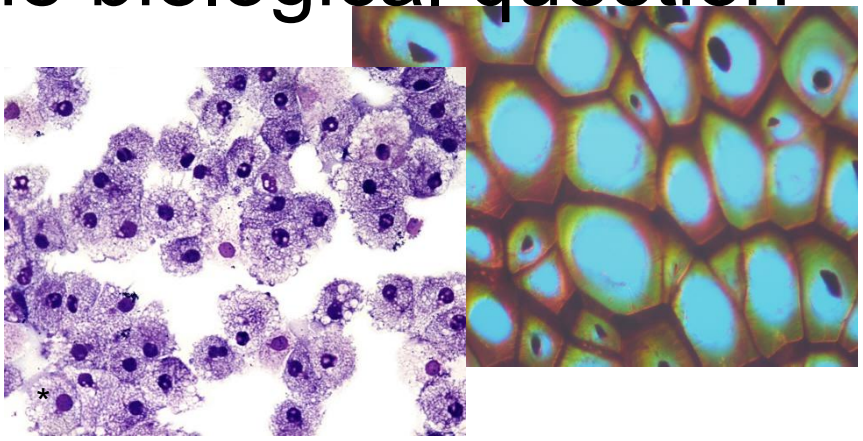
My suggestion? use both approaches

Why?

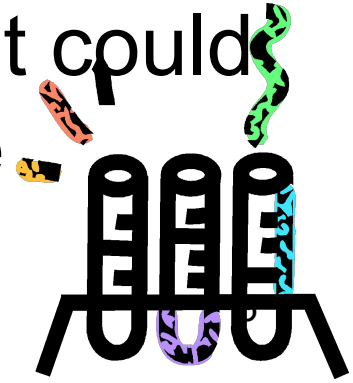
- Experimental Design does not mean how to split the samples over the experimental conditions or field

but carefully devise

the biological question



and potential effects that could impact the answer.

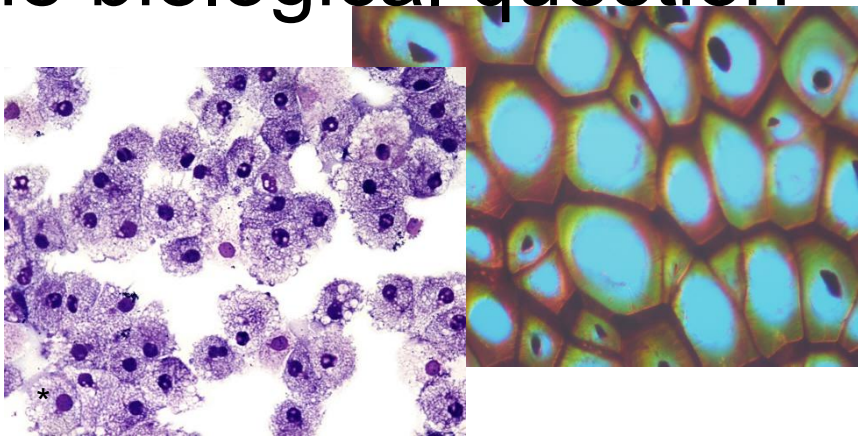


Why?

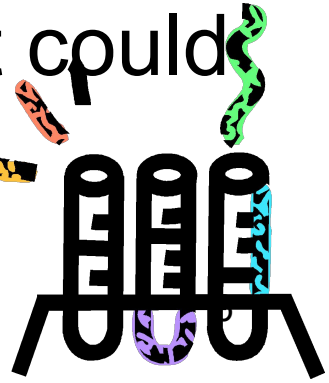
These steps are valid for both experimental scenarios

but carefully devise

the biological question



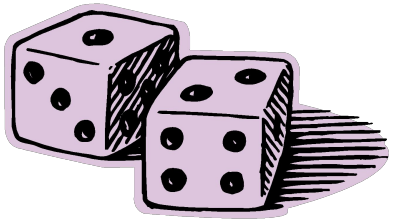
and potential effects that could impact the answer.



Why? (in scenario 1)

- This will provide clues about what **can be controlled**

and what would not



and plan in accordance

Why? (in scenario 2)

- This will provide clues about what **things should I take care about**

and figure it out of what should I pay attention to

*

So..

- Share your ideas with the whole team at the very beginning
- This means:
  - Explain, discuss and plan your ideas with the whole team
    - Current and future ones
    - Good and Bad Errors
  - Explain and discuss your ideas with your technology providers. The walls of your lab are not longer at hand
    - This will bring light to your possibilities

Experimental Design

- Proper experimental design is needed to ensure that questions of interest can be answered and that can be done accurately, given the experimental constraints, such as
  - cost of reagents
  - sample availability
  - mRNA availability
  - geographical, etc.

Experimental design involves developing a comprehensive plan that outlines the objectives, variables, sample size, procedures, and data collection methods, ensuring the experiment is NOT ONLY statistically sound and unbiased, but also **feasible**

Experimental Design, Why?

- In the absence of a proper design, it is essentially impossible to partition **biological variation** from **technical variation**.
- When these two sources of variation are **confounded**, there is no way of knowing which source is driving the observed results.
- No amount of statistical sophistication can separate confounded factors after data have been collected

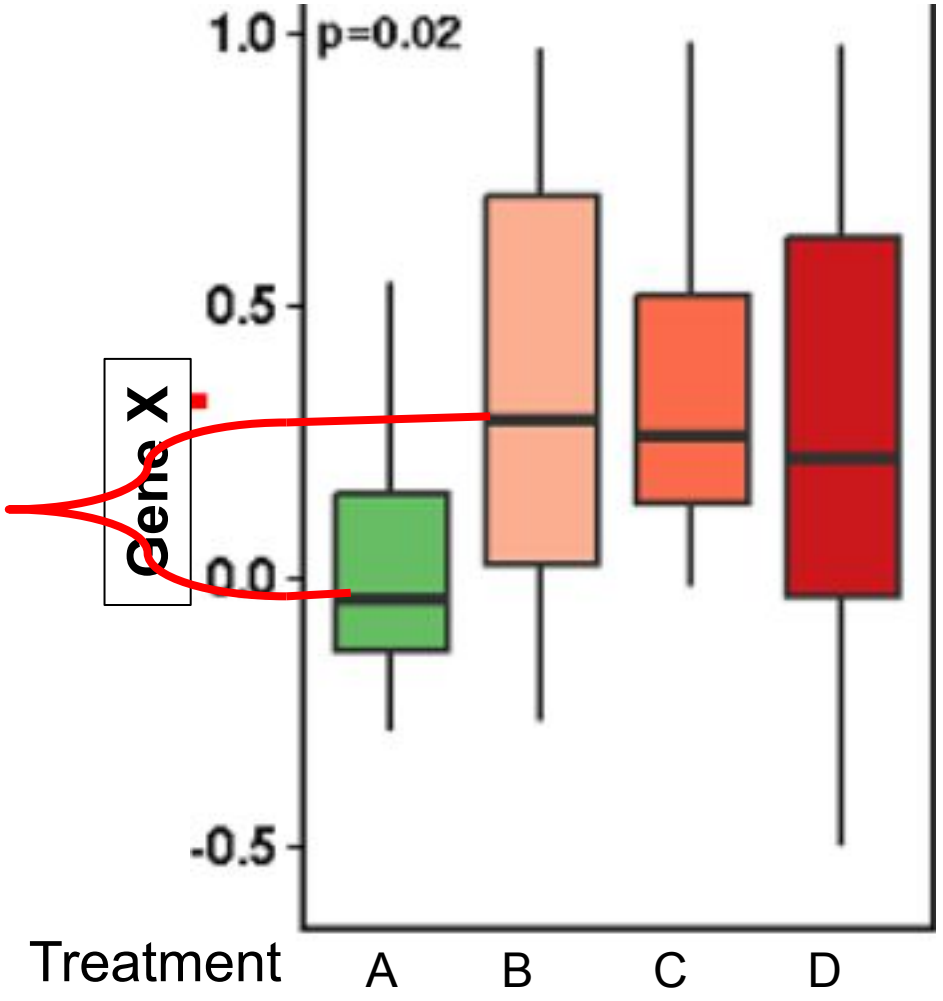
Confounding Factors in Experimental Design

- Confounding factors, also known as confounders, are variables that can influence both the independent variable (the variable being manipulated) and the dependent variable (the outcome being measured) in an experiment
- confounding factor creates a false association between the independent and dependent variables by distorting the true relationship between them

This is a key concept to account when you define your experiment in order to define your sample object

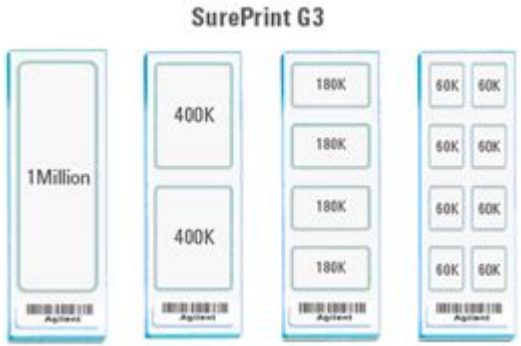
variability

LogFC



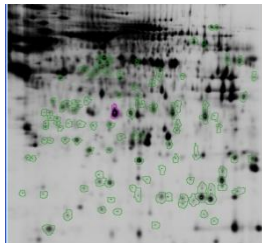
What is experimental design in "omics"

- Identify and understand the problem
- Analyze the available population characteristics. (size, weight, etc. All those characteristics that could be of interest in your hypothesis.)
- Allocate the target samples to the slides

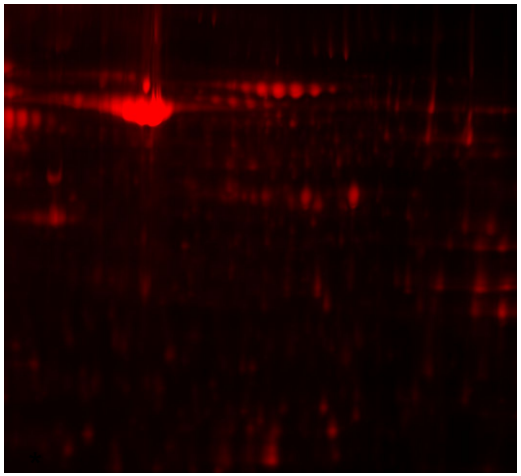


What is experimental design in "omics"

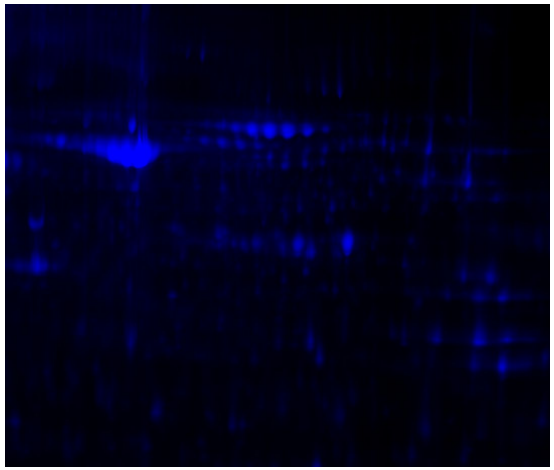
- Allocate the target samples in a gel



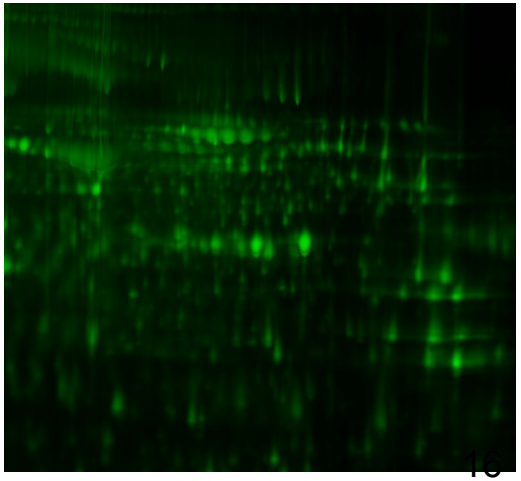
Cy3



Cy2

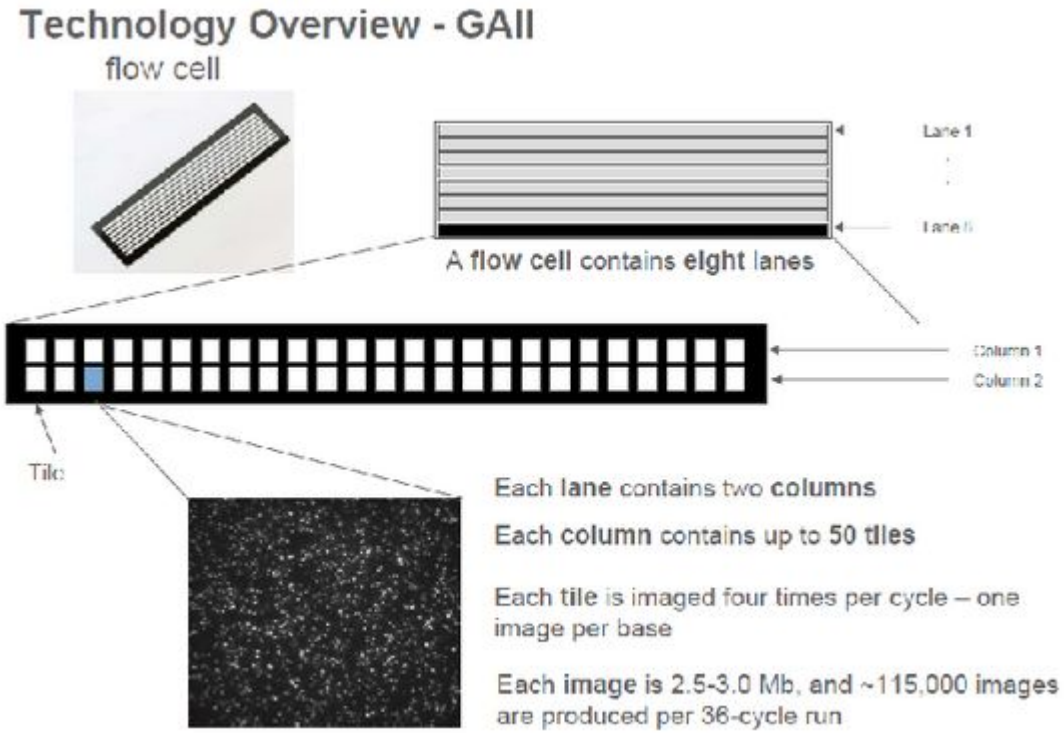


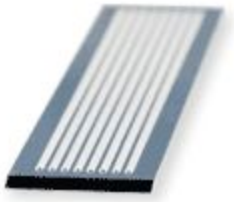
Cy5



What is experimental design in "omics"

- Allocate the target samples in a lane/picotiter plate/multiplexing





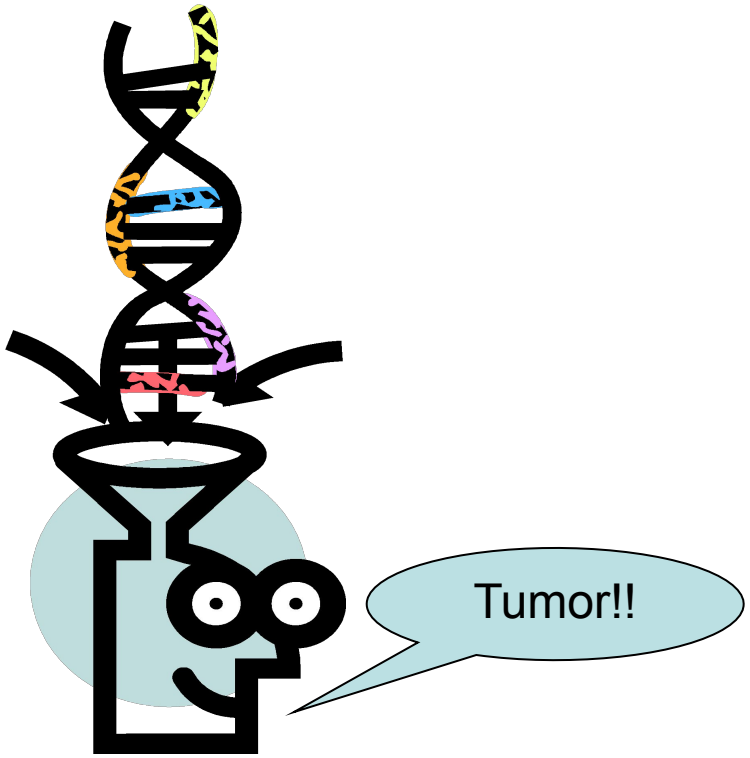
Identify the problem

- Class comparison: Treatment-Controls studies

*

Identify the problem

- Class prediction: molecular signatures

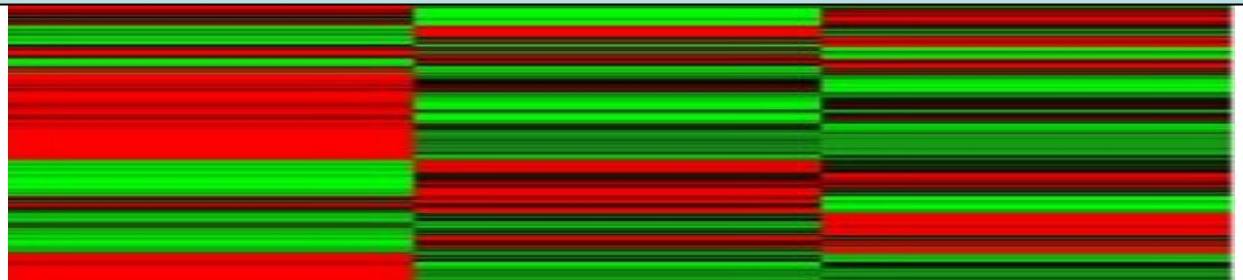


*

Identify the problem

- Class discovery

Each problem has several potential solution alternatives, assumptions and potential tools that you must know in advance



Indentify the problem: The platform

- Usually
  - Microa
    - SNPs
    - Metilation
  - Proteomics:
    - DIGE
    - Silver
    - Protein Arra
  - NGS
    - GBI
    - ILLUMINA
    - PGM-ION P

Each platform has several potential applications

The source

DNA-Seq
- Copy Number
- SNP
- Structural variants
- Whole genome sequencing
- Metagenomics
- Targeted/Amplicon Sequencing

ChIP-Seq
- Transcription Factor binding sites
- Methylation sites
- Histone modifications
- RIP-Seq (RNA-binding proteins)

RNA-Seq Transcriptome
- Differential Gene Expression
- Alternative Splicing
- SNP detection
- Indel detection
- Novel exons/genes

miRNA-Seq
- identify regulatory (non-coding) RNAs

Identify the problem: Variation Sources

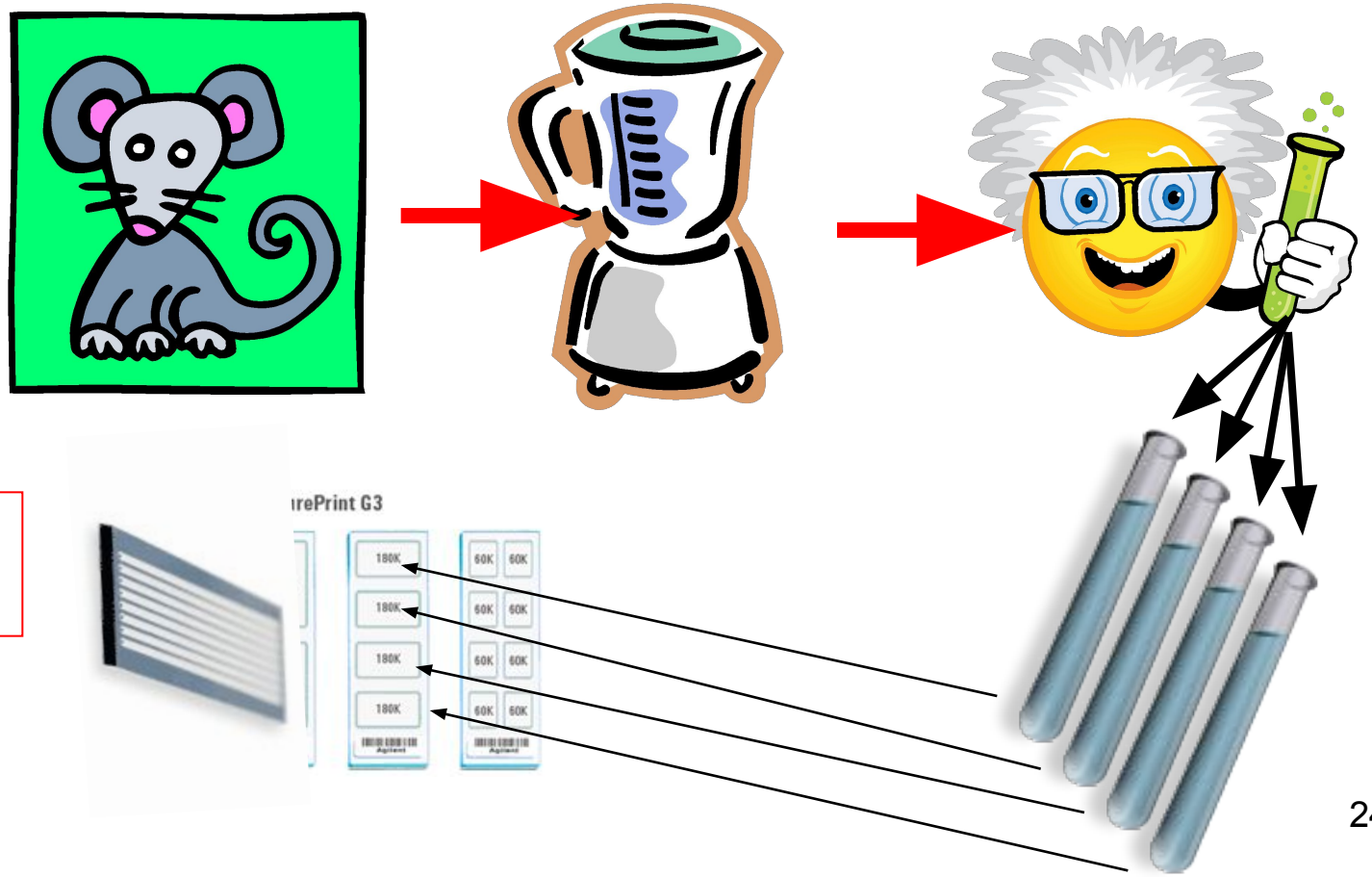
- ALL
  - Technician, day, ozone, batch,…
- Biological one is what we care about!!
- BUT ! Data are affected by both, the technical variability and the Biological one.
- By design we can diminish the technical one and allows a good estimation of the biological one.

Samples

- Replication in experimental design refers to the process of repeating an experiment or study under the same conditions
- Replication: is essential for:
  - estimating and decreasing the experimental error, and thus to detect the biological (treatment) effect more precisely.
  - Population inference
- A true replication is an **independent** repetition of the **same** experimental process and independent acquisition of the observations

Samples

- Technical re **Whats for?**



Each to one chip/gel/lane

*

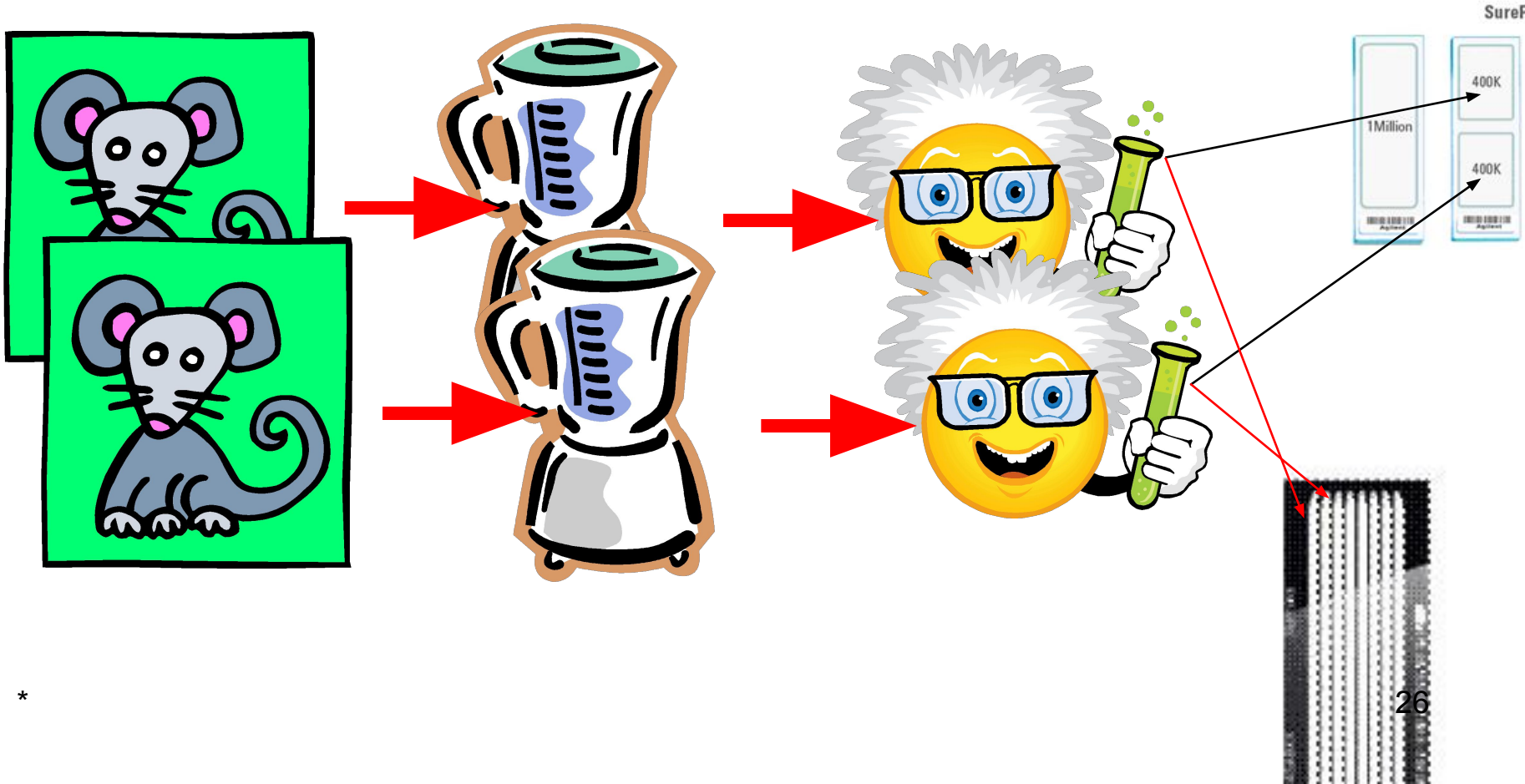
Technical replicates

Technical replication involves repeating **the same experiment** using the **same biological sample multiple times**. These replicates help assess the consistency and reliability of the experimental technique or measurement process itself, rather than biological variability.

- **Purpose:** The main goal is to ensure that the results are not due to random errors in the experimental procedure, such as pipetting errors, machine calibration issues, or other technical inconsistencies.
- **Example:** If a researcher measures the expression of a gene in a cell line using PCR, technical replication would involve running the PCR multiple times on the same RNA sample to ensure that the measurement is consistent.

Samples

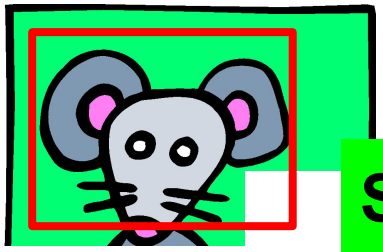
- Biological replicates

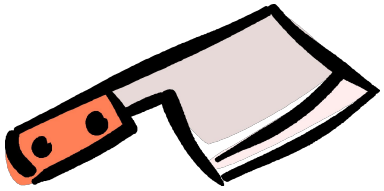


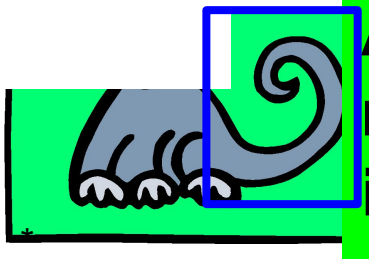
*

Samples

- Biological replicates (Paired)



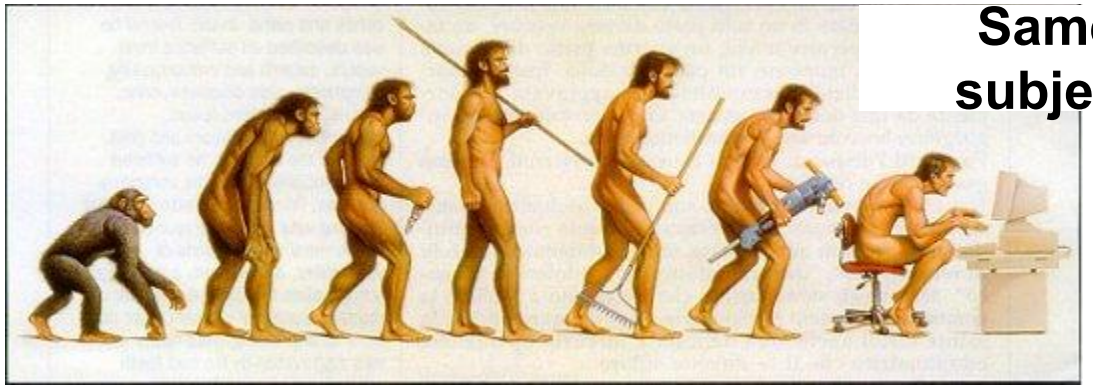




Suppose we want to test two different formulations of a soap on hair strength. In this case, each mouse has one formulation applied to the hair of the head and the other formulation applied to the hair of the tail. After the treatment period, hair strength is measured on head and tail for each individual

Samples

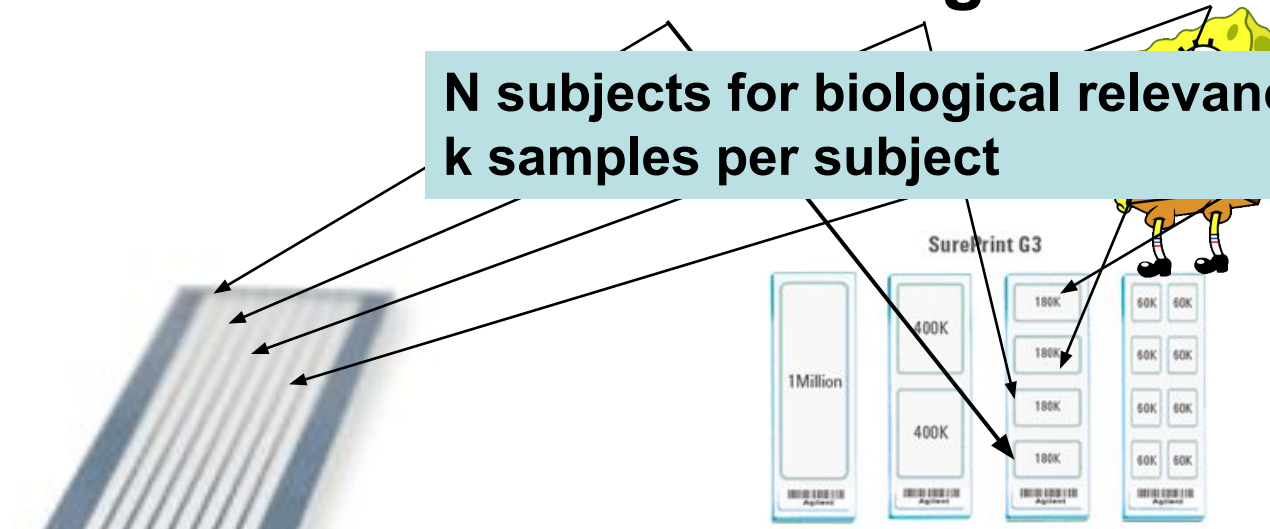
- Biological replicates (longitudinal)



Same subject

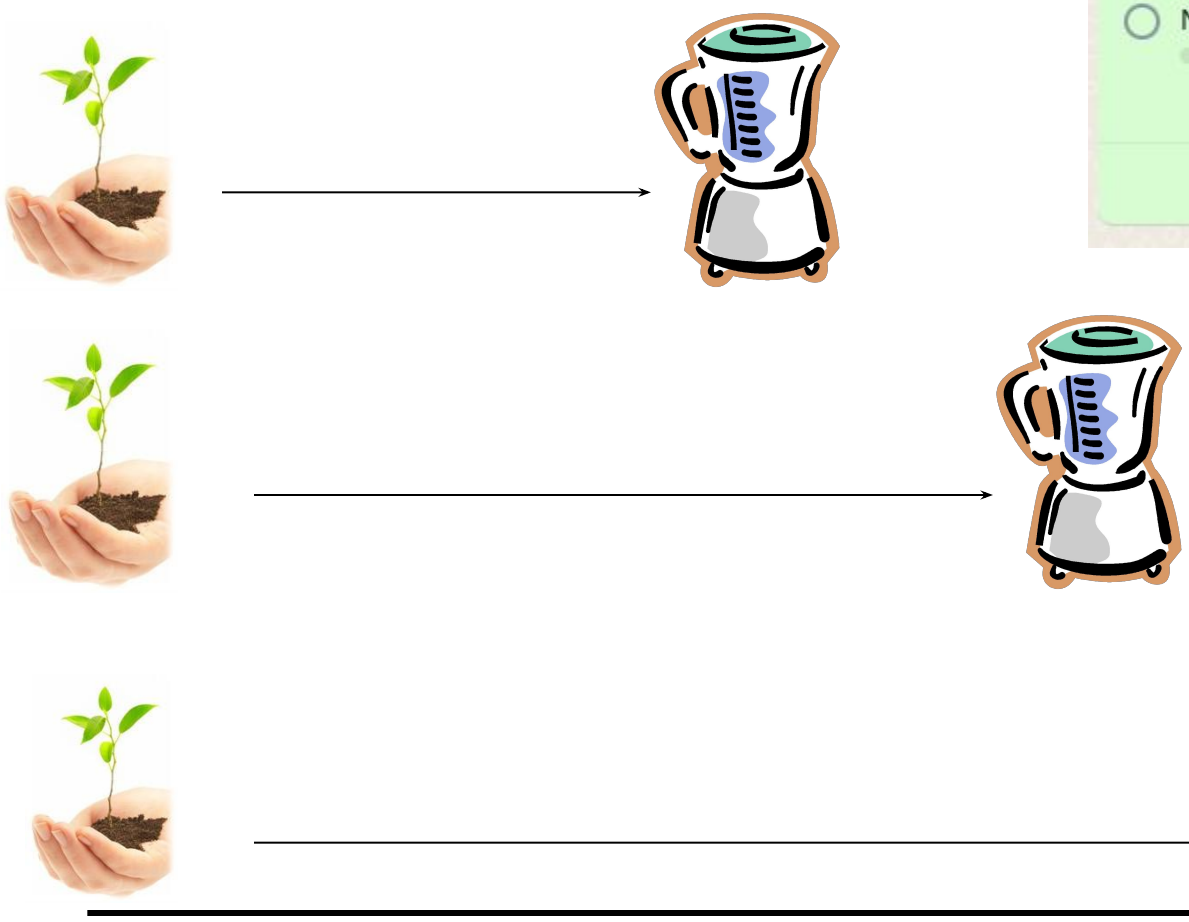
Time= birth teenager Hunting Working

N subjects for biological relevance, k samples per subject



*

Is this a longitudir



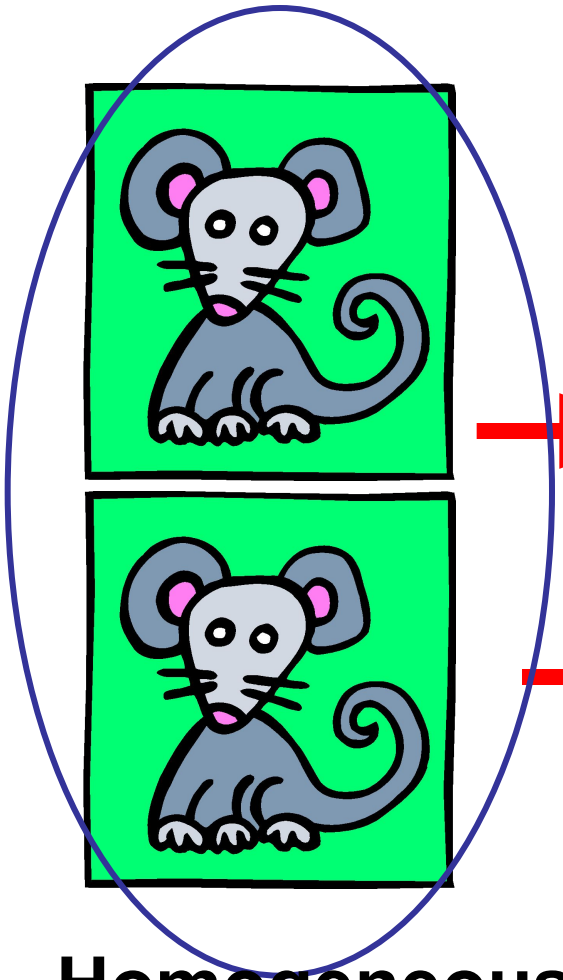
*

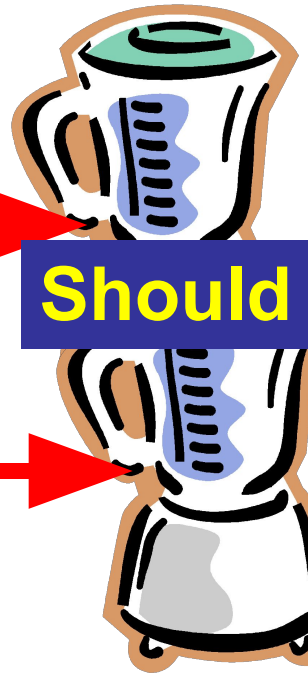
Build a longitudinal experiment

We want to evaluate exposure to some chemical over time.

*

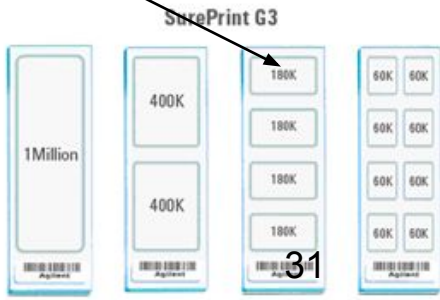
Pooling





Diminish biological variance

Should or should not ?

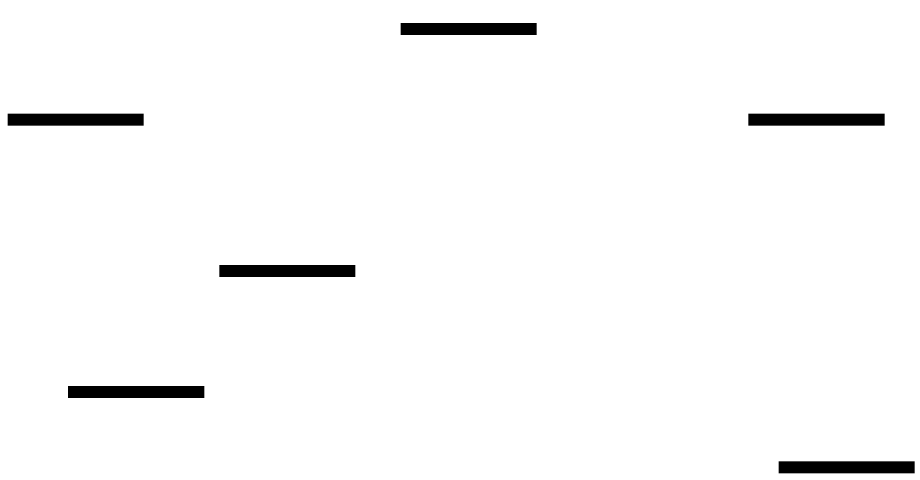


Homogeneous or Heterogeneous?*

Pooling

- Diminish biological variance, why?

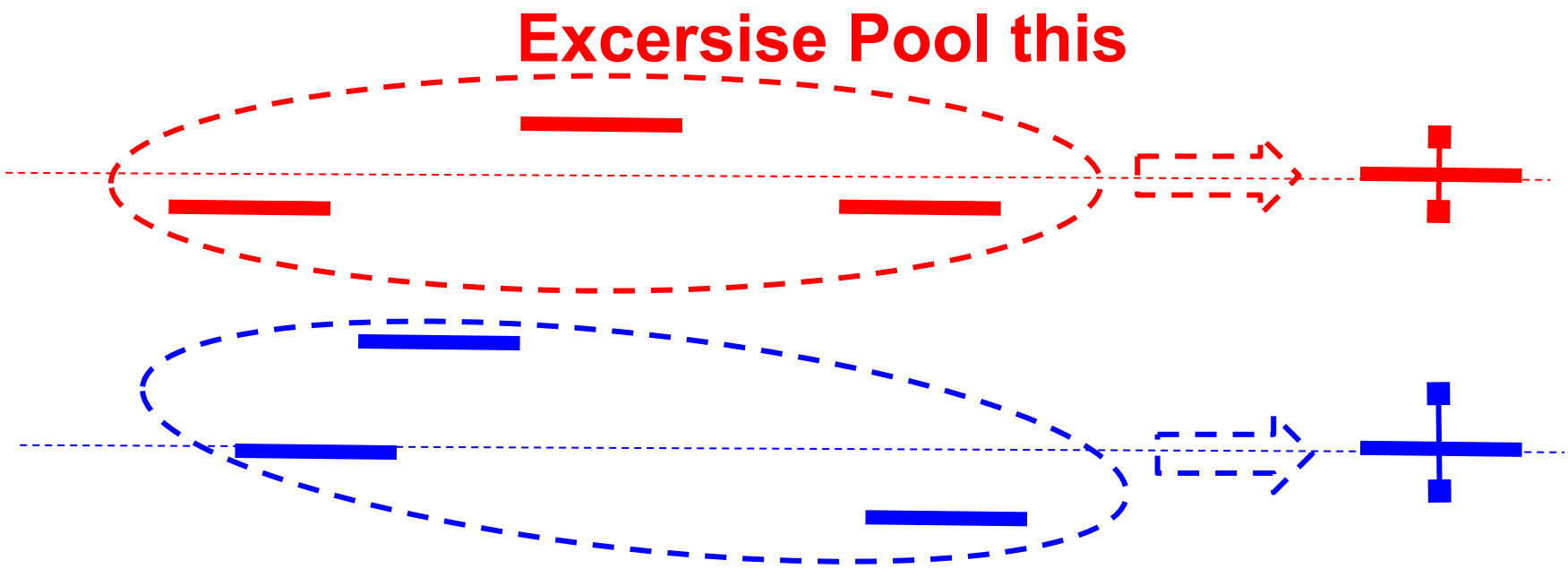
Excersise Pool this



*

Pooling

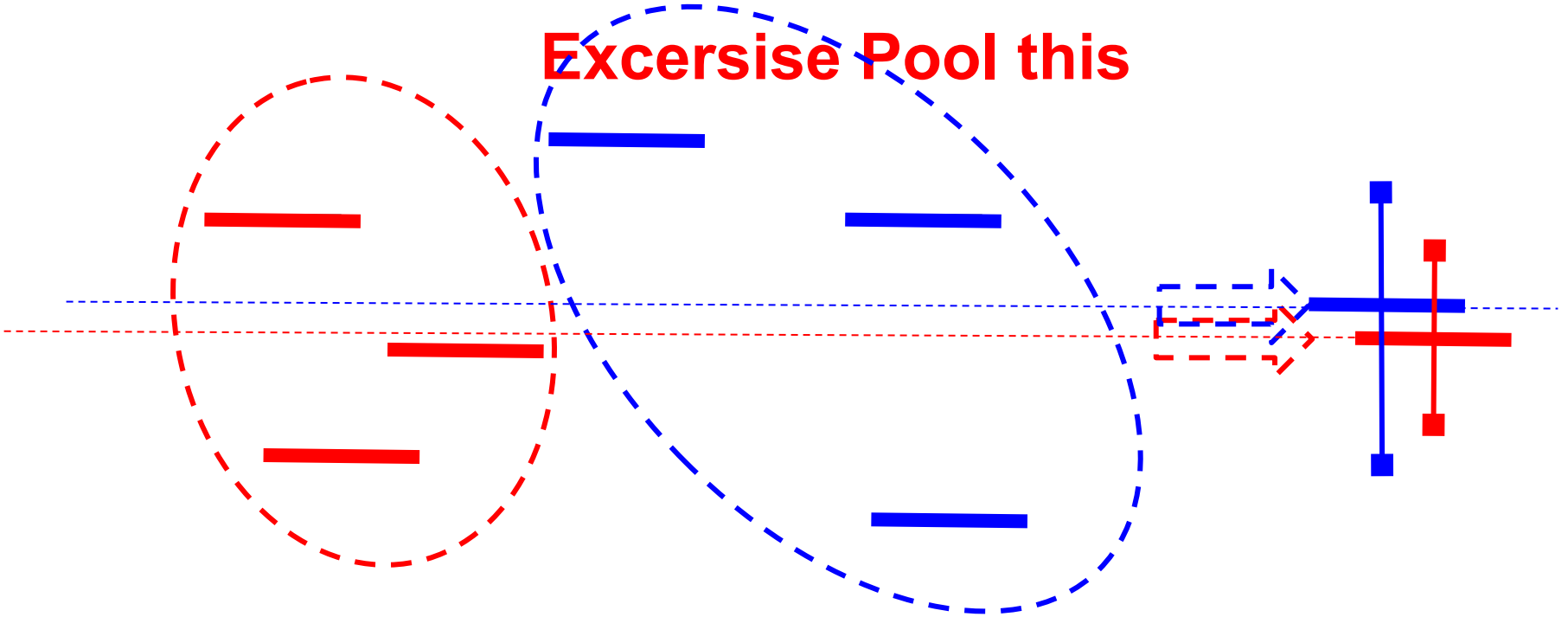
- Diminish biological variance, why?



*

Pooling

- Diminish biological variance, why?



*

Common pools

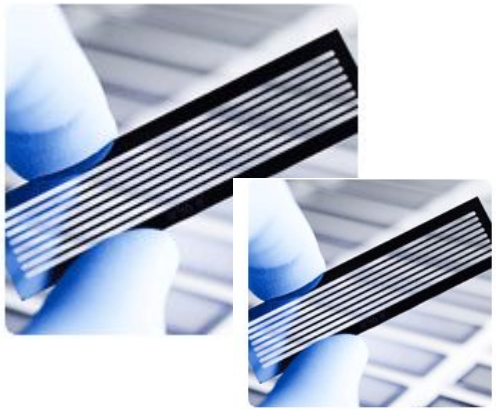
The tumor sample is a **pool** of cells, that's why single cell RNAseq emerges

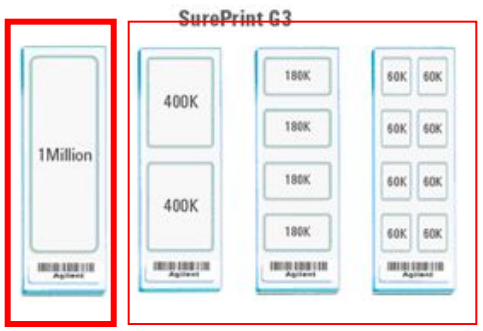
Splitting the samples

- Randomize: Randomization dictates that the experimental subjects should be randomly assigned to the treatments or conditions to be studied in order to "***eliminate"*** **unknown factors that potentially affect results**.
- Randomization should be considered all through the wet and dry lab.
  - Technician
  - Processing day or day time
  - Library preparation (one of the largest)
  - Flow cell / lane
  - Etc.

Splitting the samples

- One color/Lane (affy, agilent,NGS, Silver Gels):
  - Try to balance
  - Confounding factors
  - One combination per chip/Lane, there's not much to do



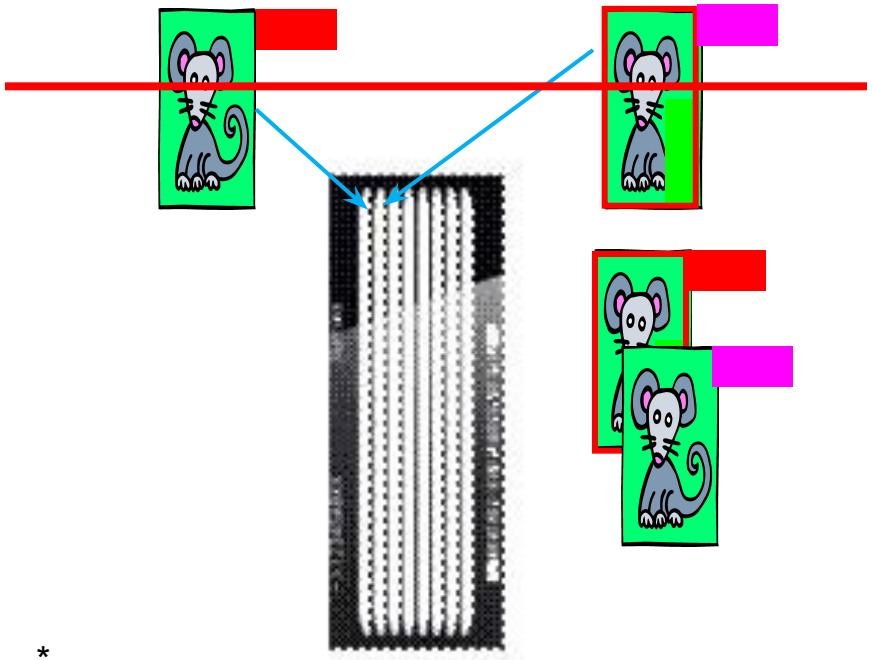


However ◄slide/Flow cell effects could happens

*

Splitting the Samples

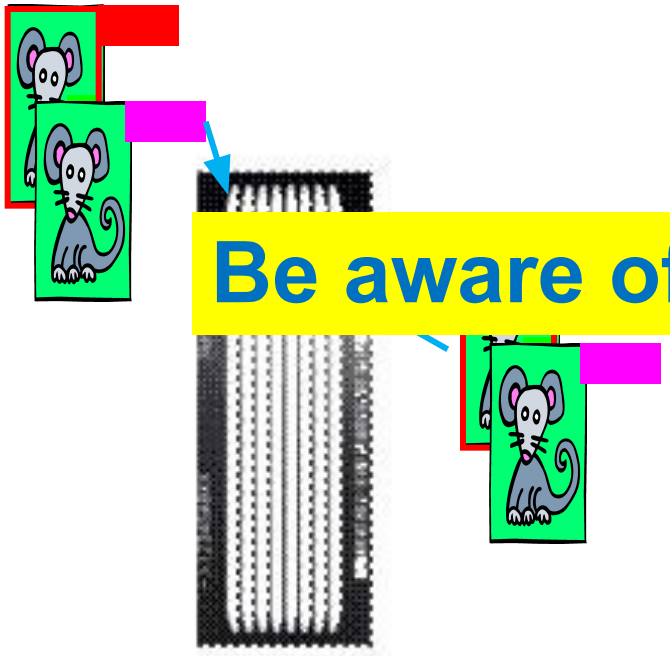
- Balanced Block Design (i. e for multiplexing or bar coding)
  - Objetive: Wild Mouse vs. “Untailed” mouse



*

Splitting the Samples

- Balanced Block Design (i. e for multiplexing or bar coding)
  - Objetive: Wild Mouse vs. “Untailed” mouse



Be aware of BATCH effects

Advantages:

- **If same effect of is**

can be estimated

The million dollar question: How many samples?

- What can be said with only one sample?
  - ▼#samples => ▲outlier chances.
  - P values?
    - Through some Fisher like test of proportions

The million dollar question: How many samples?

- *What can be said with only one sample?*
- To take into account
  - Population variances of genes ¿?
  - Expected difference ¿?
  - More the better (n>=5 the magic number)
  - Some says that Tophat/Cufflinks statistics work best with three or **<u>more</u> biological** replicates

But

But, i just can afford two samples

*

do not be miserly and

Keep samples for validation!!

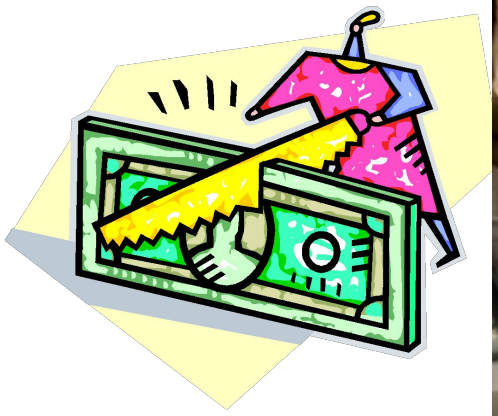
Dear and beloved Boss, Can I spend more

Can we send this work to this congress??

Other topics

- Sequencing depth (accuray?, budget?, low expressed genes is the target?)
- Everybody recommends a kind of pilot study, but who has the money for it??
  - Take a look to available data on free repositories
- Currently, sequencing technology is quite stable, so it's better to invest in biological replicates.

Think First



*

Take into account!!!

if you do what everybody does, you will get what everybody gets

The next step will be….

you will either run the experiment
or
search for data from repositories

Elmer Fernández

Independent Researcher at CONICET