Introduction to statistics, experimental design and hypothesis testing

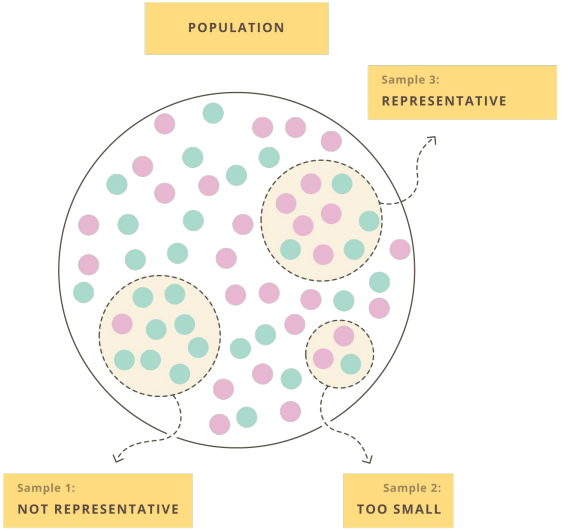
Session I-II

Michela Traglia and Reuben Thomas

Bioinformatics Core, GIDB
Gladstone Institutes

February 24th and 25th, 2025

Data collection is fundamental to make general claims



- Empirical data are noisy
- Resources are limited
- Generalize our scientific claims as much as possible

Experimental design refers to…
The organization of an experiment, to ensure that the **right type of data**, and **enough of it**, is available to answer the questions of interest as clearly and efficiently as possible.

Experimental design guides data collection

Maximum amount of relevant data for the research **at minimum resource spend**

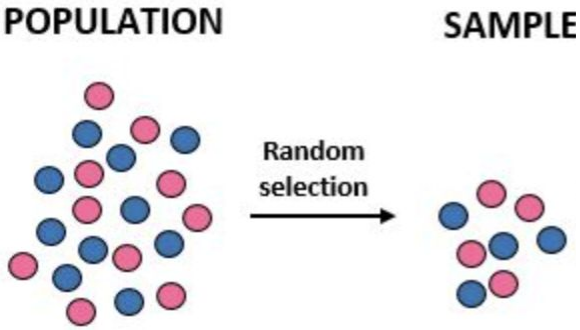
Minimize the number of experiments

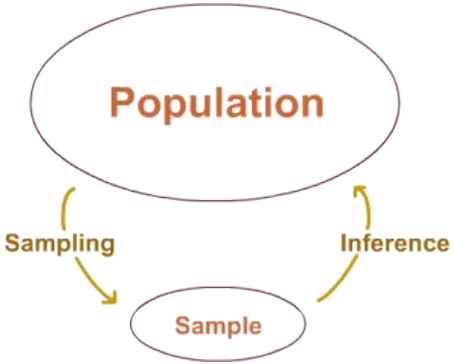
Source of variations evaluation, variables/factors that could affect the system.

Sampling the right amount of data

All subjects/units that we want base our claims/conclusions on

Data is expensive
Studying of the sample -> Conclusion on the population





The goal is to collect enough data to statistically test whether you can reasonably reject the null hypothesis in favor of the alternative hypothesis

Sample size - larger always better?

As always, it depends…

- on what we want to do (differential gene expression, variant detection, GWAS, …)
- on the variability between samples (cell lines, inbred animals, patients, …)
- on the magnitude of the expected effect

Tips

1) Set a sample size in advance and estimate the power of the experiment to detect differences of various sizes
2) Pilot experiments to guide sample size estimation

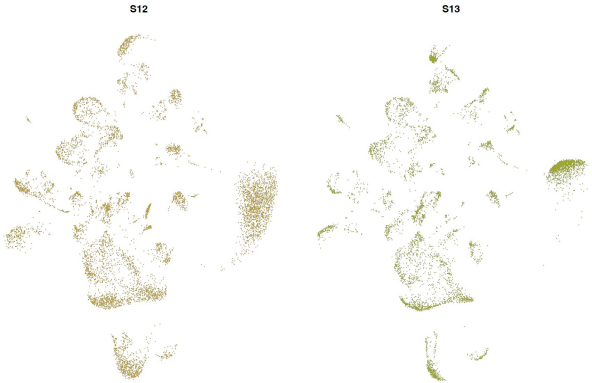
Biological variability across samples

scRNA-seq experiment – gene expression differences between WT and mutated mice

Population of animal models

Sampling

Population of animal from a given genotype



Within same genotype, two distinct animals -> biological variability

Biological vs technical replicates

- Biological replicates are biologically distinct samples (for ex. same condition) showing biological variation.
- # of replicates that give a high probability of detecting an effect of practical importance.
- At minimum you should have three biological replicates for treatment.

- Technical replicates are same sample repeated measurements (reproducibility?)
- Technical replication is essential when part of the experimental objective is to measure the measurement error.

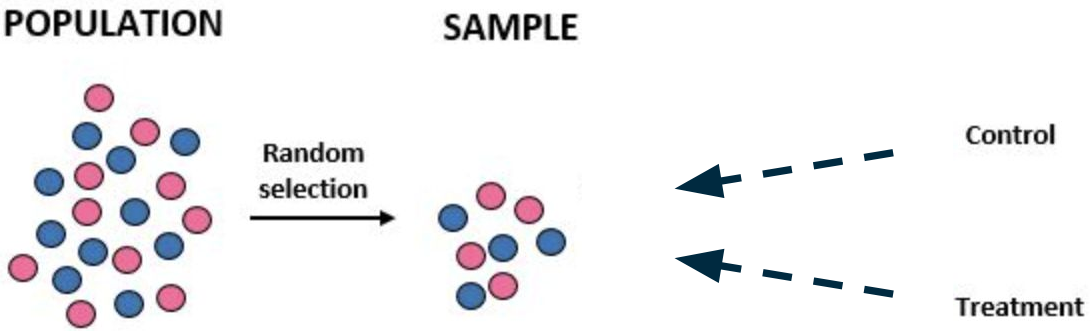
Knowing how to deal with the challenges of experimental design is central to achieving reproducible experiments.

- ✔ Identify the response and variables of interest
- ✔ Identify target population that you want to base your claims on
- ✔ Identify factors that affect the response of interest
- ✔ Choose samples from target population

When well-designed, experiments minimize any bias to make stronger inferences about the biological differences we see in the experiment.

Assign treatment to groups

Experiment: compare control and treated groups



Scenario I

Analysis batch I / Study center I / Processing protocol I ...

Analysis batch II / Study center II / Processing protocol II ...

Bad assignment!

Scenario II

Treatment I

Treatment II

Bad assignment!

Ronald Fisher

"Analysed data from Classical Field Experiments"

Overcome the large amount of variation in agricultural and biological experiments that often confused the results

This motivated him to find experimental techniques that could

- eliminate as much of the natural variation as possible
- prevent unremoved variation from confusing or biasing the effects being tested
- detect cause and effect with the minimal amount of experimental effort necessary - time consuming and costly

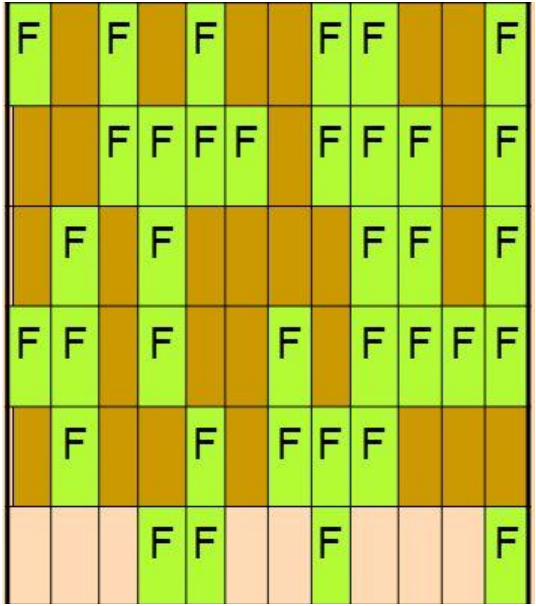
Statistical Methods for Research Workers in 1925 and *The Design of Experiments* in 1935

Well-designed experiments are characterized by three features: randomization, replication, and local control

Fisher -> helps to avoid biases due to changes in background or confounding variables.

One of the main purposes for experimental designs is to minimize the effect of experimental error.

Randomization, replication, and blocking, are methods of error control.



Year 1

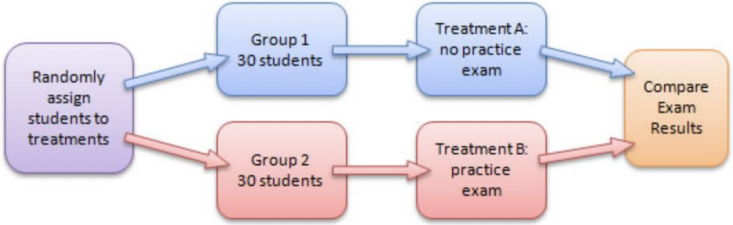
Repeat - Multiple years

Randomization

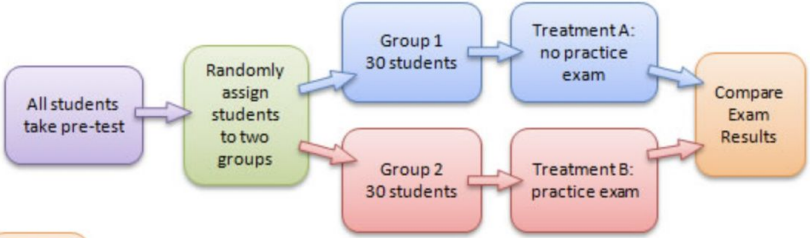
- Randomly assign subjects to treatment and control groups in order to minimize bias and moderate experimental error.
- Assign random numbers so that any experimental unit (EU) has equal chances of being assigned to treatment or control
- Can create unequal numbers between treatment and control groups.
- Appropriate only for experiments with homogeneous experimental units (e.g., mice should be of same sex, strain, age, etc.) where environmental effects, such as light or temperature, are relatively easy to control.

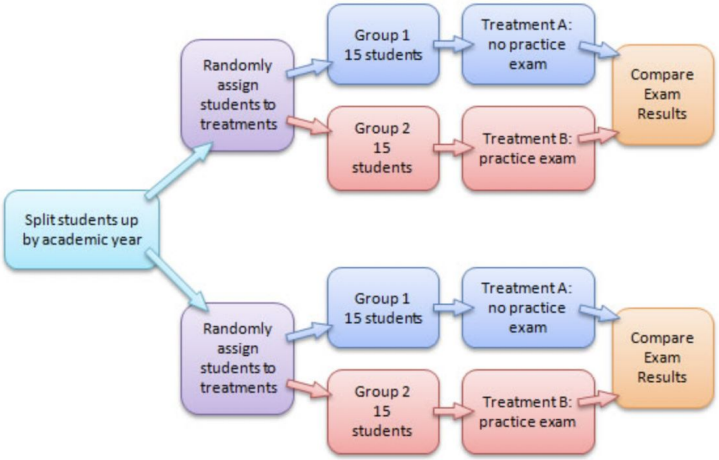
Randomization and blocking

Completely Randomized Design



Matched-Pairs Design

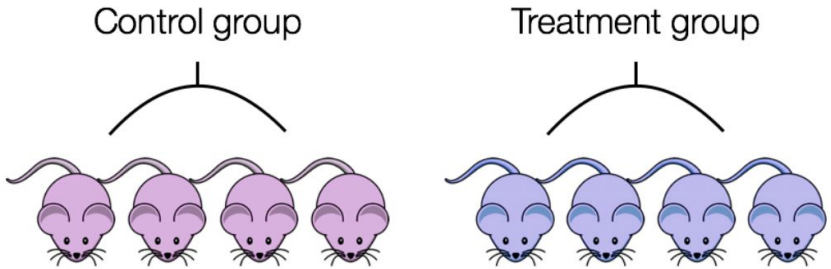




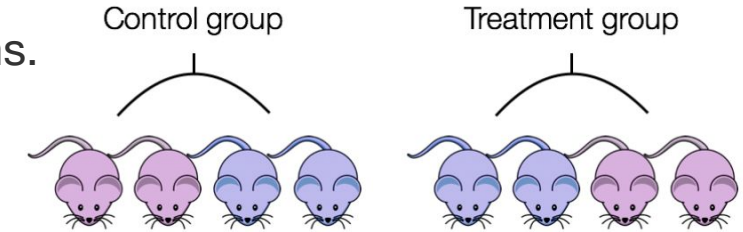
Randomized **Block** Design - stratification

Blocking

If all *control* mice were female and all of the *treatment* mice were male, then the treatment effect would be confounded by sex.

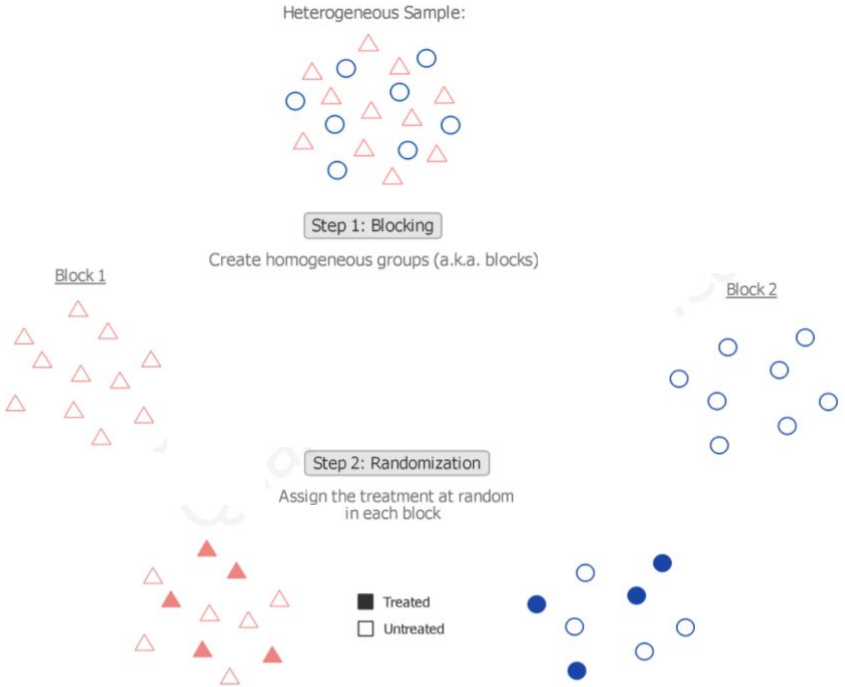


Ensure animals in each condition are all the **same age, sex, litter and batch**, if possible.
If not, split animals equally between conditions.



Blocking approach helps to reduce variability unexplained in the model

Randomized block design



Treatment	
Placebo	Vaccine
500	500

Gender	Treatment	
	Placebo	Vaccine
Male	250	250
Female	250	250

https://quantifyinghealth.com/randomized-block-design/
https://stattrek.com/experiments/experimental-design.aspx

Capture effects of interest and avoid unwanted variation in experiment

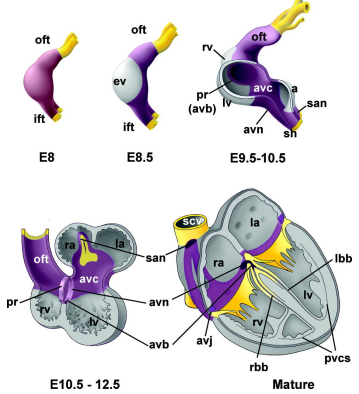
- ✔ Identify the response and variables of interest
- ✔ Identify target population that you want to base your claims on
- ✔ Identify factors that affect the response of interest
- ✔ Choose samples from target population

Randomly assign samples across different levels of factors affecting response

Block out variation that is not of interest by randomly assigning to levels of factors within a block

"Block what you can control; randomize what you cannot control"

Which one is the best design?



	Design 1 – Sample prep date	Design 2 – Sample prep date	Design 3 – Gender	Design 4 - Gender
Sample_1_E9.5	Jan 9th, 2019	Jan 11th, 2019	Male	Male
Sample_2_E9.5	Jan 9th, 2019	Jan 9th, 2019	Male	Female
Sample_3_E9.5	Jan 9th, 2019	Jan 11th, 2019	Male	Male
Sample_4_E9.5	Jan 9th, 2019	Jan 9th, 2019	Male	Female
Sample_1_E11.5	Jan 11th, 2019	Jan 11th, 2019	Female	Male
Sample_2_E11.5	Jan 11th, 2019	Jan 9th, 2019	Female	Female
Sample_3_E11.5	Jan 11th, 2019	Jan 11th, 2019	Female	Male
Sample_4_E11.5	Jan 11th, 2019	Jan 9th, 2019	Female	Female

Between-subjects vs within subjects design

- In a between-subjects design (or between-groups design)

  - every experimental unit experiences only one condition,
  - group differences between participants in various conditions

Also called **independent measures or independent-groups design** because compare unrelated measurements taken from separate groups.

Gender	Treatment	
	Placebo	Vaccine
Male	250	250
Female	250	250

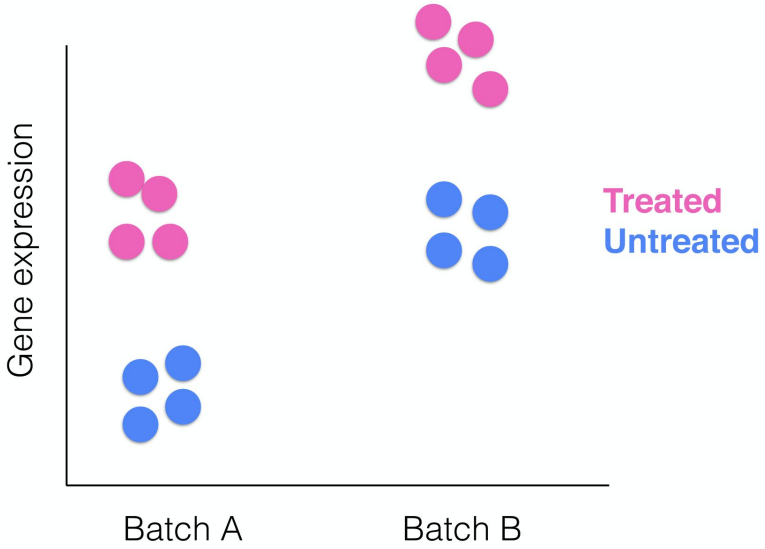
Between-subjects vs within subjects design

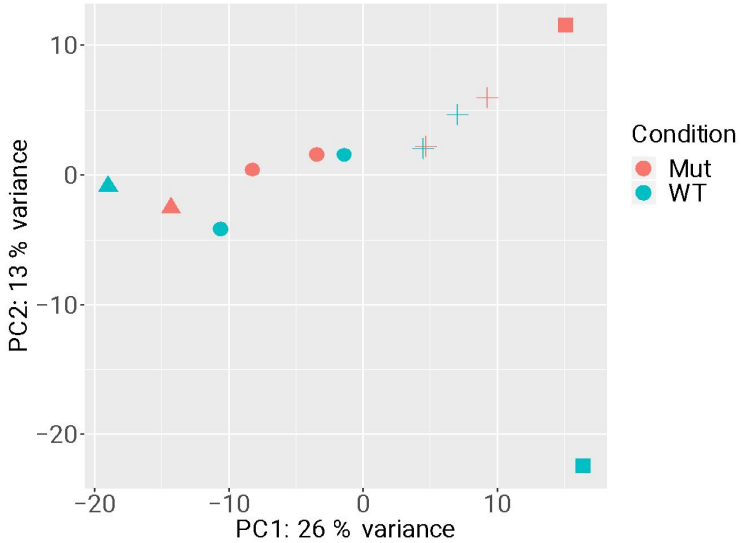
- In a within-subjects design (or repeated measurement design)

  - every experimental unit experiences all the conditions,
  - test the same individuals repeatedly to assess differences between conditions

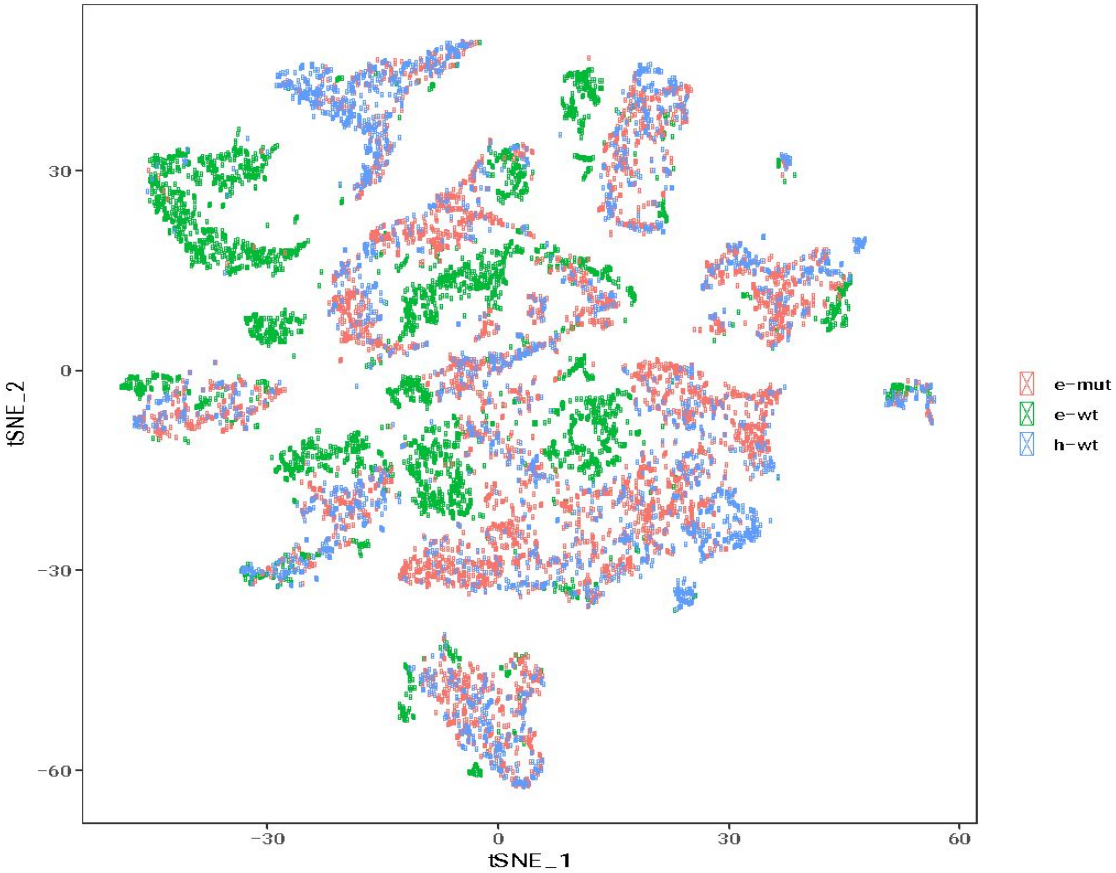
Also called a **dependent groups or repeated measures design** because compare related measures from the same individuals between different conditions.

Is the variation between groups driven by the biology?

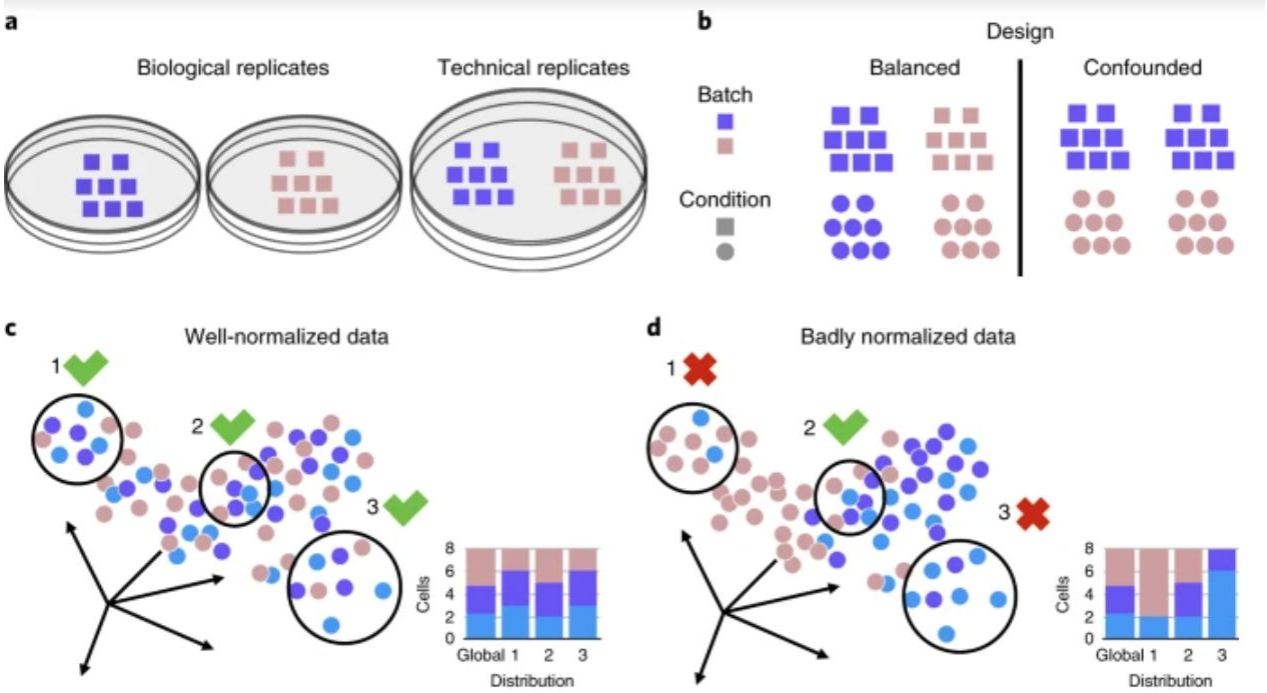




Confounding in scRNA-seq data is a problem

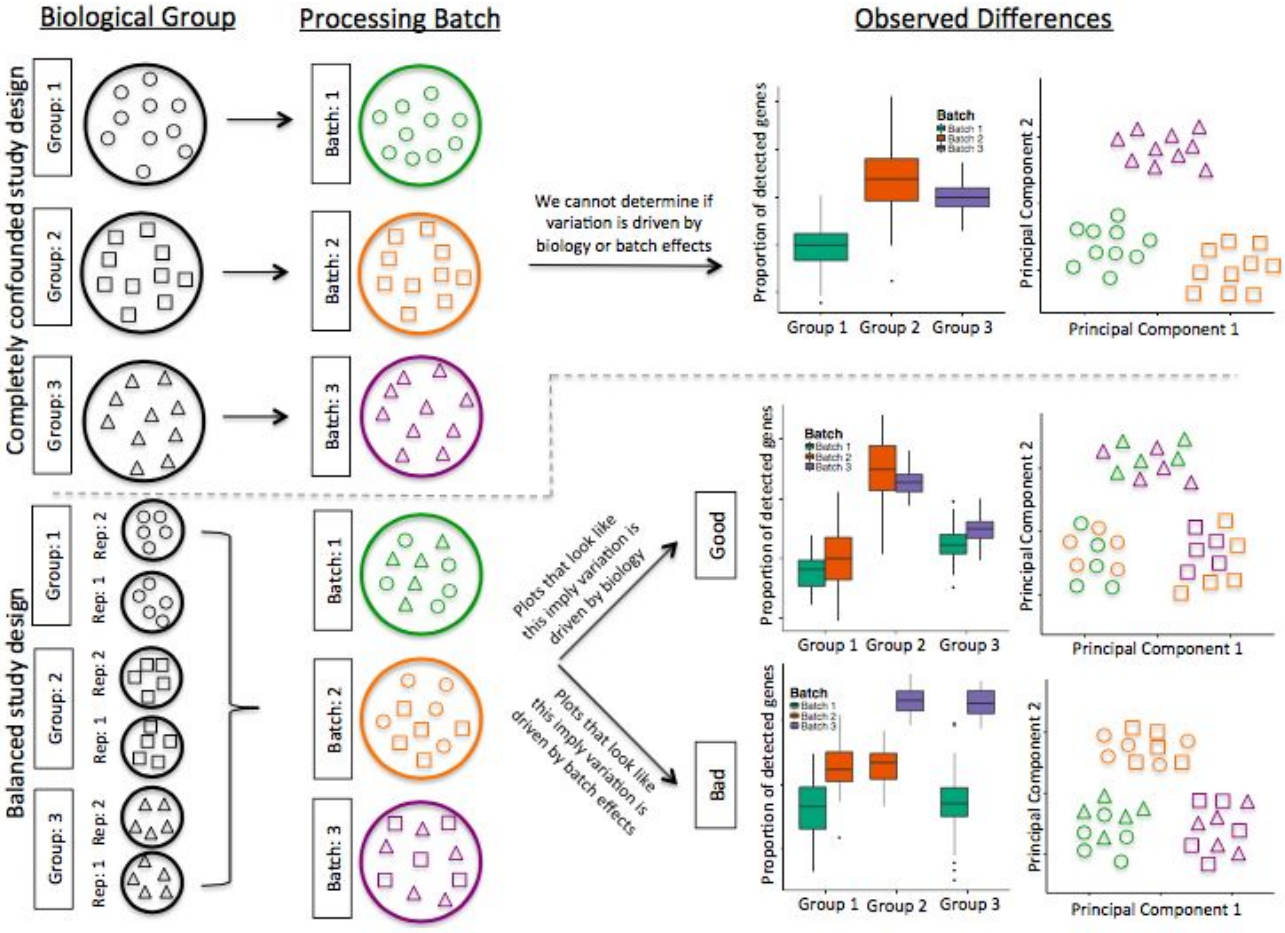


Well and badly analyzed scRNA-seq data



https://www.nature.com/articles/s41592-018-0254-1

Confounding biological variation and batch effects



Hot to know whether you have batches

- Were all RNA isolations performed on the same day?
- Were all library preparations performed on the same day?
- Did the same person perform the RNA isolation/library preparation for all samples?
- Did you use the same reagents for all samples?
- Did you perform the RNA isolation/library preparation in the same location?

If *any* of the answers is **‘No’**, then you have batches.

Isolate batch effects for RNA-seq

If unable to avoid batches:

- Split replicates of the different sample groups across batches.
- The more replicates the better (definitely more than 2).
- Include batch information in your experimental metadata.
- During the analysis regress out the variation due to batch if not confounded so it doesn’t affect the results.

Take – home messages

- Plan ahead
- Prevent bias from uncontrollable
- Randomization and balancing
- Write it down in an Experimental plan
- Follow the experimental plan
- Be careful with interpretation of results!