Introduction to statistics, experimental design and hypothesis testing

Michela Traglia and Reuben Thomas

Bioinformatics Core, GIDB
Gladstone Institutes

January 17, 2023

Introductions

Michela Traglia

Statistician III

Reuben Thomas

Associate Core Director

Motivation of the workshop

- Accessible statistical tools allow researcher to easily perform analyses
- How the program work and which setting to use?
- How to interpret the results?
- Hard to get defensible conclusions
- Understand and critically evaluate scientific publications
- Before the statistical analysis, fundamental is to plan the experimental design

Goals of this workshop

- Introduce basic concepts underpinning experimental design
- Learn how to think critically about the data you want to generate and use to make claims about
- Overview of statistical tests and concepts

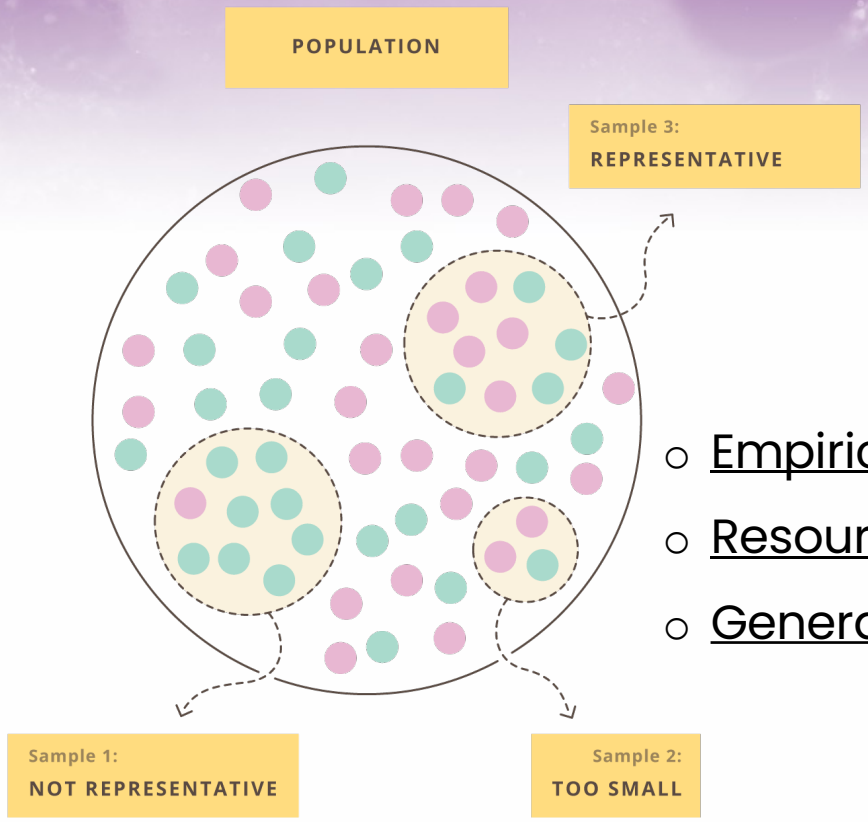
Basic level course - No prerequisites

Please feel free to interrupt with questions, speak up or use the chat!

Workshop sessions

- Day 1 – Experimental design (Michela Traglia)
- Day 2 – Hypothesis testing (Reuben Thomas)
- Day 3 – Overview of the statistical analysis (Reuben Thomas)

When we perform and experiment…



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

Statistics is defined as
the science of collecting, analyzing, and drawing conclusions from data

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Medical research has a credibility problem

- Estimated that ~75% published research findings cannot be reproduced
- ~$28 billion per year (nearly half of the annual non-clinical research budget in the US) is wasted on attempts to reproduce published studies
- Only a small percentage are due to overt fraud (intentional fabrication)
- Most are what are considered “detrimental research practices”
- Patient lives placed at risk

Thanks: Kevin Mullane
Director, Corporate Liaison & Ventures
Corporate Ventures and Translation
Gladstone Institutes

https://gladstone.org/events?series=responsible-conduct-of-research
https://rcr.ucsf.edu/

What we mean with ‘experimental design’

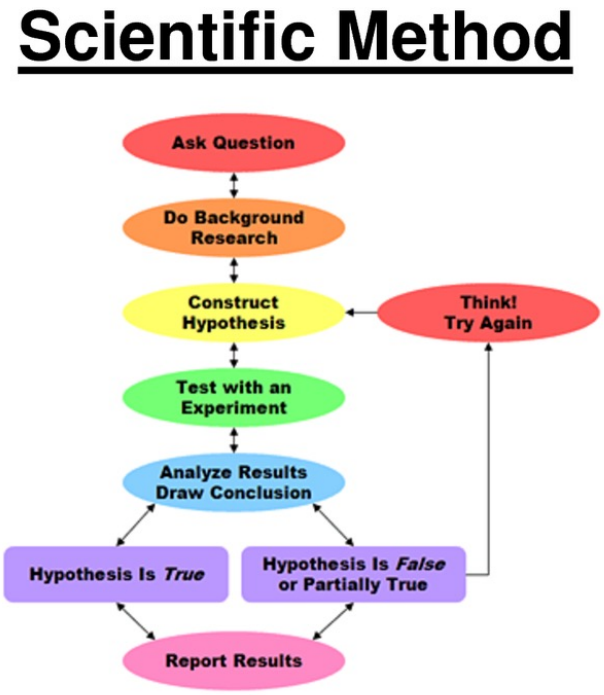
The organization of an experiment, to ensure that the right type of data, and enough of it, is available to answer the questions of interest as clearly and efficiently as possible.

Because the validity of an experiment is directly affected by its construction and execution, attention to experimental design is extremely important.

Statistical experimental designs

The scientific method consists of iterative application of the following steps:

(1) observing of the state of nature
(2) hypothesizing the mechanism for what has been observed
(3) collecting data
(4) analyzing the data to confirm or reject the hypothesis



Statistical experimental designs provide a plan for collecting data in a way that they can be analyzed statistically to corroborate the conjecture in question.

Outline of experimental design principles

1. Define the research questions
2. Understanding the system you want to study
3. Formulate your null and alternative hypothesis
4. Independent and dependent variables of interest
5. Target population, sampling, replicates
6. Confounding variables

Break 5 minutes

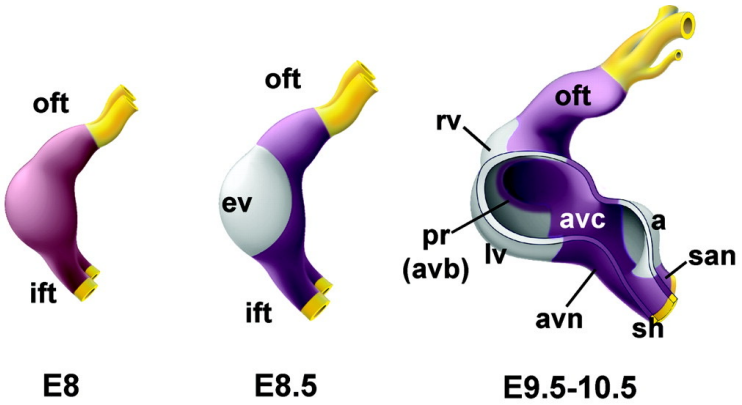
7. Assign treatment to groups
8. Batch effects

1. Define the research question

A specific issue, contradiction between two or more perspectives, or a gap in knowledge that you will aim to address in your research.

What is a scientific question that you are currently working on?

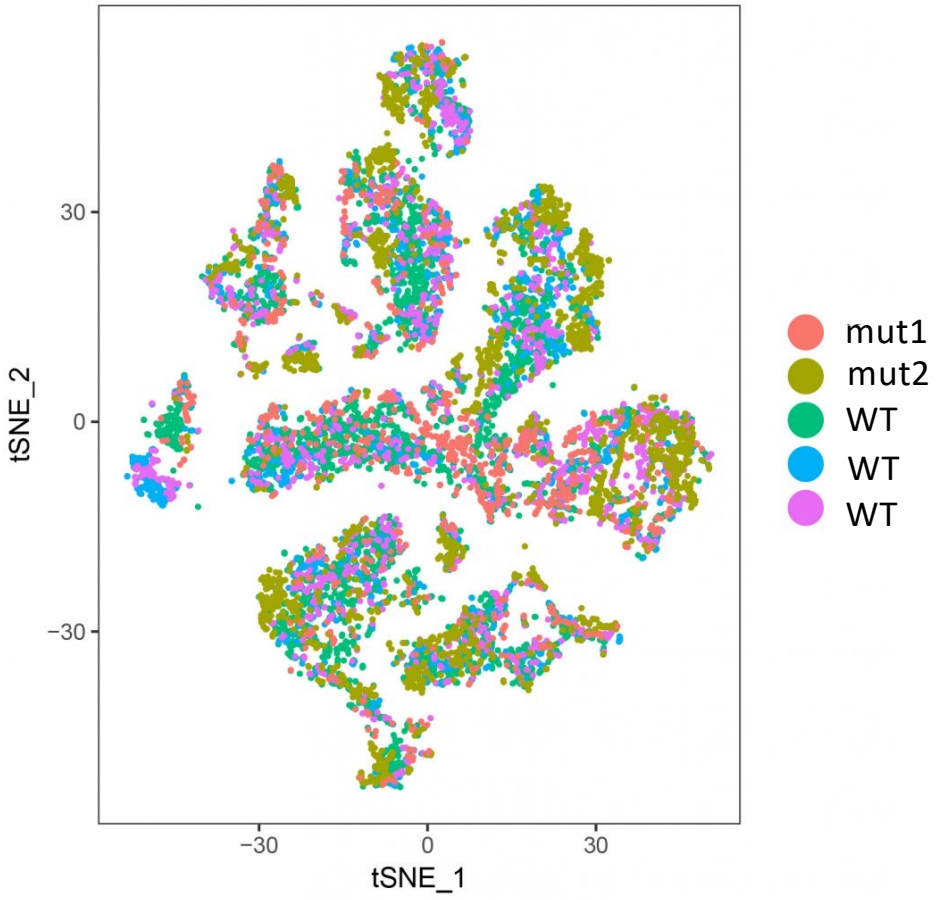
Research question 1



Gene controlling developing heart

https://www.ahajournals.org/doi/epub/10.1161/CIRCEP.108.829341

Research question 2



scRNA-seq on cells from WT and mutant mice

Research question 3

Chemical pollutants during pregnancy contributing to autism in children

Lyall *et al,* Environ Health Perspect 2017

Ask a specific question

Do the pregnant women with more chemical pollutants in the blood have more autistic children?

Lyall *et al,* Environ Health Perspect 2017

2. Understand the system you are going to study

Observational studies	Experimental studies
Researchers don't assign exposure	Researchers manipulate factors
Observing what is already happening	Create a treatment and compare the response
No establishing cause-effect	Changes cause an effect
Ex. Case-controls, cohort	Ex.Clinical trials

3. Formulate your hypothesis

1. The amount of chemical pollutants in the blood of pregnant women negatively affect the fetal brain development and increase the risk of autism in children.

2. The mother exposed to more pollutants will have more autistic children than mother less exposed to pollutants.

and then write a null and alternative hypothesis

- H_0: The pollutants exposure in the blood during pregnancy has no effect on the fetal brain development.

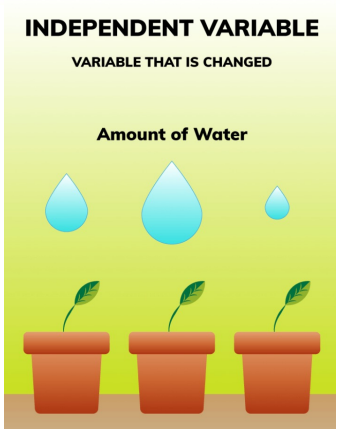
- H_1: The pollutants exposure in the blood during pregnancy has a negative effect on the fetal brain development.

4. Consider the variables of interest

Independent variable (IV)

Also called:

- Exposure variable
- Control variable
- Explanatory variable
- Manipulated variable

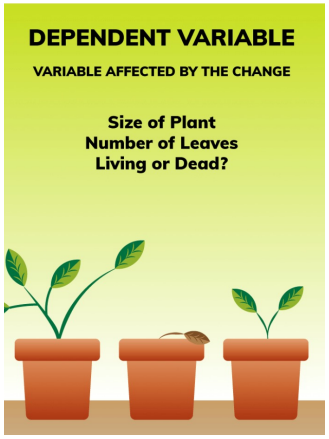


Pollutants measurement

Dependent variable (DV)

Also called:

- Outcome variable
- Controlled variable
- Explained variable
- Response variable



Autism in children

Image source: by Hannah Bonville

Experimental design is used to establish the effect an independent variable has on a dependent variable.

Independent variable (IV)	Dependent variable (DV)
Risk factors	Disease
Genotype	Gene expression
X	Y

In an experiment, you manipulate an independent variable to study its effects on a dependent variable (response)

5. From the population to a sample

Biological and experimental units

Entities to which we apply treatment and on which we make observations and inferences

- Animal or human subject
- Raw material for some processing operation
- Condition that exist at a point in time or trial?

Experimental units are the smallest entities that can be independently assigned to a treatment (e.g., animal, litter, cage, well).

Experimental design guides in

Collecting the maximum volume of relevant and required data for the subject of the research, at minimum resource spend.

Evaluating the source of variations, variables/factors that affect the system.

It is an efficient method to minimize the number of experiments and gather the maximum amount of appropriate data.

Target population - generalization

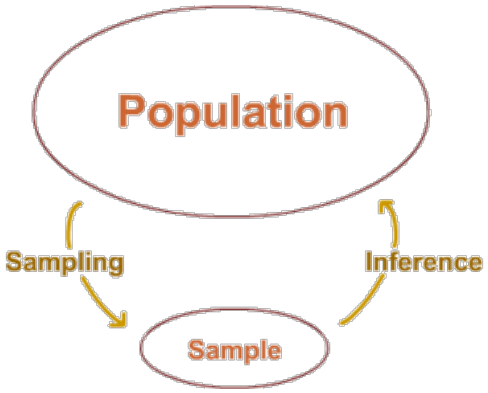
All subjects/units that we want base our claims/conclusions on

The cardiac tissue of all mice at embryonic stage E9.5

All children below 5 years old who are diagnosed with autism

All mother that were pregnant in a geographical area

Data is expensive
Studying of the sample ➔ Conclusion on the population



Sample size

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

Is a larger sample sizes always better?

Sample size -> amount of information -> precision (margin of error) / level of confidence in our sample estimates.

- High variability -> Greater uncertainty
- Larger sample size -> more information -> less uncertainty

+ Greater precision and power
- Cost more time and money

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

Replicates

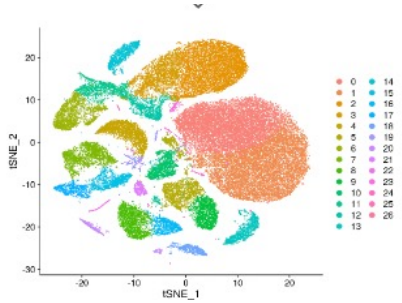
scRNA-seq experiment – gene expression differences between WT and mutated mice

H0: ?
H1: ?

Population of animal models

Sampling

Population of animal from a given genotype



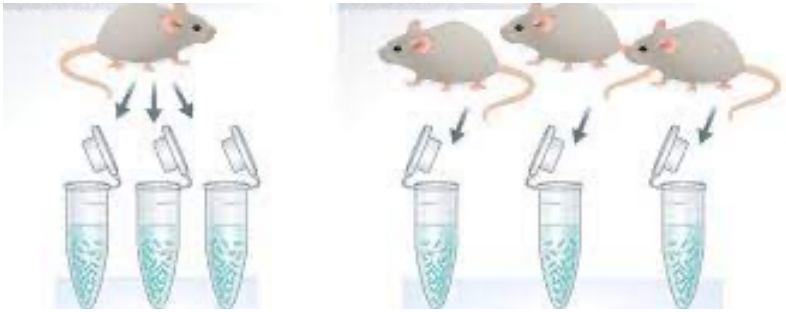
Replicates – variance estimation

- How large differences are you looking for?
- What is the expected expression difference of targeted biology in these samples?
- Will ”no change” be a desired significant result?

Biological vs technical replicates

- Number of replicate runs that will give a high probability of detecting an effect of practical importance
- Use biological replicates to answer biological questions, and technical replicates to answer technical questions

Poll

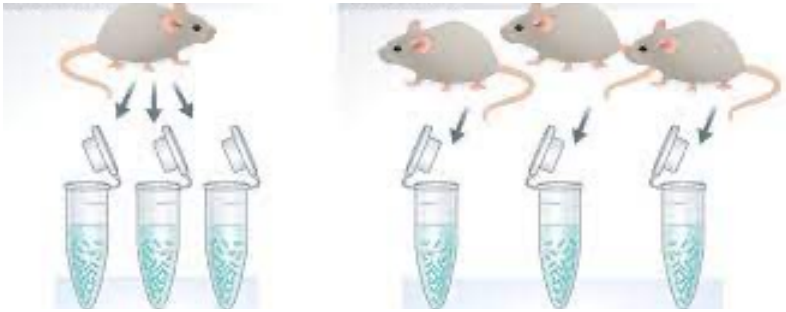


A B

Cell cultures have been originated from one mouse (Exp A) and from 3 mice (Exp B). How many biological (n) and technical replicates in experiment A?

1) n=1 biological and 0 technical replicates
2) n=1 biological replicate and 3 technical replicates
3) n=3 biological replicates and 3 technical replicates

Poll

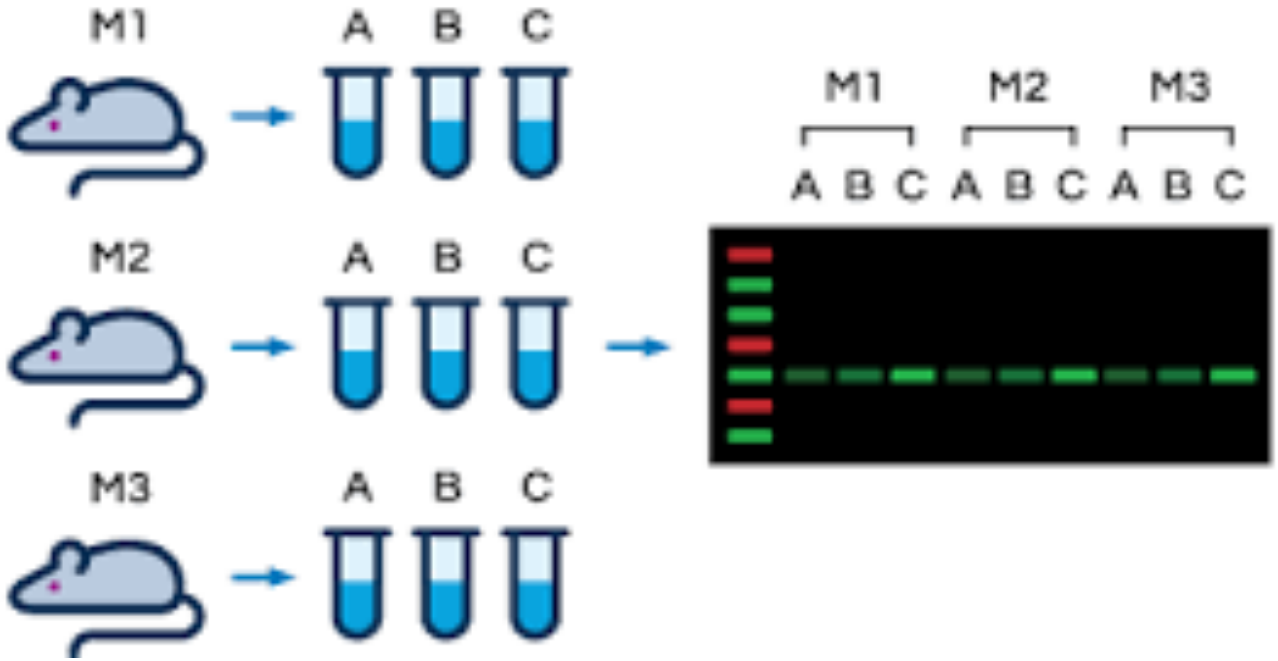


A B

Cell cultures have been originated from one mouse (Exp A) and from 3 mice (Exp B). What do you have in experiment B?

1) 3 technical replicates
2) 3 biological replicates

Biological vs technical replicates



How many n?

Do we need to perform a statistical power calculation?

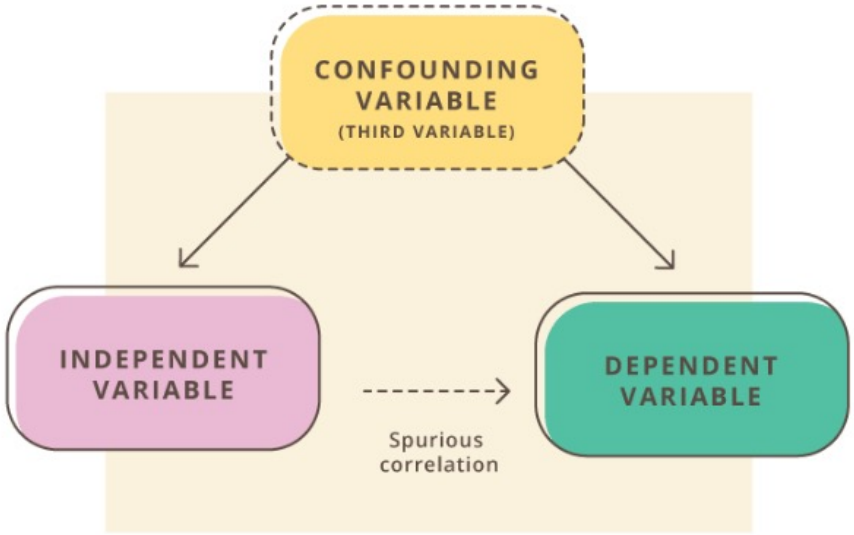
Workshop Day 2 - Reuben Thomas

6. Confounding variables

What affects the outcome of an experiment?

$$\text{Outcome} = \underbrace{\text{Treatment effects}}_{\substack{\text{Environment}\\ \text{Compound}\\ \text{Inhibitor}\\ \text{siRNA}\\ \text{Dose}\\ \text{Time}}} + \underbrace{\text{Biological effects}}_{\substack{\text{Sex}\\ \text{Age}\\ \text{Weight}\\ \text{Litter}\\ \text{Genotype}\\ \text{Species}\\ \text{Cell line}}} + \underbrace{\text{Technical effects}}_{\substack{\text{Technician}\\ \text{Batch}\\ \text{Plate}\\ \text{Cage}\\ \text{Array}\\ \text{Day}\\ \text{Order}\\ \text{Source}}} + \underbrace{\text{Error}}_{\substack{\text{Experimental}\\ \text{Treatment}\\ \text{Sampling}\\ \text{Measurement}}}$$

A third variable



What at first looks like a causal relationship between IV and DV is ultimately spurious. The confounding variable is the hidden explanation.

A variable that is not included in an experiment, yet affects the relationship between the two variables in an experiment.

Be aware of confounding factors

In observational studies (case-control, cohort studies):

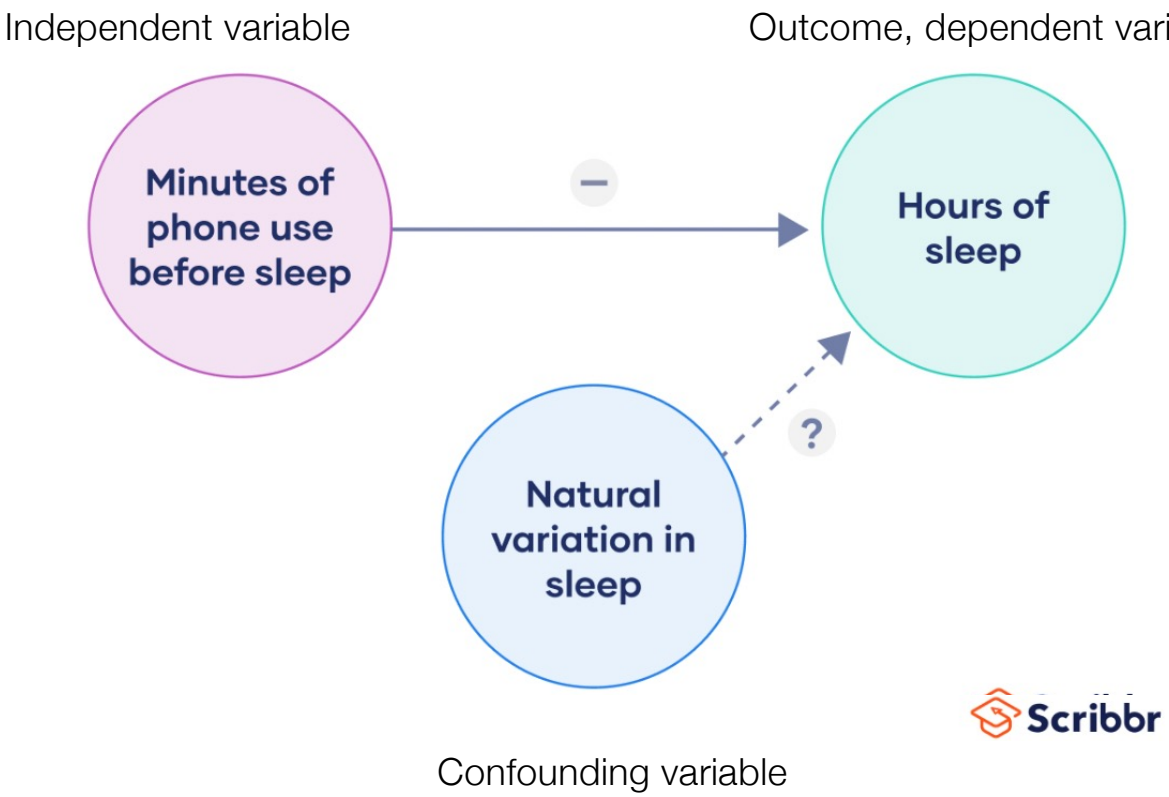
- confounding variables could cause unusual interpretations of data and the relationships between variables

In experimental studies:

- design the experiment to eliminate (as much as possible) the risk of confounding variables

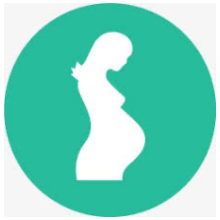
Which potential variables could be affecting the relationship between the variables in your study?

An example...

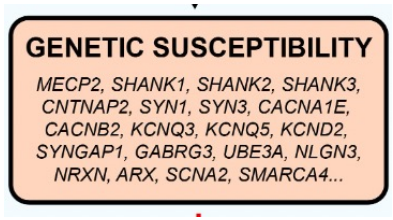

Independent variable
Outcome, dependent variable
Minutes of phone use before sleep
−
Hours of sleep
?
Natural variation in sleep
Scribbr
Confounding variable

An example...

Independent variable



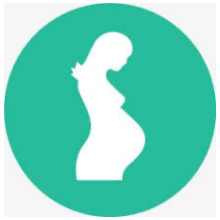
Outcome, dependent variable



Confounding variable

An example...

Independent variable



Outcome, dependent variable

Confounding variable:
- Ethnicity
- Country of birth
- Genetics

Chemical in pregnant maternal blood correlating with healthy children

Mother with high levels of chemicals blood

->

Correlation

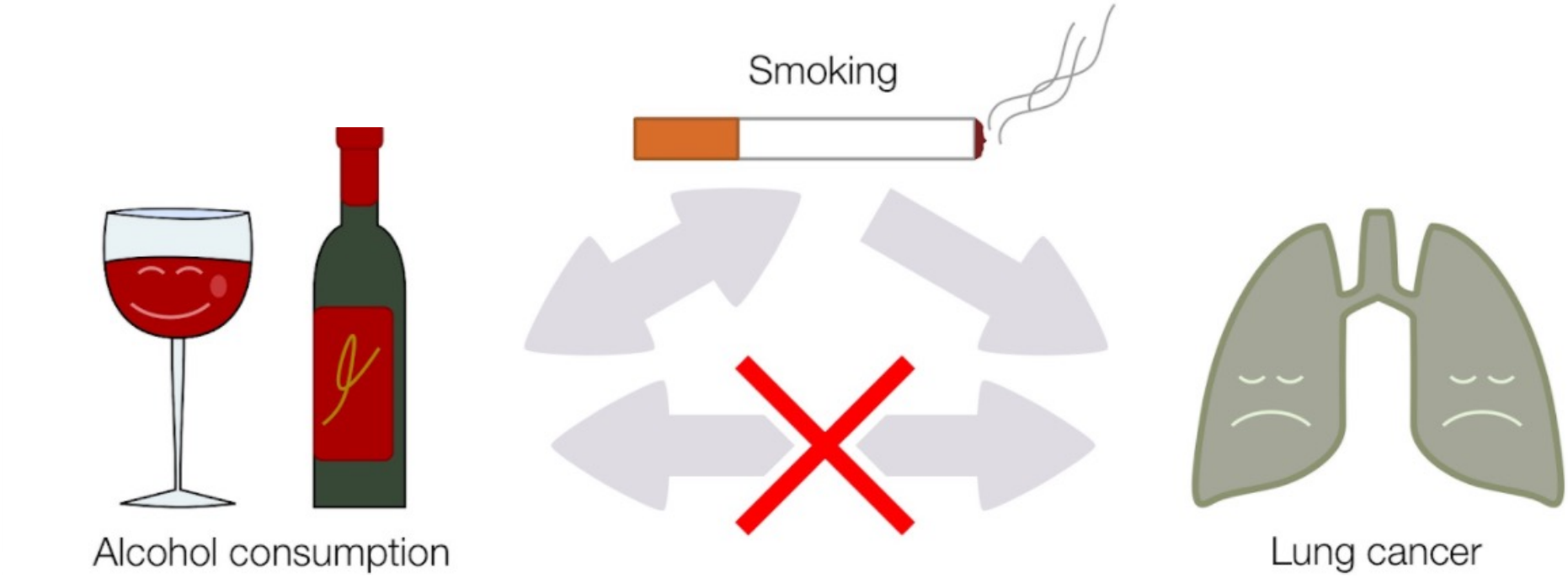
Healthy kids

High chemicals levels – not developing autism
Low chemicals levels – children developing autism

?

Genetics make-up controlling metabolism of chemical

Correlation vs Causation



Poll

Research question: *Does light exposure improve learning ability in mice?*

What can be the source of variability, confounding the outcome?
Select all that apply

1. Mouse inbred strains
2. Genetic background
3. Learning environment
4. All are ‘independent variables’

Factors potentially affecting the response

Biological factors that could affect the response

- -BMI
- -ethnicity
- -gender

Non-biological or technical factors that could affect the response.

- -time/day/month of experiment/batch
- -reagents used, reagents batch used
- -technician

Considerations about confounding variables

1. Must be correlated with the independent variable.

2. Must have a causal relationship with the dependent variable.

3. Can make it seem that cause-and-effect relationships exist when they don’t.

4. Can mask the true cause-and-effect relationship between variables.

When confounding variables are present, we can’t always say with complete confidence that the changes we observe in the dependent variable are a direct result of changes in the independent variable.

Break ~ 5 minutes

Please take the survey:
https://www.surveymonkey.com/r/DY7K5ZY

Outline of experimental design principles

1. Define the research questions
2. Understanding the system you want to study
3. Formulate your null and alternative hypothesis
4. Independent and dependent variables of interest
5. Target population, sampling, replicates
6. Confounding variables
7. Assign treatment to groups
8. Batch effects

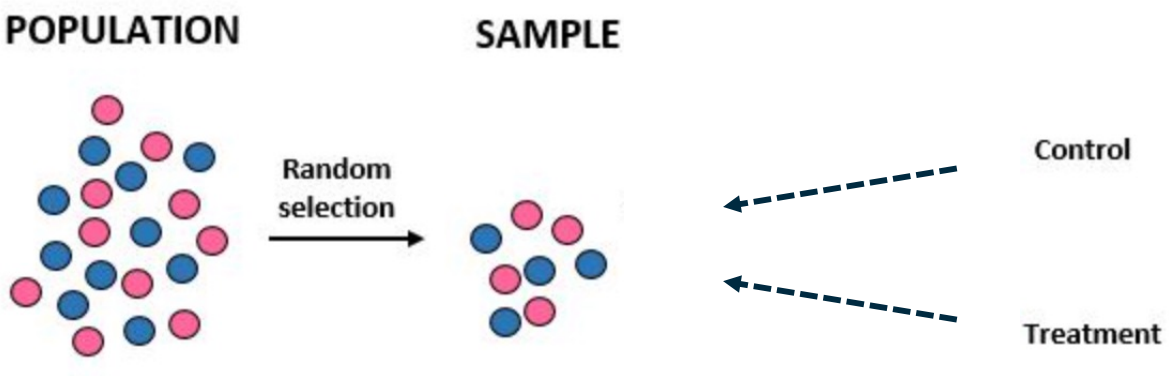
Knowing how to deal with the challenges of experimental design is central to achieving reproducible experiments.

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

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

7. Assign treatment to groups



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

Tr Tr Tr Tr Tr Tr Tr Tr

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

Ctl Ctl Ctl Ctl Ctl Ctl Ctl Ctl

Is this a good assignment?

Treatment I

Treatment II

Ronald Fisher

1920 CE, Design of 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

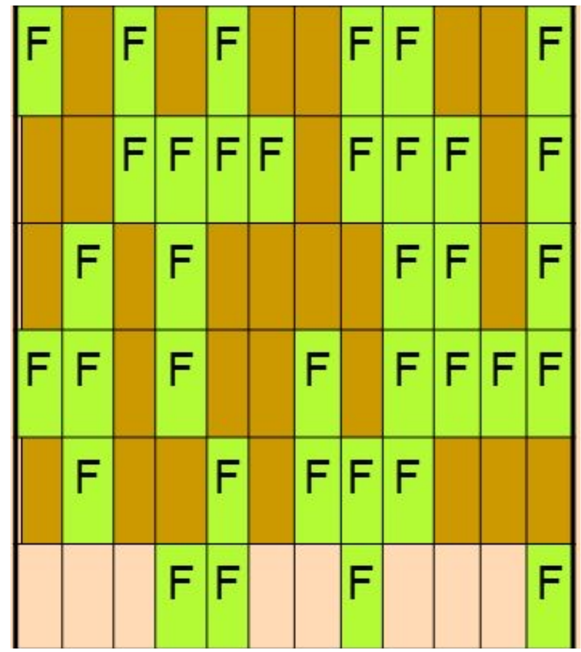
- eliminate as much of the natural variation as possible
- prevent unremoved variation from confusing or biasing the effects being tested
- detect cause and effect with the minimal amount of experimental effort necessary - time consuming and costly

Randomization

Fisher -> helps to avoid confusion or 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.



Well-designed experiments are characterized by three features: <u>randomization, replication, and local control.</u>

Randomization

- Randomly assign subjects to treatment and control groups in order to minimize bias and moderate experimental error.

- Assign random numbers to experimental so that any experimental unit (EU) has equal chances of being assigned to treatment or control (for example, odd-numbered EU in the treatment group, and even-numbered EU in the control group).

- 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.

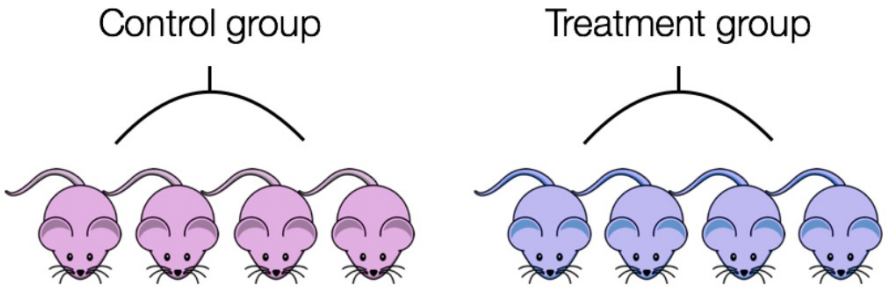
Blocking

- A rack of many mice cages is heterogeneous with respect to light exposure.
- The rack of cages can be divided into smaller blocks such that cages within each block tend to be more homogeneous (have equal light exposure).

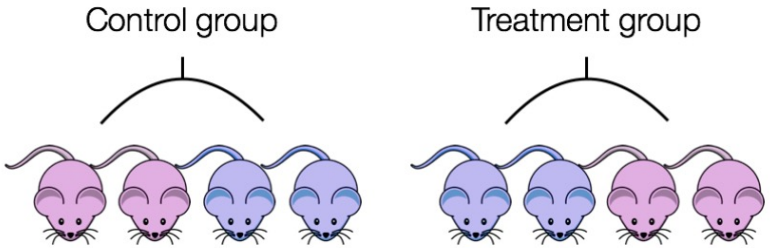
Blocking approach helps to reduce variability unexplained in the model

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.



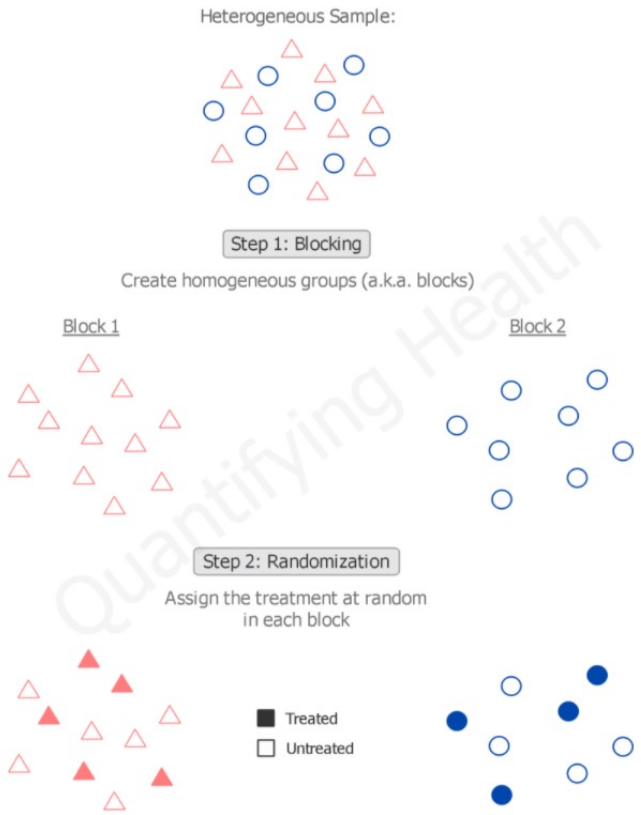
Capture effects of interest and avoid unwanted variation in experiment

- ✓ 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

Randomized block design I

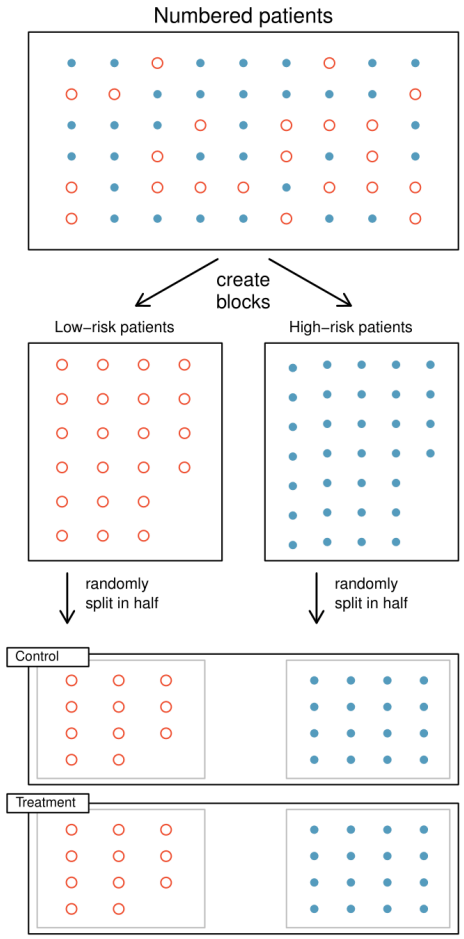


Treatment	
Placebo	Vaccine
500	500

Gender	Treatment	
	Placebo	Vaccine
Male	250	250
Female	250	250

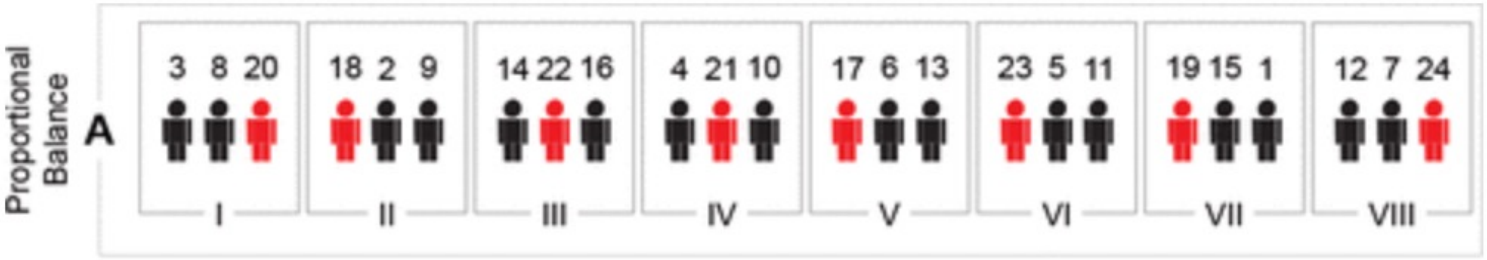
https://quantifyinghealth.com/randomized-block-design/
https://stattrek.com/experiments/experimental-design.aspx

Randomized block design II

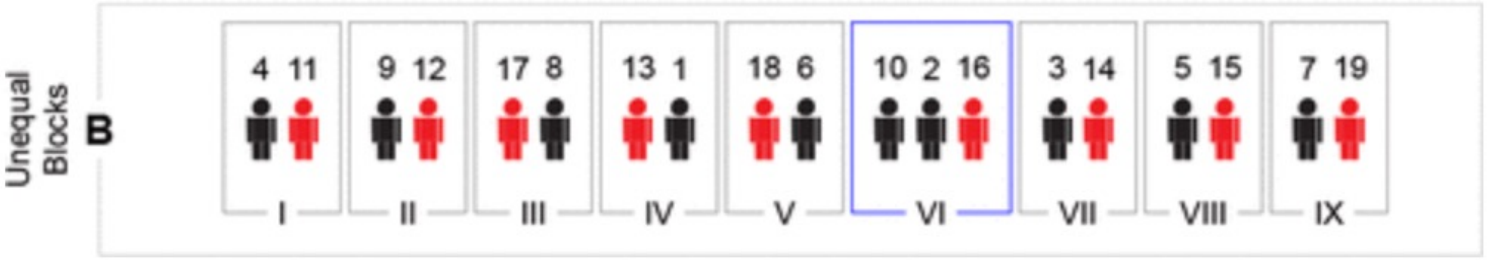


Randomized block design III: equal/unequal

Black: subjects 1–16 Placebo
Red: subjects 17–24 Treatment



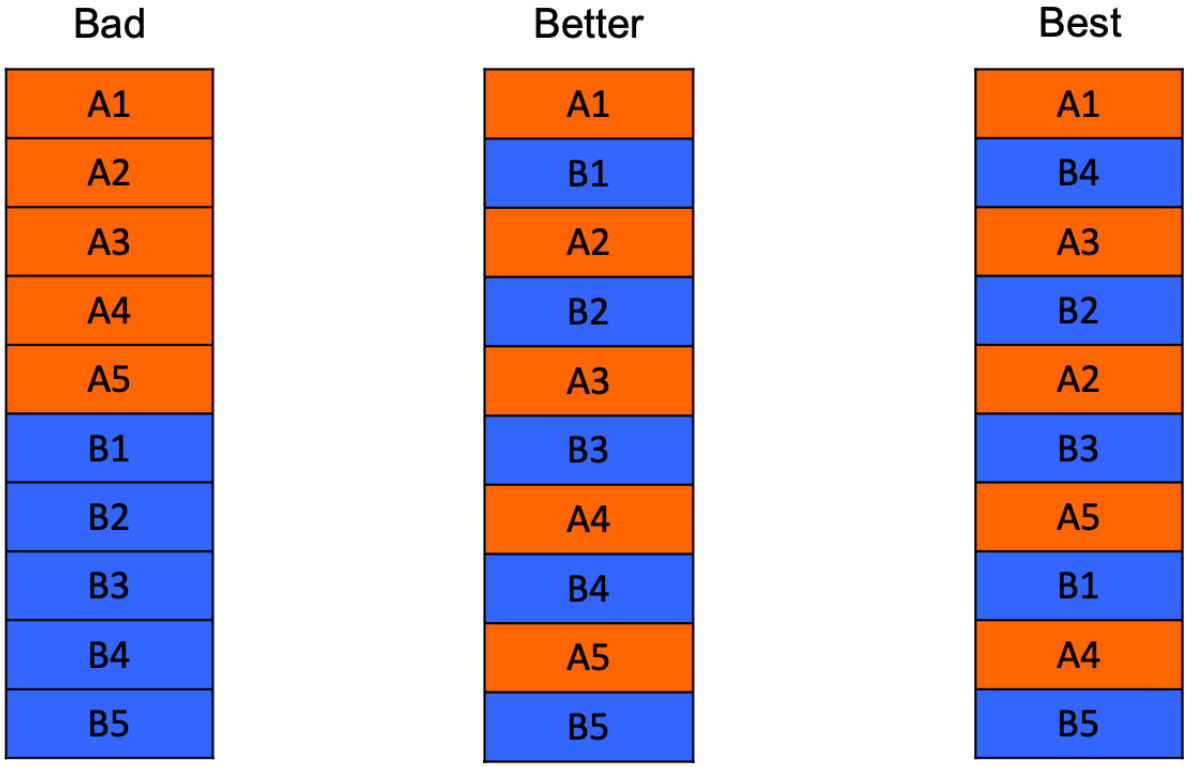
Black: subjects 1–10 Placebo
Red: subjects 11–19 Treatment



https://pubs.acs.org/doi/10.1021/acs.jproteome.0c00536

“Block what you can control; randomize what you cannot control”

Good vs Bad experimental design



- The default analysis assumes the data have come from a completely randomized design.
- In practice, this is often a false assumption.

Poll

	Design 1 – Sample prep date	Design 2 – Sample prep date
Sample_1_E9.5	Jan 9th, 2019	Jan 11th, 2019
Sample_2_E9.5	Jan 9th, 2019	Jan 9th, 2019
Sample_3_E9.5	Jan 9th, 2019	Jan 11th, 2019
Sample_4_E9.5	Jan 9th, 2019	Jan 9th, 2019
Sample_1_E11.5	Jan 11th, 2019	Jan 11th, 2019
Sample_2_E11.5	Jan 11th, 2019	Jan 9th, 2019
Sample_3_E11.5	Jan 11th, 2019	Jan 11th, 2019
Sample_4_E11.5	Jan 11th, 2019	Jan 9th, 2019

Which is better design?

1) Design 1
2) Design 2

Poll

	Design 1 – Gender	Design 2 - Gender
Sample_1_E9.5	Male	Male
Sample_2_E9.5	Male	Female
Sample_3_E9.5	Male	Male
Sample_4_E9.5	Male	Female
Sample_1_E11.5	Female	Male
Sample_2_E11.5	Female	Female
Sample_3_E11.5	Female	Male
Sample_4_E11.5	Female	Female

Which is better design?

1) Design 1
2) Design 2

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.

Between-subjects vs within subjects design

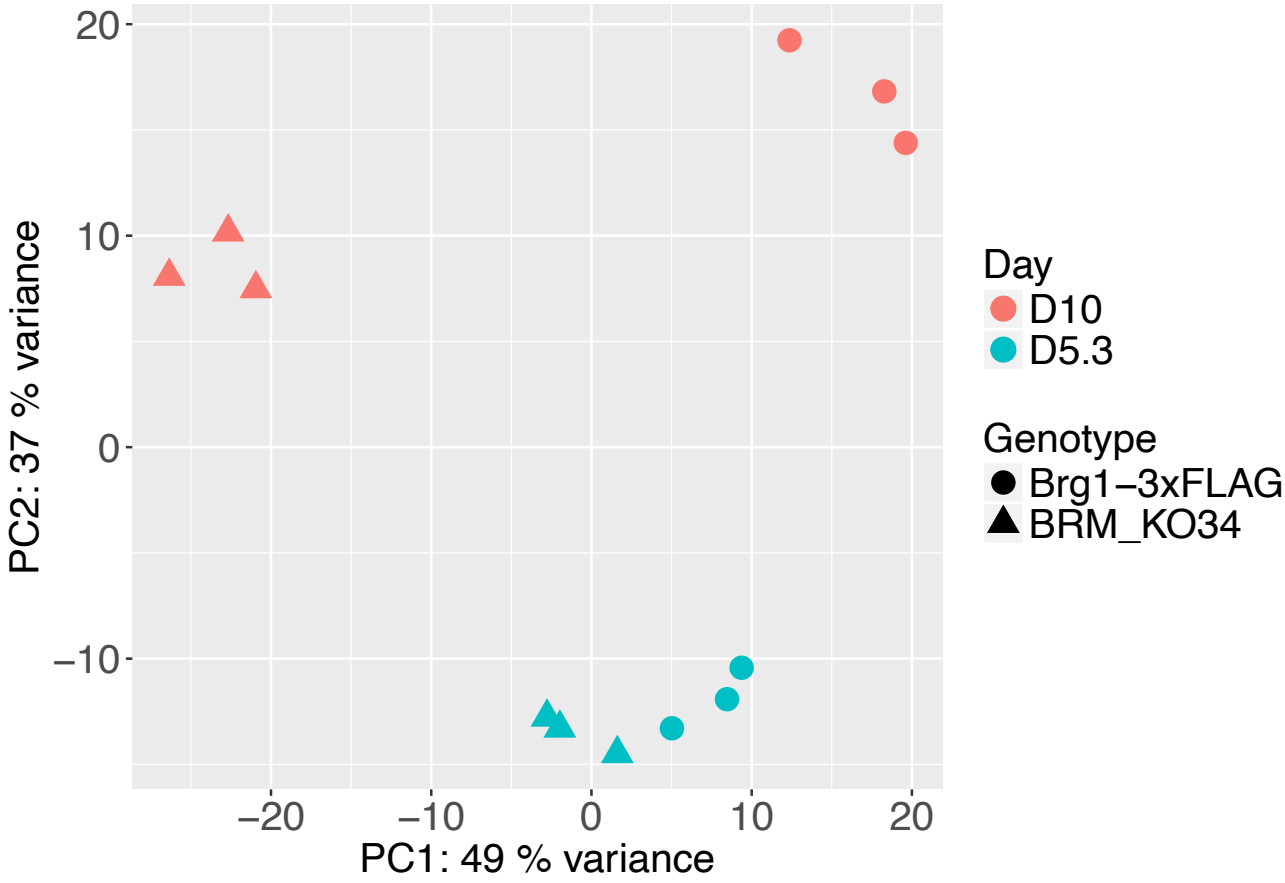
- In a within-subjects design (or between-groups design)

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

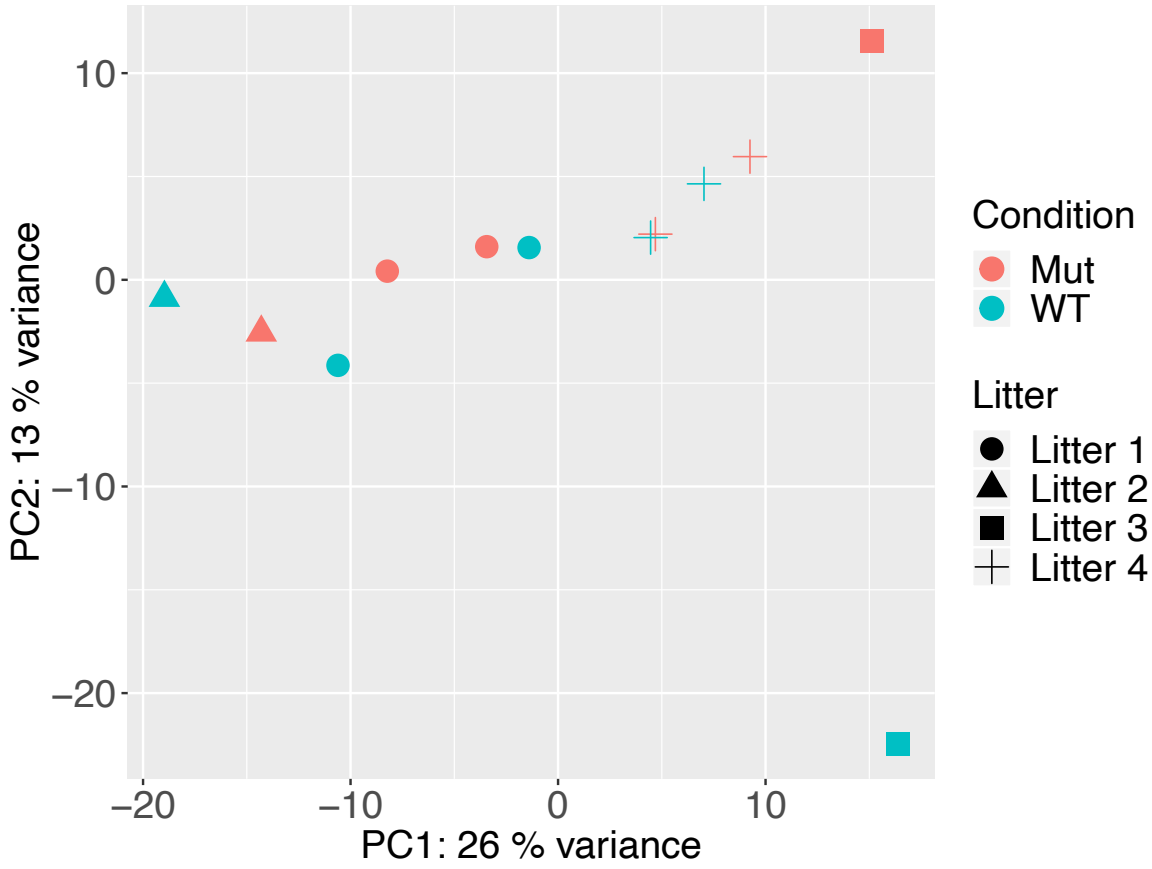
Also called a **dependent groups or repeated measures design** because compare related measures from the same individuals between different conditions.

8. Bad design introducing batch effects

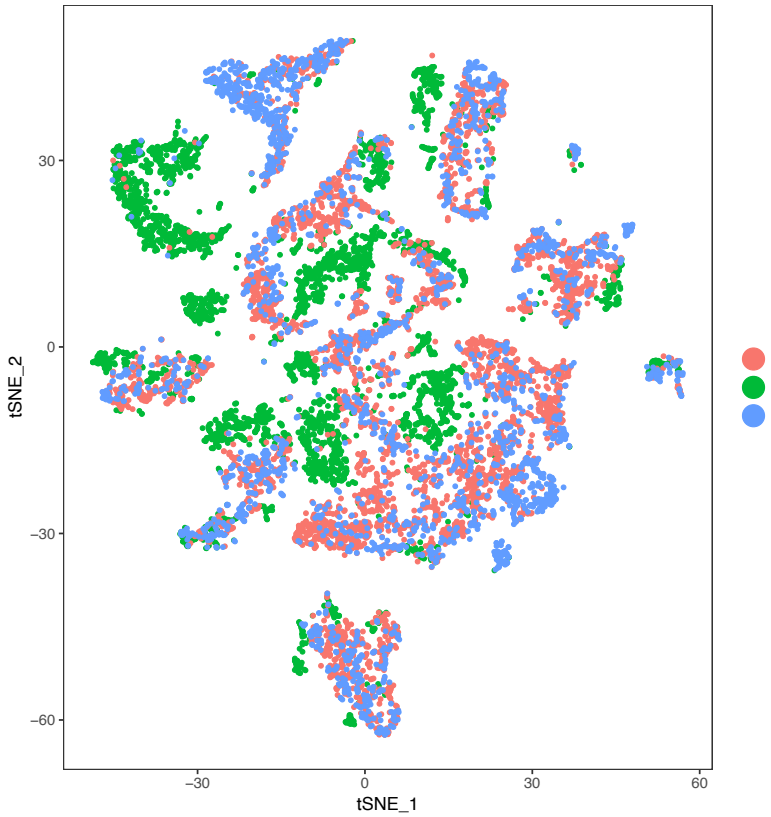
Genotype and development time effect on gene expression



Litter effect dominates the variation

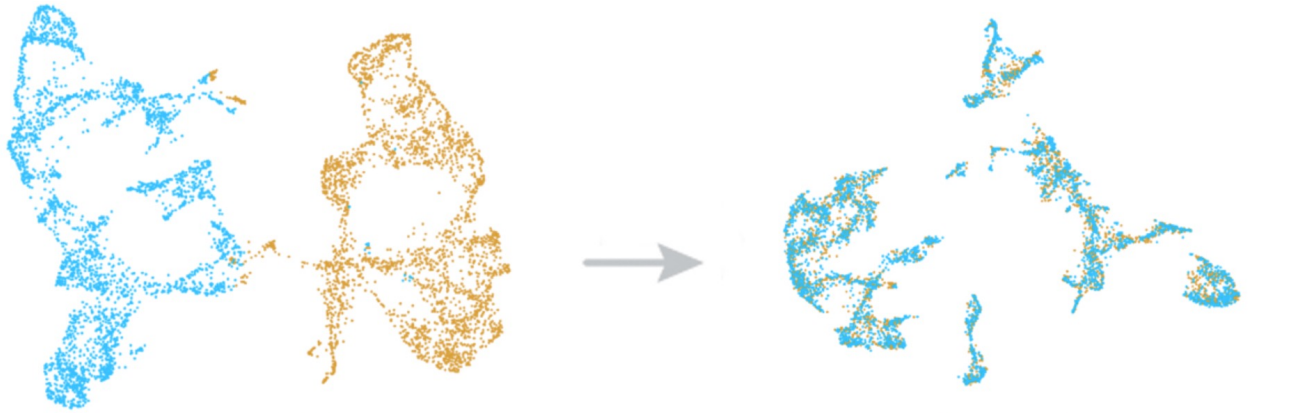


Confounding in scRNA-seq data is a problem



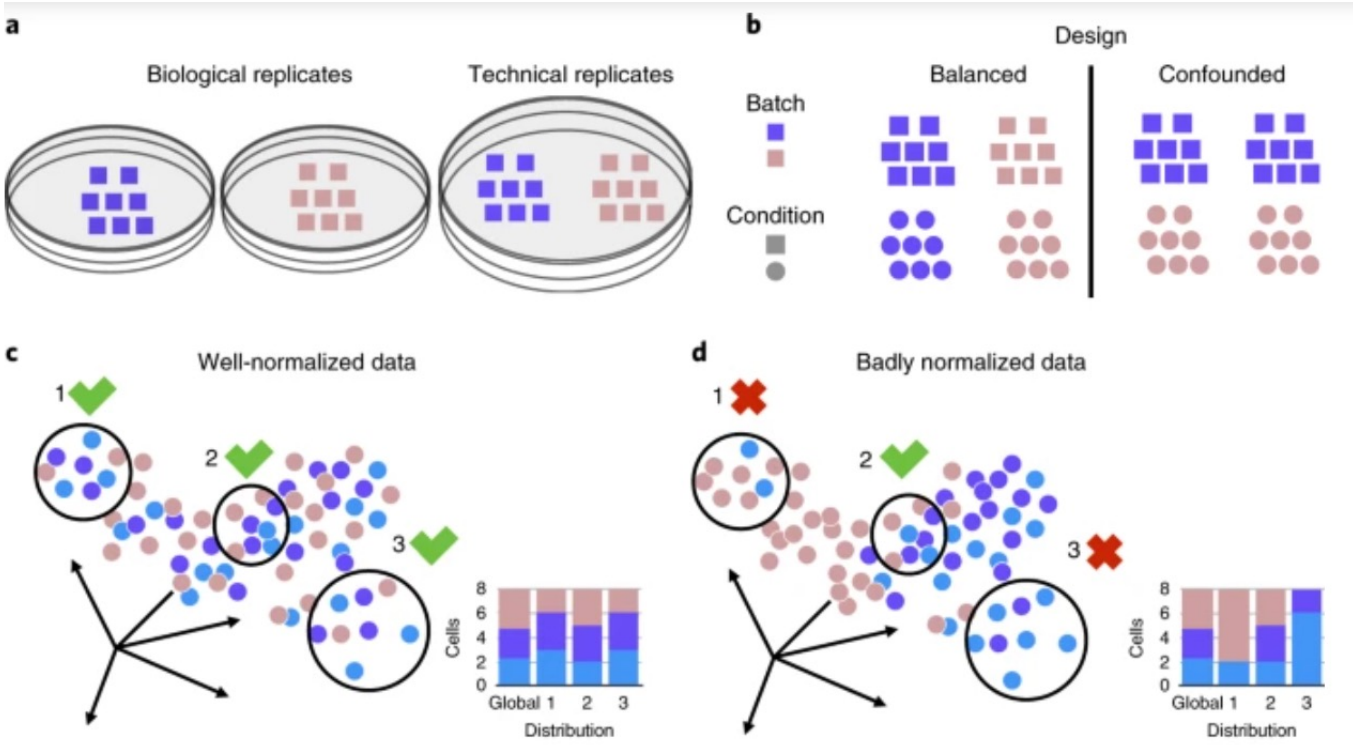
Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. Preprint available from:https://doi.org/10.1093/biostatistics/kxx053 (2017).

Confounding in scRNA-seq data is a problem



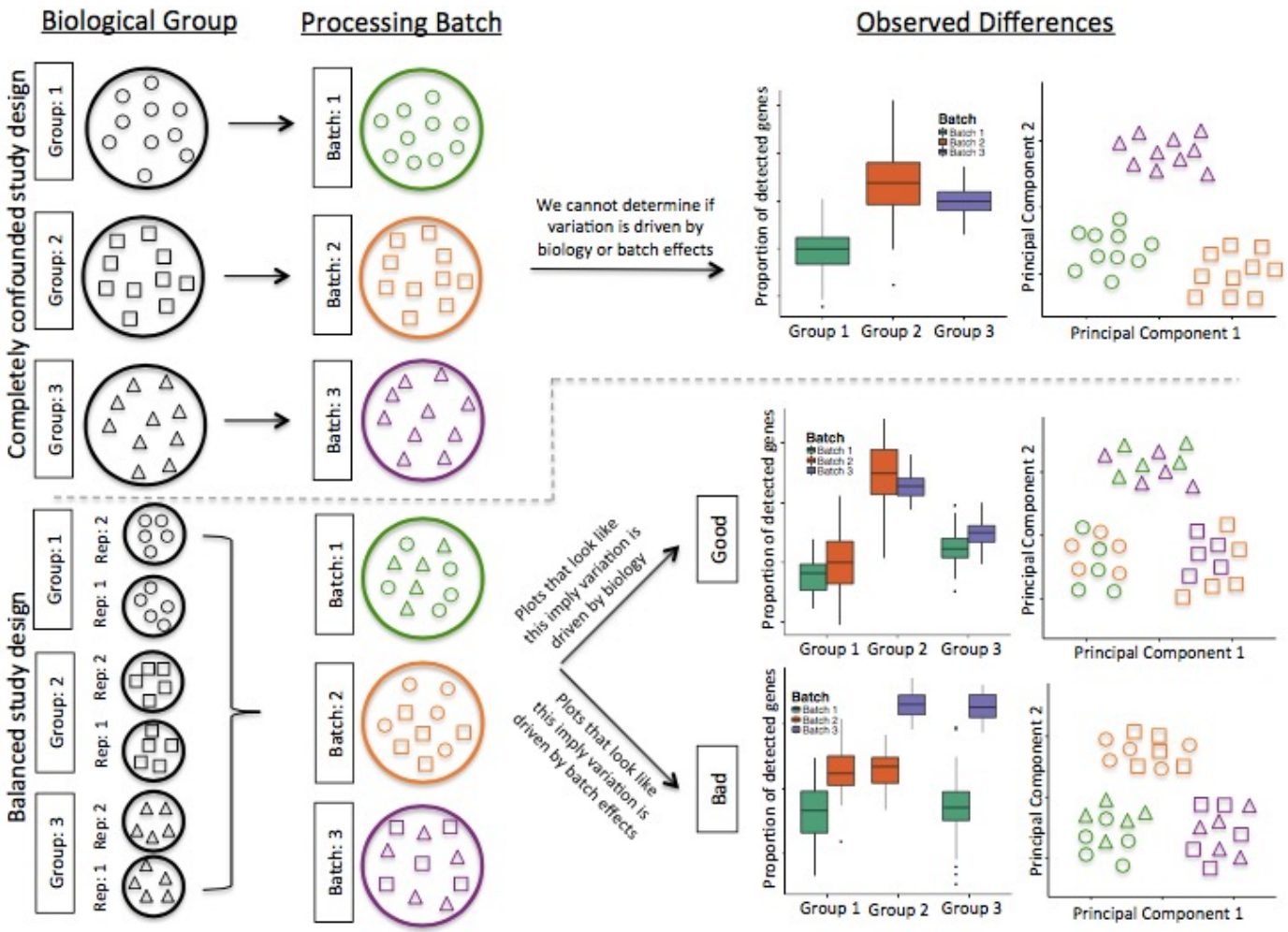
Batch Effect **Batch Effect Correction**

Well and badly analyzed scRNA-seq data

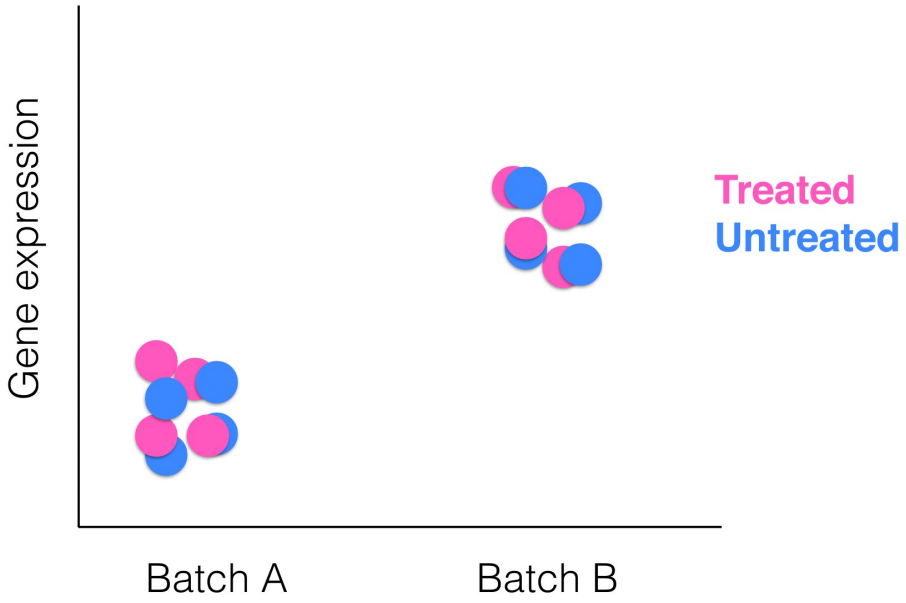


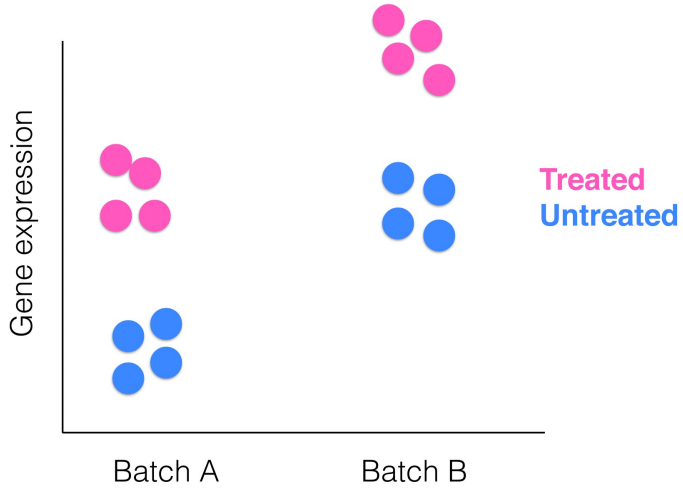
https://www.nature.com/articles/s41592-018-0254-1

Confounding biological variation and batch effects



Non confounding experiments – no 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.

scRNA-seq workshop – January 26th-27th focusing on batch effect day 2

Experimental design allow you to:

(1) choose an experimental design that is appropriate for the research problem at hand

(2) construct the design (performing randomization and determining the number of replicates)

(3) execute the plan to collect the data (or advise a colleague on how to do it)

(4) determine the model appropriate for the data

(5) fit the model to the data

(6) interpret the data and present the results in a meaningful way to answer the research question

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!

Upcoming workshop at Gladstone:

Winter series

January 26-27	Single Cell RNA-seq analysis
February 7	Introduction to Pathway Analysis
February 9-10	Single-Cell ATAC-Seq Data Analysis Part 1
February 17	Introduction to Cytoscape and Network Biology
February 23	Single-Cell ATAC-Seq Data Analysis Part 2
March 9	Intermediate Cytoscape Networks and Omics Data Visualization

For questions:

michela.traglia@gladstone.ucsf.edu

reuben.thomas@gladstone.ucsf.edu

Slack: #questionsforbioinformaticscore

Thank you!

Please take the survey:
https://www.surveymonkey.com/r/DY7K5ZY