Statistics primer: Experimental designs

Laurent Gatto

Introduction

> To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of. – Ronald Fisher

BUT

> Designing effective experiments need thinking about biology more than it does mathematical (statistical) calculations. [1]

The first quote highlights the fact that

> **Myth**: It does not matter how you collect data, there will always be a statistical 'fix' that will allow to analyse it. [1]

Poor experimental design

Badly designed experiments can lead to completely useless data, or worse, to completely erroneous conclusions that could lead to more waste of time, money, resources and scientific dead ends. Additionally, there can be ethical concerns in badly designed experiments.

Hypothesis-driven vs hypothesis-free (-generating) experiments

A **hypothesis** is a clearly defined description of how the system under study works, or is assumed to work. Ideally, use clear, single research questions, such as

- What is the effect of drug A on the transcriptome/proteome?
- Is drug A better than drug B in inducing a given effect?

NOT

- Do cells treated with drug A for 20 or 40 min express protein X but not Y? (accumulation of conditions, bad start)
- Let's test all these drugs over as many time points as we can and see if something changes. (that's not even a question!)

An experiment should then be designed to answer your hypothesis.

Hypothesis-driven vs hypothesis-free (-generating) experiments

Other experiments can be designed to explore/describe a biological system, or generate new hypotheses:

- genome sequencing
- the genome-wide binding site mapping of a transcription factor
- sub-cellular protein map

These experiments must still be carefully designed!

Experiment vs. study

Experiment: typically in a lab, highly controlled conditions, well-designed interventions, controlled times and intensities, stringent control over experimental variables and to draw very specific conclusions. Risk of lack of generalisation, such as extrapolating observations from cells or lab mice to humans.

Study: *in the wild*, much more variability in the data, un-controlled or even un-known variables (confounding effects) that affect what is studied. Requires **much bigger** sample size.

To test our **hypothesis**, we need to perform a comparison that will highlight the effect we are interested in. To do so, we will need different experimental conditions, including **controls**.

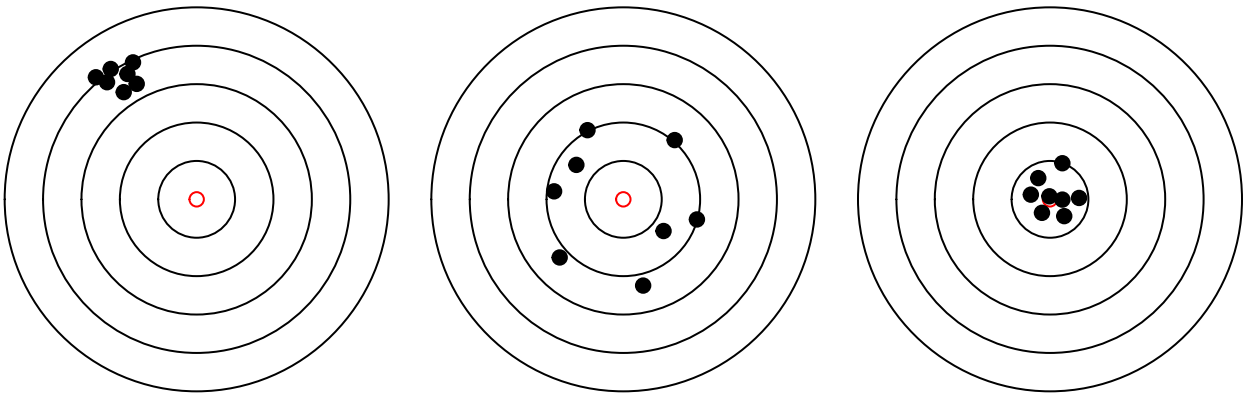
- What are the differences in expression between a control group and a group treated with drug A? Everything being equal, we will attribute observed changes to the drug.
- Effect of a drug over time. Time 0, no drug/effect vs time 1, time 2, ... (these time points will need to be defined.)
- **Positive control**: demonstrate that an experimental system works in principle.
- **Negative control**: quantify the baseline condition, assure that what we measure is not plain background.

Experimental units

- Mice, cultured cells, ...
- Time points

Precision vs accuracy

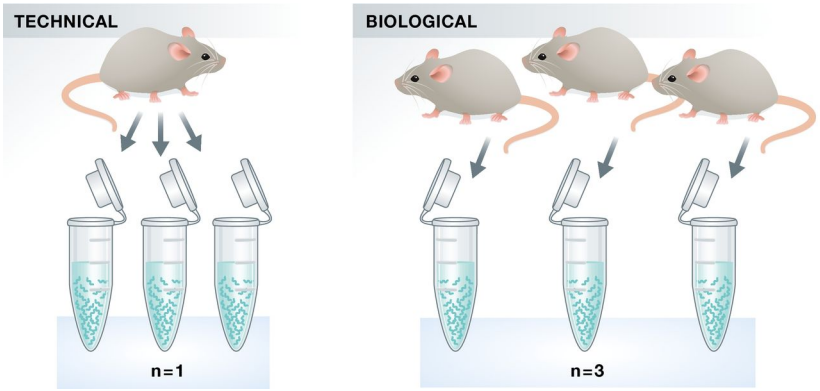
- **Precision** is how close the measured values are to each other.
- **Accuracy** is how close a measured value is to the actual (true) value.



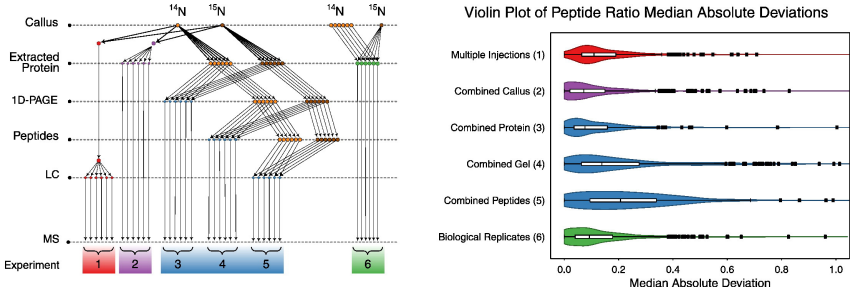
Variability and replication: what are we measuring?

- Noise/random variation vs systematic bias.
- **Technical** vs **biological** variability: the variability we measure is composed of technical and biological variability. If the former dominates, we won't learn anything about biology.

Technical vs biological replicates [2]



Assessing technical variability in N^{15} labelling [3]

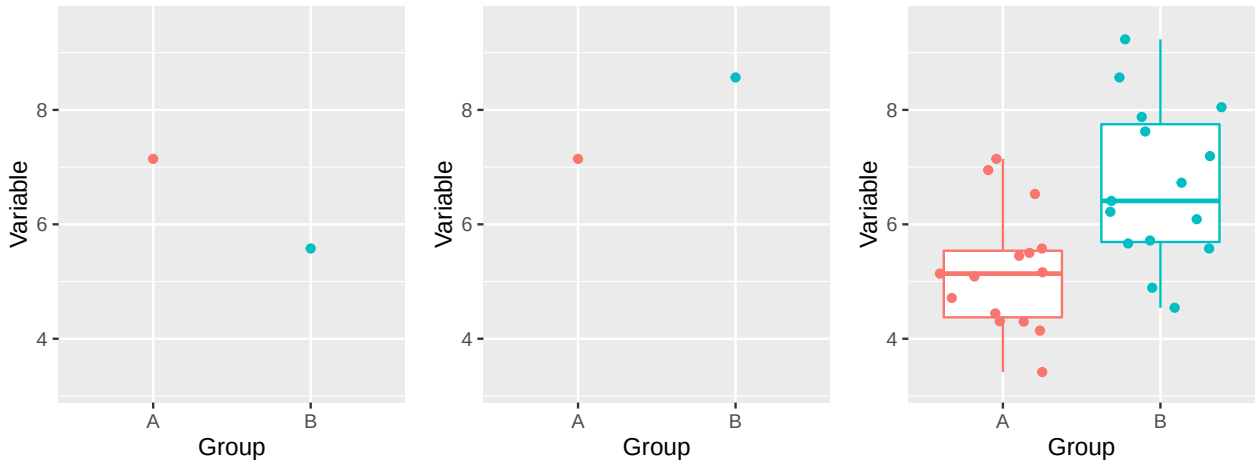


Replication

We can only assess an effect if we consider the variability in our experiment. How much change is there?

We can only observe variation if we repeat our measurement: repeat manipulation are take new measurement. Typically on different sample, because we are interested in biological variation.

Comparison of two groups: the difference is strong relative to the variability between the measurements within each group.




```
rnorm(n = 5, mean = 5, sd = 1.25)
```

```
## [1] 5.474549 4.372096 4.583491 3.726781 3.660261
```

```
summary(rnorm(n = 25, mean = 5, sd = 1.25))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.114   4.593   5.227   5.164   5.805   7.563
```

```
summary(rnorm(n = 25, mean = 7, sd = 2))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.664   5.744   6.529   6.943   8.097  11.370
```

Designing an experiment

We have 20 samples, 10 in each group. We can multiplex 10 samples (to reduce costs and variability). Never, ever do the following:

- Multiplex 10 treated samples and control 10 samples.
- Operator A runs the controls, operator B runs the treated samples.

Co-founding batch effect: any observed differences between the groups or interest could be due to technical variability between 2 data acquisitions/operators.

Assign samples to runs/operators at random.

Leek *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010

> One often overlooked complication with such studies is batch effects, which occur because measurements are affected by laboratory conditions, reagent lots and personnel differences. This becomes a major problem when batch effects are correlated with an outcome of interest and lead to incorrect conclusions. Using both published studies and our own analyses, we argue that batch effects (as well as other technical and biological artefacts) are widespread and critical to address. [5]

Blocks/batches

We aim to perform experiments within a **homogeneous group** of experimental units. These homogeneous groups, referred to as **blocks**, help to reduce the variability between the units and increase the meaning of differences between conditions (as well as the power of statistics to detect them). [2]

- If possible, measure as many (all) experimental conditions at the same time
- Minimise risks of day-to-day variability between the measurements

6 treatment conditions you want to apply to mice (the experimental units), but you can fit only 5 mice per cage (i.e. block).

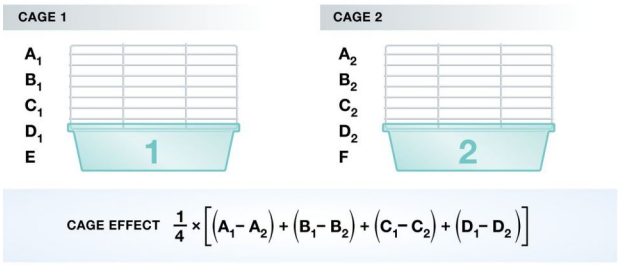


Illustration of batches and how to correct for them. All but two treatments have been applied to mice in two different cages (= batches). The batch/cage effect can now be computed based on the treatments that are shared between the cages.

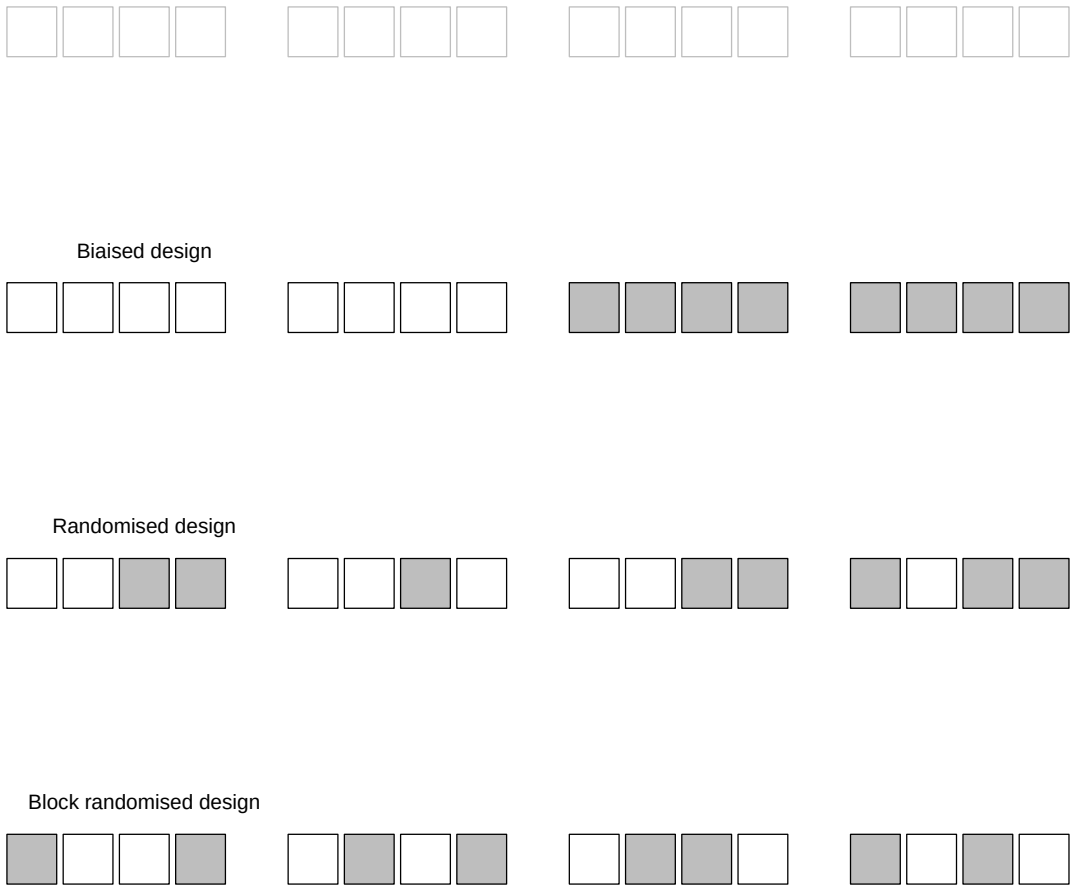
E and F are not directly comparable. However, the difference between E and F can be computed as E - F - *cage effect*.

Randomisation revisited

Even after blocking, other factors, such as age, generic background, sex differences, etc. can influence the experimental outcome.

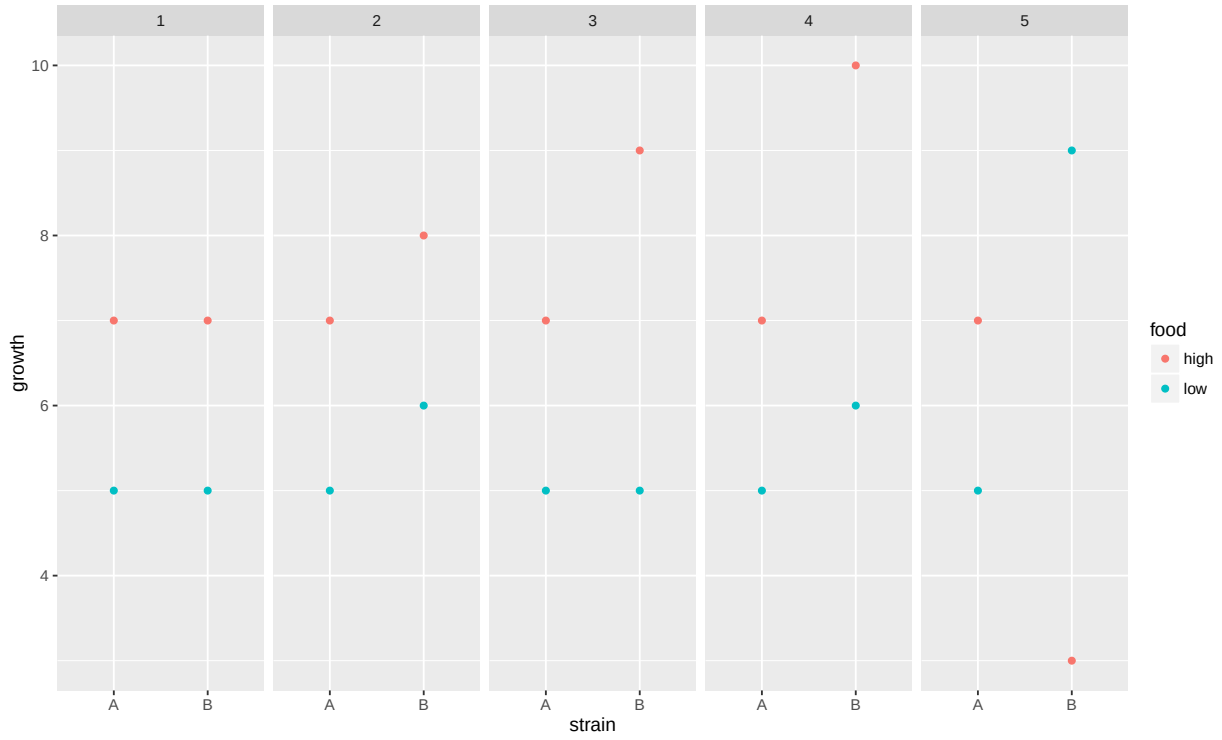
Randomise within blocks, to reduce confounding effects by equalising variables that influence experimental units and that have not been accounted for in the experimental design.

More designs



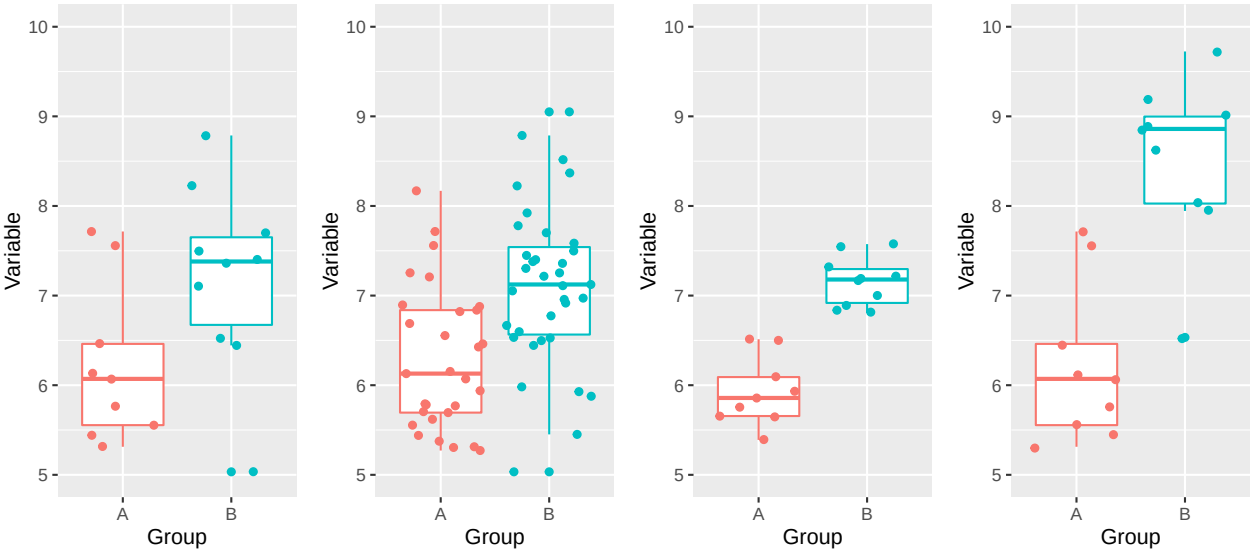
- Balanced vs. non-balanced
- One-way (one-factor) designs: one experimental factor, n levels (Treatments; drug A, drug B, control)

- Two-way design: treatment (A, B, control) x genetic background (WT, MT)
- Main effect (like 2 one-way designs) and interaction effects



How many replicates?

How variable are your groups? What effect size (magnitude) are you expecting to see?



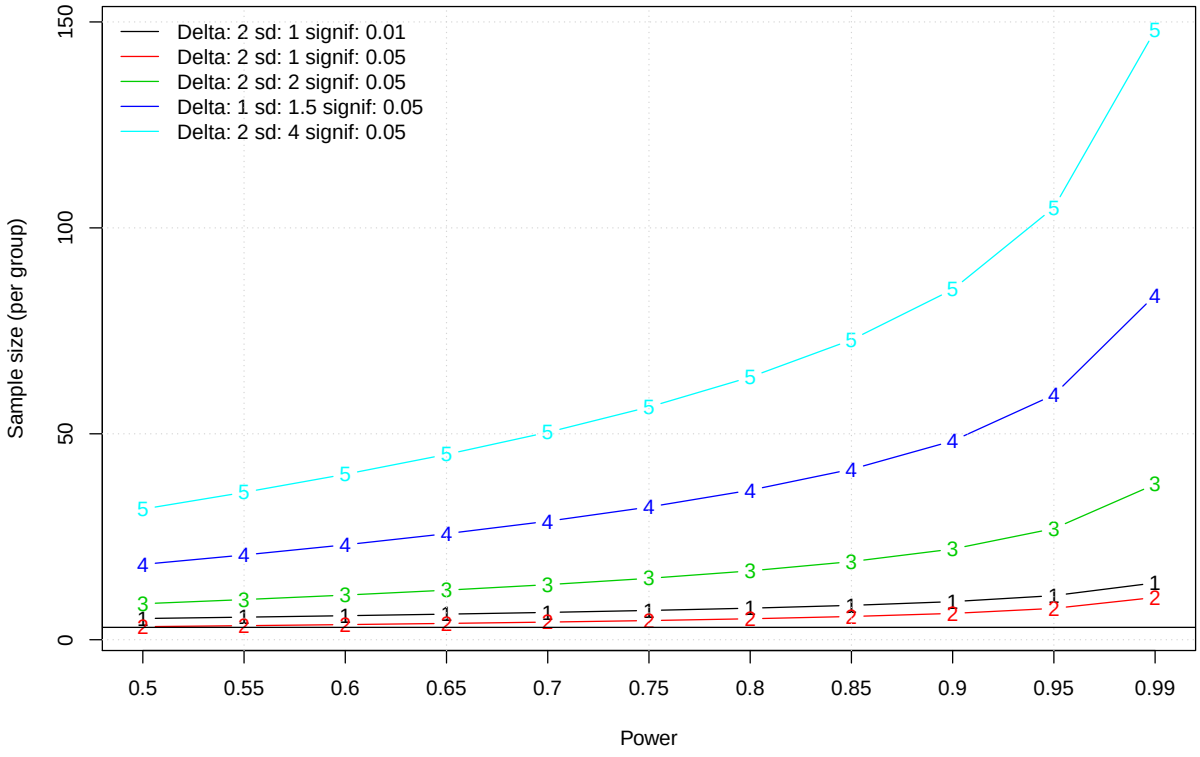
(A) 10 samples, means 6 and 7, sd 1. (B) Same as (A), but 25 additional samples. (C) Same as (A) but tighter distribution (sd 0.5). (D) Bigger effect.

How many replicates?

If too few samples are used, real difference will not be detected. The study is said to be under powered.

Statistical power is the probability of a test to detect an effect, if the effect actually exists. Power is affected by (1) effect size, (2) random variation and (3) sample size.

- As many as possible
- As many as you can afford
- Educated guesswork
- Power analysis
- Pilot study




```
power.t.test(delta = 2, sd = 1,
             sig.level = 0.01, power = 0.9)
```

```
##
##      Two-sample t test power calculation
##
##               n = 9.251528
##           delta = 2
##              sd = 1
##       sig.level = 0.01
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Next steps

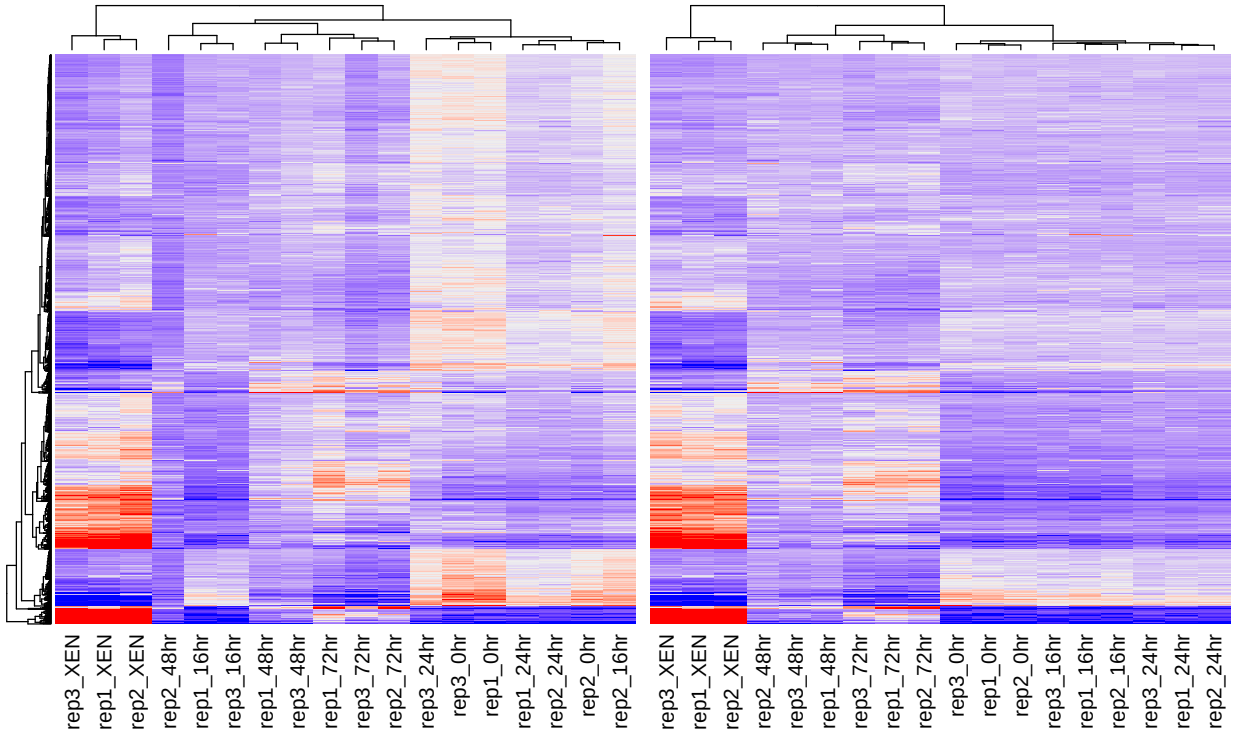
Once you have the data, explore it - Exploratory Data Analysis (EDA) using PCA plots, clustering, histograms, boxplots, . . .

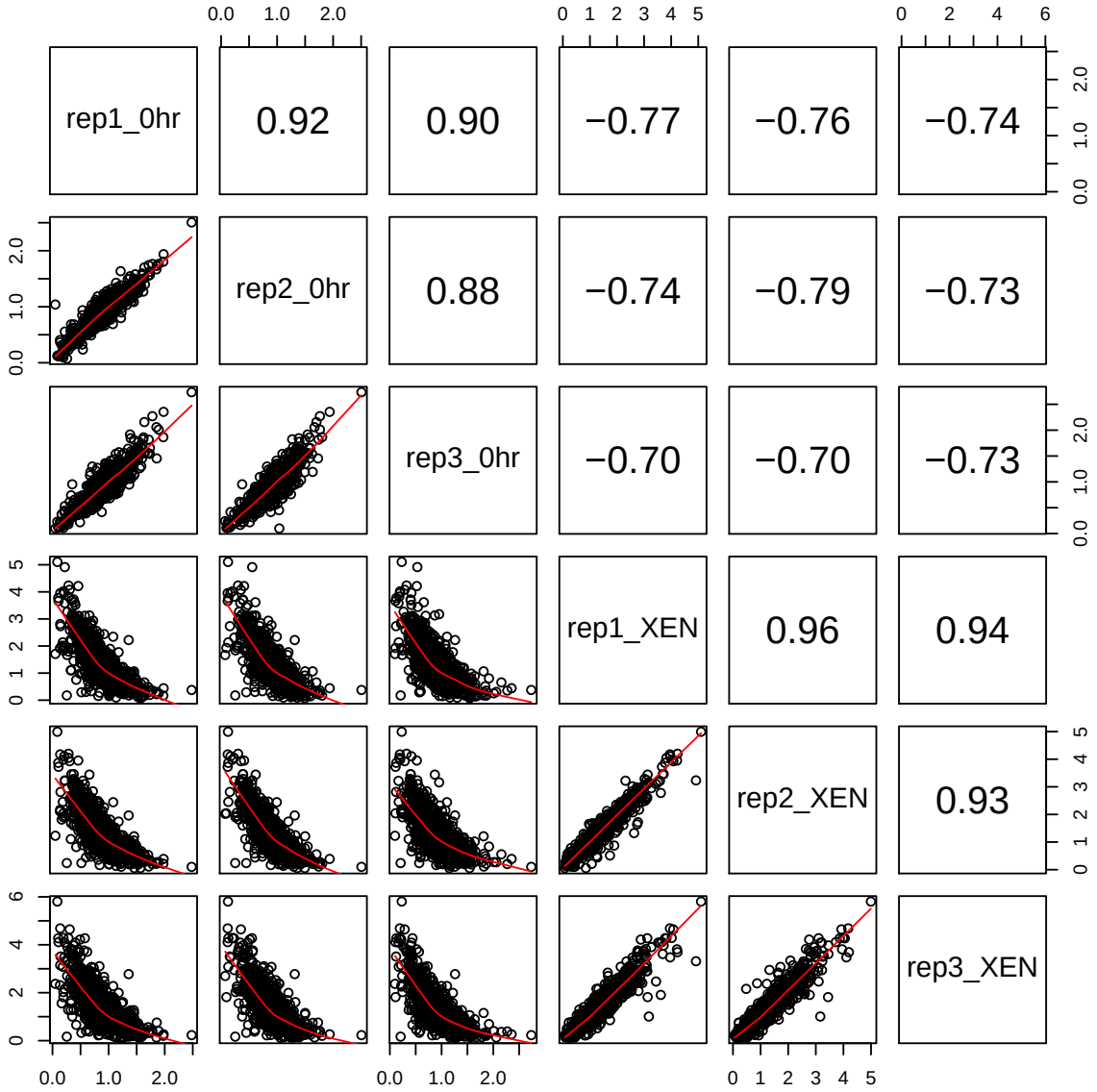
- Does the variability match your design (within/between groups)?
- Can you identify co-founding/batch effects?
- Identify significant changes

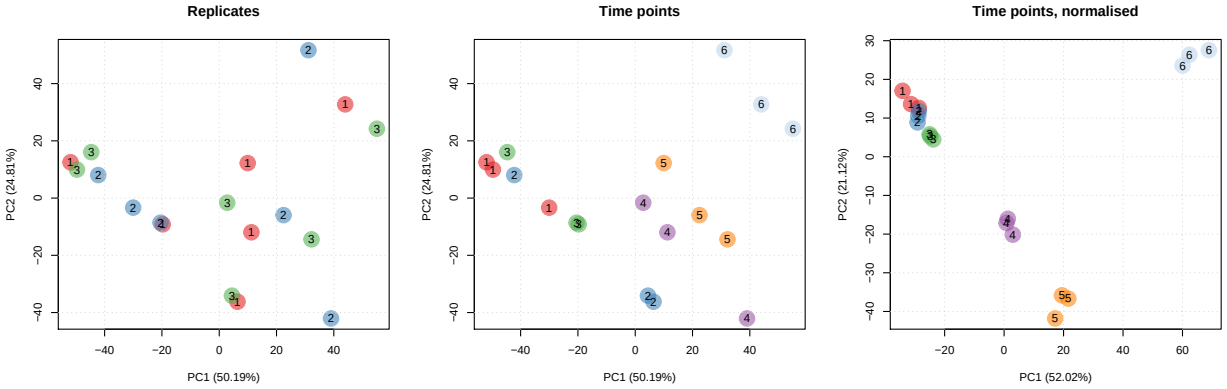
Examples

Mulvey CM, Schröter C, Gatto L, Dikicioglu D, Fidaner IB, Christoforou A, Deery MJ, Cho LT, Niakan KK, Martinez-Arias A, Lilley KS. *Dynamic Proteomic Profiling of Extra-Embryonic Endoderm Differentiation in Mouse Embryonic Stem Cells.* Stem Cells. 2015 Sep;33(9):2712-25. doi:10.1002/stem.2067.

> Here, we use this GATA-inducible system to quantitatively monitor the dynamics of global proteomic changes during the early stages of this differentiation event and also investigate the fully differentiated phenotype, as represented by embryo-derived XEN cells.




rep1_0hr
0.92
0.90
−0.77
−0.76
−0.74
rep2_0hr
0.88
−0.74
−0.79
−0.73
rep3_0hr
−0.70
−0.70
−0.73
rep1_XEN
0.96
0.94
rep2_XEN
0.93
rep3_XEN



Summary

When designing an experiments, make sure you address:

- Hypothesis and experimental groups (controls, conditions)
- Sources of variability and adequacy of replication
- Effect of technology (multiplexing to reduce/share technical variability, missing values, . . .)
- What is the question, does the design address it, how will you test it?
- Enough samples to see any desired effect?
- Keep it simple (n-way design vs one-way)
- Think about EDA (how will you verify that things went according to plan) and analysis beforehand.

References

[1] Ruxton GD, Colegrave N (2010) Experimental Design for the Life Sciences, 3rd edn. Oxford: Oxford University Press.

[2] Klaus B. Statistical relevance-relevant statistics, part I. EMBO J. 2015 Nov 12;34(22):2727-30.

[3] Russell M, Lilley KS., Pipeline to assess the greatest source of technical variance in quantitative proteomics using metabolic labelling, Journal of Proteomics, Dec 2012;77:441-454

[4] Cairns DA. Statistical issues in quality control of proteomic analyses: good experimental design and planning. Proteomics. 2011 Mar;11(6):1037-48. doi: 10.1002/pmic.201000579.

[5] Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010 Oct;11(10):733-9. doi: 10.1038/nrg2825. Epub 2010 Sep 14. PMID:20838408; PMC3880143.