Experimental Design

Mikhail Dozmorov
Fall 2017

What is 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.

What characterizes well-designed experiments?

- Effects can be estimated unambiguously and without bias.
- Estimates are precise.
- Protected from possible one-off events that might compromise the results.
- Easy to conduct.
- Easy to analyse and interpret.
- Maximum information obtained for fixed time, resources, and samples.
- Applicability of the findings to a wide variety of subjects, conditions, and situations.

Why Design an Experiment?

- The goal of an experiment dictates everything from how the samples are collected to how the data are generated
- The design of the analytical protocol should be reflected in the design
  - Do we have enough replicates?
  - Do we have sufficient controls?
  - Do we collect samples and data to avoid confounding and batch effects?

Types of Experiments

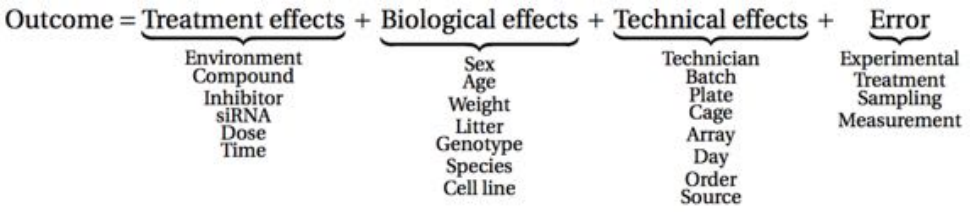
Class Comparison
- Can I find genes that distinguish between two classes, such as tumor and normal?

Class Discovery
- Given what I think is a uniform group of samples, can I find subsets that are biologically meaningful?

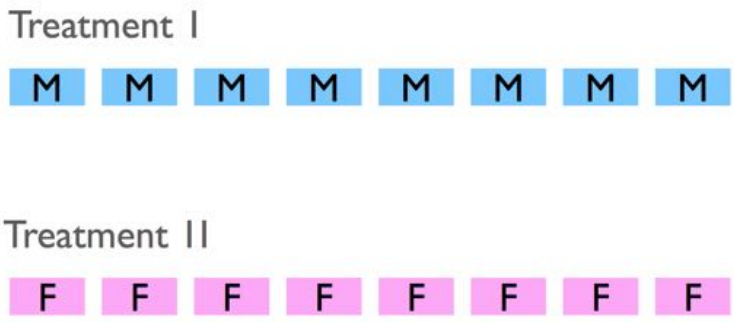
Classification
- Given a set of samples in different classes, can I assign a new, unknown sample to one of the classes?

Large-scale Functional Studies
- Can I discover a causative mechanism associated with the distinction between classes? These are often not perfectly distinct.

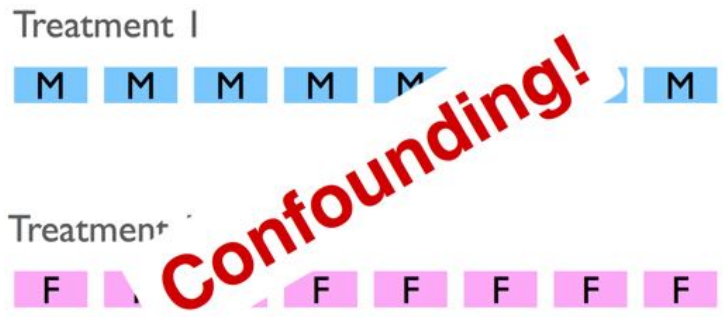
What affects the outcome of an experiment?



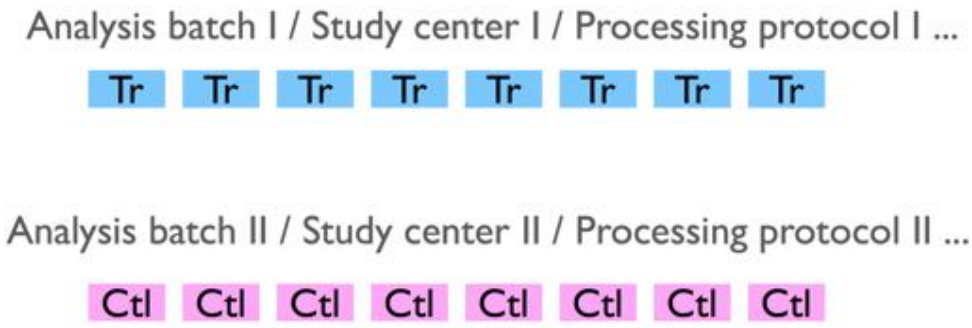
What is bad Bad experimental design - examples



What is bad Bad experimental design - examples



What is bad Bad experimental design - examples



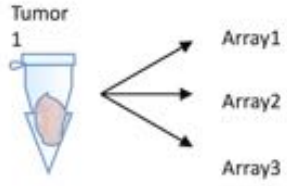
What would be a better experimental design?

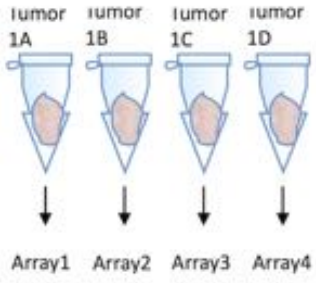
- Process all samples at the same time/in one batch (not always feasible)
- Minimize confounding as much as possible through
  - blocking
  - randomization
- The batch effect will still be there, but with an appropriate design we can account for it

Principles of experimental design

- **Replication**. It allows the experimenter to obtain an estimate of the experimental error
- **Randomization**. It requires the experimenter to use a random choice of every factor that is not of interest but might influence the outcome of the experiment. Such factors are called nuisance factors
- **Blocking**. Creating homogeneous blocks of data in which a nuisance factor is kept constant while the factor of interest is allowed to vary. Used to increase the accuracy with which the influence of the various factors is assessed in a given experiment
- **Block what you can, randomize what you cannot**

Replicates

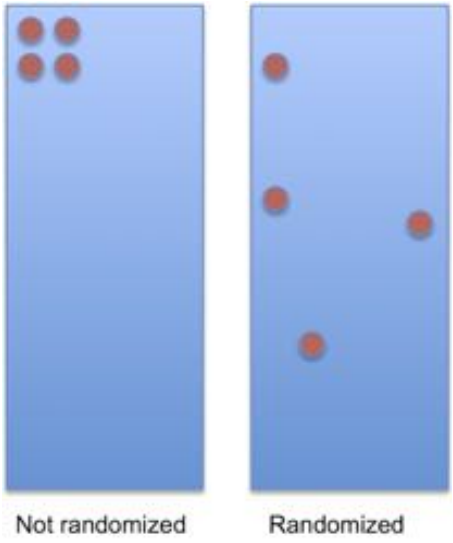




- **Technical** replicates and **Biological** replicates
- Rule of thumb: for two-fold change – use 3 replicates
- Smaller change – 5 replicates

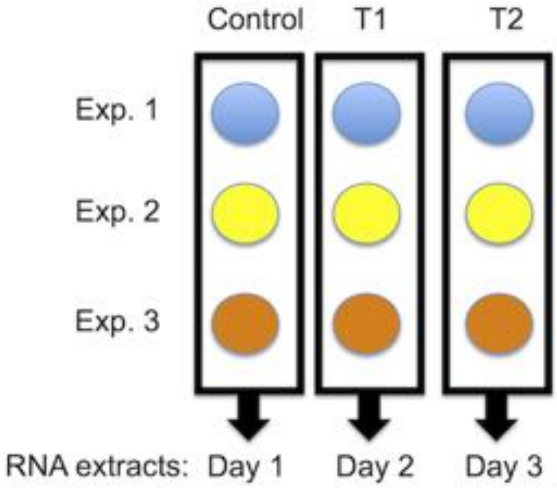
Randomization

- Each gene has multiple probes – randomize their position on the slide



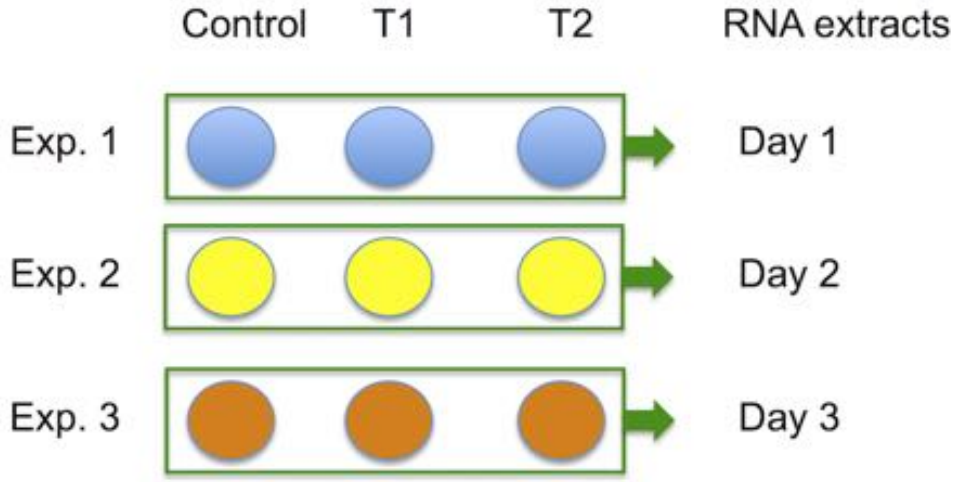
Blocking

- Treatment and RNA extraction days are confounded

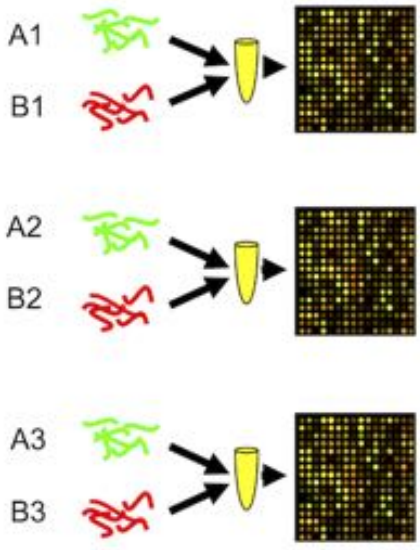


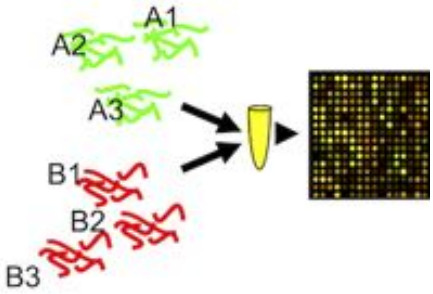
Blocking

- Block replicated experiments



Pooling





Pooling

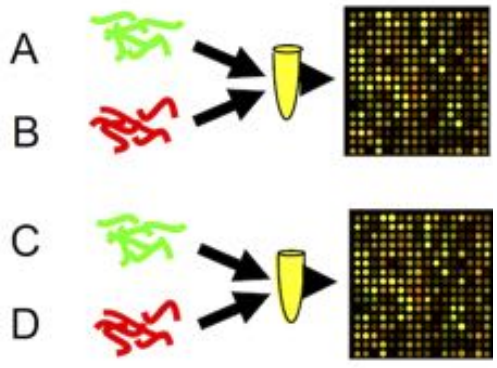
- OK when the interest is not on the individual, but on common patterns across individuals (population characteristics)
- Results in averaging -> reduces variability -> substantive features are easier to find
- Recommended when fewer than 3 arrays are used in each condition
- Beneficial when many subjects are pooled

"inference for most genes was not affected by pooling"

C. Kendziorski, R. A. Irizarry, K.-S. Chen, J. D. Haag, and M. N. Gould. "On the utility of pooling biological samples in microarray experiments". PNAS March 2005, 102(12) 4252-4257

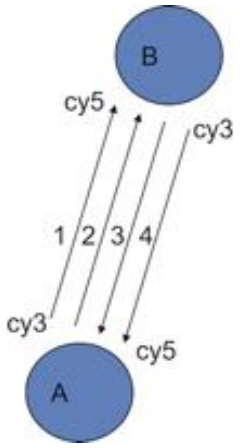
How to allocate the samples to microarrays?

- which samples should be hybridized on the same slide?
- how different experimental designs affect outcome?
- what is the optimal design?



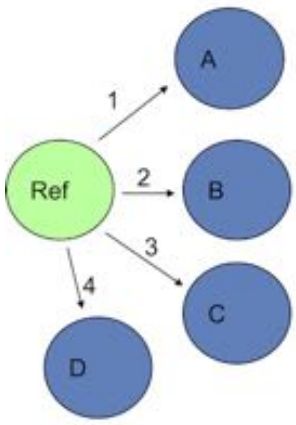
Example of four-array experiment

Dye swap



array	cy3	cy5	log(cy5/cy3)
1	A	B	log(B) – log(A)
2	A	B	log(B) – log(A)
3	B	A	log(A) – log(B)
4	B	A	log(A) – log(B)

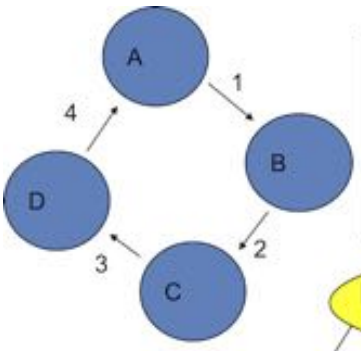
Common reference design



array	cy3	cy5	log(cy5/cy3)
1	Ref	A	log(A) – log(Ref)
2	Ref	B	log(B) – log(Ref)
3	Ref	C	log(C) – log(Ref)
4	Ref	D	log(D) – log(Ref)

log(C/A) = log(C) - log(A)
= log(C) - log(Ref) + log(Ref) - log(A)
= log(C) - log(Ref) – (log(A) - log(Ref))
= logratio(array3) - logratio(array1)

Loop design



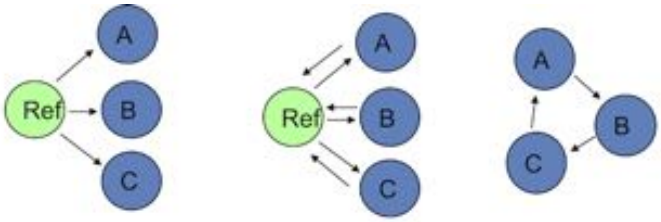
array	cy3	cy5	log(cy5/cy3)
1	A	B	log(B) – log(A)
2	B	C	log(C) – log(B)
3	C	D	log(D) – log(C)
4	D	A	log(A) – log(D)

log(C/A) = log(C) – log(B) + log(B) – log(A)
= logratio(array2) + logratio(array1)

log(C/A) = log(C) – log(D) + log(D) – log(A)
= - logratio(array3) - logratio(array4)

log(C/A)=(logratio1 + logratio2)/2

Comparing the designs



	reference design	reference design with replicates	loop design
number of arrays	3	6	3
amount of RNA required per sample	1+Ref	2+Ref	2
error	2.0	1.0	0.67

Design with all direct pairwise comparisons