RNA-seq experimental considerations

Mikhail Dozmorov

2021-03-03

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.

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?

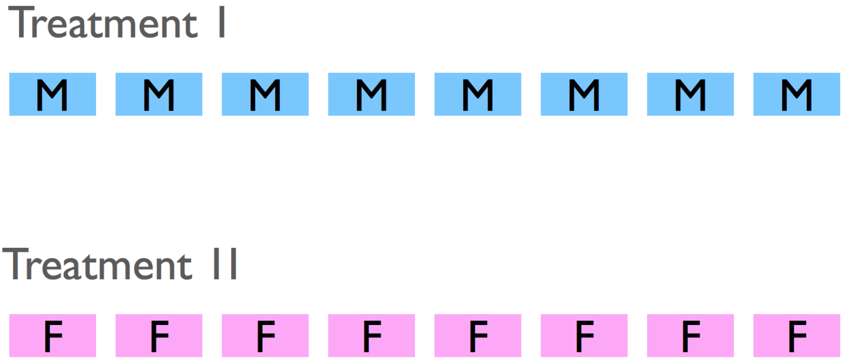
Classification
- Given a set of samples in different classes, can I assign a new, unknown sample to one of the classes?

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?

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

What is bad experimental design - examples



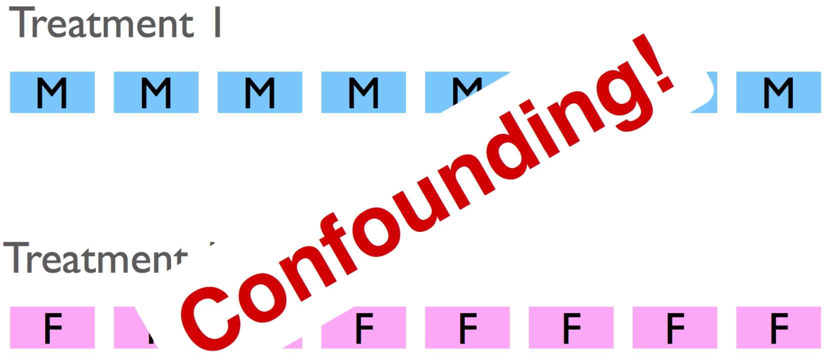
Treatment I

M M M M M M M M

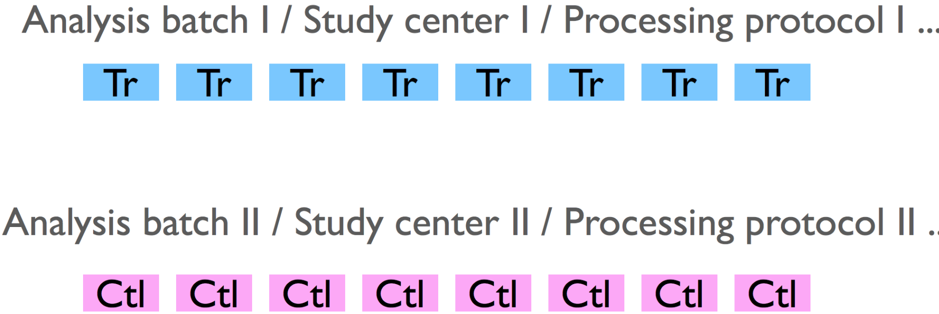
Treatment II

F F F F F F F F

What is bad experimental design - examples



What is bad experimental design - examples



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

Sources of variability in RNA-seq measures

In RNA-seq, we have multiple levels of randomness:

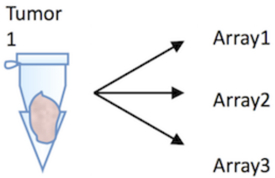
- Biological variability in samples
- Stochasticity of RNA content
- Randomness of fragments being sequenced
- Technical variability

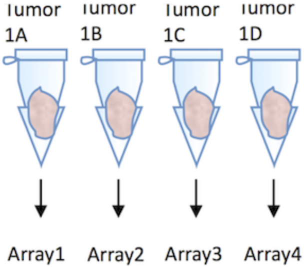
Auer, P.,RW Doerge. "Statistical Design and Analysis of RNA Sequencing Data." Genetics, 2010
http://www.genetics.org/content/185/2/405.long

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

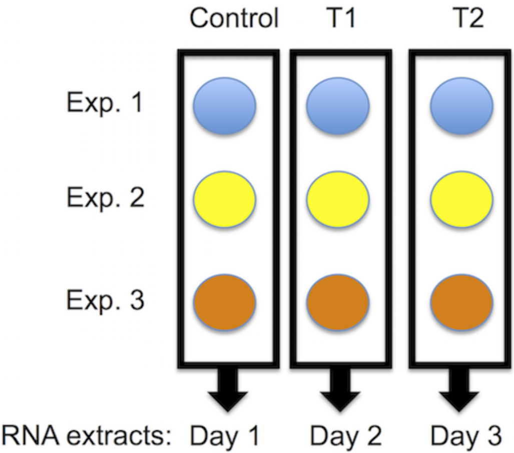




- **Technical** replicates and **Biological** replicates
- Rule of thumb: for two-fold change – use 3 replicates
- Smaller change – 5 replicates

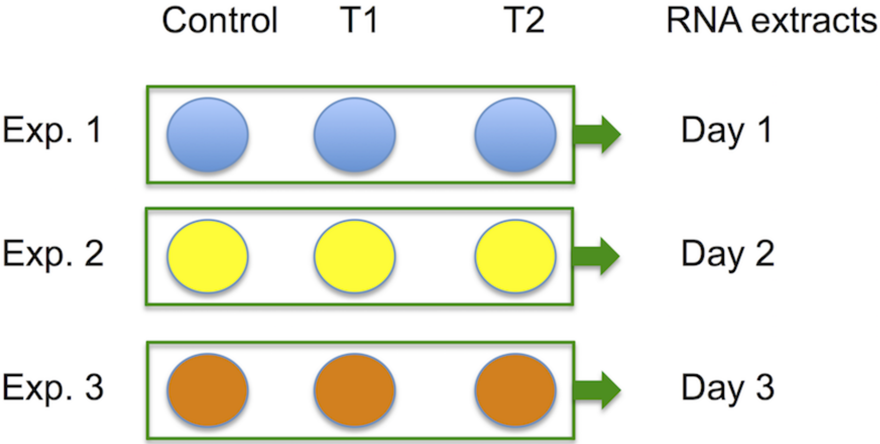
Blocking

- Treatment and RNA extraction days are confounded

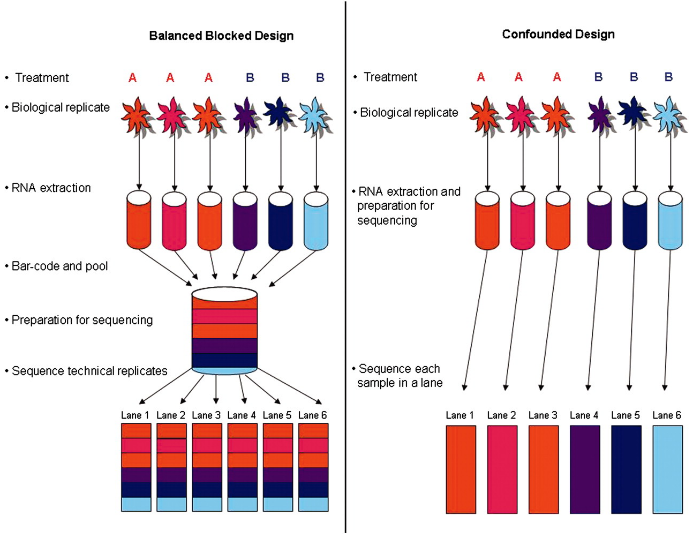


Blocking

- Block replicated experiments



Experimental design: Multiplexing balances technical variability



Sequencing length/depth

- Longer reads improve mappability and transcript quantification
- More transcripts will be detected and their quantification will be more precise as the sample is sequenced to a deeper level
- Up to 100 million reads is needed to precisely quantify low expressed transcripts
- In reality, 20-30 million reads is OK for human genome

Power calculations

- **Scotty** - Power Analysis for RNA Seq Experiments
- **powerSampleSizeCalculator** - R scripts for power analysis and sample size estimation for RNA-Seq differential expression
- **RnaSeqSampleSize** - R package and a Shiny app for RNA sequencing data sample size estimation
- **RNASeqPower** - R package for RNA-seq sample size analysis

http://scotty.genetics.utah.edu/, Busby MA, Stewart C, Miller CA, Grzeda KR, Marth GT. "Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression". *Bioinformatics* 2013 https://www.ncbi.nlm.nih.gov/pubmed/23314327

http://www2.hawaii.edu/~lgarmire/RNASeqPowerCalculator.htm, Travers C. et.al. "Power analysis and sample size estimation for RNA-Seq differential expression" *RNA* 2014 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4201821/

https://cqs.mc.vanderbilt.edu/shiny/RNAseqPS/, Guo et.al. "RNAseqPS: A Web Tool for Estimating Sample Size and Power for RNAseq Experiment" *Cancer Informatics* 2014 http://insights.sagepub.com/rnaseqps-a-web-tool-for-estimating-sample-size-and-power-for-rnaseq-ex-article-a4433

https://bioconductor.org/packages/release/bioc/html/RNASeqPower.html, Svensson, V. et.al. "Power Analysis of Single-Cell RNA-Sequencing Experiments." *Nature Methods* 2017 http://www.nature.com/nmeth/journal/v14/n4/pdf/nmeth.4220.pdf