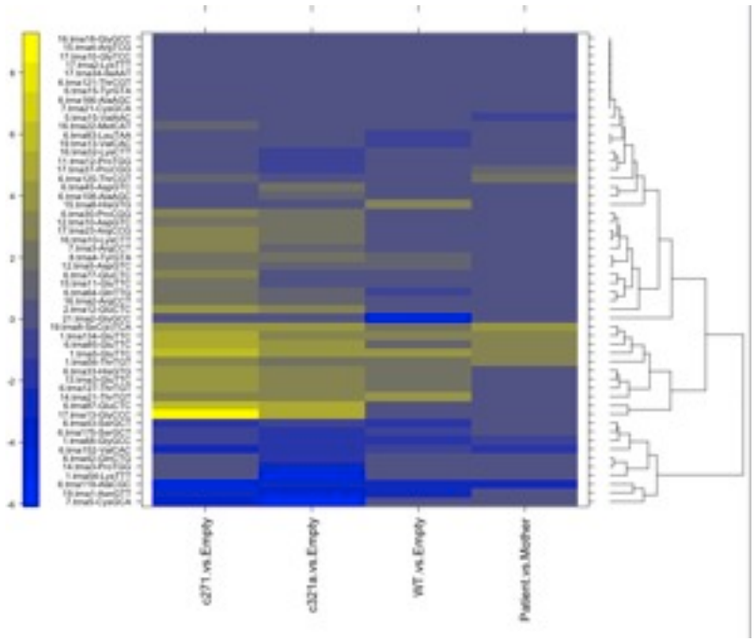
Making your science powerful: an introduction to NGS experimental design



Dr Jelena Aleksic
University of Cambridge
ja313@cam.ac.uk

Originally based on a talk by Dr Roslin Russell (CRUK CRI)

The purpose of this talk

Bioinformaticians often get asked about experimental design, as we have experience working with the data.

Good experimental design is absolutely crucial for NGS applications, because of the size of the datasets.

The purpose of this talk is to introduce the basic concepts of good NGS experimental design.

NGS Experimental Design

- **Importance**
- Specificity and power
- Accuracy, precision and bias
- Considerations for sample collection
- Considerations for sequencing
- Conclusions

Good Experimental Design

- Improves **quality** & **validity** of your science
- Saves **time** & **money**
- Obtains **meaningful** results
- Gets published in **higher impact journals**

Without a valid design, valid scientific conclusions cannot be drawn. Also particularly important for NGS because of the size of the datasets.

Good to think about early!

Ronald A. Fisher
(1890-1962)

*"To consult the statistician **after** an experiment is finished is often merely to ask them to conduct a **post mortem** examination. They can perhaps say what the experiment died of." (1938)*

Value of Planning

Rushing into experiments without thoughtful planning invites failure.

"Seventy percent of whether your experiment will work is determined before you touch the first test tube"

Tung-Tien Sun (2004).
Excessive trust in authorities and its influence on experimental design.
Nature Reviews Molecular Cell Biology

Still a serious issue in the field

- Almost 70% of all the human RNA-seq samples in GEO do not have biological replicates (Feng et al. 2012)

- More unreplicated RNA-seq data were published than replicated RNA-seq data in 2011 (Feng et al. 2012)

- ENCODE guidelines recommend two ChIP-seq biological replicates (Landt et al. 2012), but this is controversial and arguably too few

NGS Experimental Design

- Importance
- **Specificity and power**
- Accuracy, precision and bias
- Considerations for sample collection
- Considerations for sequencing
- Conclusions

Important concepts

- **Specificity:** the proportion of results that are genuine ones, as opposed to false positives

  e.g. At FDR 1%, 1 in 100 results will be a false positive.

- **Power (sensitivity):** the proportion of genuine results that are successfully identified by the experiment

  e.g. At 20% power, the experiment will successfully identify 200 out of 1000 genuine differentially expressed genes at a specified fold-change level.

Adequately Powered

*The power (sensitivity) of an experiment is the **probability** that it can detect an effect, if it is present.*

- Power is often **overlooked**.

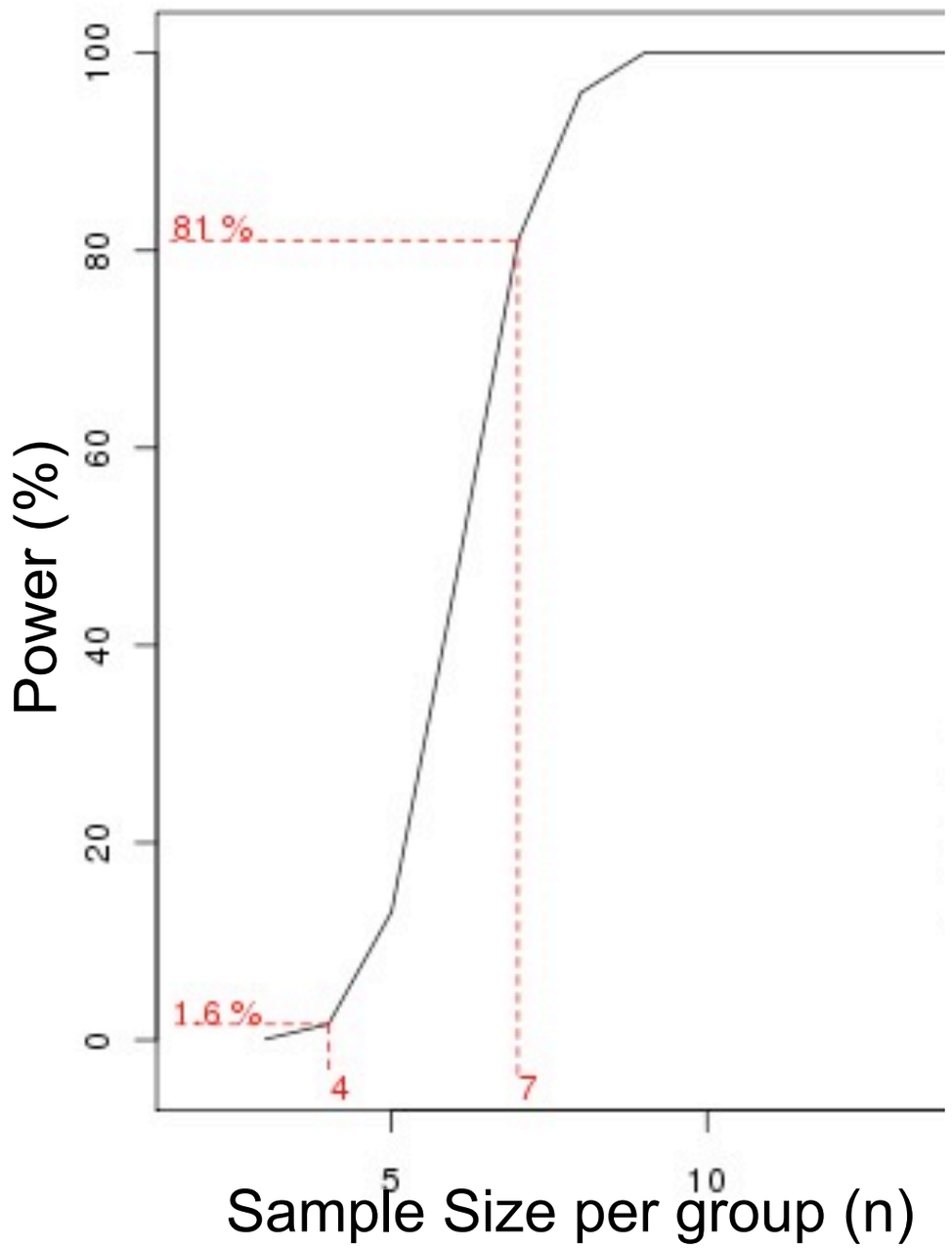
- A **probability** - any value between 0% and 100%.

- Achieved by:
  - Using **appropriate numbers** of animals / samples (sample size)
  - Controlling **sources of variation**

If you increase the **variability** when you increase the size then it won't necessarily have more power.

Adequately Powered

- **Too <u>many</u>** samples:
  - wastes resources e.g. animals (unethical), money, time and effort.
- **Too <u>few</u>** samples:
  - may <u>lack power</u> and miss a scientifically important effect.



How much power is enough?

- Depends on the experiment purpose.
  - To just get the **top few targets** = not much.
  - To get a **comprehensive list** of targets = a lot.
  - **Rare** transcripts / genomic variants = a lot.

- Also depends on variability of samples.
  - **In vitro** studies = less variability
  - **In vivo** = more variability, mixture of tissues
  - **Cancer** = hugely variable, need a lot of power

How do I tell?

- It's possible to calculate the required power of an experiment. For RNA-seq, there's even an online tool (and there are various general calculators):

Gene Expression

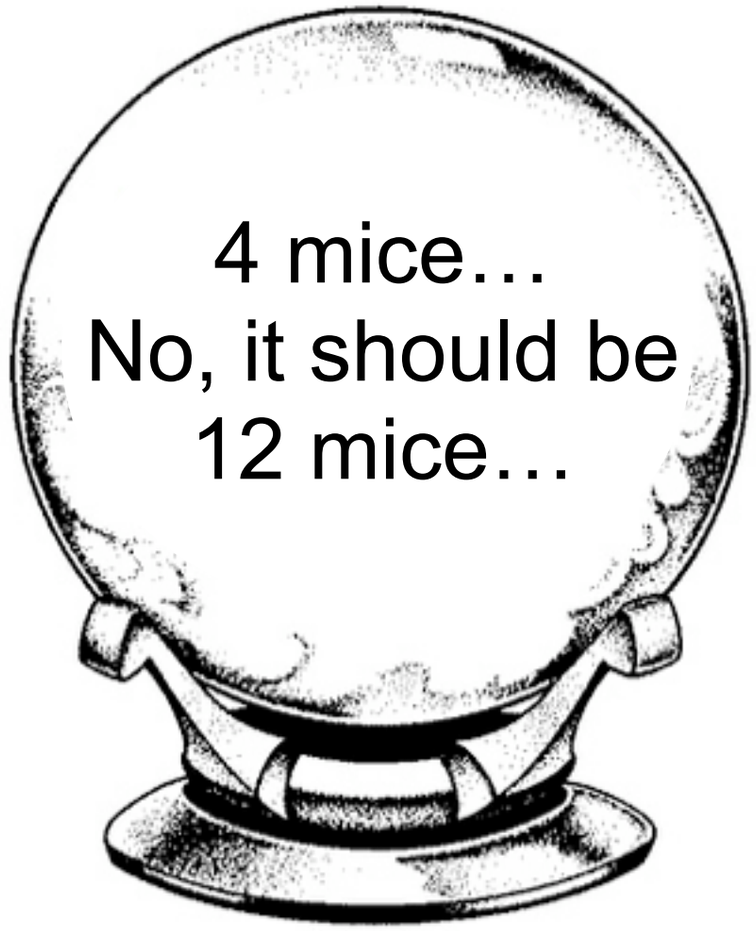
Scotty: A Web Tool For Designing RNA-Seq Experiments to Measure Differential Gene Expression

M.A. Busby, C. Stewart, C. Miller, K. Grzeda, G. Marth
Department of Biology, Boston College, 140 Commonwealth Avenue, Chestnut Hill MA, USA

- Alternatively, use rules of thumb based on other people's systematic reviews of a particular experiment type.

How many replicates?

Look into my crystal ball……



The sensitivity, ability or **power** to detect changes depends on the **sample size**.

How many replicates?

Depends on the resources, the goals of the study, and the reliability of the technology:

- How much **money** do you have?
- Can you **handle** all these samples without problem?
- What size of differences (**effect size**) to detect?
- It's an **accurate representation** of the population?
- Large enough to achieve **meaningful** results?

What is the minimum number of replicates?

Experiments with only one replicate **make puppies cry.**

- **Unless:**
  - It's a pilot experiment, e.g. to estimate data quality / antibody specificity / quantity of particular transcripts etc., and you plan to do follow-ups (further reps or other biological validation).
  - It's performed on a homogeneous and well-characterised system where a lot of other data already exists. e.g. RNA-seq of a much used cell line.
  - It's a simple confirmatory experiment

What is the minimum number of replicates?

- Rule of thumb minimum number of replicates for an exploratory study = **3**.

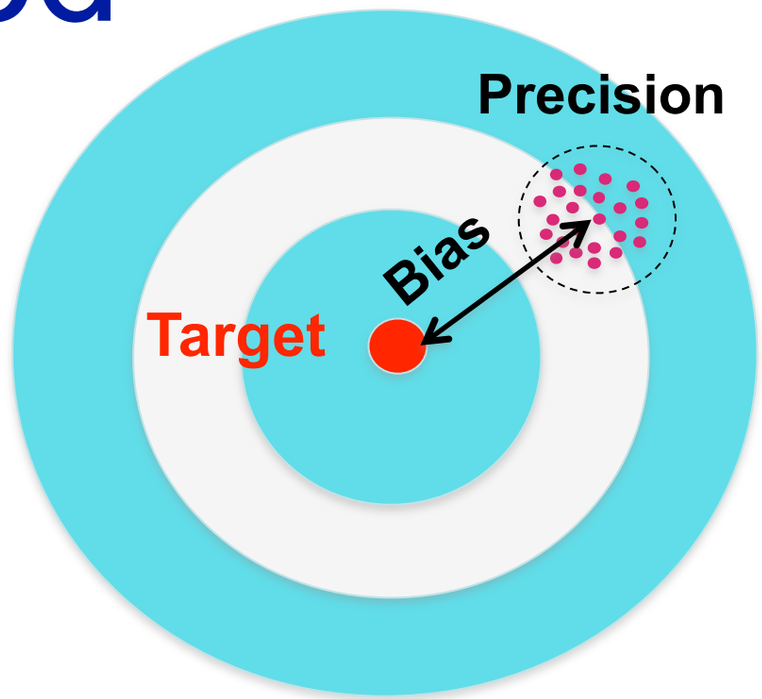
- The reasoning
  - 3 data points per gene lets you fit a statistical distribution more accurately, and gives a better estimate of variability
  - This in turn improves both the specificity and the sensitivity of the experiment.

Some labs do 4 replicates as standard, in case one fails.

NGS Experimental Design

- Importance
- Specificity and power
- **Accuracy, precision and bias**
- Considerations for sample collection
- Considerations for sequencing
- Conclusions

Precise & Unbiased



PRECISION:

- **Reproducibility** of repeated measurements.
- **Precise estimation** of the quantity of interest.
  - *Random variation* (**chance**) leads to results being **imprecise**.

UNBIASED:

- Bias can affect **accuracy**.
- Should control for **systematic differences** between the measure and some “true" value (target).
- **Doesn’t confound** that estimate with a technical effect.
- *Systematic variation* (**bias**) leads to results being **inaccurate**.

"A biased scientific result is no different from a useless one"

Daniel Sarewitz

(Beware the creeping cracks of bias. *Nature*, 2012)

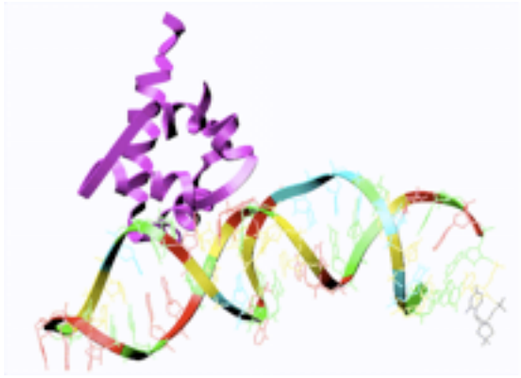
Controlling bias

- **Bias is avoided by**:
  - Correct selection of experimental units.
  - Randomisation of the experimental units.
  - Randomisation of the order in which measurements are made.
  - “Blinding” and the use of coded samples where appropriate.

- **Failure to randomise and blind** can lead to **false positive and negative results**

Confounding factors:

example



RNA Extraction

Plate1

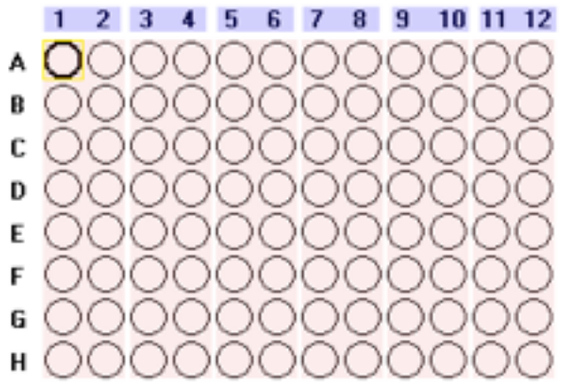


Plate2

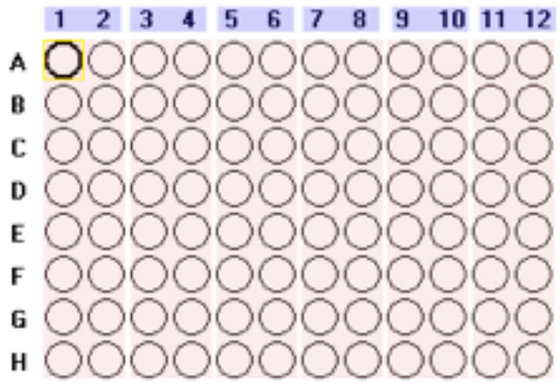
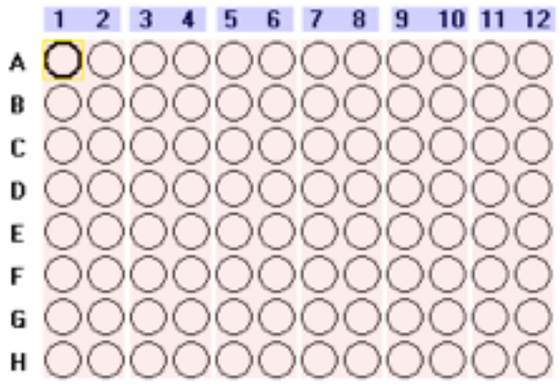


Plate3



Control

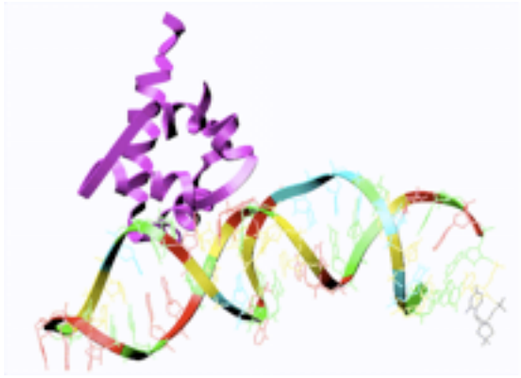
Treatment 1

Treatment 2

The difference between Control, Treatment 1 and Treatment 2 is confounded by **Plate**

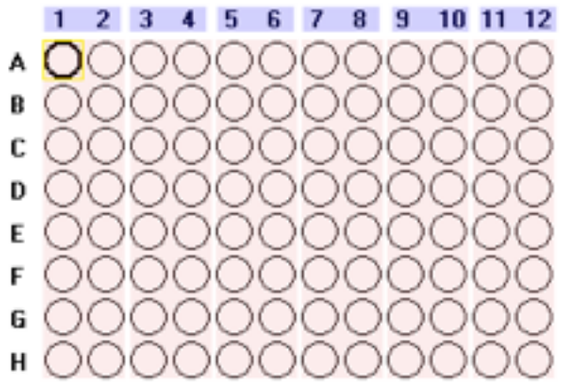
Confounding factors:

example 2

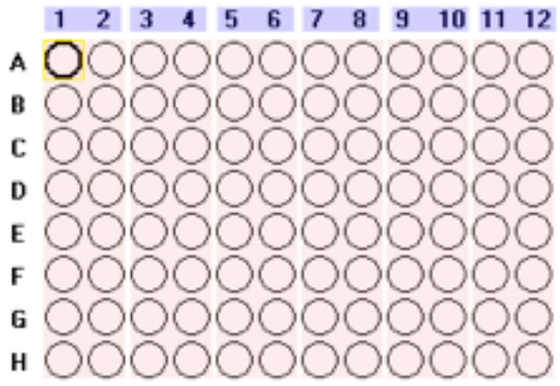


RNA Extraction

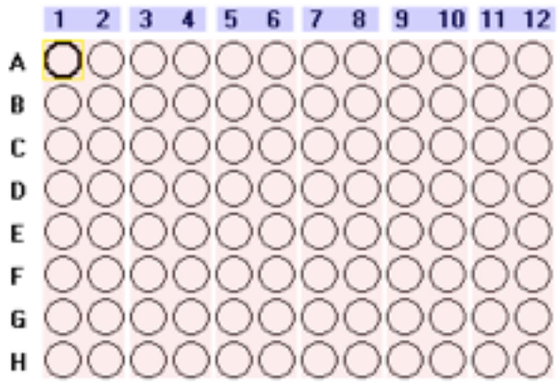
Day1, Plate 1



Day2, Plate 2



Day3, Plate 3



Control

Treatment 1

Treatment 2

The difference between Control, Treatment 1 and Treatment 2 is confounded by **day** and **plate.**

Sources of potential bias

Issues can arise if any of the experimental steps are applied in a systematically biased way between sample and control. e.g. During:

- Sample collection procedure
- Molecular biology during the main experiment (e.g. ChIP)
- Library prep
- Sequencing (different lanes on same flowcell is ok, but different flow cells can produce different results)

AN EXPRESSION OF CONCERN HAS BEEN PUBLISHED ON THIS PAPER; SEE LAST PAGE

Sciencexpress Report

Genetic Signatures of Exceptional Longevity in Humans

Paola Sebastiani,[1]* Nadia Solovieff,[1] Annibale Puca,[2] Stephen W. Hartley,[1] Efthymia Melista,[3] Stacy Andersen,[4] Daniel A. Dworkis,[3] Jemma B. Wilk,[5] Richard H. Myers,[5] Martin H. Steinberg,[6] Monty Montano,[3] Clinton T. Baldwin,[6,7] Thomas T. Perls[4]*

[1]Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA. [2]IRCCS Multimedica, Milano, Italy; Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Segrate, 20122, Italy. [3]Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA. [4]Section of Geriatrics, Department of Medicine, Boston University School of Medicine and Boston Medical Center, Boston, MA 02118, USA. [5]Department of Neurology, Boston University School of Medicine, Boston, MA 02118, USA. [6]Departments of Medicine and Pediatrics, Boston University School of Medicine and Boston Medical Center, Boston, MA 02118, USA. [7]Center for Human Genetics, Boston University School of Medicine, Boston, MA 02118, USA.

- A GWAS study of 800 centenarians against controls found 150 SNPs which can predict if a person is a centenarian with 77 % accuracy.
- Problem: they used **different SNP chips** for centenarian vs control.
- Retracted 2011 following an independent lab reviewed the data and QC applied.

http://www.the-scientist.com/blog/display/57558/

NGS Experimental Design

- Importance
- Specificity and power
- Accuracy, precision and bias
- **Considerations for sample collection**
- Considerations for sequencing
- Conclusions

Samples to collect

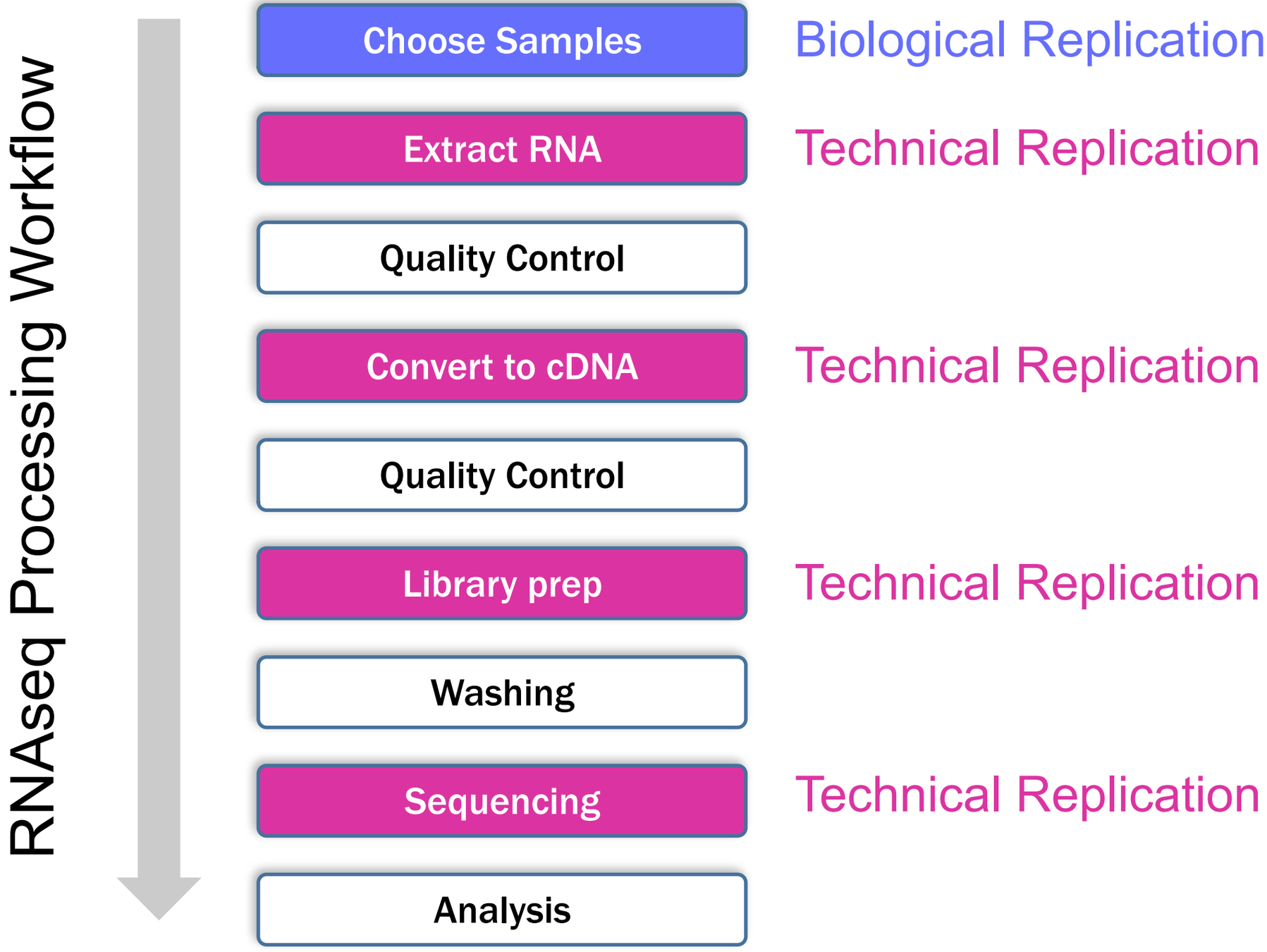
- Adequate <u>minimum</u> number of replicates for an exploratory study = **<u>3</u>**.

- Controls are also essential.

- Biological replicates should be performed - technical reps will not capture variation present in the tissues / organisms.

–> This means you need at least 3 biological replicate samples, and 3 biological replicate controls.

Biological or technical replicates?

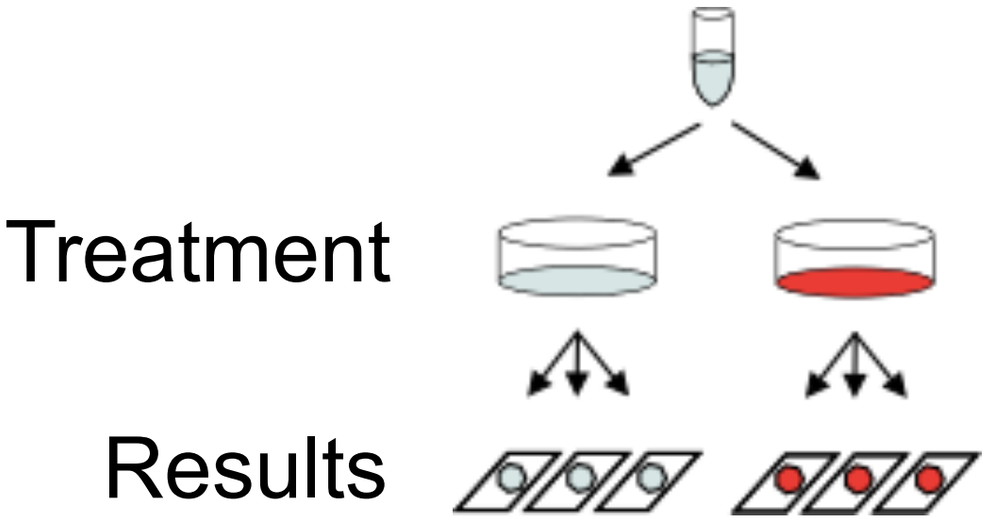


Biological replicate tricky cases

- Some biological replicate cases are obvious: e.g. tissue samples from different mice, blood samples from different people, cell cultures from different people.

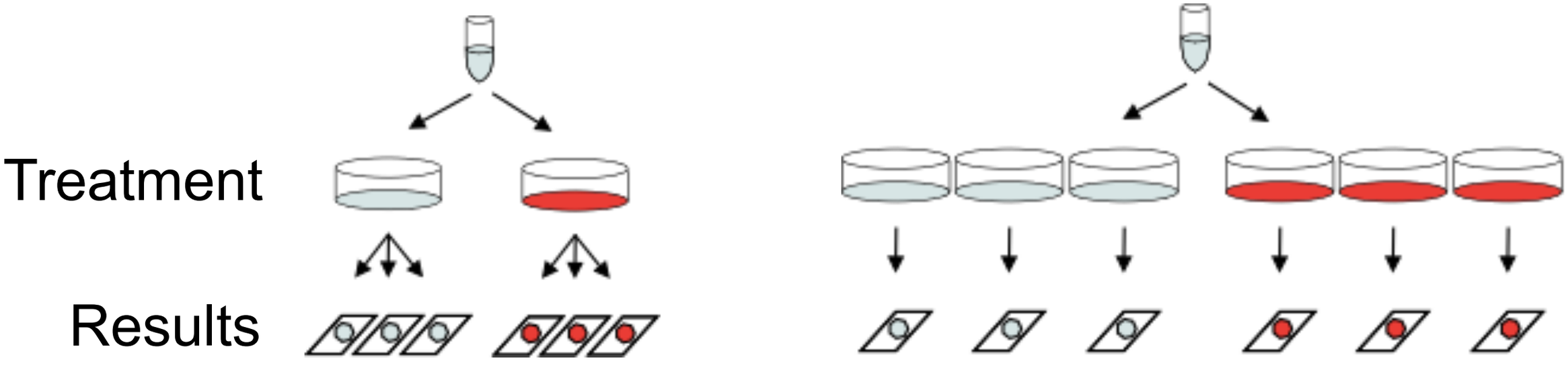
- Others are less clear cut:
    - e.g. Experiments using a single cell line

Cell line replicates



Bad example

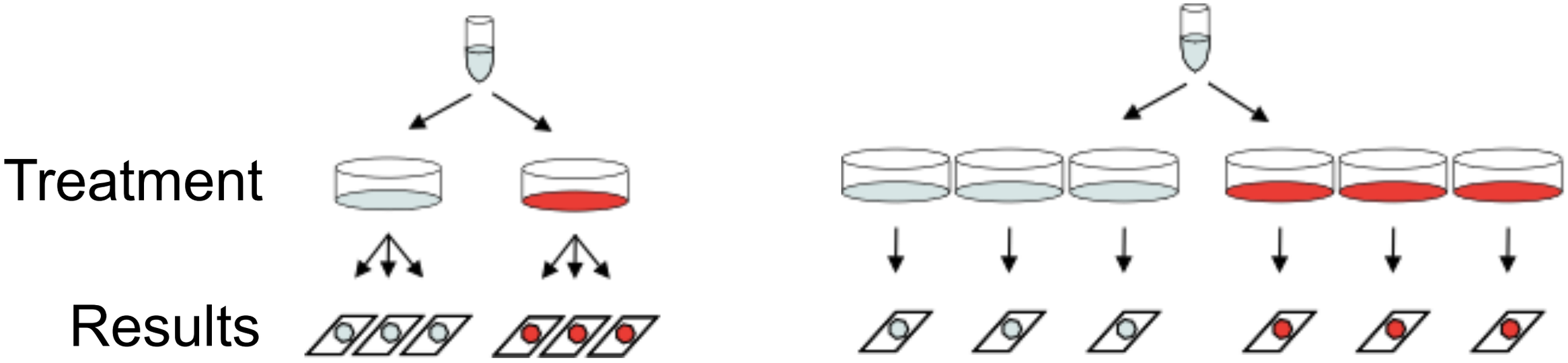
Cell line replicates



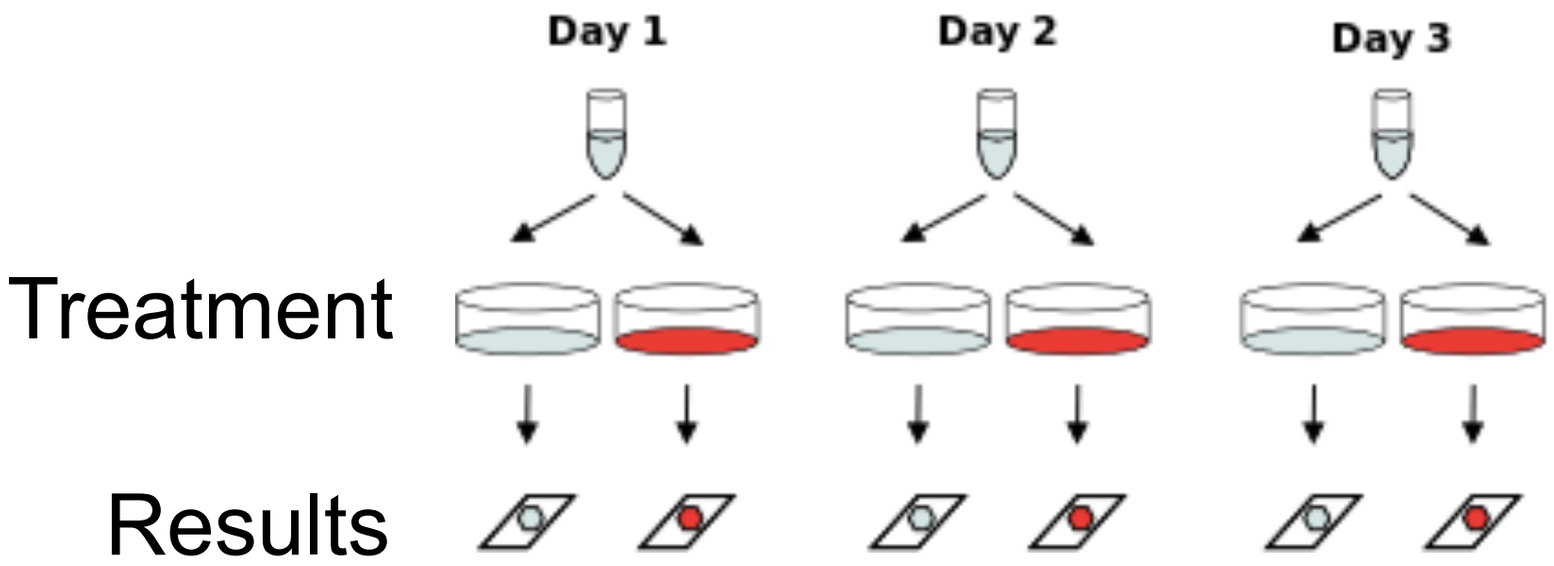
Bad example

Better

Cell line replicates

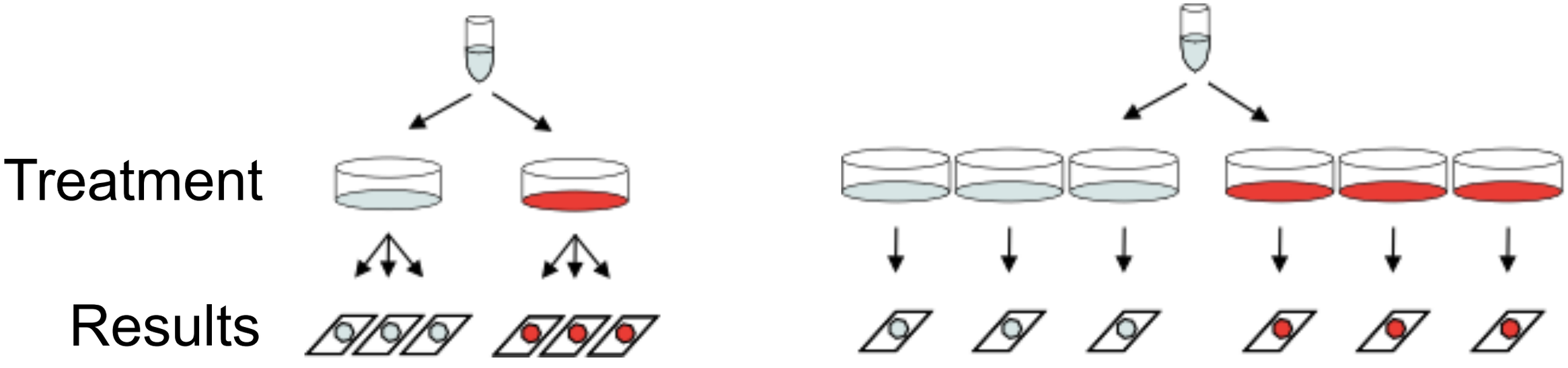


Bad example **Better**



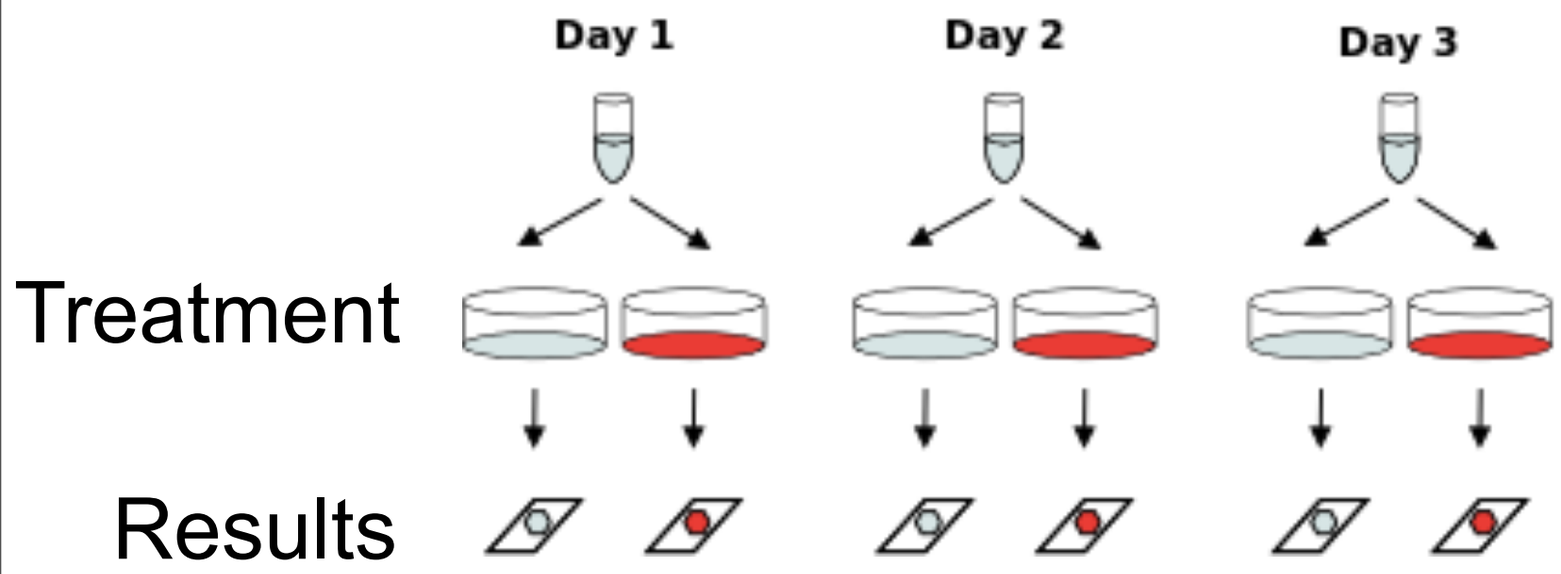
Even better

Cell line replicates



Bad example

Better



Even better

None of these are biological replicates - but we do the best we can.

Performing cell line replicates

- The aim is to assess the **replicability** of a given experiment, by performing it independently

- This means ideally they should be done on different days, with fresh medium and reagents etc. However, this might be impractical.

- It is at least important to perform the treatments independently. e.g. To get 4 replicates, do 4 separate transfections (or RNAi etc) for the experimental sample, and 4 for the control.

Experiment Controls

- A very important part in your experiment, because it's very difficult to eliminate all of the possible confounding variables & bias.
- Designing the experiment with controls in mind is absolutely crucial.
- Increases the statistical validity of your data.
- Two types: **positive** and **negative**.

Experimental Controls

- **Placebo control**

Mimic a procedure or treatment without the actual use of the procedure or test substance.

  - e.g. same surgical procedure but without X implanted

- **Vehicle control**

Used in studies where a substance is used to deliver an experimental compound.

e.g. apply EtOH to cell lines on it's own as a control since it's used as a vehicle for delivering the Tamoxifen drug.

- **Baseline biological state**

e.g. Wild type to observe the effects of a knockout.

ChIP-seq Controls

- **Particularly important**
  - ChIP reactions are extremely noisy and variable.

- **What controls to use**
  - Either input DNA or IgG antibody are commonly used. Both have pros and cons.

- **Replicates**
  - The control should have as many replicates as the sample, as it is also variable.

Should I pool samples?

*When working with very small amounts of tissue sample, pooling is a good way of **reducing the noise** while keeping the number of n reasonably small.*

e.g. where n = no. of sequencing reactions.

- Better to pool the **biological material** (tissue, cells), not the purified RNA or labeled cDNA. In this way, problems are far easier to spot.
- However, there's **no** way to estimate **variation between individuals** in a pool, which is sometimes important and often interesting.
- **Beware of outliers -** Don't include any sample that looks suspicious.
    - In some studies (pooled vs unpooled designs) the majority of DEGs turn out extreme in only one individual.

NGS Experimental Design

- Importance
- Specificity and power
- Accuracy, precision and bias
- Considerations for sample collection
- **Considerations for sequencing**
- Conclusions

Basic NGS parameters

- **Library size / sequencing depth:** the number of reads obtained for a given sample (e.g. 20 million)

- **Read length:** bp length of each read in the library (e.g. 150 bp)

- **Single end vs paired end:** single end sequencing only sequences one end of each DNA fragment, while paired end sequences both ends.

How deeply should you sequence

- **Depends on the application:** e.g. some parts of the genome will be more tricky to sequence, so full coverage for genome assembly may require very deep sequencing.

- **RNA-seq differential expression:** very roughly (for human and mouse) - at least 10 million reads per sample is required, and 30 million reads per sample should suit most applications. Can sequence deeper if low expressed genes are particularly important.

- **If in doubt:** you can run a pilot experiment and see how much coverage and saturation you get, and potentially sequence further after that.

Sequencing depth vs replicates

Replicates are good! Even if you don't sequence any more deeply.

A study looking at RNAseq (Liu et al. 2013) found that adding more replicates at 10 million reads library size was a more cost-efficient way of improving the power of the experiment, than adding more sequencing coverage. This was true for 3-7 replicates.

So, if you don't have money for a lot of sequencing, you can still multiplex a number of biological replicates!

What read length?

- **The read length should match your sample fragment length:** you want it to be around the size of and just a bit shorter than your fragments.

- **Recommended for Illumina:**

  sRNAseq, ribosomal profiling = 50 bp read length

  mRNAseq (fragment size ~80-200bp) = 150 bp read length

  longer fragments = 250 bp

- **For longer reads / niche applications:** other sequencers also exist. PacBio has an impressive read length (often >10 kb)

Single end vs paired end

- **Single end:** a bit cheaper, less sequencing required. Suitable for most general purposes, such as differential expression analysis.

- **Paired end:** contains more length and positional information about the sequenced fragment - can tell where it starts and stops.

- **Applications for which you need paired end sequencing:**
  - splice junctions
  - rearrangements: indels and inversions

NGS Experimental Design

- Importance
- Specificity and power
- Accuracy, precision and bias
- Considerations for sample collection
- Considerations for sequencing
- **Conclusions**

Summary

- **Experimental design** is key to a successful genomics experiment.
- The required **power** of an experiment should be considered when choosing the number of replicates.
- An adequate number of **biological replicates** (usually >=**3**) is crucial for a reliable statistical analysis
- Appropriate **controls** are essential
- **Systematic bias** can completely mess up an experiment and should be controlled