RNA–SEQ: DESIGNING AN EXPERIMENT

Phu Tran

OUTLINE

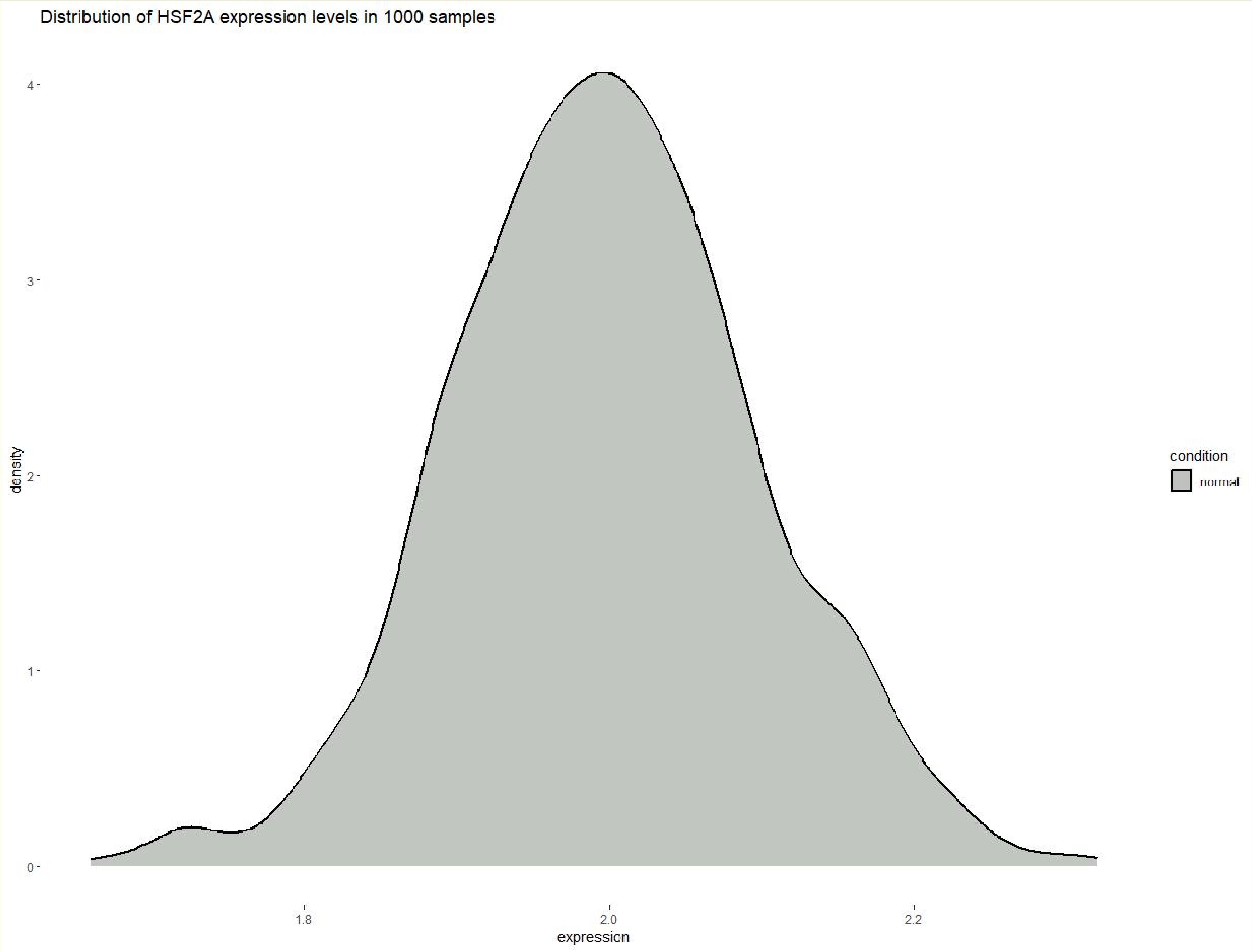
1 How many samples do we need

2 Best practices for experimental design

3 Additional points

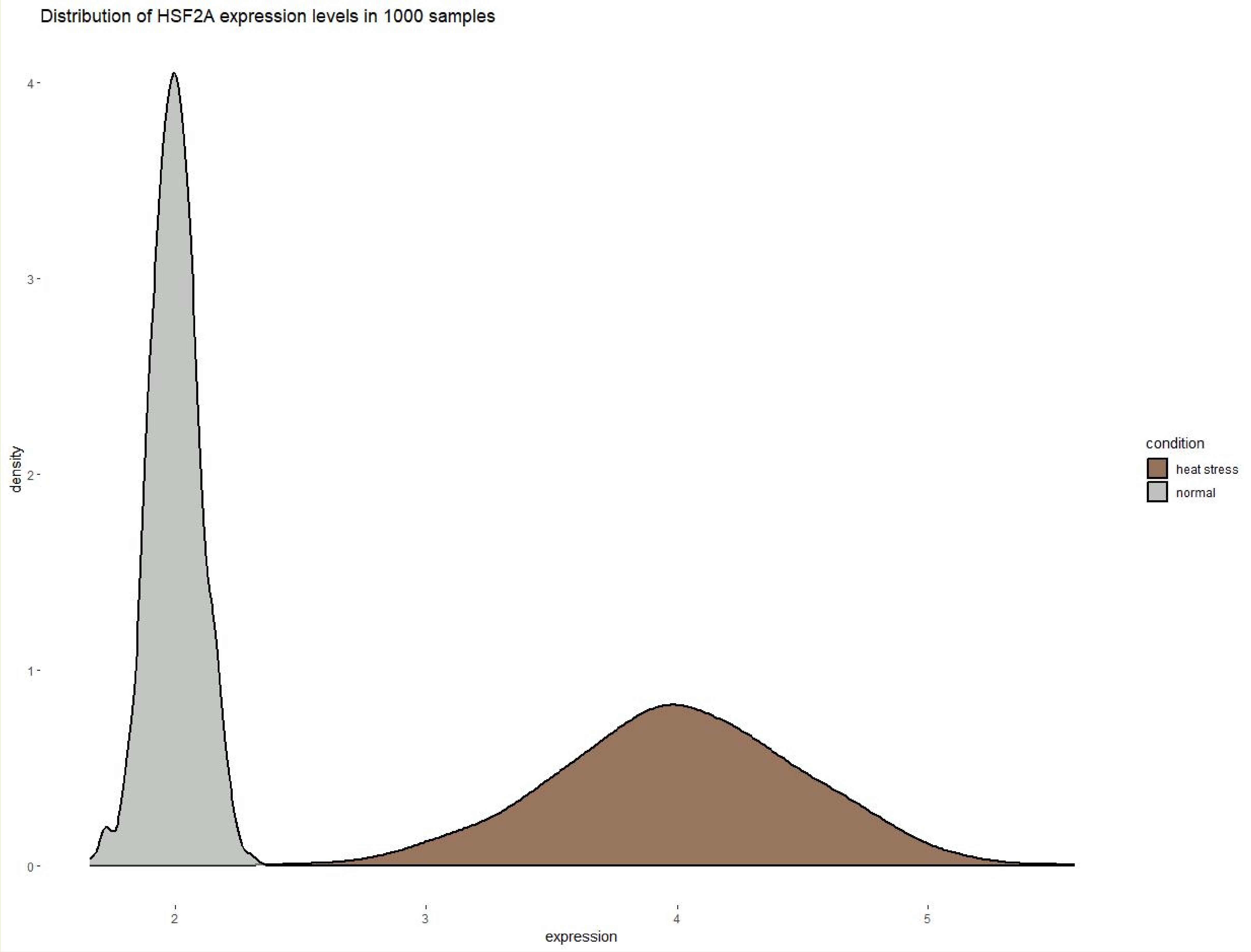
1. HOW MANY SAMPLES DO WE NEED

Heat stress transcription factor A–2” (HSFA2)



Distribution of HSF2A expression levels in 1000 samples

Heat stress transcription factor A-2” (HSFA2)



Welch Two Sample t-test
t = -126.96, df = 1081.4, p-value <2.2e-16

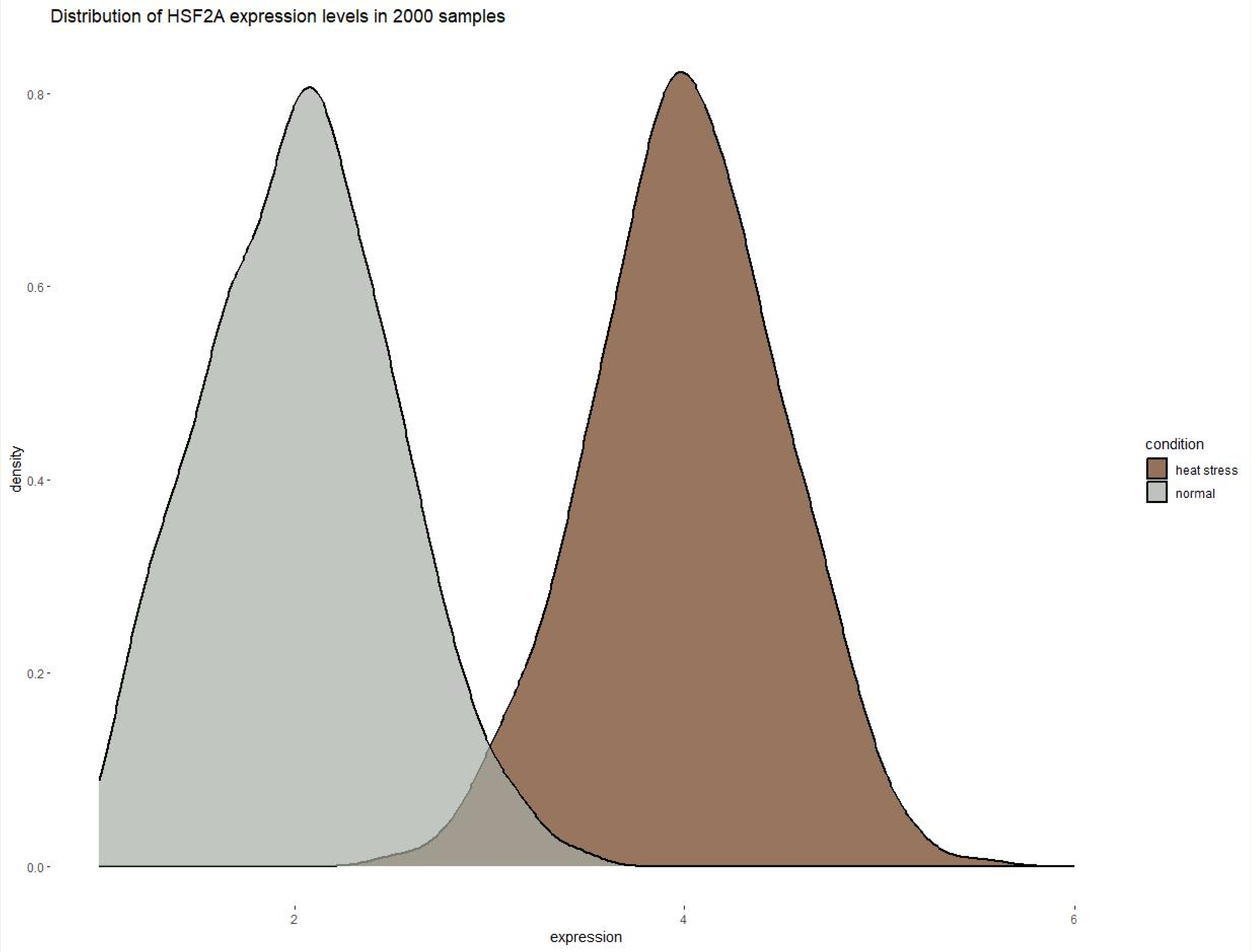
Alternative hypothesis:
true difference in means is not equal to 0

95 percent confidence interval:
-2.040977 -1.978850

sample estimates:

```
  mean of x          mean of y
  1.997340           4.007254
```

Heat stress transcription factor A-2” (HSFA2)



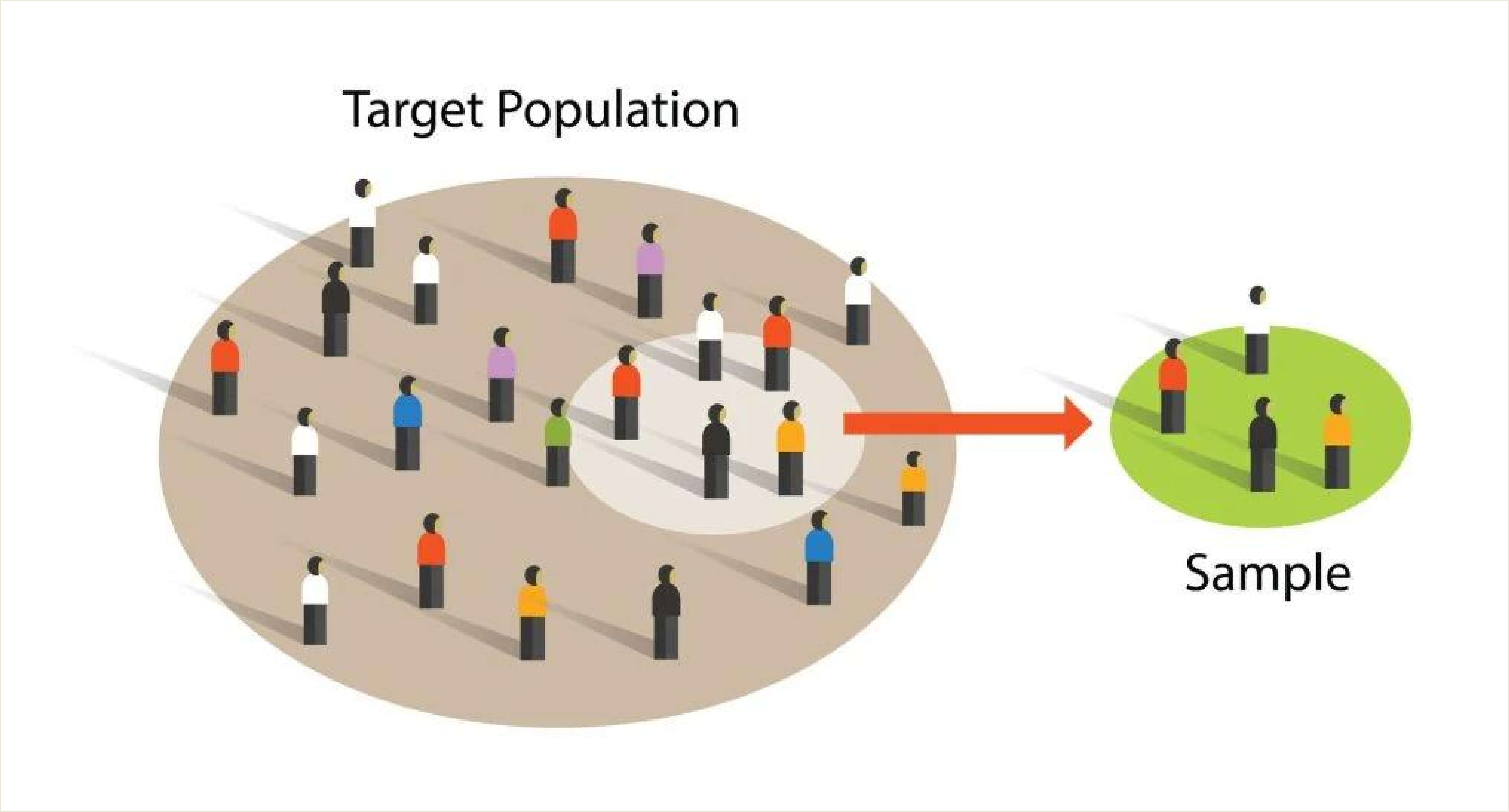
Welch Two Sample t-test

t = -89.398, df = 1996.1, p-value < 2.2e-16
alternative hypothesis:
true difference in means is not equal to 0

95 percent confidence interval:
-2.036495 -1.949062

sample estimates:

mean of x	mean of y
2.014476	4.007254

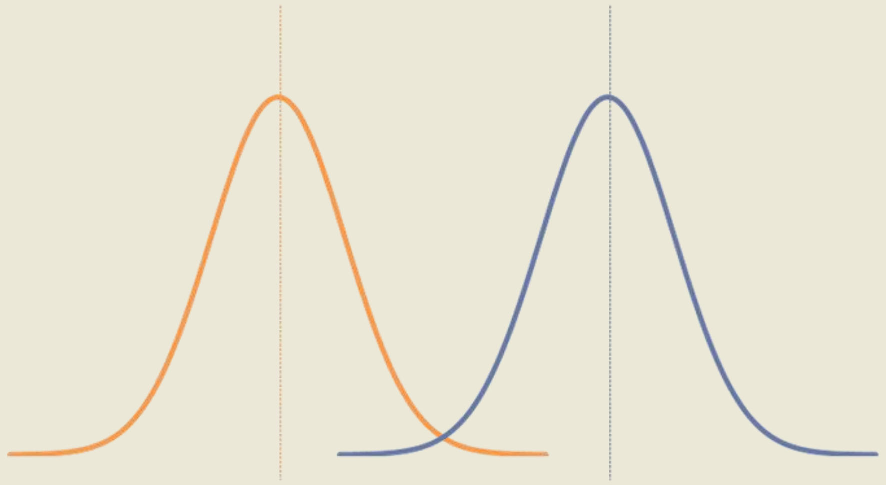

Target Population
Sample

Effect size

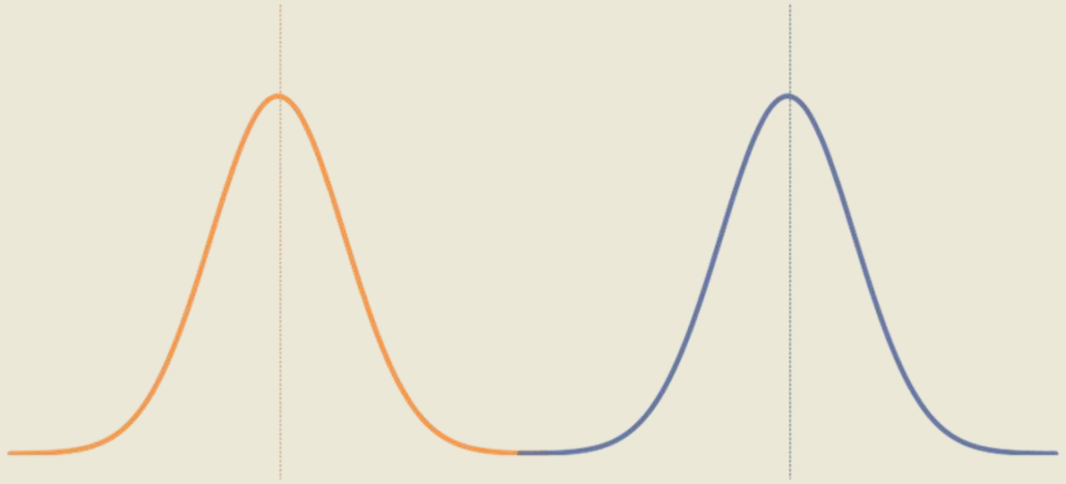
Type I error

Type II error

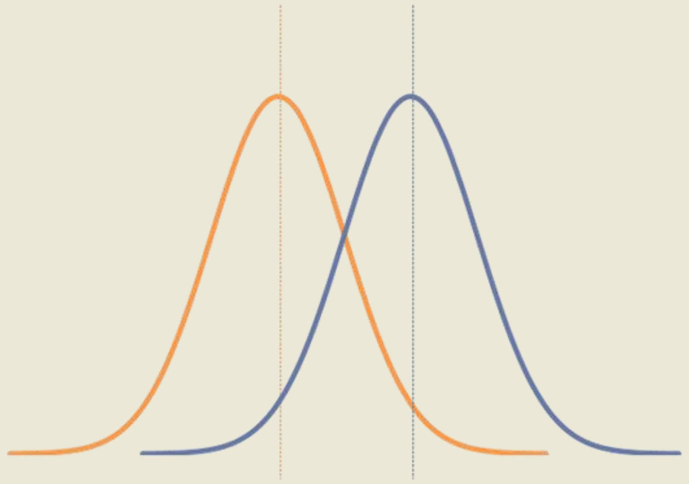
Effect size



n1=?

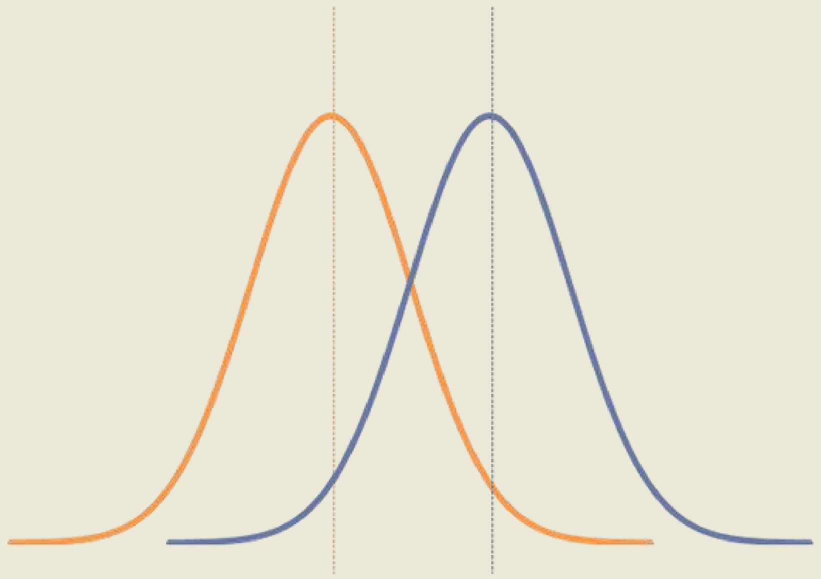


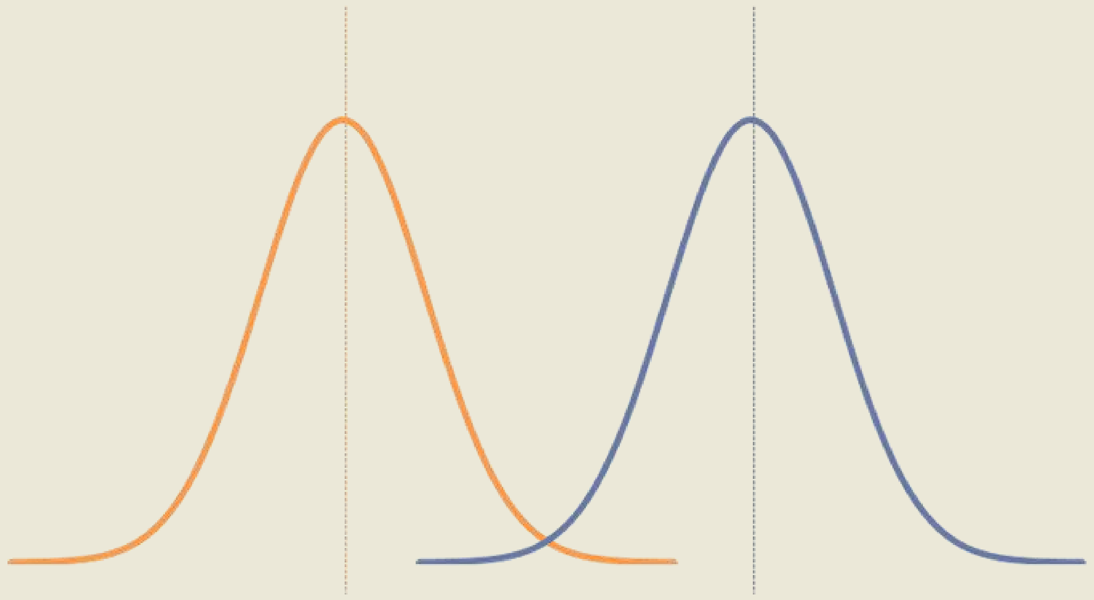
n2=?



n3=?

Error type 1 (alpha)





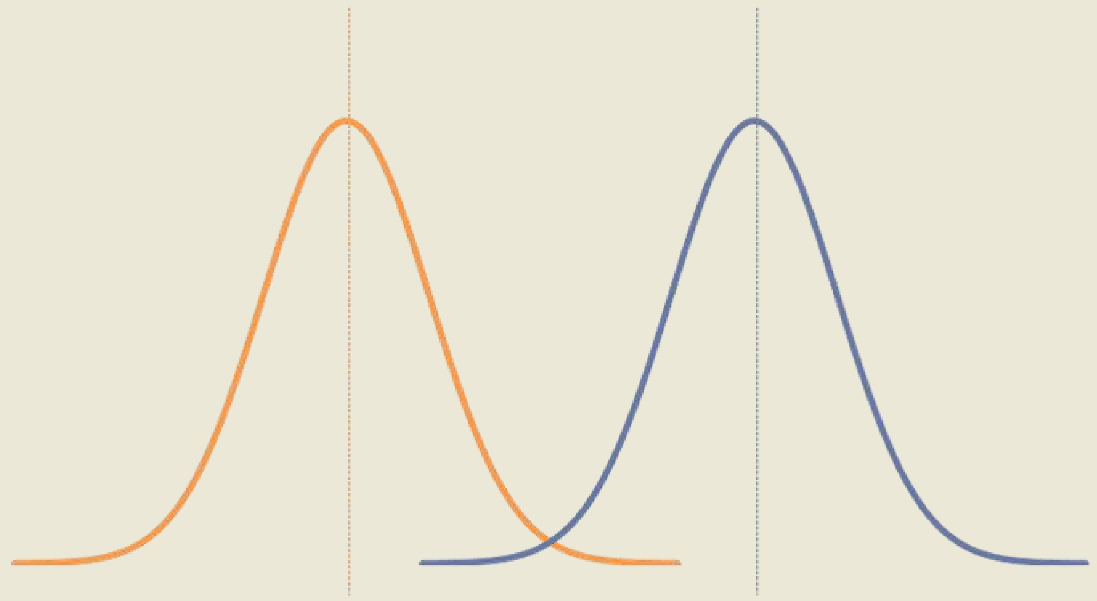
Truth in population

Insignificant

Truth in sample

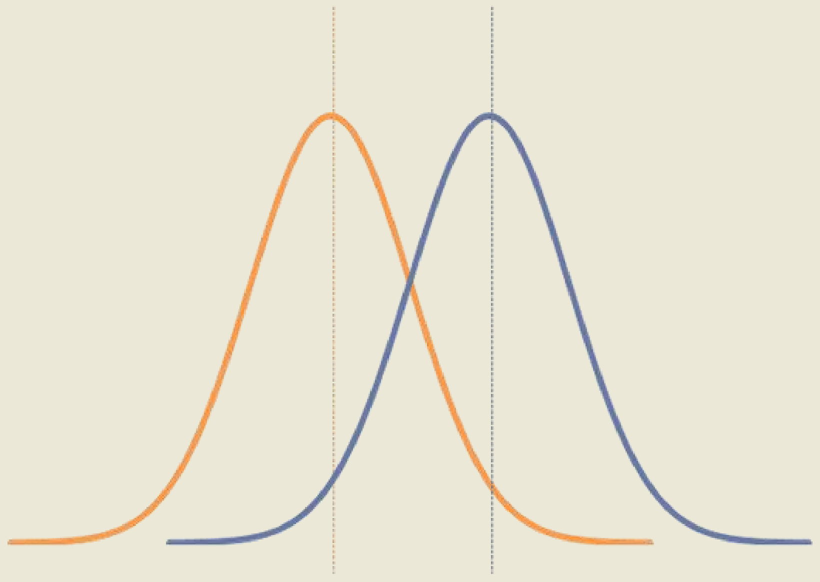
Significant

Error type 2 (beta)



Truth in population

Significant



Truth in sample

Insignificant

Parameters needed to calculate sample size:

Type I error: α-value.
Often set to 0.01 (1%) or 0.001 (0.1%) in RNA-seq experiments.

Type II error: β-value. $(1-\beta)$ (the power of your analysis).
Should be set to 70 or 80% to detect 70 or 80% of the differentially expressed genes.
The number of biological replicates might be hard to reach in practice for RNA-seq experiments.

Effect size: this is a parameter you will set.
For instance, if you want to investigate genes that differ between treatments with a difference of their mean of 2 then the effect size is equal to 2.

Sample size: the quantity you want to calculate.

Are there any significant different expressions of the genes between two conditions?

If the RNA seq can detect the expression of ~10.000 gene

We are conducting t-test 10.000 times

At alpha = 0.05

There could be 500 significantly different genes due to chance alone (false positive)

We need to set a lower alpha

is there any significant different in the expression of gene A ?

```
      pwr.t.test(d = 1,
            power = .8,
        sig.level = .05,
type = ‘two.sample‘,
       alternative =
          ‘two.sided‘)
```

```
Two–sample t test power calculation

          n = 16.71472
          d = 1
  sig.level = 0.05
      power = 0.8
alternative = two.sided

NOTE: n is number in *each* group
```

Effect size

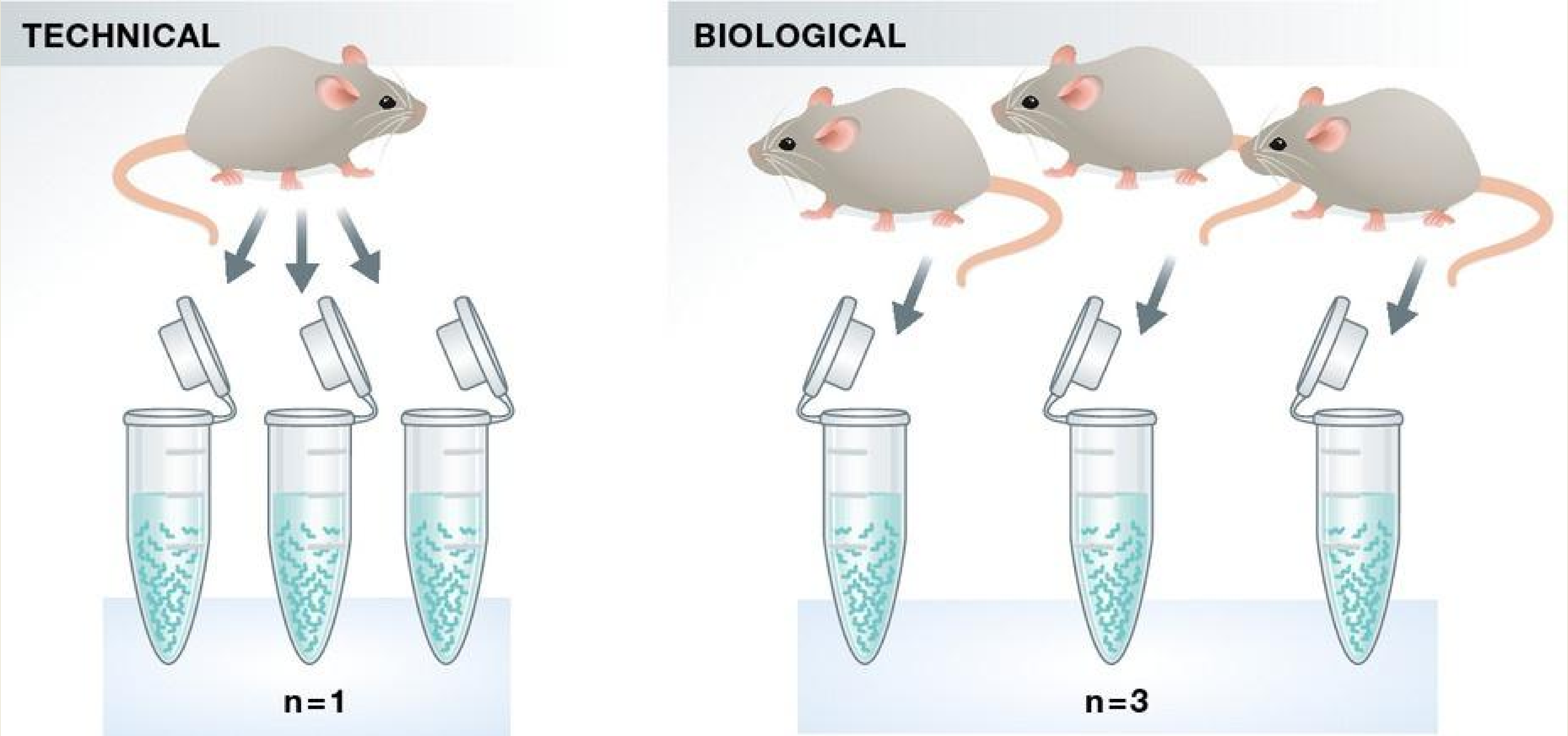
Type I error

Type II error

POWER OF AN ANALYSIS

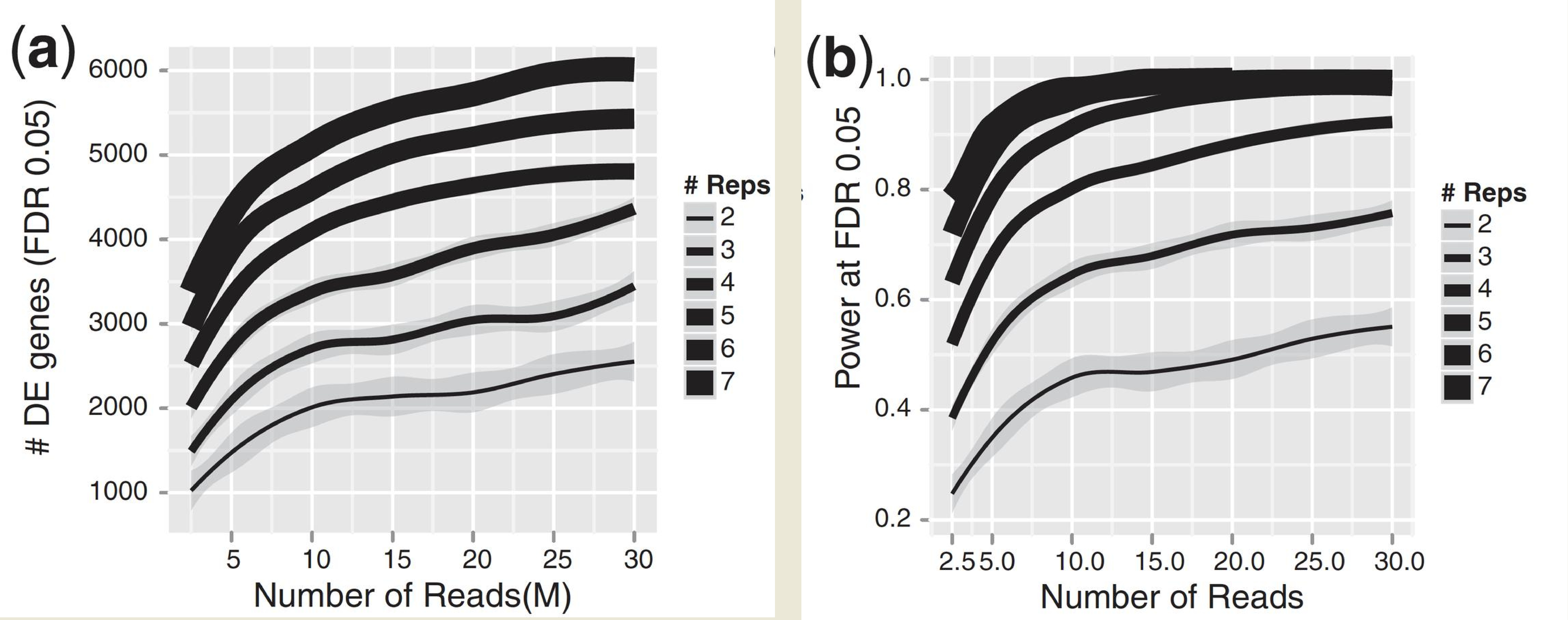
UNDERPOWER

REPLICATES



Klaus B., EMBO J (2015) 34: 2727–

REPLICATES



Liu, Y., et al., Bioinformatics (2014) 30(3): 301–304

REPLICATES

General gene-level differential expression:

- ENCODE guidelines suggest 30 million SE reads per sample (stranded).
- 15 million reads per sample is often sufficient, if there are a good number of replicates (|3).
- Spend money on more biological replicates, if possible.
- Generally recommended to have read length |= 50 bp

Gene-level differential expression with detection of lowly-expressed genes:

- Similarly benefits from replicates more than sequencing depth.
- Sequence deeper with at least 30-60 million reads depending on level of expression (start with 30 million with a good number of replicates).

REPLICATES

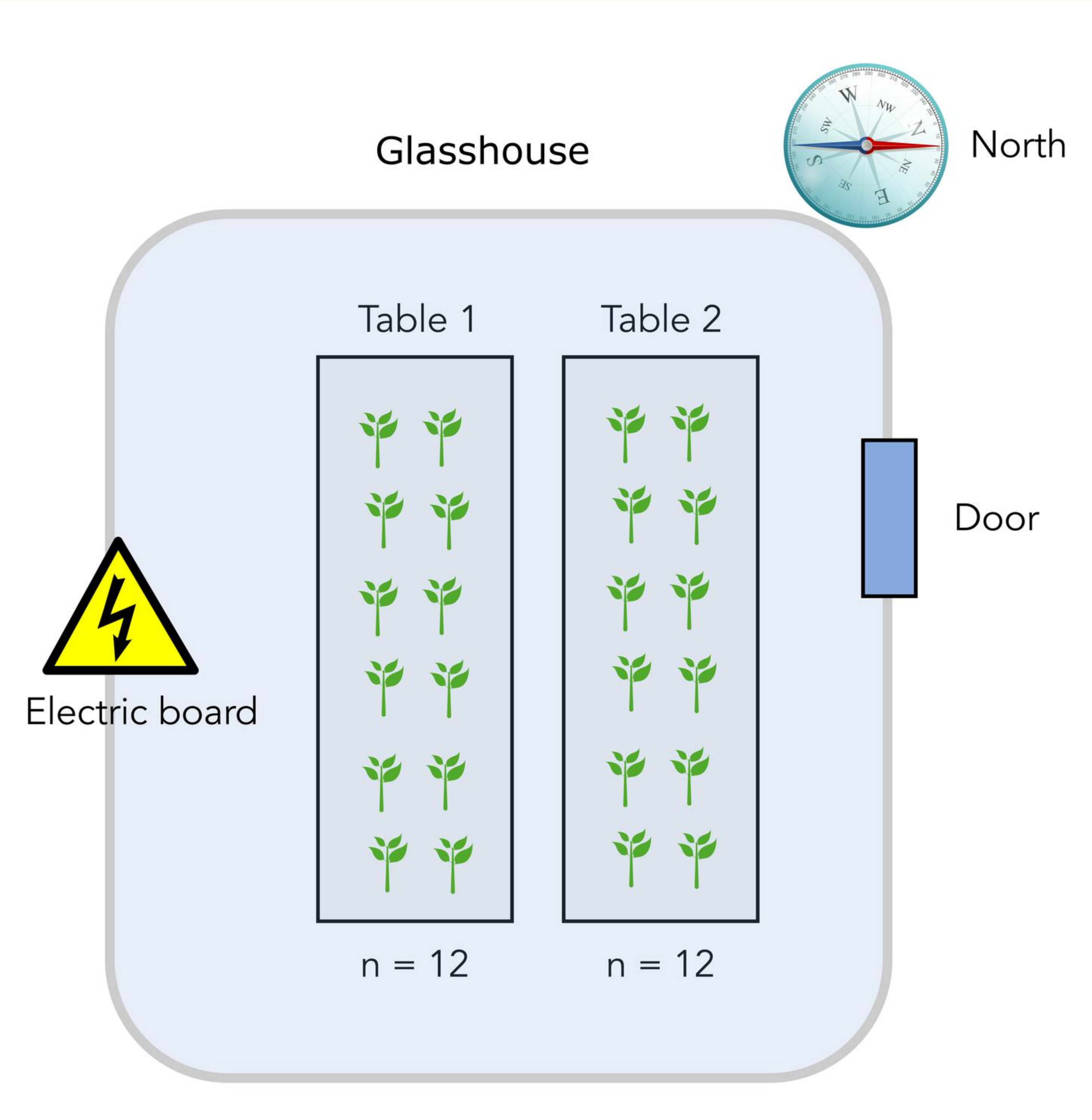
Isoform–level differential expression:

- Of known isoforms, suggested to have a depth of at least 30 million reads per sample and paired–end reads.
- Of novel isoforms should have more depth (ı 60 million reads per sample).
- Choose biological replicates over paired/deeper sequencing.
- Generally recommended to have read length ı= 50 bp, but longer is better as the reads will be more likely to cross exon junctions
- Perform careful QC of RNA quality. Be careful to use high quality preparation methods and restrict analysis to high quality RIN # samples.

Other types of RNA analyses (intron retention, small RNA–Seq, etc.):

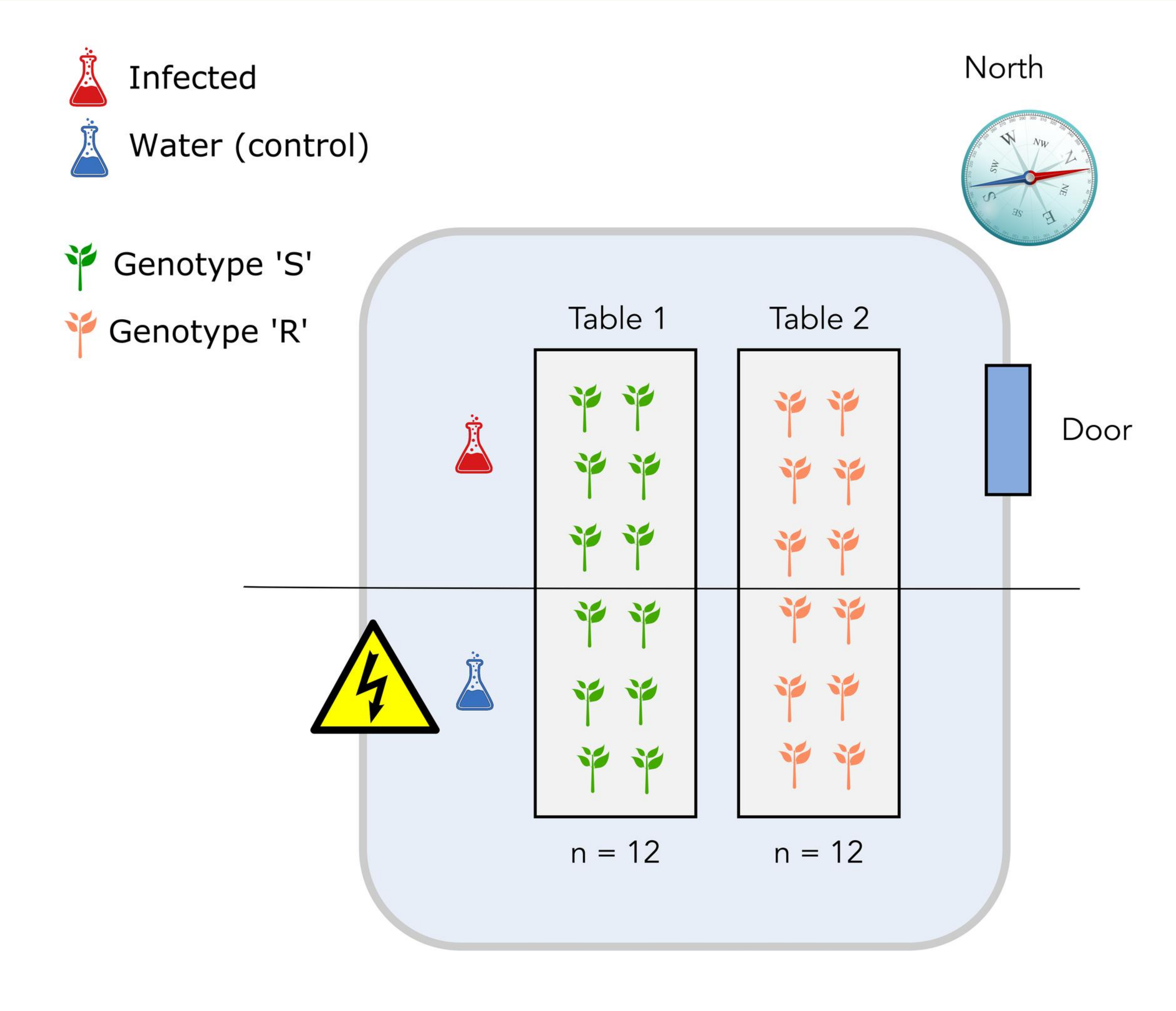
- Different recommendations depending on the analysis.
- Almost always more biological replicates are better!

2. BEST PRACTICES FOR EXPERIMENTAL DESIGN



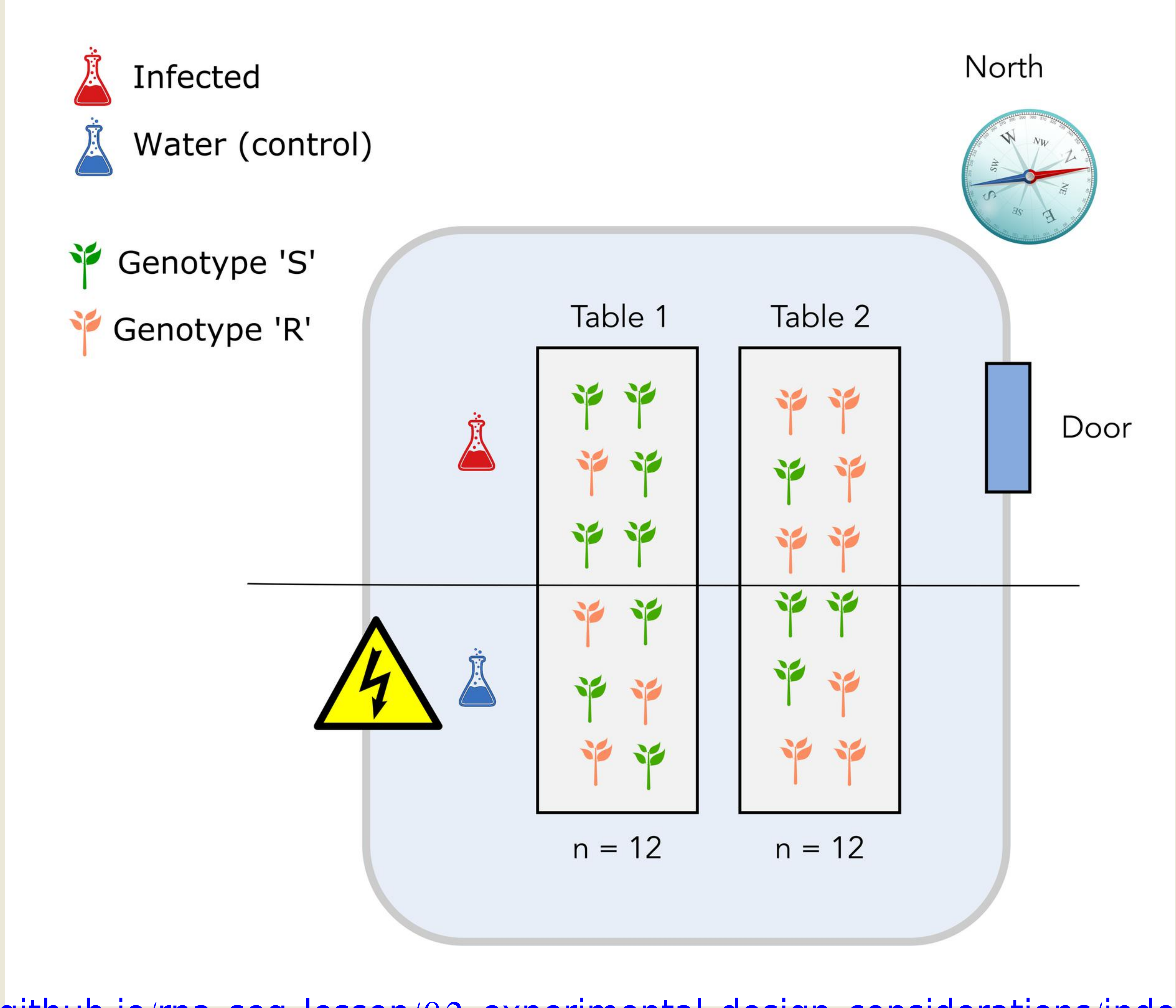
https://scienceparkstudygroup.github.io/rna-seq-lesson/02-experimental-design-considerations/index.html

DESIGN NUMBER 1



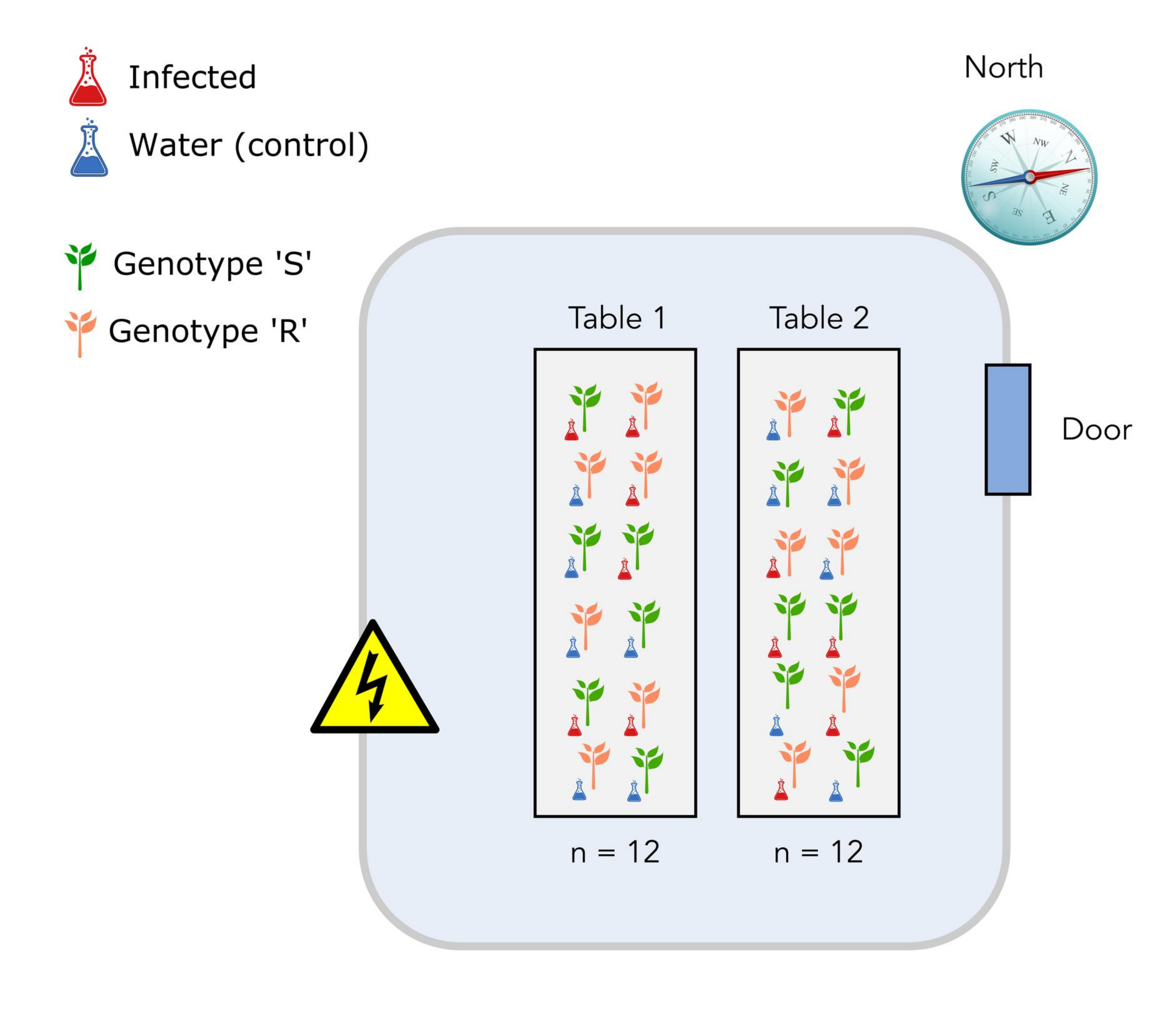
https://scienceparkstudygroup.github.io/rna-seq-lesson/02-experimental-design-considerations/index.html

DESIGN NUMBER 2



https://scienceparkstudygroup.github.io/rna-seq-lesson/02-experimental-design-considerations/index.html

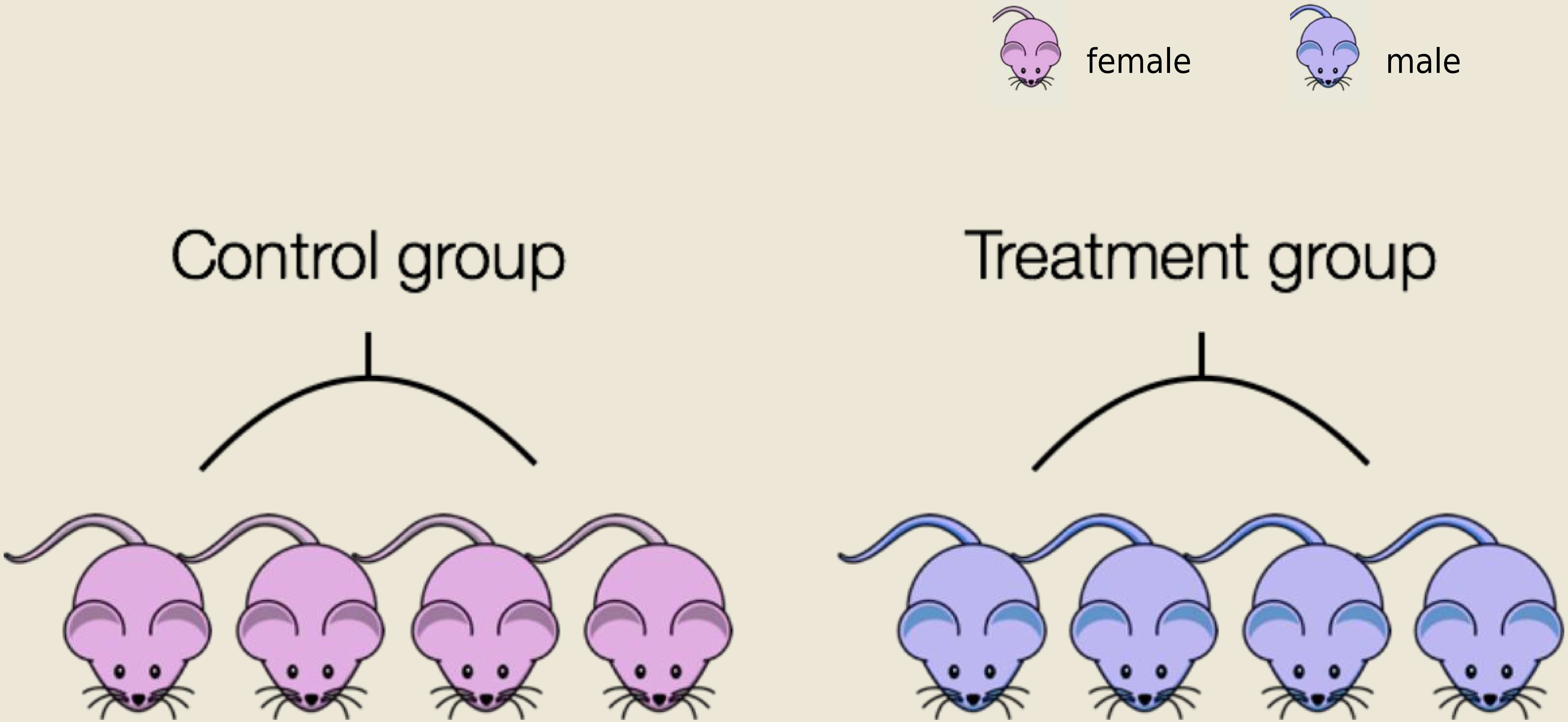
DESIGN NUMBER 3



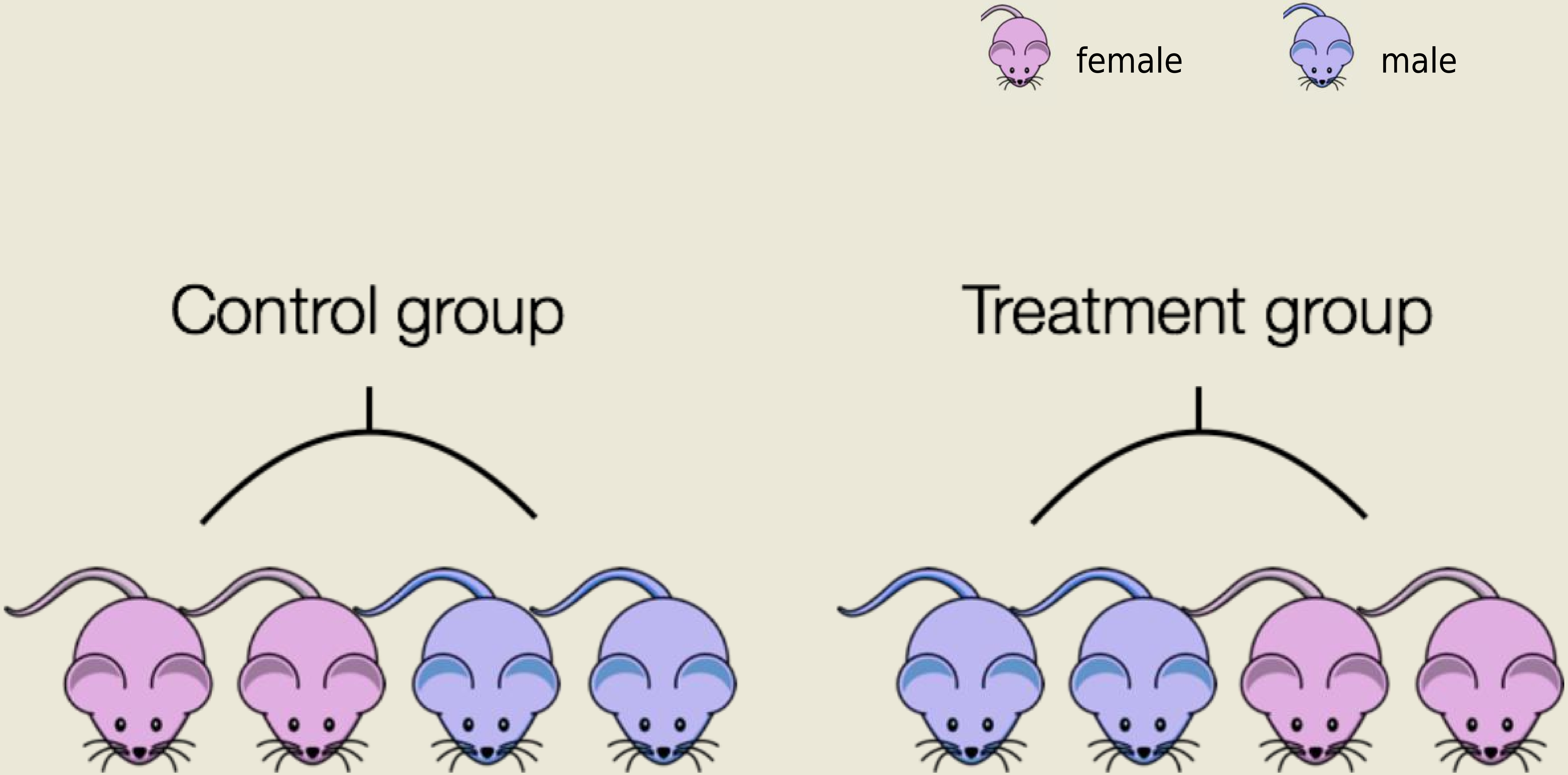
In a typical biological experiment, you will encounter various sources of variation that are either:

- desirable because they are part of your experimental factors. These are typically the one you first think of when you design your experiment.

- undesirable (unwanted) because you are not interested in them. Although you might not list them, they will still affect the outcome of your experiment.



https://scienceparkstudygroup.github.io/rna-seq-lesson/02-experimental-design-considerations/index.html



female

male

Control group

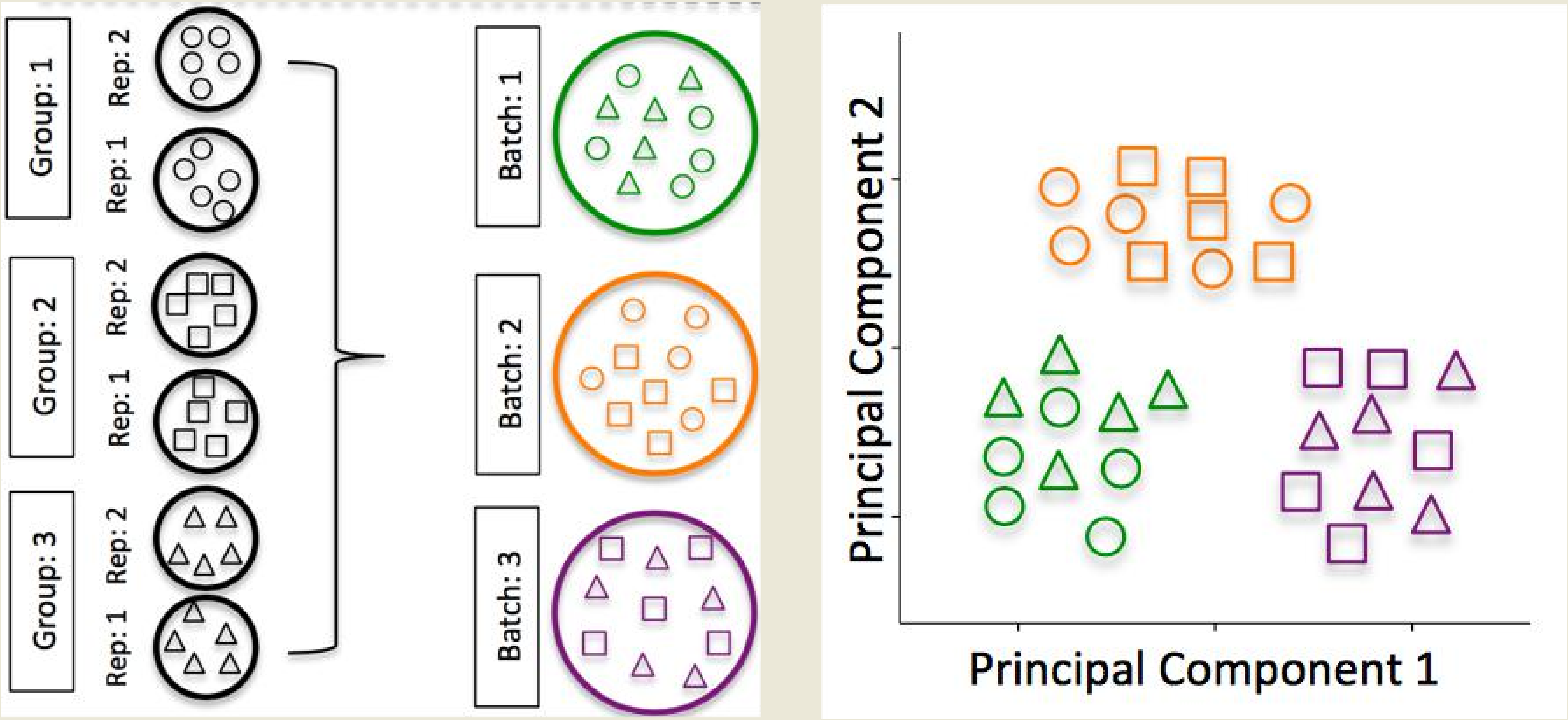
Treatment group

To AVOID confounding:

Ensure animals in each condition are all the same sex, age, litter, and batch, if possible.

If not possible, then ensure to split the animals equally between conditions

BATCH EFFFECT

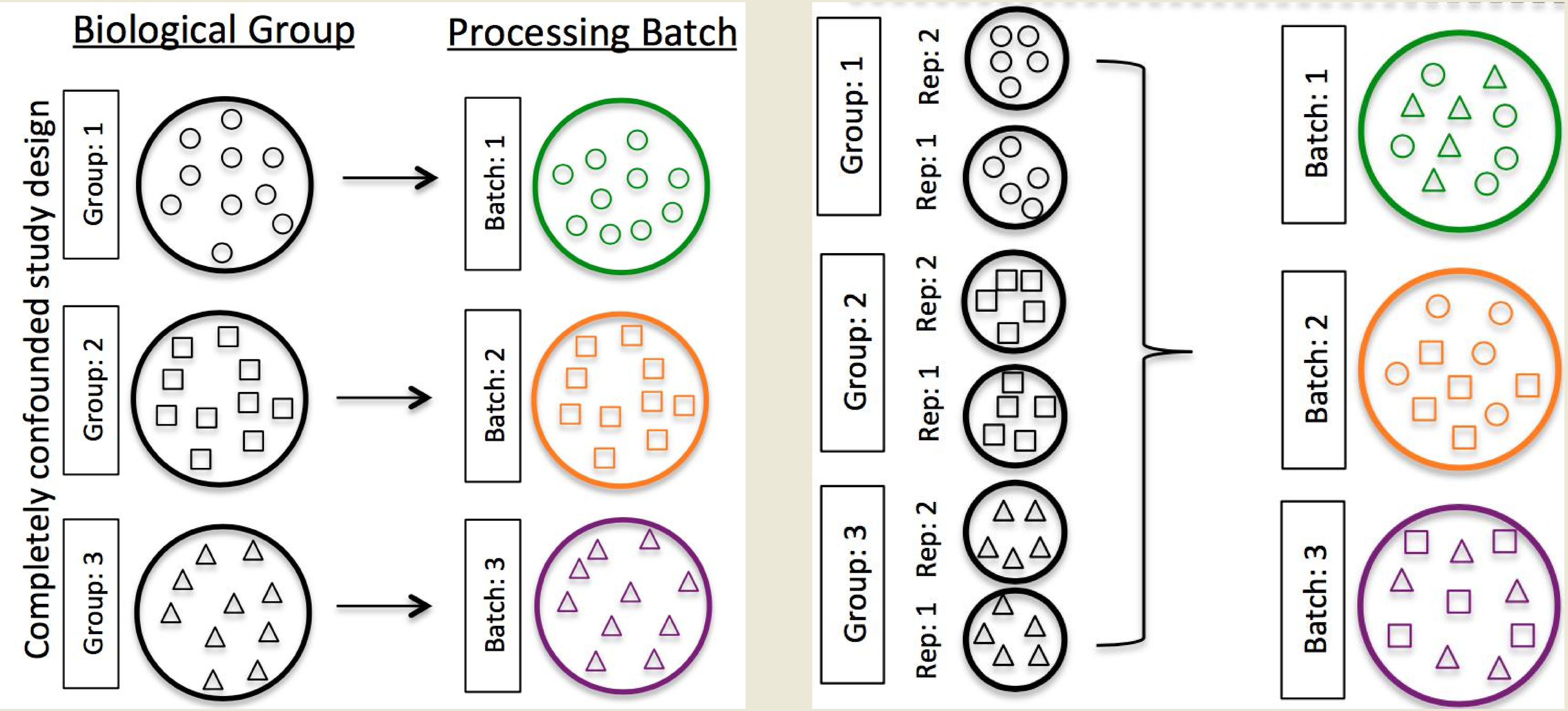


Hicks SC, et al., bioRxiv
(2015)

How 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.



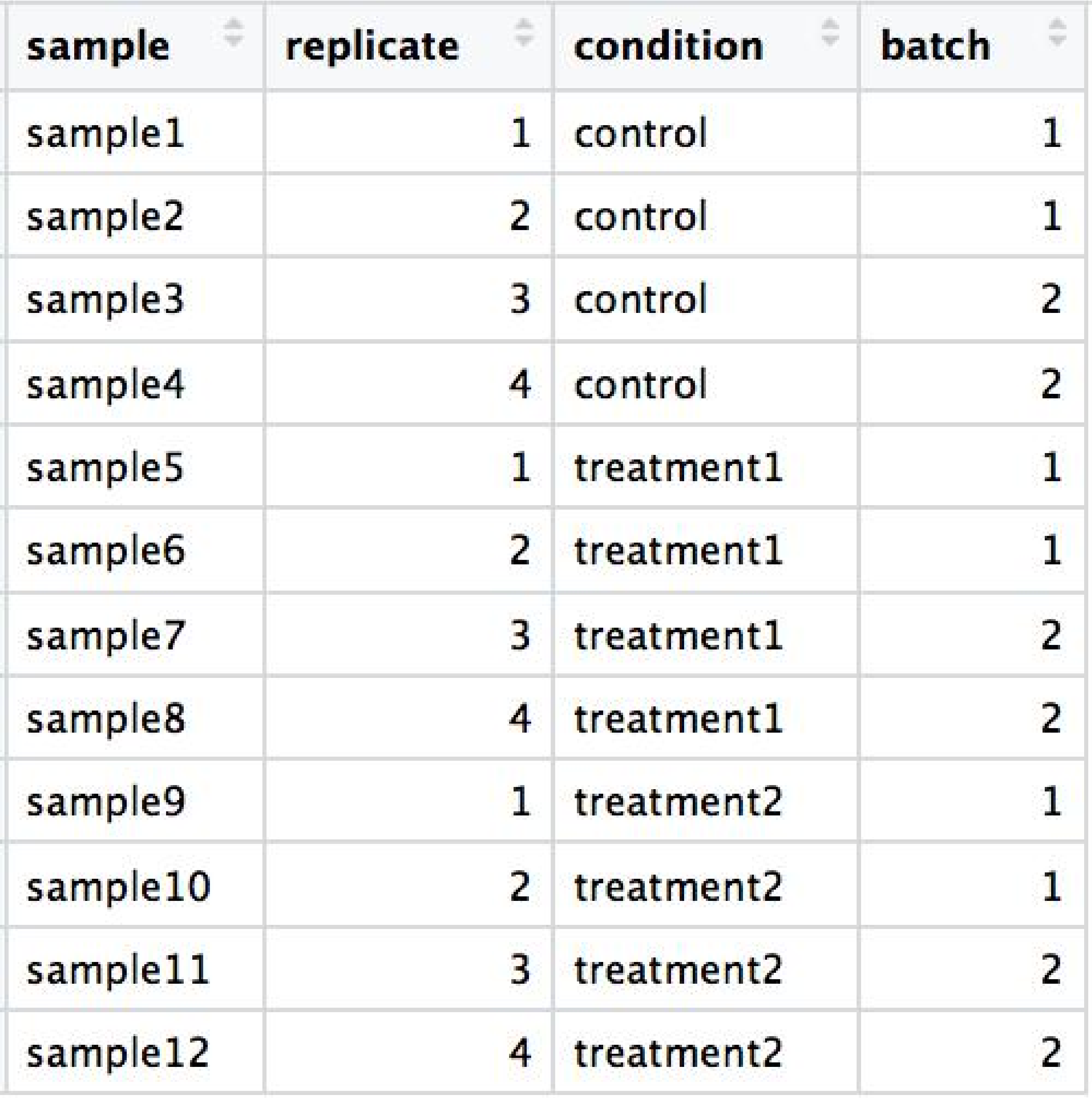
Hicks SC, et al., bioRxiv
(2015)

Best practices regarding batches:

Design the experiment in a way to avoid batches, if possible.

If unable to avoid batches:

- Do NOT confound your experiment by batch:
- DO split replicates of the different sample groups across batches. The more replicates the better (definitely more than 2).
- DO include batch information in your experimental metadata.



sample	replicate	condition	batch
sample1	1	control	1
sample2	2	control	1
sample3	3	control	2
sample4	4	control	2
sample5	1	treatment1	1
sample6	2	treatment1	1
sample7	3	treatment1	2
sample8	4	treatment1	2
sample9	1	treatment2	1
sample10	2	treatment2	1
sample11	3	treatment2	2
sample12	4	treatment2	2

PRINCIPLES OF A GOOD EXPERIMENTAL DESIGN

01

Randomization

when you assign treatments to experimental units

02

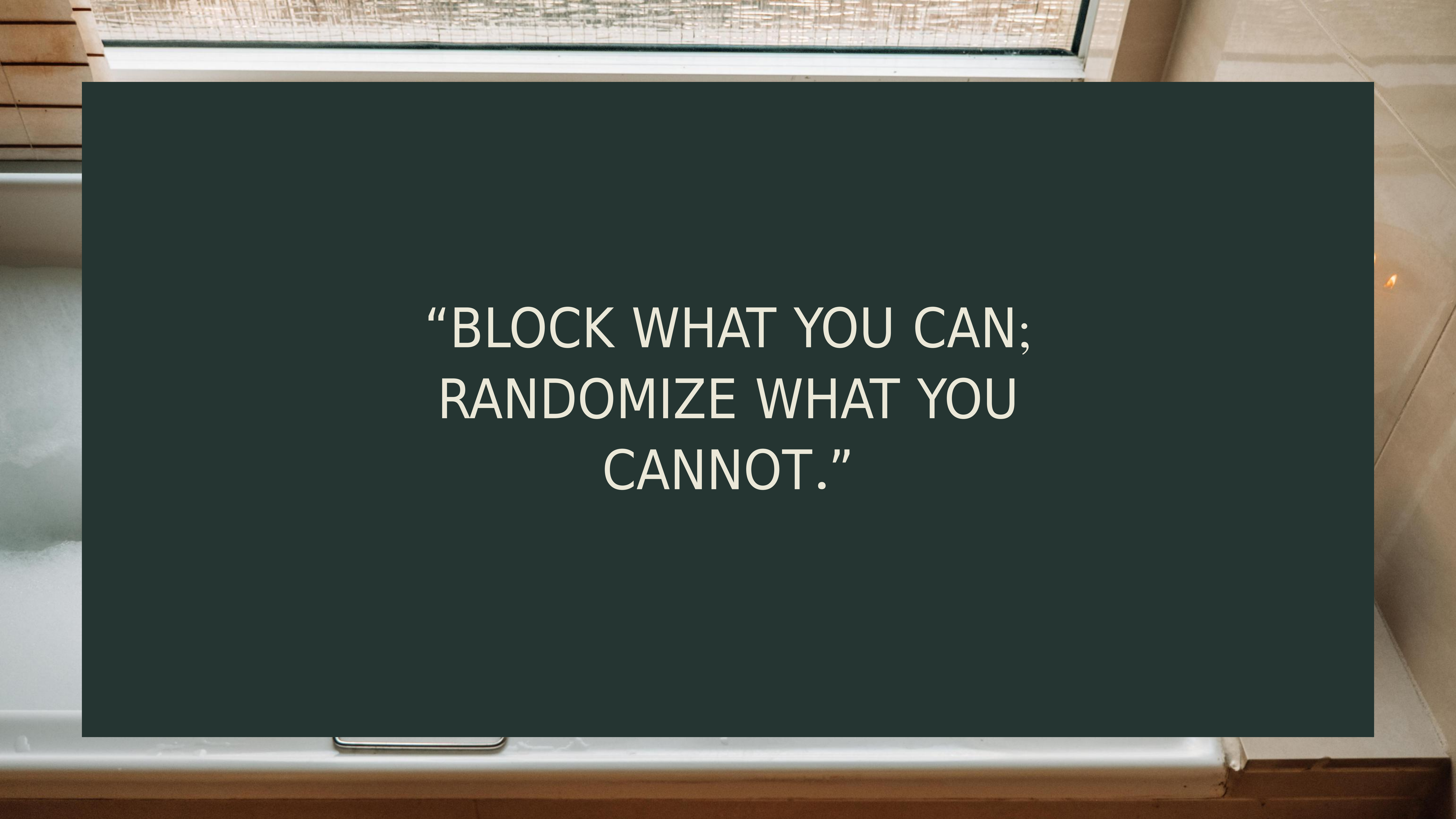
Replication

estimate the error due to the experimenter manipulation, increase the precision by which you estimate the variable of interest

03

Blocking

can help to reduce variability unexplained in one’s model

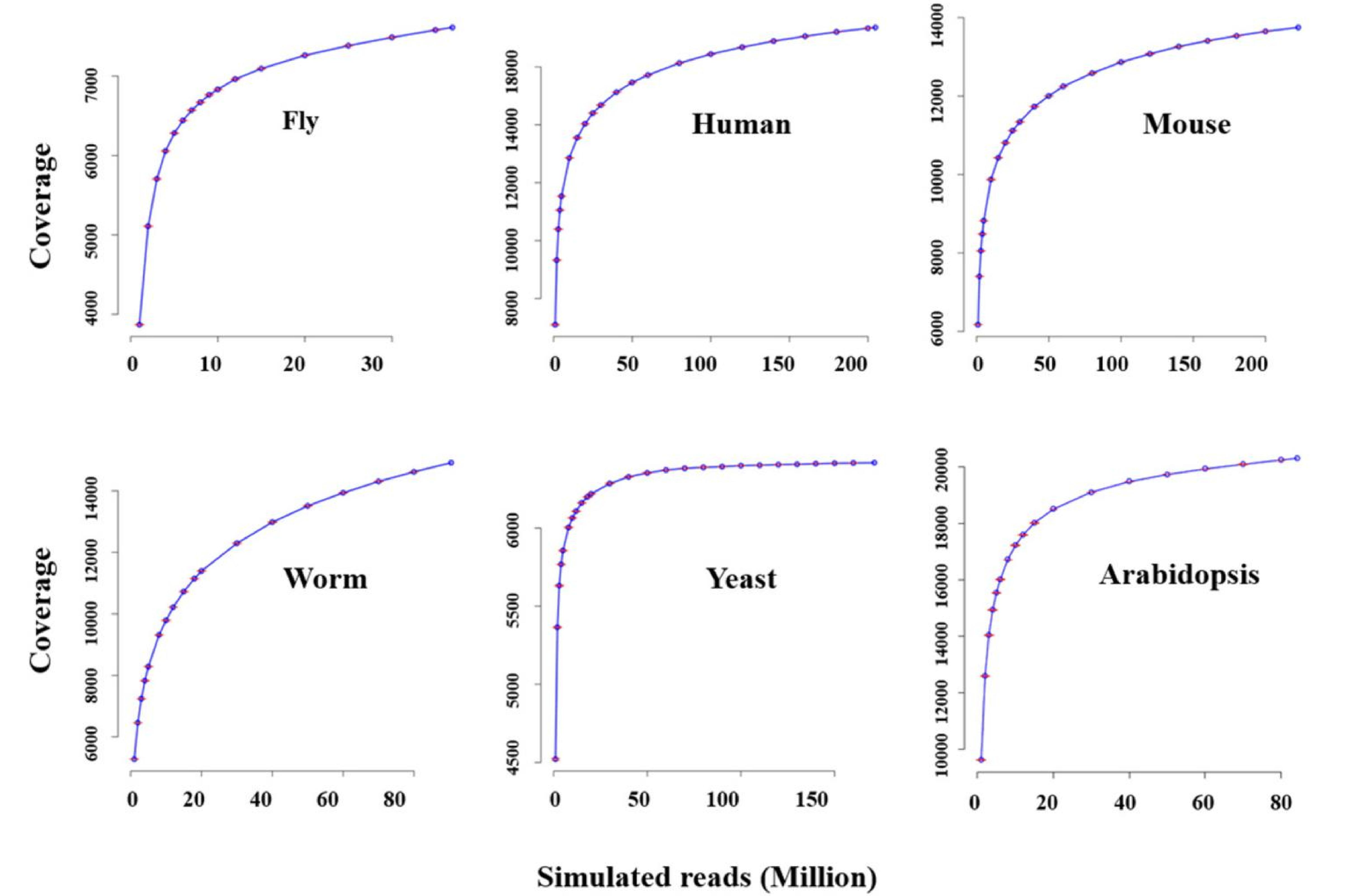


“BLOCK WHAT YOU CAN; RANDOMIZE WHAT YOU CANNOT.”

3. AND SOME OTHER POINTS

SEQUENCING DEPTH

R. Lei et al. / Gene 557 (2015) 82–87



Lei et al.
(2014)

POOLING OF SAMPLES

you need at least 3 individuals to get enough material for your control replicate and at least 5 individuals to get enough material for your treatment replicate?

pool 5 individuals for the control and 5 individuals for the treatment conditions. You would also make sure that the individuals that are pooled in both conditions are similar in sex, age, etc.

TAKE–HOME MESSAGE

- Low statistical power reduces the chance of detecting a true effect.

- Replication, randomization and blocking are the three core principles of proper experimental design.

- Confounding happens when two sources of variation cannot be distinguished from one another.

- Randomize what you cannot control, block what you can control.

- Maximizing the number of biological replicates in RNA–seq experiments is key to increase statistical power and lower the number of false negatives.

Further references:

https://scienceparkstudygroup.github.io/rna-seq-lesson/02-experimental-design-considerations/index.html
https://www.youtube.com/watch ?v=7ECKDDaUyQE&ab_channel=XiaoleShirleyLiu
https://www.youtube.com/watch ?v=ogk6CvuroYY&ab_channel=DNALearningCenter
https://www.youtube.com/watch ?v=qVjiaX80cb4&ab_channel=XiaoleShirleyLiu
https://www.youtube.com/watch ?v=7ECKDDaUyQE&ab_channel=XiaoleShirleyLiu
https://www.youtube.com/watch ?v=ogk6CvuroYY&ab_channel=DNALearningCenter
https://www.youtube.com/watch ?v=qVjiaX80cb4&ab_channel=XiaoleShirleyLiu