# Day 05

Pseudoreplication, confounding variables, sampling biases

- Pseudoreplication
- Confounding variables
- Sampling biases

Science relies on replicate measurements.

- Var(measurement) = External factor variability
  - + Natural biological variability
  - + Measurement error
- Additional replicates → more accurate & reliable summary statistics.
- Replicates can be used to:
  - Assess & isolate sources of variation in measurements
  - Limit the effect of spurious variation on hypothesis testing & parameter estimation.

#### <u>Biological</u> replicates:

- Parallel measurements of biologically distinct samples
- Capture random biological variation (could be subject of study or a noise source).

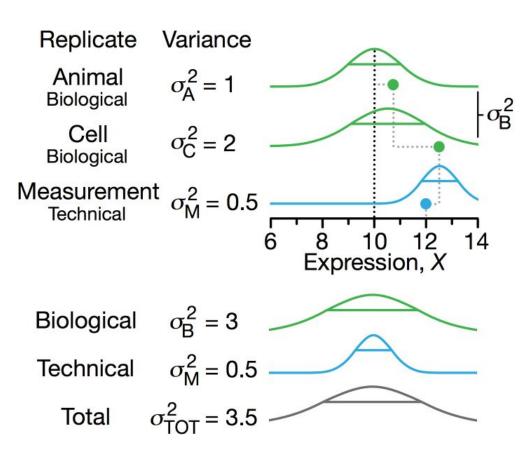
#### <u>Technical</u> replicates:

- Repeated measurements of the same sample
- Represent independent measures of the random noise associated with protocols or equipment.

Which sources of variation are being studied & which are considered noise? B: biological, T: technical

	Replicate type	Replicate category
Animal	Colonies	В
study subjects	Strains	В
	Cohoused groups	В
	Gender	В
	Individuals	В
Sample preparation	Organs from sacrificed animals	В
	Methods for dissociating cells from tissue	T
	Dissociation runs from given tissue sample	T
	Individual cells	В
	RNA-seq library construction	T
Sequencing	Runs from the library of a given cell	T
	Reads from different transcript molecules	$\Lambda_p$
	Reads with unique molecular identifier (UMI) from a given transcript molecule	T

https://www.nature.com/articles/nmeth.3091



Which sources of variation are being studied & which are considered noise? B: biological, T: technical

	Replicate type	Replicate category <sup>a</sup>
Animal	Colonies	В
study subjects	Strains	В
	Cohoused groups	В
	Gender	В
	Individuals	В
Sample preparation	Organs from sacrificed animals	В
	Methods for dissociating cells from tissue	T
	Dissociation runs from given tissue sample	T
	Individual cells	В
	RNA-seq library construction	T
Sequencing	Runs from the library of a given cell	T
	Reads from different transcript molecules	$\Lambda_p$
	Reads with unique molecular identifier (UMI) from a given transcript molecule	T

• Sample size (n)

$$Var(\overline{X}) = \sigma^2 / n$$

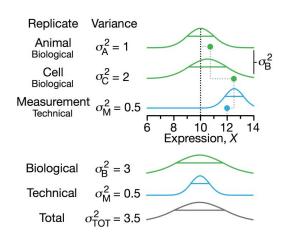
• Effective sample size  $(n_{eff})$ 

$$Var(\overline{X}) = \sigma^2 / n_{eff}$$

If  $\rho$  is the correlation between samples,

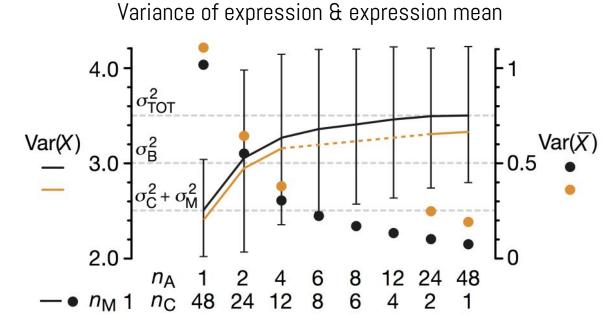
$$n_{ ext{eff}} = rac{n}{1+(n-1)
ho}$$

$$n_{\rm eff} \neq n$$
: Pseudoreplication



#### **Simulation**

$$n = n_{\rm A} n_{\rm C} n_{\rm M} = 48$$
  
 $n_{\rm A} = 1.48$ ,  $n_{\rm C} = 1.48$ ,  $n_{\rm M} = 1$ , 3  
 $n_{\rm eff} = 2.48 = {\rm Var}({\rm X})/{\rm Var}({\rm X}_{\rm mean})$ 



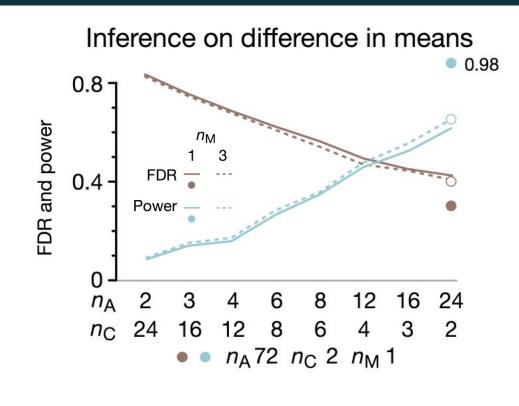
$$Var(X_{mean}) = \sigma_A^2/n_A + \sigma_C^2/n_A n_C + \sigma_M^2/n_A n_C n_M$$

 $n_{\rm eff}$  2.4 5.5 11 15 19 25 37 48

Num. replicates has a practical effect on inference errors in analysis of differences of means or variances.

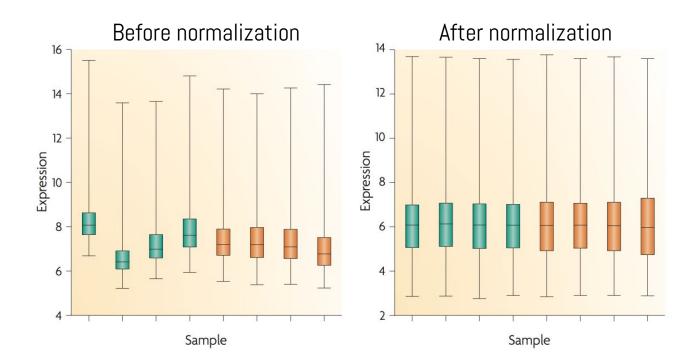
Simulation of 10% effect in mean

- More animals the better.
- $(n_A, n_C, n_M)$  from (24,2,3) to (72,2,1): 50% inc. in power (0.66 $\rightarrow$ 0.98).
- Consider cost difference between biological and technical replicates.

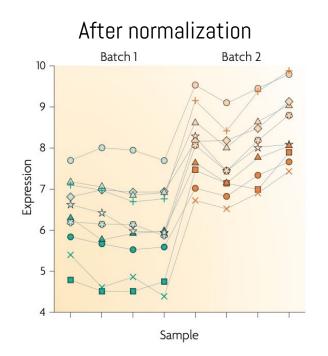


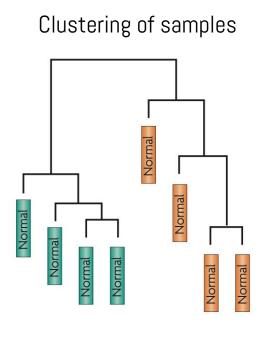
- Typically, biological variability >> technical variability.
  - Commit resources to sampling biologically relevant variables (unless measures
    of technical variability are themselves of interest).
- Planning for replication:
  - 1. Identify the question the experiment aims to answer.
  - Determine proportion of variability induced by each step.
  - 3. Distribute the capacity for replication of the experiment across steps.
  - 4. Be aware of the potential for pseudoreplication and aim to design statistically independent replicates.
- As capacity for higher-throughput assays increases: more is not always better.

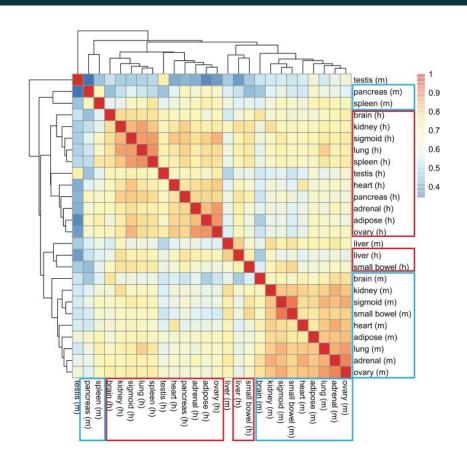
Extraneous variables (e.g. processing data) can be *confounded* with the outcome of interest (e.g. disease state) when it correlates both with the outcome and with an independent variable of interest (e.g. gene expression).



Extraneous variables (e.g. processing data) can be *confounded* with the outcome of interest (e.g. disease state) when it correlates both with the outcome and with an independent variable of interest (e.g. gene expression).



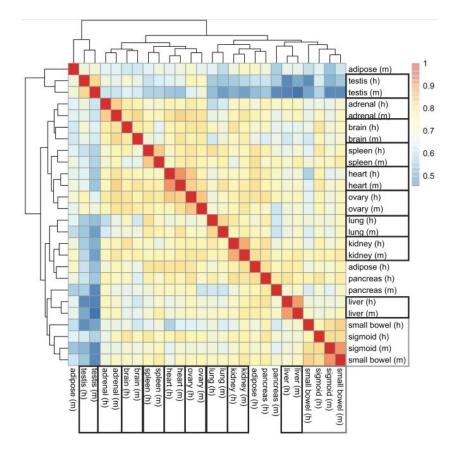




# Mouse ENCODE comparative gene expression data

Lin et al. (2011) Comparison of the transcriptional landscapes between human and mouse tissues. PNAS 111:17224.

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX, lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX, lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX, lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	Human
testis		pancreas		Mouse



Re-analysis of the data after correcting for batch-effects.

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX, lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX, lane 4)	MONK (run 312, flow cell C2GR3ACXX, lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX, lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	Human
testis		pancreas		Mouse

#### **Exploratory analyses**

Hierarchically cluster the samples and label them with biological variables and batch surrogates (such as laboratory and processing time)



Plot individual features versus biological variables and batch surrogates



Calculate principal components of the high-throughput data and identify components that correlate with batch surrogates

#### Downstream analyses

Do you believe that measured batch surrogates (processing time, laboratory, etc.) represent the only potential artefacts in the data?

Ves

Use measured technical variables as surrogates for batch and other technical artefacts

Estimate artefacts from the high-throughput data directly using surrogate variable analysis (SVA)

Perform downstream analyses, such as regressions, t-tests or clustering, and adjust for surrogate or estimated batch effects. The estimated/surrogate variables should be treated as standard covariates, such as sex or age, in subsequent analyses or adjusted for use with tools such as ComBat

#### Diagnostic analyses

Use of SVA and ComBat does not guarantee that batch effects have been addressed. After fitting models, including processing time and date or surrogate variables estimated with SVA, re-cluster the data to ensure that the clusters are not still driven by batch effects