

Day 04

Pseudoreplication, confounding variables

- Pseudoreplication
- Confounding variables

Replication & Pseudoreplication

Science relies on replicate measurements.

- $\text{Var}(\text{measurement}) = \text{External factor variability}$
 - + Natural biological variability
 - + Measurement error
- Additional replicates \rightarrow 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.

Replication & Pseudoreplication

Biological replicates:

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

Technical replicates:

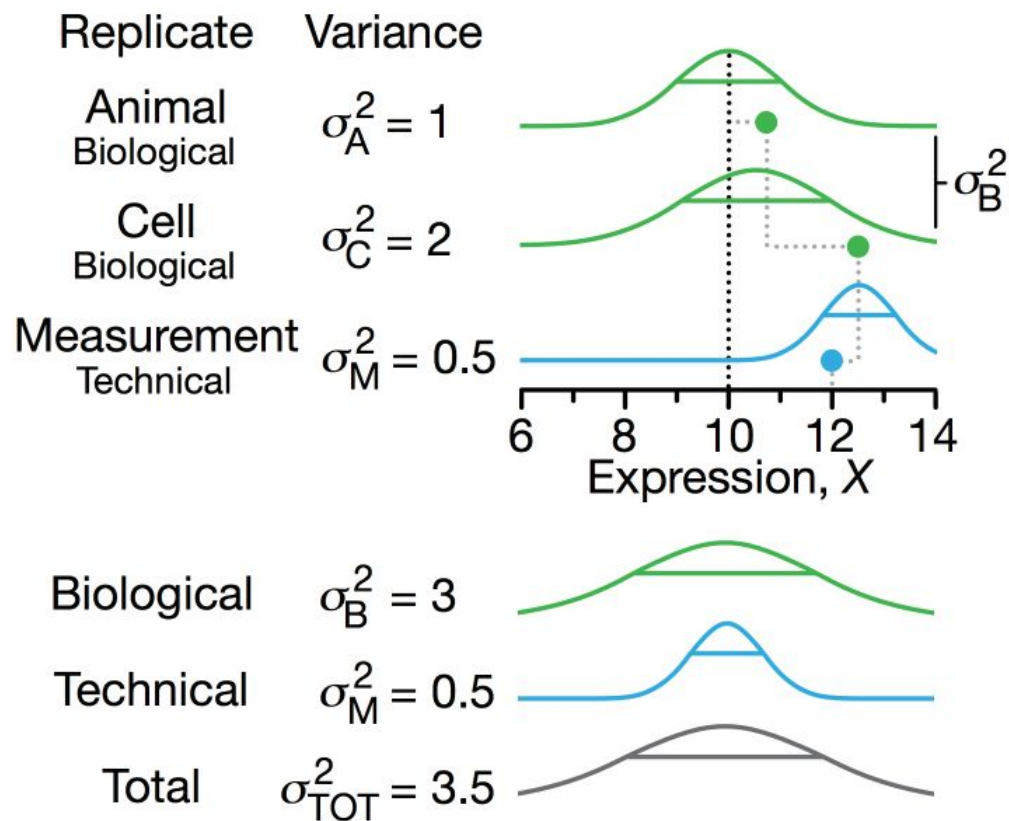
- Repeated measurements of the same sample
- Represent independent measures of the random noise associated with protocols or equipment.

Which sources of variation are being studied & which are considered noise?

B: biological, T: technical

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

Replication & Pseudoreplication



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

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

Replication & Pseudoreplication

- Sample size (n)

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

- Effective sample size (n_{eff})

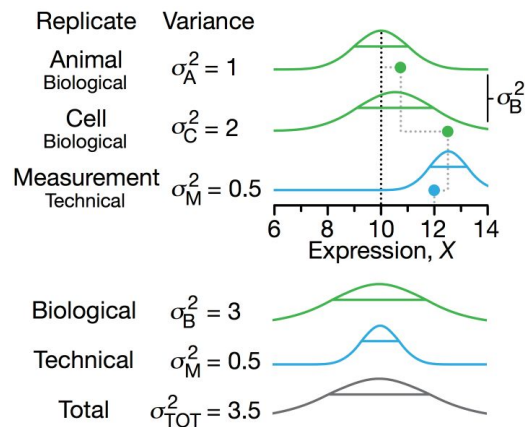
$$\text{Var}(\bar{X}) = \sigma^2 / n_{\text{eff}}$$

If ρ is the correlation between samples,

$$n_{\text{eff}} = \frac{n}{1 + (n - 1)\rho}$$

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

Replication & Pseudoreplication



Variance of expression & expression mean

Simulation

$$n = n_A n_C n_M = 48$$

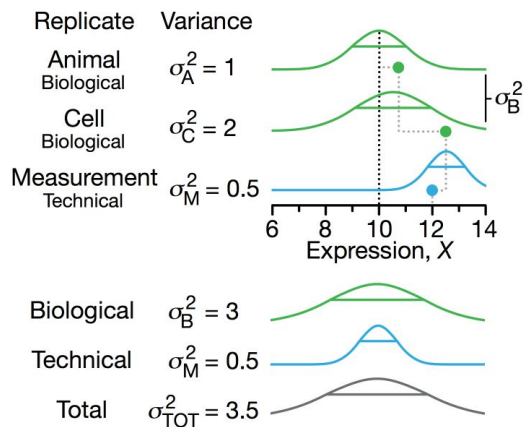
$$n_A = 1:48, n_C = 1:48, n_M = 1, 3$$

$$n_{\text{eff}} = 2:48 = \text{Var}(X)/\text{Var}(X_{\text{mean}})$$

	n_A	1	2	4	6	8	12	24	48	
— ● n_M	1	n_C	48	24	12	8	6	4	2	1
		n_{eff}	2.4	5.5	11	15	19	25	37	48

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

Replication & Pseudoreplication

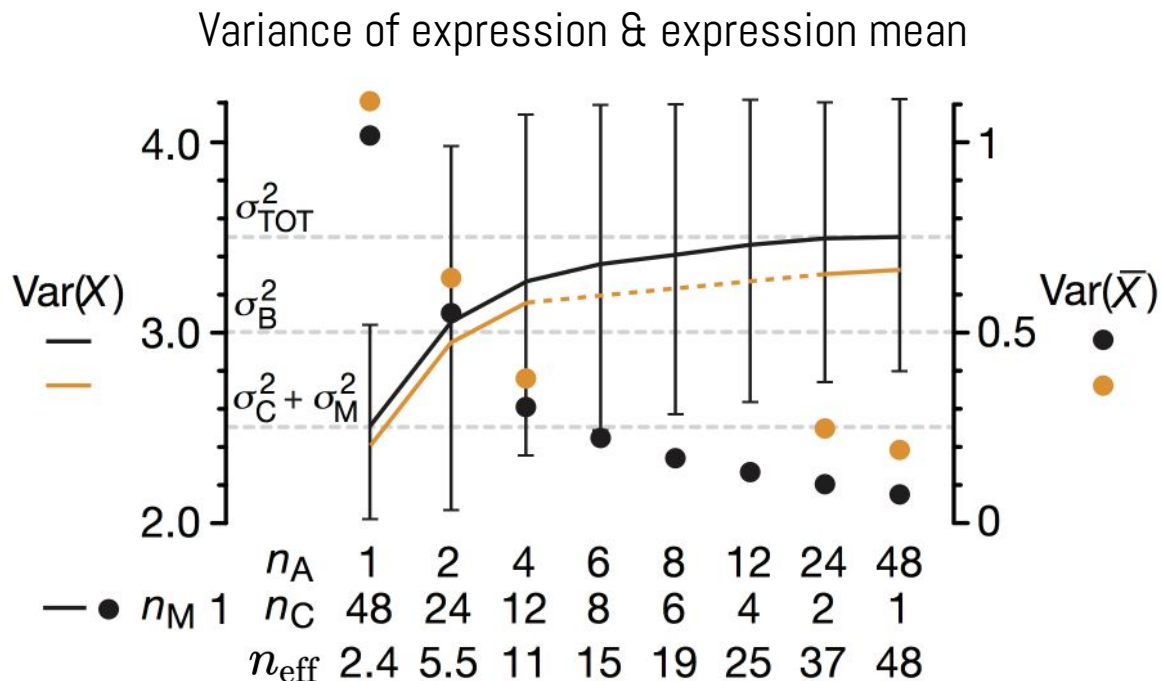


Simulation

$$n = n_A n_C n_M = 48$$

$$n_A = 1:48, n_C = 1:48, n_M = 1, 3$$

$$n_{\text{eff}} = 2:48 = \text{Var}(X)/\text{Var}(X_{\text{mean}})$$



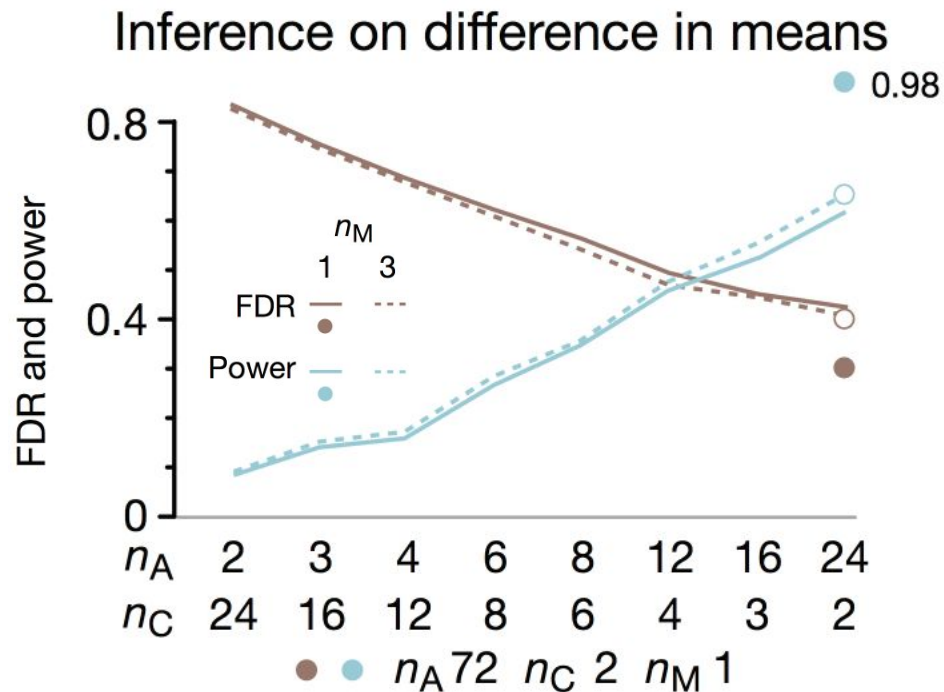
$$\text{Var}(X_{\text{mean}}) = \sigma_A^2/n_A + \sigma_C^2/n_A n_C + \sigma_M^2/n_A n_C n_M$$

Replication & Pseudoreplication

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 → 0.98).
- Consider cost difference between biological and technical replicates.

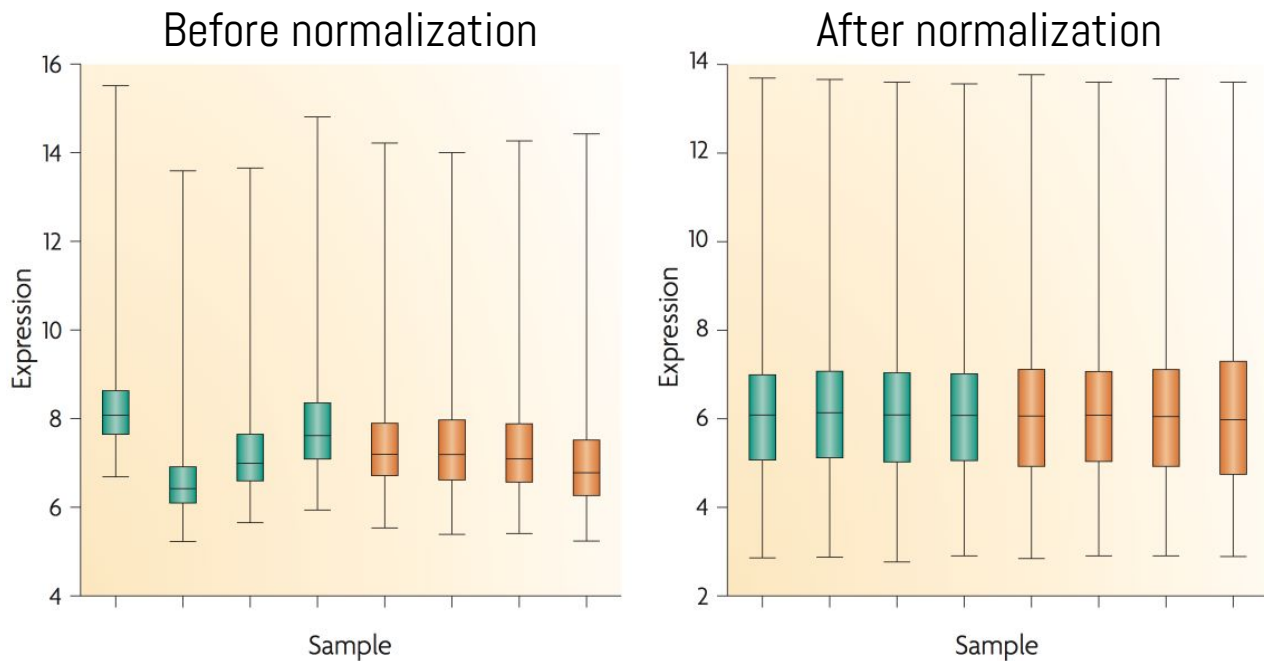


Replication & Pseudoreplication

- Typically, biological variability \gg 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.
 2. 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.

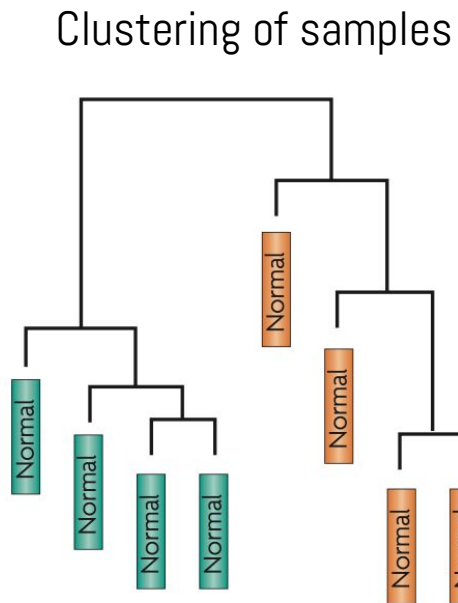
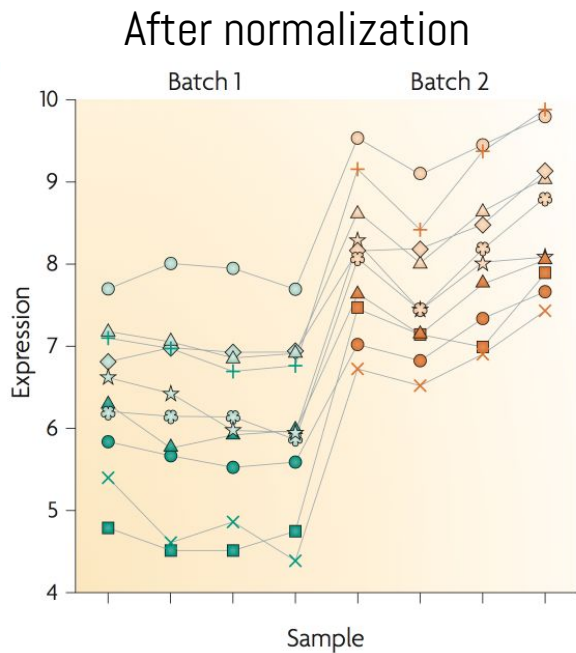
Confounding variables

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).

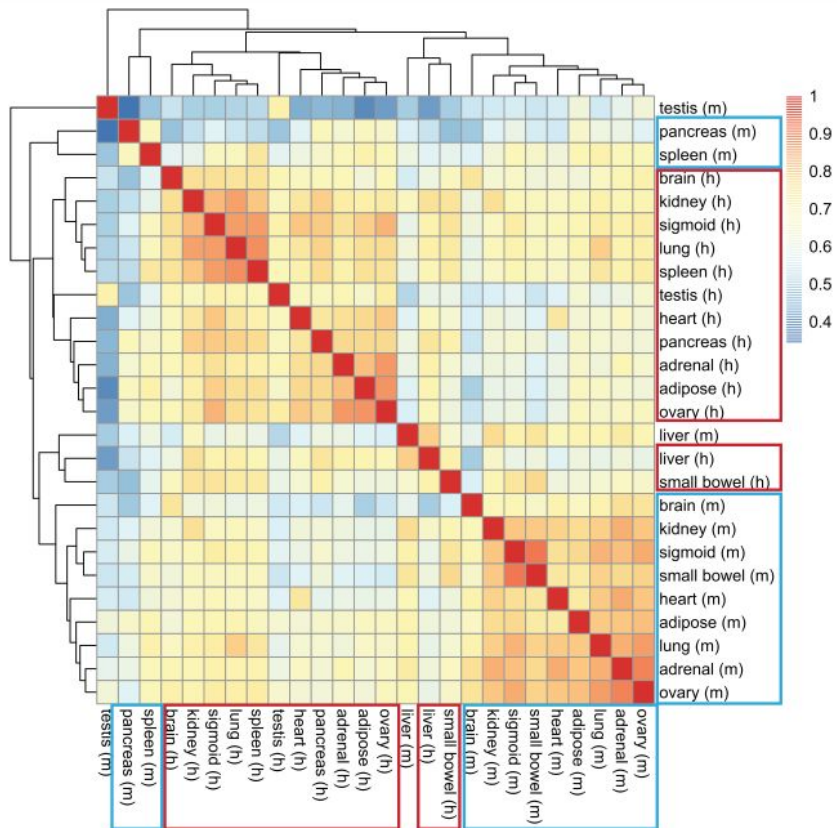


Confounding variables

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).



Confounding variables

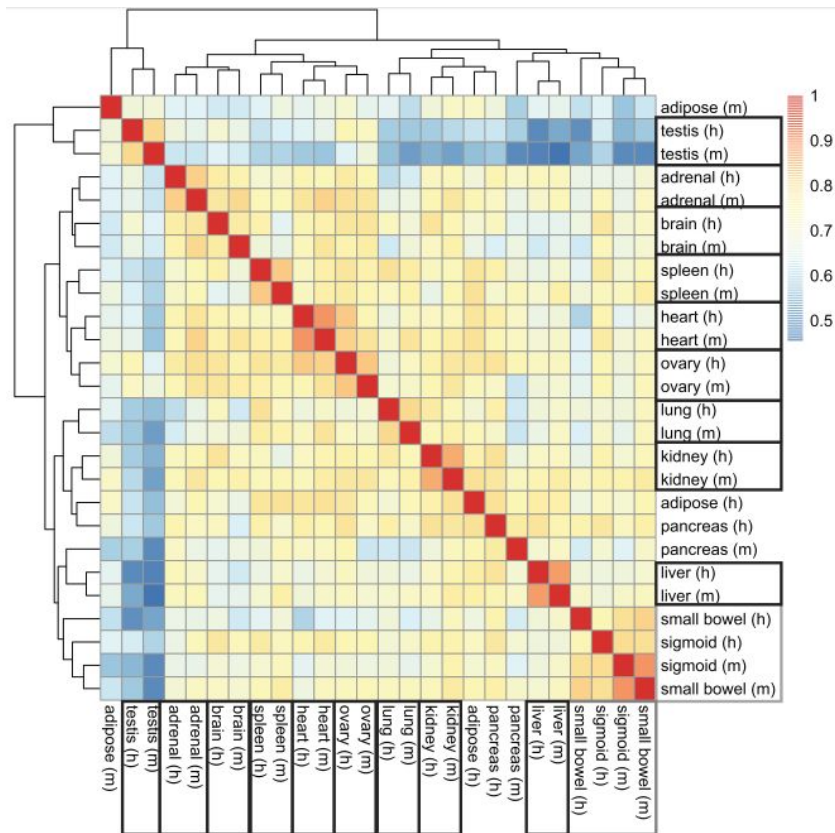


Mouse ENCODE comparative gene expression data

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

Confounding variables



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

Confounding variables

Exploratory analyses

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



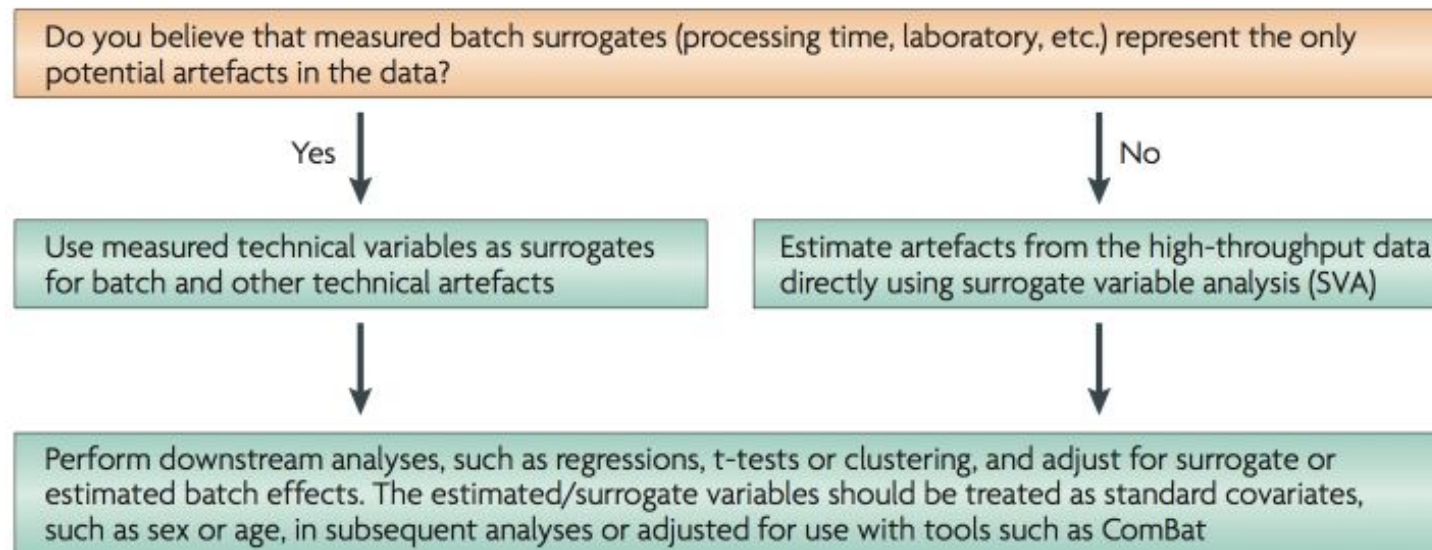
Plot individual features versus biological variables and batch surrogates



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

Confounding variables

Downstream analyses



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

Sampling biases

Selection bias

- Sampling bias
- Rejection of “bad” data on arbitrary grounds (esp. post hoc)
- Cherry-picking data subsets or experimental run
- Data chunking & analysis - P-hacking or Data dredging

Confirmation bias

Reporting bias