Day 05

Pseudoreplication, confounding variables, sampling biases

- Pseudoreplication
- Confounding variables
- Sampling biases

Replication & Pseudoreplication

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.

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:

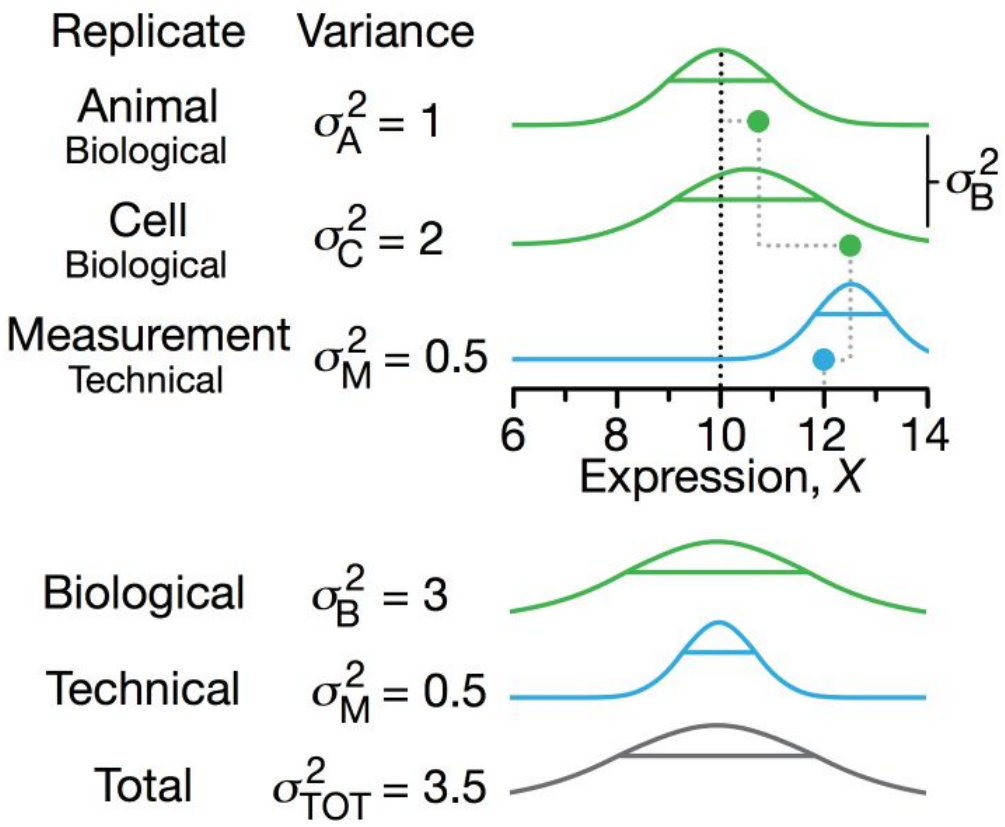
- Repeated measurements of the same sample
- Represent independent measures of the random noise associated with protocols or equipment.

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

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

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

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

Replication & Pseudoreplication

- Sample size (n)

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

- Effective sample size (n_{eff})

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

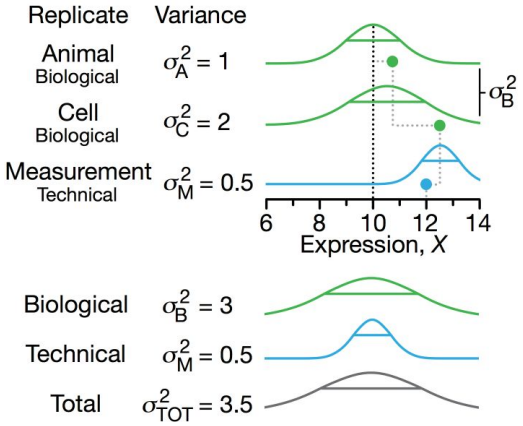
  If ρ is the correlation between samples,

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

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

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

Replication & Pseudoreplication



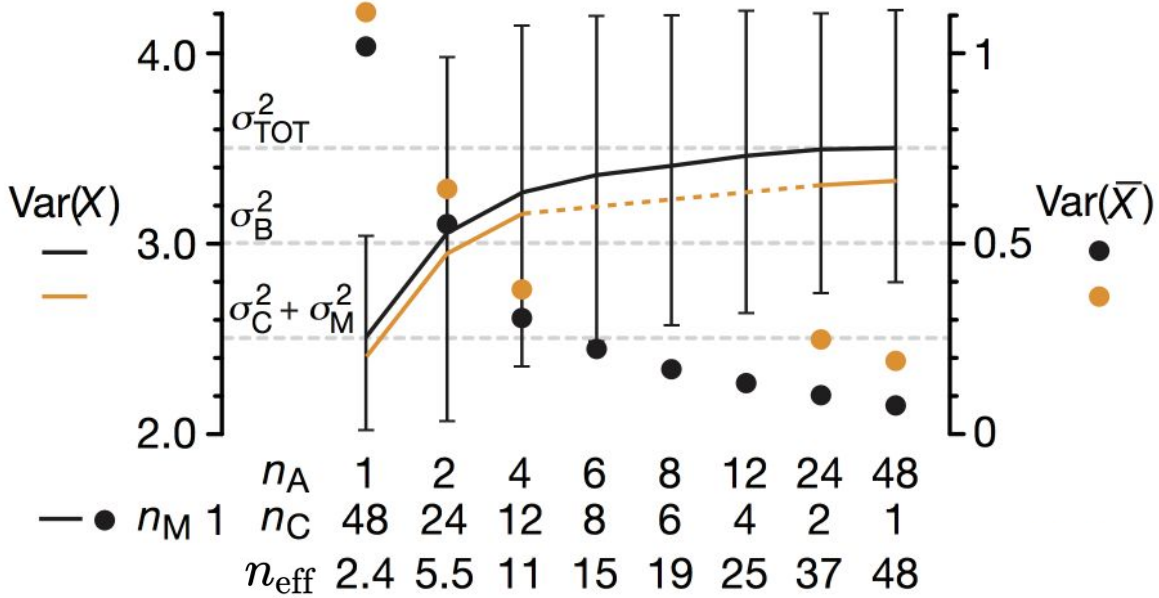
Simulation

$n = n_A n_C n_M = 48$

$n_A = 1{:}48,\ n_C = 1{:}48,\ n_M = 1, 3$

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

Variance of expression & expression mean



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

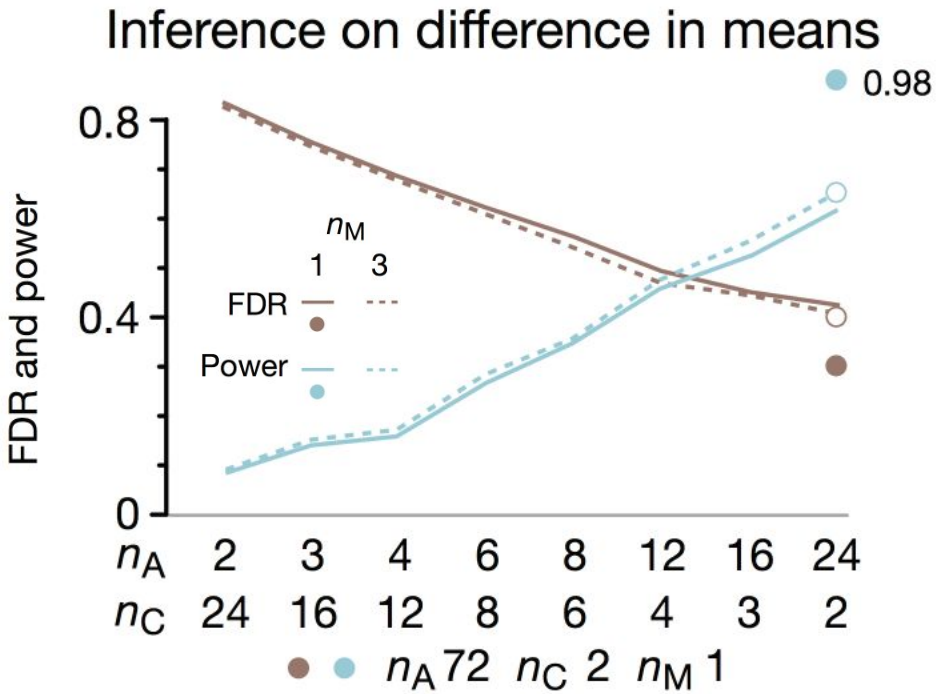
https://www.nature.com/articles/nmeth.3091

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.



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

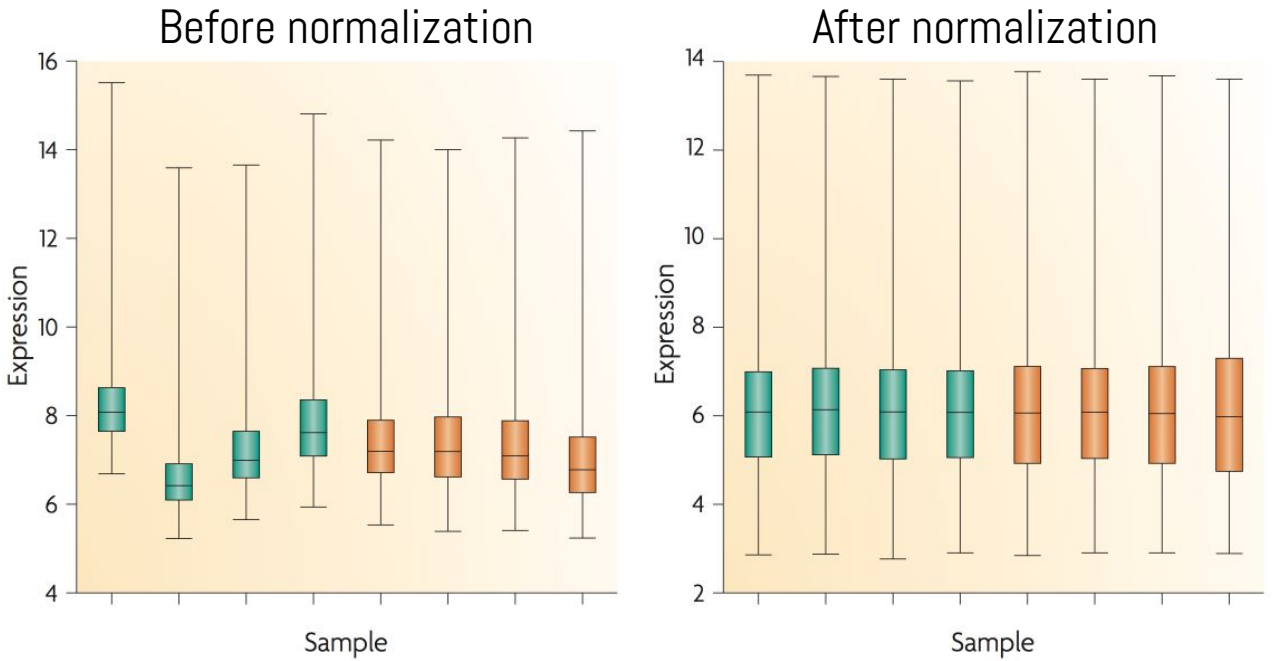
Replication & Pseudoreplication

- 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.
  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.

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

Confounding variables

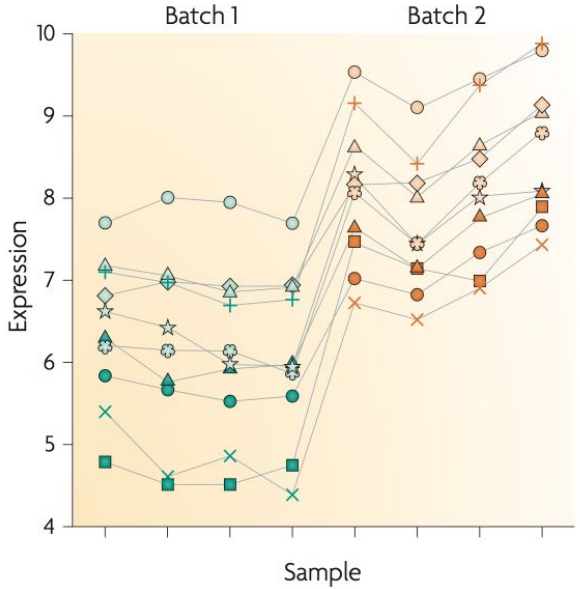
Extraneous variables (e.g. processing data) can be *confounded* with the outcome of interest (e.g. disease state) when it correlates both with the outcome and with an independent variable of interest (e.g. gene expression).



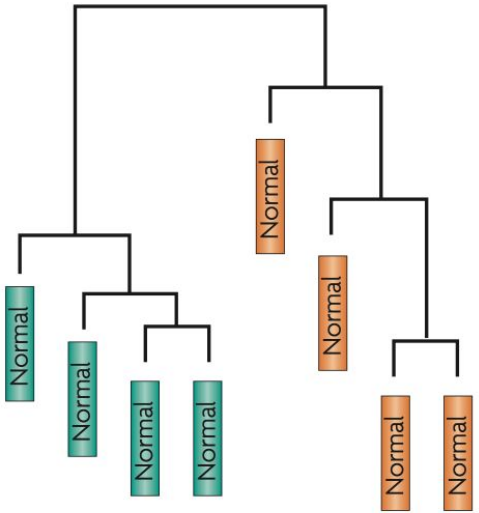
Confounding variables

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).

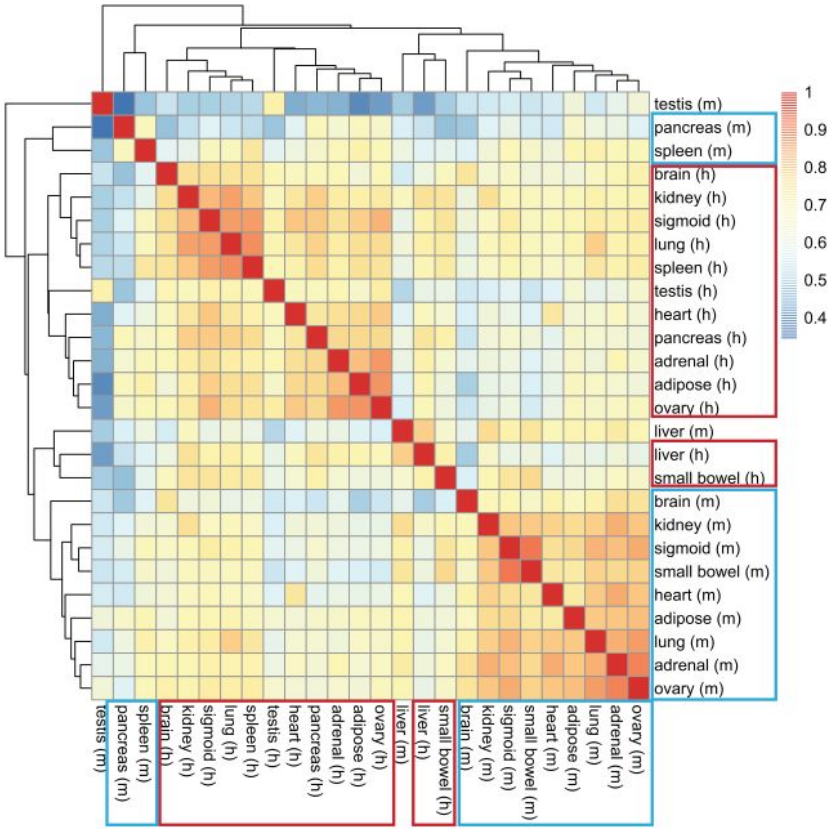
After normalization



Clustering of samples



Confounding variables



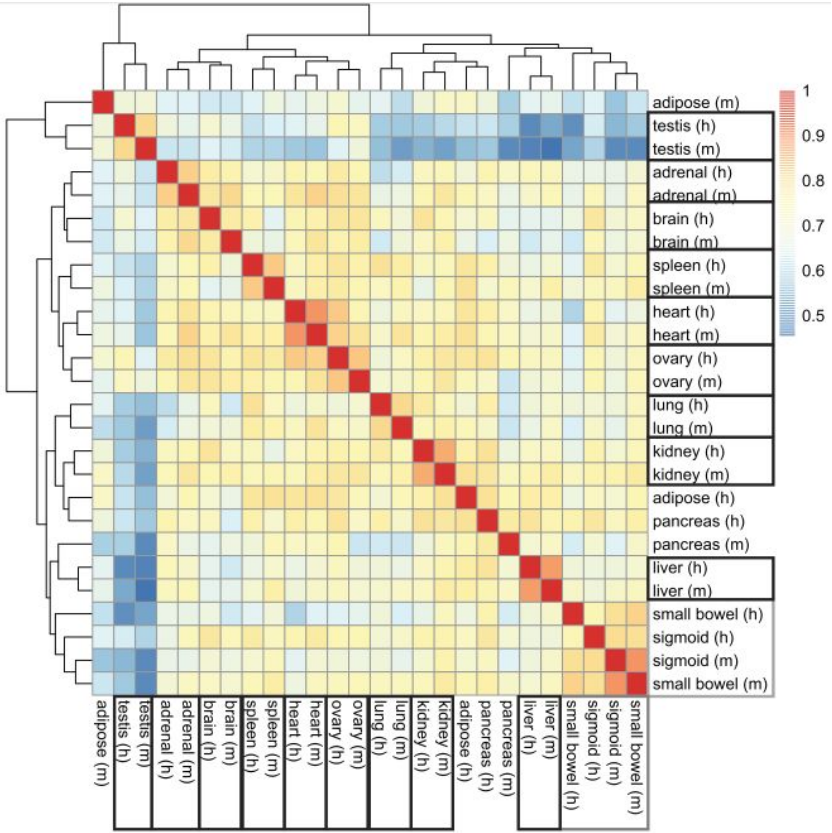
Mouse ENCODE comparative gene expression data

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

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4516019/

Confounding variables

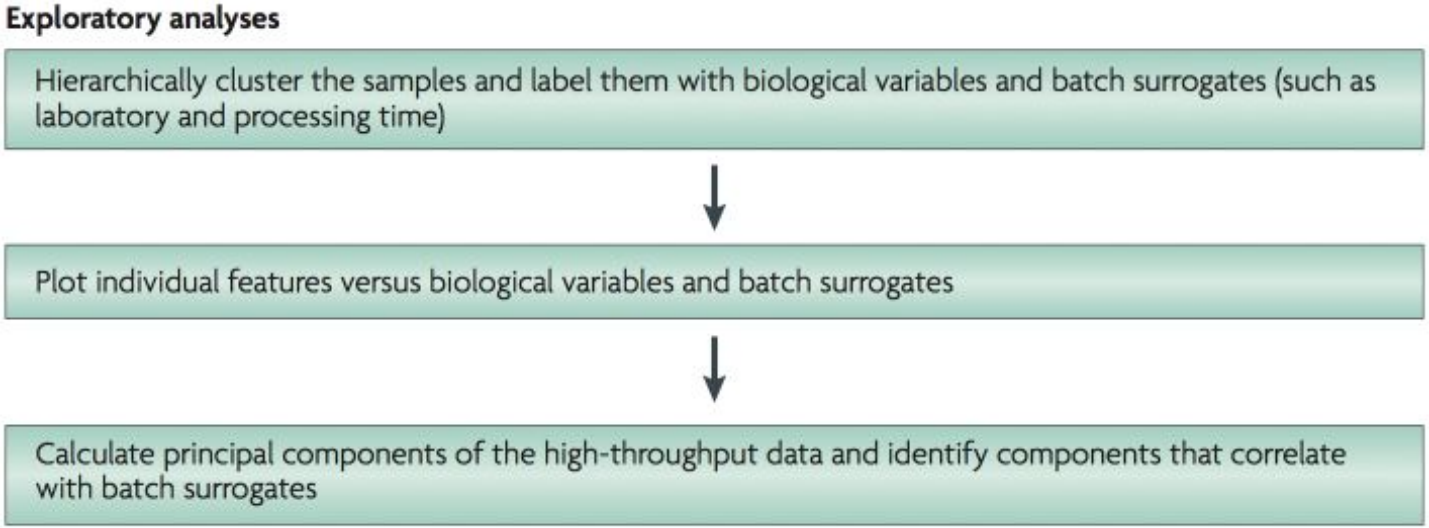


Re-analysis of the data after correcting for batch-effects.

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

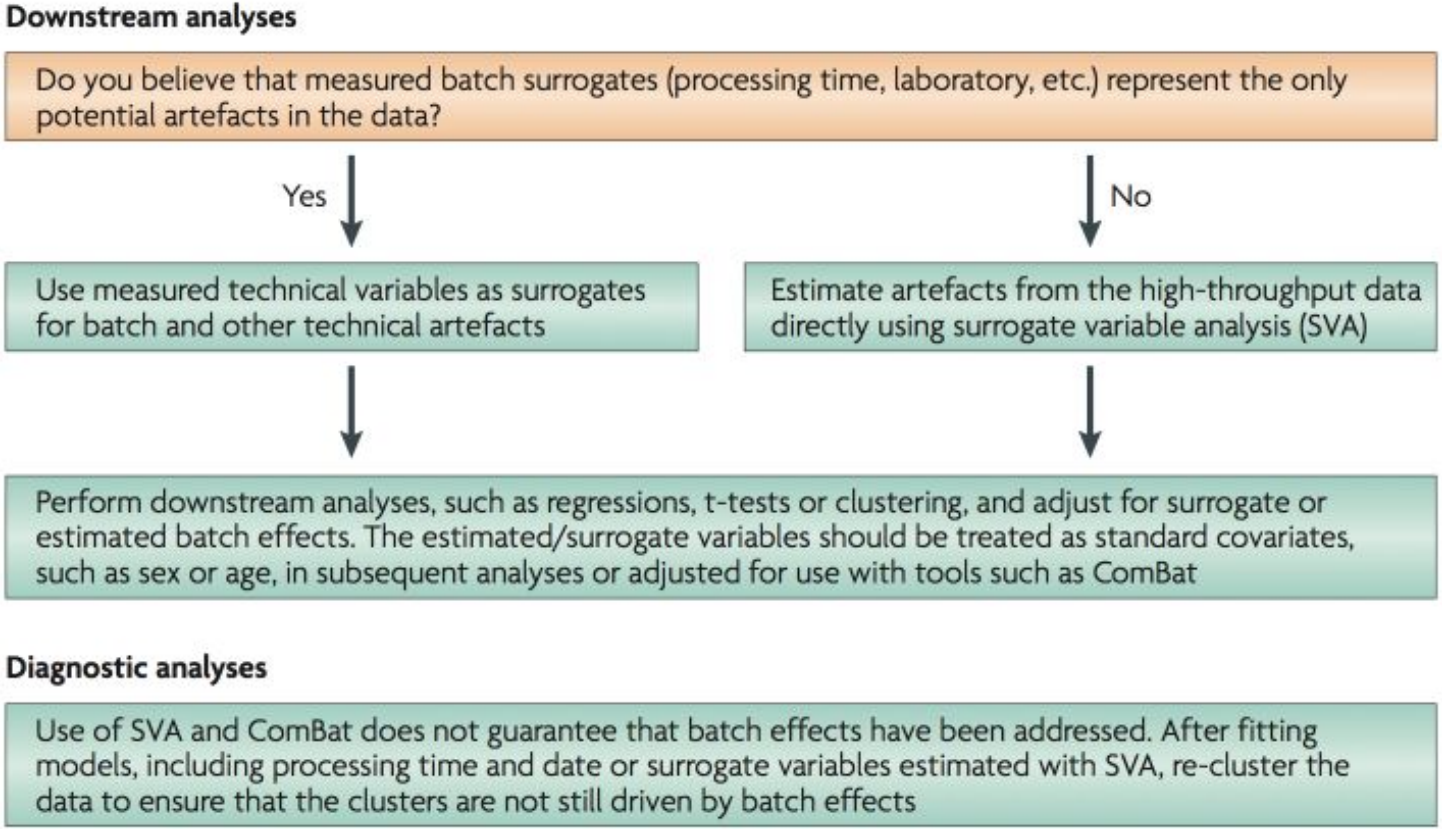
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4516019/

Confounding variables



https://www.nature.com/articles/nrg2825

Confounding variables



https://www.nature.com/articles/nrg2825