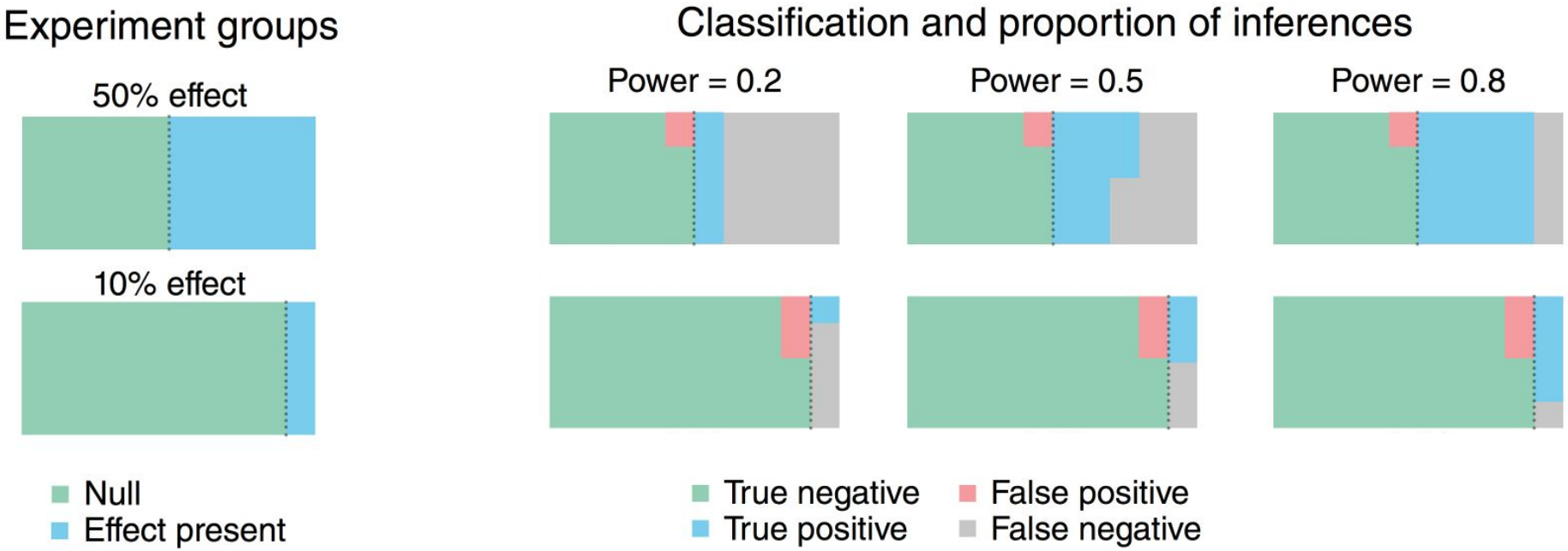
Topic 2: Power, replication, design

Lectures 4 & 5

- Statistical power
- Sample size
- Pseudoreplication & Confounding variables
- Experimental design

Statistical power



- Power of a study is the probability that it can distinguish an effect of a given size from random chance; Power = True positive rate = Sensitivity = Recall
- Most studies are underpowered → Waste of resources & Unethical

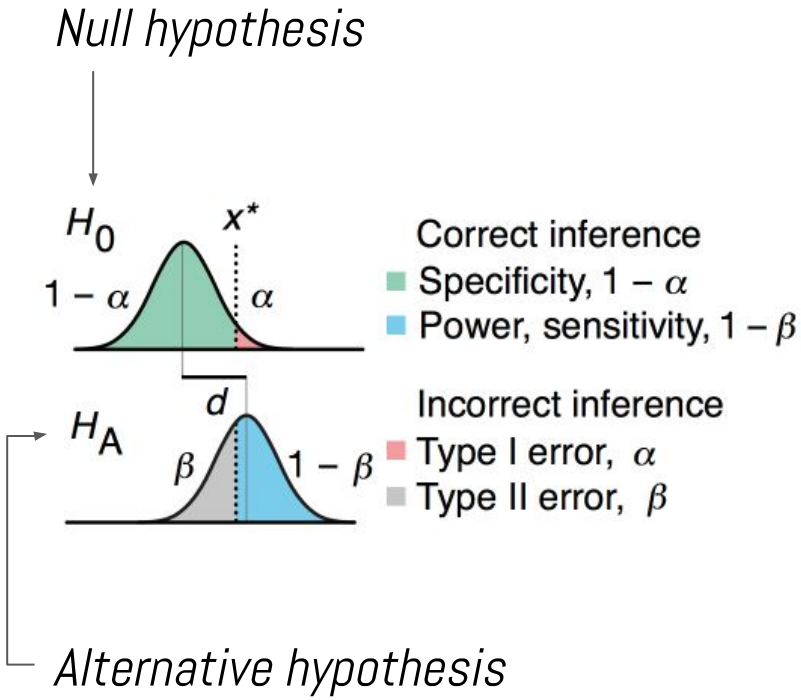
http://www.nature.com/doifinder/10.1038/nmeth.2738

Statistical power

The power is the probability that the test correctly rejects the null hypothesis (H_0) when a specific alternative hypothesis (H_1) is true.

- Pr(reject H_0 | H_1 is true)
- H_1 has to be specific (cannot just be negation of H_0)
- The probability that it will yield a statistically significant outcome.

Power = 1 – β: As power increases → Probability of making type II error (β) decreases.



http://www.nature.com/doifinder/10.1038/nmeth.2738

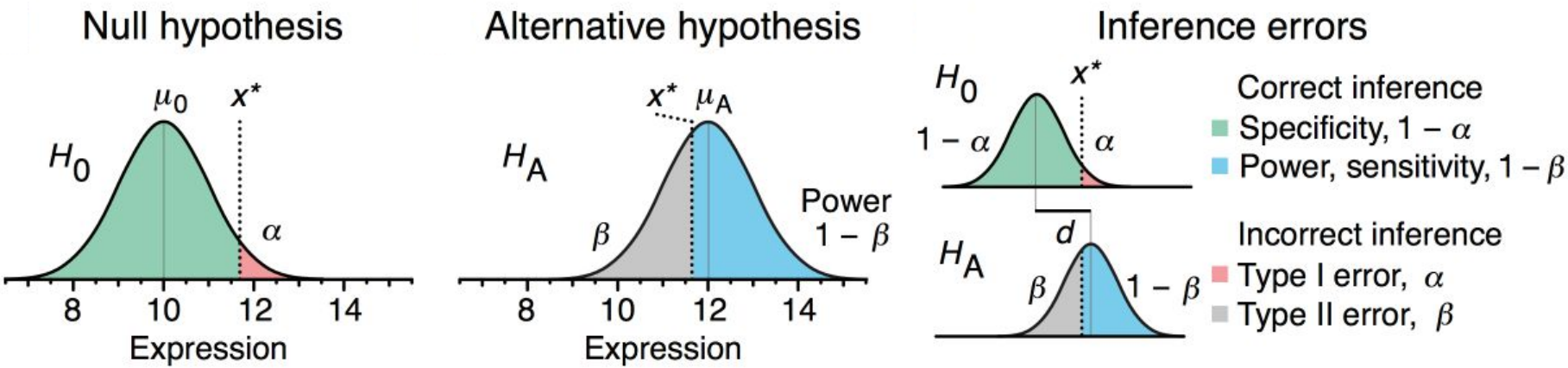
Statistical power

Power depends on:

- Size of the effect: larger the effect, easier it is to detect; standardized effect size is better
- Sample size: collecting more data → easier to detect small effects; relates to the efficiency of a given testing procedure, experimental design, or an estimator (sample size required for a given power)
- Statistical significance criterion: lesser conservative test (larger significance criterion) → more power
- Measurement error: counting cells vs. estimating level of fatigue/depression
- Experimental design: e.g. in a two-sample setting, optimal to have equal number

Write code to generate a power curve to detect unfair coins.

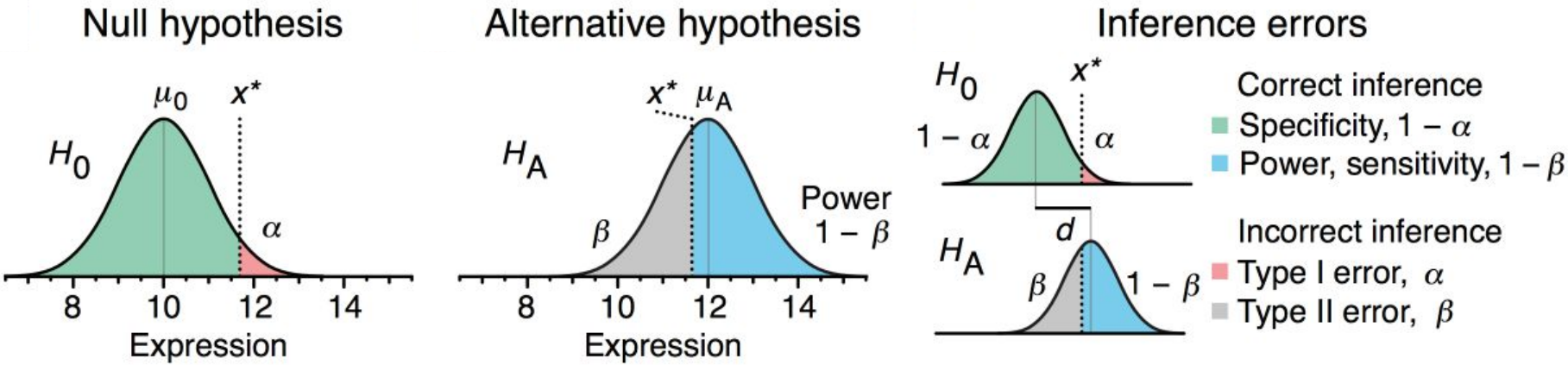
Statistical power



- Values sampled from $H_A < x^*$ do not trigger rejection of H_0 and occur at a rate β.
- Power (sensitivity; TRP) = 1 – β (blue area).
- Good to have low α (FPR) & low β (FNR), but:
  - The α and β rates are inversely related: ↓α → ↑β (& reduces power).

http://www.nature.com/doifinder/10.1038/nmeth.2738

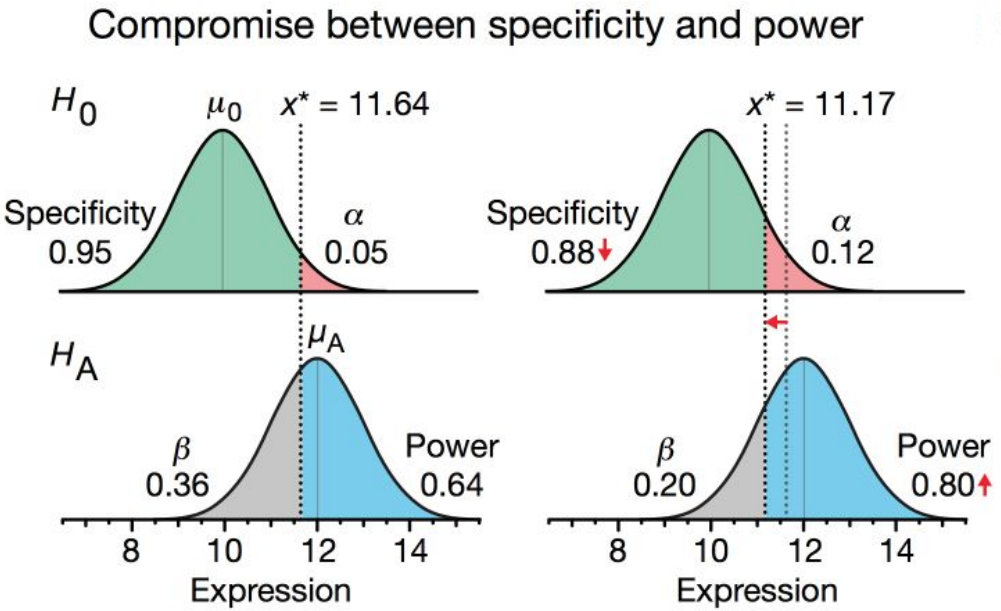
Statistical power

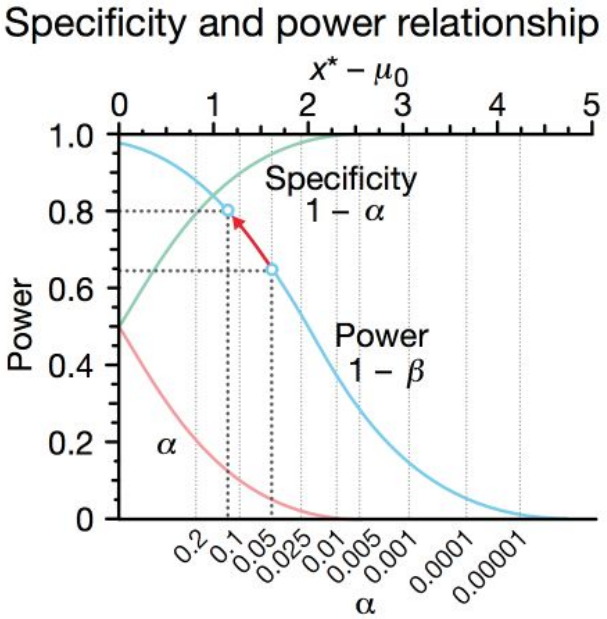


- Typically, $\alpha < \beta$: consequences of FP (in an extreme case, a retracted paper) are more serious than those of FN (a missed opportunity to publish).
- But, the balance between α and β depends on the objectives:
  - If FP are subject to another round of testing but FN are discarded, β should be kept low.

http://www.nature.com/doifinder/10.1038/nmeth.2738

Statistical power

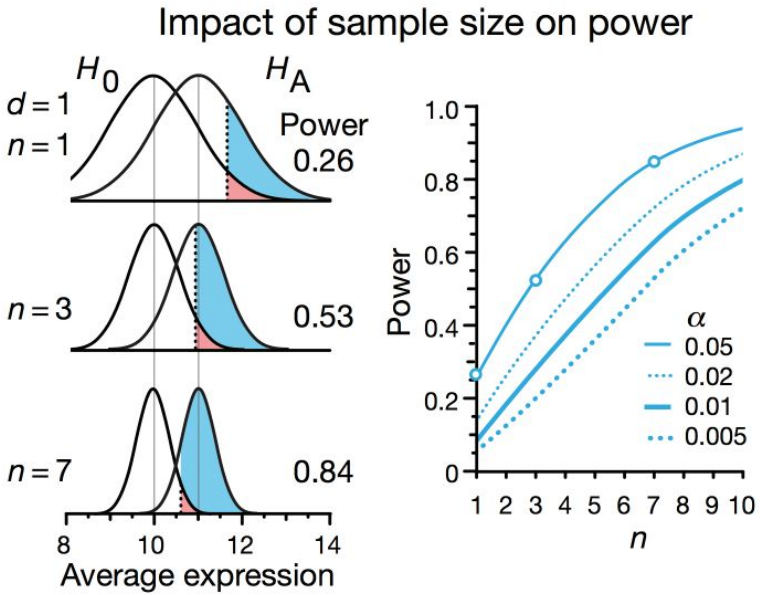


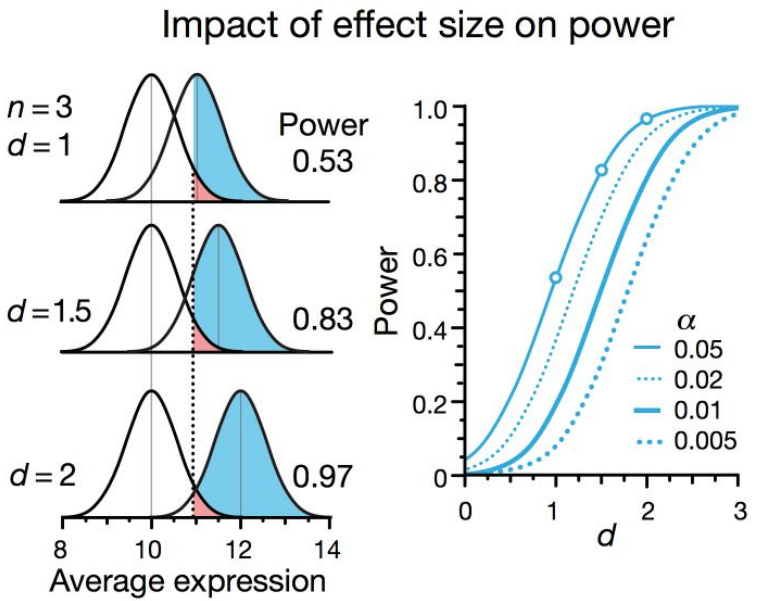


- Decreasing specificity (TNR) increases power (TPR)
- Can we improve our chance to detect increased expression from H_A (increase power) without compromising **α** (increasing FP)?

http://www.nature.com/doifinder/10.1038/nmeth.2738

Statistical power





- Impact of sample size and effect size on power.
- In practice, because we estimate population σ from the samples, power is decreased and we need a slightly larger sample size to achieve the desired power.

http://www.nature.com/doifinder/10.1038/nmeth.2738

Statistical power

Balancing sample size, effect size, and power is critical to good study design.

- First, set the values of type I error (α) and power ($1 - \beta$) to be statistically adequate:
  - Traditionally 0.05 and 0.80, respectively.
- Then determine sample size (n) on the basis of the smallest effect we wish to measure.
  - If the required sample size is too large → may need to reassess objectives or more tightly control the experimental conditions to reduce the variance.
- When the power is low, only large effects can be detected, and negative results cannot be reliably interpreted.

Sample size

- Clinical research (behavioral or drug treatments):
  - Need enough participants to represent all subtypes for which treatment might be used.
  - Some issues: lack reliable methods for diagnosis.
  - Rough rule of thumb: at least 100 people.
    - The actual number needed to find a valid effect depends on a range of factors, including the magnitude and frequency of the effect in the general population.

https://www.spectrumnews.org/news/studies-size-matters-let-us-count-ways/

Sample size

- Brain imaging studies:
  - Historically included 20 or fewer participants. In the past 10 years, closer to 100 participants.
  - Studies that aim to trace developmental trajectories should also track the same few individuals over time, scanning their brains at regular intervals, rather than examining a cross-section of people of different ages at different sites.

https://www.spectrumnews.org/news/studies-size-matters-let-us-count-ways/

Sample size

- Genetic studies (large no. of variants/genes, each making a small contribution):
  - Rare variants in coding regions: order of thousands of people.
  - Risk variants across the whole-genome: tens of thousands of individuals.
    - Millions of statistical tests, one per variant → increases FPR.
  - GWAS: hundreds of thousands of individuals
    - Common gene variants that contribute to the risk of a condition.

https://www.spectrumnews.org/news/studies-size-matters-let-us-count-ways/

Sample size

- Preclinical research:
  - Underpowered animal studies for decades (cost and ethical issues).
  - Make up for their low numbers by analyzing a large number of cells or other samples from each animal → 'pseudoreplication.'
  - Can control lab animals' diets, ages and housing conditions, and scale doses or treatments by weight → sample sizes on the order of 10 animals to be acceptable. Should ≥15 per group to identify important biological effects.
  - In the past few years, push for larger numbers in animal studies.

https://www.spectrumnews.org/news/studies-size-matters-let-us-count-ways/

Sample size

- Biomarker studies (physiological characteristics, such as patterns of eye movements, brain waves or activity, or blood chemistry):
  - Candidate biomarkers have often failed in subsequent studies.
  - Must draw samples from at least 100 individuals.
  - Clinical trials of biomarkers designed to flag people with disease → ≥ 1,000 participants. Researchers should also replicate the efficacy of a biomarker in an independent sample.
  - Some scientists are designing biomarker studies of thousands of participants that combine data from behavioral, imaging and genetic studies.

https://www.spectrumnews.org/news/studies-size-matters-let-us-count-ways/

Sample size

- Field trials:
  - Variables that are hard to control, and so must include hundreds of individuals to yield meaningful results.
  - Needs more than an appropriate number of participants.
    - Representative mix of sexes and ages.

https://www.spectrumnews.org/news/studies-size-matters-let-us-count-ways/

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:

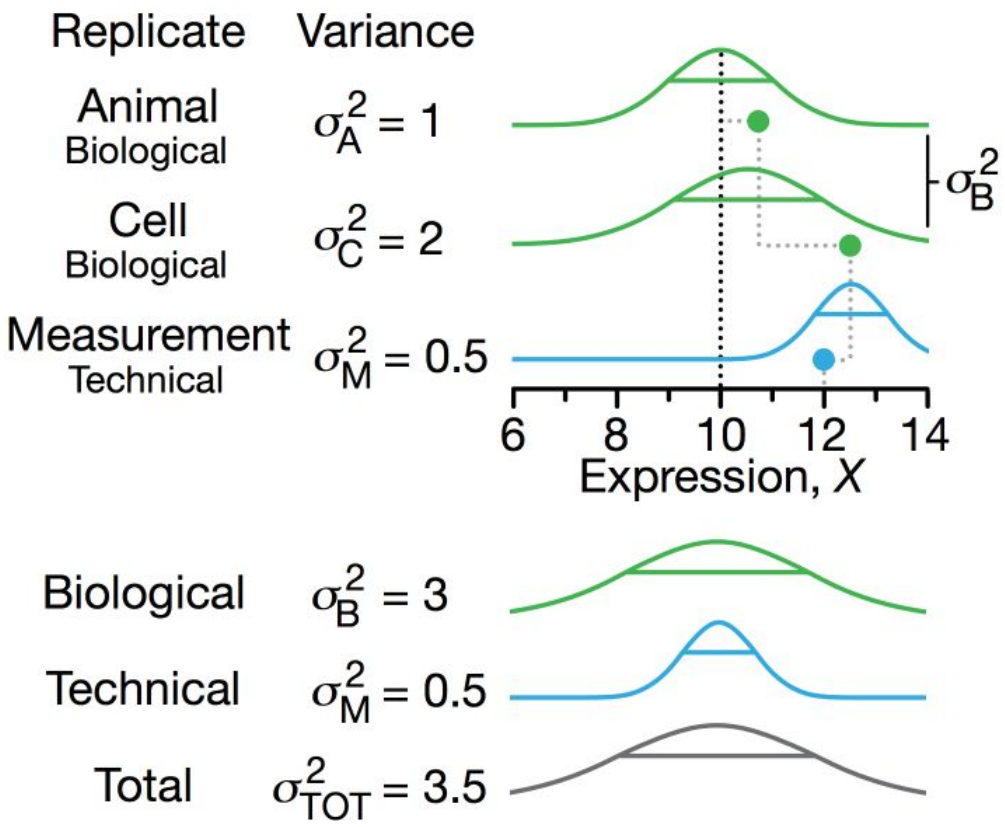
- Repeated measurements of the same sample
- Represent independent measures of the random noise associated with protocols or equipment.

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,

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

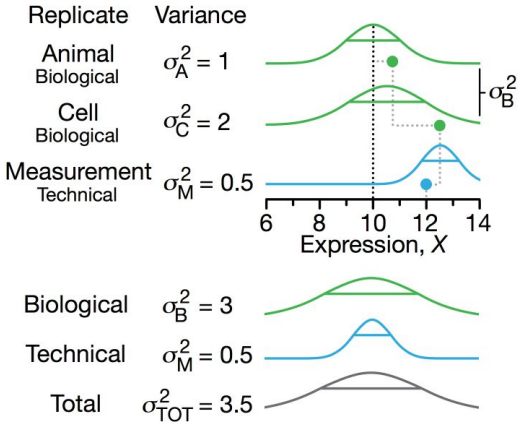
$n_{\mathrm{eff}} \neq n$: Pseudoreplication

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



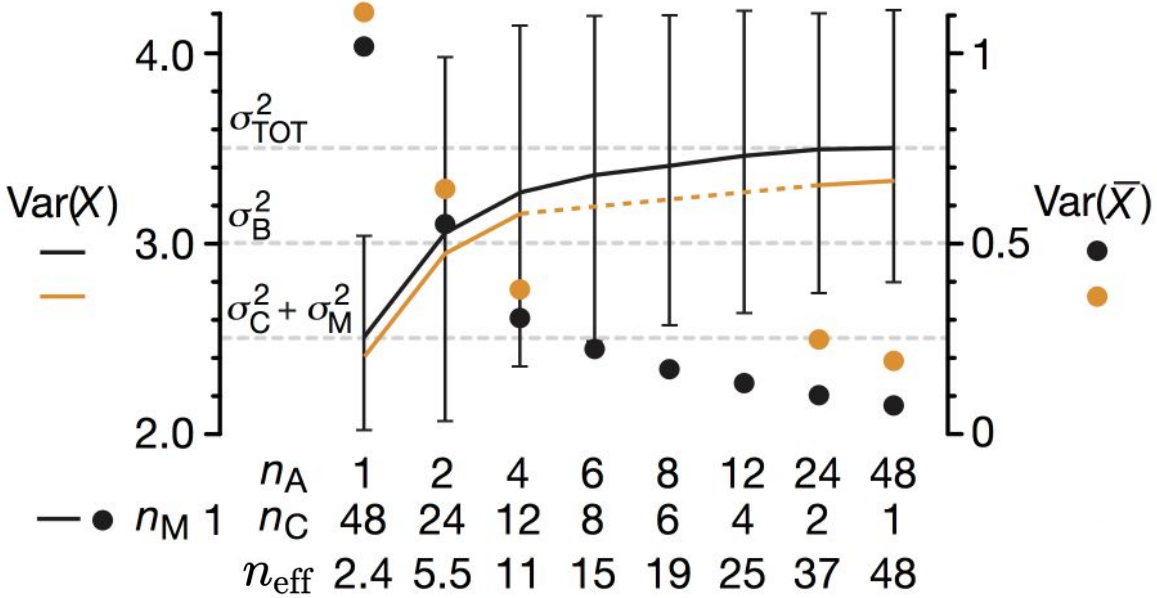
Simulation

$n = n_A n_C n_M = 48$

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

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

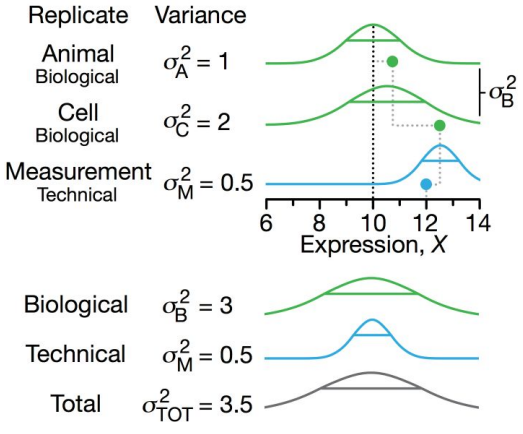
Variance of expression & expression mean



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

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

Replication & Pseudoreplication



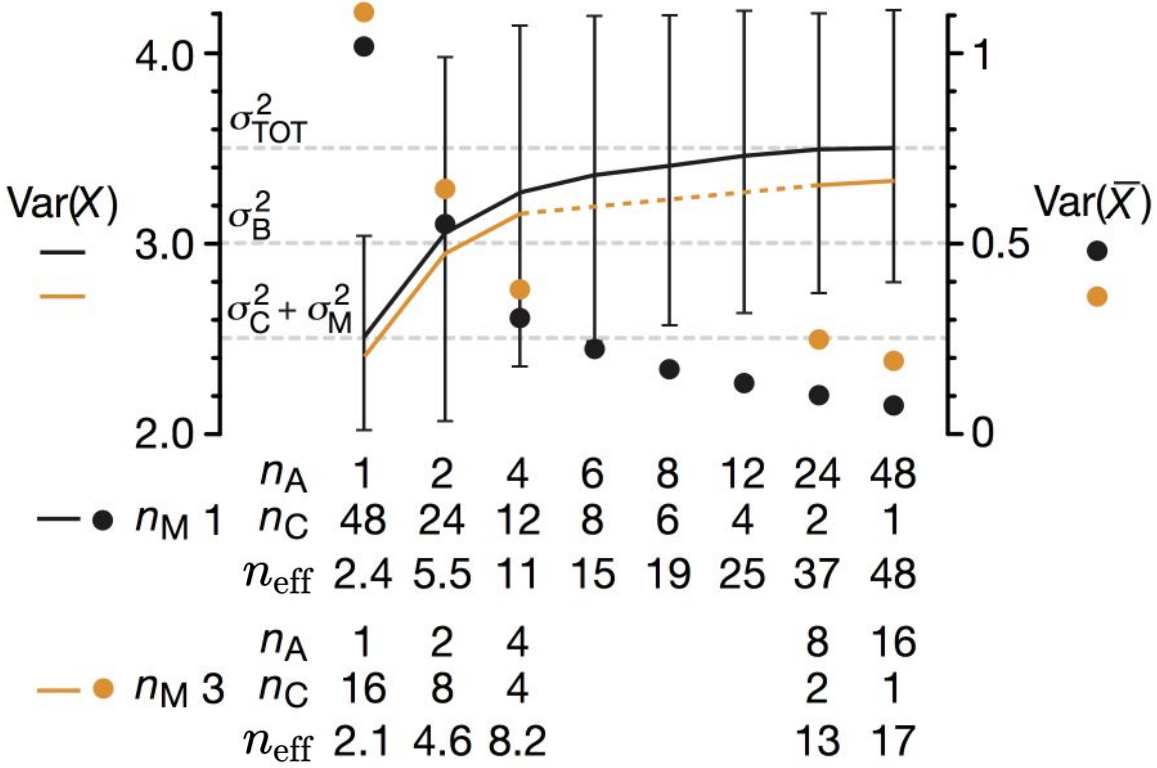
Simulation

$n = n_A n_C n_M = 48$

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

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

Variance of expression & expression mean



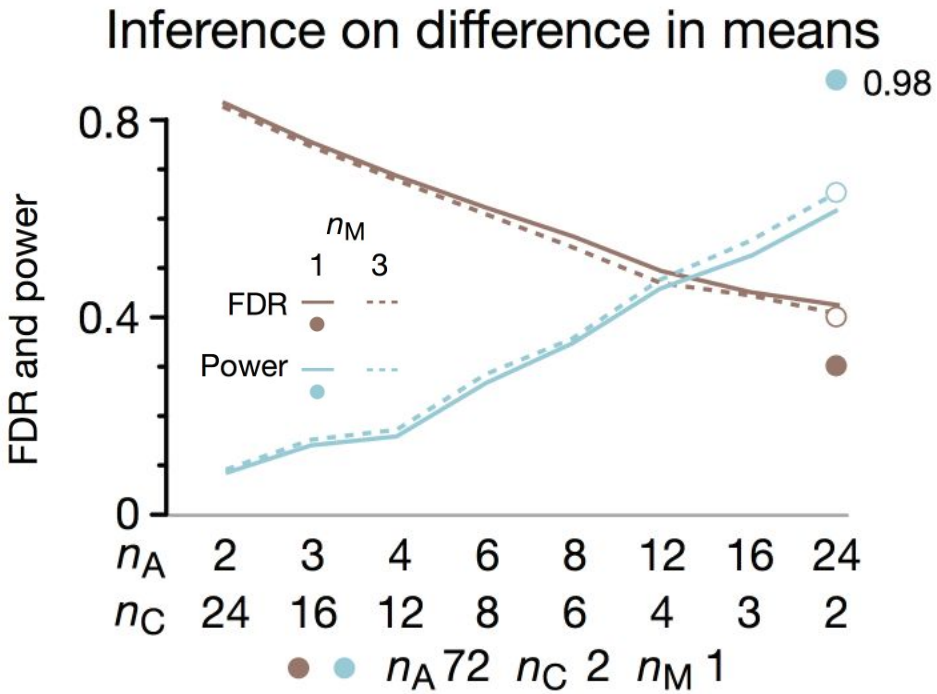
https://www.nature.com/articles/nmeth.3091

Replication & Pseudoreplication

No. 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.
- (nA,nC,nM) 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

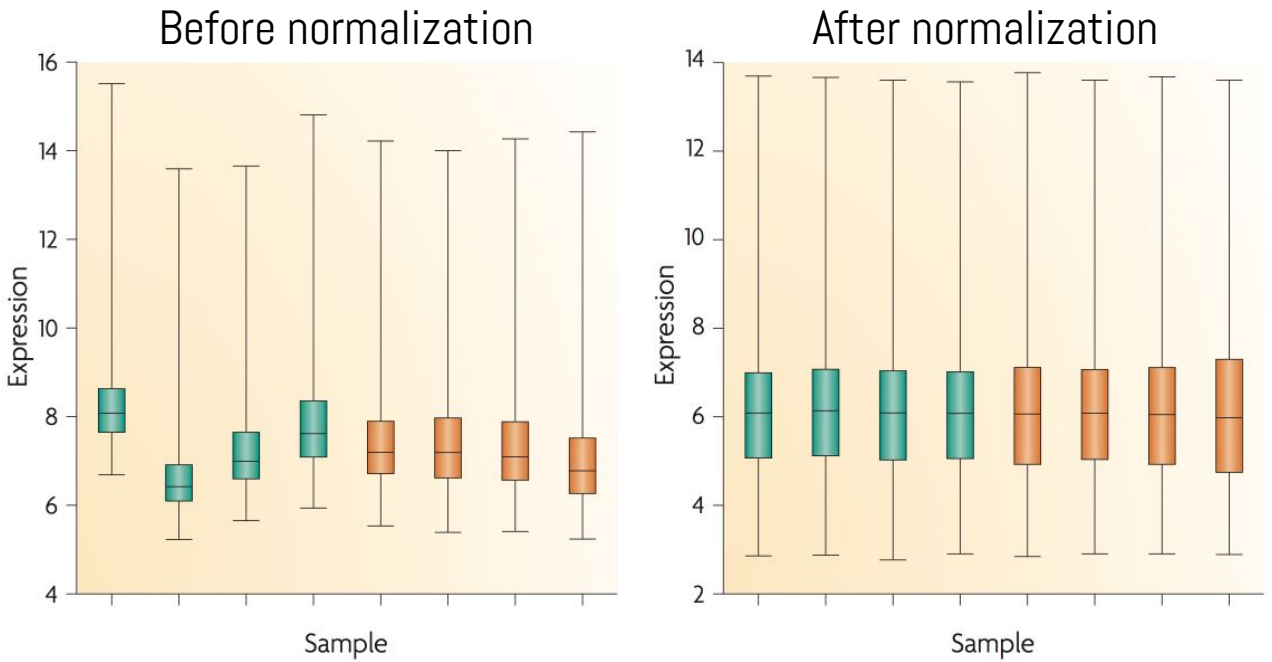
Replication & Pseudoreplication

- 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

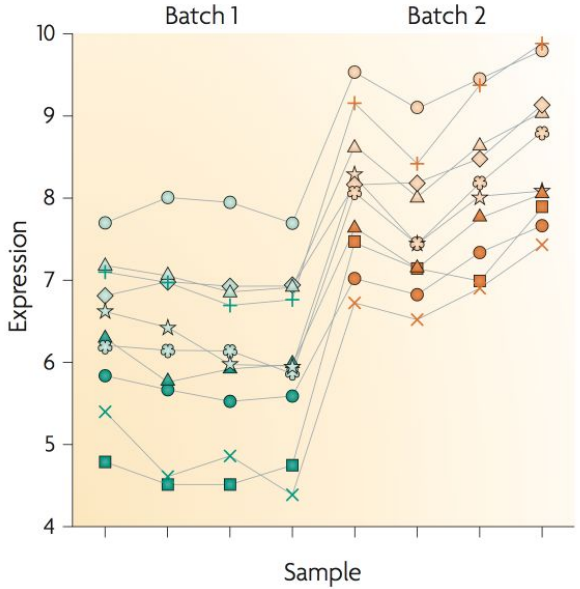
Extraneous variables (e.g. processing data) can be *confounded* with the outcome of interest (e.g. disease state) when it correlates both with the outcome and with an independent variable of interest (e.g. gene expression).



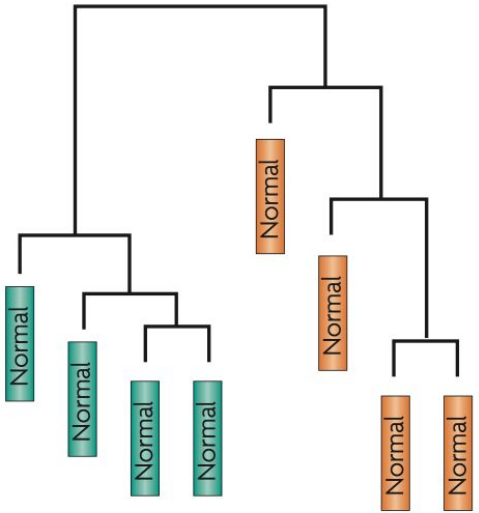
Confounding variables

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).

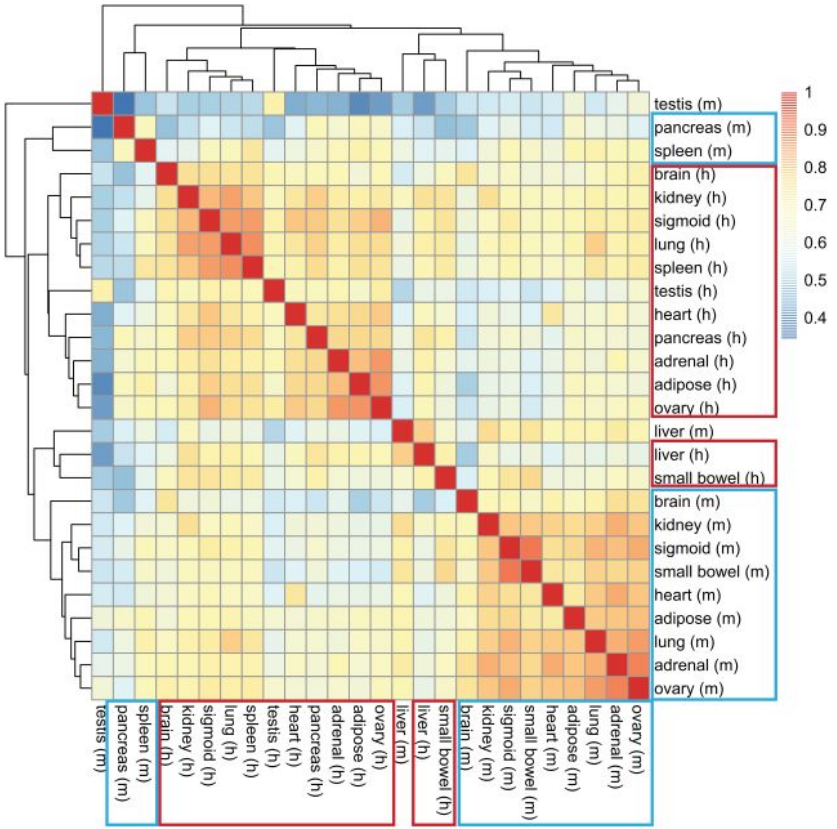
After normalization



Clustering of samples



Confounding variables



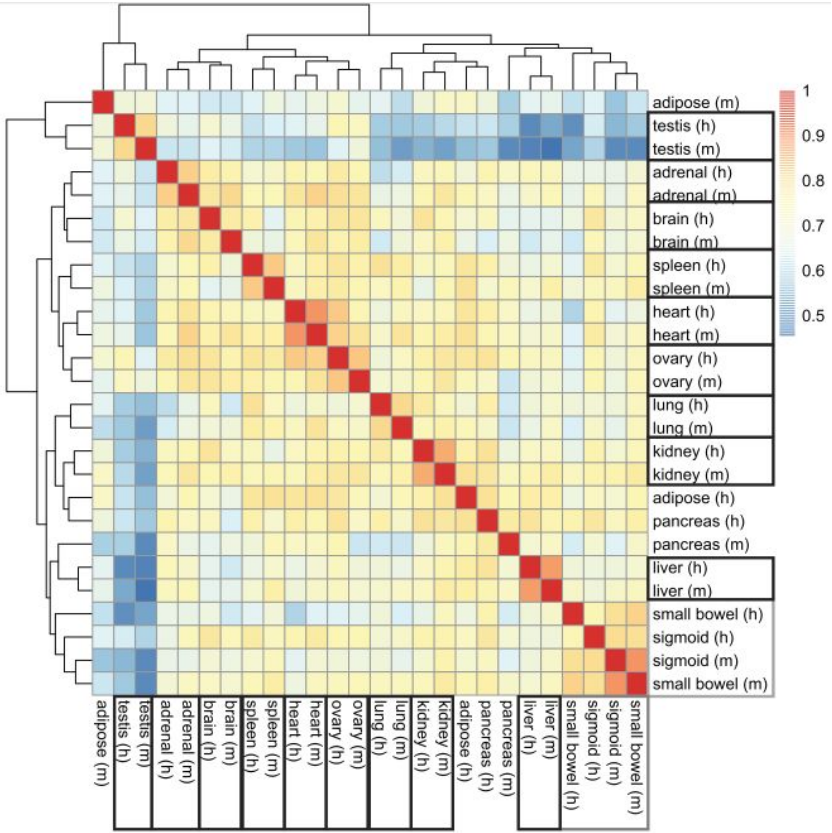
Mouse ENCODE comparative gene expression data

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

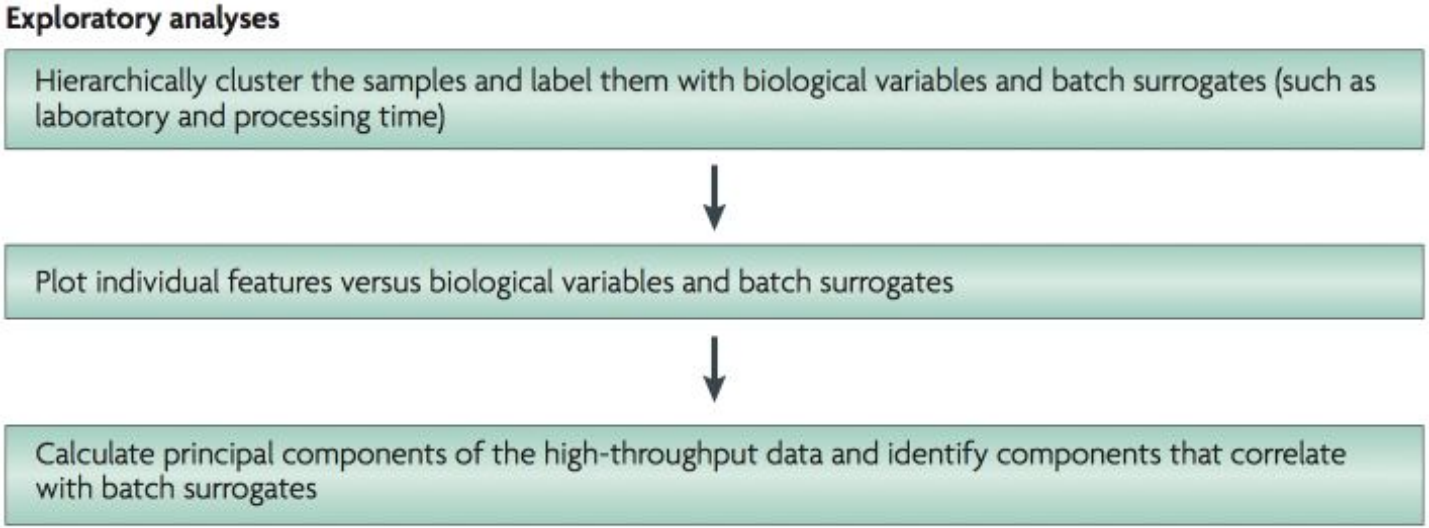


Re-analysis of the data after correcting for batch-effects.

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

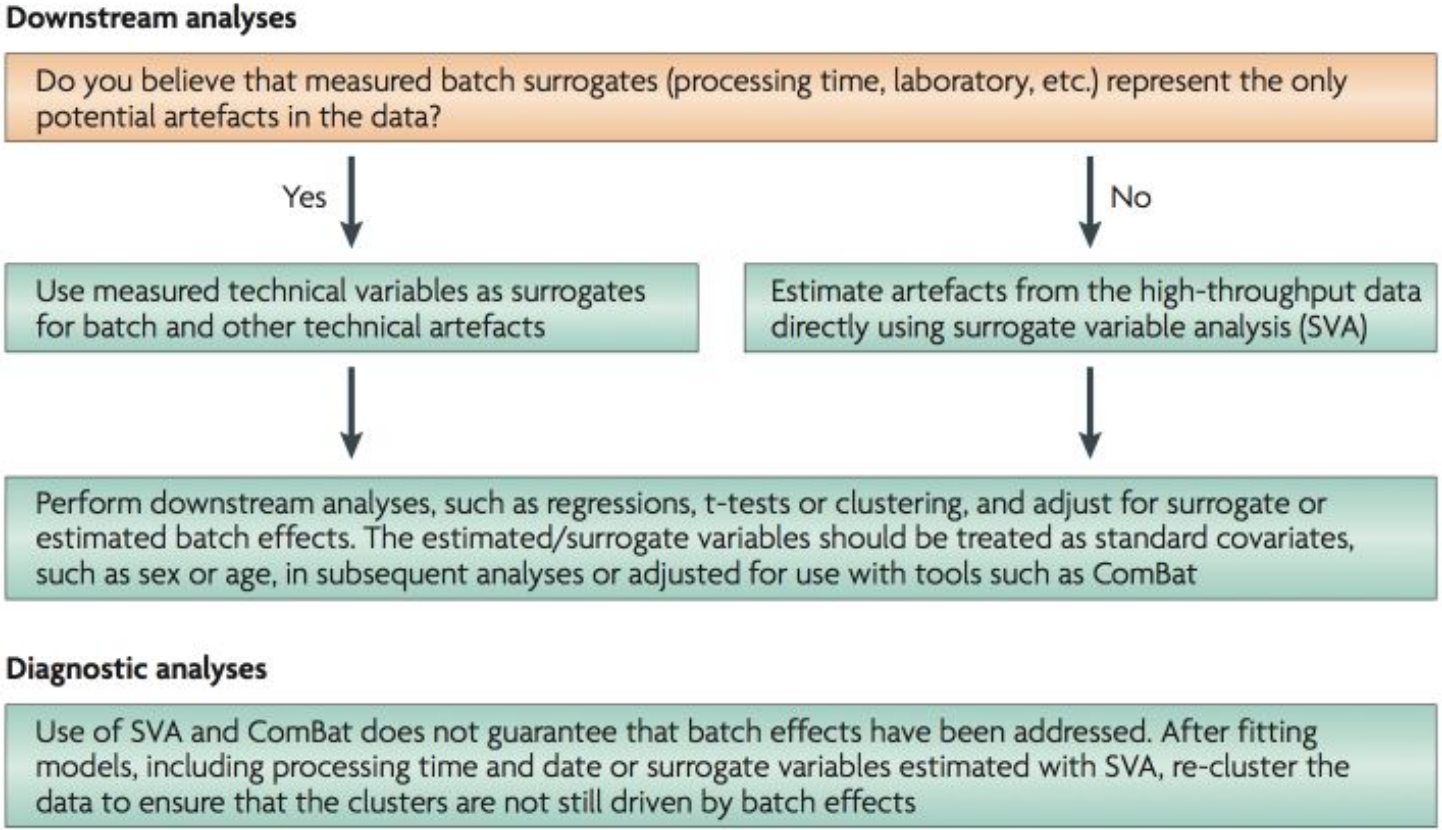
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4516019/

Confounding variables



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

Confounding variables



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

Topic 3: Circularity & Regression

Lectures 6 & 7

- Circular analysis
- Regression to the mean
- Stopping rules