Statistical Methods for High Dimensional Biology

STAT/BIOF/GSAT 540

Confounding & batch effects

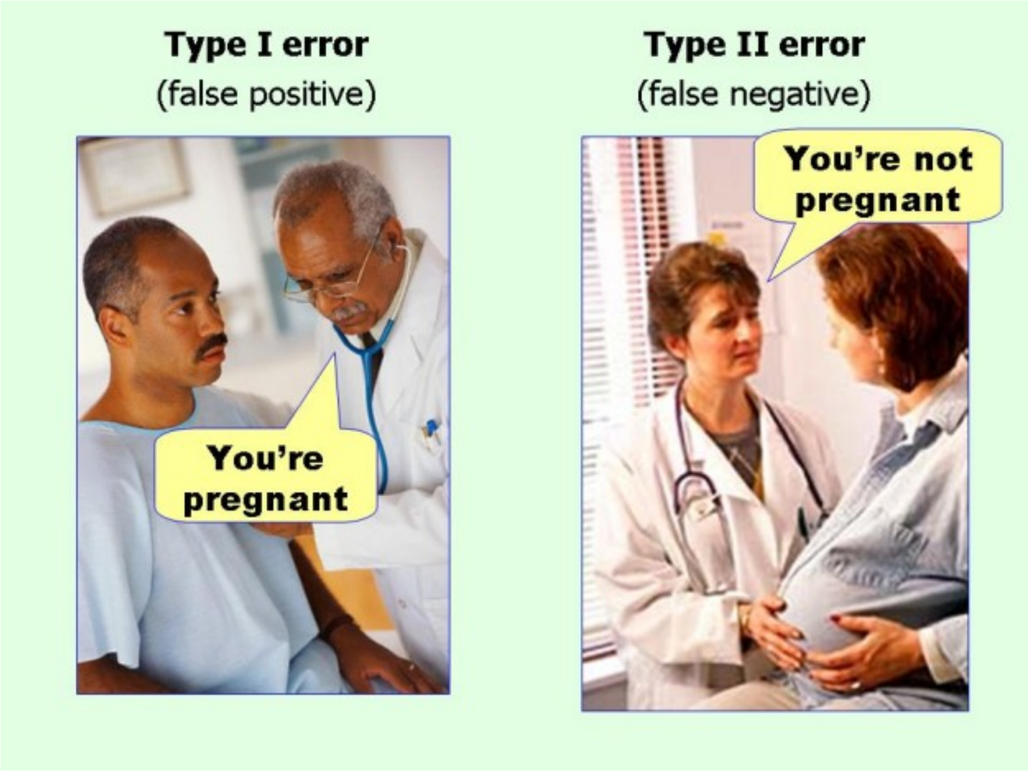
Keegan Korthauer

10 February 2021

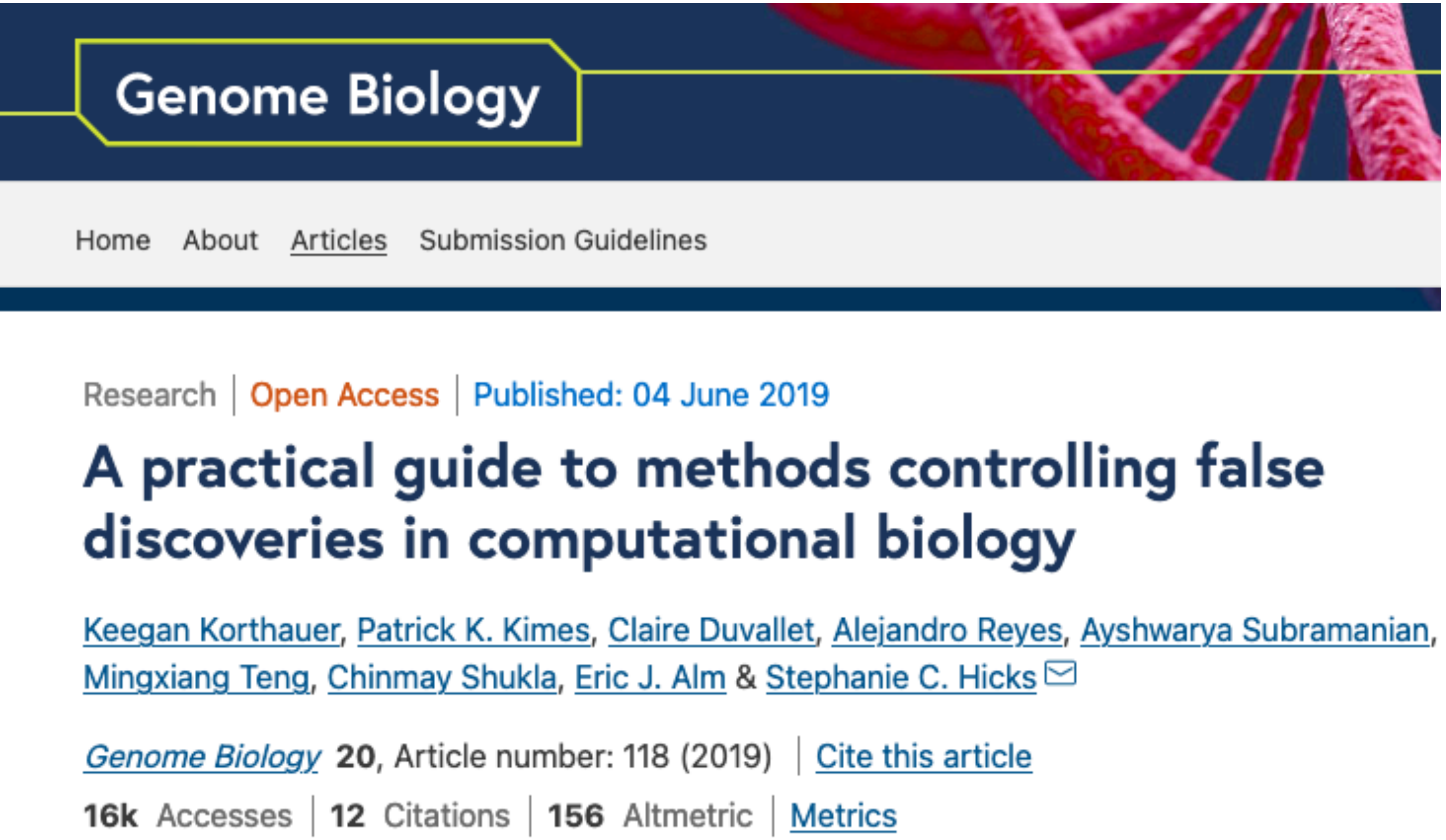
With slide contributions from Sara Mostafavi, Jenny Bryan, Su-In Lee, and Doug Fowler

Recall: adjusting for multiple comparisons

- Need to control the rate of false positives
- Trade-off between Type I error and Type II error
  - Classical statistics: control FWER (probability at least 1error)
  - High-throughput biology: more common to control FDR (proportion of rejections that are false positives)
- False Discovery Rate control is an active area of research
  - Recent work: methods that use more than the p-value
    - e.g. p-value + effect size, or p-value + ‘independent covariate’



Example: recent evaluation of FDR methods



Genome Biology

Home About Articles Submission Guidelines

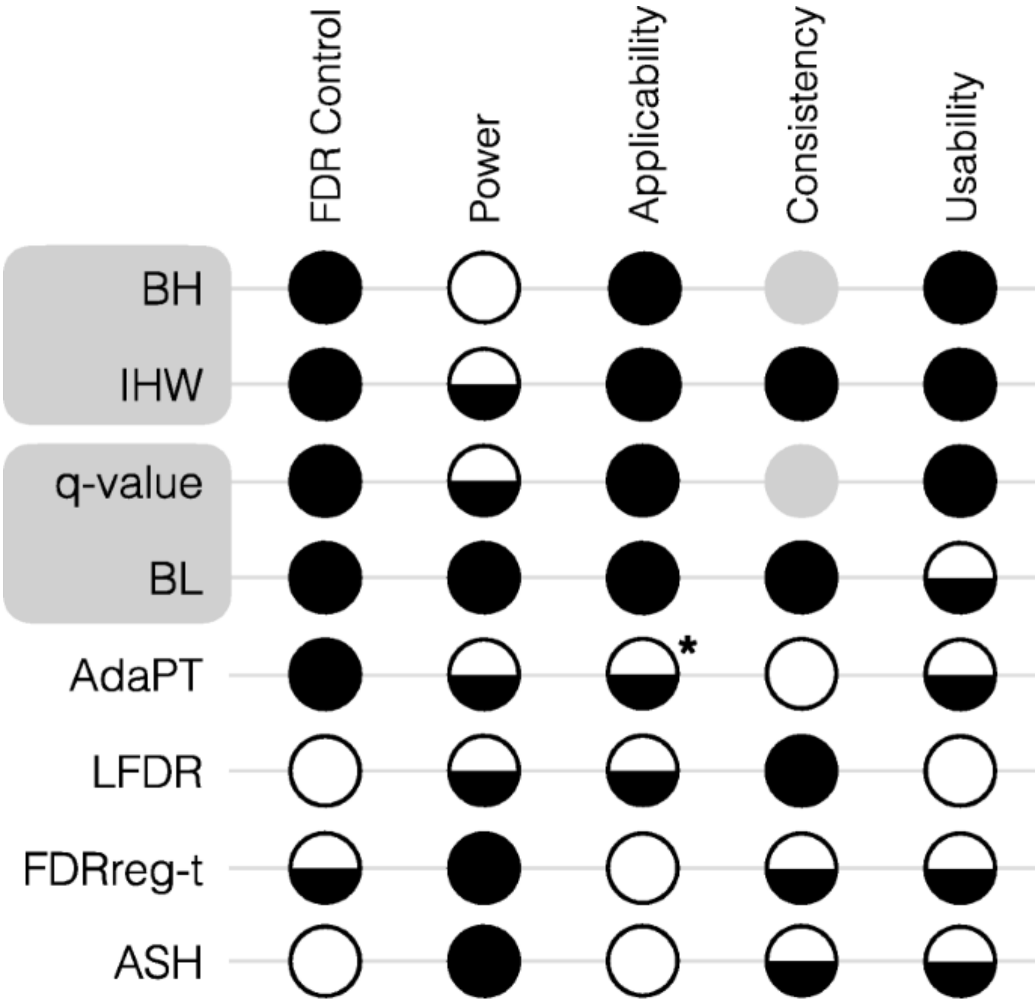
Research | Open Access | Published: 04 June 2019

A practical guide to methods controlling false discoveries in computational biology

Keegan Korthauer, Patrick K. Kimes, Claire Duvallet, Alejandro Reyes, Ayshwarya Subramanian, Mingxiang Teng, Chinmay Shukla, Eric J. Alm & Stephanie C. Hicks

Genome Biology **20**, Article number: 118 (2019) | Cite this article

16k Accesses | **12** Citations | **156** Altmetric | Metrics



	FDR Control	Power	Applicability	Consistency	Usability
BH	●	○	●	◌	●
IHW	●	◒	●	●	●
q-value	●	◒	●	◌	●
BL	●	●	●	●	◒
AdaPT	●	◒	◒*	○	◒
LFDR	○	◒	◒	●	○
FDRreg-t	◒	●	○	◒	◒
ASH	○	●	○	◒	◒

Figure 6 (Korthauer*, Kimes* et al. 2019)

Today's Learning objectives

- Understand how **confounding variables** may influence the observed association between a predictor and outcome
- Be able to assess the impact of measured confounders (e.g. batch)
- Apply general techniques to adjust for the impact of measured and unmeasured confounders, when possible
  - e.g. ComBat (measured), SVA (unmeasured)
- Understand how **normalization** can be seen as a specific example of adjusting for confounding variation

Confounding

- **Confounding**: a situation in which a measure of association or relationship between response and explanatory variables is distorted by presence of another variable

- **Confounder**: an extraneous or external variable that wholly or partially accounts for your observed effect

Example: 9 month weight gain from pickles?

shutterstock.com · 1400121137

Example: deadly ice cream?

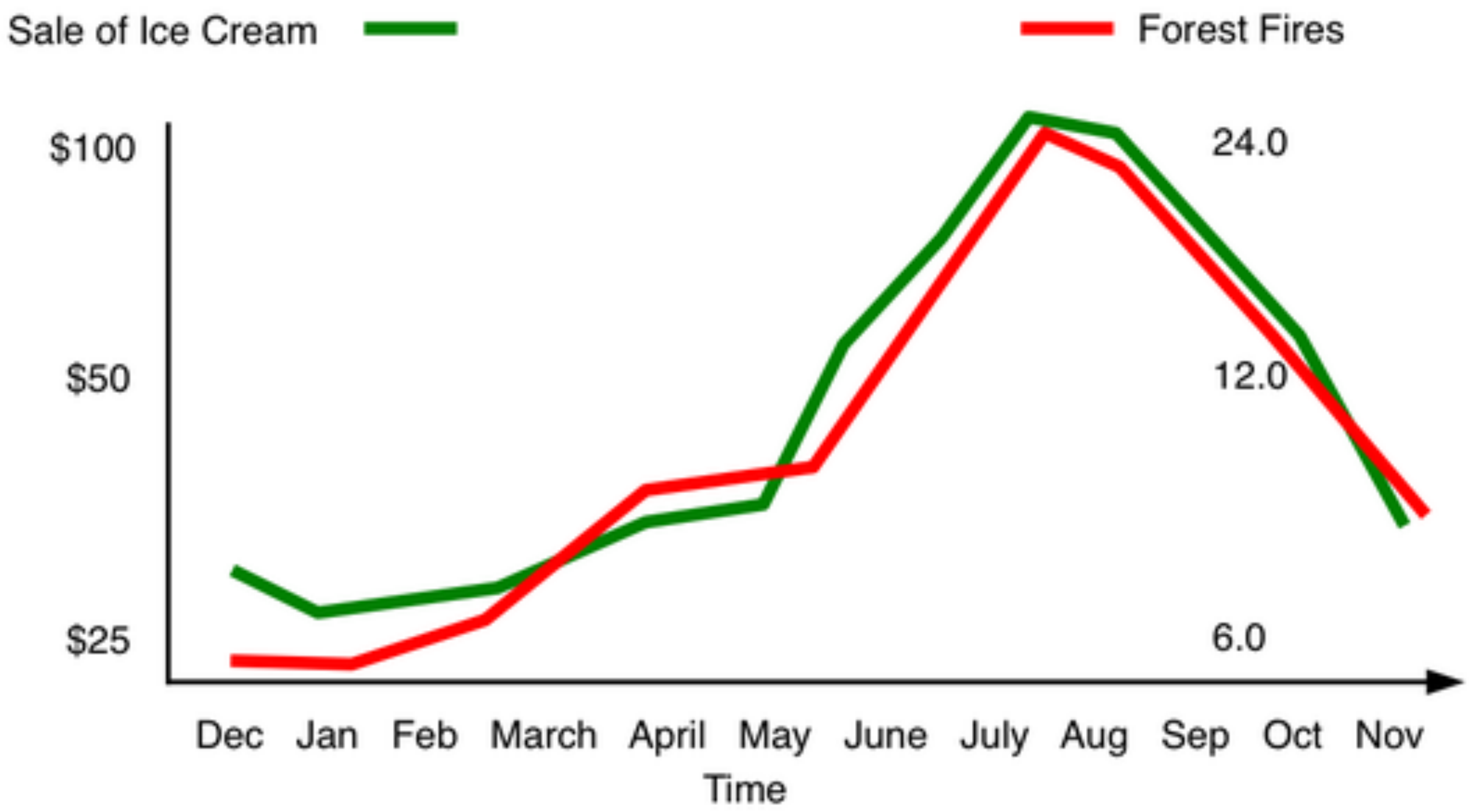


image source: https://www.decisionskills.com/blog/how-ice-cream-kills-understanding-cause-and-effect

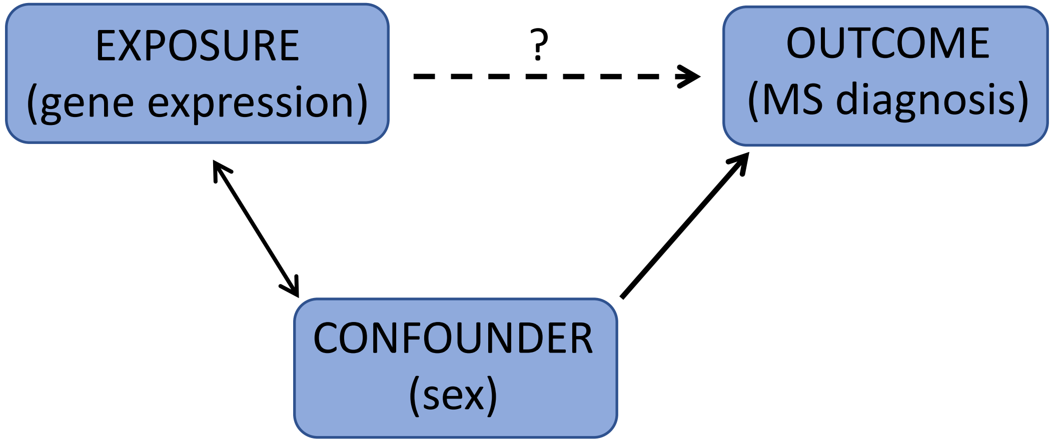
Example: Multiple Sclerosis

- MS roughly 3 times more prevalent in females vs males
- Hypothetical study: differential gene expression cases vs controls
  - Cases: females with MS
  - Controls: males without MS
- What do you expect to find in gene expression analysis?

Definition of a confounder

For a variable to be a confounder it should meet three conditions:

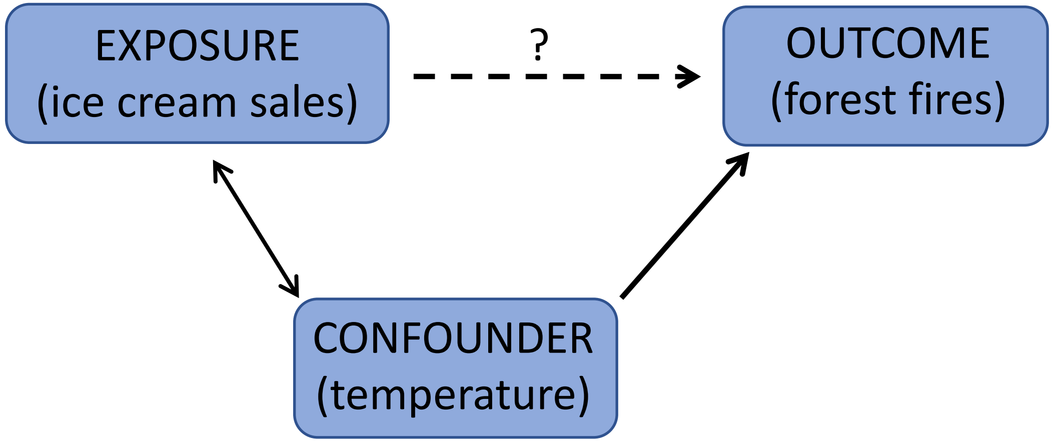
1. The factor must be associated with the exposure being investigated
2. Must be independently associated with the outcome being investigated
3. Not be in the causal pathway between exposure and outcome



Definition of a confounder

For a variable to be a confounder it should meet three conditions:

1. The factor must be associated with the exposure being investigated
2. Must be independently associated with the outcome being investigated
3. Not be in the causal pathway between exposure and outcome



Confounding factors in genomics studies

- **Observational studies**:
  - Independent variable is not under the control of the researcher (e.g., ethical reasons)
    - e.g., case/control study: which subjects are case and which are controls are out of the control of the investigator
  - **Selection bias:** typically many variables/factors are correlated with the independent variable of interest
- **Interventional studies**:
  - e.g., randomized study: investigator can randomly assign individuals to groups and so control the assignment of the independent variable
  - Minimizes the **selection bias** problem

Example: Observational study of MS

Factors that may be associated with gene expression & differ between cases and controls

- Age
- Sex
- Smoking status
- Medication intake
- Latitude
- ….

Simpson’s paradox

What effect does the drug have on response?

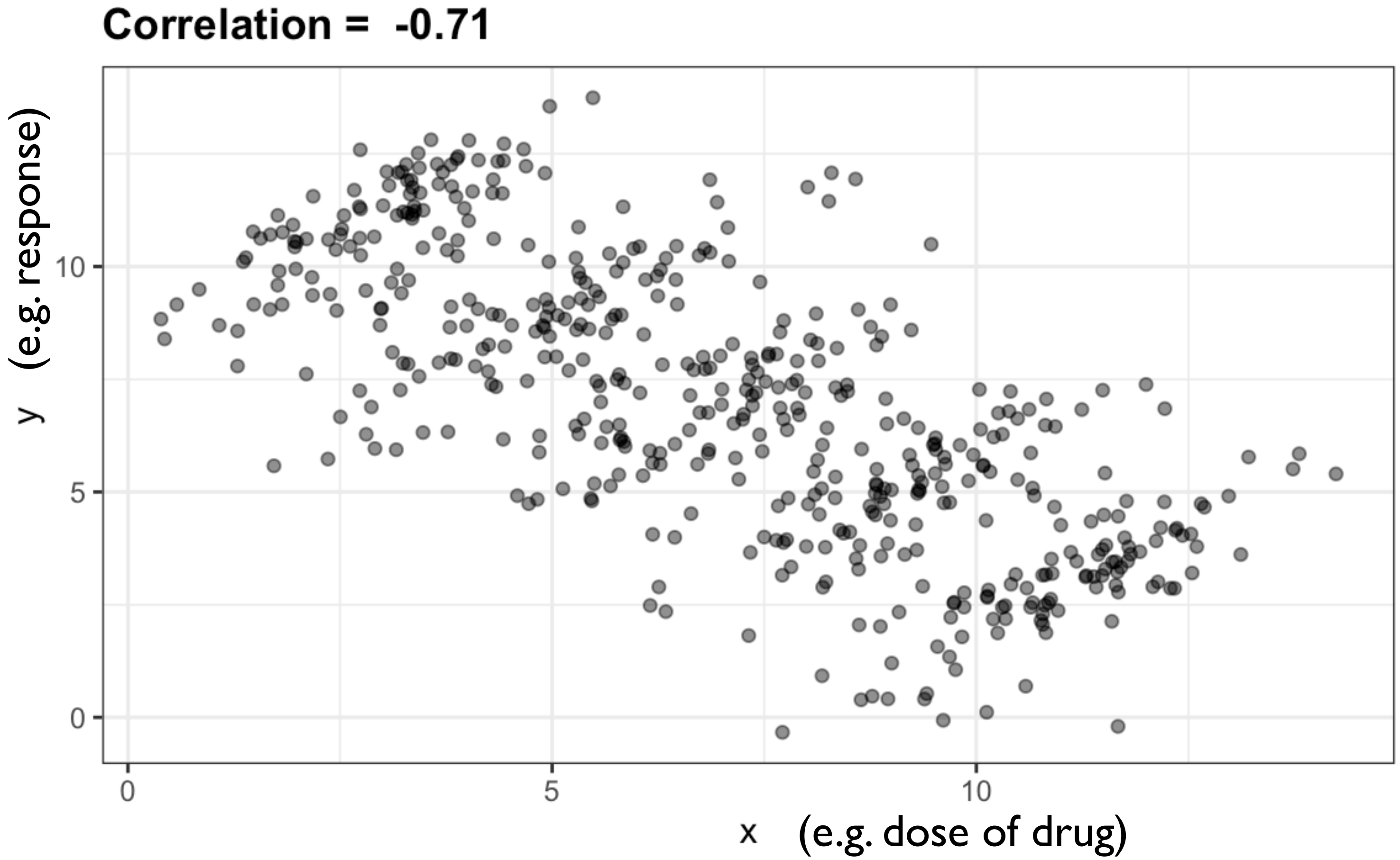


Image source: https://rafalab.github.io/dsbook/association-is-not-causation.html

Simpson’s paradox

What effect does the drug have on response in each group Z?

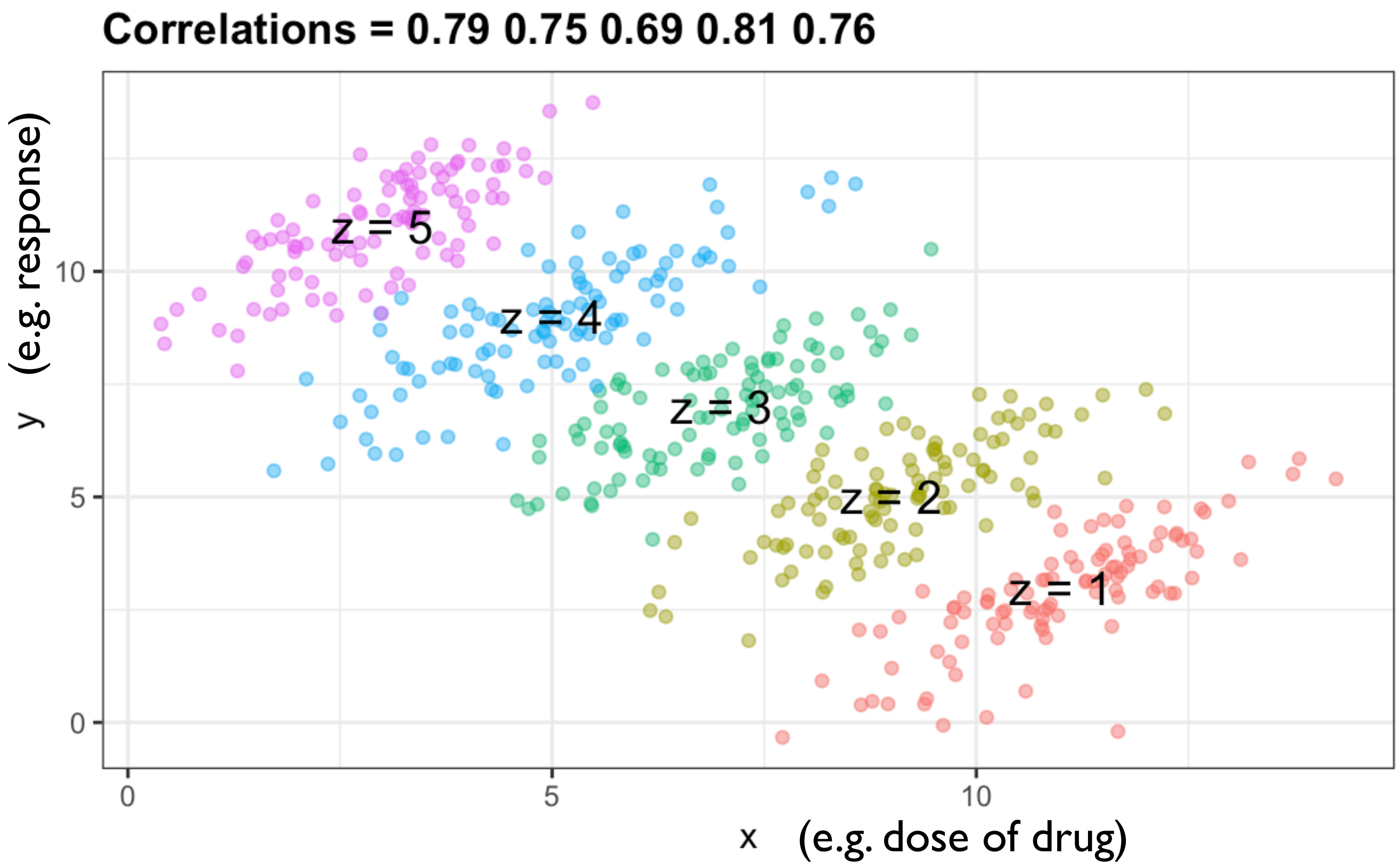
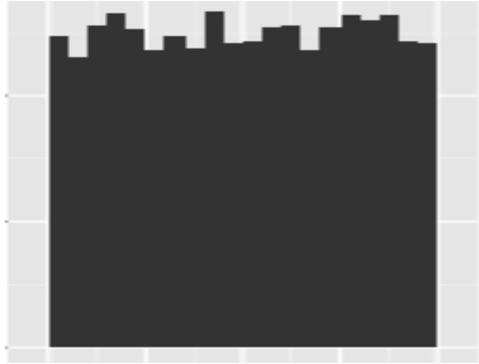


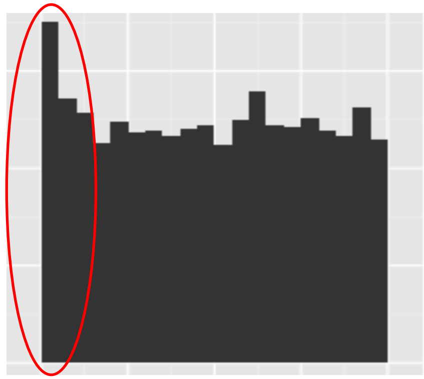
Image source: https://rafalab.github.io/dsbook/association-is-not-causation.html

Diagnosing potential confounding effects

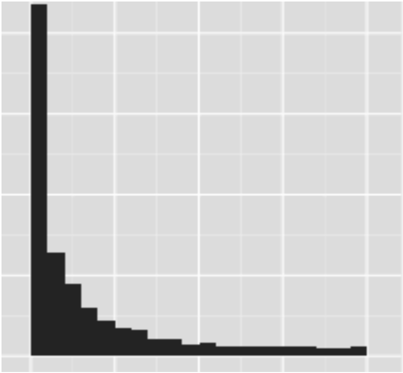
P-value distribution under the null should look uniform



Spike of small p-values near zero are potential non-nulls



Too much signal should (usually) alarm you



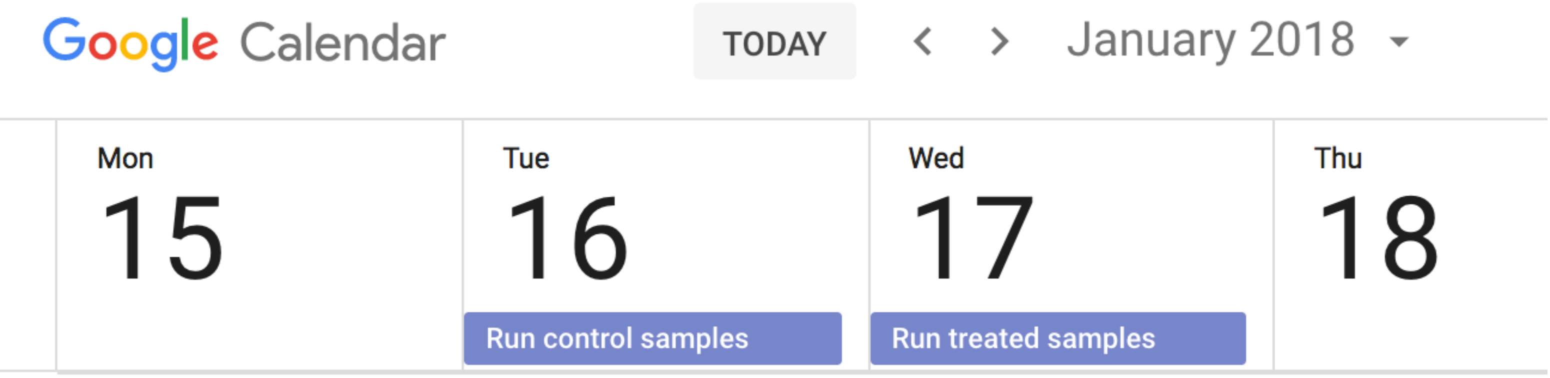
More ways things can go wrong: http://varianceexplained.org/statistics/interpreting-pvalue-histogram/

Types of confounding

- Biological heterogeneity:
  - e.g., sex, age, …
- Environmental heterogeneity:
  - e.g., smoking status, alcohol use, diet, …
- Genetic heterogeneity:
  - e.g., population stratification
- Experimental heterogeneity (batch effects/systematic artifacts):
  - e.g., processing date, technician, probe position, sequencing lane, library size

Batch effects are a huge problem in genomics

- Generation of data depends on: complicated reagents + software used by highly trained personnel
- If some of these conditions vary in the course of experiment: measurements for MANY genes/features will be affected
  - e.g., controls were run on Tuesday and treated samples on Wednesday



Avoid this type of design (from the Chd8 study discussed in Lecture 3)

```
> table(m$SeqRun,m$DPC)

    12.5 14.5 17.5 21 77
  A    8    0    0  0  0
  B    0    9    0  0  0
  C    0    0    5  0  0
  D    0    0    5  0  0
  E    0    0    0 11  0
  H    0    0    0  0  6
```

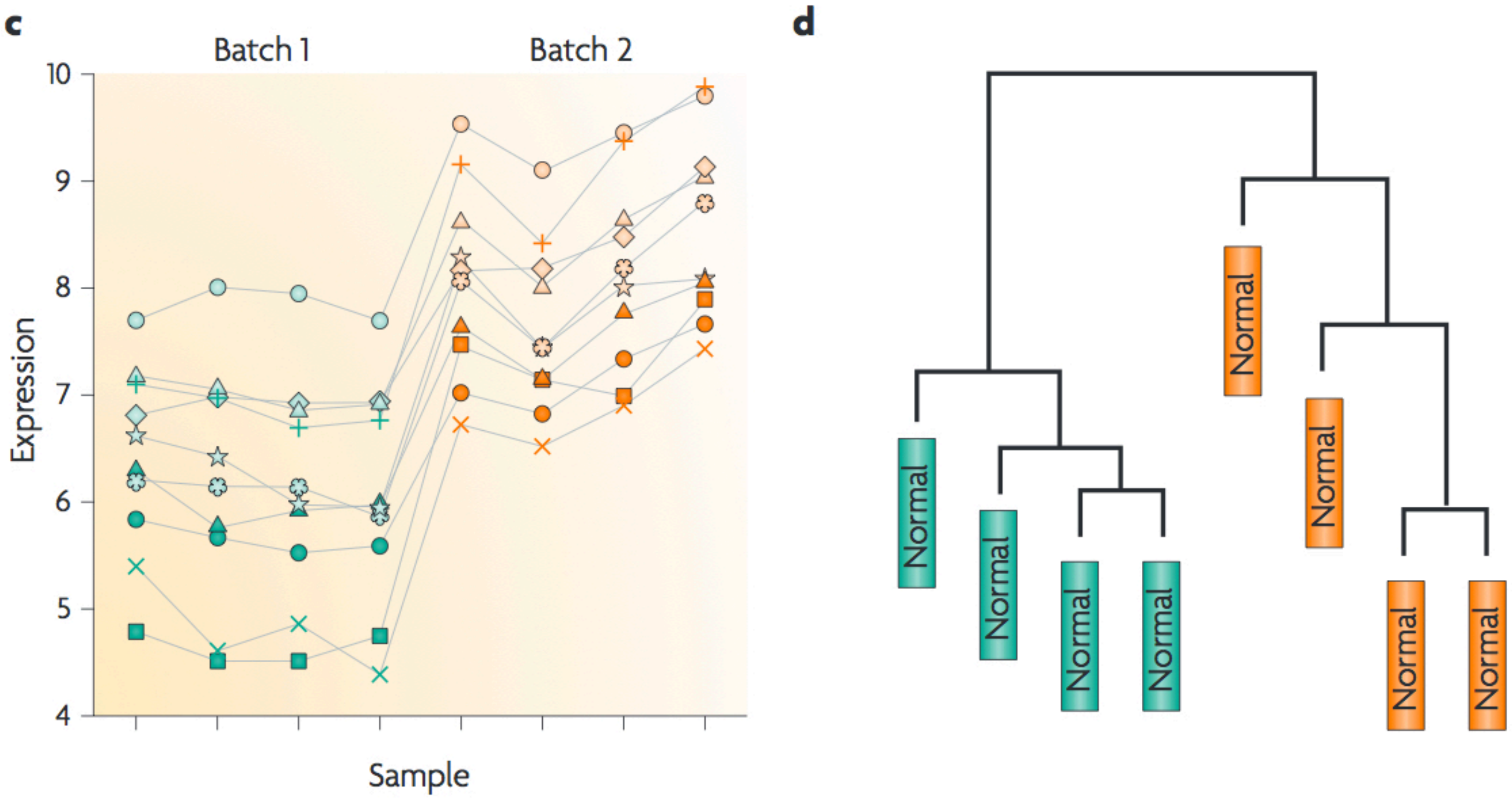
“Batch effects are sub-groups of measurements that have qualitatively different behaviour across conditions and are unrelated to the biological or scientific variables in a study”

Leek et al. 2010 Nature Rev. Genetics 11:733

Opinion

Tackling the widespread and critical impact of batch effects in high-throughput data

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly & Rafael A. Irizarry



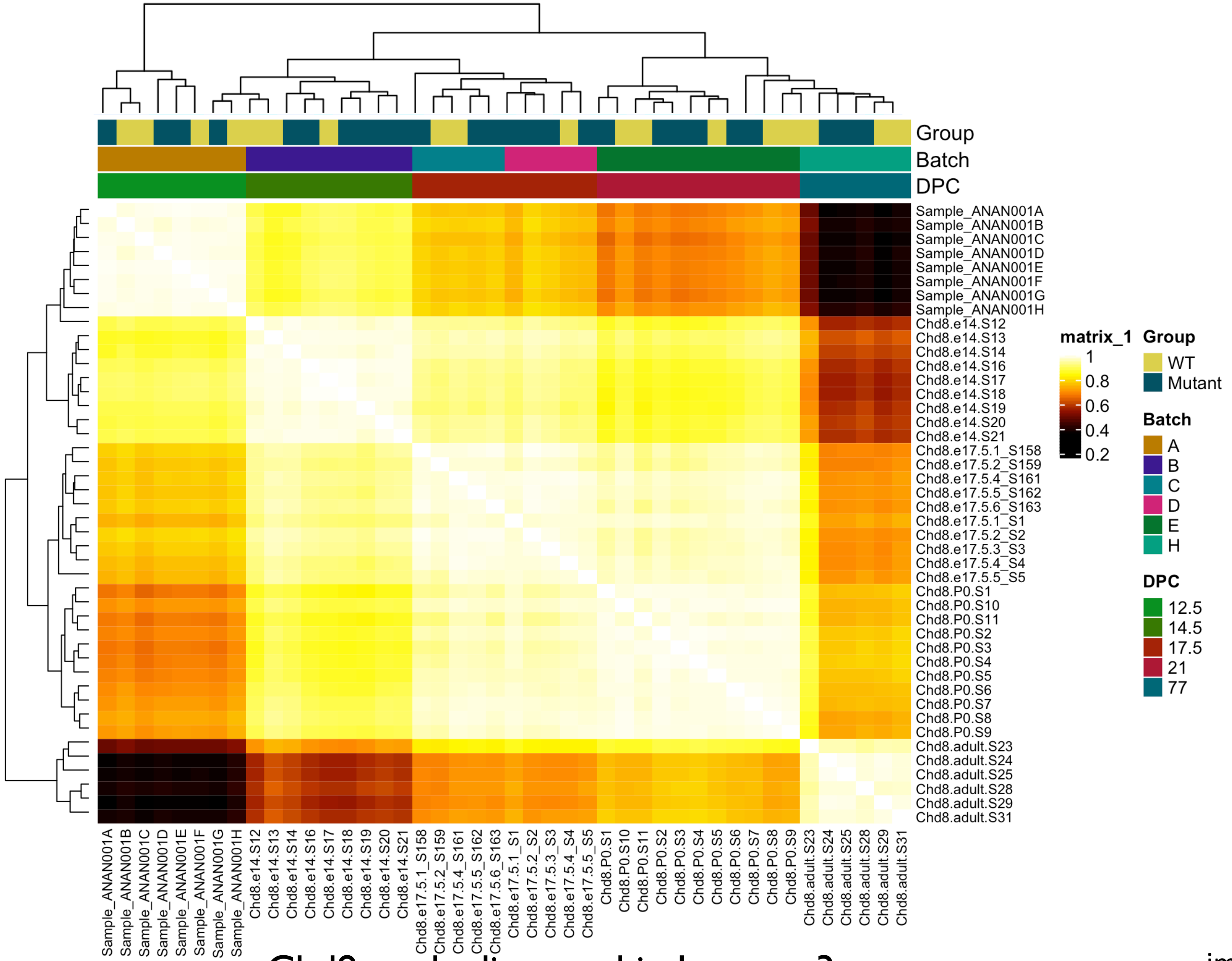
Consequences of batch effects

- Reduced statistical power: **false negatives**

- Confounding / spurious associations: **false positives**

Visualizing batch effects

Sample-sample covariance matrix (clustering)



Chd8 study discussed in Lecture 3

Principal Component Analysis (dimensionality reduction) – more on this in a later lecture

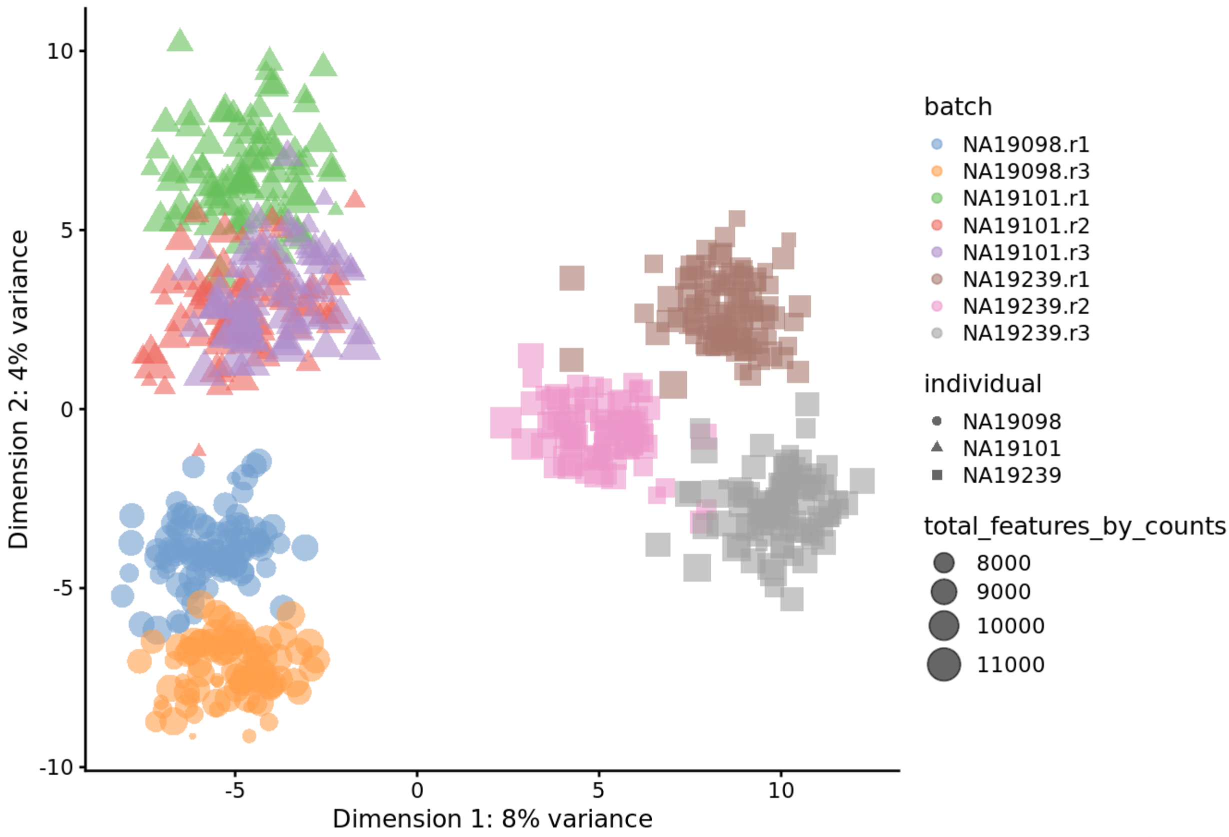
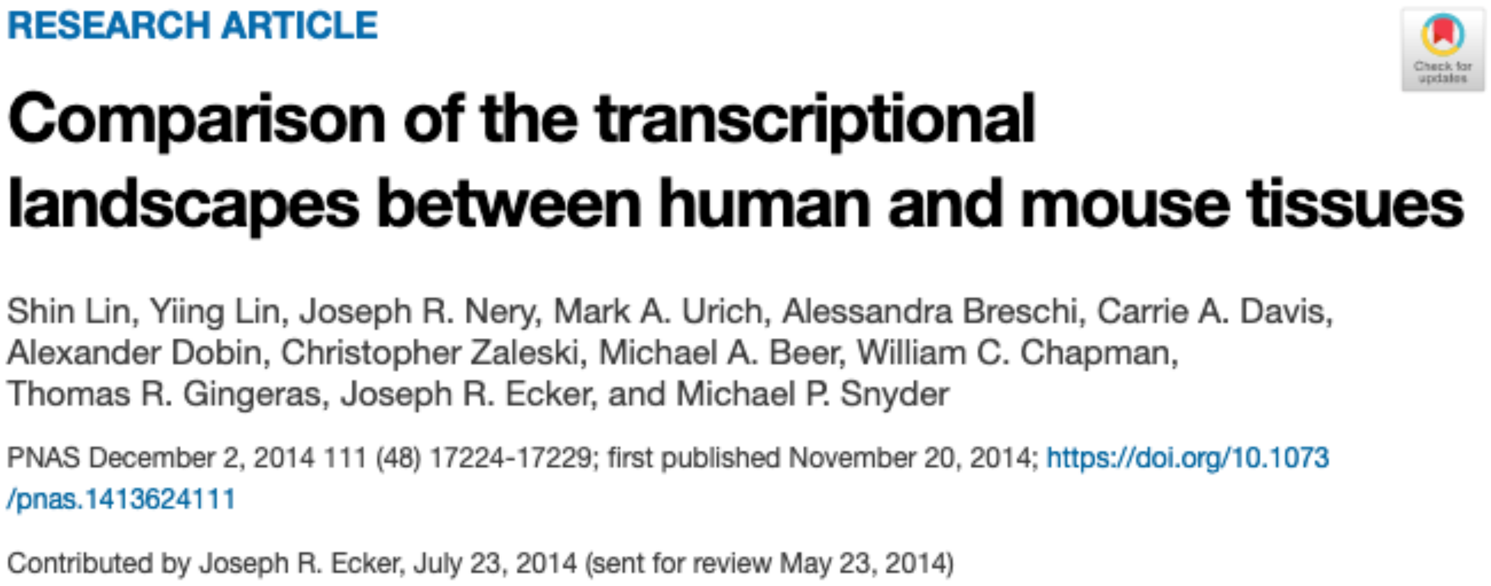


image source: https://biocellgen-public.svi.edu.au/mig_2019_scrnaseq-workshop/public/

Case study: tissue vs organism differences



RESEARCH ARTICLE

Comparison of the transcriptional landscapes between human and mouse tissues

Shin Lin, Yiing Lin, Joseph R. Nery, Mark A. Urich, Alessandra Breschi, Carrie A. Davis,
Alexander Dobin, Christopher Zaleski, Michael A. Beer, William C. Chapman,
Thomas R. Gingeras, Joseph R. Ecker, and Michael P. Snyder

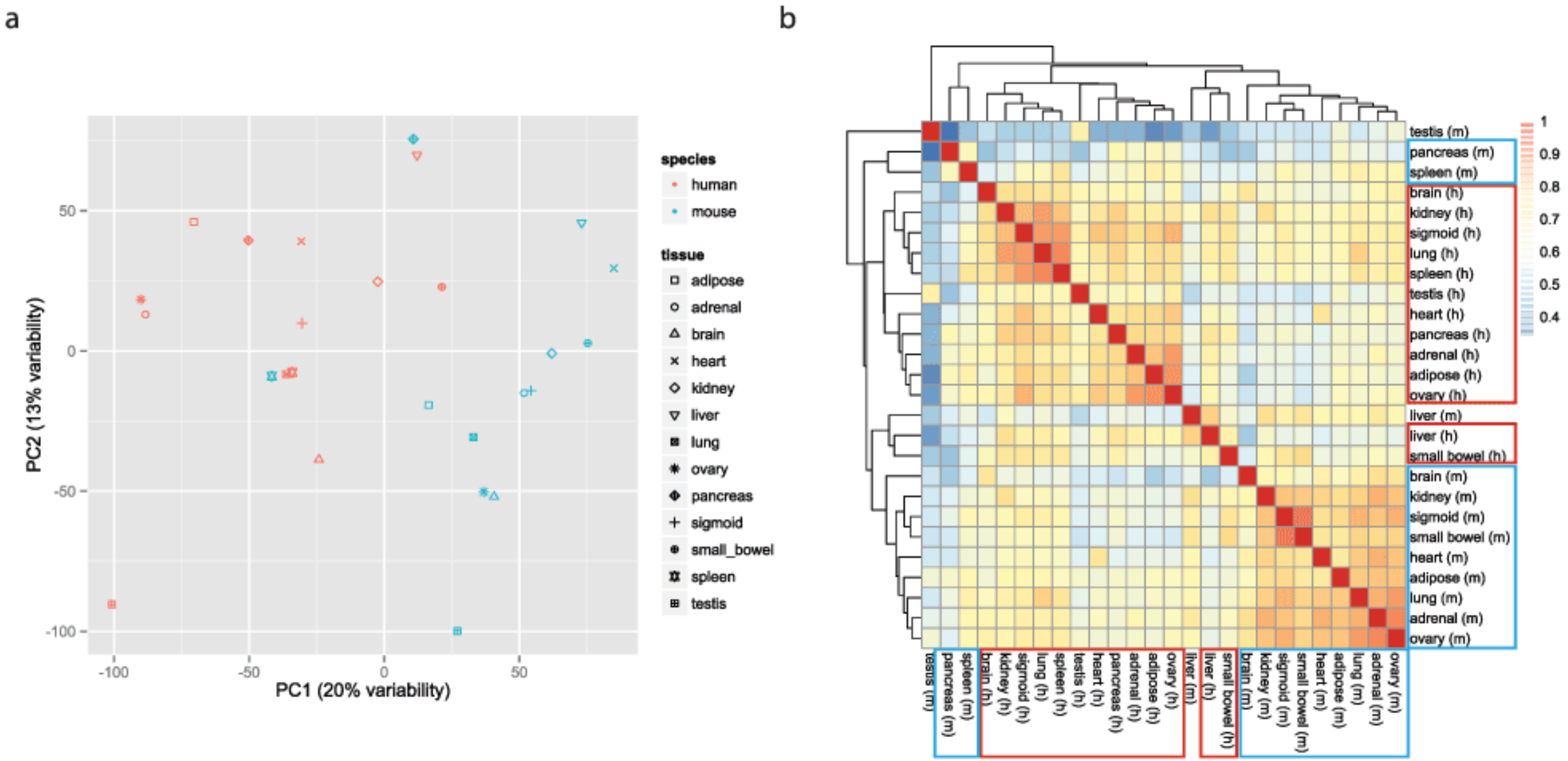
PNAS December 2, 2014 111 (48) 17224-17229; first published November 20, 2014; https://doi.org/10.1073/pnas.1413624111

Contributed by Joseph R. Ecker, July 23, 2014 (sent for review May 23, 2014)

Although commonalities are evident in the expression of tissue-specific genes between the two species, the expression for many sets of genes was found to be **more similar in different tissues within the same species than between species.**

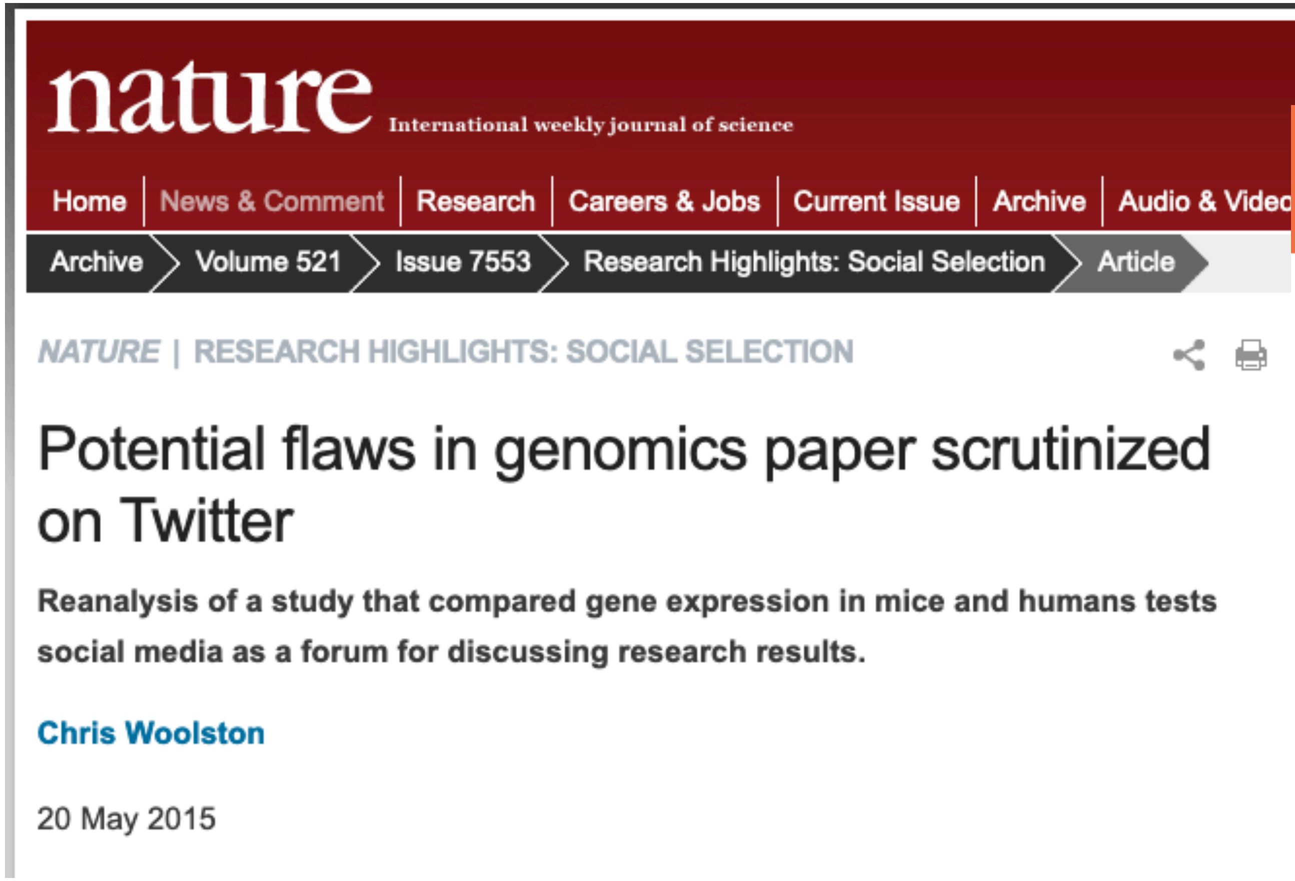
https://www.pnas.org/content/111/48/17224/

Findings of the original study (Lin et al. 2014)



Gilad Y and Mizrahi-Man O. A reanalysis of mouse ENCODE comparative gene expression data. F1000Research 2015, 4:121
https://f1000research.com/articles/4-121/v1

Others looked closer at the surprising results



nature
International weekly journal of science
Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video
Archive > Volume 521 > Issue 7553 > Research Highlights: Social Selection > Article

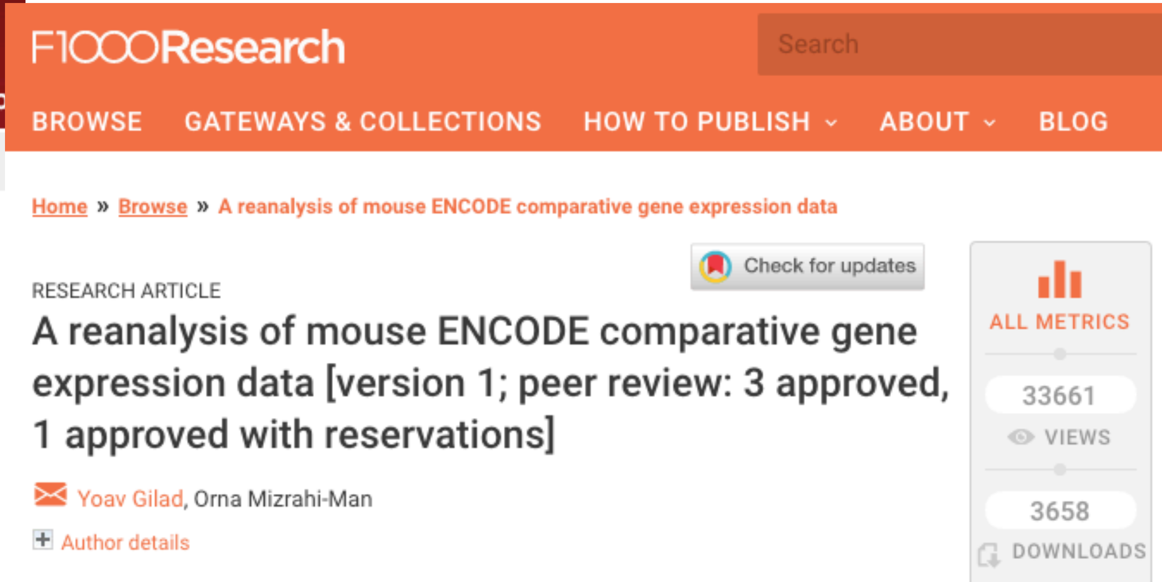
NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Potential flaws in genomics paper scrutinized on Twitter

Reanalysis of a study that compared gene expression in mice and humans tests social media as a forum for discussing research results.

Chris Woolston

20 May 2015



F1000Research
Search
BROWSE GATEWAYS & COLLECTIONS HOW TO PUBLISH ABOUT BLOG
Home » Browse » A reanalysis of mouse ENCODE comparative gene expression data
Check for updates
ALL METRICS
33661 VIEWS
3658 DOWNLOADS

RESEARCH ARTICLE

A reanalysis of mouse ENCODE comparative gene expression data [version 1; peer review: 3 approved, 1 approved with reservations]

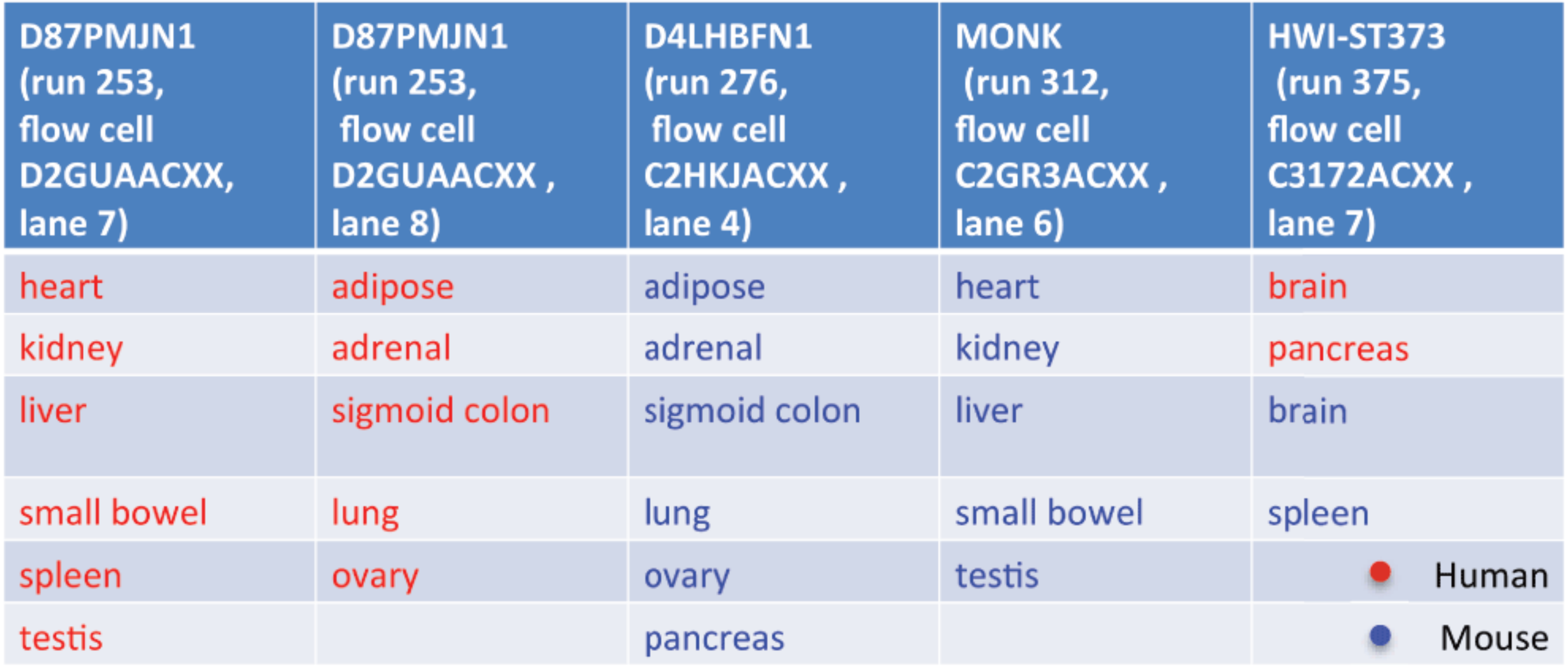
Yoav Gilad, Orna Mizrahi-Man

Author details

https://f1000research.com/articles/4-121/v1

https://www.nature.com/news/potential-flaws-in-genomics-paper-scrutinized-on-twitter-1.17591

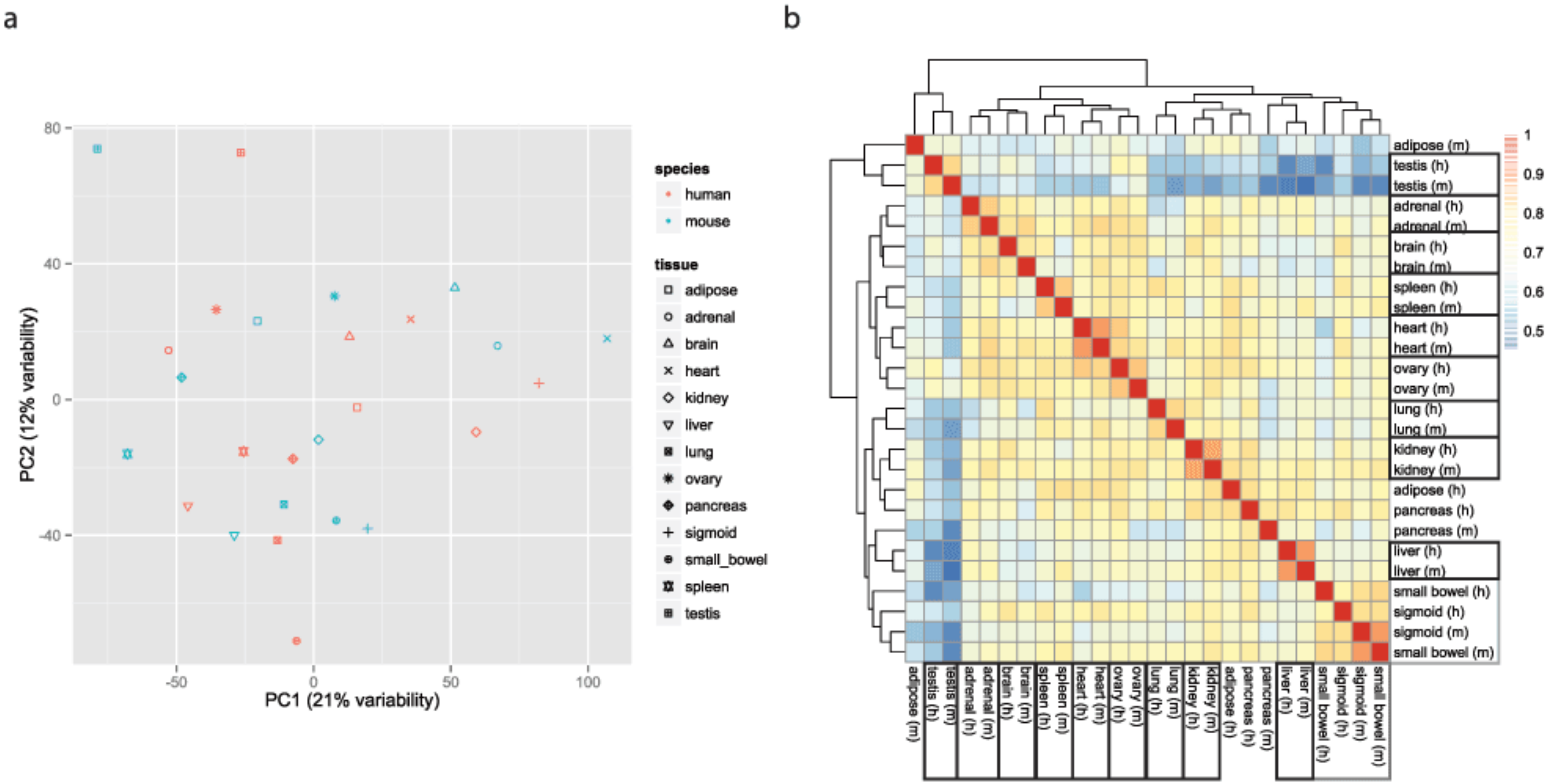
They found batch effects



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

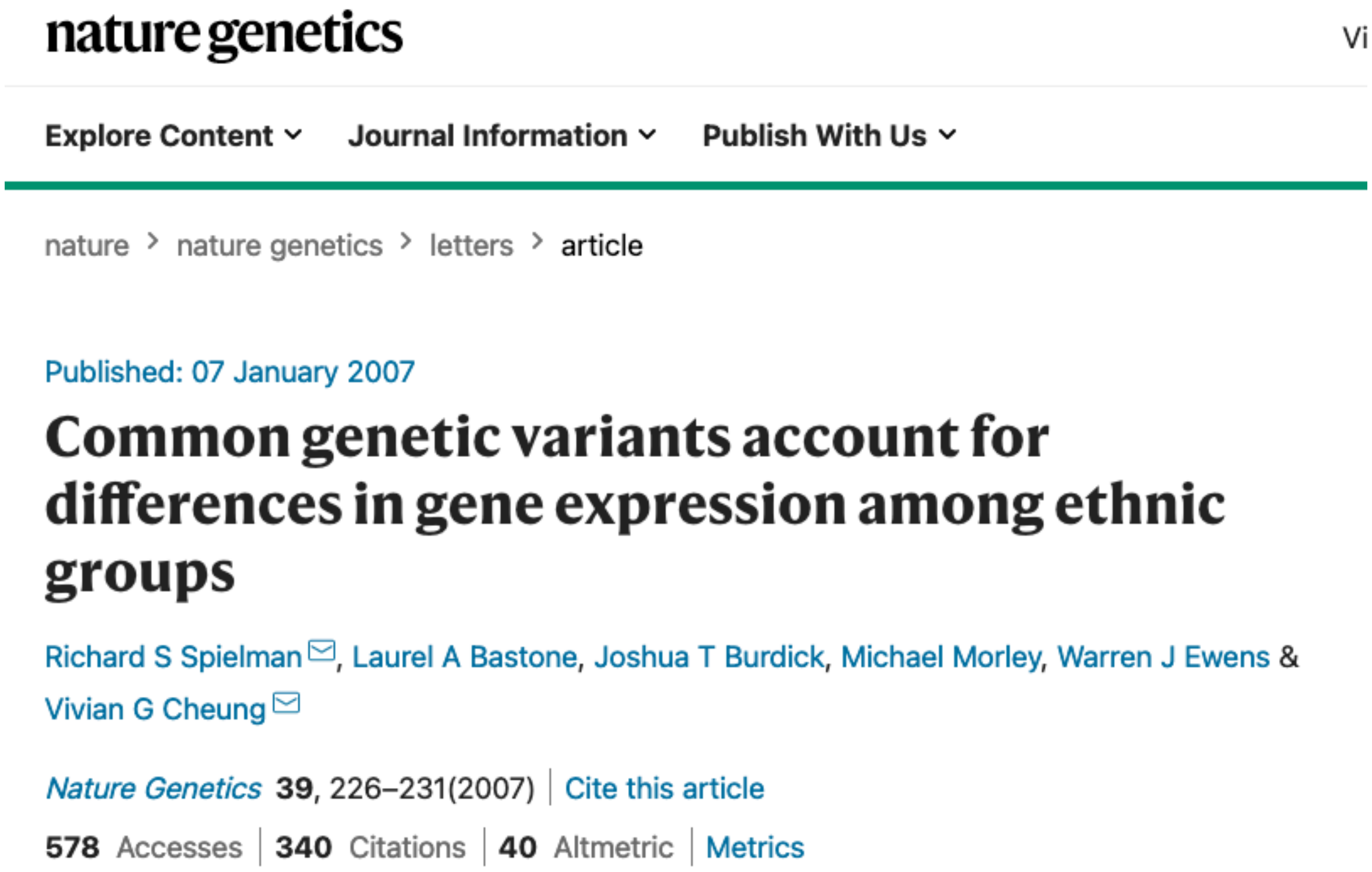
Gilad Y and Mizrahi-Man O. A reanalysis of mouse ENCODE comparative gene expression data. F1000Research 2015, 4:121
https://f1000research.com/articles/4-121/v1

Reanalysis adjusting for batch



Gilad Y and Mizrahi-Man O. A reanalysis of mouse ENCODE comparative gene expression data. F1000Research 2015, 4:121
https://f1000research.com/articles/4-121/v1

Case study: ancestry vs processing date



nature genetics   Vi

Explore Content ⌄   Journal Information ⌄   Publish With Us ⌄

nature › nature genetics › letters › article

Published: 07 January 2007

Common genetic variants account for differences in gene expression among ethnic groups

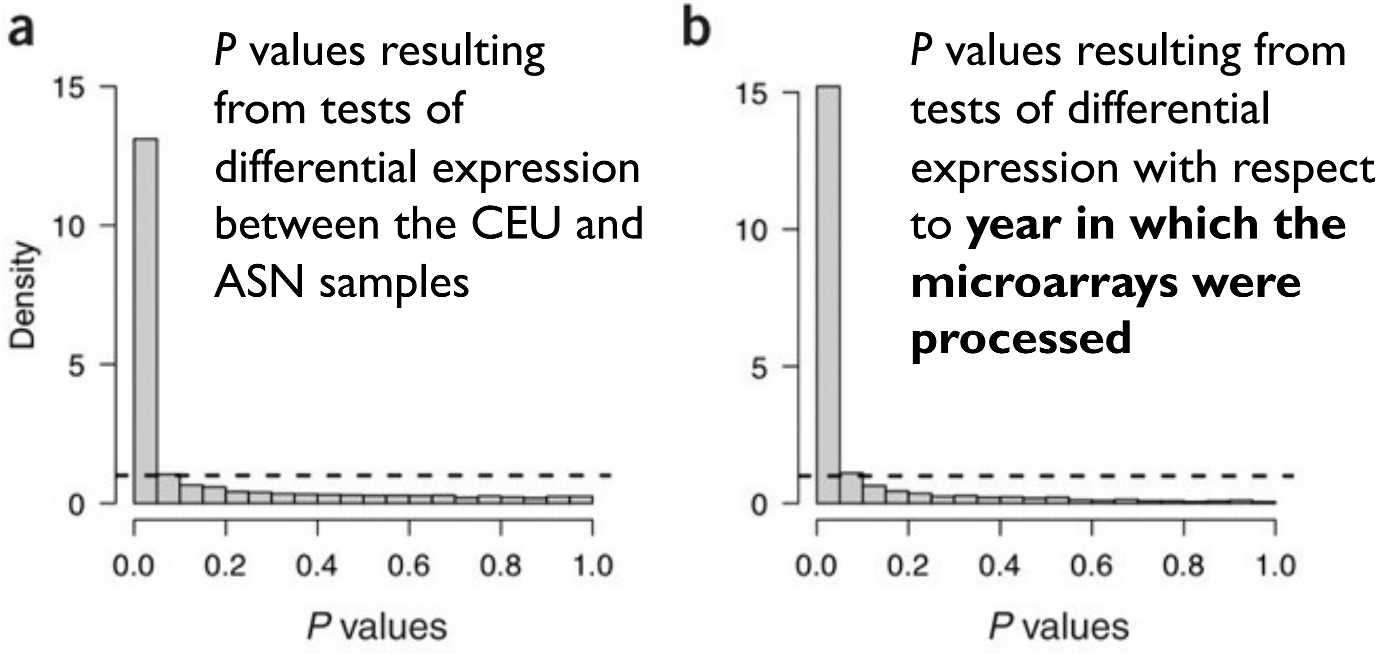
Richard S Spielman✉, Laurel A Bastone, Joshua T Burdick, Michael Morley, Warren J Ewens & Vivian G Cheung✉

Nature Genetics **39**, 226–231(2007) | Cite this article

578 Accesses | **340** Citations | **40** Altmetric | Metrics

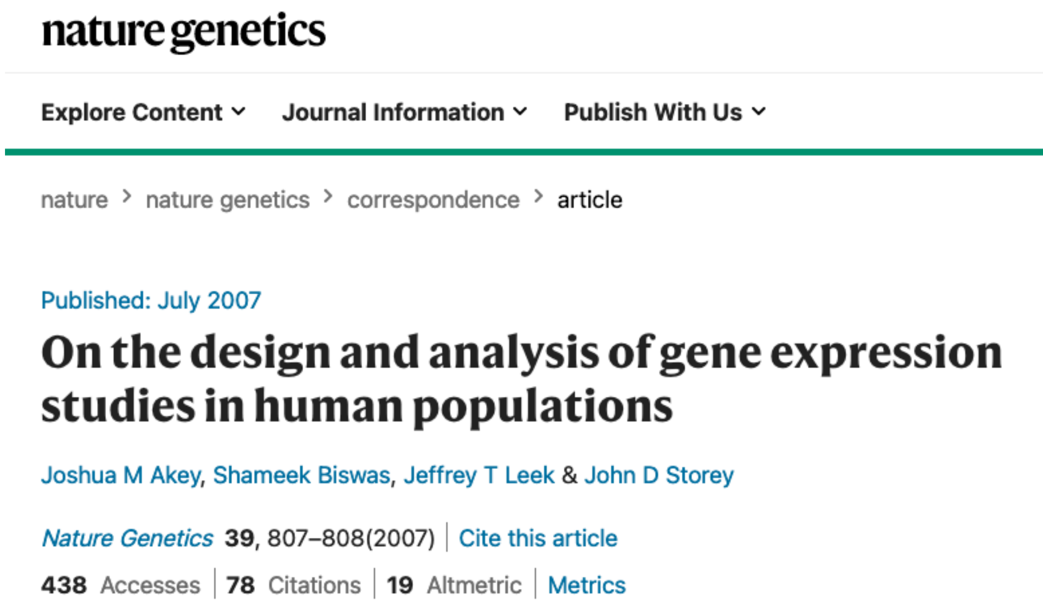
https://www.nature.com/articles/ng1955

Reanalysis by Akey et al. (2007):



P values resulting from tests of differential expression between the CEU and ASN samples

P values resulting from tests of differential expression with respect to **year in which the microarrays were processed**



nature genetics

nature › nature genetics › correspondence › article

Published: July 2007

On the design and analysis of gene expression studies in human populations

Joshua M Akey, Shameek Biswas, Jeffrey T Leek & John D Storey

Nature Genetics **39**, 807–808(2007) | Cite this article

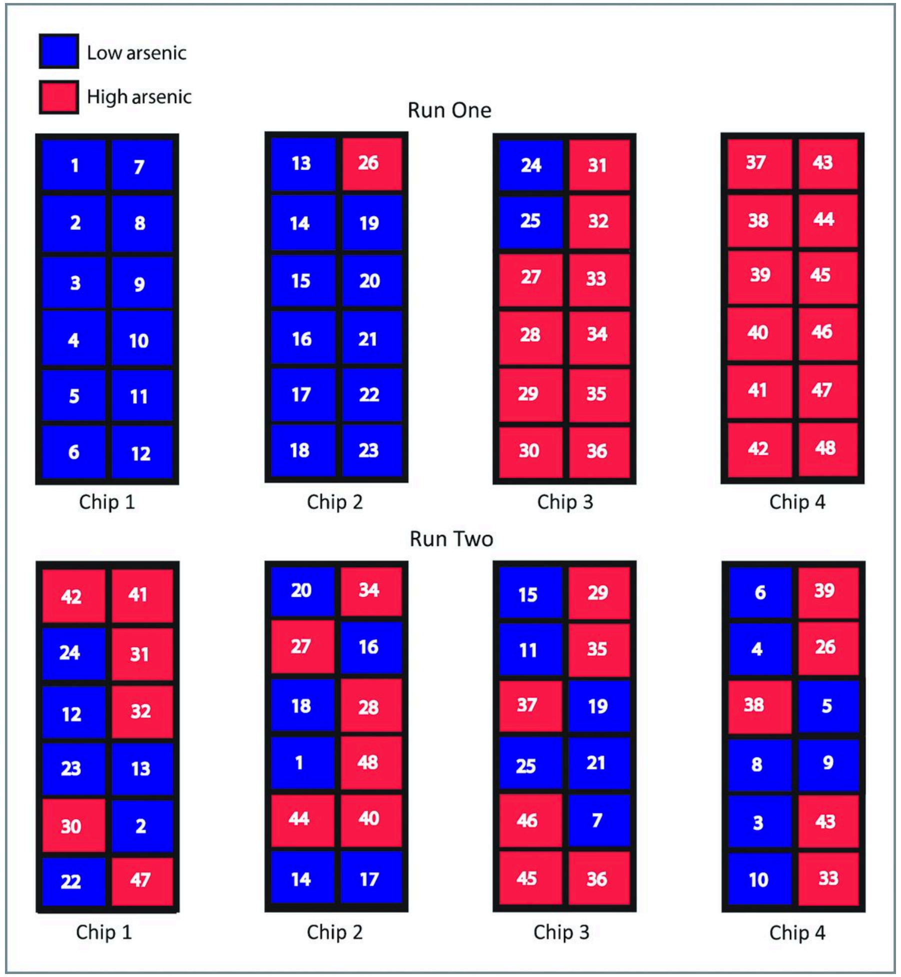
438 Accesses | **78** Citations | **19** Altmetric | Metrics

"When we used a standard method to [account for processing date], we find no evidence for differential expression between populations"

https://www.nature.com/articles/ng0707-807

Avoiding batch/confounding effects: **design**

- **Randomization:**
  - Don’t run in batches (can be hard to avoid)
  - Balance or randomize design with respect to batches
- **Record keeping:**
  - Record potential sources of artifacts (e.g. new tube of a reagent, technician, processing date, etc.)
- **Technical replicates:**
  - Run same biological sample in multiple batches



Run One

Chip 1		Chip 2		Chip 3		Chip 4	
🟦 1	🟦 7	🟦 13	🟥 26	🟦 24	🟥 31	🟥 37	🟥 43
🟦 2	🟦 8	🟦 14	🟦 19	🟦 25	🟥 32	🟥 38	🟥 44
🟦 3	🟦 9	🟦 15	🟦 20	🟥 27	🟥 33	🟥 39	🟥 45
🟦 4	🟦 10	🟦 16	🟦 21	🟥 28	🟥 34	🟥 40	🟥 46
🟦 5	🟦 11	🟦 17	🟦 22	🟥 29	🟥 35	🟥 41	🟥 47
🟦 6	🟦 12	🟦 18	🟦 23	🟥 30	🟥 36	🟥 42	🟥 48

Run Two

Chip 1		Chip 2		Chip 3		Chip 4	
🟥 42	🟥 41	🟦 20	🟥 34	🟦 15	🟥 29	🟦 6	🟥 39
🟦 24	🟥 31	🟥 27	🟦 16	🟦 11	🟥 35	🟦 4	🟥 26
🟦 12	🟥 32	🟦 18	🟥 28	🟥 37	🟦 19	🟥 38	🟦 5
🟦 23	🟦 13	🟦 1	🟥 48	🟦 25	🟦 21	🟦 8	🟦 9
🟥 30	🟦 2	🟥 44	🟥 40	🟥 46	🟦 7	🟦 3	🟥 43
🟦 22	🟥 47	🟦 14	🟦 17	🟥 45	🟥 36	🟦 10	🟥 33

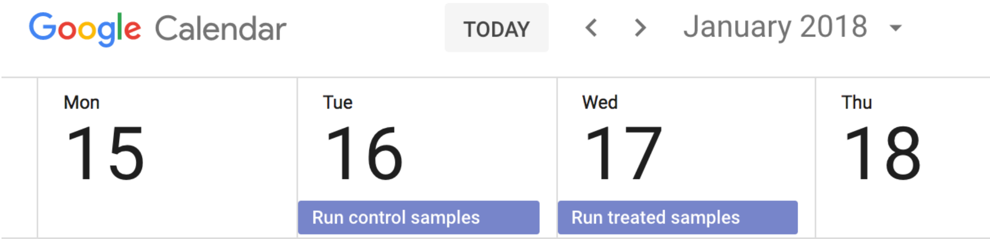
Harper, Peters, and Gamble (2013)

Avoiding confounding effects – public data

- Sometimes potential measured confounders can be found in provided metadata, but not always
  - e.g. processing date, sequencing run, clinical data
- Don’t use public data on “auto-pilot”, e.g.:
  - TCGA: are samples pre- or post-chemo? (not documented in most cases)
  - MDD/SCZ expression profiling: who is taking which medication?

Correcting for batch effects

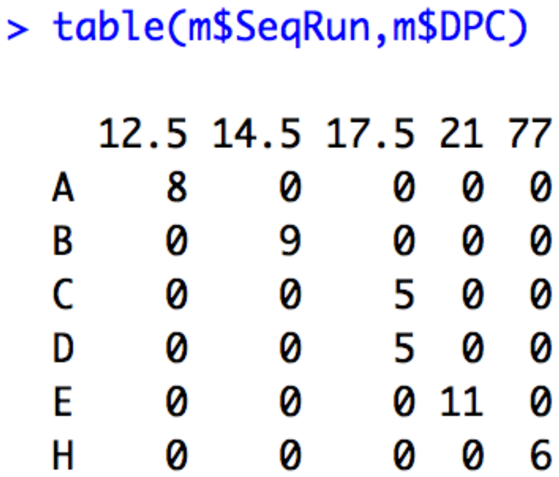
- Batch effects are not *always* large
  - Identify and collect "batch-related" variables
  - Assess the effect of batch using data visualization and dimensionality reduction
- Consider correcting for them if possible
  - not always possible – e.g. if batch is confounded with biology



Run control samples

Run treated samples

Chd8 study discussed in Lecture 3)



```
> table(m$SeqRun,m$DPC)

   12.5 14.5 17.5 21 77
 A    8    0    0  0  0
 B    0    9    0  0  0
 C    0    0    5  0  0
 D    0    0    5  0  0
 E    0    0    0 11  0
 H    0    0    0  0  6
```

Correcting for batch effects

- We can do this with linear models
- Add a ‘batch’ variable as a covariate
  - **Option 1 (preferred)**: Estimate/test effects of interest in the presence of the batch effect – in this way you are ‘adjusting’ for the variation attributable to batch
  - **Option 2**: Fit model with only batch, and proceed with residuals as ‘adjusted’ data (variation due to ‘batch’ is removed)
- But… should we do this for every gene separately??

ComBat: adjust for known batches

Biostatistics. 2007 Jan;8(1):118-27. Epub 2006 Apr 21.

Adjusting batch effects in microarray expression data using empirical Bayes methods.

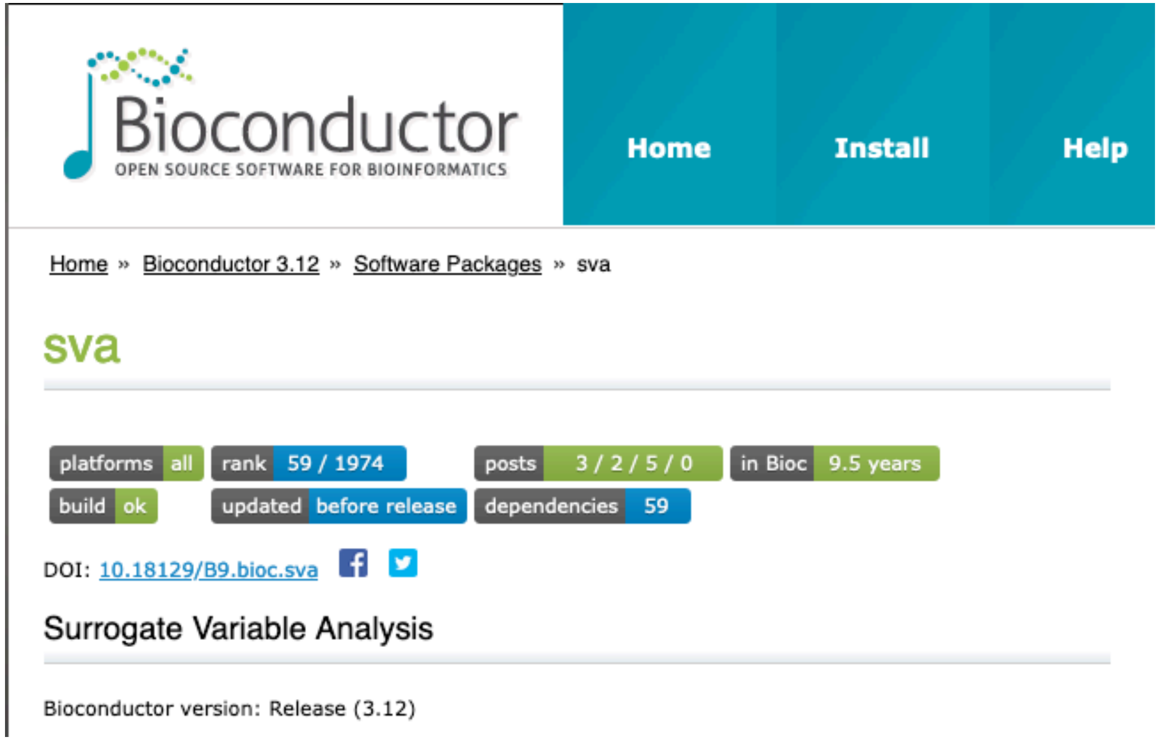
Johnson WE[1], Li C, Rabinovic A.

“This method incorporates systematic batch biases **common across genes** in making adjustments, assuming that phenomena resulting in batch effects often **affect many genes in similar ways**”

- uses *empirical Bayes* techniques to “*shrink*” the batch effect parameter estimates toward the overall mean of the batch effect estimates (across genes) – sound familiar?
- Model for sample i, batch j, gene g: $Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg},$
  - additive and multiplicative effects of batch are assumed to be RVs with prior distribution & estimated with empirical Bayes technique

ComBat: adjust for known batches

Implemented in Bioconductor package **sva**



Home » Bioconductor 3.12 » Software Packages » sva

sva

platforms all | rank 59 / 1974 | posts 3 / 2 / 5 / 0 | in Bioc 9.5 years
build ok | updated before release | dependencies 59

DOI: 10.18129/B9.bioc.sva

Surrogate Variable Analysis

Bioconductor version: Release (3.12)

- Vignette describes how to carry out both option 1 (including batch in the linear model) and option 2 (using residuals after adjusting for batch)
- Special considerations for array and sequencing data

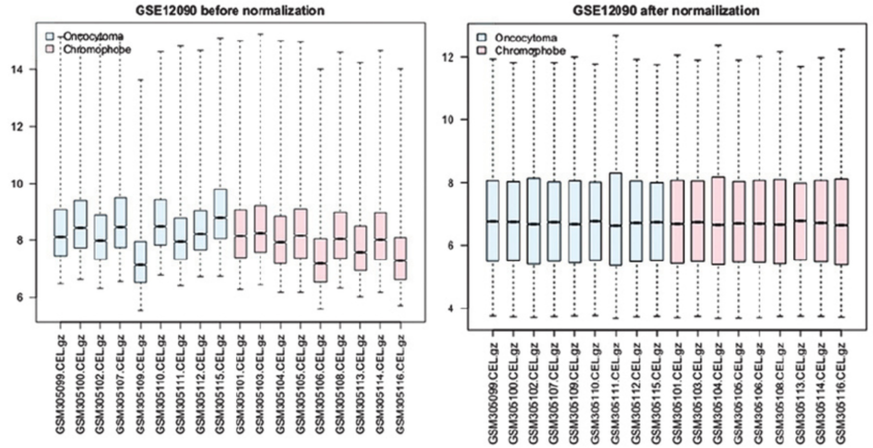
https://bioconductor.org/packages/release/bioc/html/sva.html

What about unmeasured confounders?

- Is there any hope to adjust for unwanted variation due to unmeasured confounders?
- Enter: **Surrogate Variable Analysis** (also called Latent Factor Analysis)
  - Main idea: estimate and adjust for common factors of extra, unwanted variation that are *separate* from that attributable to the effect(s) of interest
  - **Use with CAUTION!** Sometimes latent variables represent important biological sources of variation
    - e.g. unknown subgroups of different cancer types
    - One possible workaround – use only 'control' (housekeeping) genes to estimate the latent factors
  - Bioconductor packages: **sva, RUVseq**

Normalization

- Normalization in genomics does **not** mean "make normally distributed"
- Broadly, the goal of normalization:
  - standardize / make measurements comparable to one another by reducing variation attributable to technical factors
  - specifically, by modifying the scale or distribution of samples so they are comparable across the whole experiment – **requires assumptions!**
- Between-sample normalization
  - e.g. differential expression between sample groups
- Within-sample normalization
  - e.g. differential expression between genes



https://doi.org/10.3892/mmr.2014.2766

Normalization approaches - Microarrays

- Within sample:
  - Background intensity correction
  - Probe affinity correction
- Between sample:
  - **Scaling normalization** – scale so that all samples have the same mean or median
  - **Quantile normalization** – standardize quantiles of distribution across samples
  - **Loess normalization** – estimate and remove nonlinear bias
- Bioconductor Implementation: affy package

Normalization approaches - RNAseq

- Within sample:
  - Gene length correction
  - GC bias correction
- Between sample:
  - **Scaling normalization** – scale so that all samples have the same mean or median
  - **Quantile normalization** – standardize quantiles of distribution across samples
  - **Variance-stabilizing transformations** – estimate and remove nonlinear bias
- Many different implementations!
- More on this in our lectures on RNAseq

Normalization vs experimental design

- Batch effects and confounding are an **experimental design problem**
- Normalization doesn't address all unmeasured confounding and can in fact exacerbate it
  - If confounding / batch effects are present and not adjusted for, assumptions of normalization may be violated – e.g. overcorrection