

# Study design, batch effects, and confounding

Jeff Leek

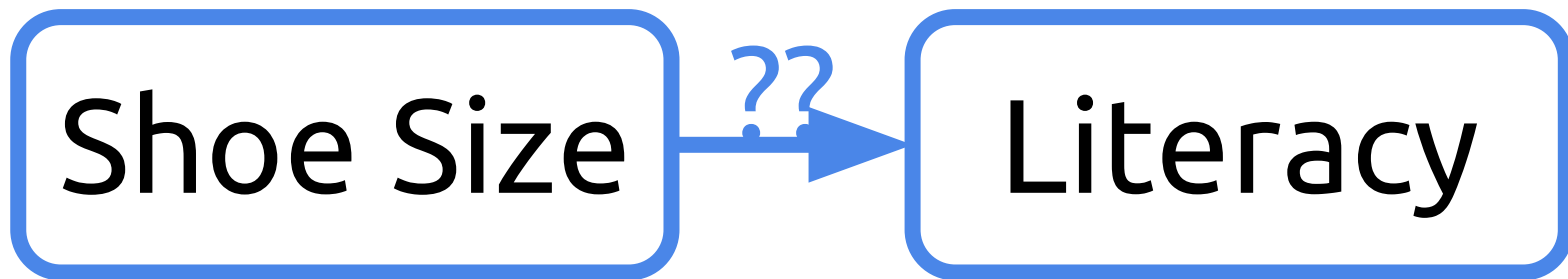
@jtleek

**What is confounding?**


A photograph of a man and a young child in a grassy yard. The man, wearing a light blue t-shirt and black shorts, is walking towards the right. The child, wearing an orange t-shirt and dark pants, is walking towards the left. A red and black ladybug-shaped ball is on the grass between them. In the background, there is a wooden fence and some green plants. Two callout boxes with white text are overlaid on the image. One box points to the child, and the other points to the man.

Small shoes  
Not literate

Big shoes  
Somewhat literate





A photograph of a man and a young child in a grassy yard. The man, wearing a light blue t-shirt and black shorts, is walking towards the right. The child, wearing an orange t-shirt and dark pants, is walking towards the left. A red ball is on the grass between them. In the background, there is a wooden fence and some green plants. Two text boxes are overlaid on the image: one in the upper left and one in the lower right.

Small shoes  
Not literate  
Young

Big shoes  
Somewhat literate  
Middle aged

Shoe Size

Literacy

Age

Shoe Size

Literacy

Age

```
graph TD; A[Shoe Size] --> C[Age]; B[Literacy] --> C;
```

The diagram illustrates a relationship where 'Shoe Size' and 'Literacy' are solid variables that point to 'Age', which is represented by a dashed box. This suggests that 'Age' is a latent variable or a common factor influencing both 'Shoe Size' and 'Literacy'.

Variable1

Variable2

Confounder

```
graph TD; C[Confounder] --> V1[Variable1]; C --> V2[Variable2];
```

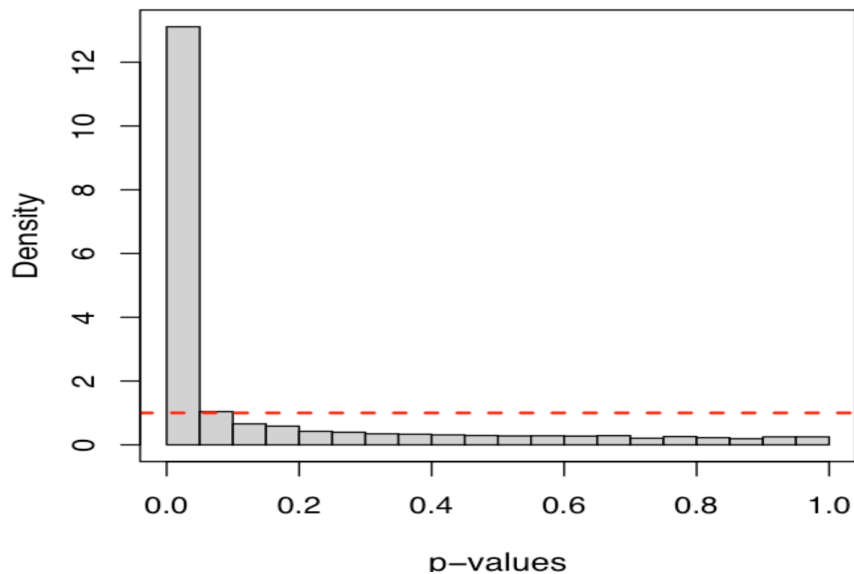
The diagram illustrates a causal relationship where a single factor, labeled 'Confounder', influences two separate variables, 'Variable1' and 'Variable2'. The 'Confounder' is represented by a dashed blue box at the bottom, while 'Variable1' and 'Variable2' are in solid blue boxes at the top. Two solid blue arrows originate from the top of the 'Confounder' box and point towards the bottom of the 'Variable1' and 'Variable2' boxes, respectively, indicating a direct causal effect.



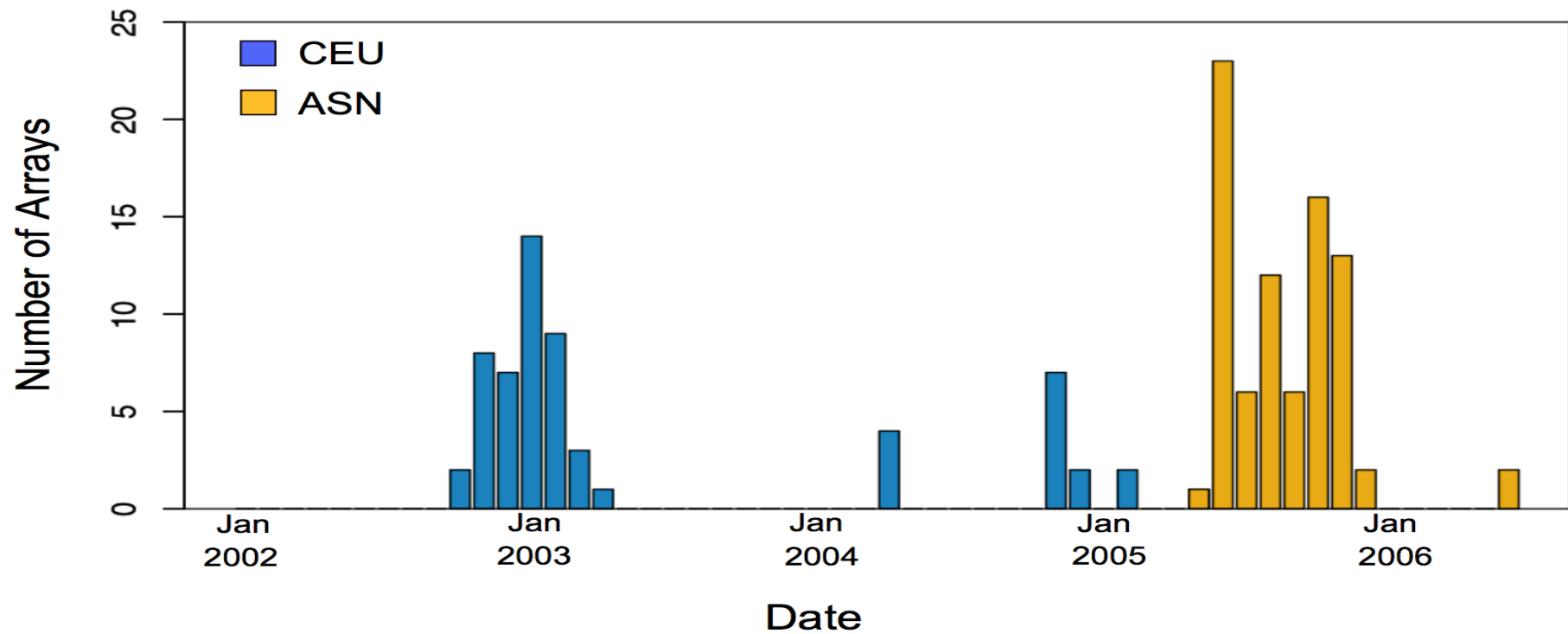
**Most common confounder: batch effects**

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

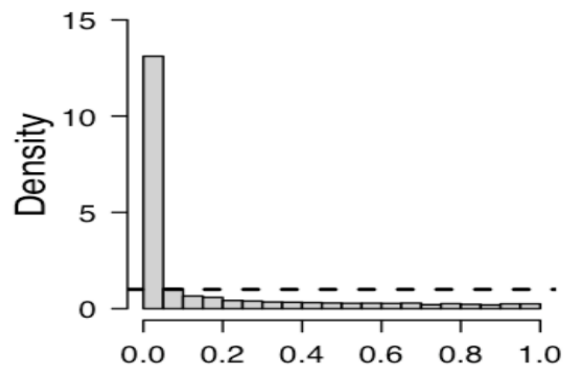
Richard S Spielman<sup>1</sup>, Laurel A Bastone<sup>2</sup>, Joshua T Burdick<sup>3</sup>, Michael Morley<sup>3</sup>, Warren J Ewens<sup>4</sup> & Vivian G Cheung<sup>1,3,5</sup>



78% of genes differentially expressed

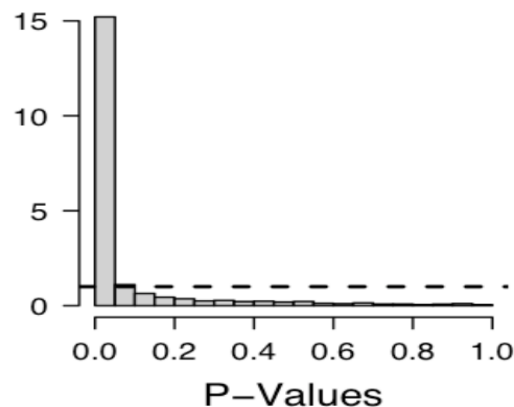


### Between Population



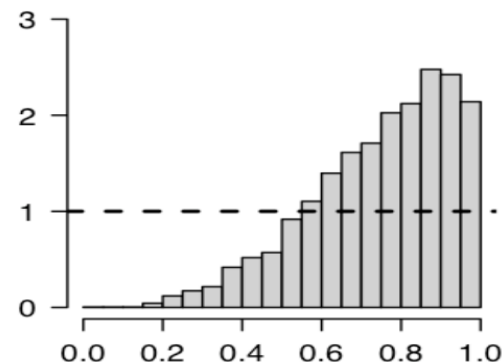
**78%** of genes estimated  
to be differentially

### Between Years



**96%** of genes estimated  
to be differentially

### Between Populations, Adjusting For Years



**0%** of genes estimated to  
be differentially

**Extremely common**

**Science** The World's Leading Journal of Original Scientific Research, Global News, and Commentary.[Science Home](#)[Current Issue](#)[Previous Issues](#)[Science Express](#)[Science Products](#)[My Science](#)[About the Journal](#)[Home](#) > [Science Magazine](#) > [Science Express](#) > [Sebastiani et al.](#)

## Article Views

- ▶ **Abstract**
- ▶ Full Text (PDF)
- ▶ Supporting Online Material

## VERSION HISTORY

- ▶ [science.1190532v4](#)  
(most recent)
- ▶ [science.1190532v3](#)
- ▶ [science.1190532v2](#)
- ▶ [science.1190532v1](#)

Published Online July 1 2010

Science DOI: 10.1126/science.1190532

< [Science Express Index](#)

## REPORT

**Genetic Signatures of Exceptional Longevity in Humans**

Paola Sebastiani<sup>1,2\*</sup>, Nadia Solovieff<sup>1</sup>, Annibale Puca<sup>2</sup>, Stephen W. Hartley<sup>1</sup>, Efthymia Melista<sup>3</sup>, Stacy Andersen<sup>4</sup>, Daniel A. Dworkis<sup>3</sup>, Jemma B. Wilk<sup>5</sup>, Richard H. Myers<sup>5</sup>, Martin H. Steinberg<sup>6</sup>, Monty Montano<sup>3</sup>, Clinton T. Baldwin<sup>6,7</sup> and Thomas T. Perls<sup>4,\*</sup>

[±](#) Author Affiliations

\*To whom correspondence should be addressed. E-mail: [sebas@bu.edu](mailto:sebas@bu.edu) (P.S.); [thperls@bu.edu](mailto:thperls@bu.edu) (T.H.P.)



**Science** The World's Leading Journal of Original Scientific Research, Global News, and Commentary.[Science Home](#)[Current Issue](#)[Previous Issues](#)[Science Express](#)[Science Products](#)[My Science](#)[About the Journal](#)[Home](#) > [Science Magazine](#) > [Science Express](#) > [Sebastiani et al.](#)

## Article Views

- ▶ Abstract
- ▶ Full Text (PDF)
- ▶ Supporting Online Material

## VERSION HISTORY

- ▶ [science.1190532v4](#)  
(most recent)
- ▶ [science.1190532v3](#)
- ▶ [science.1190532v2](#)
- ▶ [science.1190532v1](#)

**This article has been retracted**

Published Online July 1 2010

Science DOI: 10.1126/science.1190532

< [Science Express Index](#)

## REPORT

**Genetic Signatures of Exceptional Longevity in Humans**

Paola Sebastiani<sup>1,\*</sup>, Nadia Solovieff<sup>1</sup>, Annibale Puca<sup>2</sup>, Stephen W. Hartley<sup>1</sup>, Efthymia Melista<sup>3</sup>, Stacy Andersen<sup>4</sup>, Daniel A. Dworkis<sup>3</sup>, Jemma B. Wilk<sup>5</sup>, Richard H. Myers<sup>5</sup>, Martin H. Steinberg<sup>6</sup>, Monty Montano<sup>3</sup>, Clinton T. Baldwin<sup>6,7</sup> and Thomas T. Perls<sup>4,\*</sup>

[±](#) Author Affiliations

\*To whom correspondence should be addressed. E-mail: [sebas@bu.edu](mailto:sebas@bu.edu) (P.S.); [thperls@bu.edu](mailto:thperls@bu.edu) (T.H.P.)



doi:10.1016/S0140-6736(02)07746-2 | [How to Cite or Link Using DOI](#)

[Permissions & Reprints](#)

## Fast track — Mechanisms of Disease

# Use of proteomic patterns in serum to identify ovarian cancer

## How Bright Promise in Cancer Testing Fell Apart



Keith Baggerly, left, and Kevin Coombes, statisticians at M. D. Anderson Cancer Center, found flaws in research on tumors.

By JENNIFER A. GIL

## Perspective

*Nature Reviews Genetics* **11**, 733–739 (1 October 2010) | doi:10.1038/nr

### 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**

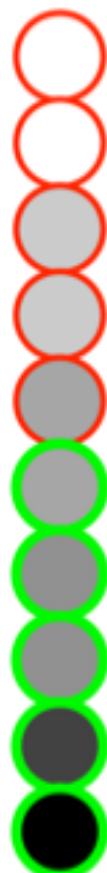
High-throughput technologies are widely used, for example to assay genetic variants, gene and protein expression, and epigenetic modifications. One often overlooked complication with such studies is batch effects, which occur because measurements are affected by laboratory conditions, reagent lots and personnel differences. This becomes a major problem when batch effects are correlated with an outcome of interest and lead to incorrect conclusions. Using both published studies and our own analyses, we argue that batch effects (as well as other technical and biological artefacts) are widespread and critical to address. We review experimental and computational approaches for doing so.

# Randomization

Confounding variable

Experimental units

Treatments



Without randomization, the confounding variable differs among treatments

Confound variable

Experimental units

Treatments



With randomization, the confounding variable does not differ among treatments



# Stratification example

### Example:

- ▶ 20 males and 20 females.
- ▶ Half to be treated; the other half left untreated.
- ▶ Can only work with 4 individuals per day.

### Question:

How to assign individuals to treatment groups and to days?

**A bad design**

## Week One

M Tu W Th F

C	C	C	C	C
C	C	C	C	C
C	C	C	C	C
C	C	C	C	C

## Week Two

M Tu W Th F

T	T	T	T	T
T	T	T	T	T
T	T	T	T	T
T	T	T	T	T

T = treated. C = control. pink = female. blue = male

# Stratifying

## Week One

M	Tu	W	Th	F
C	T	T	T	T
T	C	C	C	T
C	C	C	T	C
T	T	T	C	C

## Week Two

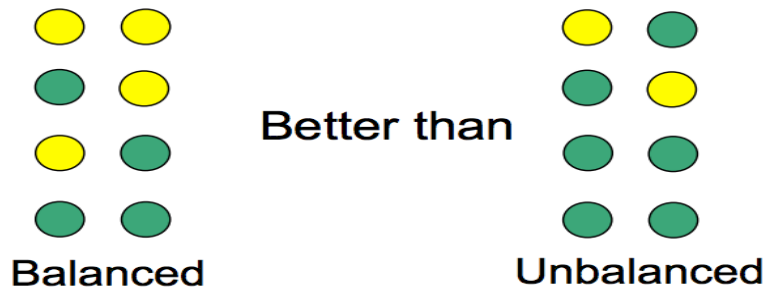
M	Tu	W	Th	F
T	T	T	C	T
C	C	C	T	T
C	C	T	T	C
T	T	C	C	C

T = treated, C = control, pink = female, blue = male



**More good study characteristics**

- Balanced



- Replicated
- Has Controls