

Experimental design: confounding and randomization

Jeff Leek

@jtleek

www.jtleek.com

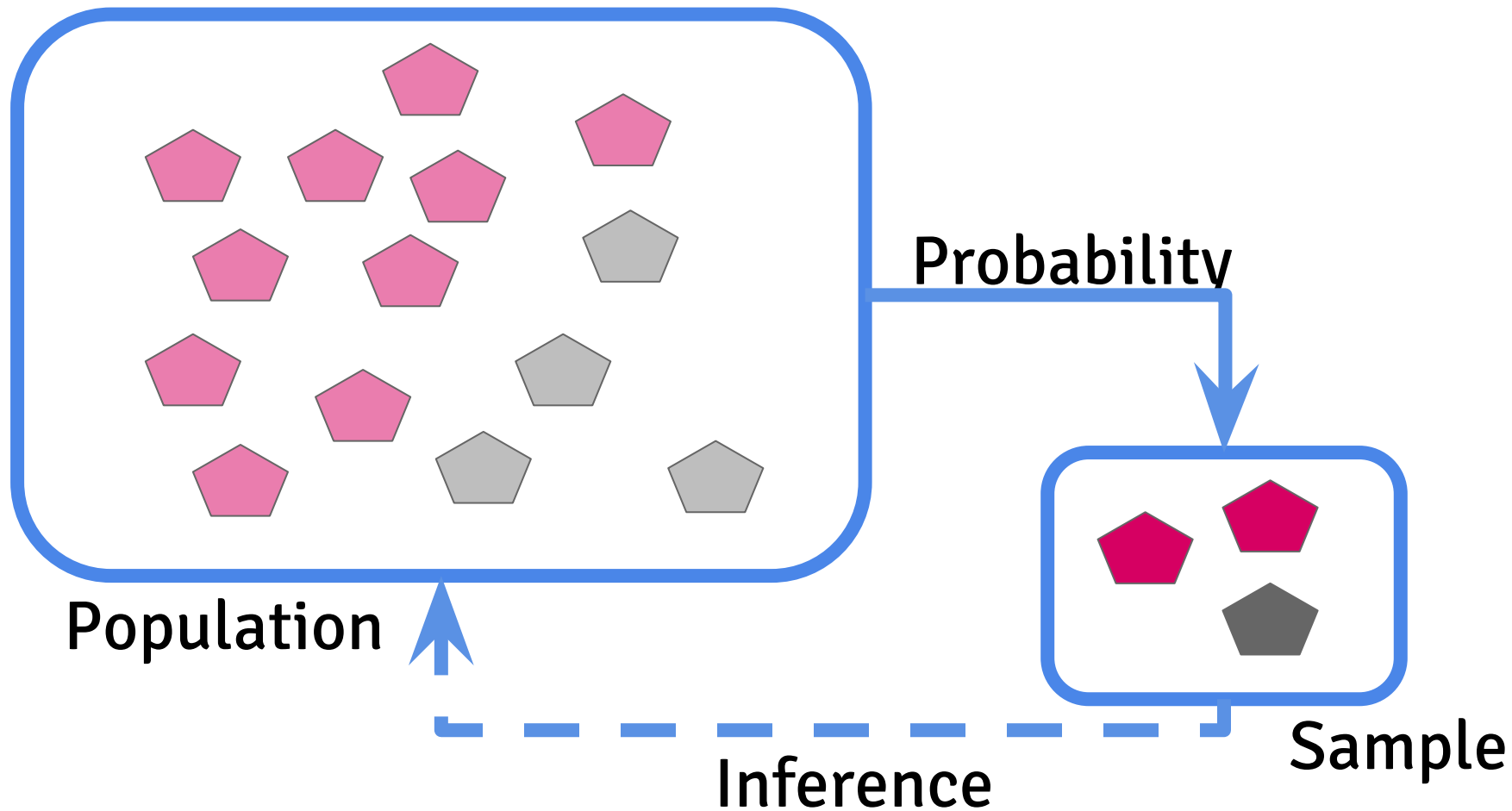
Key ideas

Confounding

Batch effects

Randomization

Central dogma of statistics

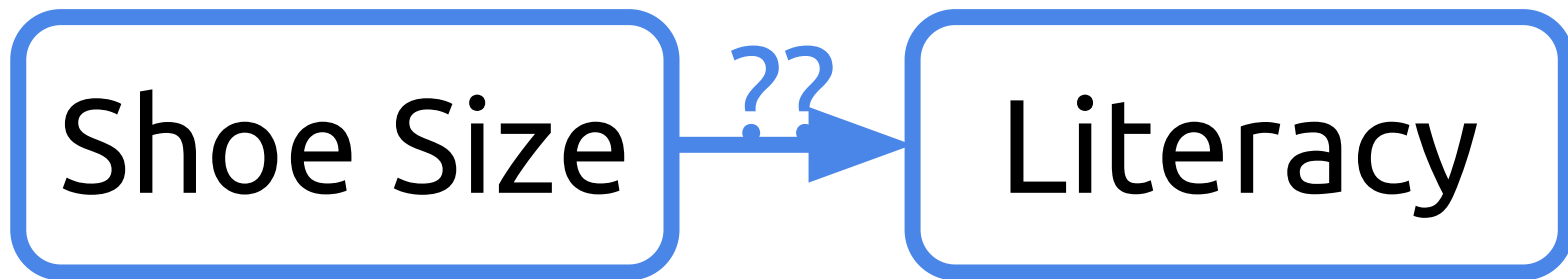



What is confounding?

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. The first box, in the upper left, contains the text 'Small shoes', 'Not literate', and 'Young'. The second box, in the lower right, contains the text 'Big shoes', 'Somewhat literate', and 'Middle aged'.

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 causal relationship. At the top, two solid blue-outlined boxes labeled 'Shoe Size' and 'Literacy' have arrows pointing downwards to a dashed blue-outlined box labeled 'Age'. This suggests that both shoe size and literacy are influenced by age.

Variable1

Variable2

Confounder

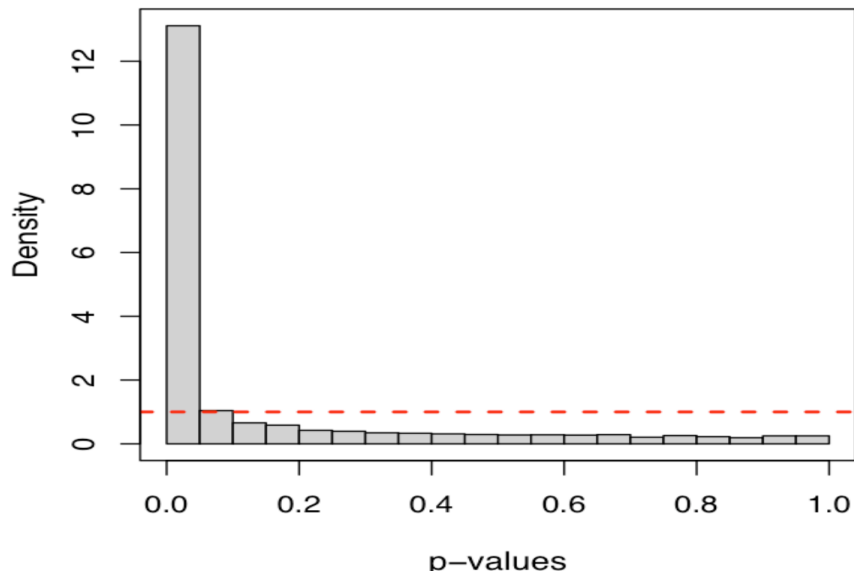
```
graph TD; C[Confounder] --> V1[Variable1]; C --> V2[Variable2];
```

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.

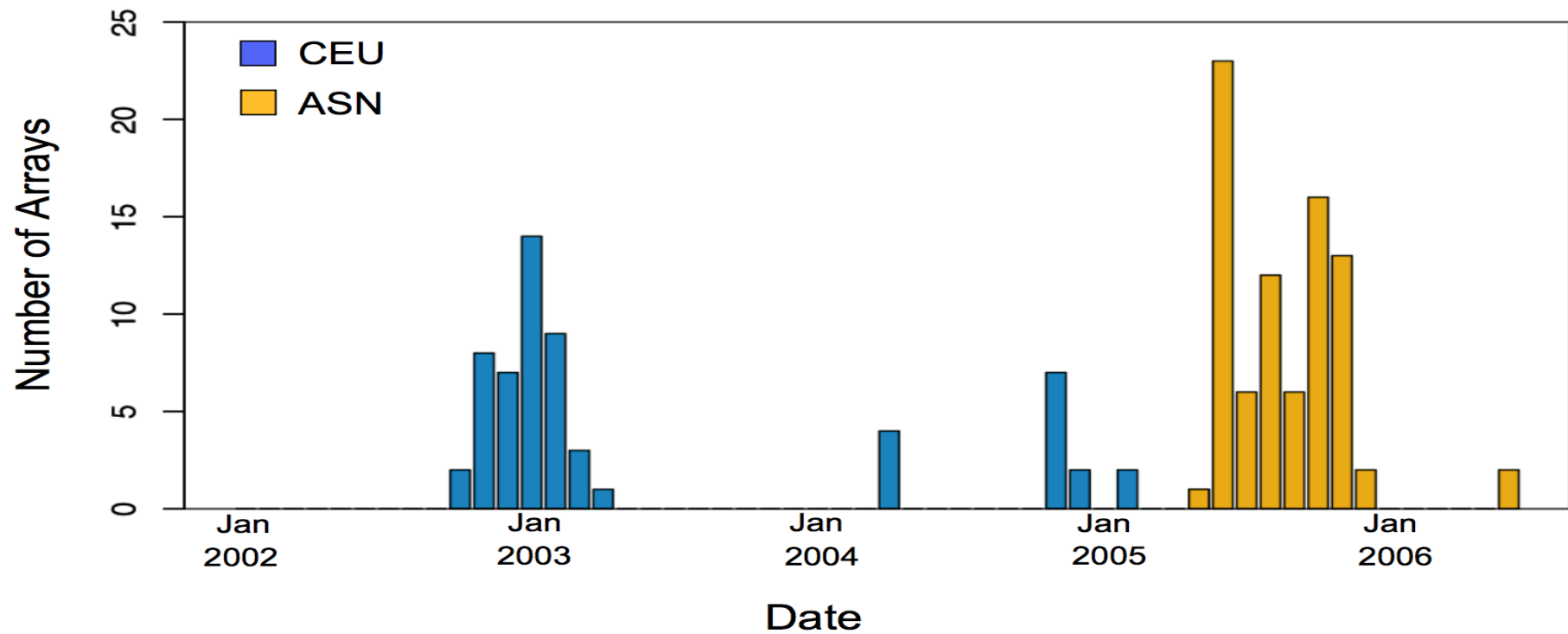
Most common confounder: batch effects

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

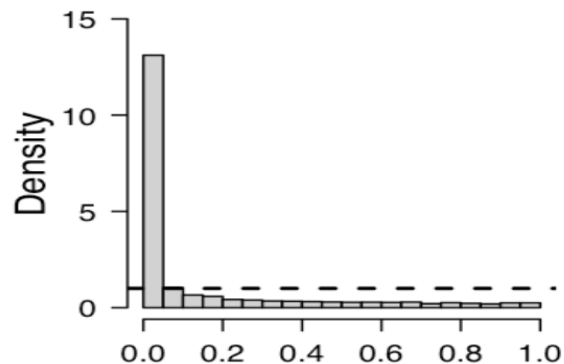
Richard S Spielman¹, Laurel A Bastone², Joshua T Burdick³, Michael Morley³, Warren J Ewens⁴ & Vivian G Cheung^{1,3,5}



78% of genes differentially expressed

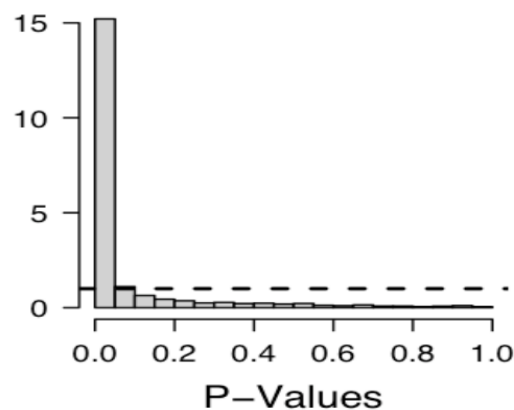


Between Population



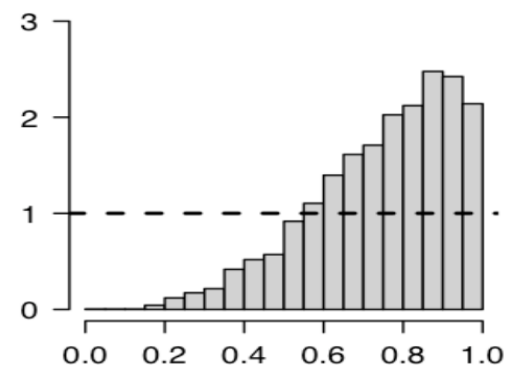
78% of genes estimated
to be differentially

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^{1,2*}, Nadia Solovieff¹, Annibale Puca², Stephen W. Hartley¹, Efthymia Melista³, Stacy Andersen⁴, Daniel A. Dworkis³, Jemma B. Wilk⁵, Richard H. Myers⁵, Martin H. Steinberg⁶, Monty Montano³, Clinton T. Baldwin^{6,7} and Thomas T. Perls^{4,*}

[±](#) Author Affiliations

*To whom correspondence should be addressed. E-mail: sebas@bu.edu (P.S.); 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^{1,2*}, Nadia Solovieff¹, Annibale Puca², Stephen W. Hartley¹, Efthymia Melista³, Stacy Andersen⁴, Daniel A. Dworkis³, Jemma B. Wilk⁵, Richard H. Myers⁵, Martin H. Steinberg⁶, Monty Montano³, Clinton T. Baldwin^{6,7} and Thomas T. Perls^{4,*}

[±](#) Author Affiliations

*To whom correspondence should be addressed. E-mail: sebas@bu.edu (P.S.); 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. WEAVER

Perspective

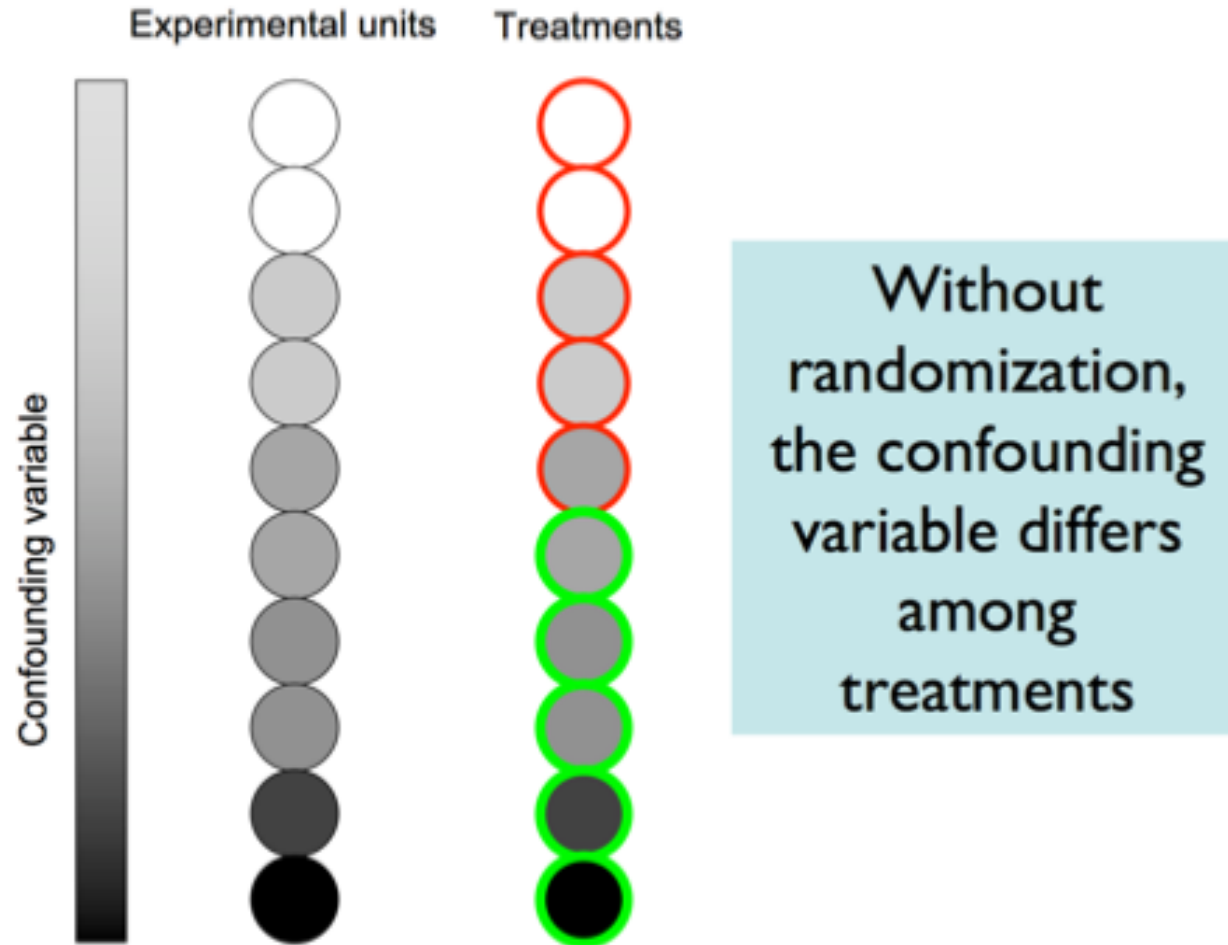
Nature Reviews Genetics **11**, 733–739 (1 October 2010) | doi:10.1038/nrg2825

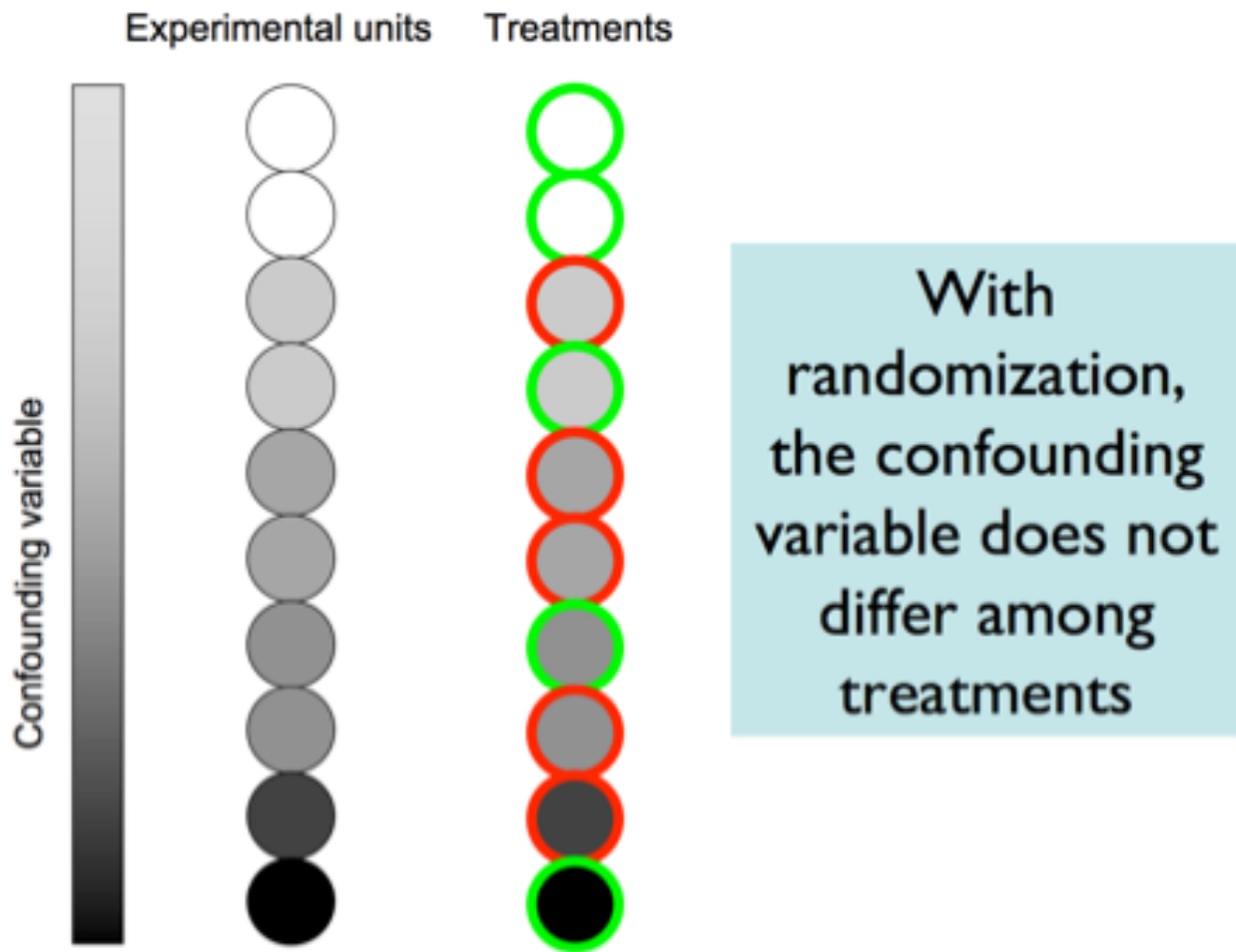
Tackling the widespread and critical impact of batch effects in high-throughput data

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





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 |

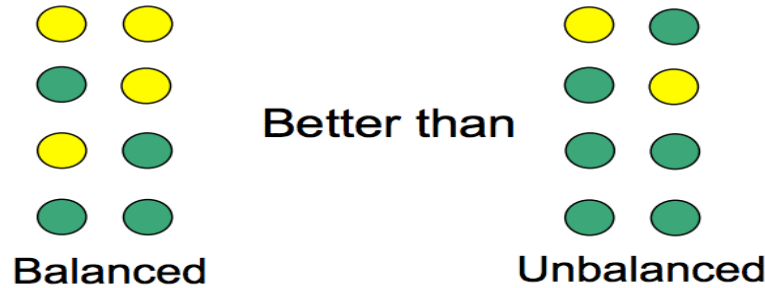
Week Two

| 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