@CSoneson

Experimental design

Charlotte Soneson

Friedrich Miescher Institute for Biomedical Research &
SIB Swiss Institute of Bioinformatics

SIB Swiss Institute of Bioinformatics

FMI Friedrich Miescher Institute
for Biomedical Research

"To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of."

Sir Ronald Fisher, Indian Statistical Congress, Sankhya, around 1938

Stephen John Senn
@stephensenn

⚙ 👤+ Follow

Statisticians are the bad fairies of research. People forget to invite them until it's too late, at which point they send everyone to sleep.

RETWEETS
92

LIKES
93

11:22 AM - 21 Feb 2016

Different types of experiments

Learning experiment questions	Confirming experiment questions
• Does the drug have toxic side effects (at what dose, given for how long, in which tissue)?	• Does 5 mg/kg of the drug given once a day for 5 days increase blood creatinine[a] concentration?
• Does stress affect rodent behaviour (what kind of stress, for how long, on what behavioural tasks)?	• Does fox urine odour (a stressor) affect the amount of food Wistar rats consume during the first 24 hours after exposure?
• How dose exercise affect cognitive functioning of older people (what type of exercise, how much, which aspect of cognition)?	• Does 30 min of aerobic activity (treadmill running) at 60% VO_2 max[b], 3 days a week for 6 weeks, in males between 55–70 years of age, improve performance on a mental rotation task?

[a] Increased creatinine indicates kidney damage.
[b] VO_2 max is the maximal oxygen uptake and is a measure of a person's aerobic fitness.

[Lazic, 2016]

What is experimental design?

The organization of an experiment, to ensure that the **right type** of data, and **enough** of it, is available to answer the **questions of interest** as clearly and efficiently as possible.

What is **bad** experimental design?

Analysis batch I / Study center I / Processing protocol I ...

| Tr | Tr | Tr | Tr | Tr | Tr | Tr | Tr |

Analysis batch II / Study center II / Processing protocol II ...

| Ctl | Ctl | Ctl | Ctl | Ctl | Ctl | Ctl | Ctl |

What is **bad** experimental design?

Analysis batch I / Study center I / P...essing protocol I ...

Tr Tr Tr Tr Tr Tr

Confounding!

Analysis batch ... enter II / Processing protocol II ...

Ctl Ctl Ctl Ctl Ctl Ctl Ctl Ctl

What can happen with bad experimental design?

- Example: gene expression study comparing 60 CEU and 82 ASN HapMap individuals

- 26% of the genes were found to be significantly differentially expressed (78% with less restrictive multiple testing correction)

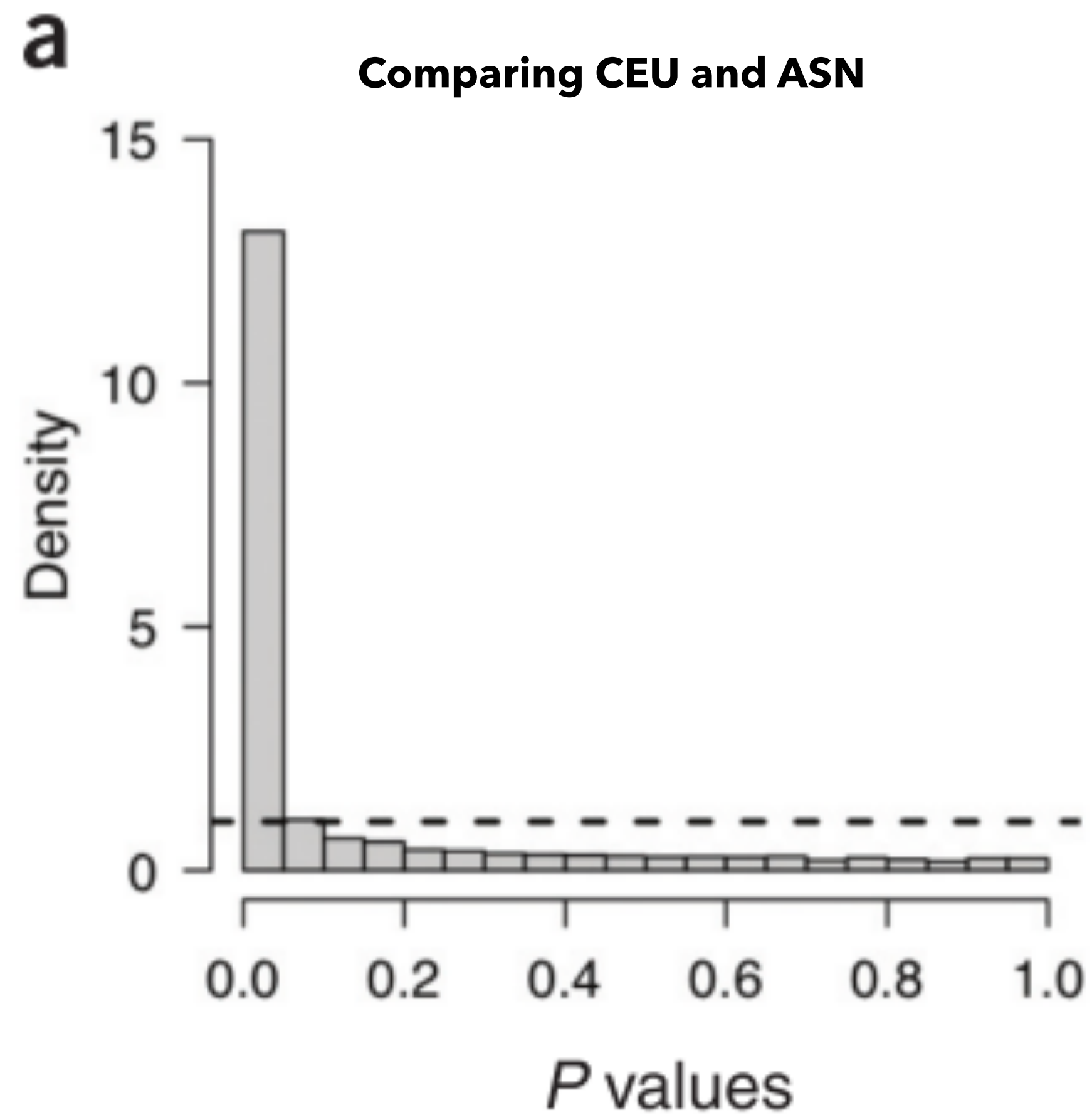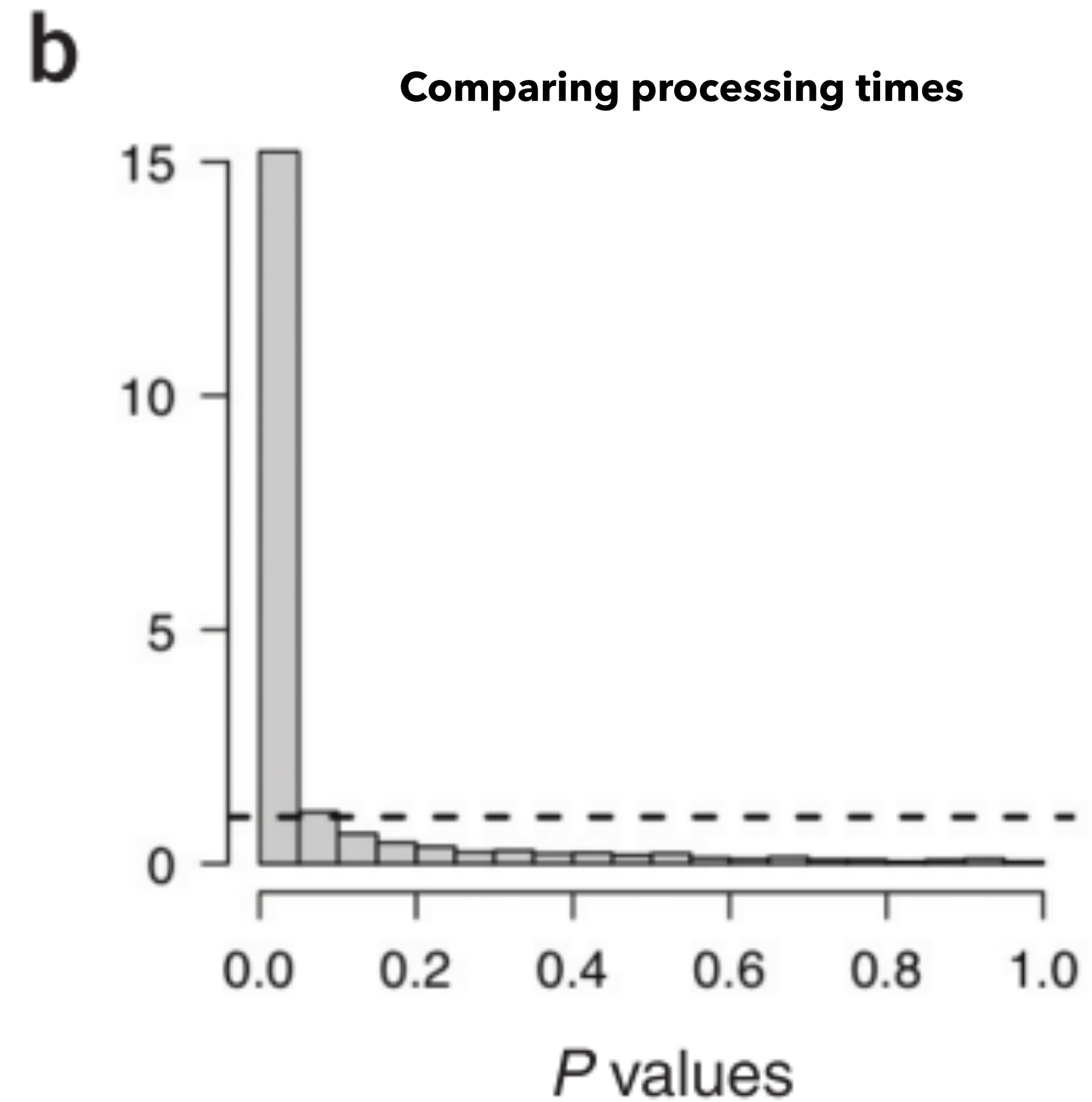- **But**: all CEU samples were processed (sometimes years) before all the ASN samples!

What can happen with bad experimental design?

- Example: gene expression study compar̶i̶s̶o̶n̶ ̶C̶E̶U and 82 ASN HapMap individuals

- 26% of the genes were f̶o̶u̶n̶d̶ ̶t̶o̶ ̶b̶e̶ ̶s̶ignificantly differentially expressed (78% with less res̶t̶r̶i̶c̶t̶i̶v̶e̶ ̶m̶u̶l̶t̶i̶p̶l̶e testing correction)

- **<u>But</u>**: all CEU samples were processed (sometimes years) before all the ASN samples!

Confounding!

Akey et al., Nature Genetics 2007; Spielman et al., Nature Genetics 2007

What can happen with bad experimental design?

Akey et al., Nature Genetics 2007; Spielman et al., Nature Genetics 2007

What would be a better experimental design?

- Process all samples at the same time/in one batch (not always feasible)

- Minimize confounding as much as possible through

  - blocking

  - randomization

- Batch effects may still be present, but with an appropriate design we can account for them
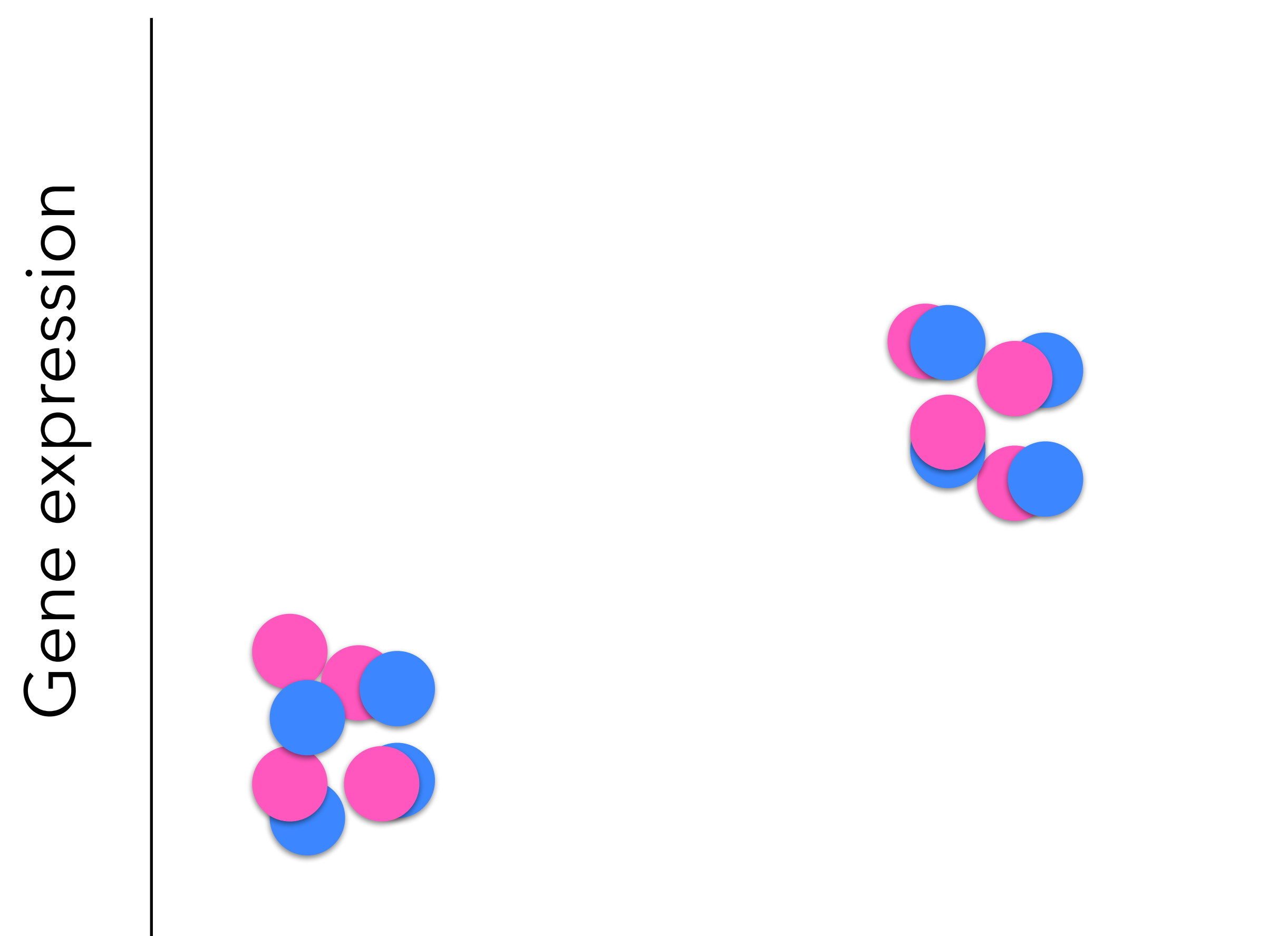
Nonzero batch effect
Zero treatment effect

Treated
Untreated

✗ Confounded design

✓ Non-confounded design

Gene expression

Gene expression
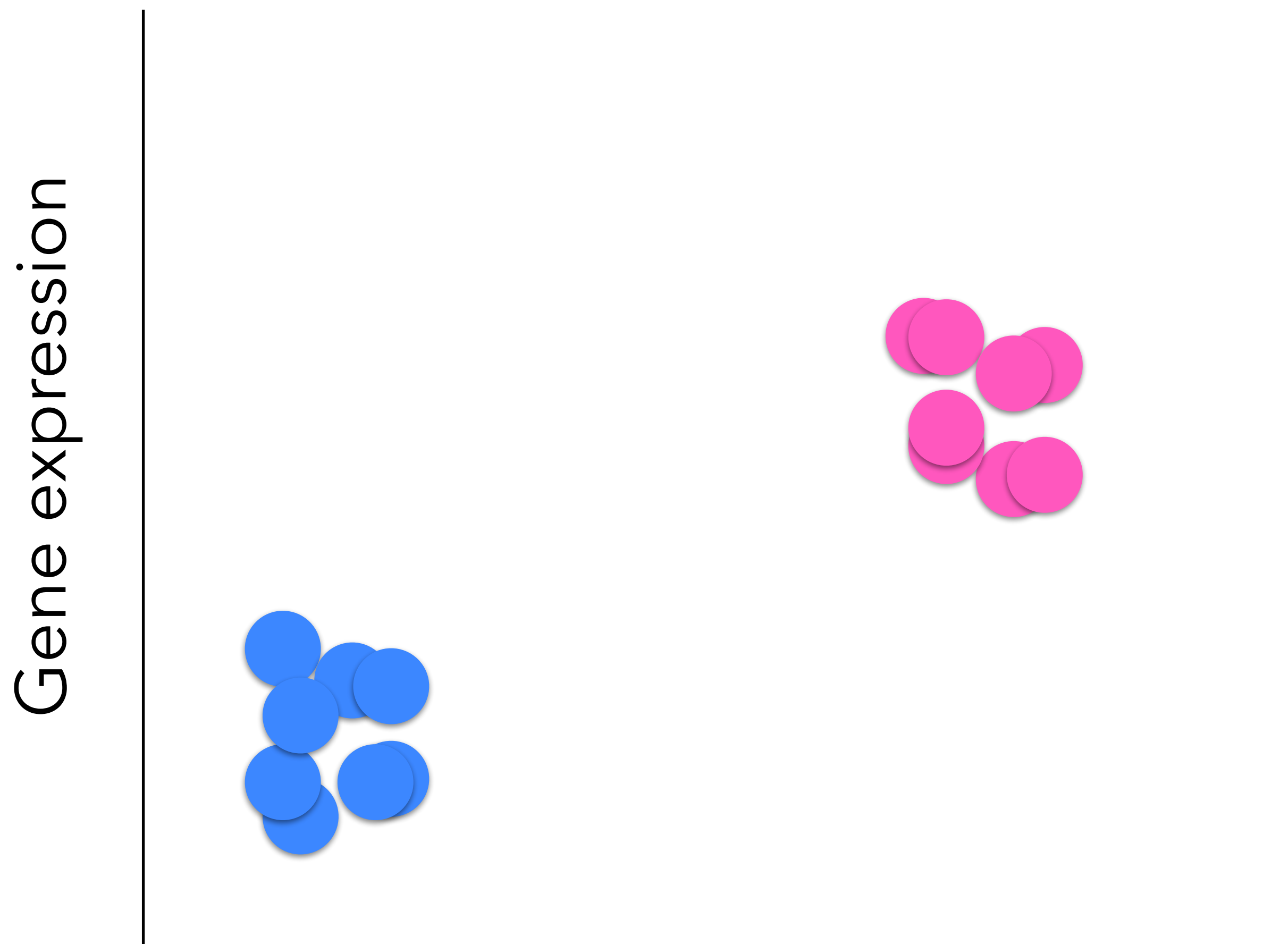
Batch A   Batch B

Batch A   Batch B

Nonzero batch effect
Nonzero treatment effect

Treated
Untreated

❌ Confounded design

✅ Non-confounded design

Gene expression

Gene expression
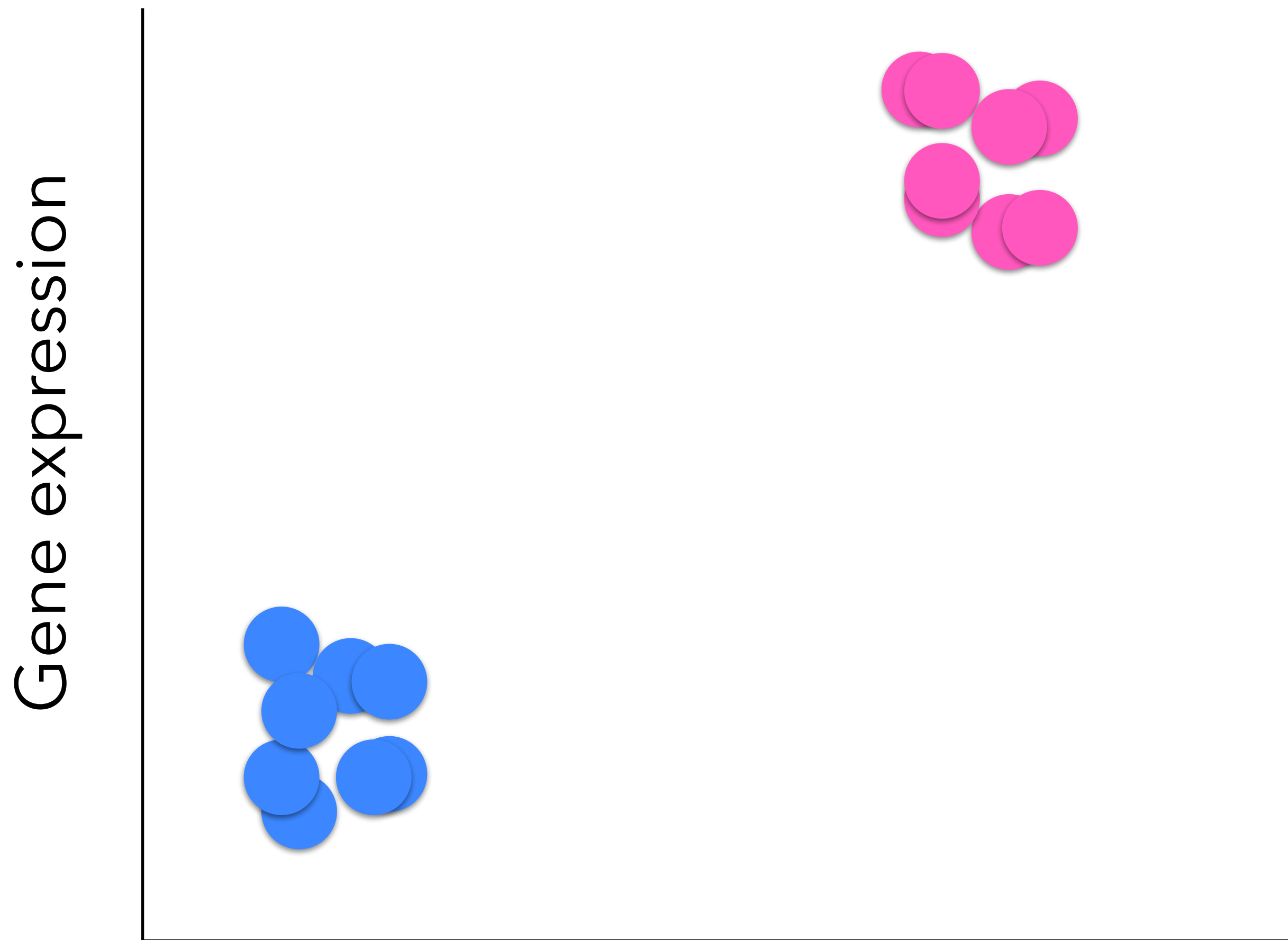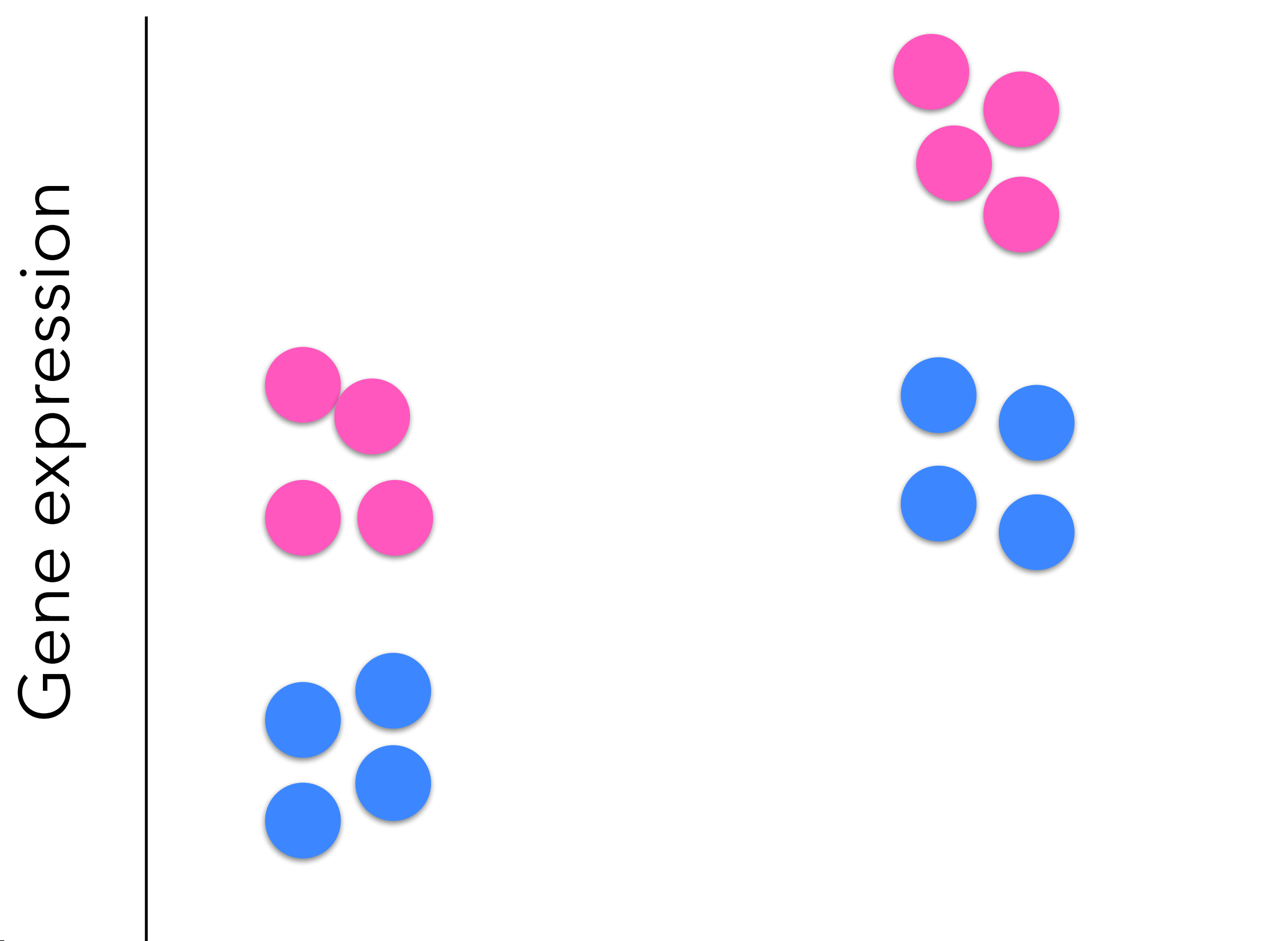
Batch A   Batch B

Batch A   Batch B

Dealing with batch effects

- In statistical modeling, batch effects can be included as **covariates** (additional predictors) in the model.

- For exploratory analysis, we often attempt to "eliminate" or "adjust for" such unwanted variation in advance, by subtracting the estimated effect from each variable (e.g. the expression of a gene).

- Even partial confounding between batch and signal of interest can lead to problems.

What can happen with bad experimental design?

a Comparing CEU and ASN

78% differentially expressed

b Comparing processing times

96% differentially expressed

Akey et al., Nature Genetics 2007; Spielman et al., Nature Genetics 2007
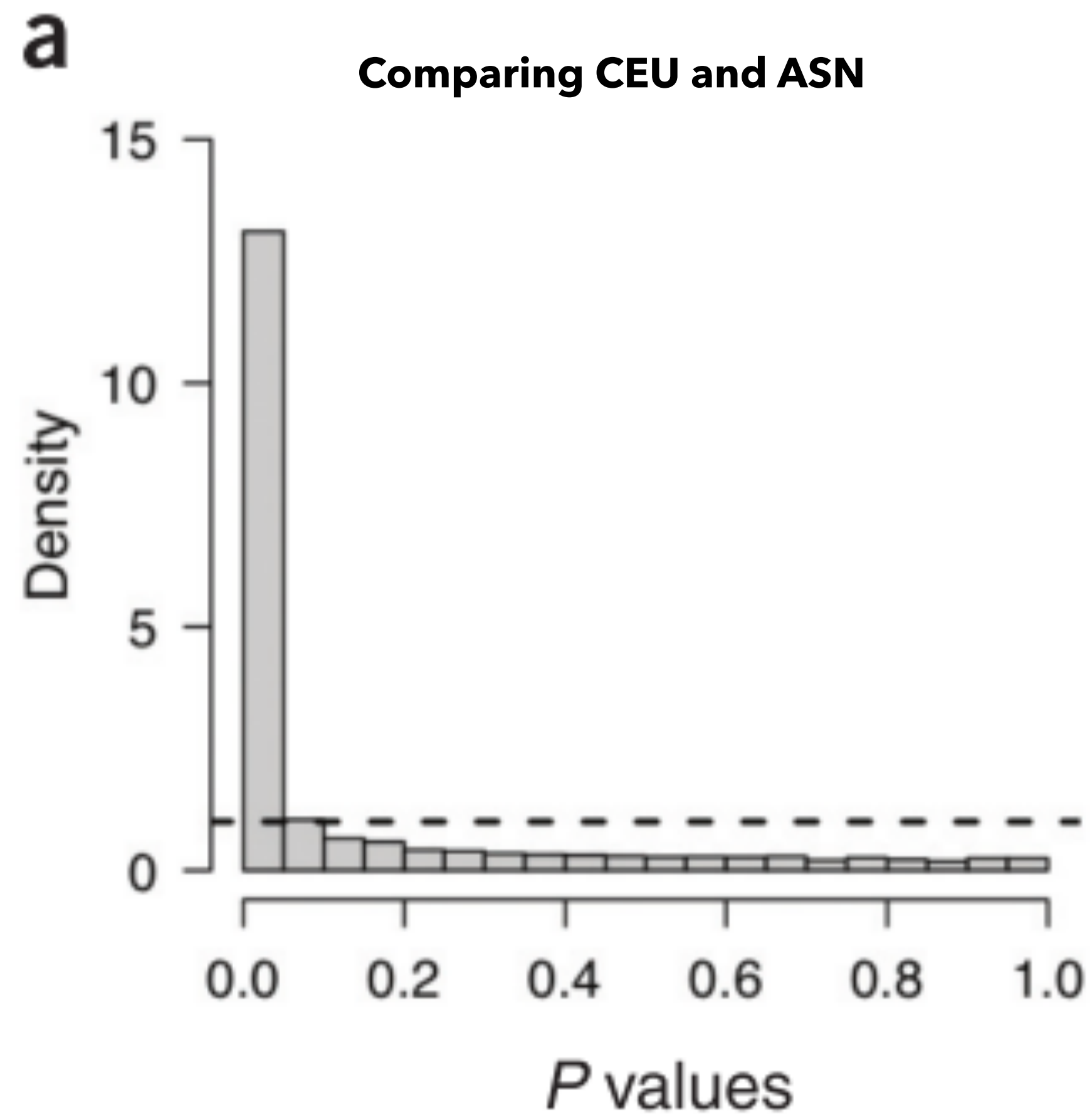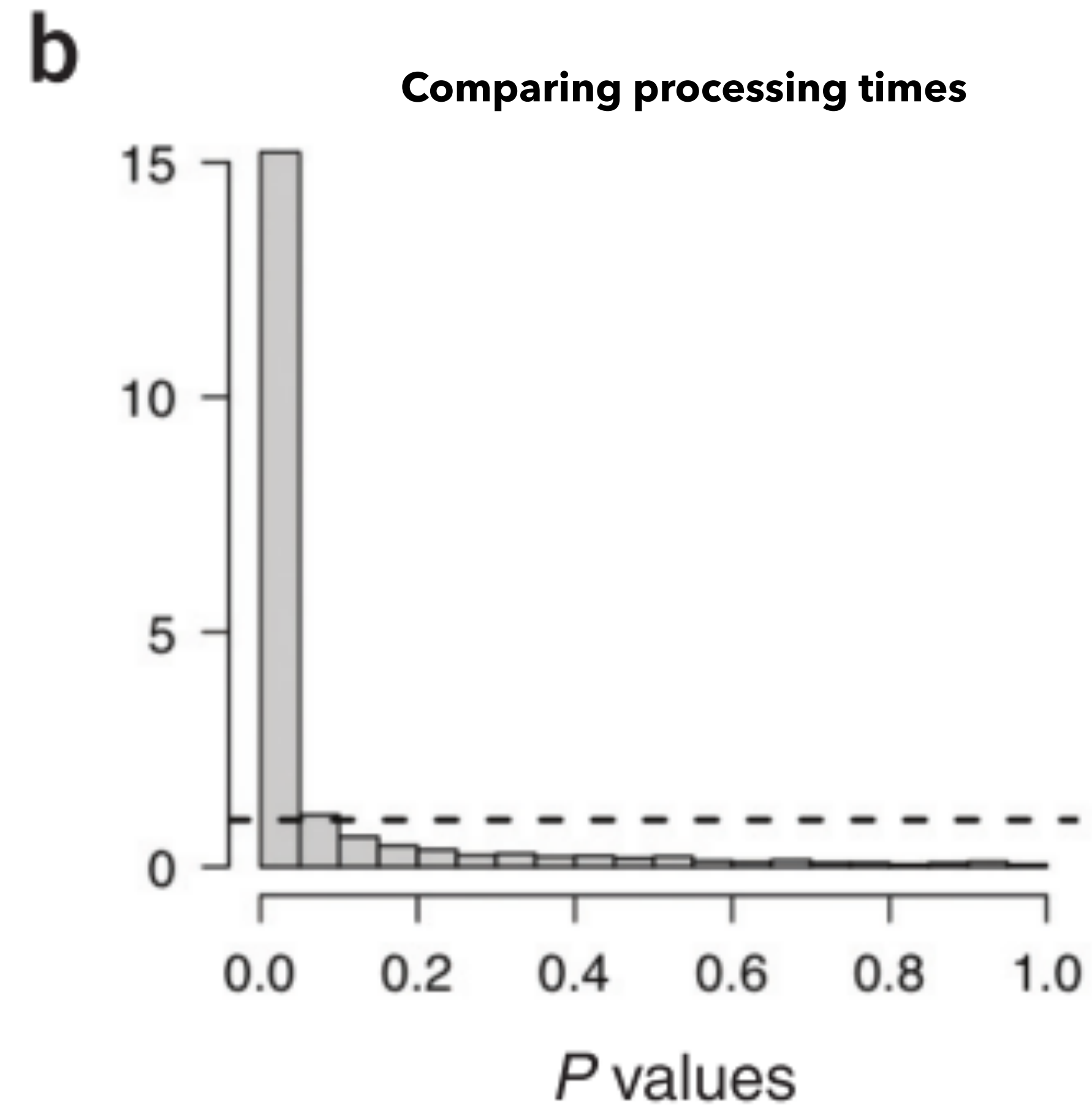
"Batch effect correction" won't work here

p-values from test comparing CEU and ASN, after controlling for the processing year
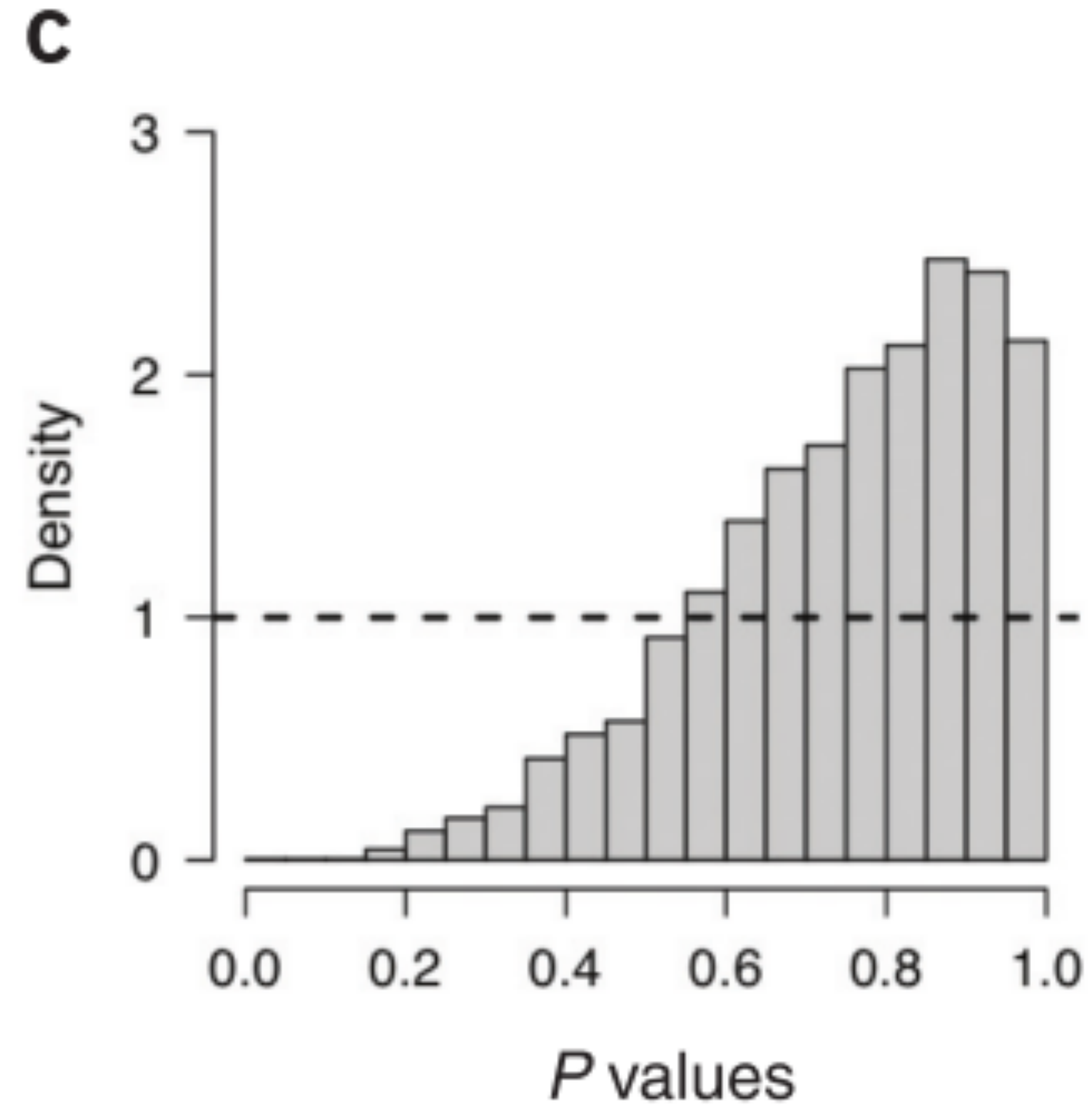
0% differentially expressed

Accounting for batch effects in practice

Public, processed RNA-seq data from 3 tissues, 4 studies show strong "study" (=batch) signal
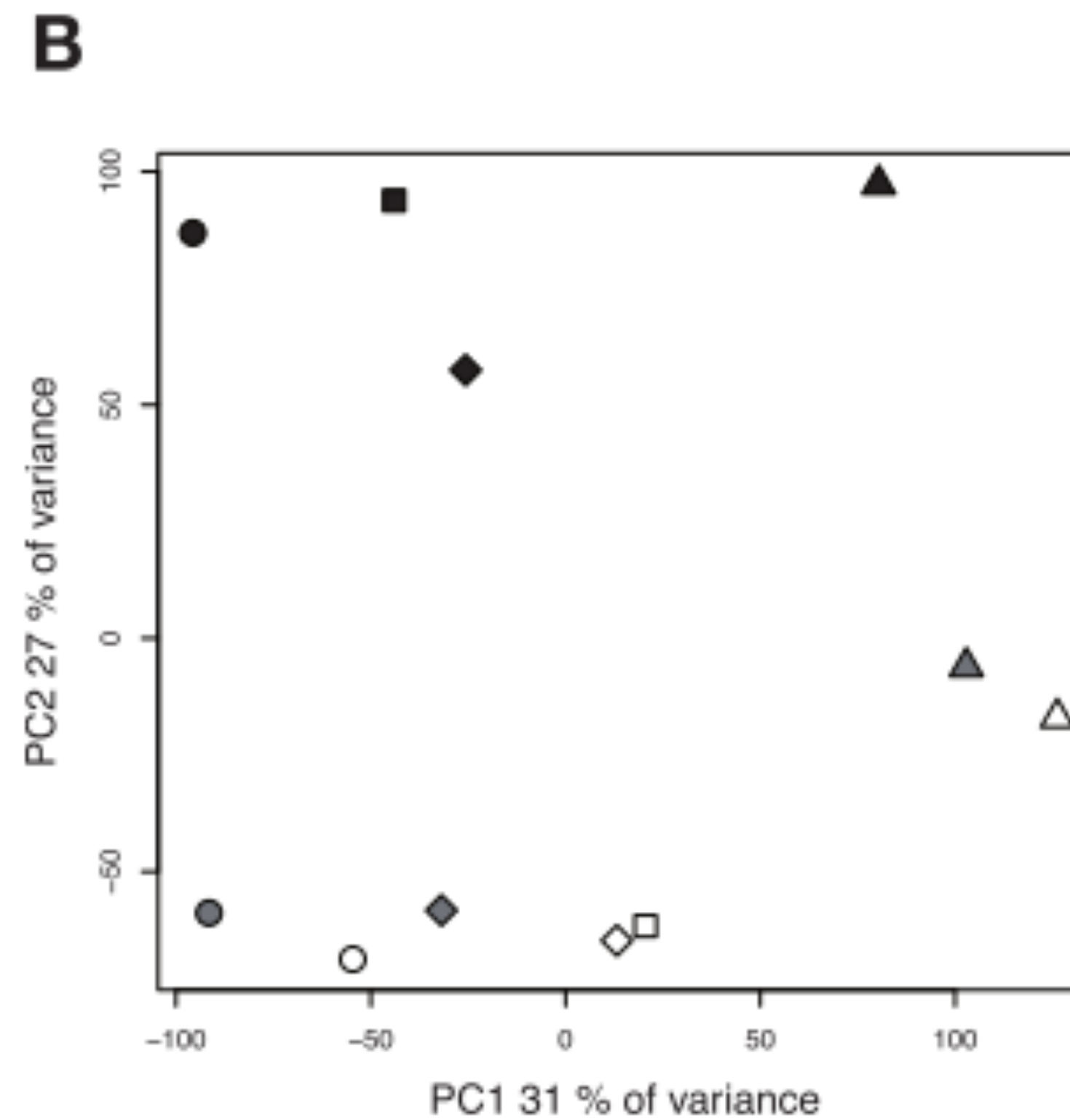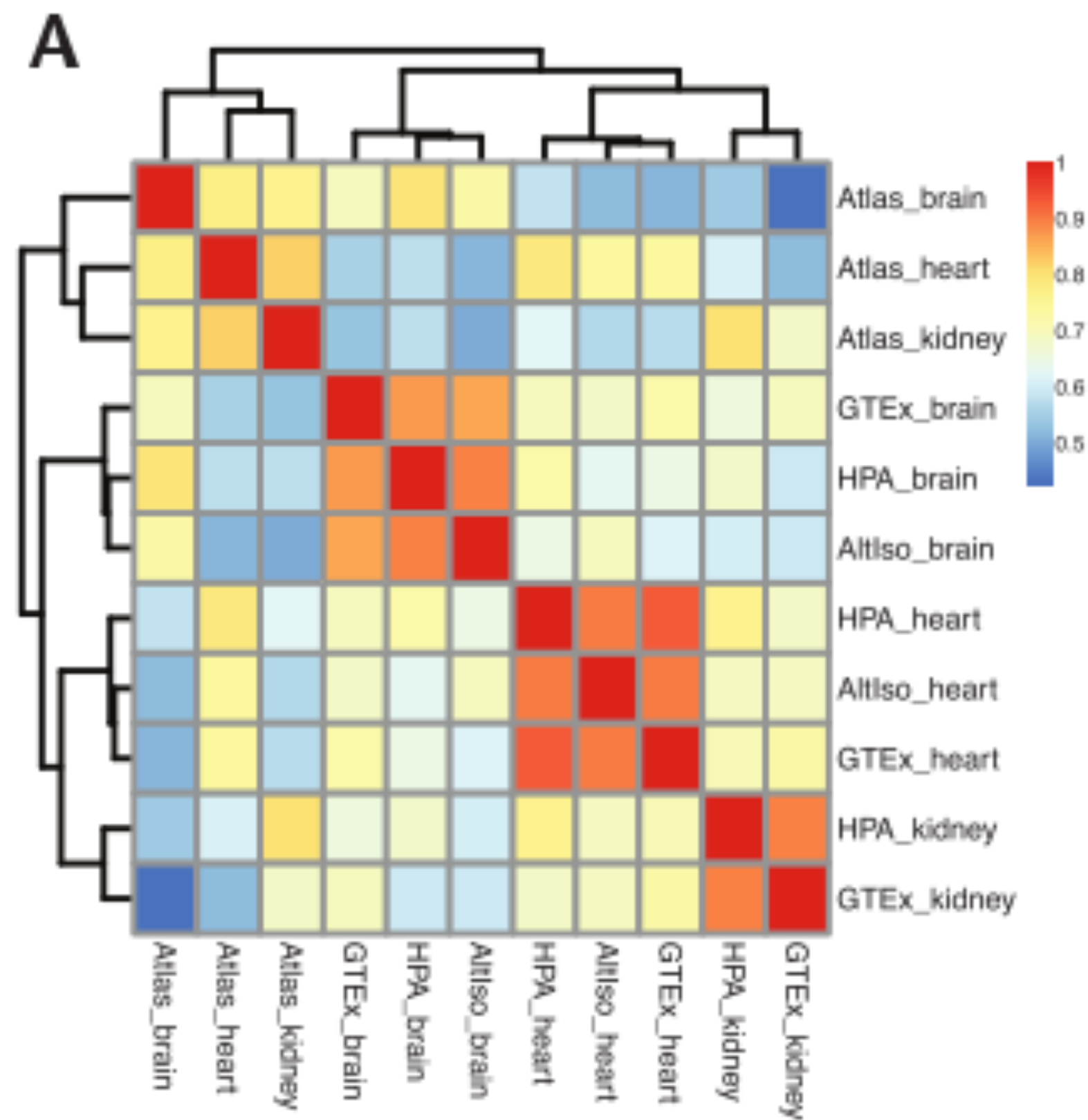
color = tissue; symbol = study (batch)

Accounting for batch effects in practice

Accounting for the batch effect brings out the signal of interest
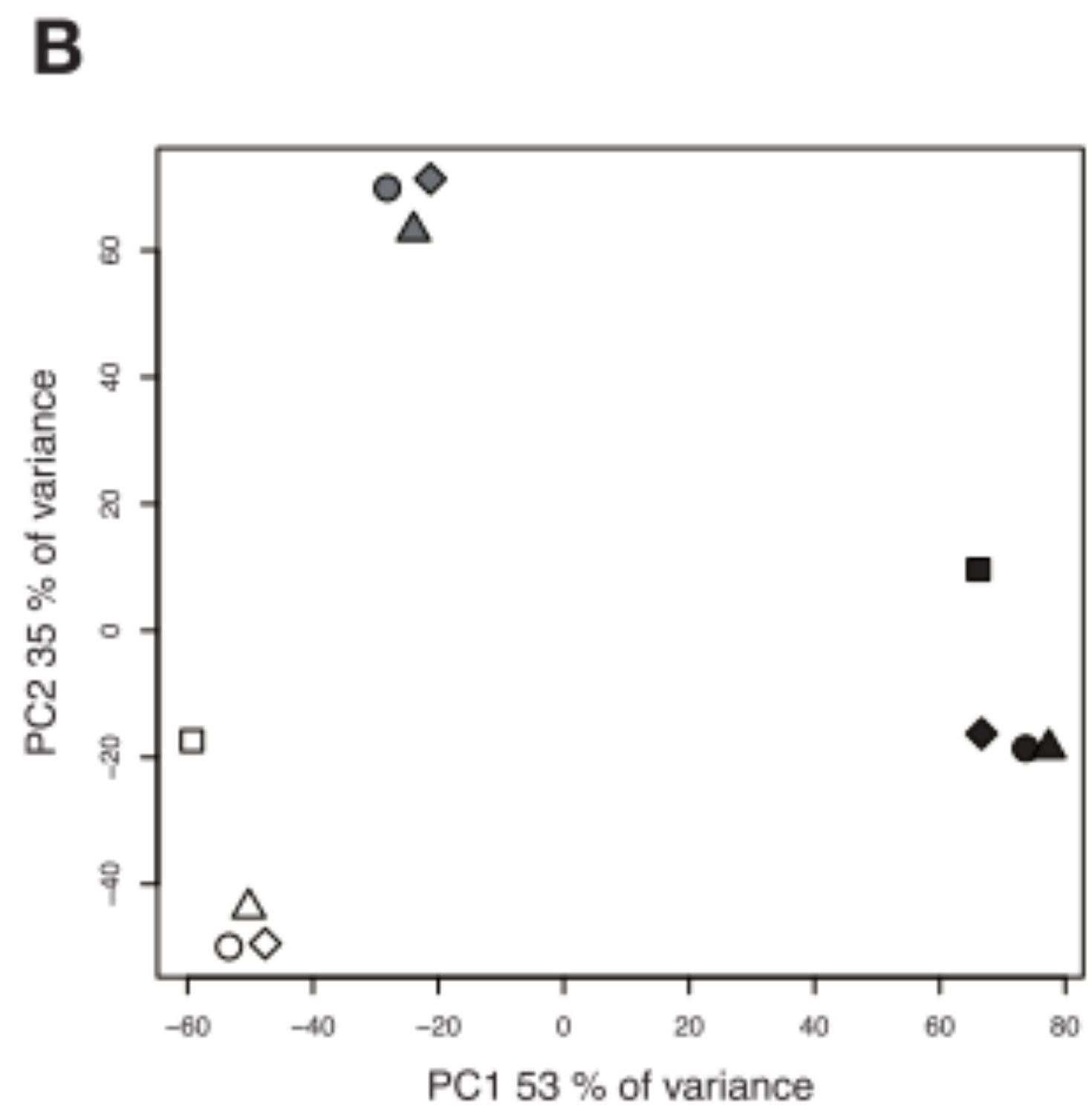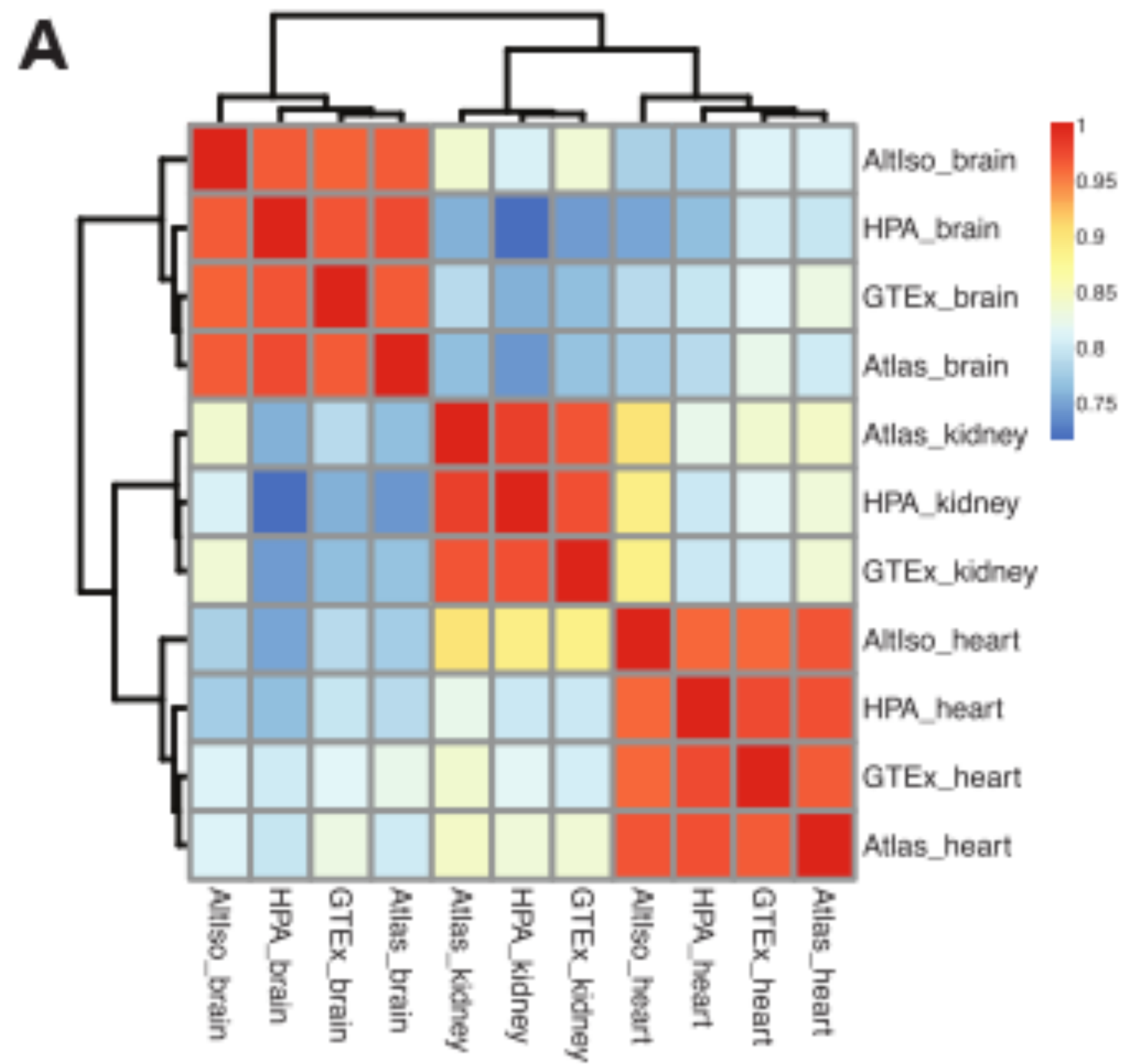
color = tissue; symbol = study (batch)

Batch effect adjustment vs normalization

Batch effect adjustment goes *beyond* the "global" between-sample
normalization methods
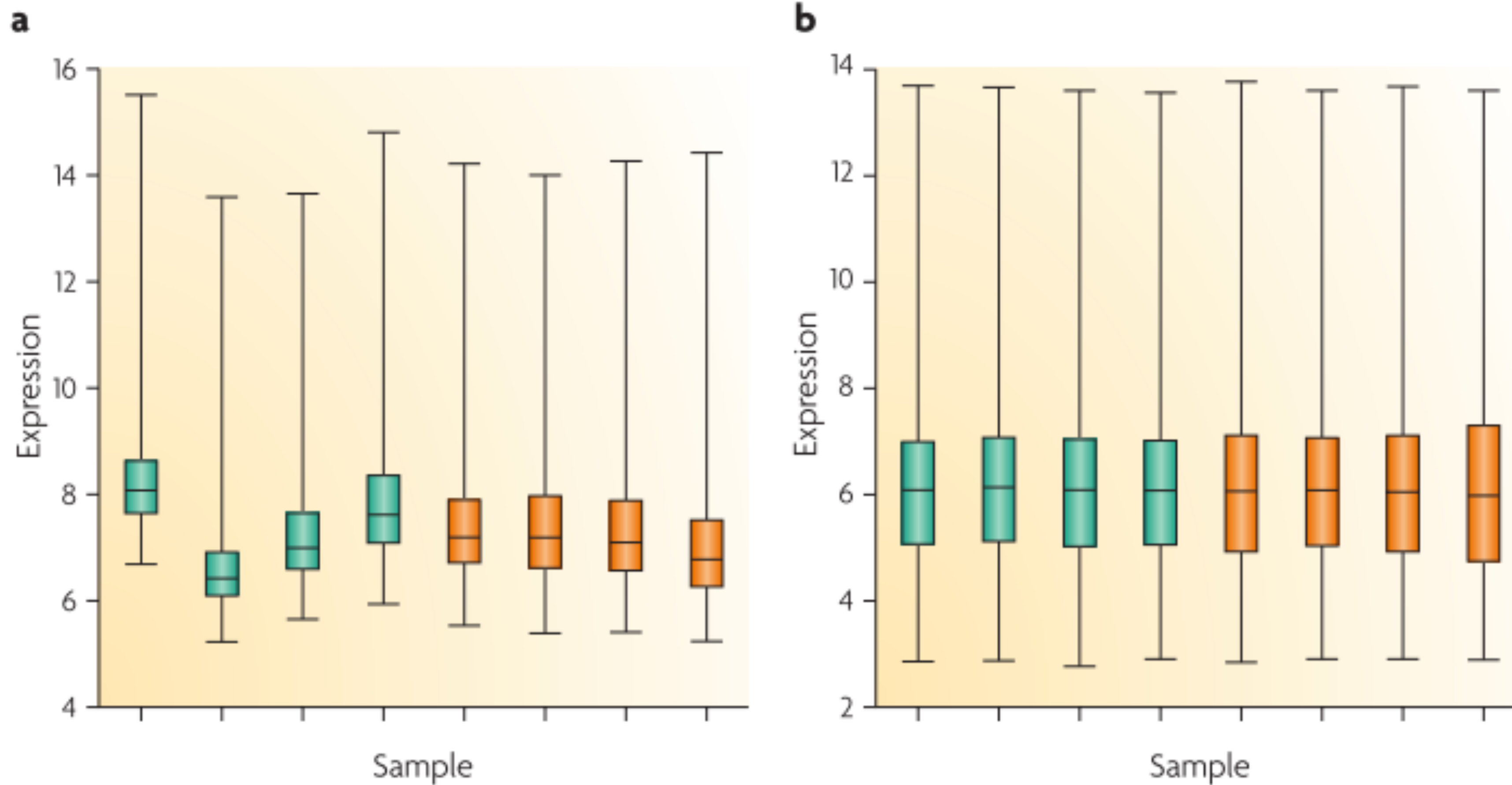
Batch effect adjustment vs normalization

Batch effect adjustment goes *beyond* the "global" between-sample normalization methods

Other design issues: replication

- Replicates are **necessary** to estimate within-condition variability.

- Variability estimates are, in turn, **vital** for statistical testing.
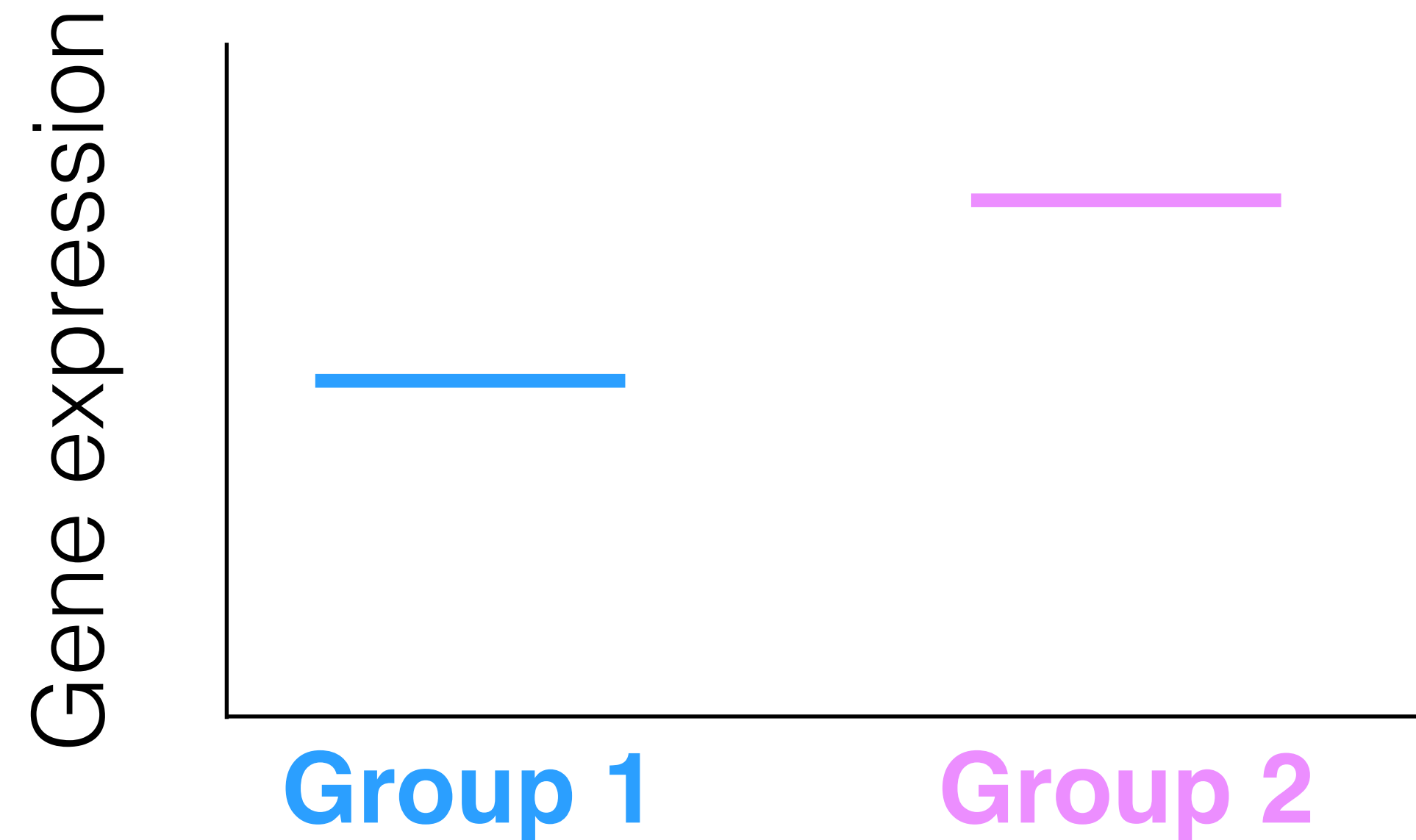
Other design issues: replication

- Replicates are **necessary** to estimate within-condition variability.

- Variability estimates are, in turn, **vital** for statistical testing.
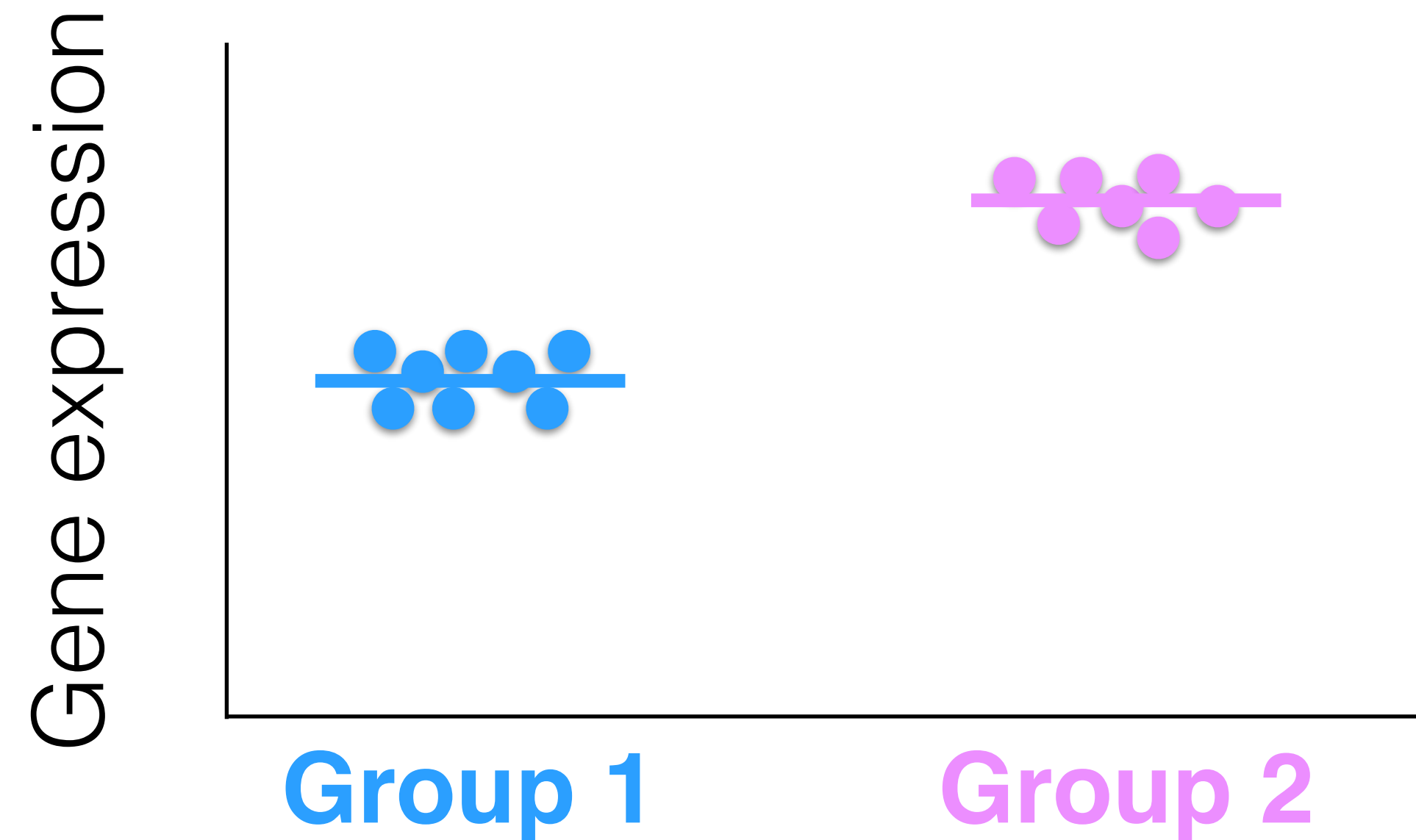
Other design issues: replication

- Replicates are **necessary** to estimate within-condition variability.

- Variability estimates are, in turn, **vital** for statistical testing.
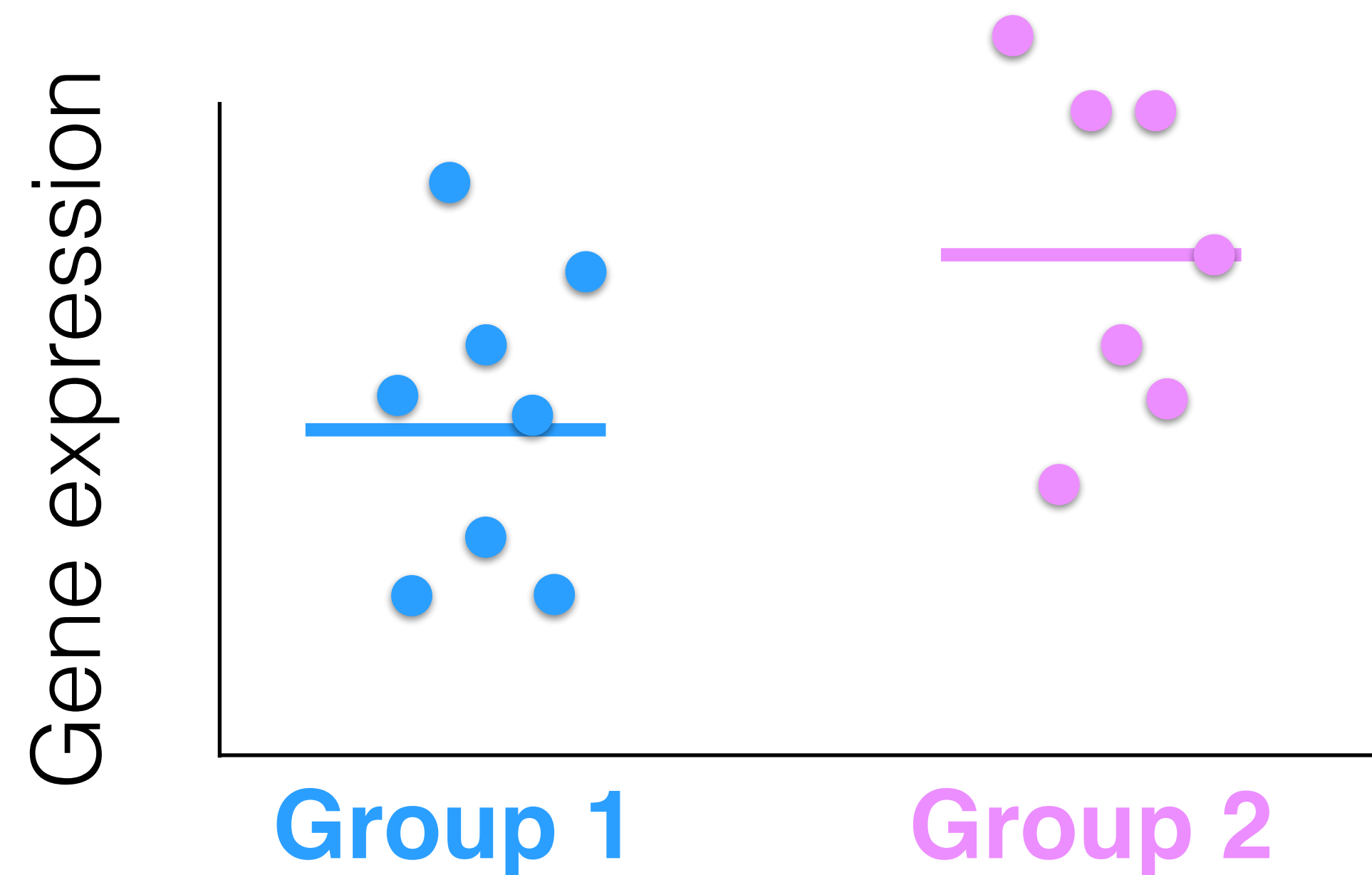
Different types of units

- Biological units (BU) - entities we want to make inferences about (e.g., animal, person)

- Experimental units (EU) - smallest entities that can be independently assigned to a treatment (e.g., animal, litter, cage, well)

- Observational units (OU) - entities at which measurements are made

Biological vs experimental units
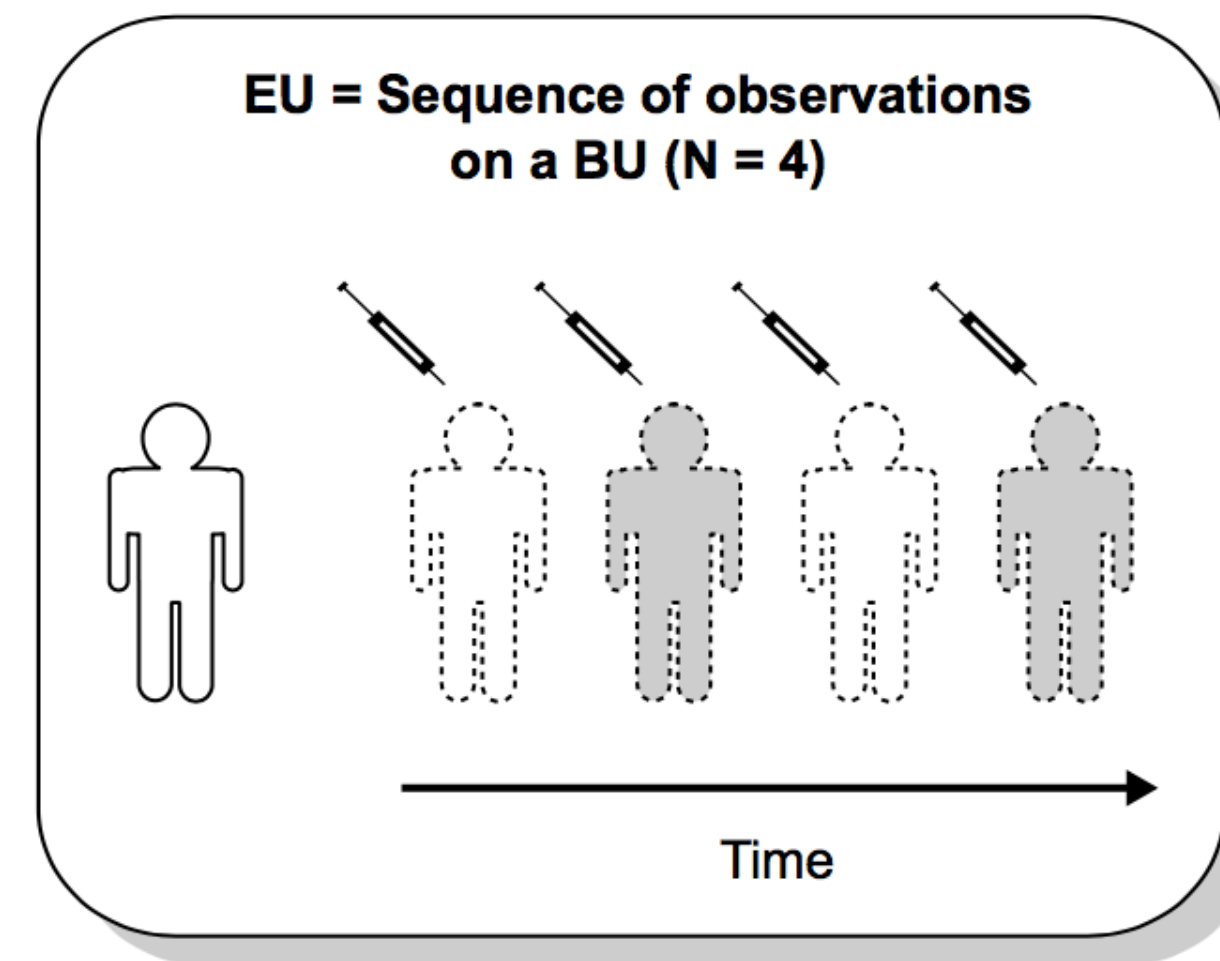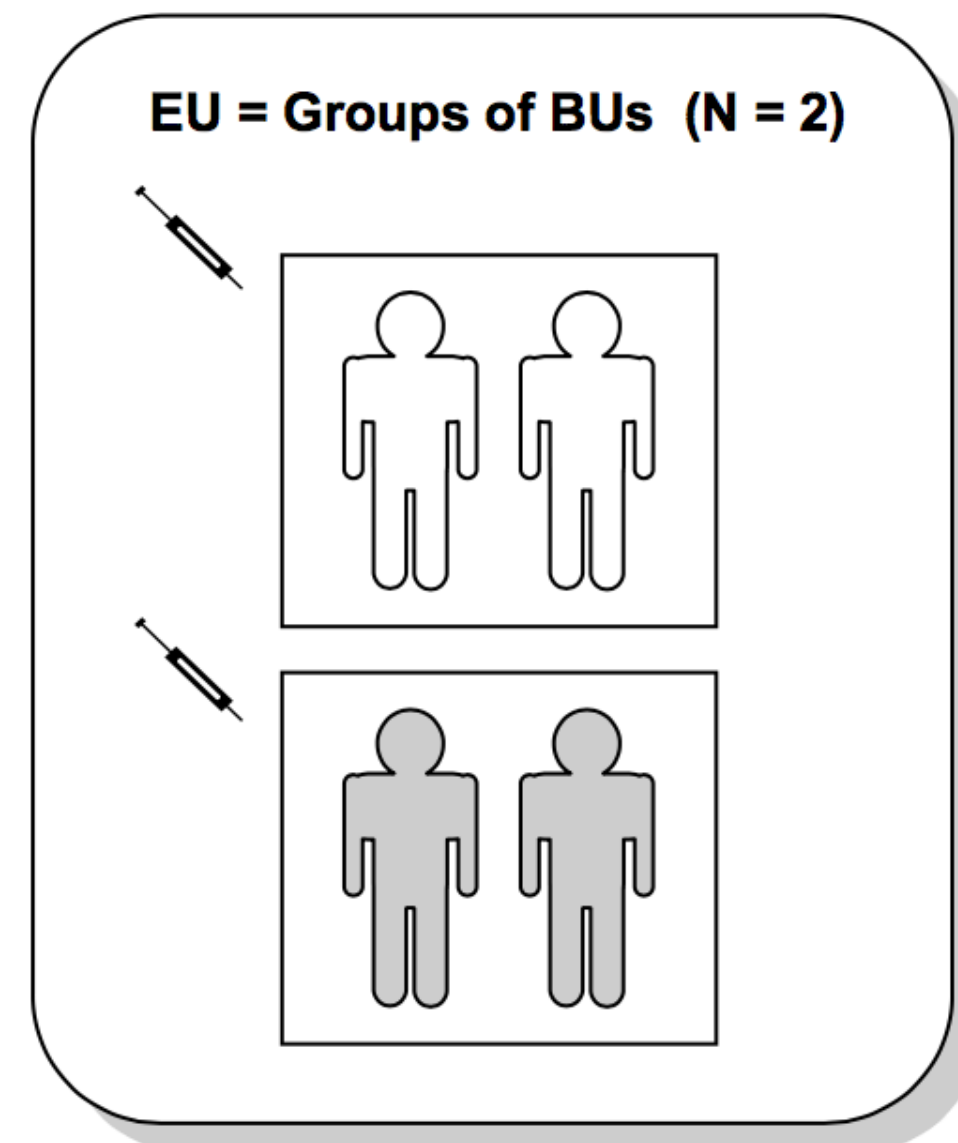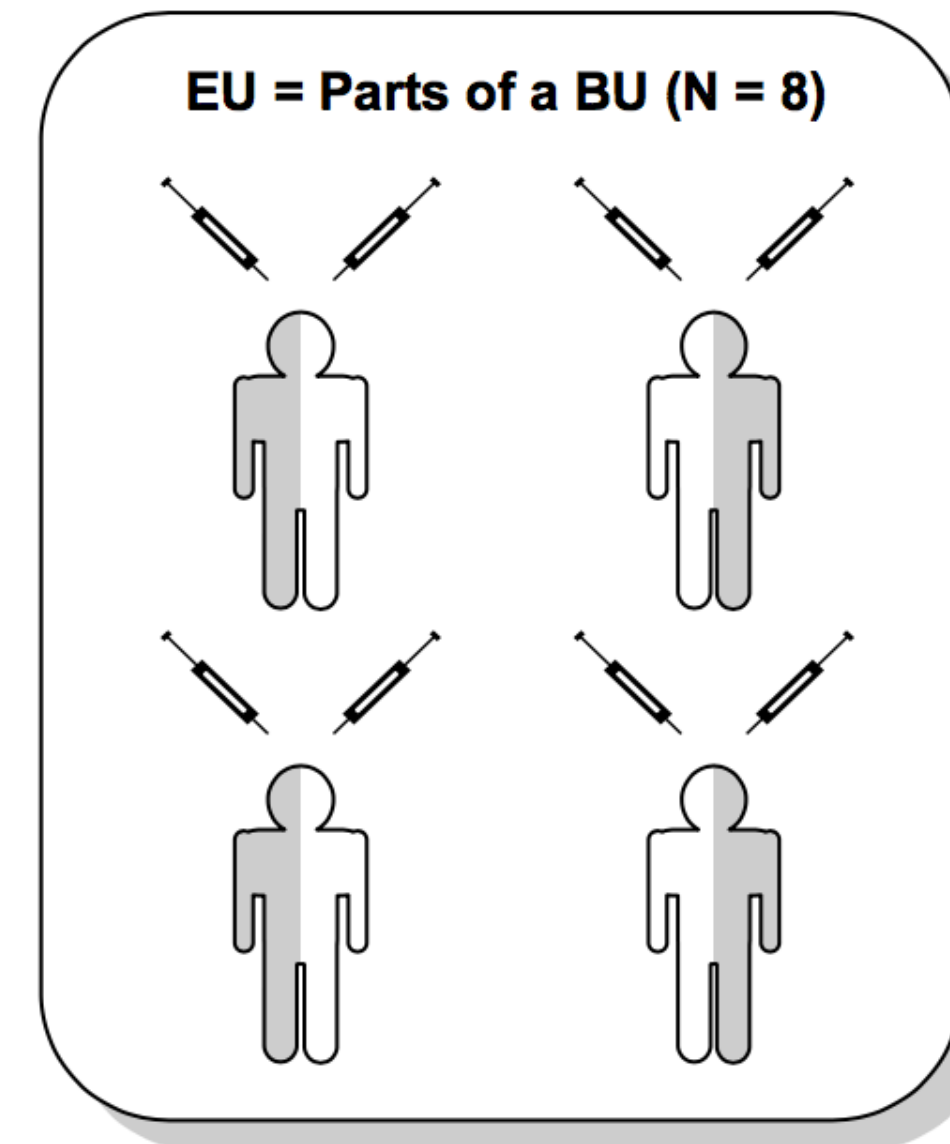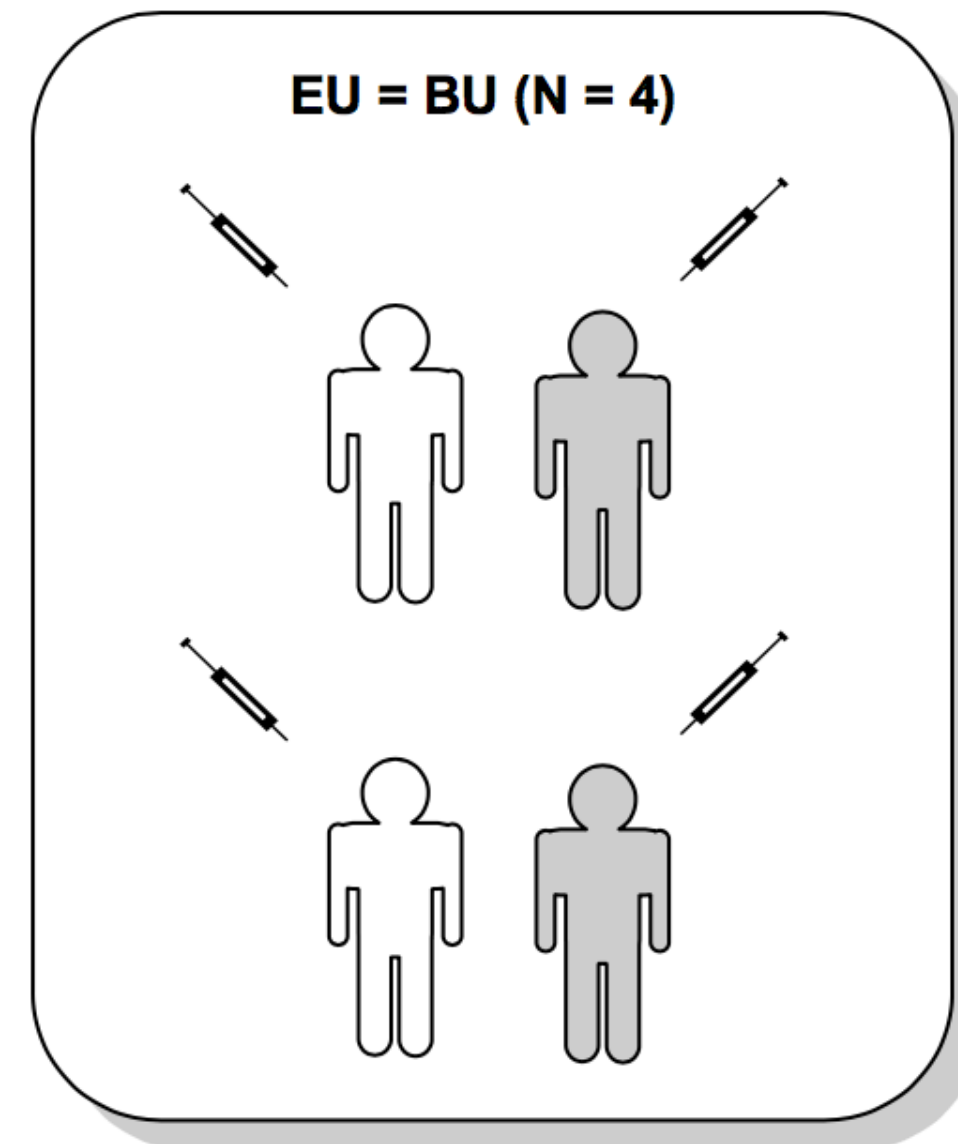
Pseudoreplication

- "**Artificial inflation** of the sample size, that usually occurs when the biological unit of interest differs from the experimental unit or observational unit."

- Only replication of experimental units is true replication

- To make a general statement about the effect of an intervention on a biological unit, we need to replicate the number of such units

[Lazic, 2016]

Model formulas and design matrices

- Testing is done separately for each gene

- We must tell the packages **which model** to fit (e.g. which predictors to use)

- The design does *not* follow "automatically" from having the sample annotation table - many different designs are often possible

- Model formulas in R:

$$\text{response variable} \ \sim \ \text{predictors}$$

- Fit a separate model for each gene - response variable changes. Specify only predictors

Examples

```r
## Linear model, mtcars data
lm(mpg ~ cyl, data = mtcars)

## Linear model (limma), gene expression data
lmFit(object = y, design = model.matrix(~ group))

## GLM (edgeR), RNA-seq data
fit <- glmFit(y = d, design = model.matrix(~ time))

## DESeq2, RNA-seq data
dds <- DESeqDataSetFromMatrix(countData = countData,
                              colData = DataFrame(condition),
                              design = ~ condition)
```

Testing and contrasts

- After fitting the model(s), we must decide *which* coefficient (or combination thereof) we want to apply a hypothesis test for.

- Combinations of coefficients are called *contrasts*.

- Design matrices can often be defined in many equivalent ways - important that the contrast is defined accordingly!

Examples

```r
## GLM (edgeR), RNA-seq data
glmLRT(fit, coef = 2)
glmLRT(fit, contrast = c(-1, 1))

## DESeq2, RNA-seq data
results(dds, contrast = c("condition", "B", "A"))
results(dds, contrast = c(0, -1, 1))
results(dds)
```

Model formulas and design matrices

- A design matrix contains the values of the predictor variables for each sample

coefficients

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix} = X \beta + \varepsilon$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

e.g.: (log) expression values for a given gene

Many ways of modeling the same expected values

- 1 predictor, 2 groups

group 1	group 2	
X	b0	b0 + b1	~X

group 1	group 2	
X	b0	b1	~0 + X

the coefficients mean different things in the different cases!

- 2 predictors, 2*2 groups

Y

	b0	b0+b1
X	b0+b2	b0+b1+b2+b3

~X*Y
~X + Y + X:Y

Y

	b0	b1
X	b2	b3

~0 + XY

New variable, combining X and Y

Y

	b0	b0+b1
X	b0+b2	b0+b2+b3

~X + X:Y
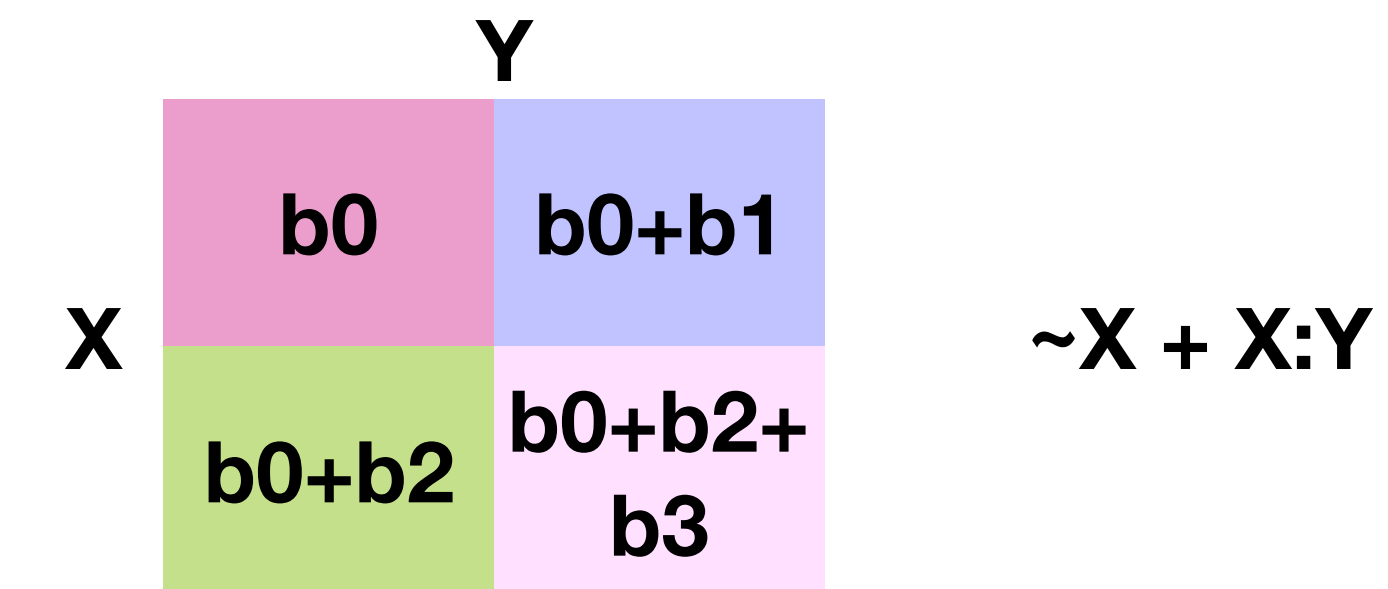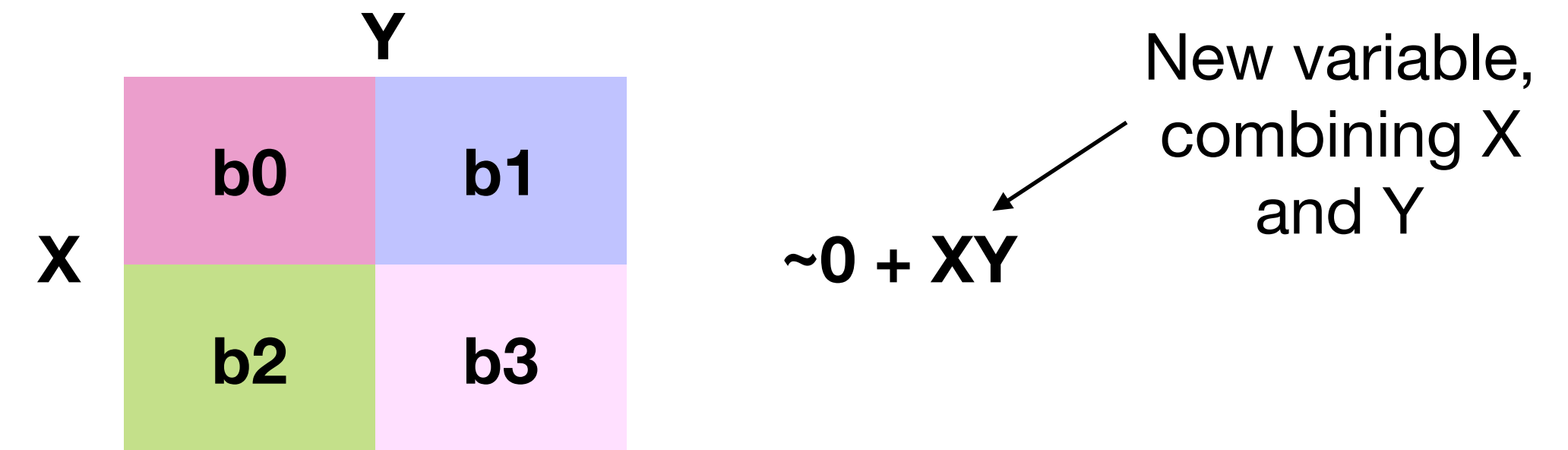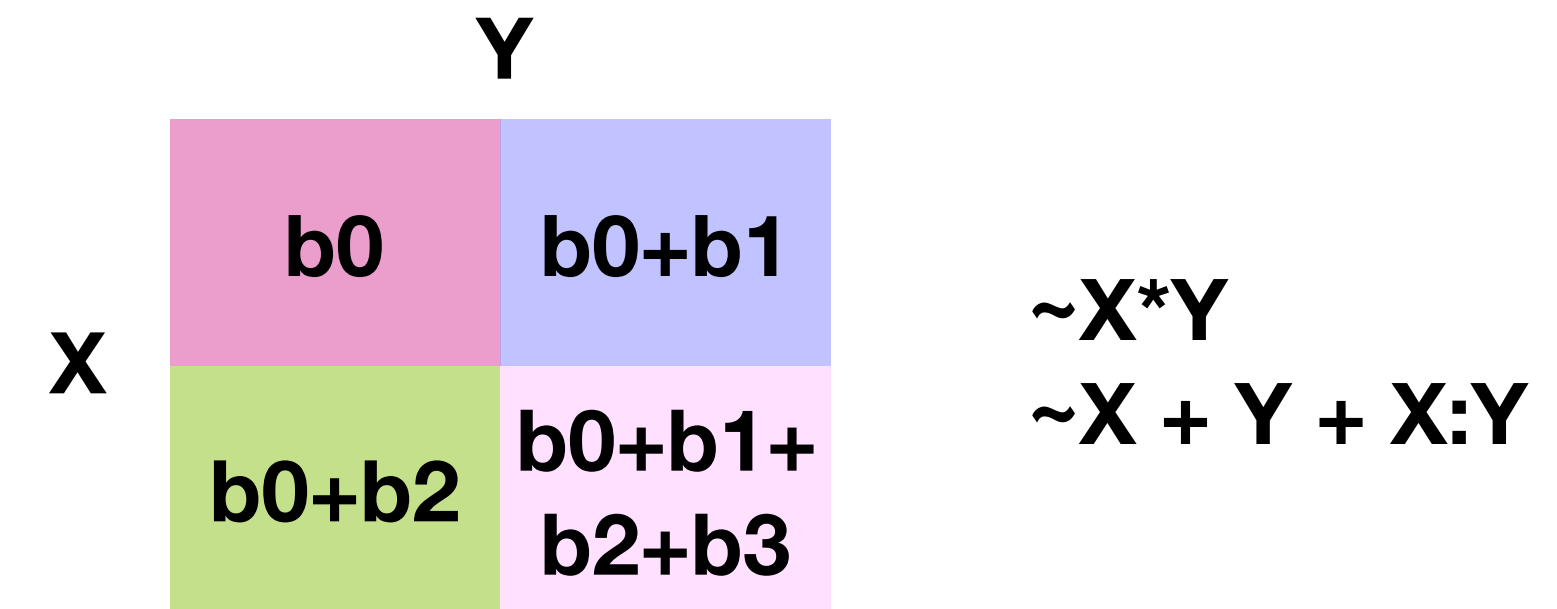
Model formulas and design matrices - example 1
One predictor, two levels (without intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

Design matrix:

	treatmentcontrol	treatmenttreated
1	1	0
2	1	0
3	1	0
4	0	1
5	0	1
6	0	1

Formula:

$$\sim 0 + treatment$$

Modeled values:

control	treated
treatmentcontrol	**treatmenttreated**

Model formulas and design matrices - example 1
One predictor, two levels (with intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

Design matrix:

	(Intercept)	treatmenttreated
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Formula:

$$\sim treatment$$

Modeled values:

control	treated
1 * Intercept + 0 * treatmenttreated	1 * Intercept + 1 * treatmenttreated

Model formulas and design matrices - example 1
One predictor, two levels (with intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

Design matrix:

	(Intercept)	treatmenttreated
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Formula:

$\sim$ treatment

Modeled values:

control	treated
1 * Intercept + 0 * treatmenttreated	1 * Intercept + 1 * treatmenttreated

Model formulas and design matrices - example 1
One predictor, two levels (with intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

Design matrix:

	(Intercept)	treatmenttreated
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Formula:

$\sim$ treatment

Modeled values:

control	treated
1 * Intercept + **0 * treatmenttreated**	**1 * Intercept +** **1 * treatmenttreated**

Model formulas and design matrices - example 1
One predictor, two levels (with intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

Design matrix:

	(Intercept)	treatmenttreated
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Formula:

$\sim \text{treatment}$

Modeled values:

control	treated
1 * Intercept + **0 * treatmenttreated**	**1 * Intercept +** **1 * treatmenttreated**

Model formulas and design matrices - example 1
One predictor, two levels (with intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

Design matrix:

	(Intercept)	treatmenttreated
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Formula:

$\sim treatment$

Modeled values:

control	treated
1 * Intercept + 0 * treatmenttreated	1 * Intercept + 1 * treatmenttreated

Model formulas and design matrices - example 1
One predictor, two levels (with intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

Design matrix:

	(Intercept)	treatmenttreated
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Formula:

$\sim$ treatment

Modeled values:

control	treated
1 * Intercept + 0 * treatmenttreated	1 * Intercept + 1 * treatmenttreated

Model formulas and design matrices - example 1
One predictor, two levels (with intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

Design matrix:

	(Intercept)	treatmenttreated
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Formula:

$\sim treatment$

Modeled values:

control	treated
1 * Intercept + 0 * treatmenttreated	1 * Intercept + 1 * treatmenttreated

Model formulas and design matrices - example 1
One predictor, two levels (with intercept)

Sample table:

```
  sample treatment
1     s1    control
2     s2    control
3     s3    control
4     s4    treated
5     s5    treated
6     s6    treated
```

Design matrix:

```
  (Intercept) treatmenttreated
1           1                0
2           1                0
3           1                0
4           1                1
5           1                1
6           1                1
```

Formula:

$\sim$ treatment

Modeled values:

control	treated
Intercept	Intercept + treatmenttreated
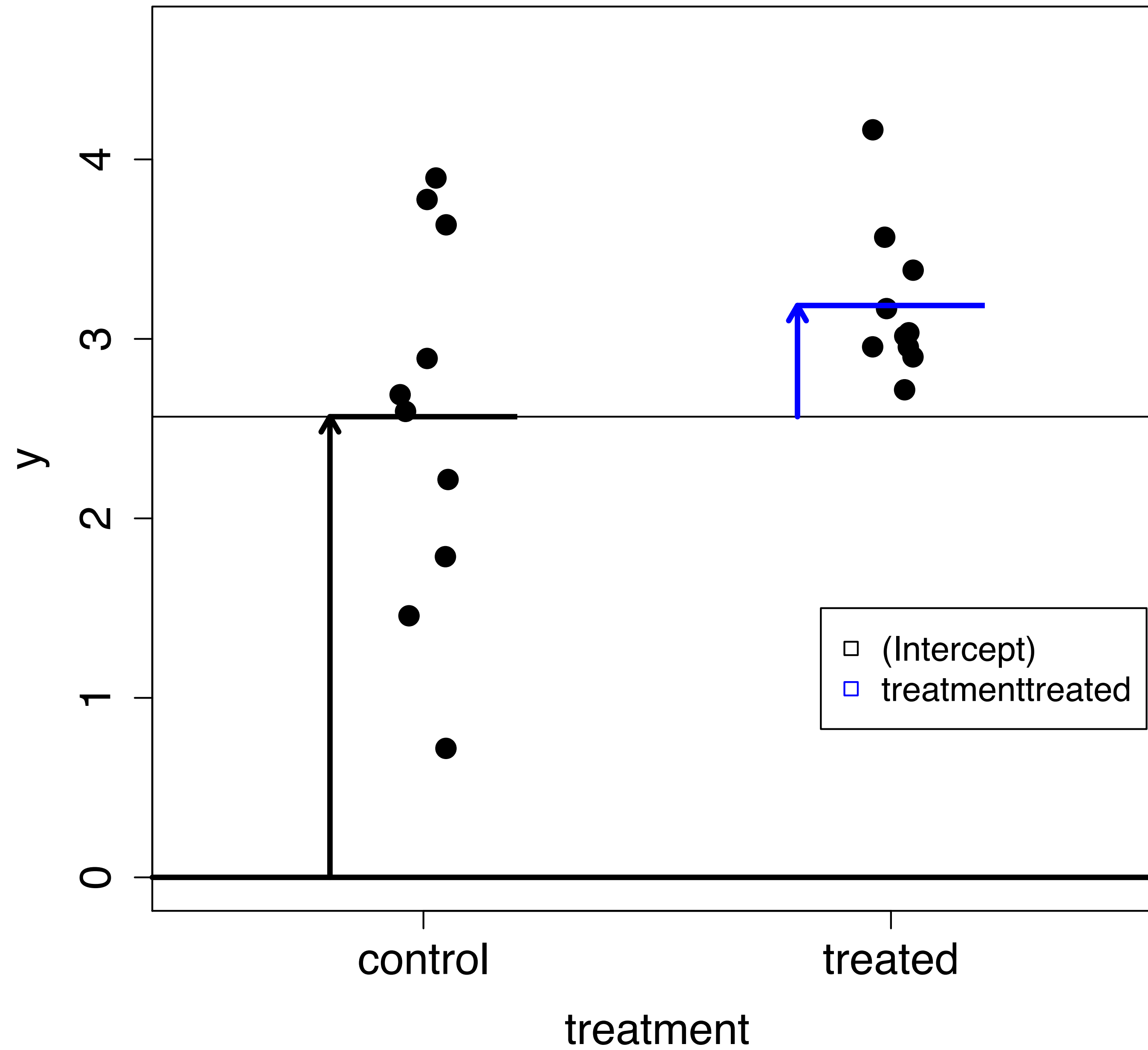
Model formulas and design matrices - example 2
One continuous predictor

Sample table:

```
   sample  age
1      s1   21
2      s2   12
3      s3   64
4      s4   44
5      s5   19
6      s6   26
```

Design matrix:

```
   (Intercept)  age
1            1   21
2            1   12
3            1   64
4            1   44
5            1   19
6            1   26
```

Formula:

$\sim$ age

Modeled values:

s1	s2	s3	s4	s5	s6
Intercept + 21 * age	Intercept + 12 * age	Intercept + 64 * age	Intercept + 44 * age	Intercept + 19 * age	Intercept + 26 * age

Model formulas and design matrices - example 3
One predictor, three levels

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	treatA
4	s4	treatA
5	s5	treatB
6	s6	treatB

Design matrix:

	(Intercept)	treatmenttreatA	treatmenttreatB
1	1	0	0
2	1	0	0
3	1	1	0
4	1	1	0
5	1	0	1
6	1	0	1

Formula:

$\sim$ treatment

Modeled values:

control	treatA	treatB
Intercept	**Intercept + treatmenttreatA**	**Intercept + treatmenttreatB**

Model formulas and design matrices - example 4
One predictor, paired data (or two predictors)

Sample table:

	sample	treatment
1	s1	control
2	s1	treated
3	s2	control
4	s2	treated
5	s3	control
6	s3	treated

Design matrix:

	(Intercept)	samples2	samples3	treatmenttreated
1	1	0	0	0
2	1	0	0	1
3	1	1	0	0
4	1	1	0	1
5	1	0	1	0
6	1	0	1	1

Formula:

$\sim \text{sample} + \text{treatment}$

Modeled values:

	s1	s2	s3
control	**Intercept**	**Intercept + samples2**	**Intercept + samples3**
treated	**Intercept + treatmenttreated**	**Intercept + samples2 + treatmenttreated**	**Intercept + samples3 + treatmenttreated**

Model formulas and design matrices - example 4
One predictor, paired data (or two predictors)

Sample table:

```
    genotype treatment
1          A   control
2          A   control
3          A   treated
4          A   treated
5          B   control
6          B   control
7          B   treated
8          B   treated
```

Design matrix:

```
  (Intercept)  genotypeB  treatmenttreated
1           1          0                 0
2           1          0                 0
3           1          0                 1
4           1          0                 1
5           1          1                 0
6           1          1                 0
7           1          1                 1
8           1          1                 1
```
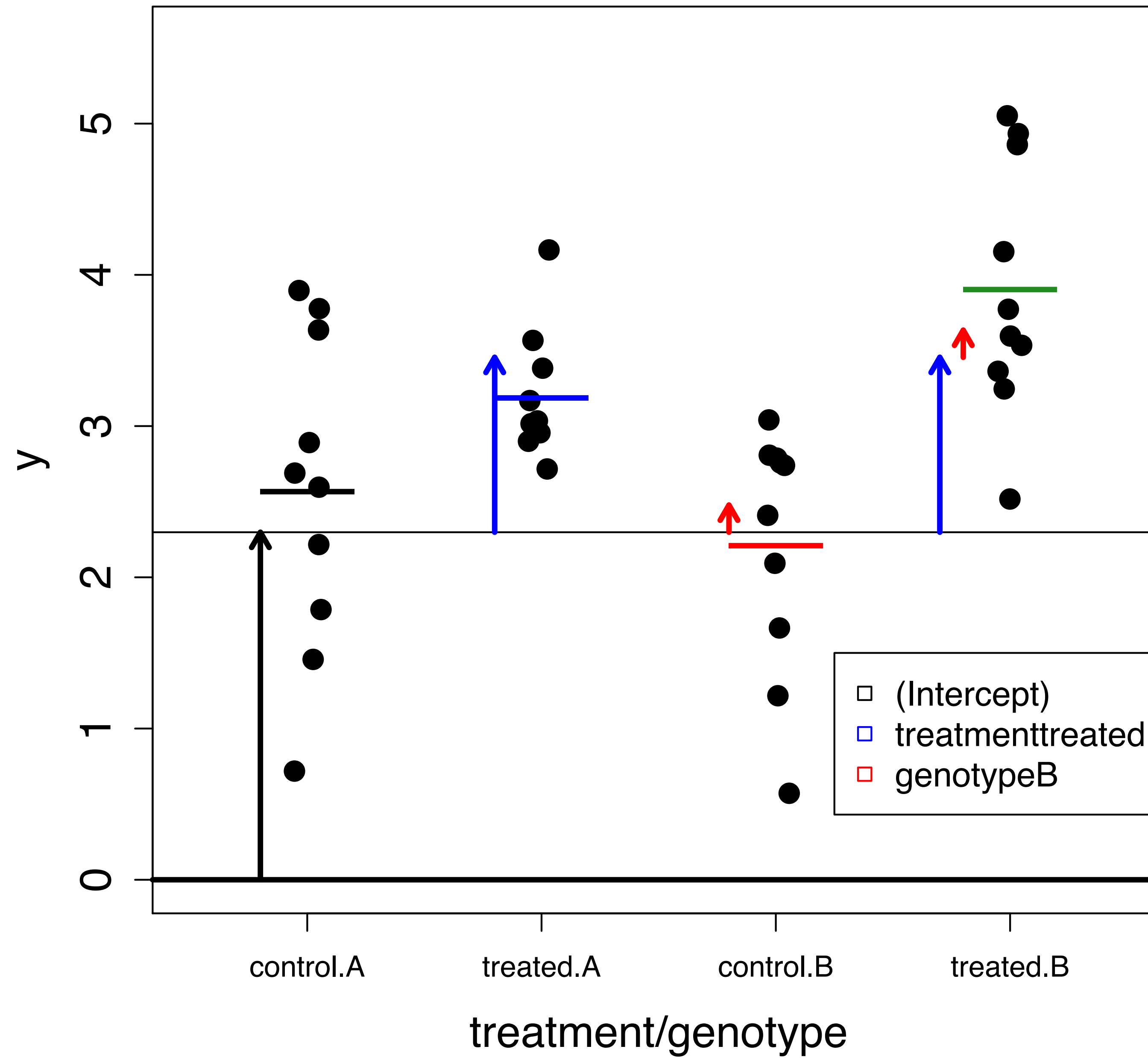
Formula:

$$\sim \text{genotype} \ + \ \text{treatment}$$

Modeled values:

	genotype A	genotype B
control	**Intercept**	**Intercept + genotypeB**
treated	**Intercept + treatmenttreated**	**Intercept + genotypeB + treatmenttreated**

Model formulas and design matrices - example 5
Two predictors, with interaction

Sample table:

	genotype	treatment
1	A	control
2	A	control
3	A	treated
4	A	treated
5	B	control
6	B	control
7	B	treated
8	B	treated

Design matrix:

	(Intercept)	genotypeB	treatmenttreated	genotypeB:treatmenttreated
1	1	0	0	0
2	1	0	0	0
3	1	0	1	0
4	1	0	1	0
5	1	1	0	0
6	1	1	0	0
7	1	1	1	1
8	1	1	1	1

Formula:

~ genotype * treatment

~ genotype + treatment + genotype:treatment

Modeled values:

	genotype A	genotype B
control	Intercept	Intercept + genotypeB
treated	Intercept + treatmenttreated	Intercept + genotypeB + treatmenttreated + genotypeB:treatmenttreated

Model formulas and design matrices - example 6
Two predictors, with interaction

Sample table:

```
       treat.gt
1  control.A
2  control.A
3  treated.A
4  treated.A
5  control.B
6  control.B
7  treated.B
8  treated.B
```
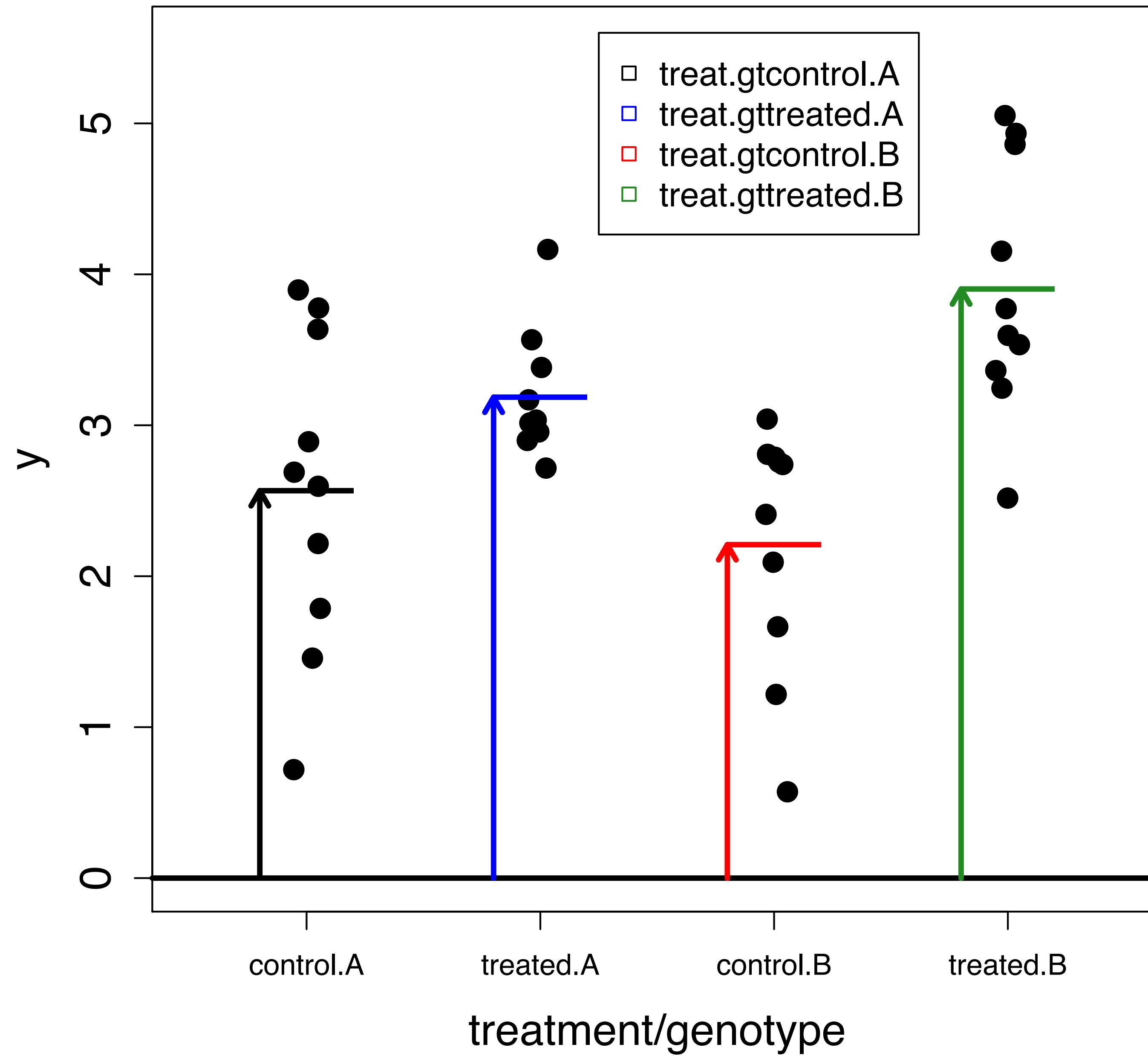
Design matrix:

	treat.gtcontrol.A	treat.gttreated.A	treat.gtcontrol.B	treat.gttreated.B
1	1	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	1	0	0
5	0	0	1	0
6	0	0	1	0
7	0	0	0	1
8	0	0	0	1

Formula:

$$\sim 0 + \text{treat.gt}$$

Modeled values:

	genotype A	genotype B
control	**treat.gtcontrol.A**	**treat.gtcontrol.B**
treated	**treat.gttreated.A**	**treat.gttreated.B**

References

- Akay et al.: On the design and analysis of gene expression studies in human populations. Nature Genetics 39(7):807-808 (2007)
- Nygaard et al.: Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics 17(1):29-39 (2016)
- Danielsson et al.: Assessing the consistency of public human tissue RNA-seq data sets. Briefings in Bioinformatics 16(6):941-949 (2015)
- Leek et al.: Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics 11(10:733-739 (2010)
- Schurch et al.: How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA (2016)
- Lazic: Experimental Design for Laboratory Biologists: Maximising Information and Improving Reproducibility. Cambridge University Press (2016).