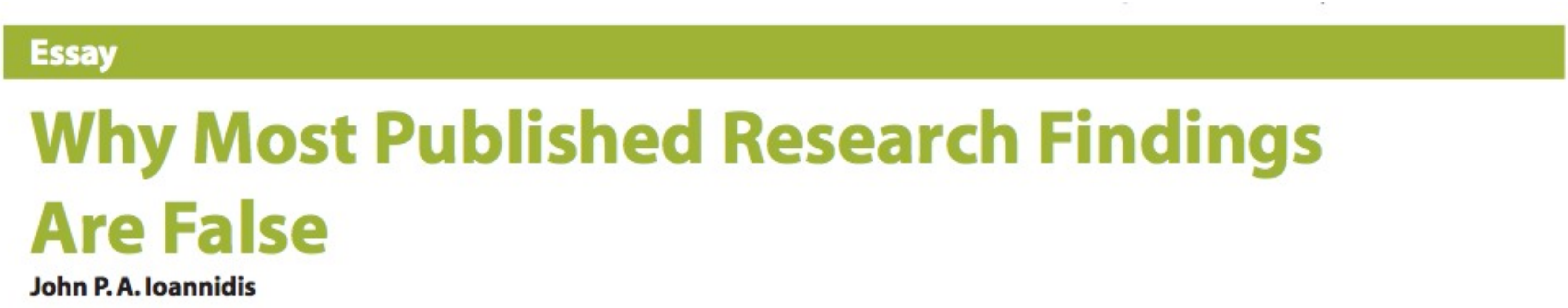
umcg | Department of genetics

Lude Franke >

Experimental design, dealing with confounders, multiple-testing correction and platform specific issues that can cause false-positives

Department of Genetics, UMC Groningen

Experimental design



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

http://www.youtube.com/watch?v=hBNeuG10-ac

PLoS Medicine 2005

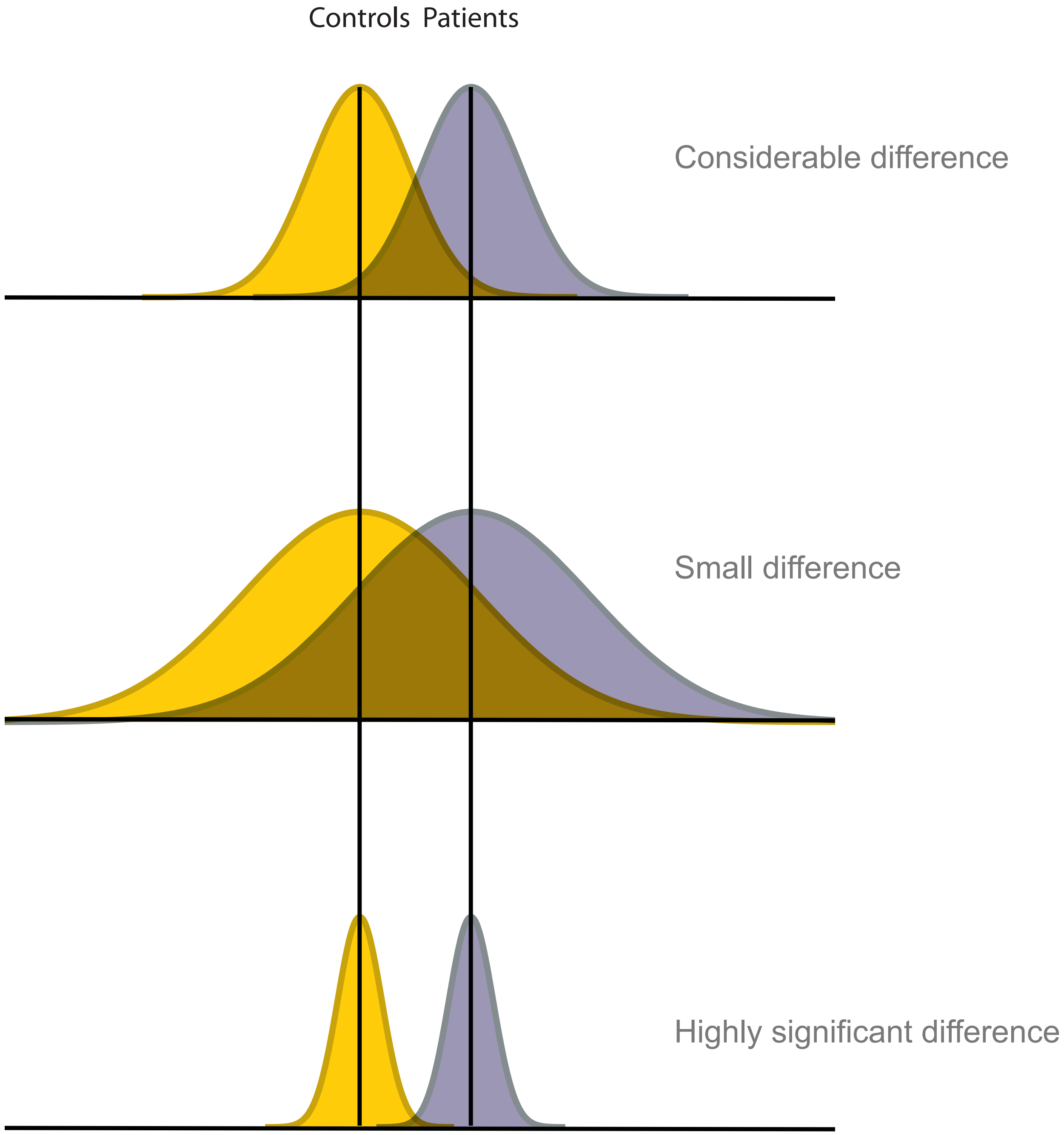
Experimental design

- We often compare two groups with each other (e.g. clinical trial: treat patients with drug or placebo, ascertain whether drug has an effect)
- The traditional scientific strategy is to change one parameter (the independent variable) and assess whether that variable has an effect on the dependent variable
- However, when dealing with genomic data we typically measure thousands of parameters, we can continue testing whatever we think is interesting.
- But how do we then correct for multiple testing?
- However, many confounders exist, but sometimes it is not even evident they exist, can we identify them?

Multiple testing correction

Multiple testing correction

Experimental design



In words:

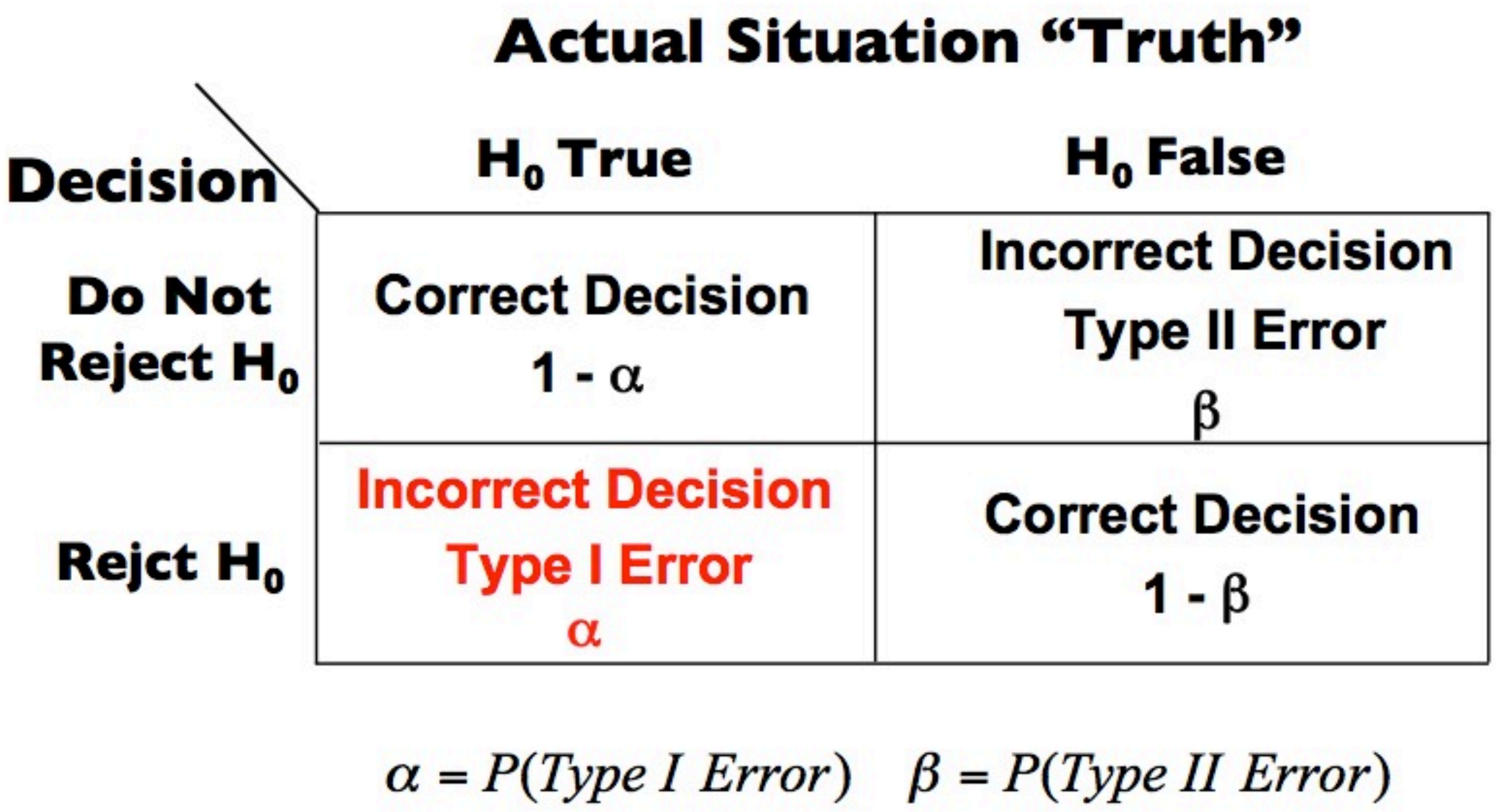
$$t = \frac{\text{difference of means}}{\text{variability}}$$

In an equation
(T = patients, C = controls):

$$t = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}}$$

→ P-value (using a T - distribution)

Type I and Type II error



Decision	Actual Situation "Truth" H_0 True	 H_0 False
Do Not Reject H_0	Correct Decision $1 - \alpha$	Incorrect Decision Type II Error β
Rejct H_0	Incorrect Decision Type I Error α	Correct Decision $1 - \beta$

$\alpha = P(Type\ I\ Error)$ $\beta = P(Type\ II\ Error)$

Why multiple testing matters

Genomics = Lots of Data = Lots of Hypothesis Tests
A typical microarray experiment might result in performing 10,000 separate hypothesis tests. If we use a standard p-value cut-off of 0.05, we'd expect ***500*** genes to be deemed "significant" by **chance**.

In general, if we perform m hypothesis tests, what is the probability of at least 1 false positive?

P(Making an error) = α

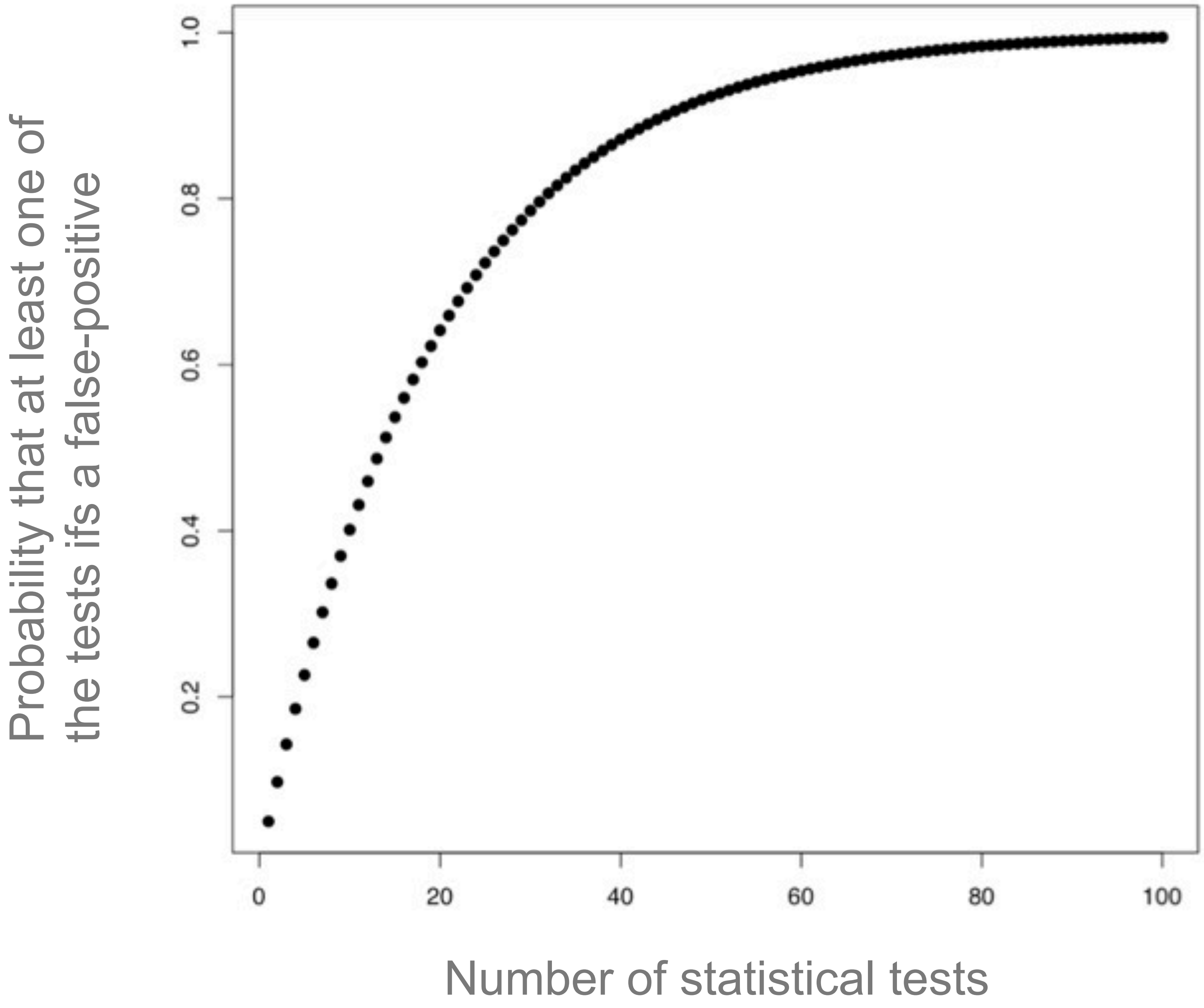
P(Not making an error) = 1 - α

P(Not making an error in m tests) = $(1 - \alpha)^m$

P(Making at least 1 error in m tests) = $1 - (1 - \alpha)^m$

Probability of at least one false-positive

When assuming that a test with P < 0.05 is significant:



Correcting for multiple testing

Bonferroni correction: Correct for the number of tests, by multiplying each P-Value with the number of statistical tests (overly stringent: High probability of type 2 errors, i.e. of not rejecting the general null hypothesis when important effects exist)

Holms method:
Order the unadjusted *p-values* such that $p1 \leq p2 \leq ... \leq pm$

- Holm adjusted p-values are: *p* ˜*j* =min[(m"j+1)•pj,1]
- The point here is that we don't multiply every *pi* by the same factor m:

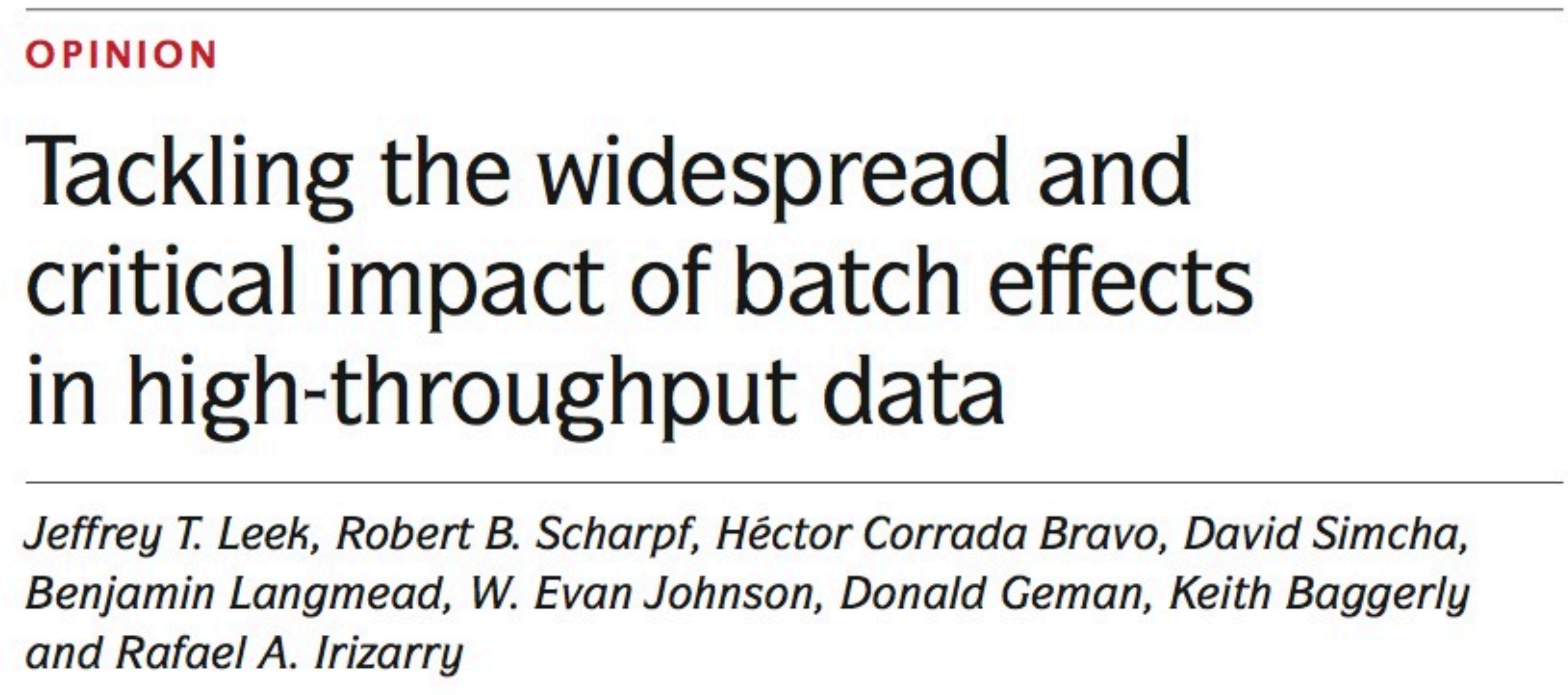
$\tilde{p}1 = 10000 * p1, \tilde{p}2 = 9999 * p2, \ldots , \tilde{p}m = 1 * pm$

Many other methods exist:
- False discovery rate (FDR)
- Benjamini and Hochberg FDR
- Storey's positive FDR
- Permutation based methods to account for correlated tests

Batch effects

Batch effects

Batch effects



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 and Rafael A. Irizarry

NATURE REVIEWS | GENETICS
VOLUME 11 | OCTOBER 2010 | 733

Batch effects

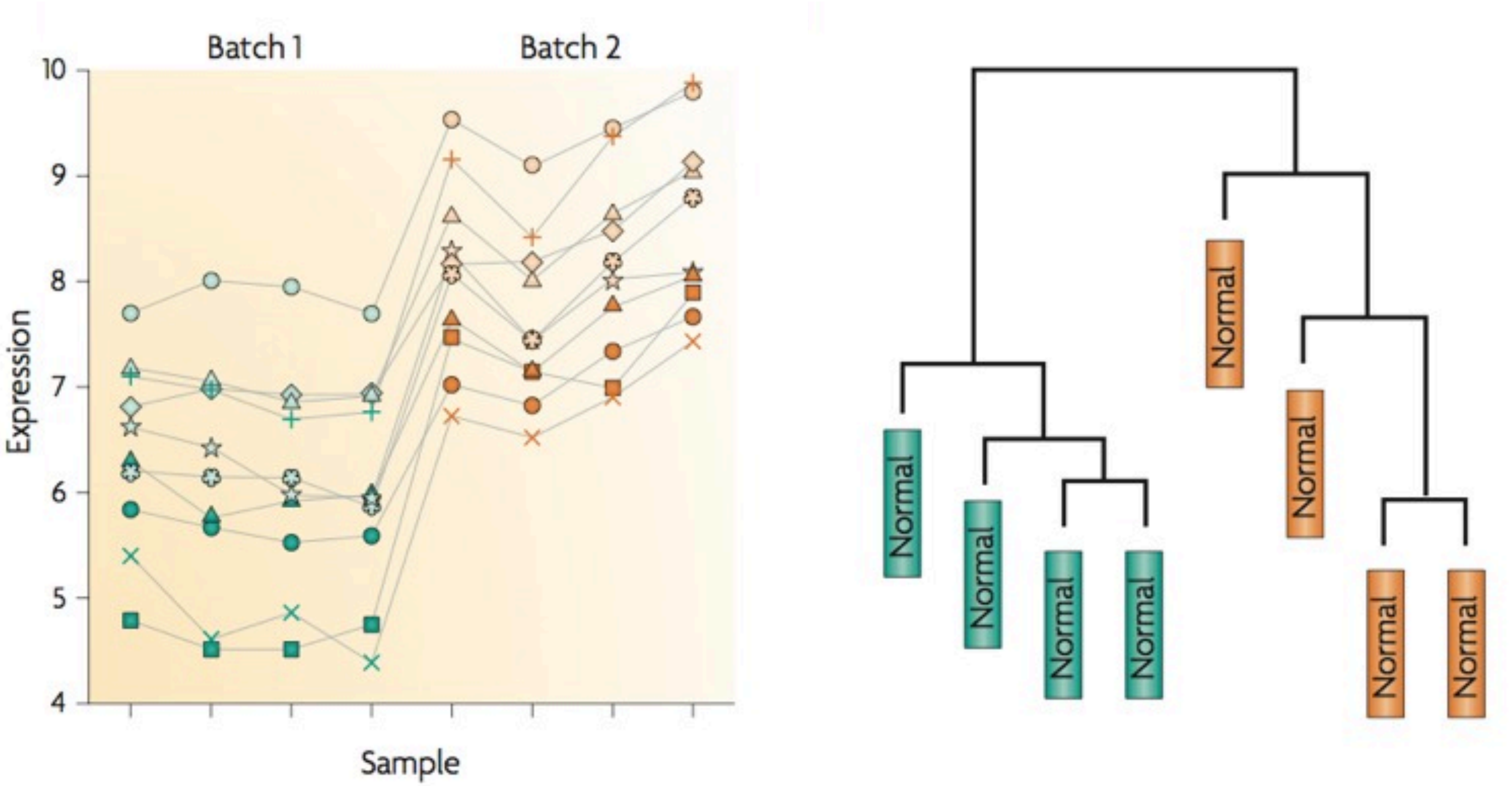
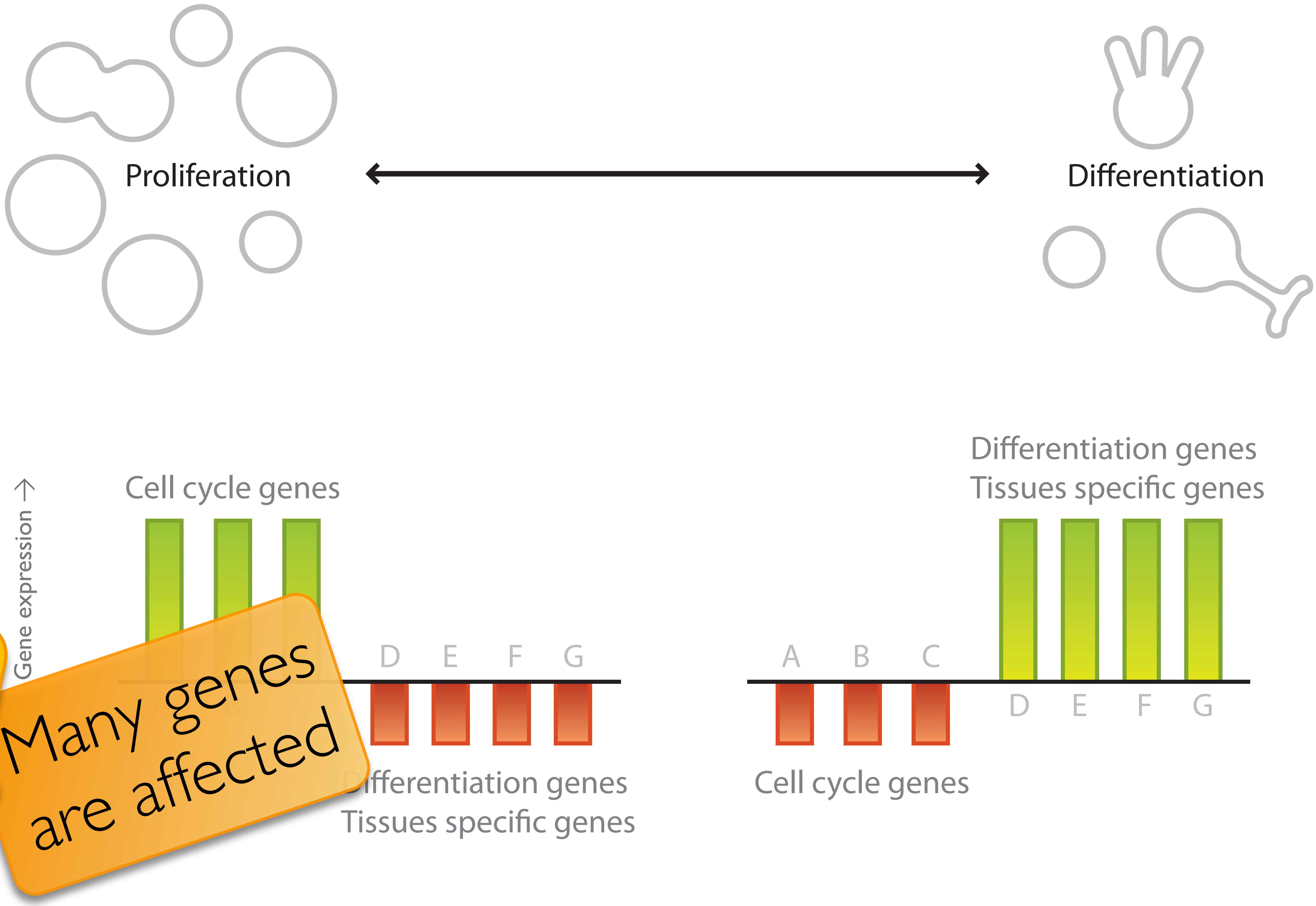


Figure 1 | **Demonstration of normalization and surviving batch effects.** For a published bladder cancer microarray data set obtained using an Affymetrix platform[9], we obtained the raw data for only the normal samples. Here, green and orange represent two different processing dates. **a** | Box plot of raw gene expression data (log base 2). **b** | Box plot of data processed with RMA, a widely used preprocessing algorithm for Affymetrix data[27]. RMA applies quantile normalization — a technique that forces the distribution of the raw signal intensities from the microarray data to be the same in all samples[28]. **c** | Example of ten genes that are susceptible to batch effects even after normalization. Hundreds of genes show similar behaviour but, for clarity, are not shown. **d** | Clustering of samples after normalization. Note that the samples perfectly cluster by processing date.

NATURE REVIEWS | GENETICS
VOLUME 11 | OCTOBER 2010 | 733

Batch effects: principal component analysis



Proliferation

Differentiation

Cell cycle genes

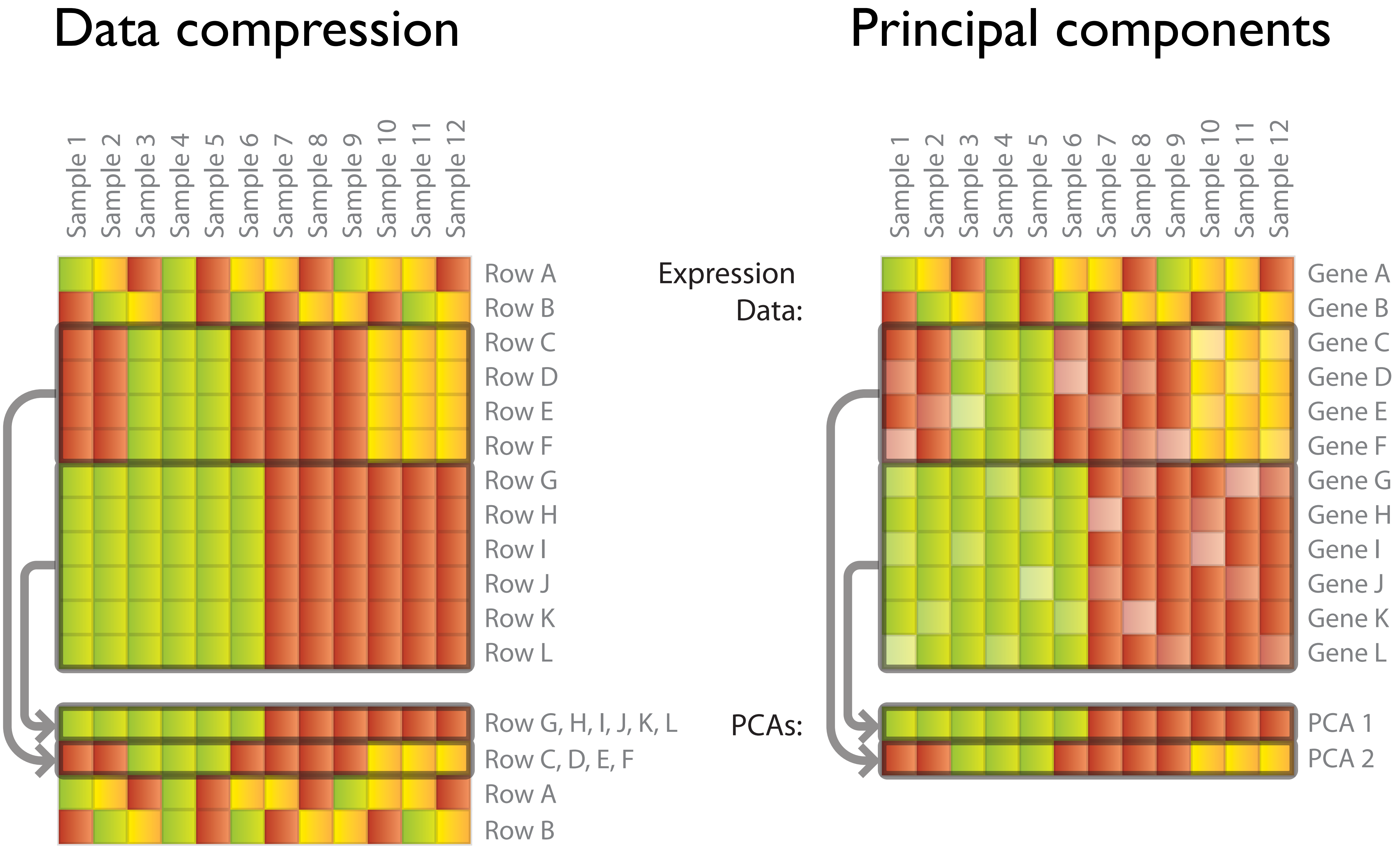
Differentiation genes
Tissues specific genes

Differentiation genes
Tissues specific genes

Cell cycle genes

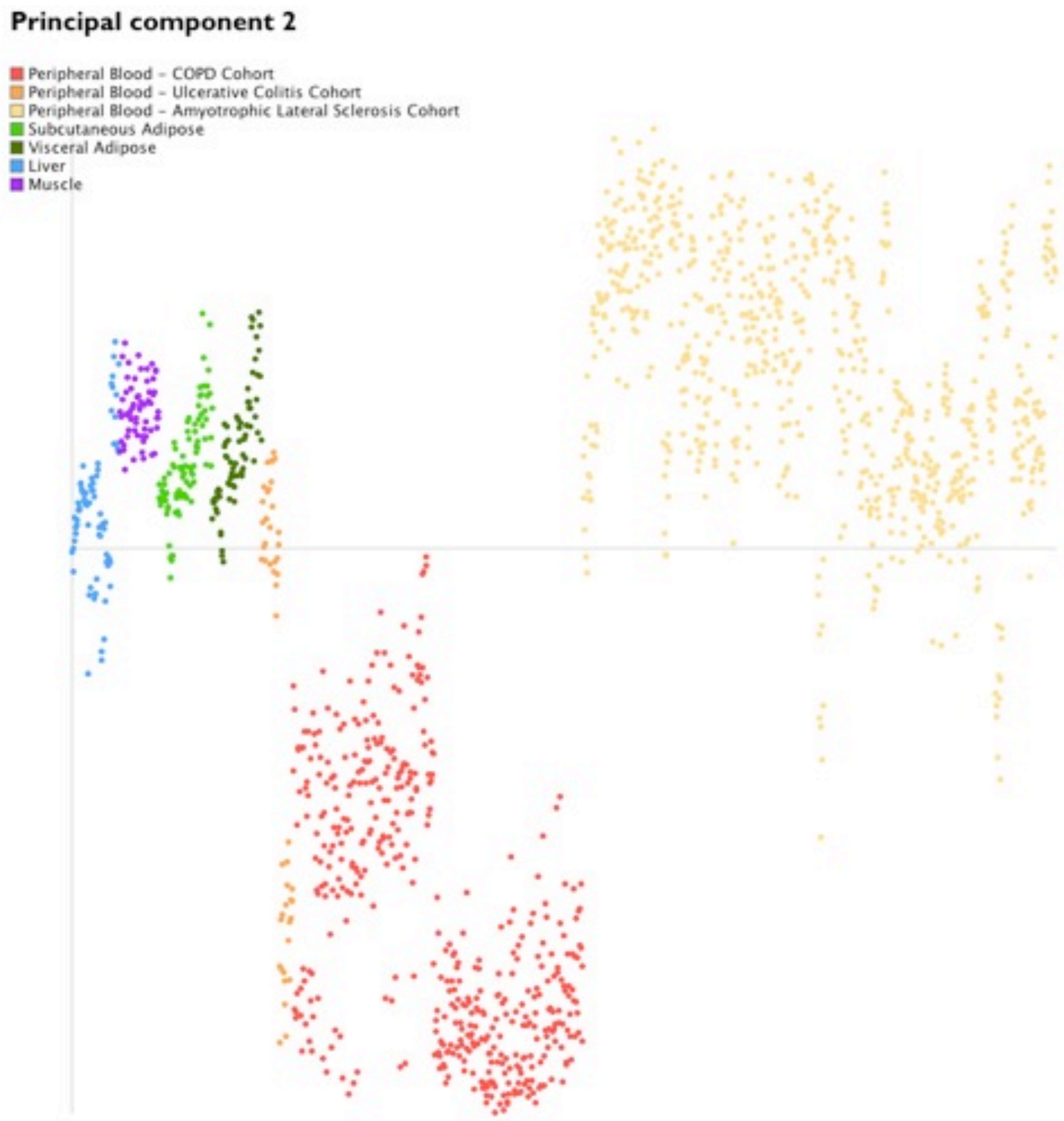
Many genes are affected

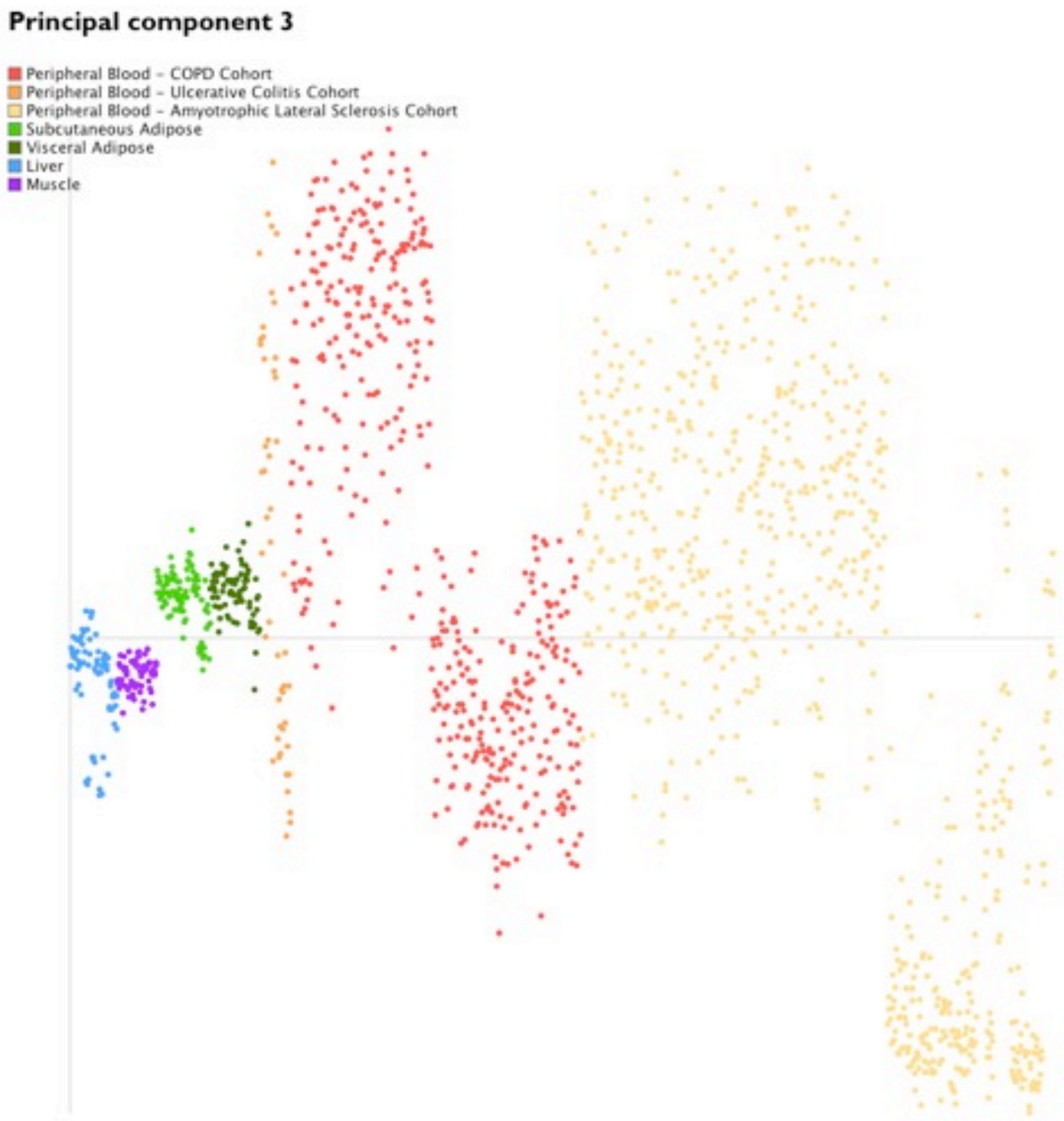
Batch effects: principal component analysis


Data compression

Principal components

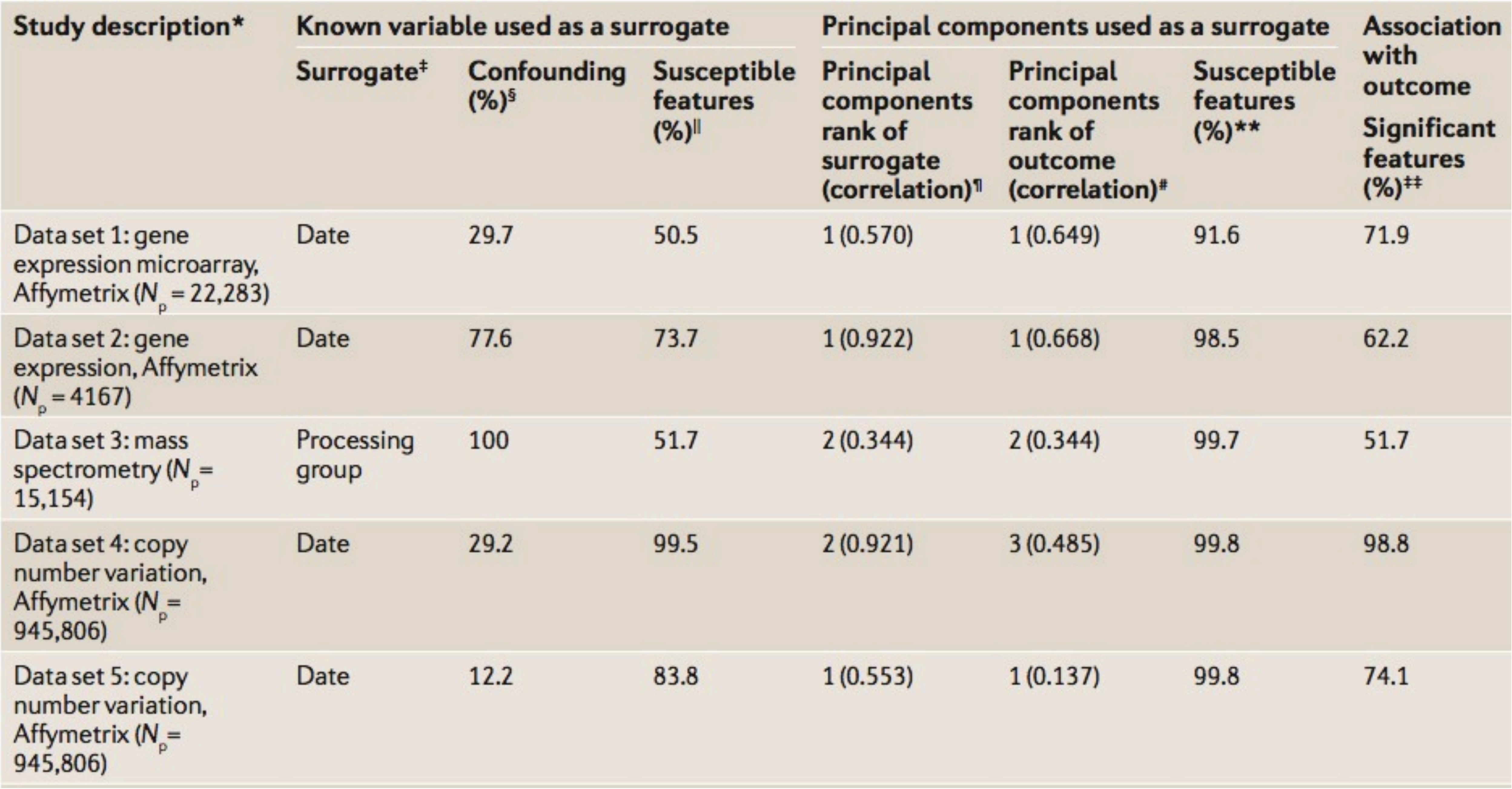
Batch effects in Groningen expression data





Fehrmann *et al*, PLoS Genetics 2011

Batch effects



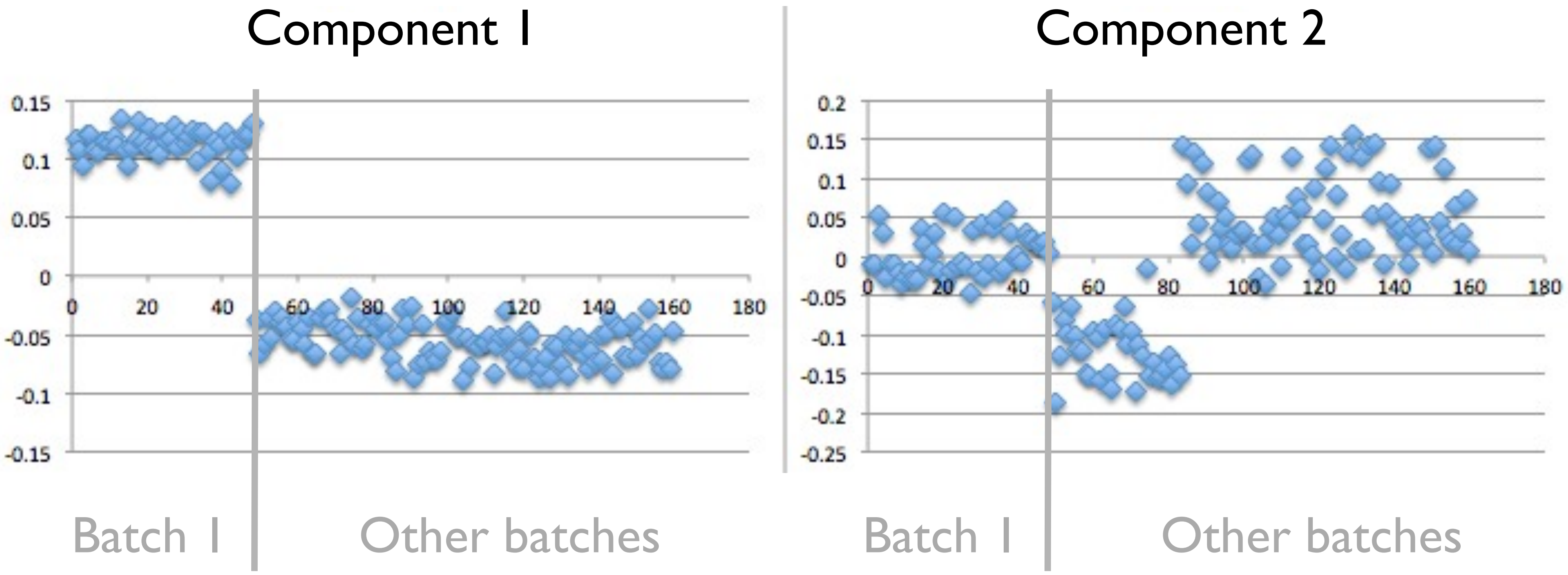
Study description*	Known variable used as a surrogate			Principal components used as a surrogate			Association with outcome
	Surrogate‡	Confounding (%)§	Susceptible features (%)‖	Principal components rank of surrogate (correlation)¶	Principal components rank of outcome (correlation)#	Susceptible features (%)**	Significant features (%)‡‡
Data set 1: gene expression microarray, Affymetrix ($N_p = 22,283$)	Date	29.7	50.5	1 (0.570)	1 (0.649)	91.6	71.9
Data set 2: gene expression, Affymetrix ($N_p = 4167$)	Date	77.6	73.7	1 (0.922)	1 (0.668)	98.5	62.2
Data set 3: mass spectrometry ($N_p =$ 15,154)	Processing group	100	51.7	2 (0.344)	2 (0.344)	99.7	51.7
Data set 4: copy number variation, Affymetrix ($N_p =$ 945,806)	Date	29.2	99.5	2 (0.921)	3 (0.485)	99.8	98.8
Data set 5: copy number variation, Affymetrix ($N_p =$ 945,806)	Date	12.2	83.8	1 (0.553)	1 (0.137)	99.8	74.1

NATURE REVIEWS | GENETICS
VOLUME 11 | OCTOBER 2010 | 733

Batch effect in recent methylation paper

DNA Methylation patterns associate with genetic and gene expression variation in HapMap cell lines:

Conclusion in paper: SNP rs10876043 does strongly influence many methylation levels (affects component 1)



Component 1

Component 2

Batch 1

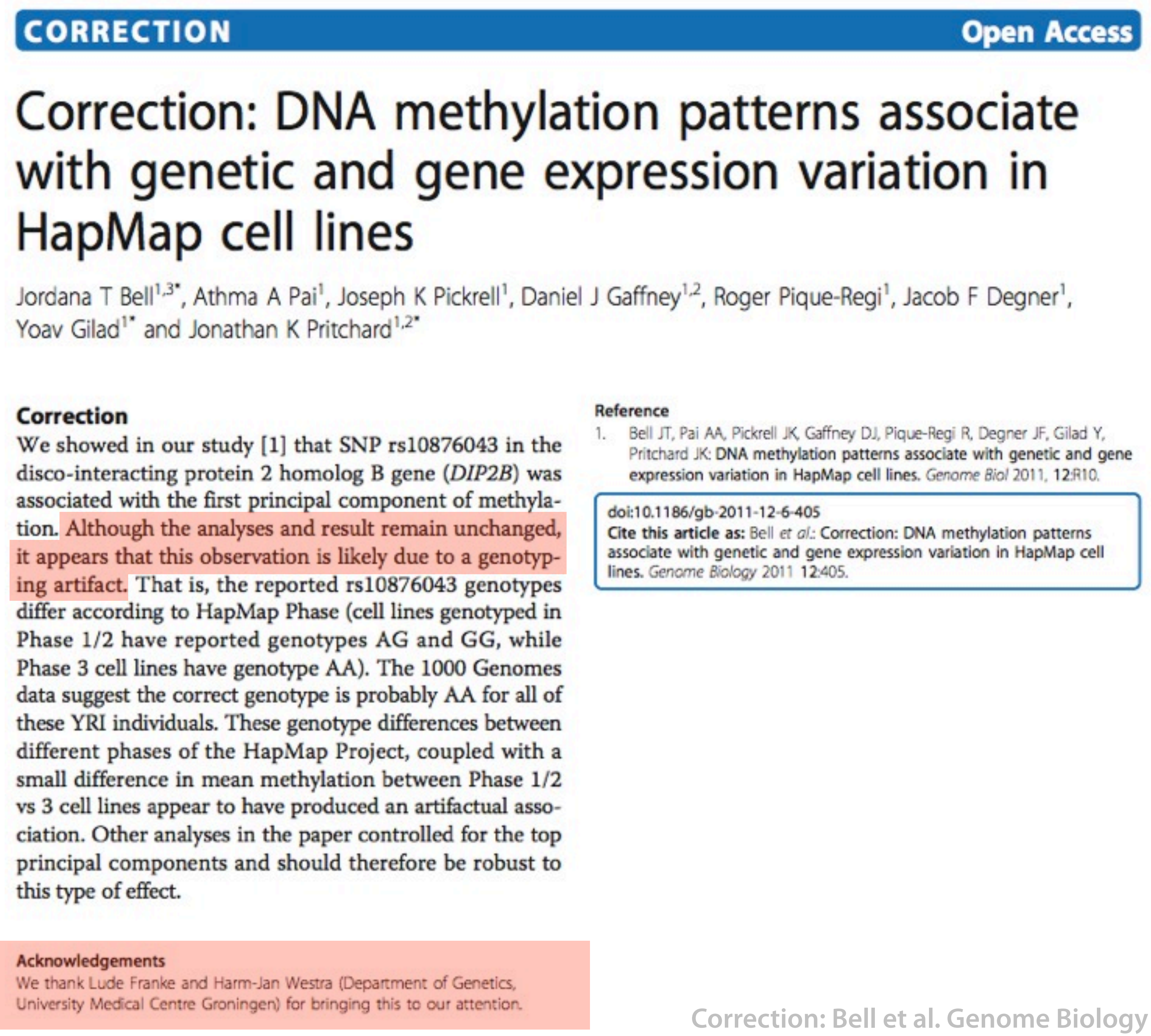
Other batches

Batch 1

Other batches

Bell *et al*, Genome Biology 2011, 12:R10
Pai *et al*, PLoS Genetics 2011

Batch effect in recent methylation paper



CORRECTION Open Access

Correction: DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines

Jordana T Bell[1,3*], Athma A Pai[1], Joseph K Pickrell[1], Daniel J Gaffney[1,2], Roger Pique-Regi[1], Jacob F Degner[1], Yoav Gilad[1*] and Jonathan K Pritchard[1,2*]

Correction

We showed in our study [1] that SNP rs10876043 in the disco-interacting protein 2 homolog B gene (*DIP2B*) was associated with the first principal component of methylation. Although the analyses and result remain unchanged, it appears that this observation is likely due to a genotyping artifact. That is, the reported rs10876043 genotypes differ according to HapMap Phase (cell lines genotyped in Phase 1/2 have reported genotypes AG and GG, while Phase 3 cell lines have genotype AA). The 1000 Genomes data suggest the correct genotype is probably AA for all of these YRI individuals. These genotype differences between different phases of the HapMap Project, coupled with a small difference in mean methylation between Phase 1/2 vs 3 cell lines appear to have produced an artifactual association. Other analyses in the paper controlled for the top principal components and should therefore be robust to this type of effect.

Acknowledgements
We thank Lude Franke and Harm-Jan Westra (Department of Genetics, University Medical Centre Groningen) for bringing this to our attention.

Reference

1. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK: **DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines.** *Genome Biol* 2011, 12:R10.

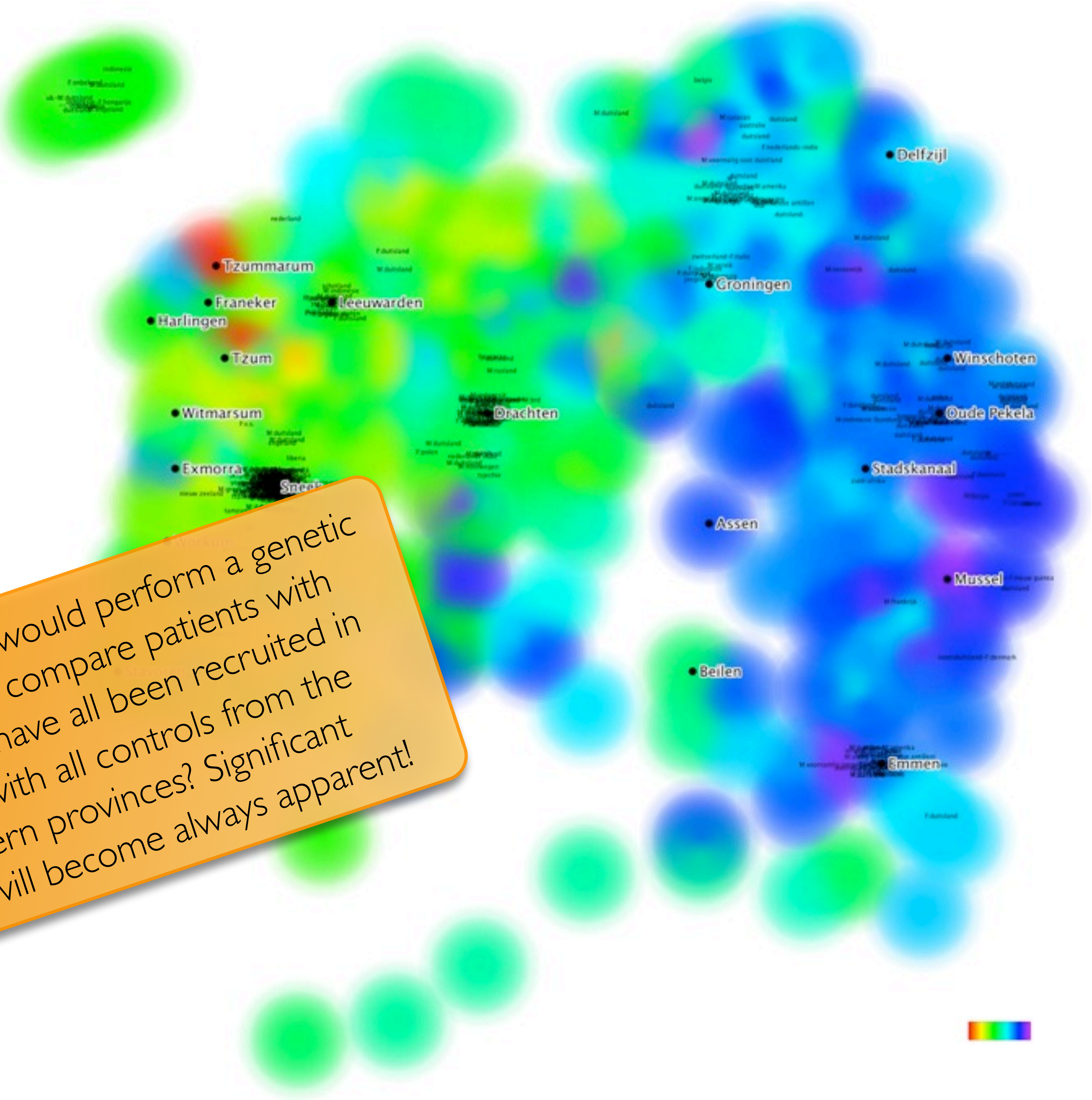
doi:10.1186/gb-2011-12-6-405
Cite this article as: Bell *et al.*: Correction: DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology* 2011 12:405.

Correction: Bell et al. Genome Biology 2011, 12:405

Systematic differences between cases and controls

Systematic differences between cases and controls

Systematic differences between cases and controls

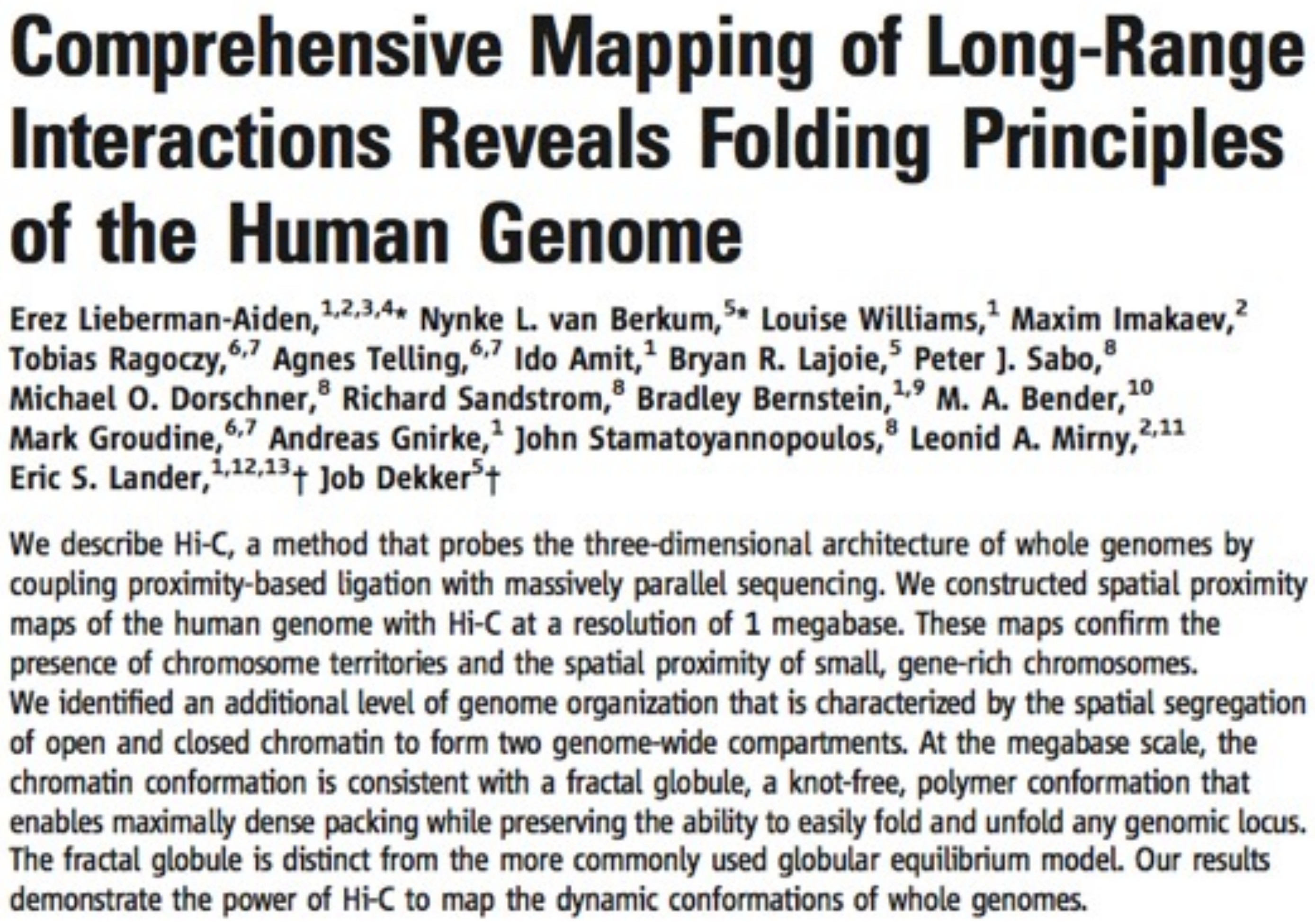


What now if we would perform a genetic study where we compare patients with disease X that have all been recruited in Leeuwarden with all controls from the three Northern provinces? Significant differences will become always apparent!

GC content

GC content

Batch effects: 3D chromosome organization



Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome

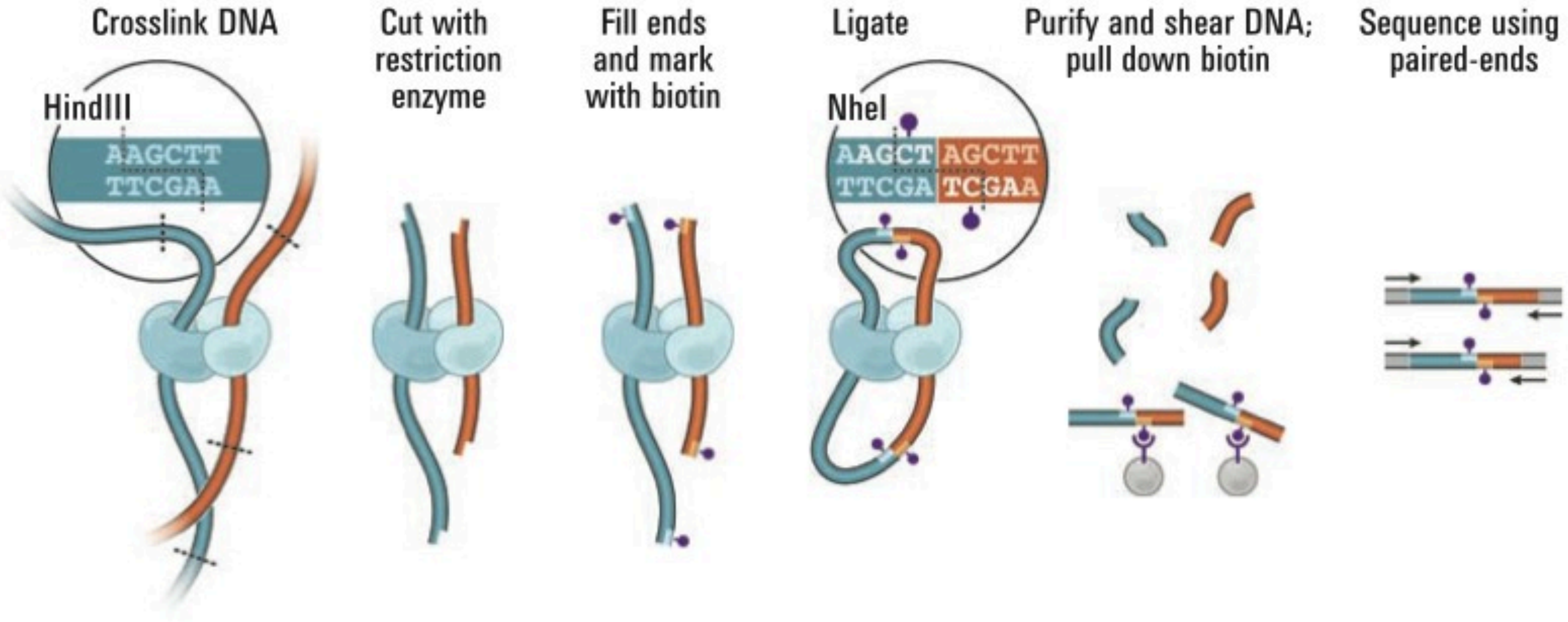
Erez Lieberman-Aiden,[1,2,3,4]* Nynke L. van Berkum,[5]* Louise Williams,[1] Maxim Imakaev,[2] Tobias Ragoczy,[6,7] Agnes Telling,[6,7] Ido Amit,[1] Bryan R. Lajoie,[5] Peter J. Sabo,[8] Michael O. Dorschner,[8] Richard Sandstrom,[8] Bradley Bernstein,[1,9] M. A. Bender,[10] Mark Groudine,[6,7] Andreas Gnirke,[1] John Stamatoyannopoulos,[8] Leonid A. Mirny,[2,11] Eric S. Lander,[1,12,13]† Job Dekker[5]†

We describe Hi-C, a method that probes the three-dimensional architecture of whole genomes by coupling proximity-based ligation with massively parallel sequencing. We constructed spatial proximity maps of the human genome with Hi-C at a resolution of 1 megabase. These maps confirm the presence of chromosome territories and the spatial proximity of small, gene-rich chromosomes. We identified an additional level of genome organization that is characterized by the spatial segregation of open and closed chromatin to form two genome-wide compartments. At the megabase scale, the chromatin conformation is consistent with a fractal globule, a knot-free, polymer conformation that enables maximally dense packing while preserving the ability to easily fold and unfold any genomic locus. The fractal globule is distinct from the more commonly used globular equilibrium model. Our results demonstrate the power of Hi-C to map the dynamic conformations of whole genomes.

Status May 2013: Cited over 700 times

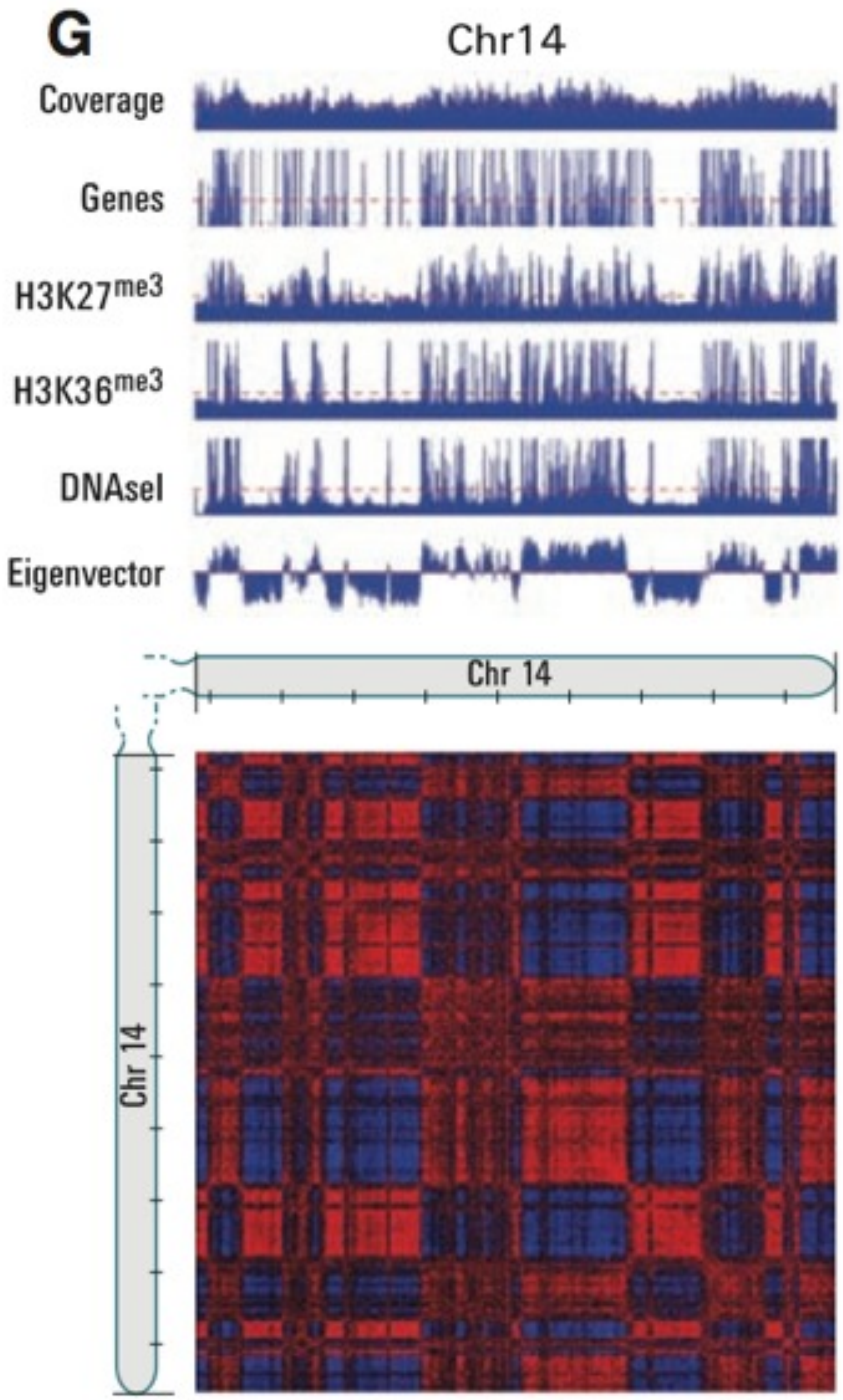
Lieberman-Aiden *et al*, Science 2009

Batch effects: 3D chromosome organization


Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends

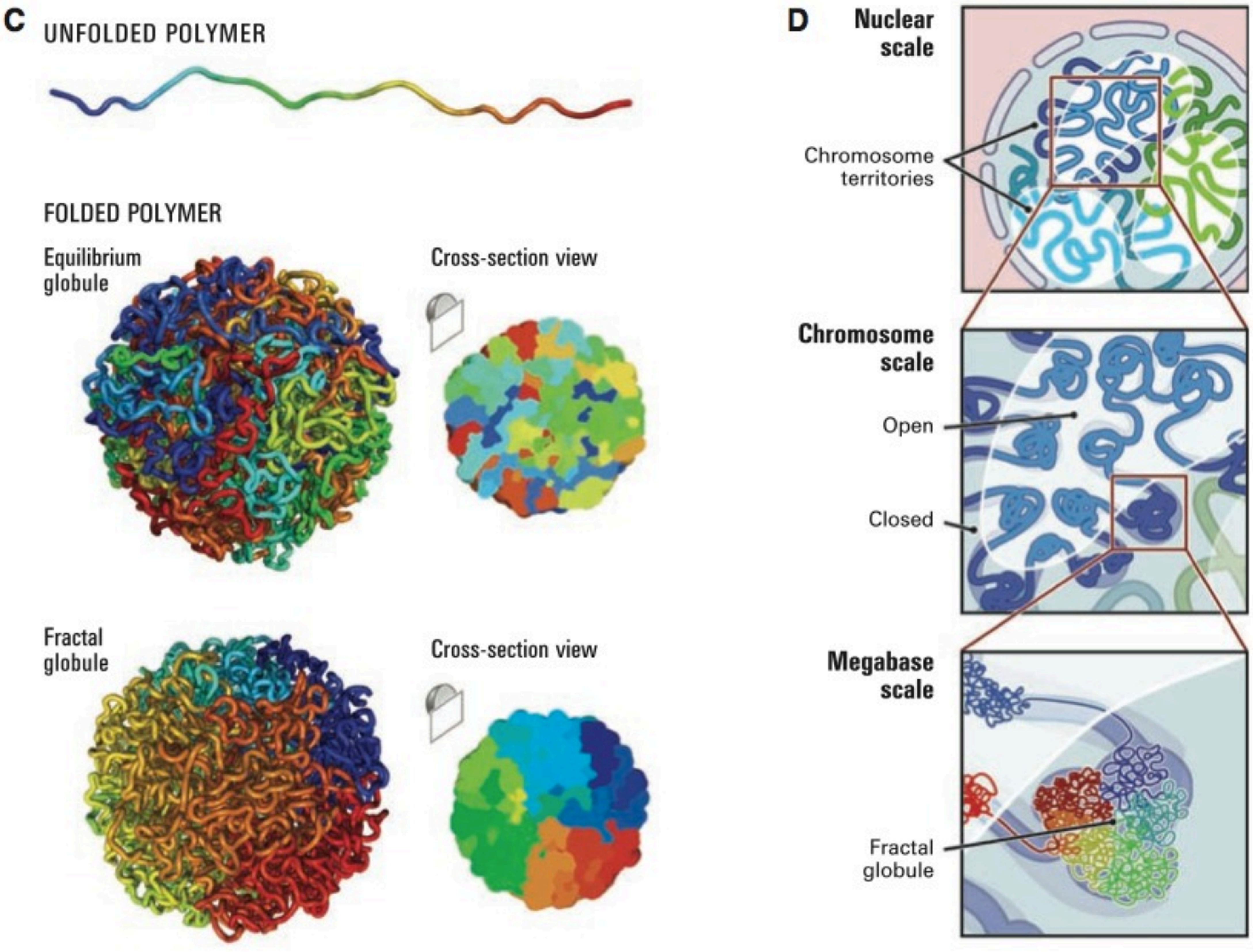
Lieberman-Aiden *et al*, Science 2009

Batch effects: 3D chromosome organization



Lieberman-Aiden *et al*, Science 2009

Batch effects: 3D chromosome organization

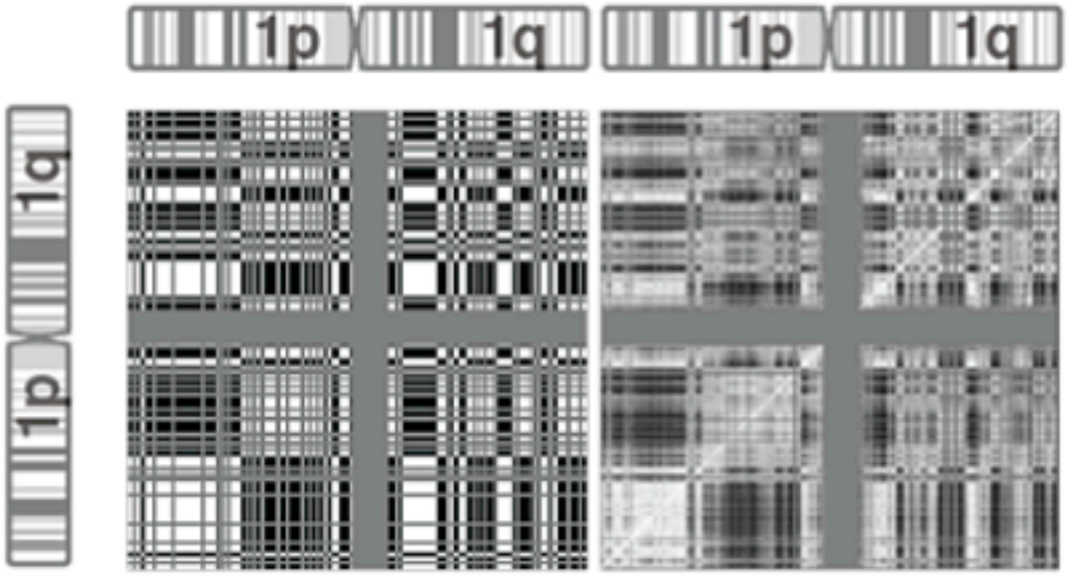


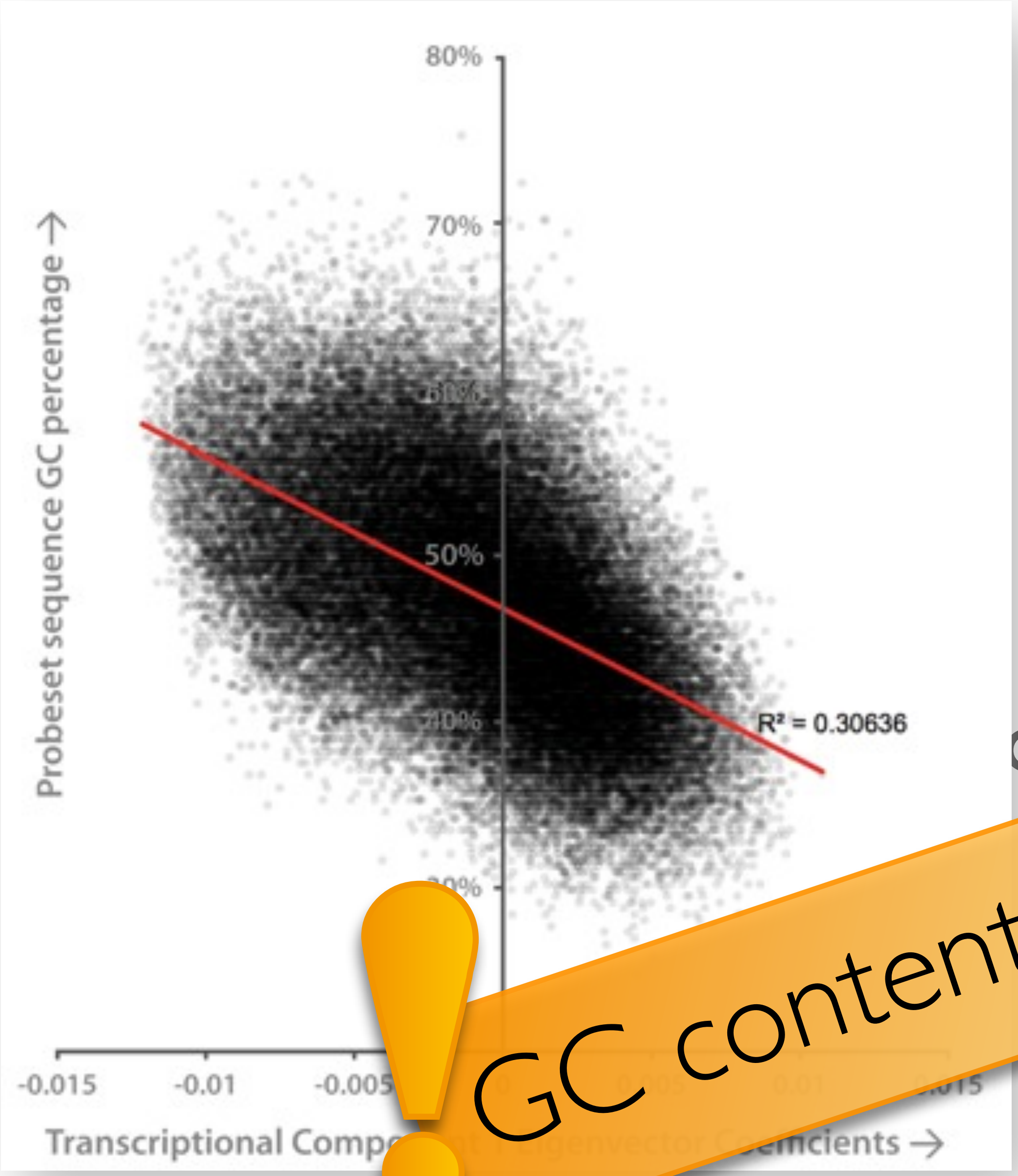
Lieberman-Aiden *et al*, Science 2009

Batch effects: 3D chromosome organization

Chromosome 1

Gene expression data   Hi-C data Science paper

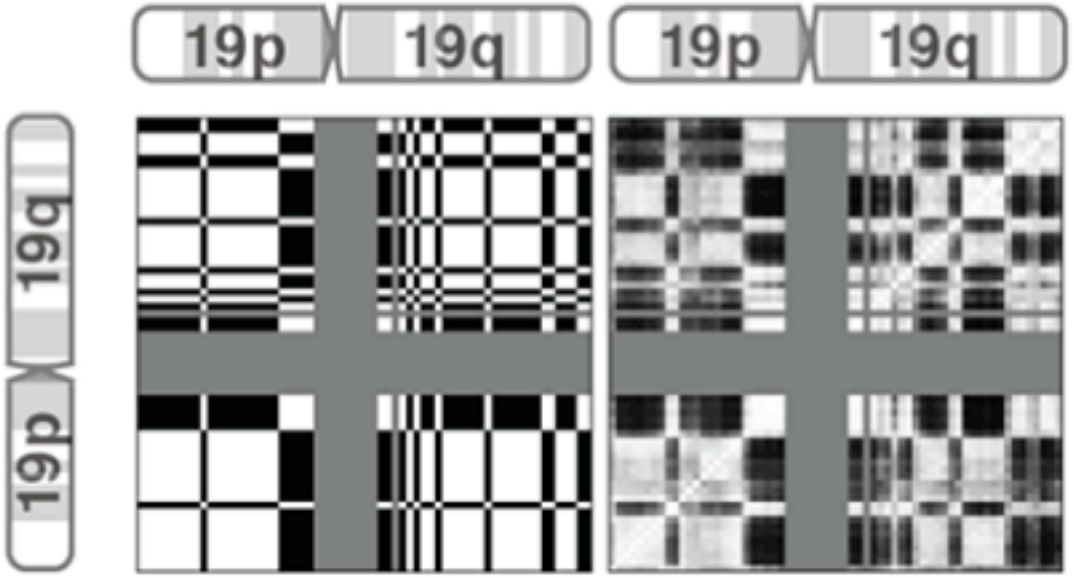




GC content!

ome 19

Gene expression data   Hi-C data Science paper



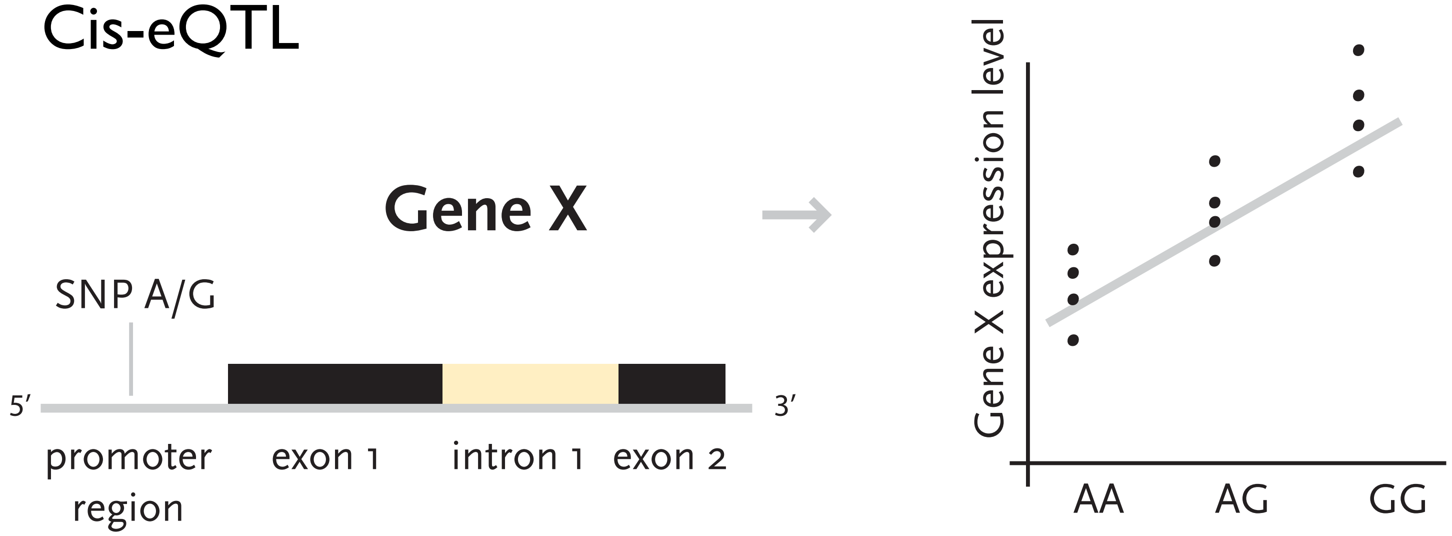
Lieberman-Aiden *et al*, Science 2009

Sample mix-ups

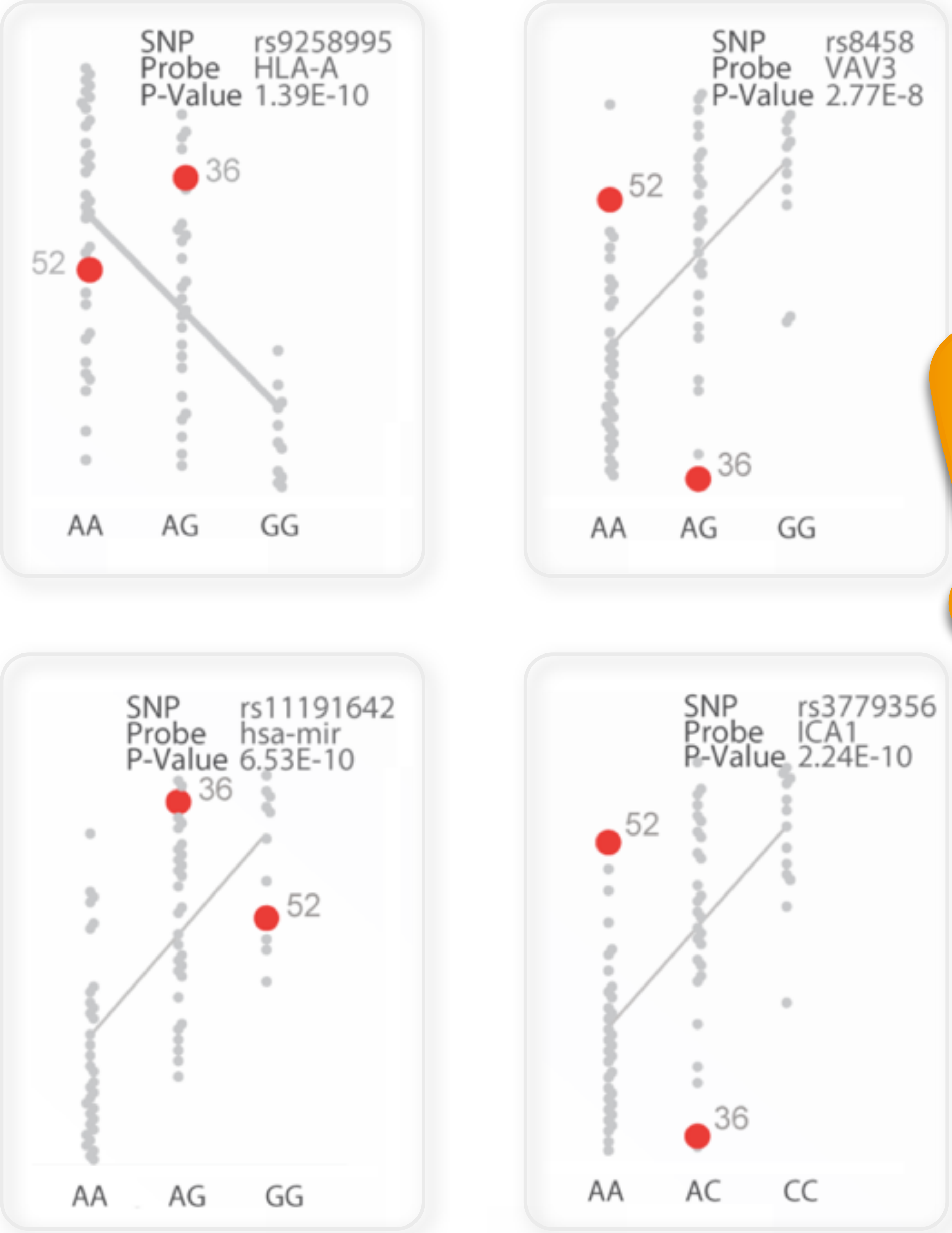
Sample mix-ups

Genetic variants can affect expression levels

Cis-eQTL



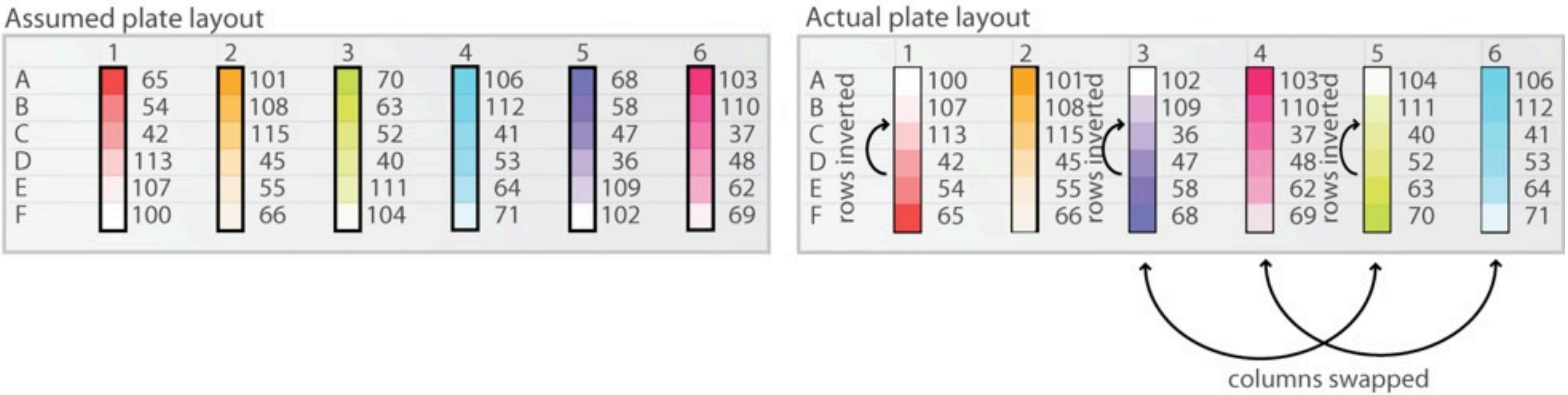
Sample mix-ups: how to identify them



What is going on with sample 36 and 52? Sample mix-up?

Westra et al, Bioinformatics, 2011

Sample mix-ups: What happened to our data



Assumed plate layout

	1	2	3	4	5	6
A	65	101	70	106	68	103
B	54	108	63	112	58	110
C	42	115	52	41	47	37
D	113	45	40	53	36	48
E	107	55	111	64	109	62
F	100	66	104	71	102	69

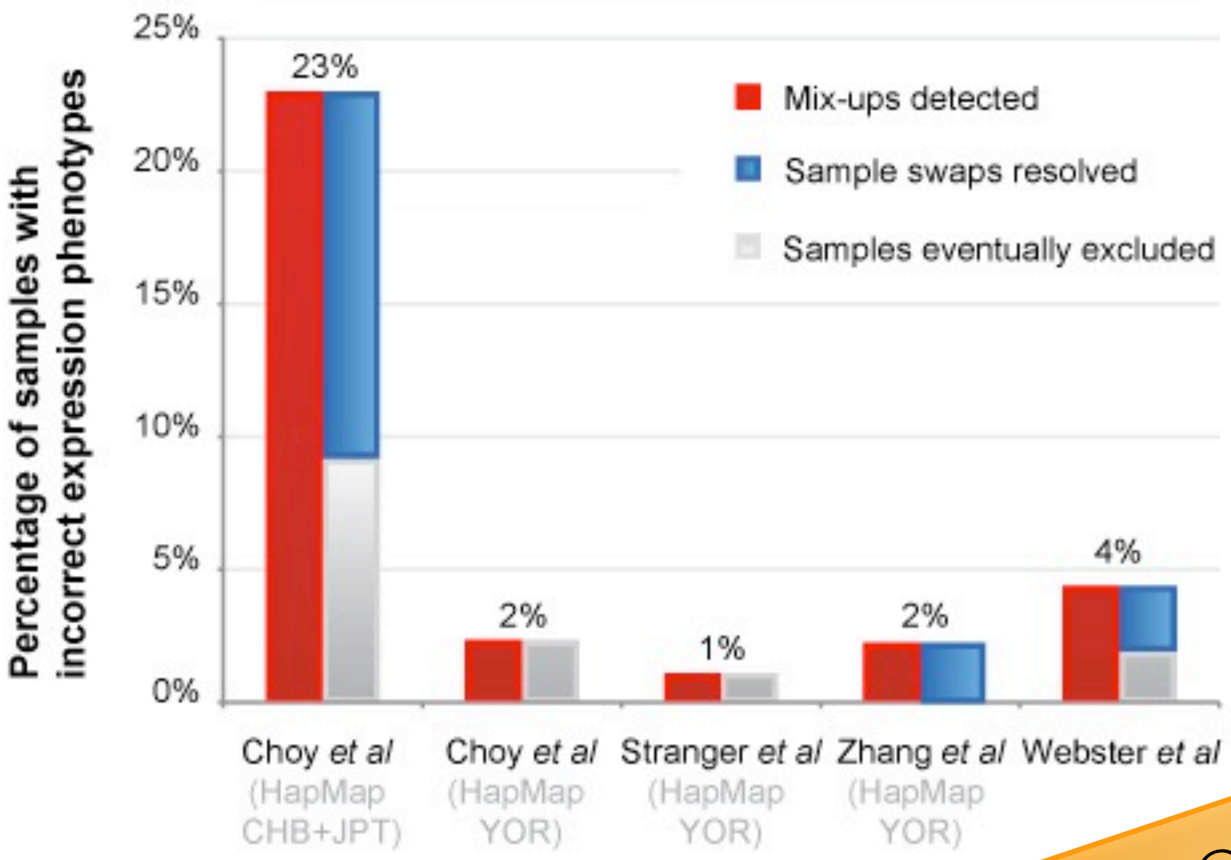
Actual plate layout

	1	2	3	4	5	6
A	100	101	102	103	104	106
B	107	108	109	110	111	112
C	113	115	36	37	40	41
D	42	45	47	48	52	53
E	54	55	58	62	63	64
F	65	66	68	69	70	71

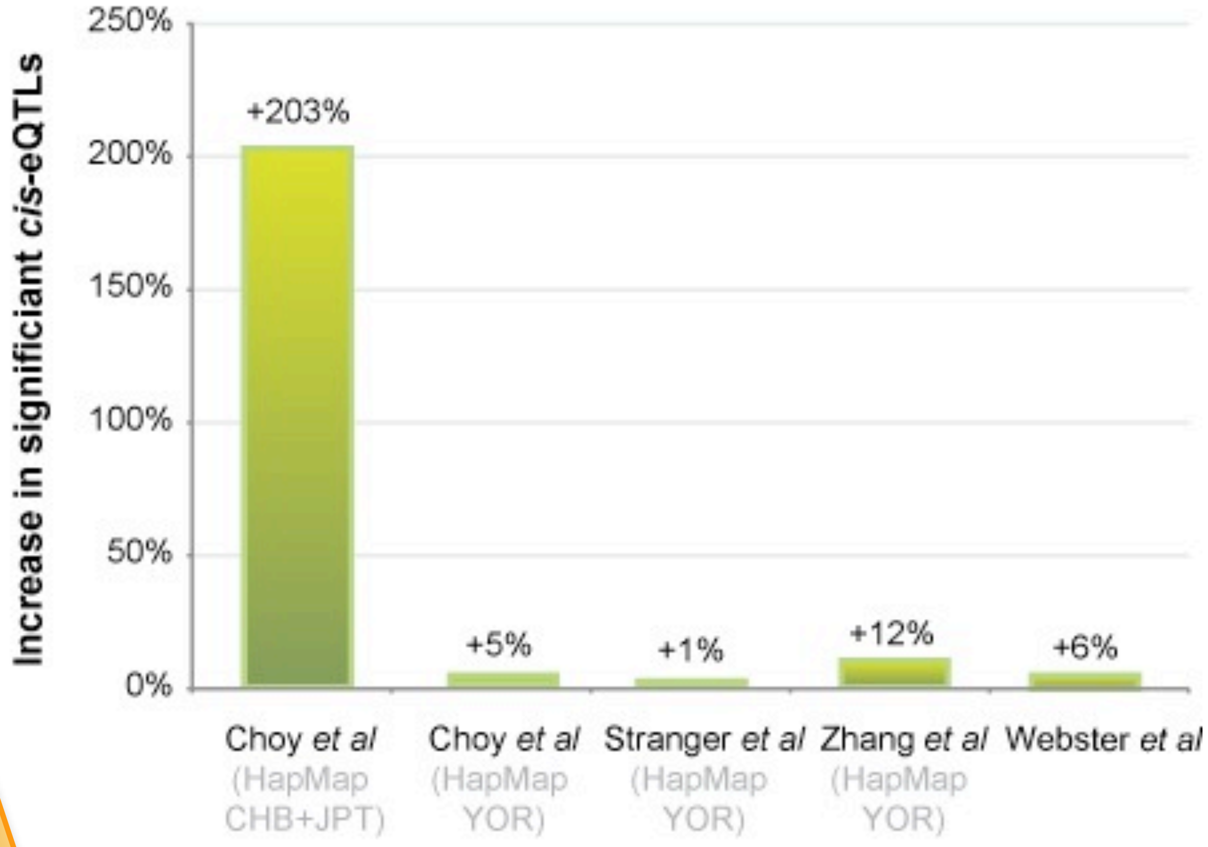
Westra et al, Bioinformatics, 2011

Sample mix-ups: do they happen?

eQTL datasets with mix-ups



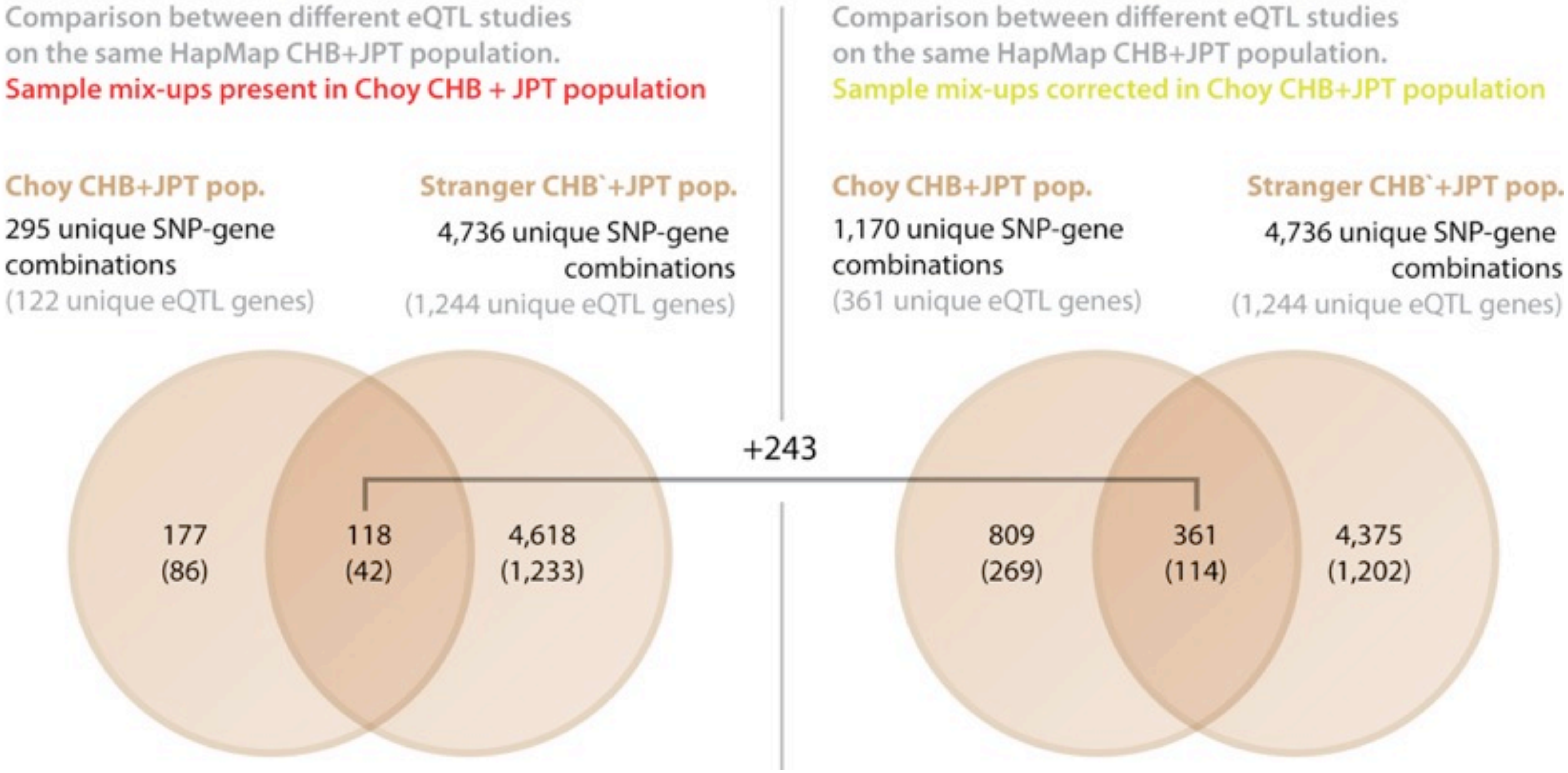
Effect of correcting for these mix-ups



On average 3% of eQTL samples are mixed-up

Westra et al, Bioinformatics, 2011

Comparing same samples using different platforms



Comparison between different eQTL studies on the same HapMap CHB+JPT population.
Sample mix-ups present in Choy CHB + JPT population

Choy CHB+JPT pop.
295 unique SNP-gene combinations
(122 unique eQTL genes)

Stranger CHB`+JPT pop.
4,736 unique SNP-gene combinations
(1,244 unique eQTL genes)

Comparison between different eQTL studies on the same HapMap CHB+JPT population.
Sample mix-ups corrected in Choy CHB+JPT population

Choy CHB+JPT pop.
1,170 unique SNP-gene combinations
(361 unique eQTL genes)

Stranger CHB`+JPT pop.
4,736 unique SNP-gene combinations
(1,244 unique eQTL genes)

Stranger *et al*, Science 2007
Choy *et al*, PLoS Genetics 2009
Westra et al, Bioinformatics, 2011

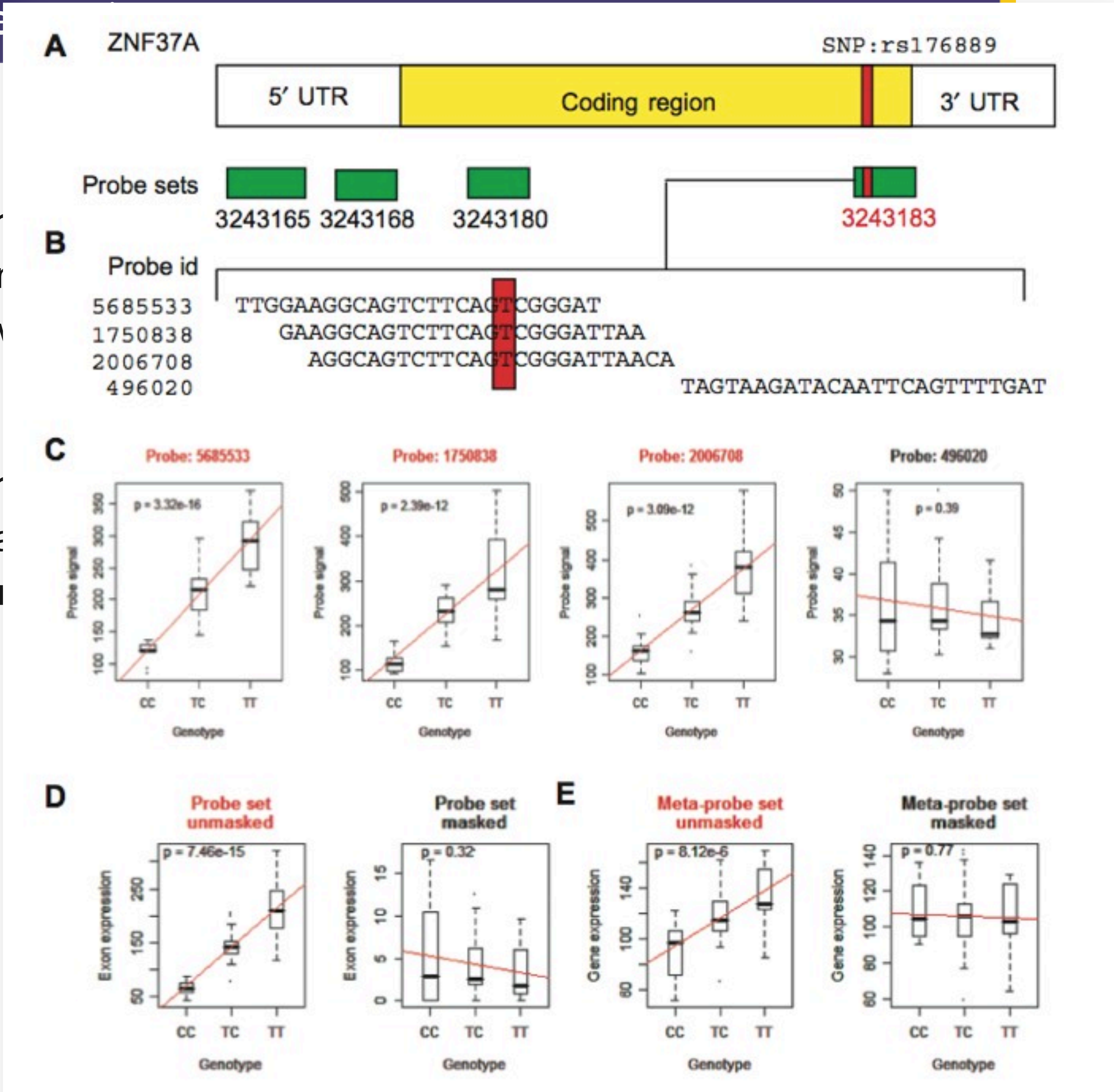
What happens when studies go wrong

Two personal experiences

Personal e

- **Goal: Ider**
- Initially sor
- However, \
  came out:

- **Goal: Ider**
- Many varia
- However, I



Benovoy et al, Nucleic Acids Research, 2008, Vol. 36, No. 13

Conclusions

- Correcting for multiple testing is very important
- Confounders often exist
- It is often unknown what these confounders are
- Principal component analysis can reveal these confounders
- GC content has a major effect, both in genetic, expression, methylation and ChIP-seq studies. Please check whether it might confound your results
- Keep in mind, it is usually possible to correct for these confounders
- **Be careful: Results that seem too good to be true, should worry you!**