AMSI-SSAI Lecture, ABS, August 26, 2013

Removing Unwanted Variation from High-throughput Omic Data

With Johann Gagnon-Bartsch & Laurent Jacob
UC Berkeley, WEHI & CNRS, Lyon

The problem

Genomic and other omic data can be affected by unwanted variation.

For example, batch effects due to time, space, equipment, operators, reagents, sample source, sample quality, environmental conditions,…the list goes on…

Also we often wish to combine data, both within and across platforms. Differences between studies and platforms need to be dealt with.

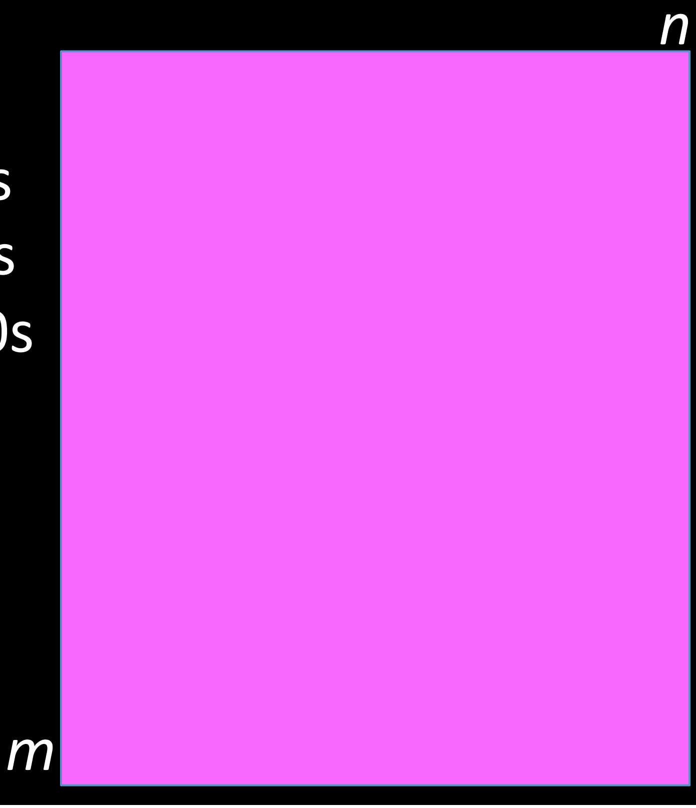
A few examples

Data structure

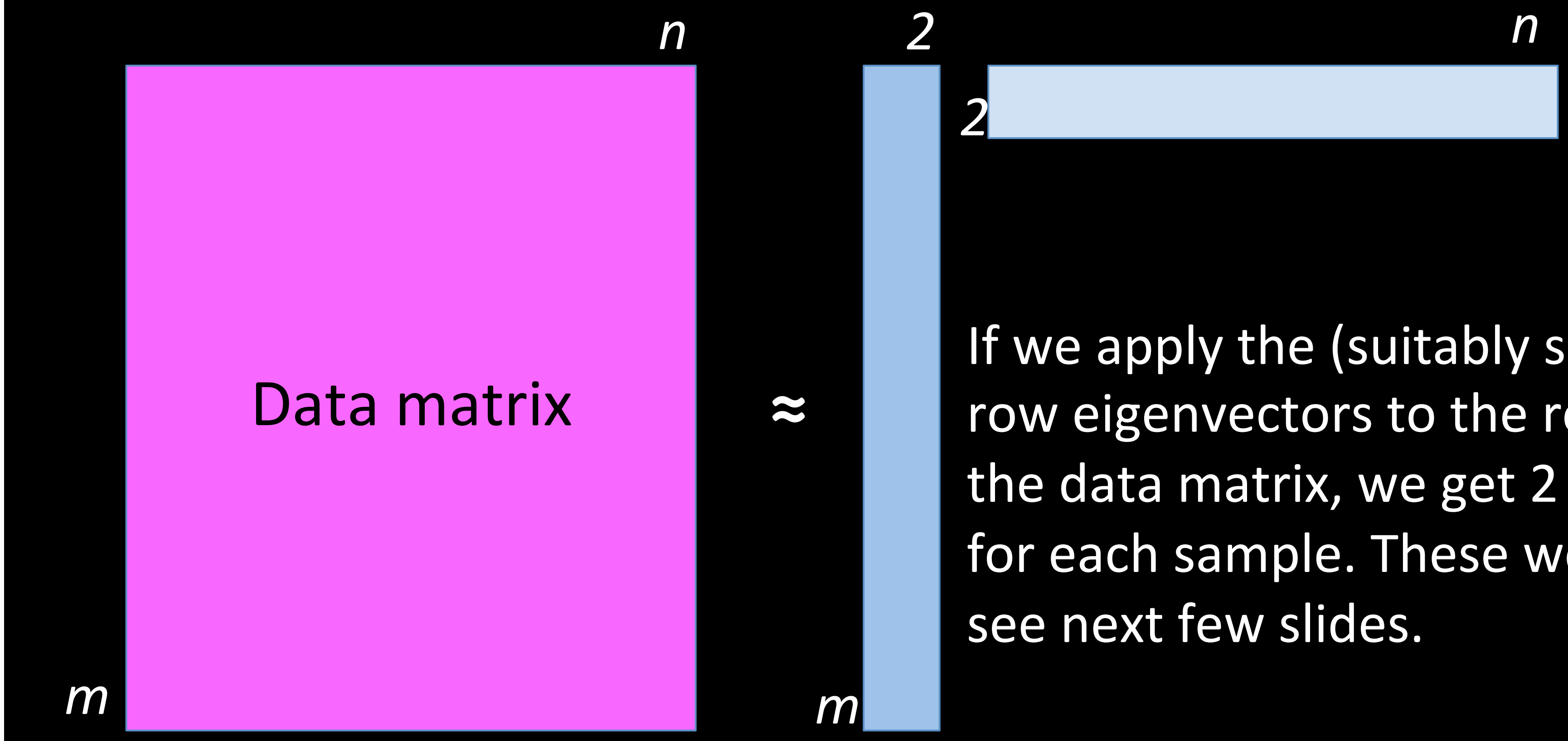
In each of following examples, our data has the form

n columns = genes (~20,000), or SNPs
= DNA variants (up to 2 million) , or …

m rows = samples
typically 10s, 100s
and at times 1,000s



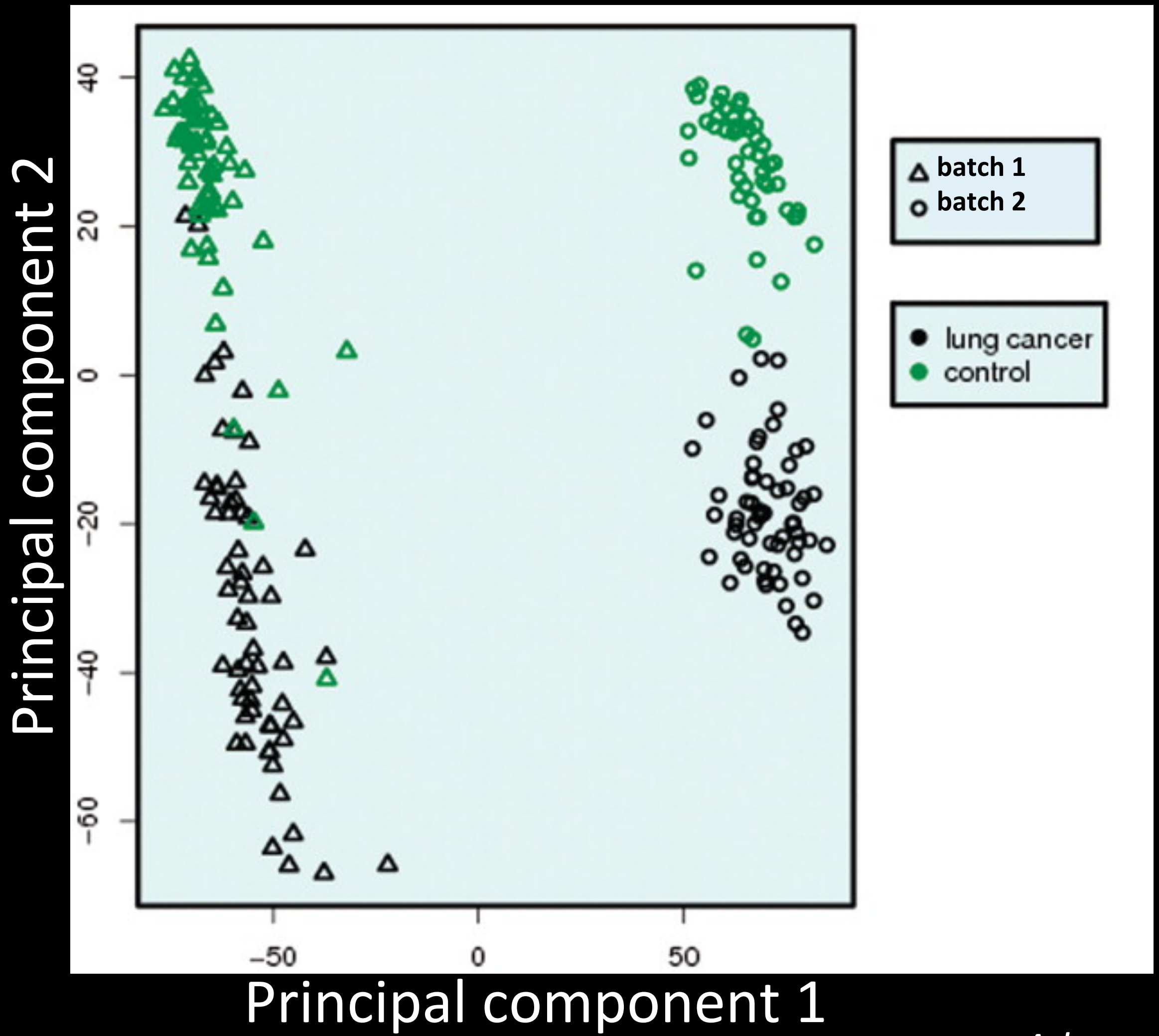
Snapshot view
(SVD, PCA, MDS…)



If we apply the (suitably scaled) row eigenvectors to the rows of the data matrix, we get 2 values for each sample. These we plot, see next few slides.

Artifact (batch) can overwhelm biology

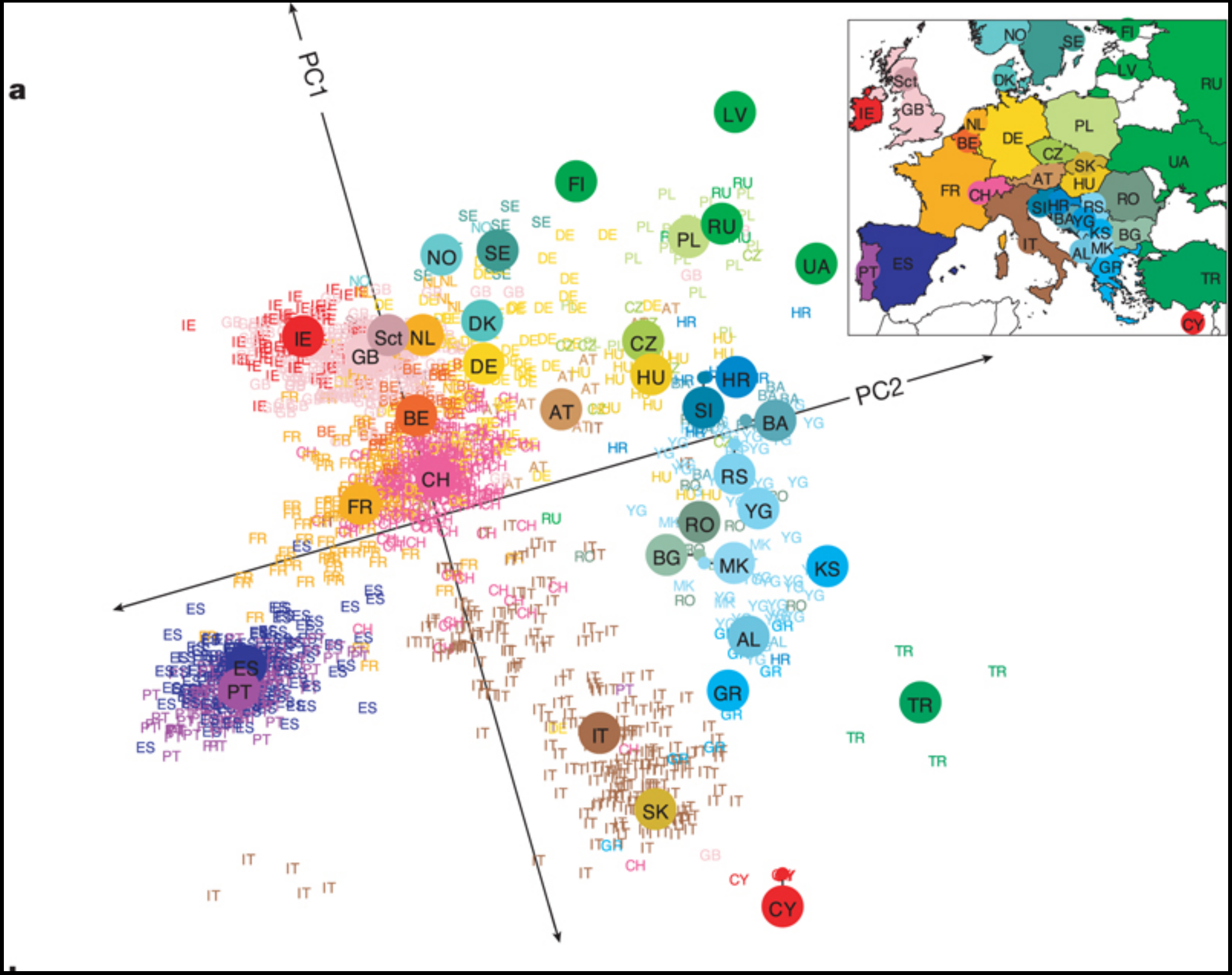
Gene expression microarrays



Adapted from Lazar C et al. Brief Bioinform 2013

SNP genotypes: population structure within Europe.

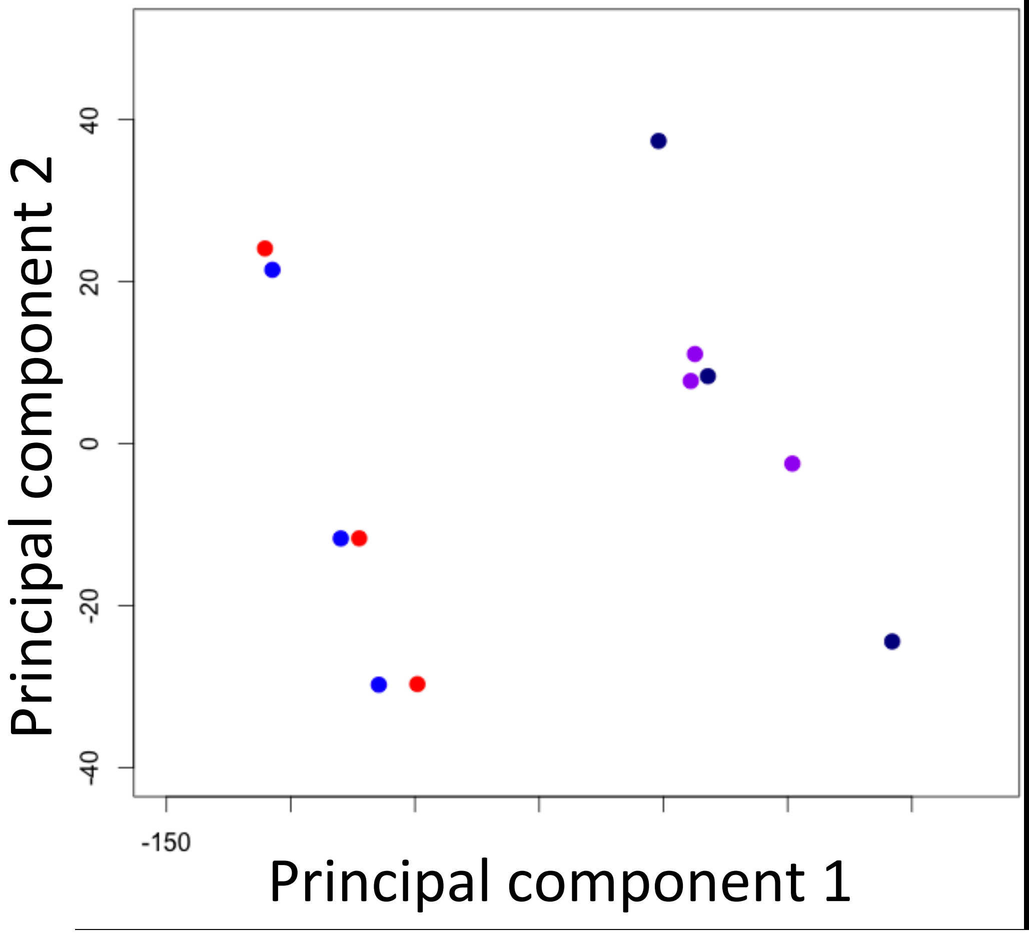
There are situations in which we would like to remove such structure!



From: J Novembre *et al.* *Nature* 456 (2008)

A microarray experiment with central retina tissue from the *rd1* mouse: *4 times x 3*

rd1 is a mouse model of *retinitis pigmentosa:* loss of rod photoreceptors, followed by that of cone photoreceptors



Light blue: 2 months

Dark blue: 4 months

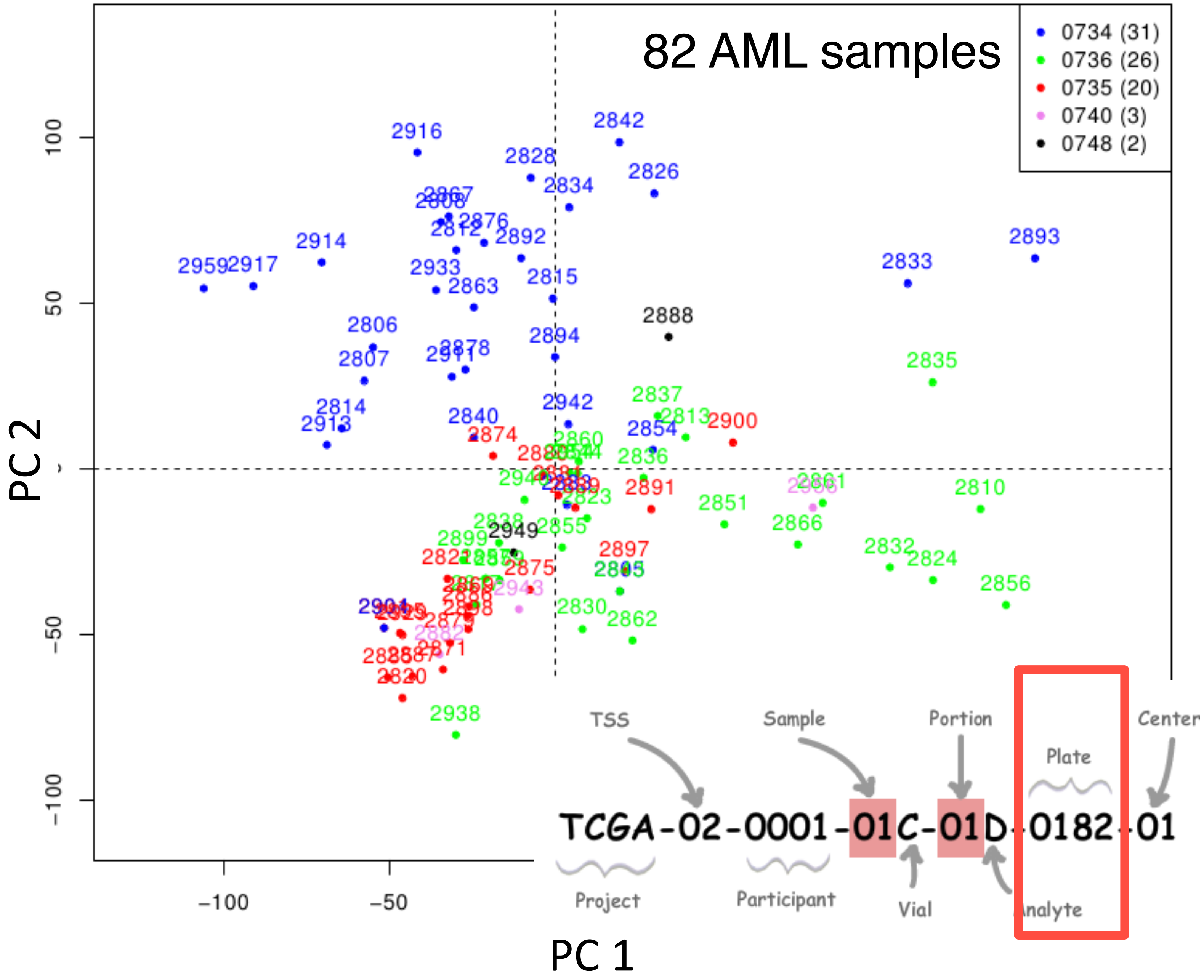
Purple: 6 months

Red: 8 months

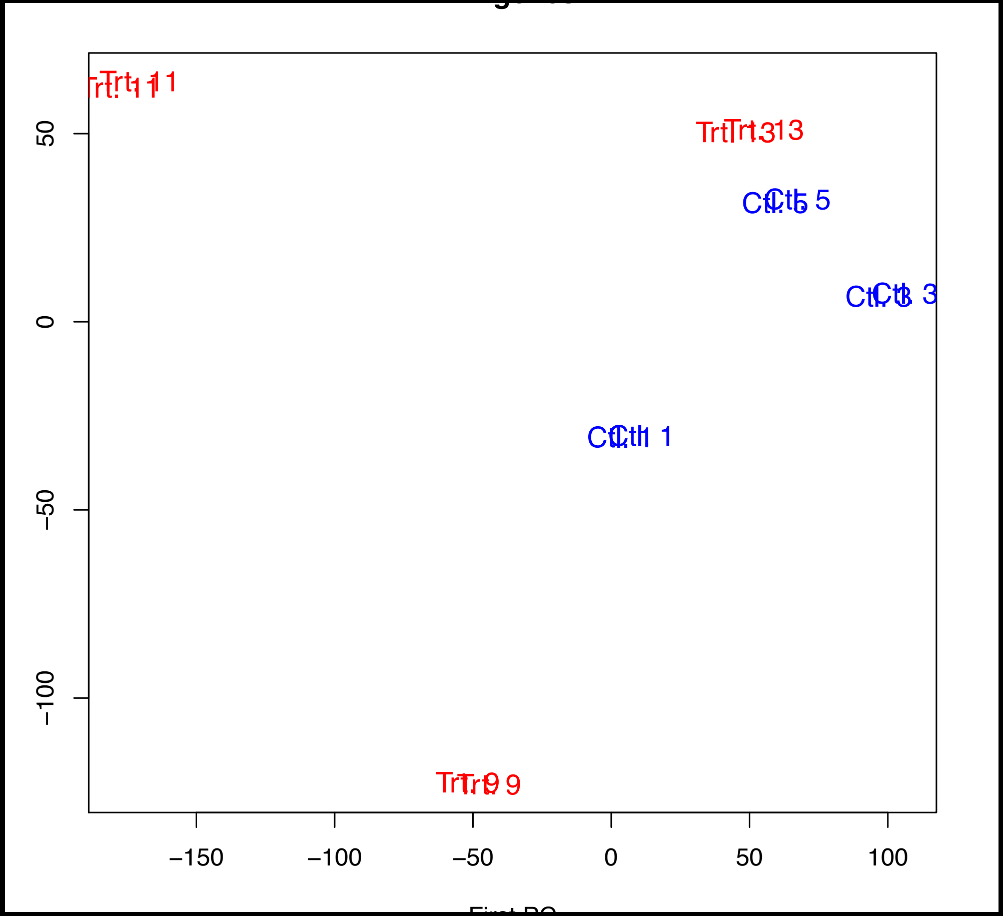
Very severe batch effects

Ideally we would have seen 4 tight groups of 3 ●, ●, ● and ● resp.

RNA-seq data: batch corresponds to plate barcode



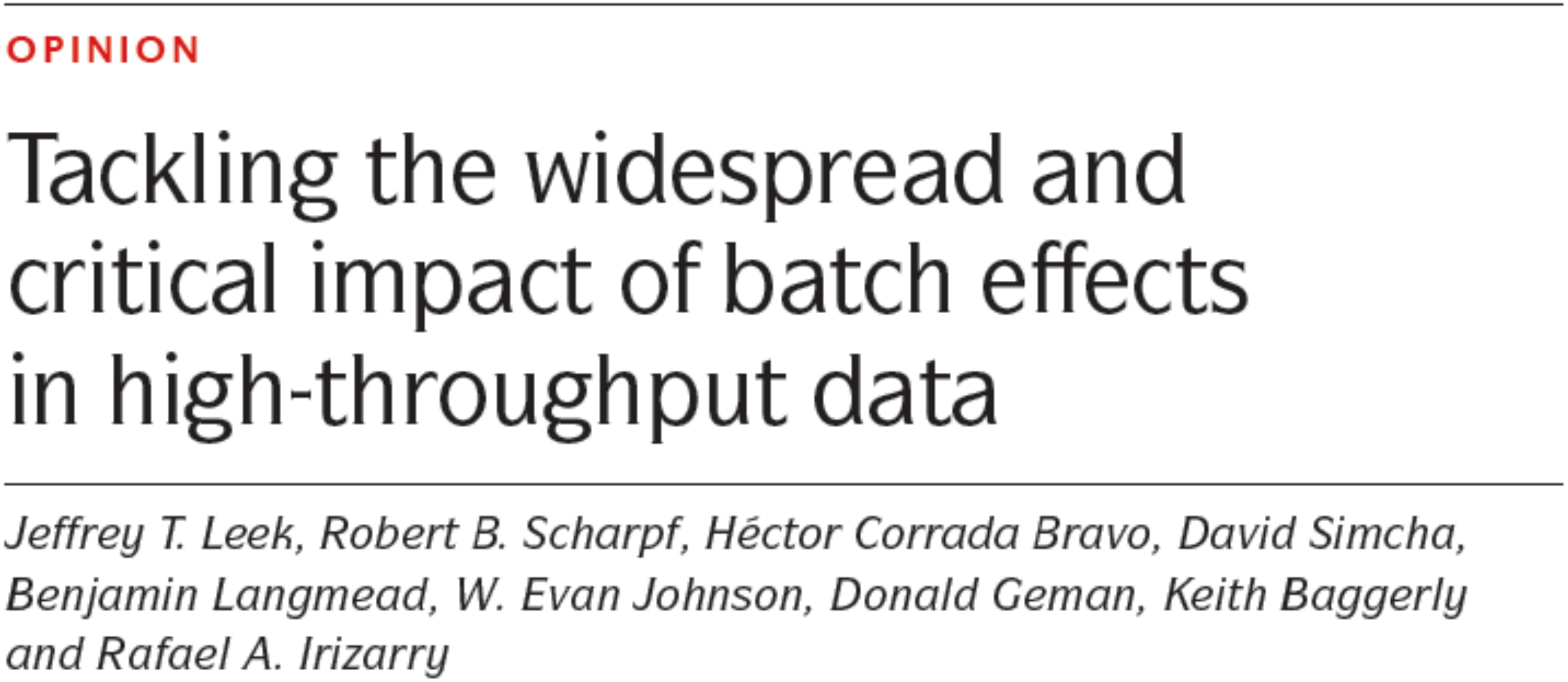
PC2 vs PC1 for 12 zebrafish RNA-seq runs: 3 treated vs 3 control (in duplicate)



Second principal component

First principal component

The biology is not evident in the first 2 PCs



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, vol **11**, October 2010, p. 733

They identify fatally flawed studies!

Combining 3 experiments

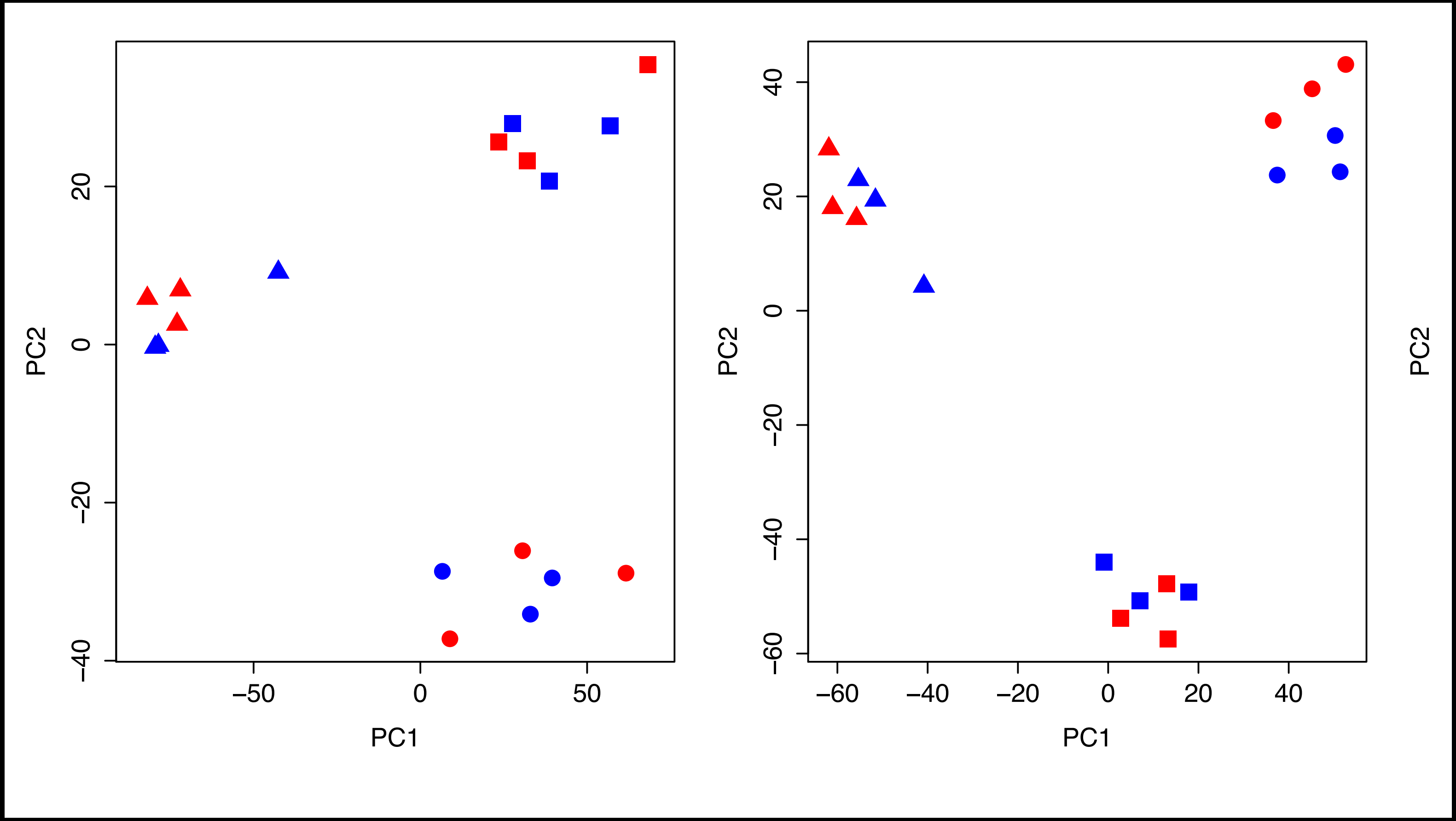
- Three microarray gene expression experiments carried out at different times are all comparisons of the form

  Knock-Out (3X) vs Wild-Type (3X)

- All are in T-cells, and while the three KOs differ (Id2, Tbet, Blimp), the WT mice are the same.
- The idea is to combine the three experiments into one, to benefit from the increased WT replication, and to compare the different KOs.

Raw

Quantile-normalized



Blue: wild-type, Red: knock-out.

Shapes: Different experiments (KOs)

Affymetrix

Agilent

Illumina

Microarrays

GC-MS

LC-MS

Illumina GA-2

HiSeq

MiSeq

Illumina Infinium Human Methylation Beadchips : a special problem

27k

450k

The 27k probes are on the 450k chip.

Wanted: to combine data from these two arrays.

Some scientific goals sought using gene expression microarrays and analogous platforms

- Quantification of expression
- Differential Expression (DE)
- Classification
- Clustering
- Correlating

Some consequences of Unwanted Variation

- Poor quantification of expression
- False discoveries (type 1 errors)
- Missed discoveries (type 2 errors)
- Incorrect predictions
- Artificial clusters
- Wrong correlations

Aim for today

To describe some ways of

- identifying and removing (i.e. adjusting for) unwanted factors, when aiming to achieve these goals, and

- telling whether or not it helped.

I will begin with Differential Expression

The model we and others use

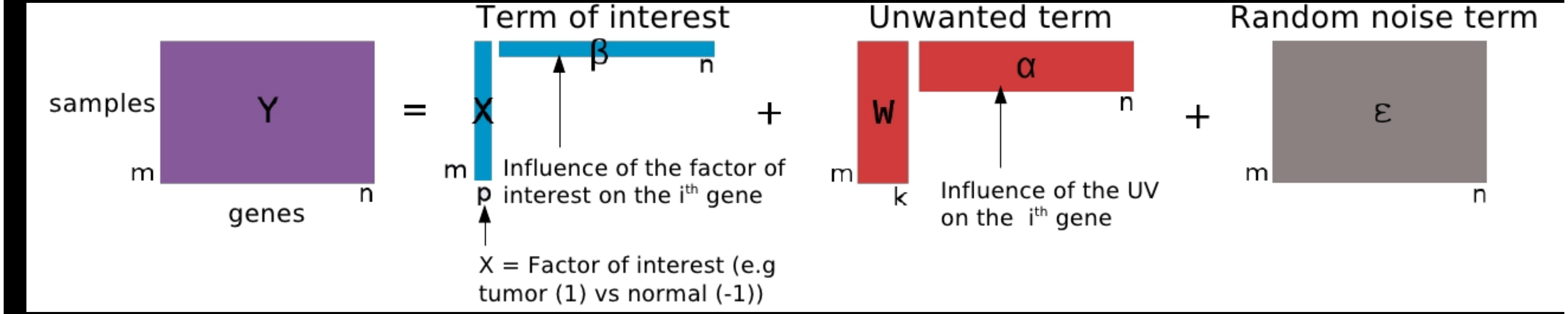
m samples, *n* genes, *k* unwanted factors

$$Y_{m\times n} = X_{m\times p}\beta_{p\times n} + W_{m\times k}\alpha_{k\times n} + \varepsilon_{m\times n}$$

where
Y is a matrix of gene expression measurments, observed,
X carries the factors of interest, observed,
β are gene coefficients, unobserved,
W carries unwanted factors, unobserved,
α are gene coefficients, unobserved,
ε are errors, unobserved.

The model we use in pictures

$$y_{ij} \quad = \quad x_i\beta_j \quad + \quad w_i\alpha_j \quad + \quad \varepsilon_{ij}$$



Influence of the factor of interest on the ith gene

X = Factor of interest (e.g tumor (1) vs normal (-1))

Influence of the UV on the ith gene

Relation to an econometric model

$$Y_{it} = X_{it}'\beta + u_{it} \;,$$

where X_{it} is a $p{\times}1$ vector of observable regressors, β is a $p{\times}1$ vector of unknown coefficients, and u_{it} has a common factor structure

$$u_{it} = \lambda_i'F_t + \varepsilon_{it} \;,$$

where λ_i is a vector of factor loadings and F_t is a vector of common factors, and the ε_{it} are idiosyncratic errors, $i=1,\ldots N$ cross-sectional units, $t=1,\ldots,T$ time periods. This is a model for panel data, Bai (2005), where interest is in estimating β. Often N >> T. Note the difference between the 2 models.

The model, 2

Our goal: for differential expression, to estimate *β.*

Note: *W* is unobserved, Otherwise, this is a standard linear model.

Our strategy: use factor analysis to estimate *W*

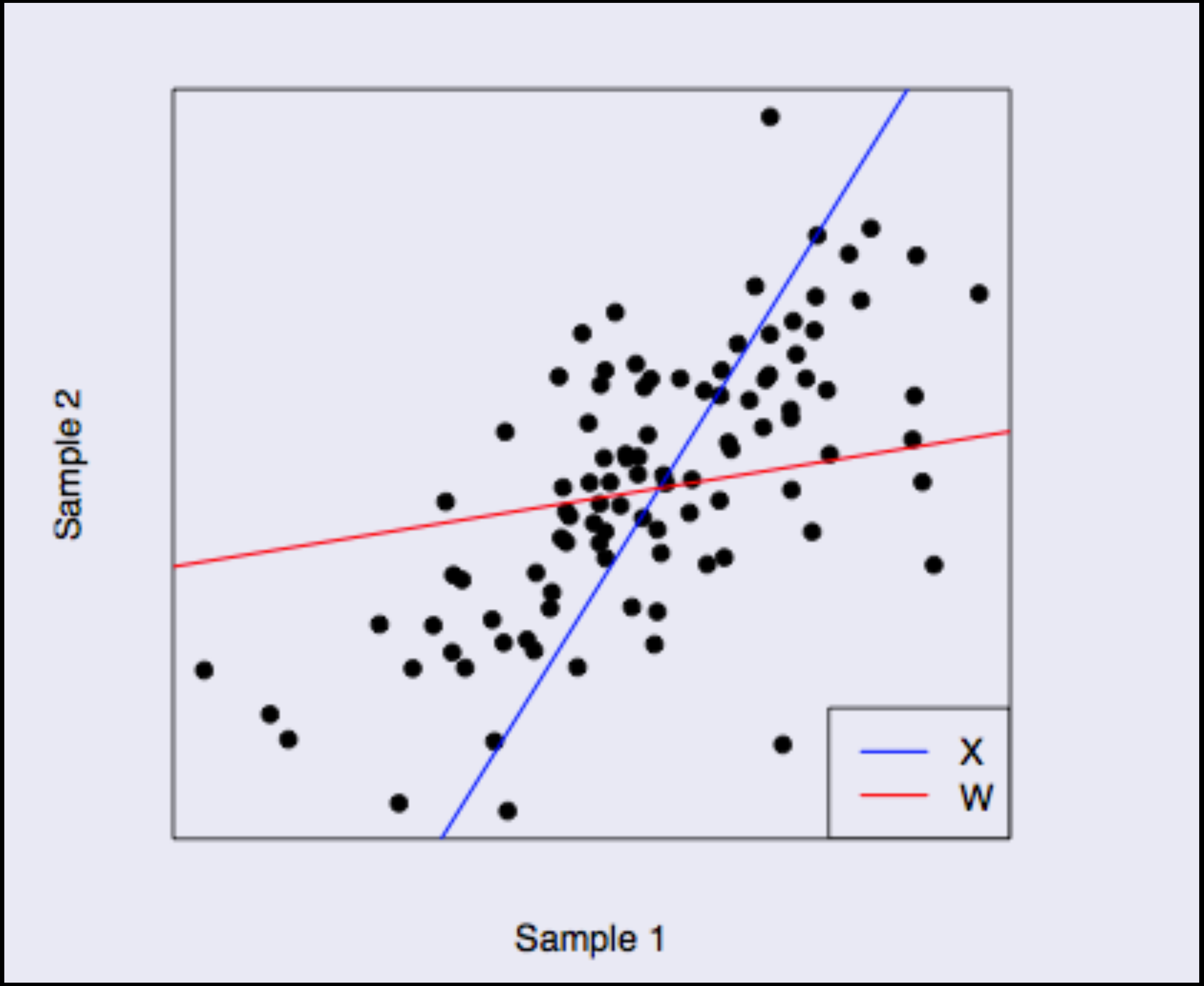
There are identifiability issues

- The correlation between *X* and *W* is unknown
- *β* and *α* are not identifiable

(The examples we use below have *p=1.*)

Identifiability: we don't know the correlation of *W* (*k=1*) with *X*

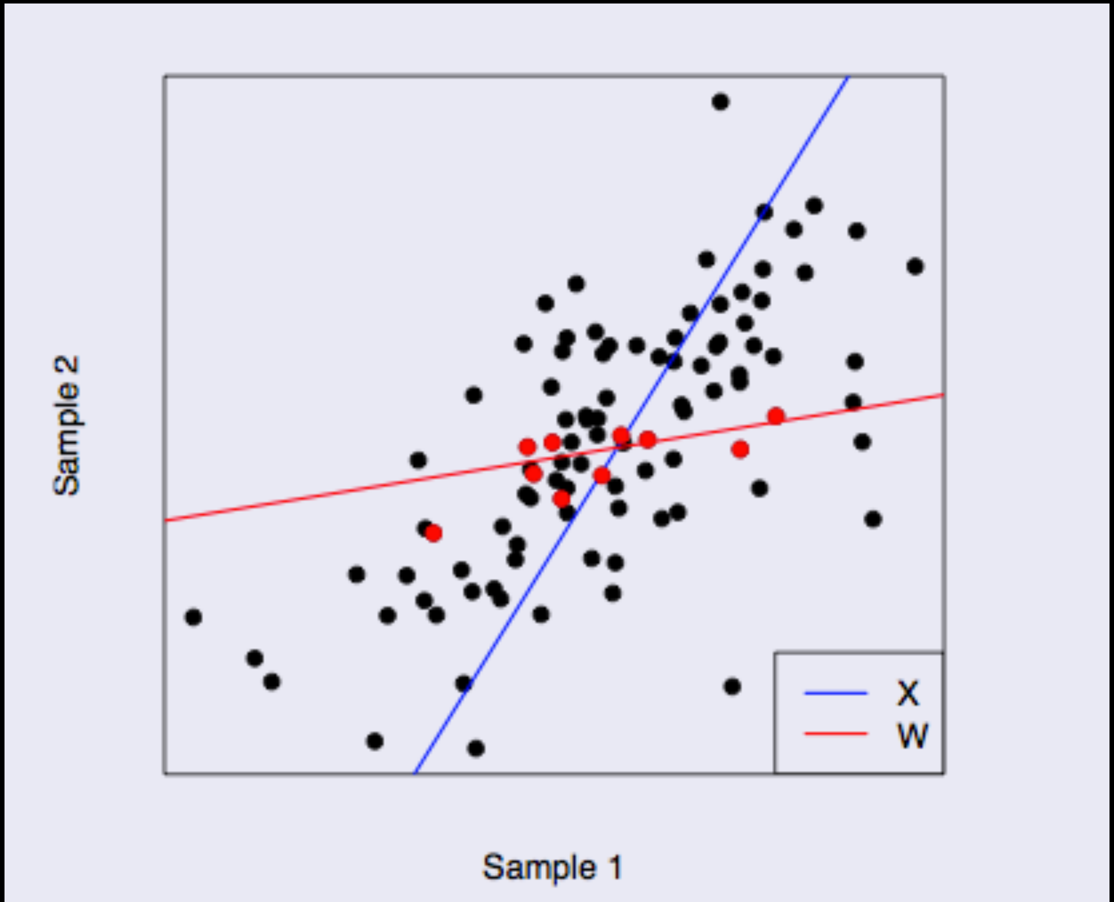
Two samples
Each dot a gene



Some ways of dealing with these problems with gene expression microarrays

- Standard linear regression
- EB linear regression (ComBat)
- Naïve factor analysis (SVD)
- Full Bayes using MCMC
- Variational Bayes (VIBES, Infer.NET, PEER)
- Surrogate Variable Analysis (SVA)
- Linear model with sparsity (LEAPP)
- Mixed model analysis (ICE)

We might have genes not affected by *X*



Call such genes negative controls.

Our solution: Use control genes

Negative controls: Assume $\beta_j = 0$.

Positive controls: Assume $\beta_j \neq 0$.

"controls" in this context means
"controls w.r.t. differential expression"

Some history

- Lucas *et al* (2006) *Sparse Statistical Modelling in Gene Expression Genomics*, created covariates from PCA based on signal from control and housekeeping probes

- Behzadi et al, (2007) *A component based noise correction method (CompCor) for BOLD and perfusion based fMRI* Neuroimaging.Ccreated covariates from PCA based on signal from "noise ROI" (white matter, CSF)

- Tradition in analytical chemistry/metabolomics: use of "internal standards"

Using the negative controls c

$$Y_c = W\alpha_c + \varepsilon_c$$

Just do a factor analysis on the negative controls!

Examples of negative controls

- housekeeping genes,
- spiked-in controls
- genes chosen carefully

This works!

Introducing the two-step: RUV-2

1. Do a factor analysis on Y_c to estimate W.
2. Then regress Y on X and the estimated W to get an estimate of β adjusted for W.

There are many ways to do the factor analysis, including SVD, the EM-algorithm, and using Infer.NET (variational Bayes), the last two needing a probability model.

SVD: Write $Y_c = U\Lambda V^T$, then put $W^\wedge = U\Lambda_k$, $\Lambda_k = k$ largest.

Ex: gender differences in the brain

(Vawter *et al*, **Neuropsychopharmacology** 2004)

- 5 men, 5 women
- 3 brain regions (AnCing, DLPFC, Cb)
- Each sample done in 3 labs
- 2 Affymetrix chip types: HGU95a, HGU95av2
- There should be (5+5) × 3 × 3 = 90 arrays, but 6 are missing, so there are just 84.

We'll ignore regions, and focus on gender.

Ex: gender differences in the brain, 2

- 12,685 probe sets
- 799 housekeeping genes, 33 spike-in negative controls
- Positive controls: genes on the Y and X chromosomes

There's no connection between Y and X here and the *Y* and *X* in my model – they are italicised, and colored!

Gender differences in the brain
# X/Y genes in the top 40

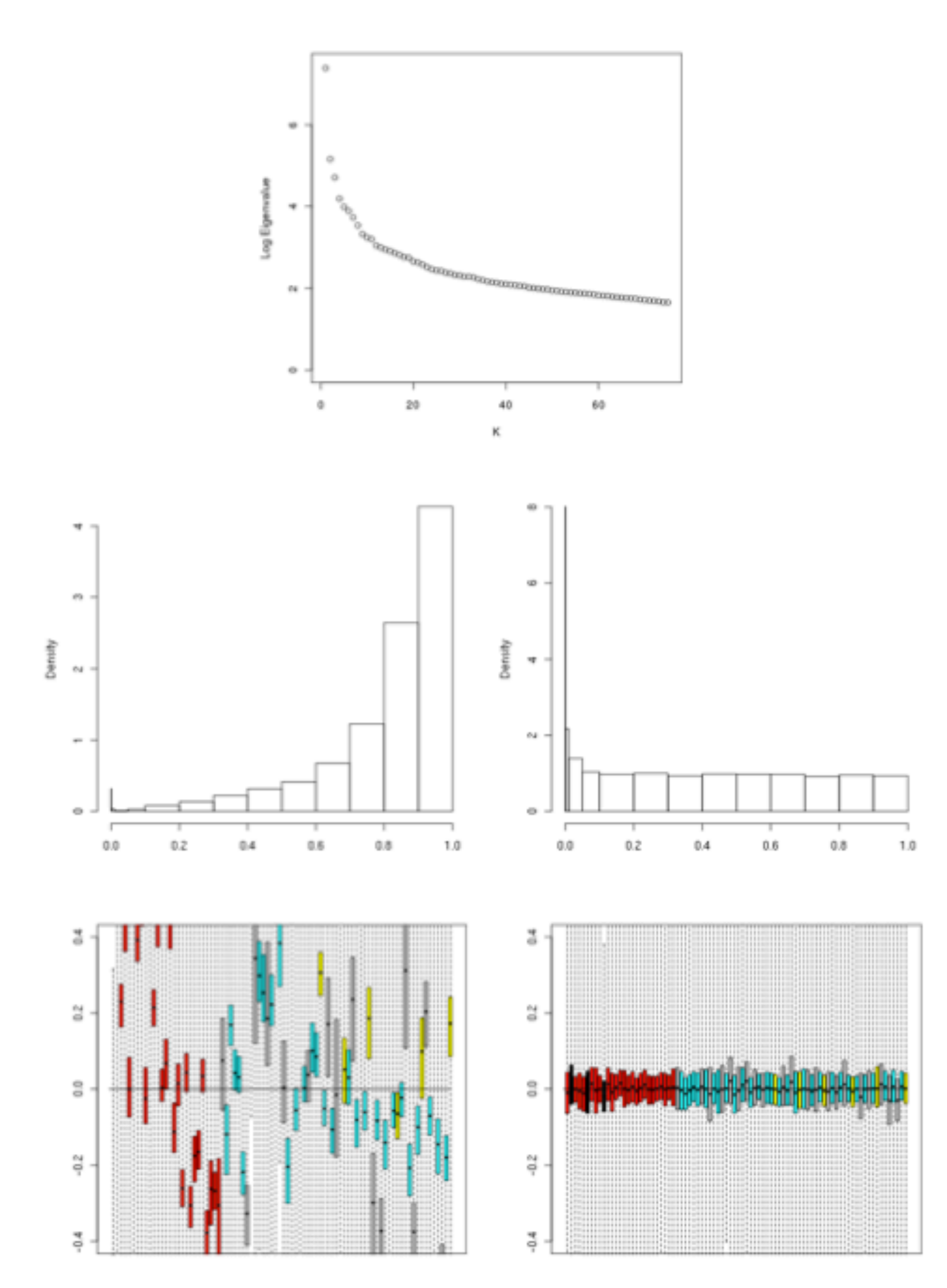
Method	W/o preprocessing	With preprocessing
No	7	13
Regression	6	16
SVA (IRW)	6	17
ComBat	14	17
RUV2-SVD	22	20
RUV2-EM	22	22

Preprocessing = standard RMA

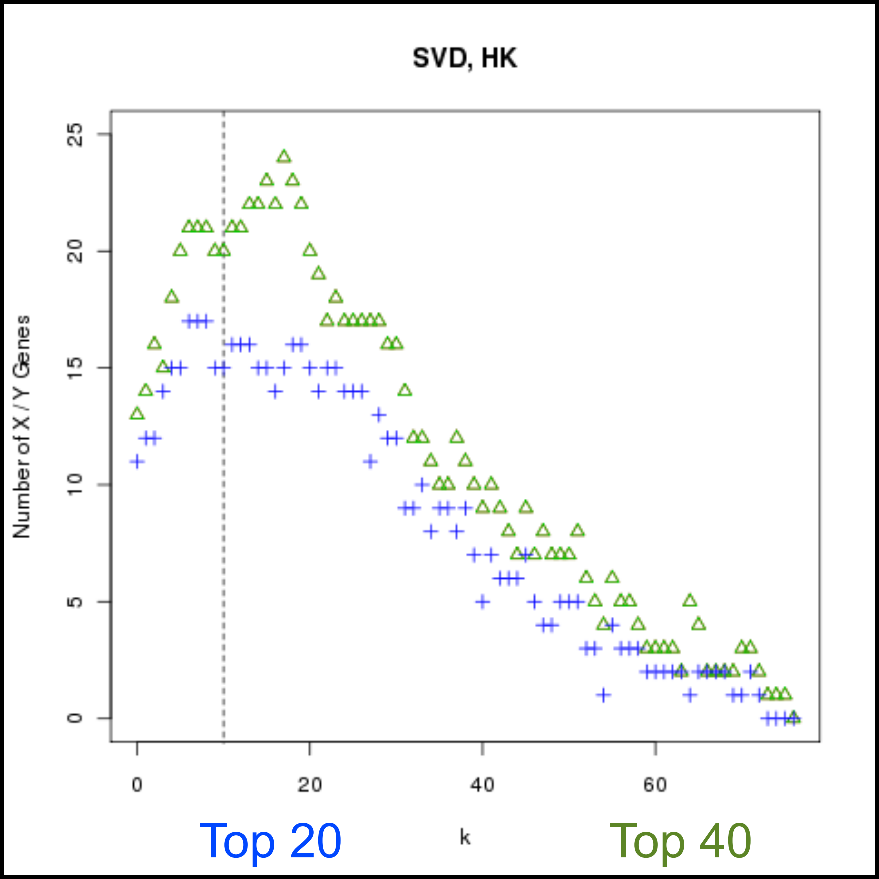
How did we find *k*?

Possible ways to determine *k*

- Scree plots
- Quality measures/plots
  - *p*-value histograms
  - RLE plots
- More math
  - hypothesis tests
  - move beyond factor analysis
- Positive controls



Number of X/Y genes in Top 20 /40



What next?

- We have an alternative to RUV2 called RUV4, which has some advantages.
- We have a form of RUV4 called RUVinv for which we do not need to estimate *k.*
- In all applications, the main issue is: what do we use as negative controls ? We can derive empirical negative control genes.
- We can ridge to improve conditioning
- We can smooth the gene-specific variances and get better Type 1 error control

Details in UC Berkeley Statistics Technical Report #820

Gender data, 4: not preprocessed

Method	#X/Y in top 100	Type 1 error × 100
Unadjusted	10	0
SVA-IRW	12	0
LEAPP	19	1
ICE	17	0
RUV4 (HK)	29	12
RUVinv (HK)	26	7
RUVinv-evar (HK)	26	6
RUVrinv-evar (HK)	28	6
RUVrinv-evar (full)	32	6
RUVrinv-evar (emp)	30	6

Relation of negative controls to instrumental variables

Instruments are variables that are correlated with the factor of interest but uncorrelated with the error term (or in our case, the unwanted variation).

They can be used to obtain unbiased estimates of the effect of interest (in our case, β).

Let V be a full rank $m{\times}r$ matrix of instruments such that m > r ≥ p, such that $V'W = 0$, and such that $V'X$ is full rank. The IVLS estimator of β would be

$$[X'V(V'V)^{-1}V'X]^{-1}X'V(V'V)^{-1}V'Y$$

Analogous formulae

Alternatively, we may write the IVLS estimator as

$$(X'P_V X)^{-1} X'P_V Y$$

Compare this to the RUV-2 estimator

$$(X'R_{\hat{W}} X)^{-1} X'R_{\hat{W}} Y$$

Comparison

With IVLS we identify a “safe” subspace using instruments. Instruments are variables that we assume lie within the “safe” subspace.

With RUV-2 we identify a “safe” subspace using negative controls. Negative controls are variables that we assume lie within the “dangerous” subspace that is the orthogonal complement of the “safe” subspace.

With both IVLS and RUV-2 there is the caveat that *X* must not be orthogonal to the “safe” subspace.

In the case of IVLS, this means that *V* must be reasonably correlated with *X*; we want to avoid weak instruments.

In the case of RUV-2, this means that *X* must lie outside **R**(W^); the control genes must not be influenced by *X*.

What next?

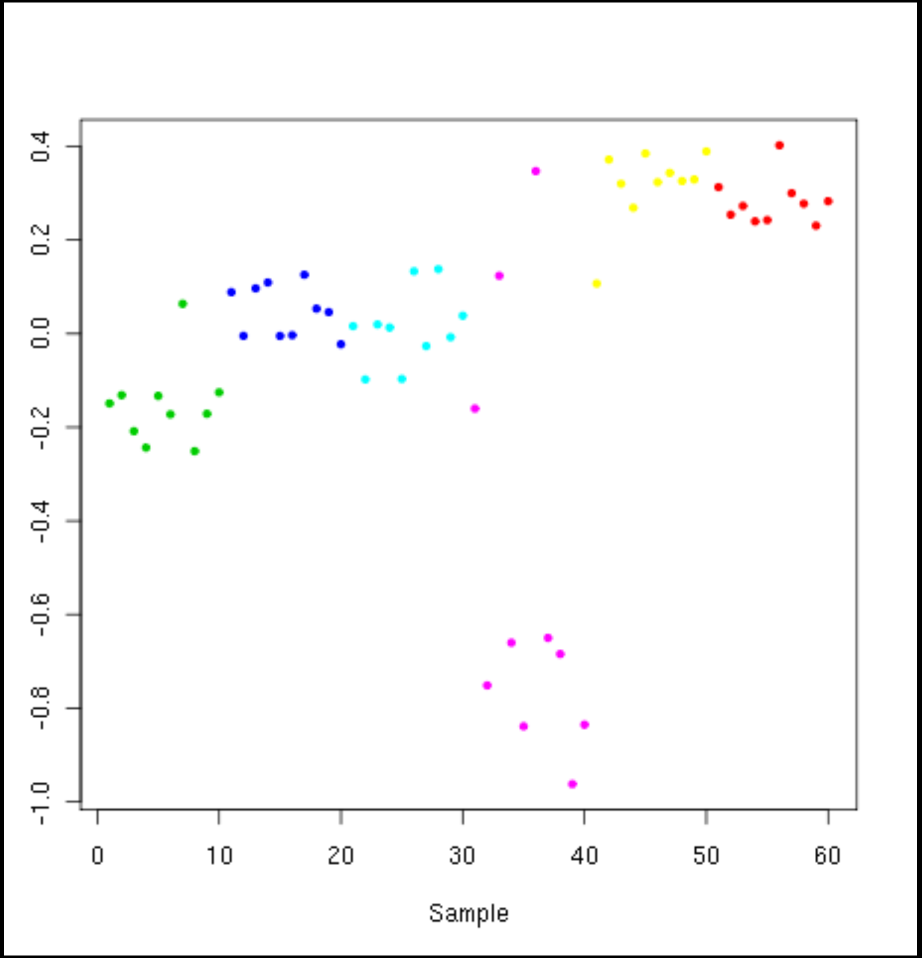
- Next I'll give a quick look at some applications of these ideas to various examples.
- In all applications, the main issue is: what do we use as negative controls and positive controls, if any.

MicroArray Quality Control dataset

- Two mRNA samples (Stratagene Universal Human Reference RNA, and Ambion Human Brain RNA)
- Each sample was assayed 5 times at each of 6 sites on the Affymetrix HU133Plus2.0 platform: 60 arrays in all.
- The labs at the different sites have all done a pretty good job on their assays. However, one lab lacked experience.
- Here we let our approach discover the site effects, not including them as dummy variables (you will see why not).

The figure (w_1) shows clear site effects

(different colors represent different sites)



Note the purple: whatever factor is varying from site to site is also varying within this site.

Dummy variables would not have worked as well here.

The effects are small.

Removing severe batch effects

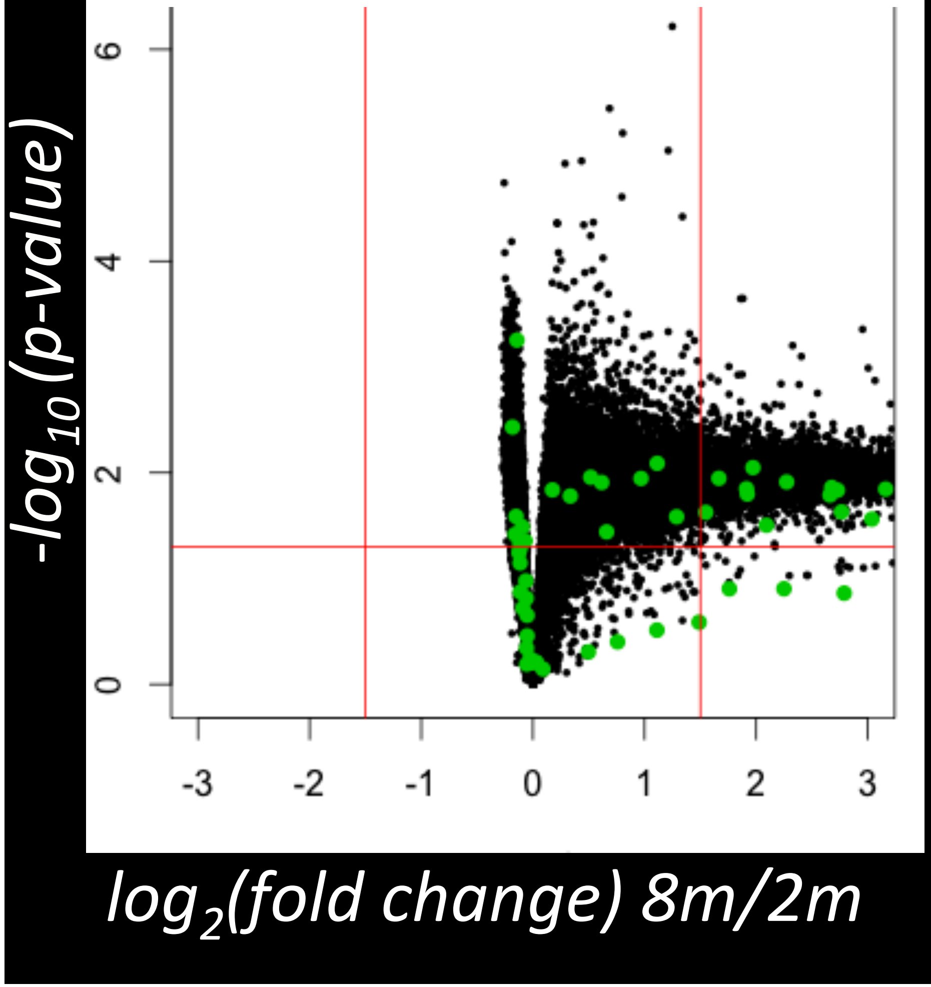
- Back to our mouse model of *retinitis pigmentosa* (loss of rod and later cone photoreceptors).

- Initially no significantly downregulated retinal genes were found between 2 and 8 months (left *volcano plot* on the next slide).

- Using RUV (right plot on the next slide), we were able to find several significantly down-regulated retinal, even cone-specific genes, which were later confirmed.

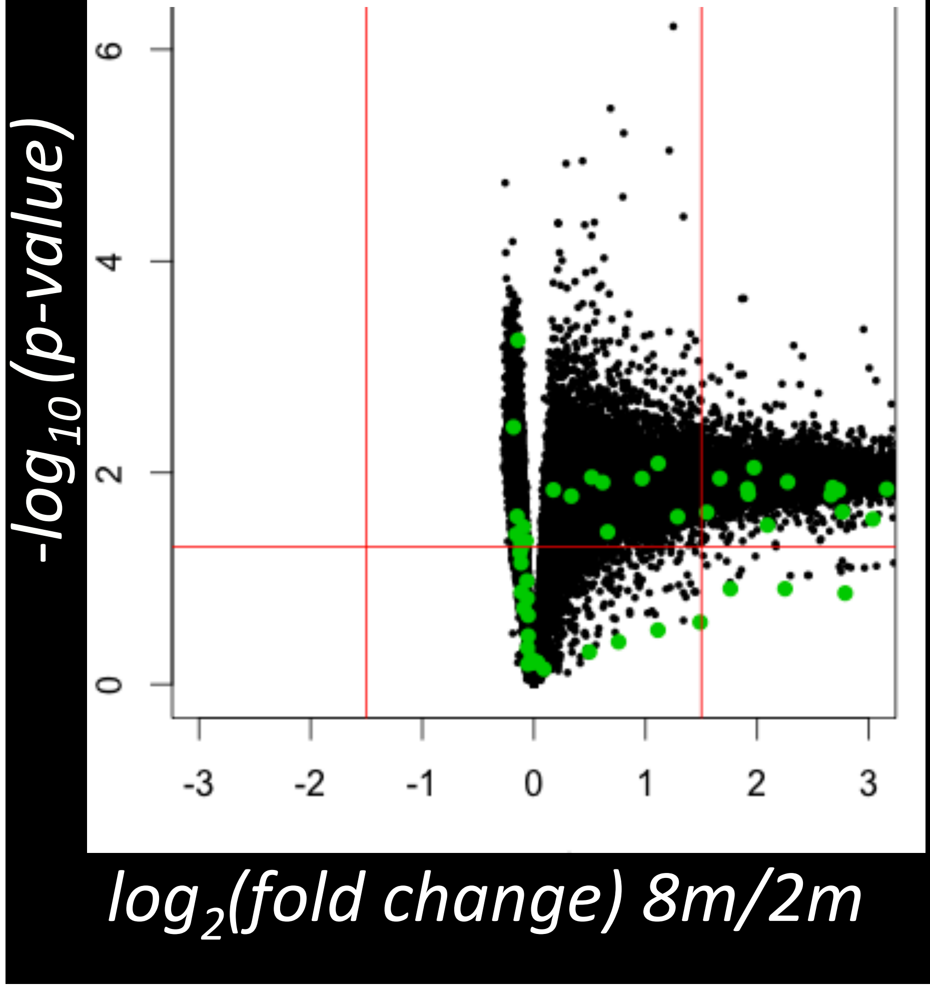
Standard analysis

Green dots: genes expressed in the retina

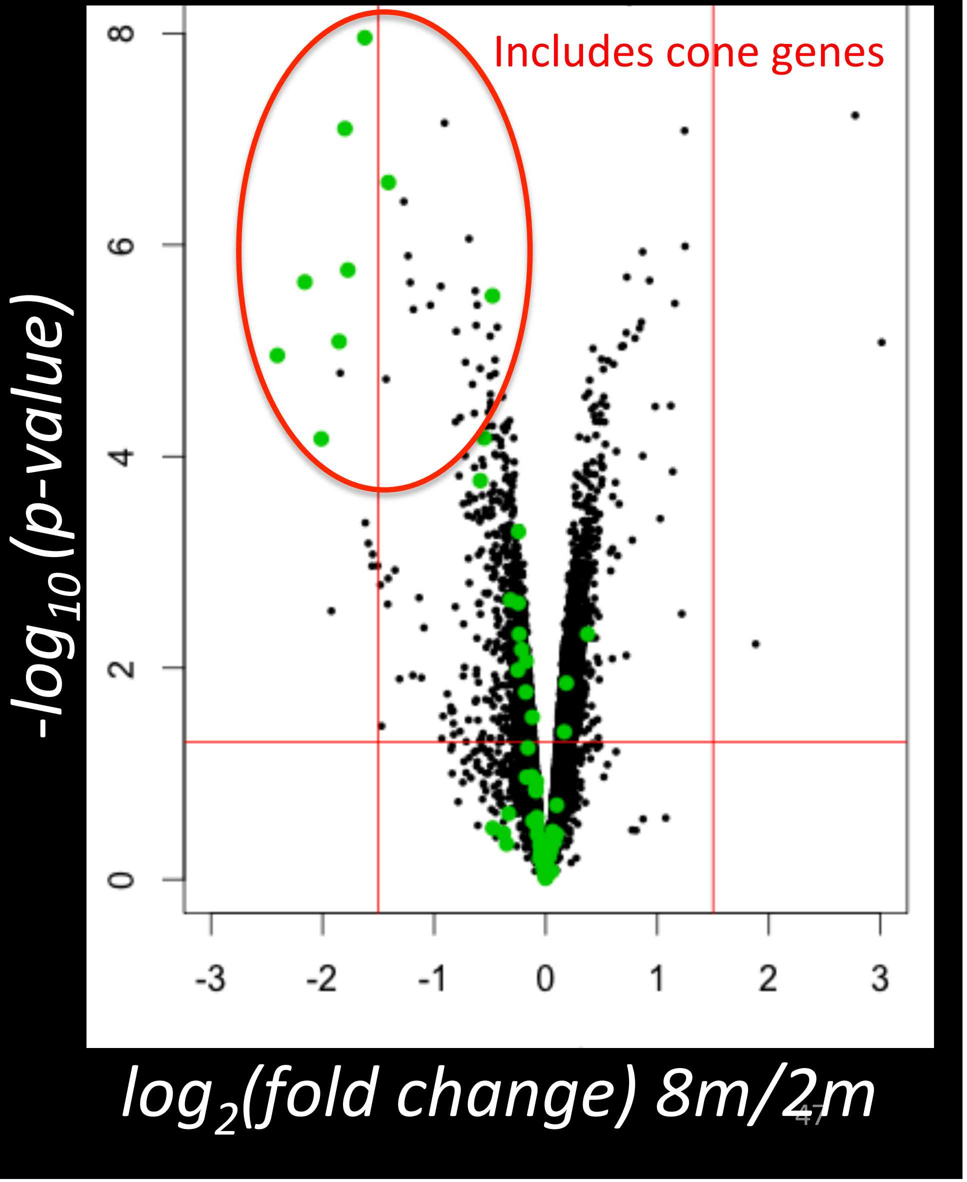


Standard analysis

Green dots: genes expressed in the retina

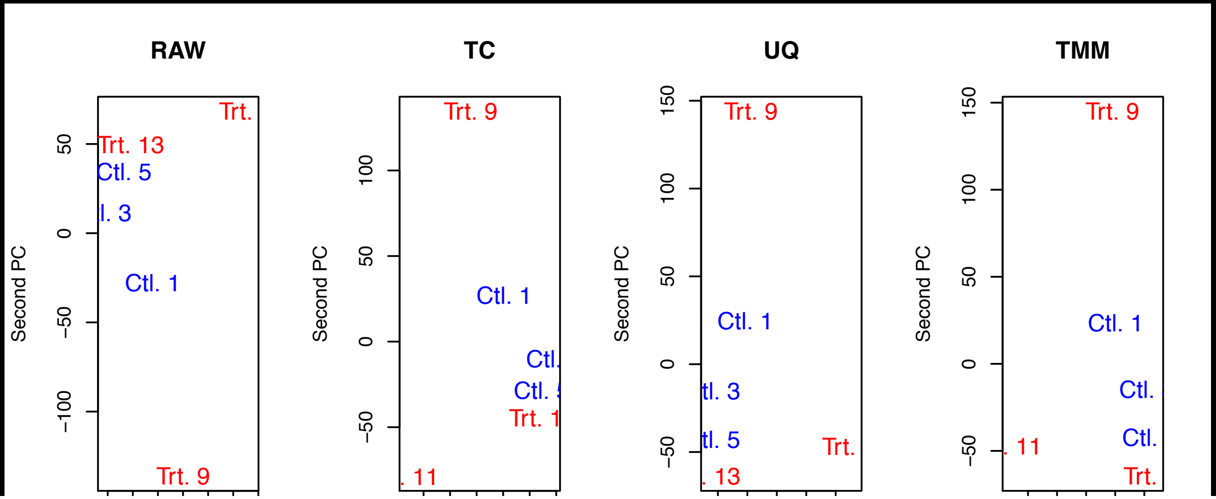


Analysis with RUV

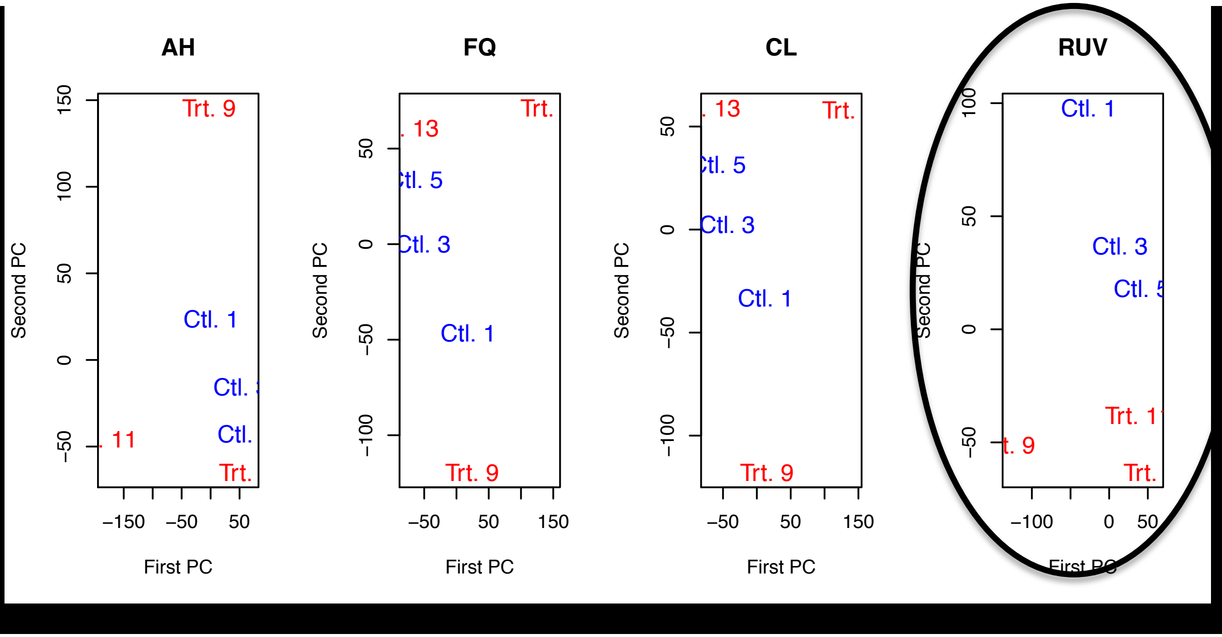

Includes cone genes

Back to our 3 treatment vs 3 control (in duplicate) RNA-seq zf experiment

PC2 vs PC1 of normalized data

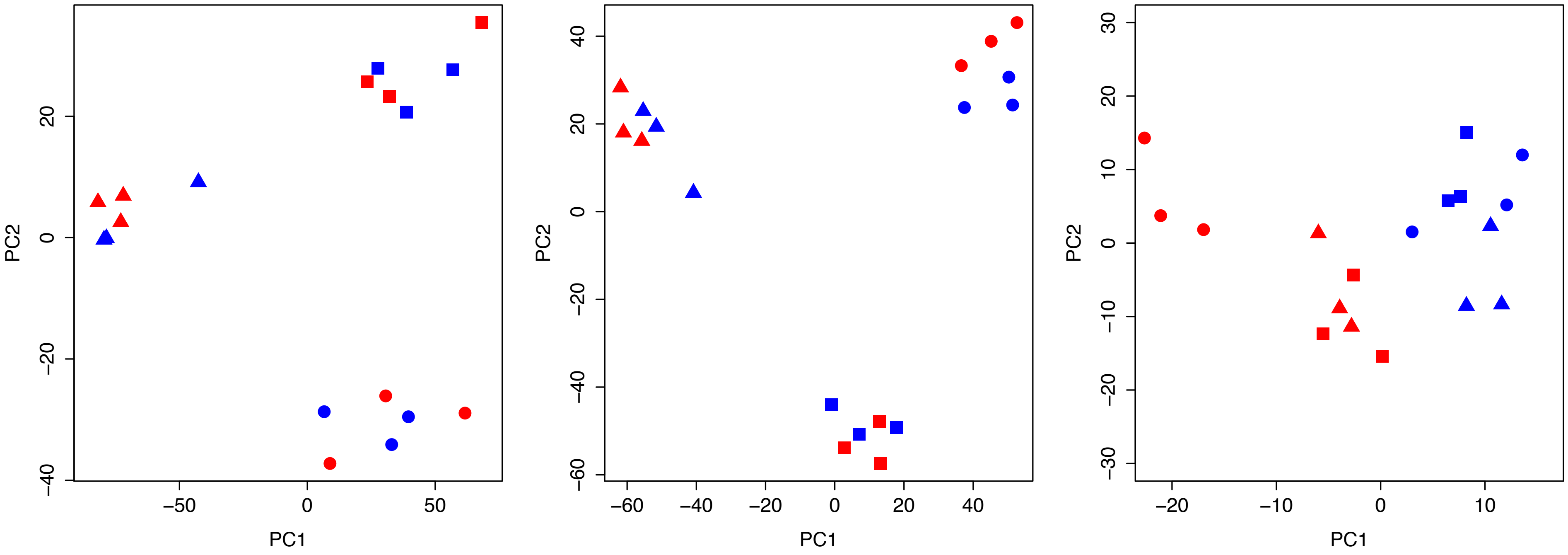


We'd hope to see the trt vs. ctl difference wouldn't we?



Back to combining 3 sets of 3 KO vs 3 WT T-cell microarray experiments (with same WT)

Raw Q-norm RUVrandom



Blue: wild-type, Red: knock-out.

Shapes: Different experiments (KOs)

Summary

With *very simple* statistical methods, we can:

- Use negative control genes to estimate the unwanted factors,
- Use positive control genes or other methods to estimate the number of unwanted factors.

With *slightly more complex* statistical methods, we can avoid estimating the number of unwanted factors, and relax the control gene assumption.

In later work we

- Apply these differential expression ideas in other contexts; microarray methylation data, mass spec metabolomic data, RNA-seq gene expression data,…
- We have analogous results for prediction (classification), clustering and correlating
- We can combine different studies on the same platform (e.g. two or more Affymetrix studies), on similar but distinct platforms (e.g. Affymetrix, Agilent and Illumina microarray studies), and studies on totally different platforms, e.g. GC-MS and LC-MS metabolomic data, microarray and RNA-seq data.

Acknowledgements

Johann Gagnon-Bartsch

Department of Statistics,

University of California at Berkeley

Laurent Jacob

Laboratoire Biométrie et Biologie Evolutive, Université Lyon, CNRS, UMR 5558, Villeurbanne

Francois Collin, Genomic Health

Moshe Olshansky, WEHI

Alysha De Livera, Univ Melbourne

Jovana Maksimovic, MCRI

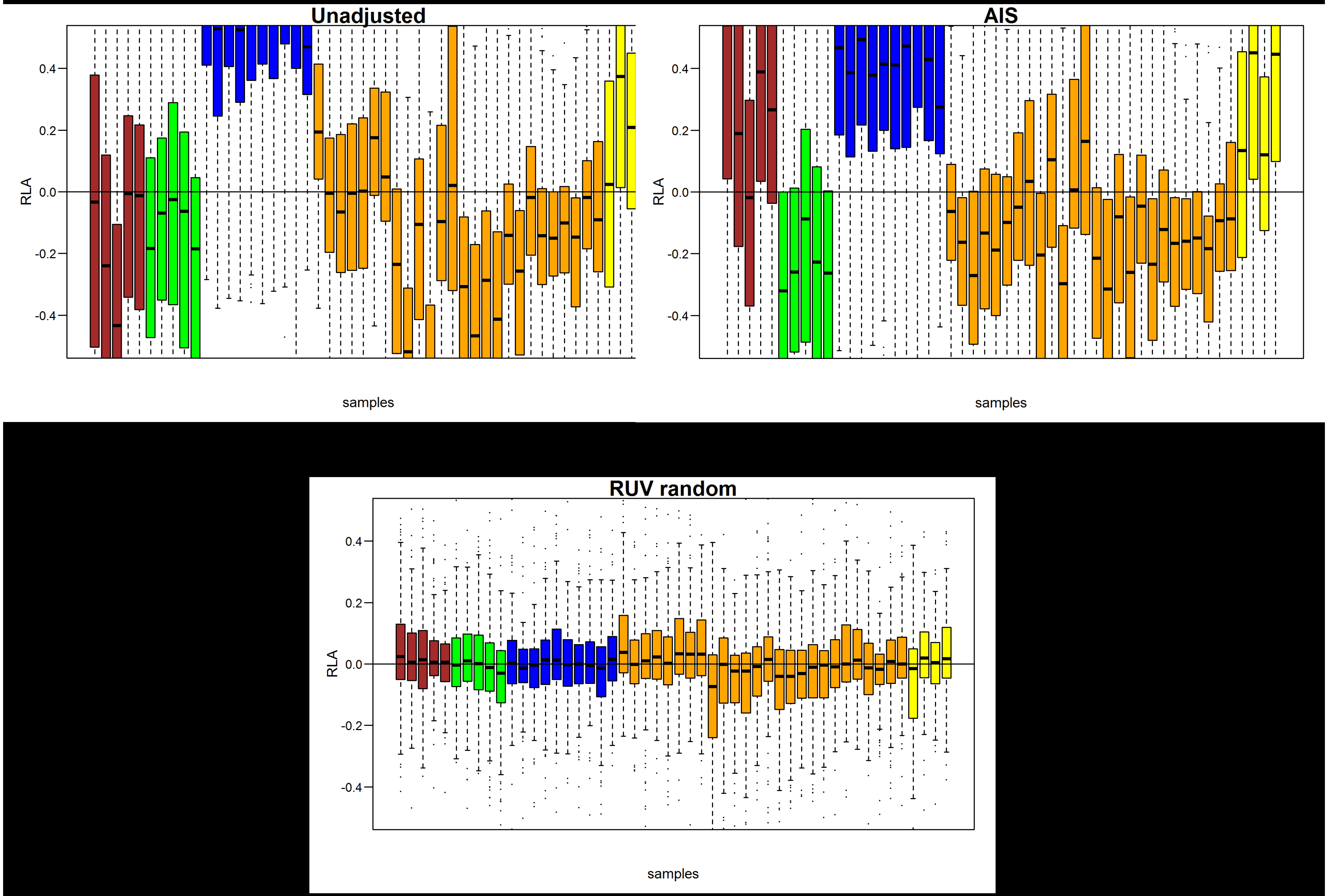
Clustering or “cleaning”

The problem

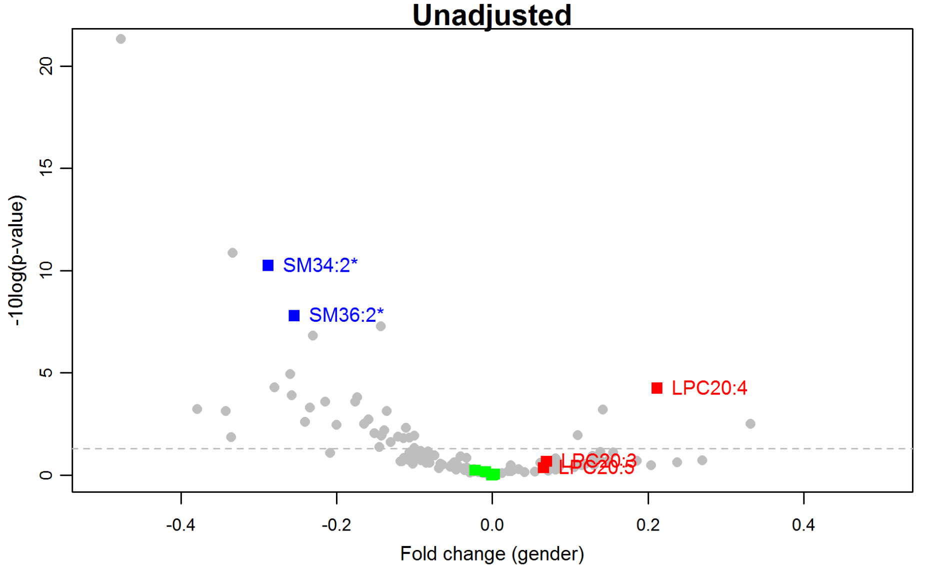
We now assume we don't know X any more, e.g. for clustering, or cleaning a dataset.

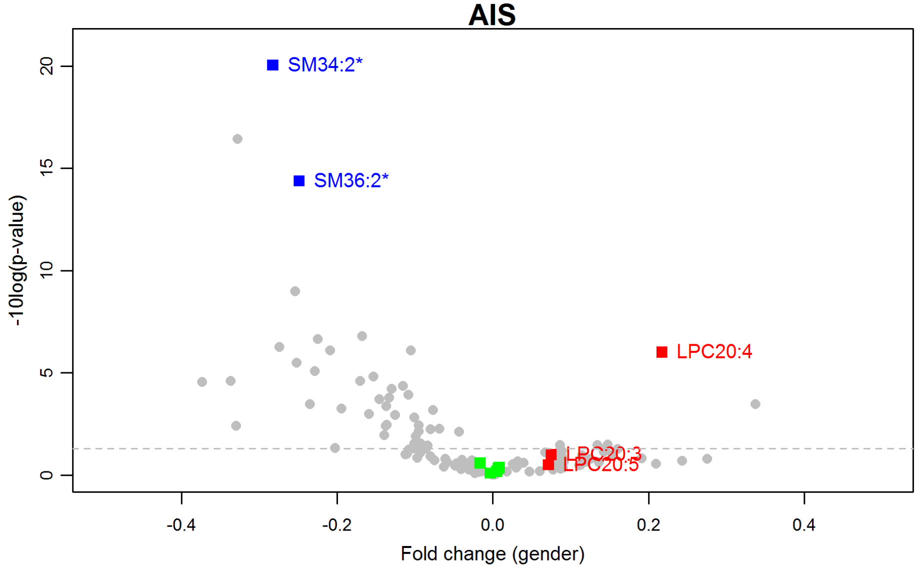
We can still estimate W as before, using Y_c, but then we can't do the regression step.

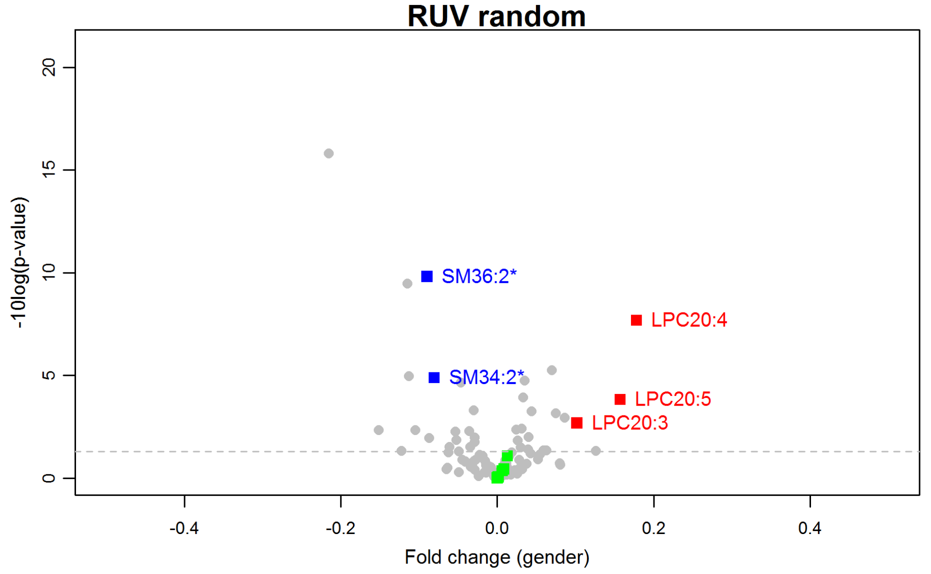
We have several statistical approaches to this problem, details omitted. One is RUV-random.


Unadjusted
RLA
0.4
0.2
0.0
-0.2
-0.4
samples
AIS
RLA
0.4
0.2
0.0
-0.2
-0.4
samples
RUV random
RLA
0.4
0.2
0.0
-0.2
-0.4
samples

We know of 5 age-related mets: 3 going up, 2 going down.
Look at volcano plots of age effects, adjusted for sex and BMI





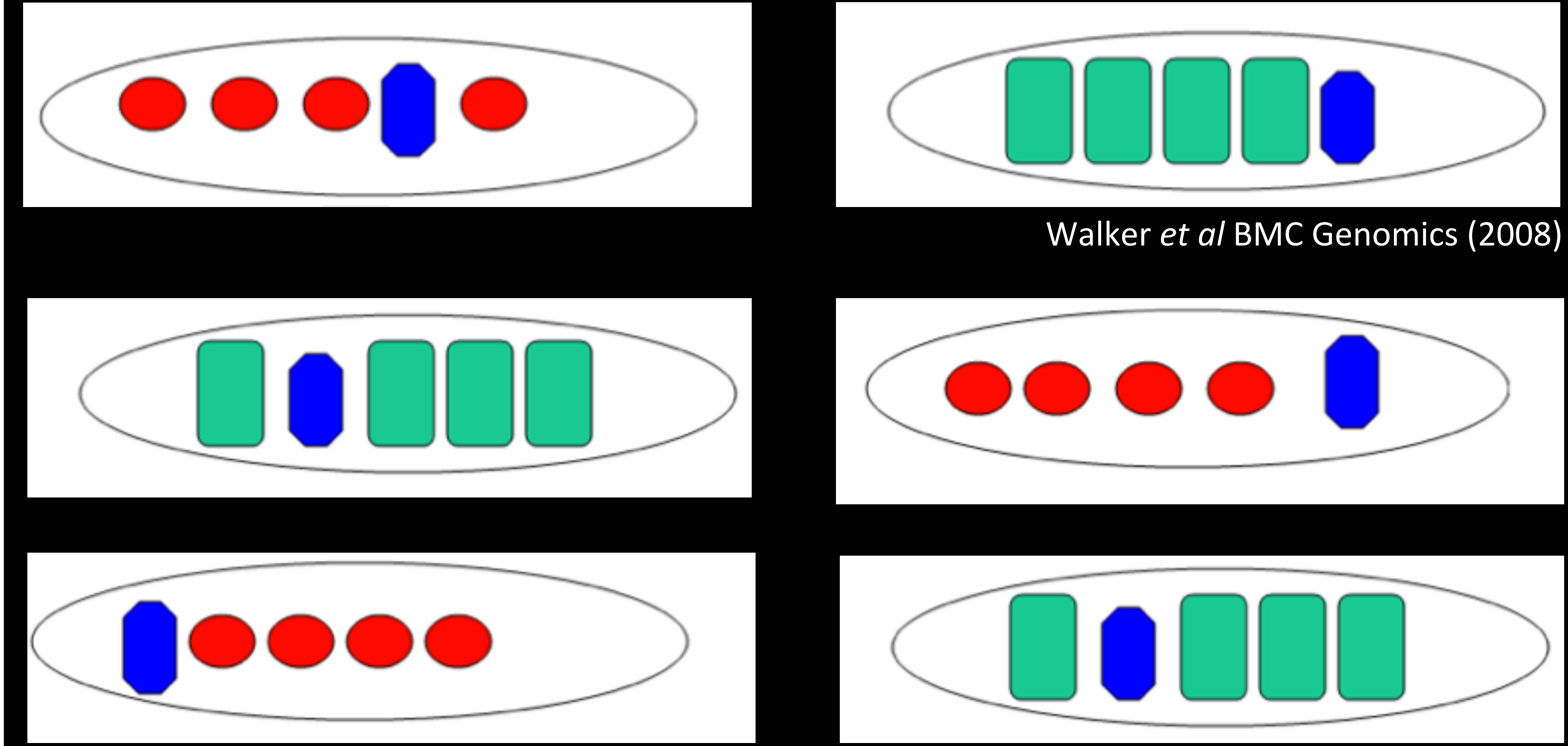


RUV-random pulls two mets up out of the pool

The biological solution

Reference controls are simply technical replicates, but replicates whose variation might well be representative of the very unwanted variation we wish to remove. That’s going to be our hope when we use them. (We’ll check the results, of course.) *Any* replicates will help, but *reference controls* have a better chance of spanning the space of UV.

Diagram illustrating a *reference control* in 6 batches of 5 samples of 2 types



Walker *et al* BMC Genomics (2008)

Note that a naïve batch adjustment here would equalize red and green, on average.

How do we use the reference control replicates? Simplest version.

- Note that the reference control *Ys* have the same (unknown) *X*, and so their row differences Y^d satisfy

$$Y^d = W^d\alpha + \varepsilon^d$$

- Estimate α from the svd of the left hand side, say α^ $= E_kQ^T$, where $Y^d = PEQ^T$.
- Plug α^ into the formula $Y_c = W\alpha_c + \varepsilon_c$ for negative control genes, and estimate *W* by linear regression.
- Once *W* and α have been estimated, subtract *W*^α^.

This too works! (but we can do better)