# GS01 0163
# Analysis of Microarray Data

Keith Baggerly and Brad Broom
Department of Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
bmbroom@mdanderson.org

11 November 2010

# Lecture 21: Batch Effects

- What are batch effects? How do we plan for them?
- How do we detect them?
- How do we correct for them?
- DWD – Benito et al, Bioinf 20:105-14, 2004
- COMBAT – Johnson et al, Biostat 8:118-27, 2007
- Review – Leek et al, Nat Rev Gen 11:733-9, 2010
- TCGA batches and genes

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Batch Effect Sources and Planning

This is a free-for-all slide, in that we've talked about these several times now.

Some causes?

Some ways of planning?

An ovarian cancer case study:
Dressman et al, JCO 2007
Baggerly et al, JCO 2008.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Signatures and Pathways for Response in Ovarian Cancer

An Integrated Genomic-Based Approach to Individualized Treatment of Patients With Advanced-Stage Ovarian Cancer

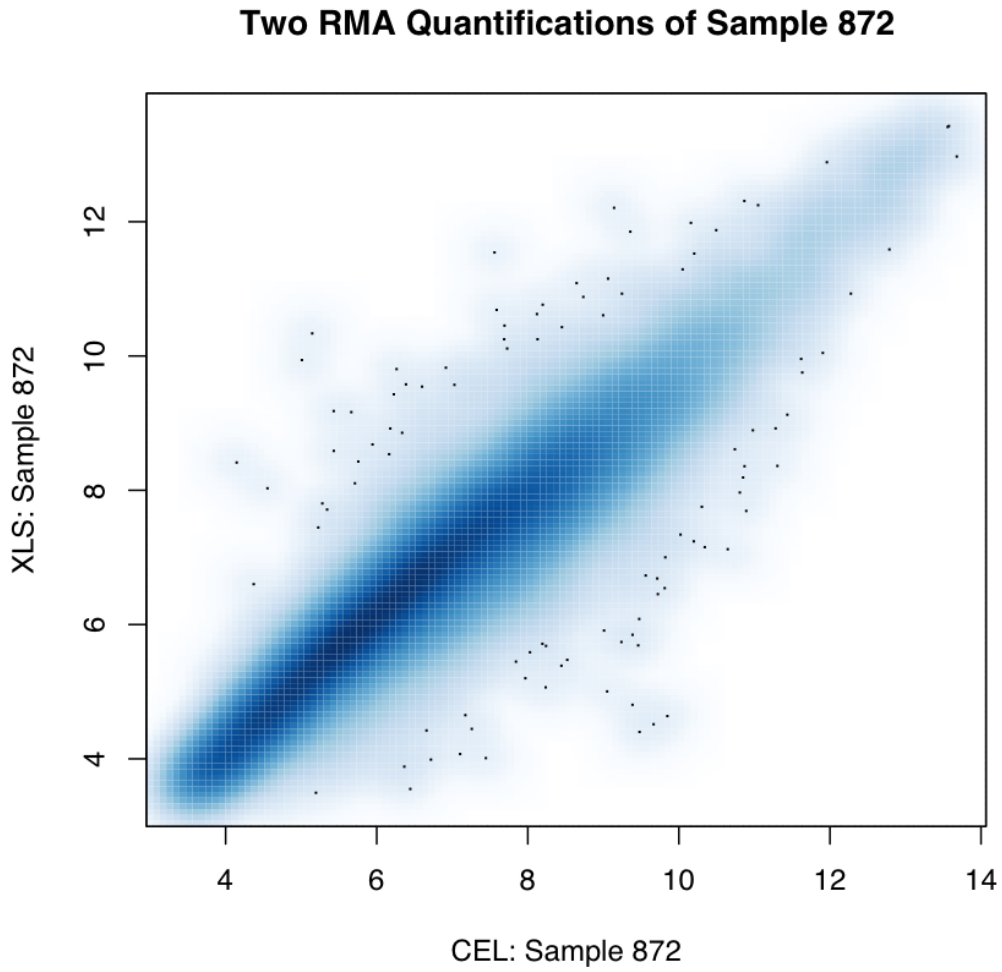
*Holly K. Dressman, Andrew Berchuck, Gina Chan, Jun Zhai, Andrea Bild, Robyn Sayer, Janiel Cragun, Jennifer Clarke, Regina S. Whitaker, LiHua Li, Jonathan Gray, Jeffrey Marks, Geoffrey S. Ginsburg, Anil Potti, Mike West, Joseph R. Nevins, and Johnathan M. Lancaster*

JCO, Feb 10, 2007.

Using profiles of 119 ovarian tumors, they looked for signatures of response to cisplatin-based chemo. The also looked at the deregulation levels of 5 pathways (Src, $\beta$-catenin, Myc, E2F3, and Ras), trying to relate them to survival.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

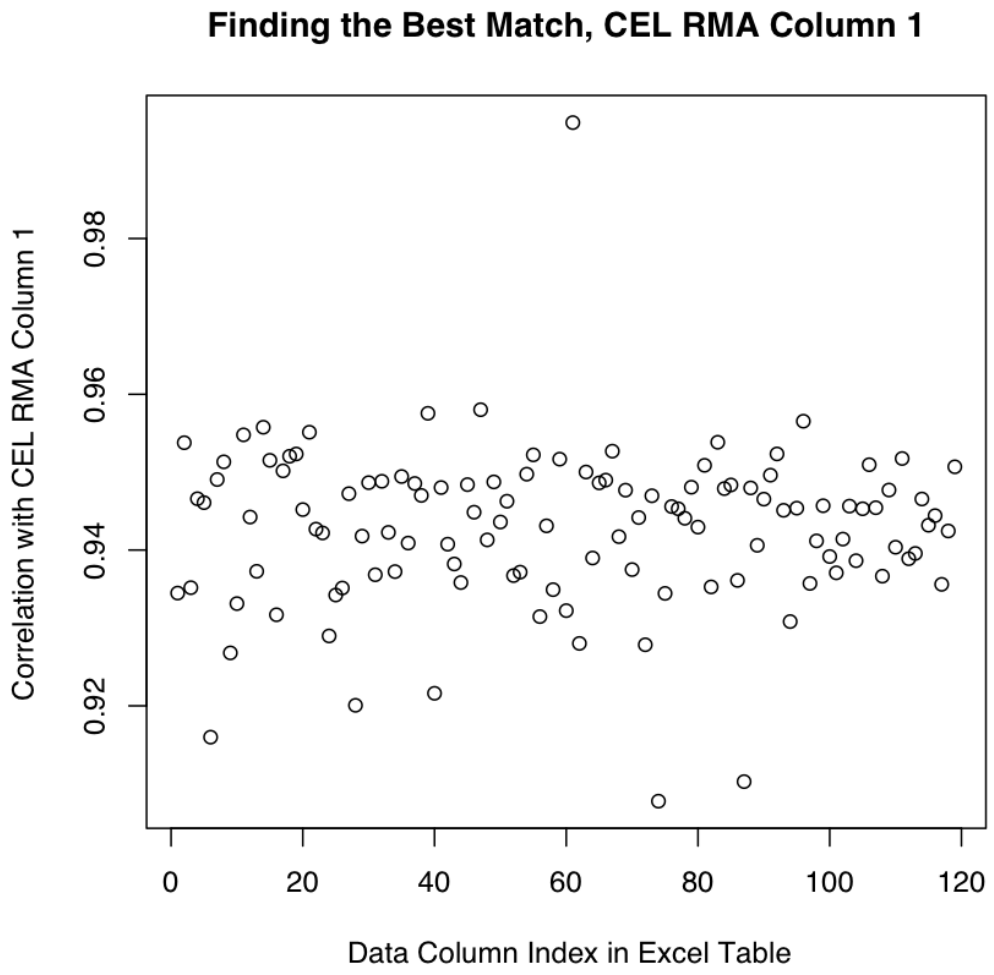
# Checking Agreement




Ours vs theirs. We expected better (fewer outliers).

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

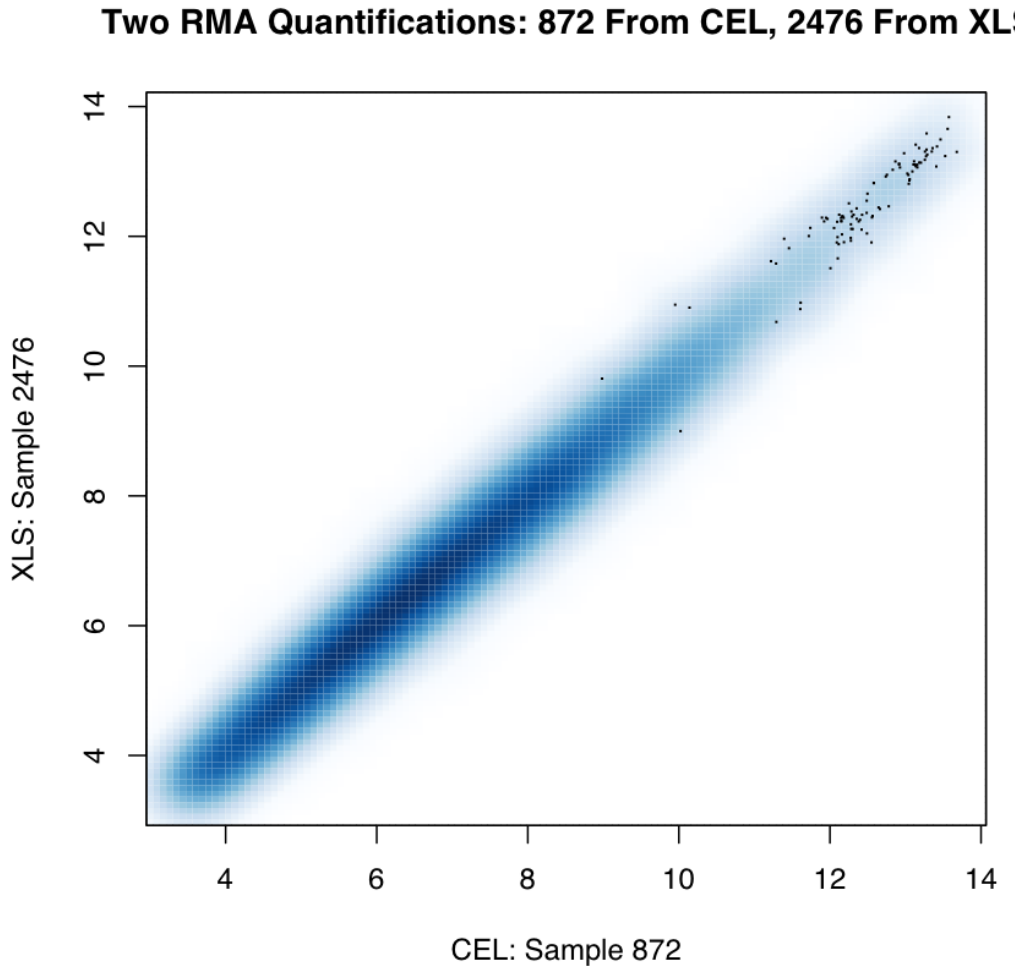
# Looking at Their Other Quants




Which one would you pick?

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

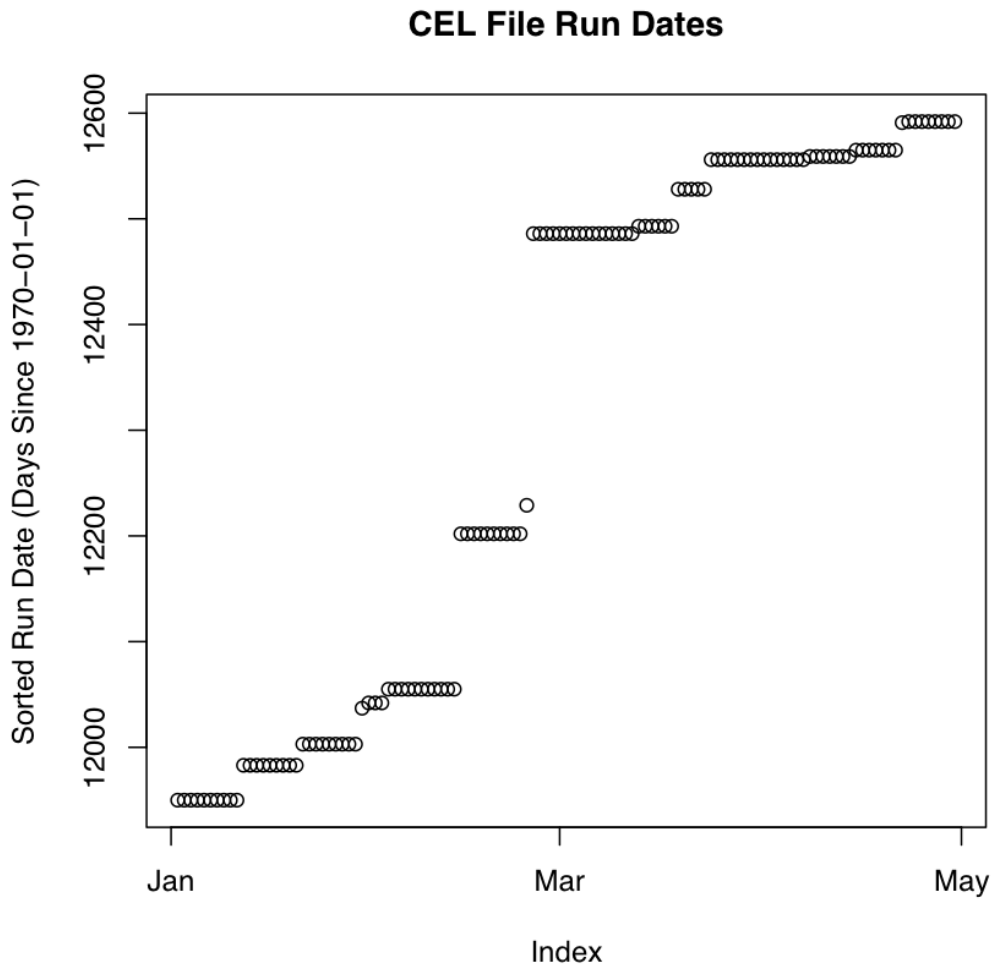
# Looking at The "Best" Fit




Same array. *Different* names (2476 from XLS, 872 from CEL).

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Grab One More Thing: Run Date




Should we be worried?

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Some Berchuck et al Clinical Info

| | |
|---|---|
| 2392 | Early Stage |
| 2393 | Early Stage |
| 1772 | Long |
| 1773 | Long |
| 1774 | Long |
| 1775 | Long |
| 1776 | Long |
| 1777 | Long |
| 1778 | Long |
| 1779 | Long |
| 1780 | Long |
| 1781 | Long |
| 1900 | Long |

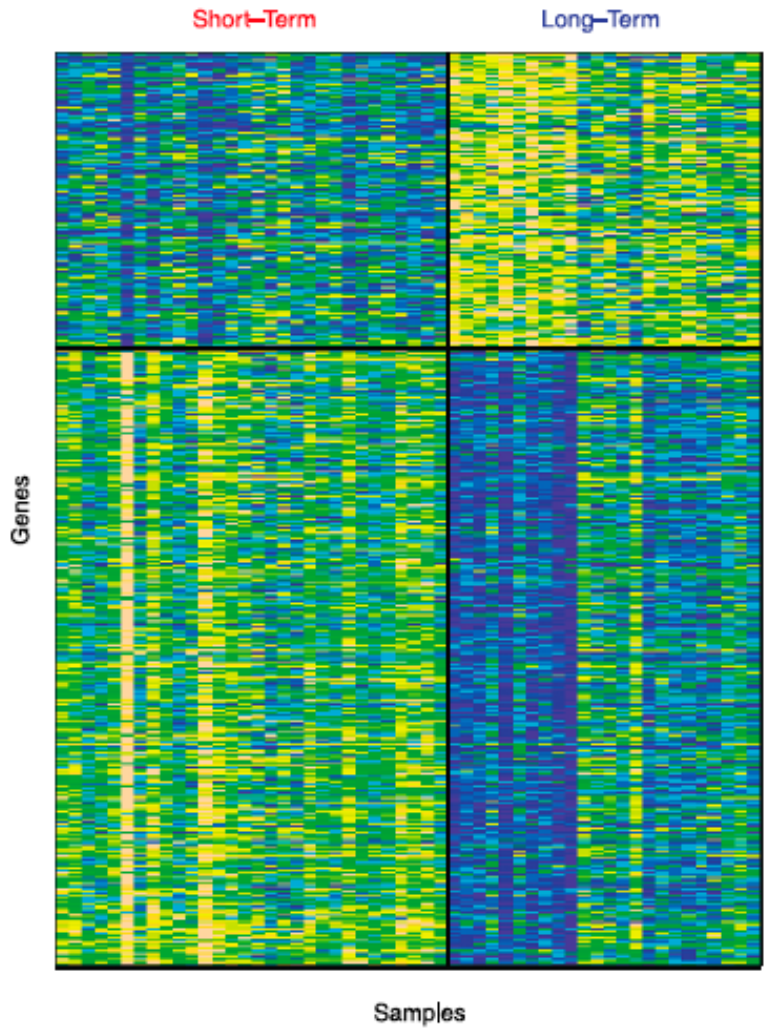
© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# The Berchuck et al Dat Headers

```
0074_1772_h133a_872.cel:DatHeader=[0..37764] .. 09/20/02 11
0074_1773_h133a_922.cel:DatHeader=[0..33251] .. 09/20/02 11
0074_1774_h133a_1451.cel:DatHeader=[0..43335] .. 09/20/02 1
0074_1775_h133a_1526.cel:DatHeader=[0..45012] .. 09/20/02 1
0074_1776_h133a_1784.cel:DatHeader=[0..46104] .. 09/20/02 1
0074_1777_h133a_1834.cel:DatHeader=[0..42469] .. 09/20/02 1
0074_1778_h133a_1846.cel:DatHeader=[0..36713] .. 09/20/02 1
0074_1779_h133a_2075.cel:DatHeader=[0..37459] .. 09/20/02 1
0074_1780_h133a_2204.cel:DatHeader=[0..43583] .. 09/20/02 1
0074_1781_h133a_2419.cel:DatHeader=[0..46101] .. 09/20/02 1
0074_1827_h133a_.08.cel:DatHeader=[0..46104] .. 10/23/02 12
0074_1828_h133a_860.cel:DatHeader=[0..46102] .. 10/23/02 12
```

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

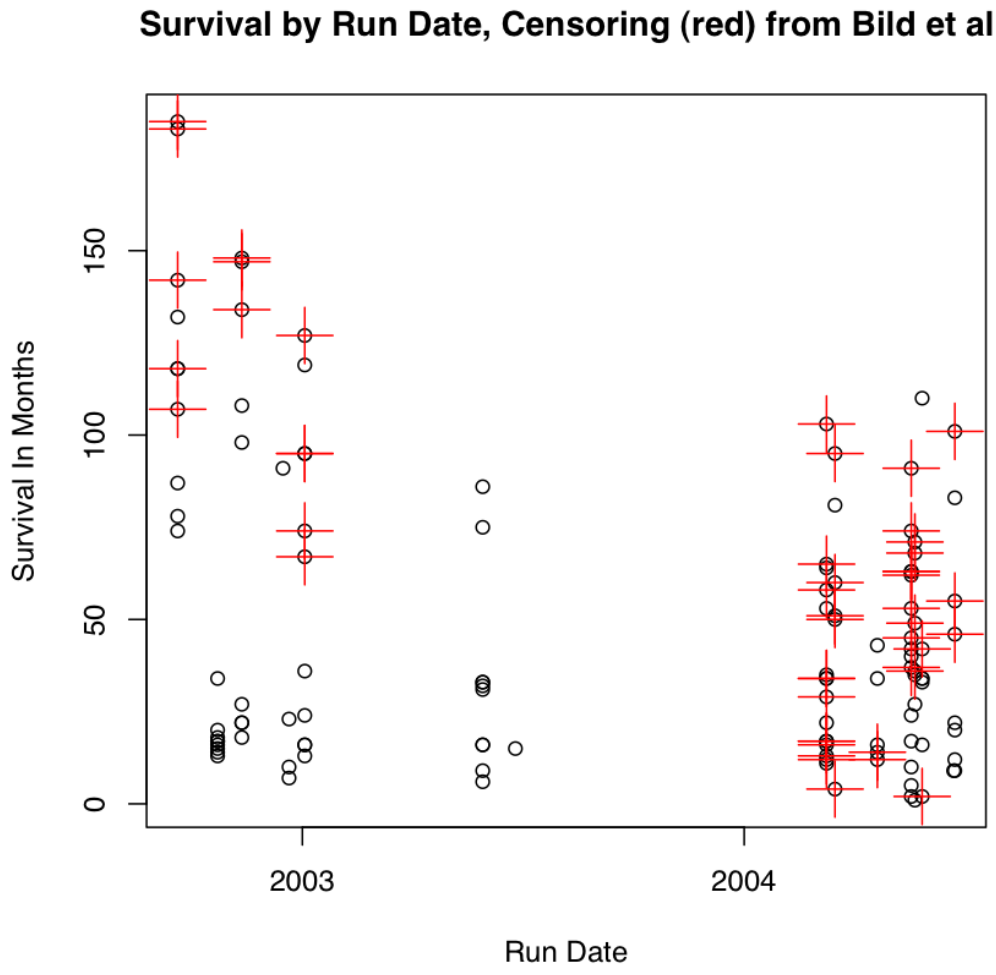
# The Berchuck et al Heatmap




Can you spot clusters? Expression tracks with batch.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

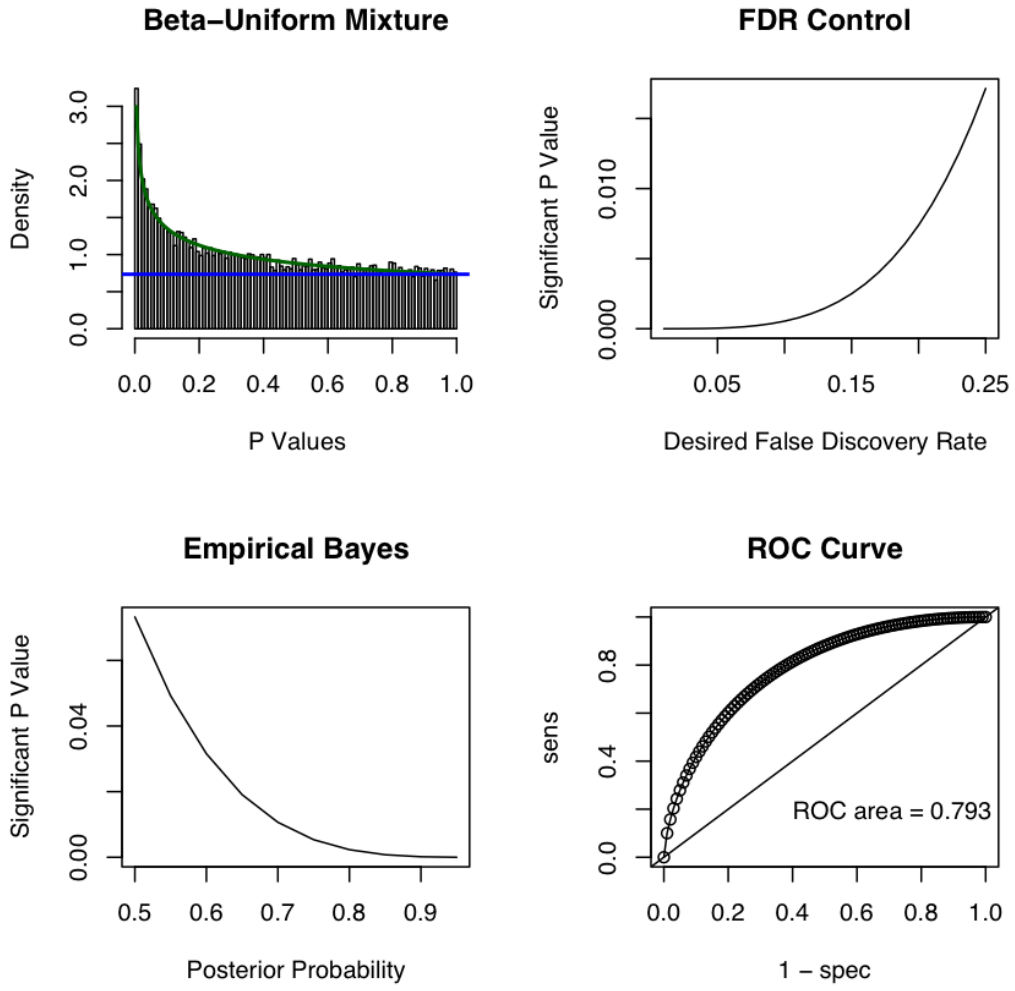
# Survival By Date (Bild Censoring)




Survival is confounded with date.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

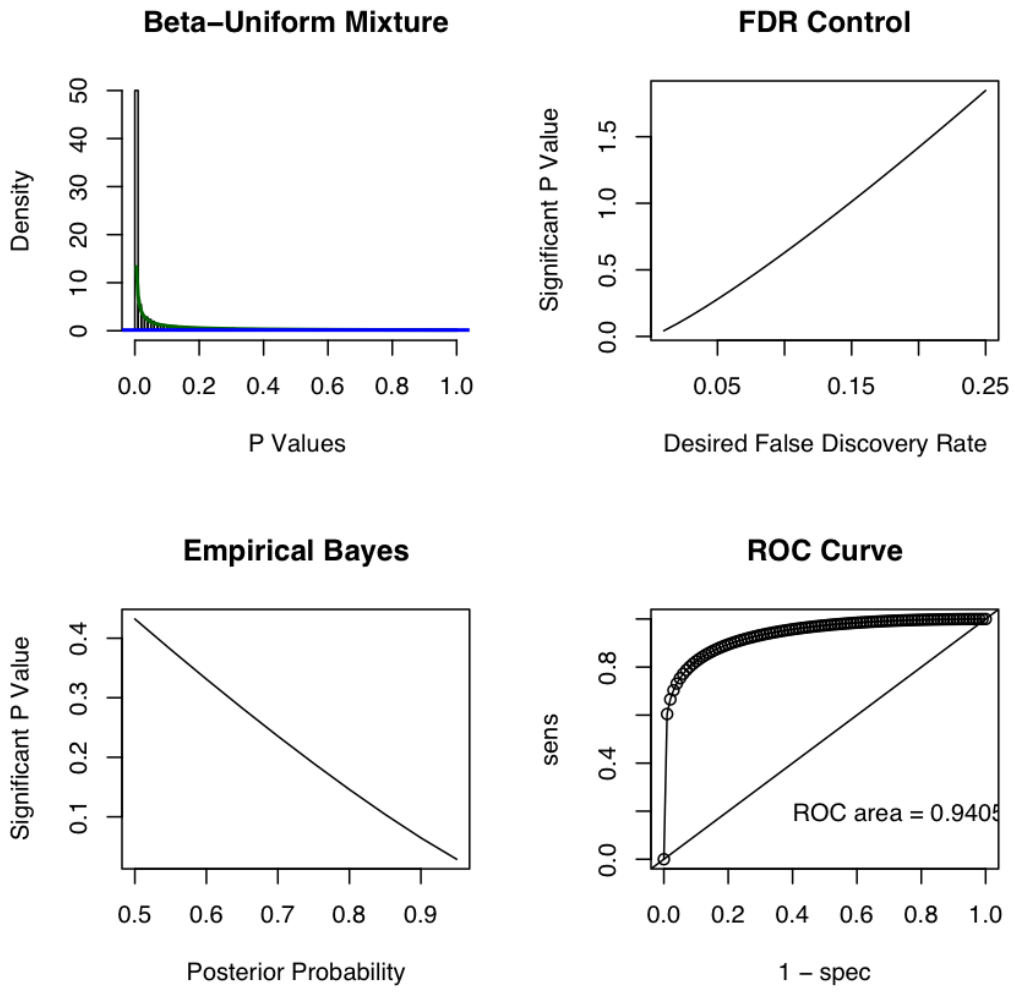
# Is There Division by CR/NR?




Maybe something.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

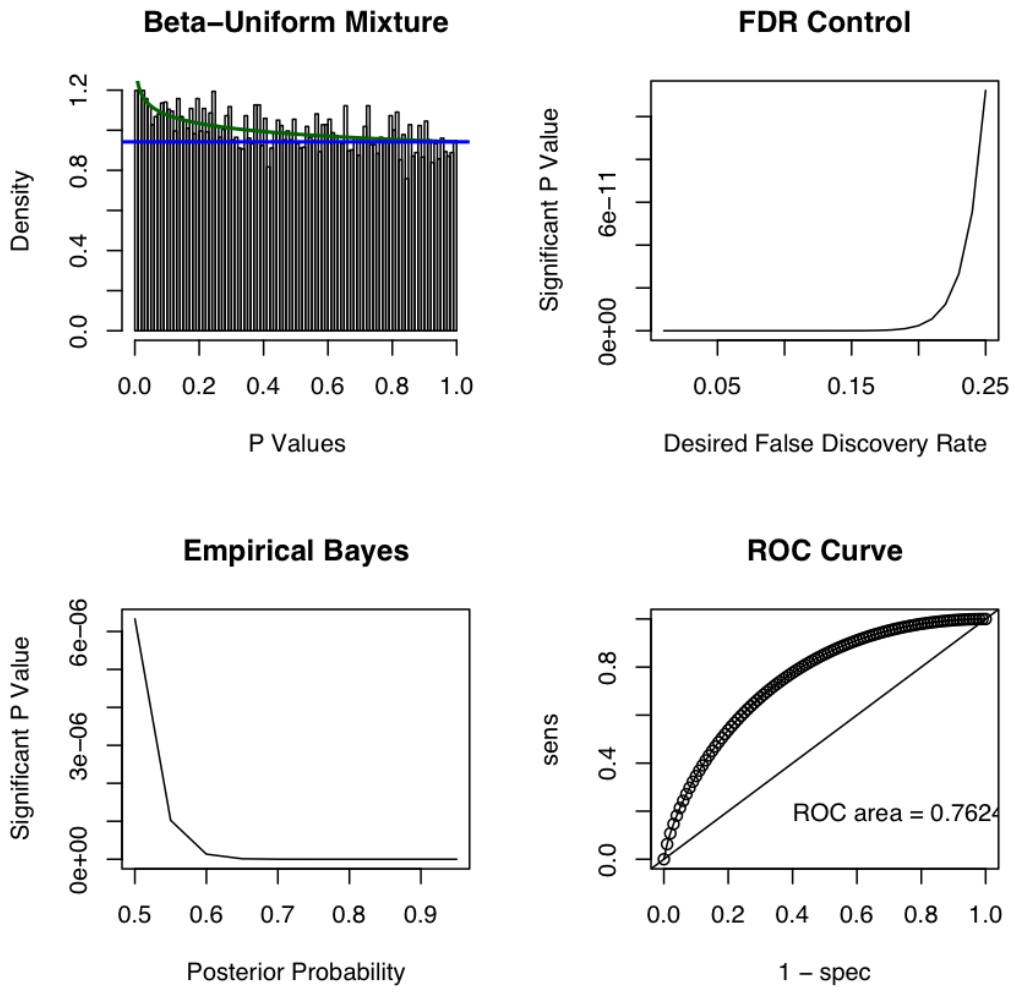
# Is There Division by Date?




Erm, yep.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

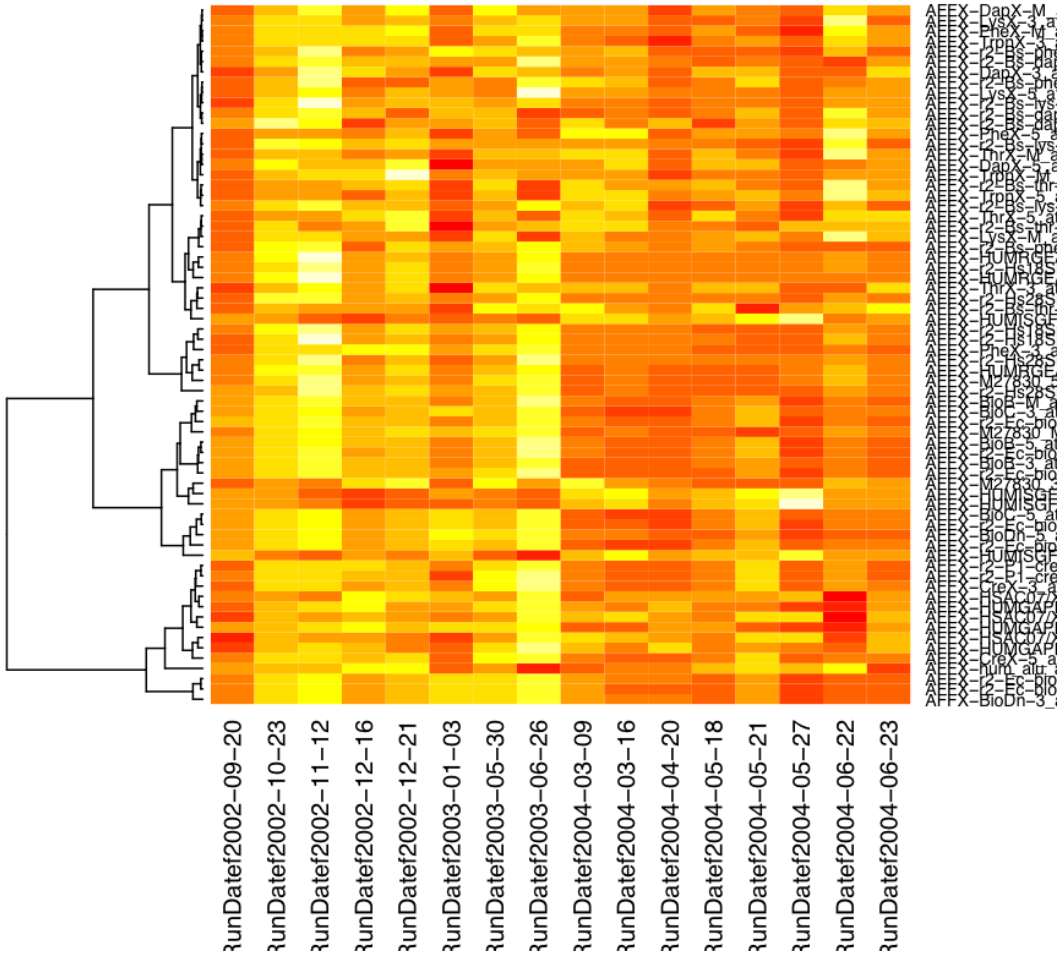
# Division by CR/NR After Date?




Not really.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

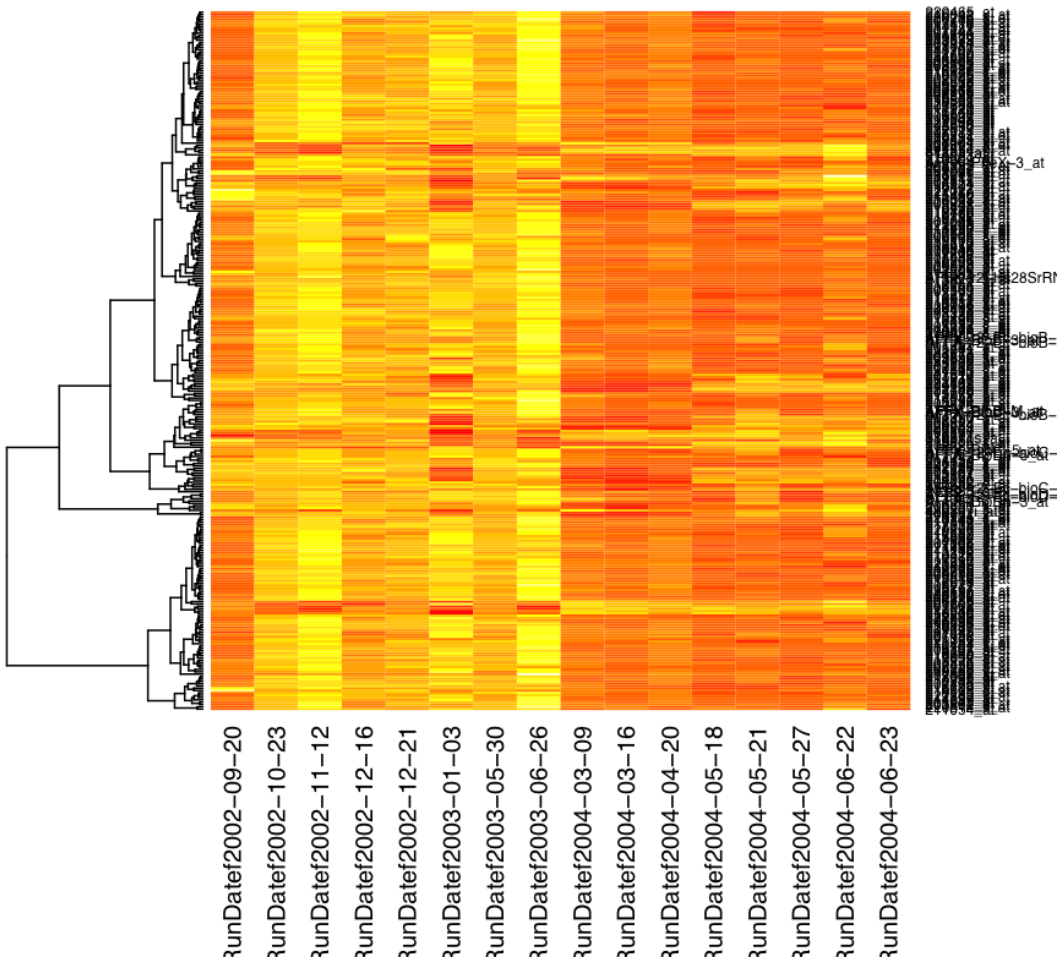
# How Many Batches? (Controls)




Using Affy controls, we see 7 blocks.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

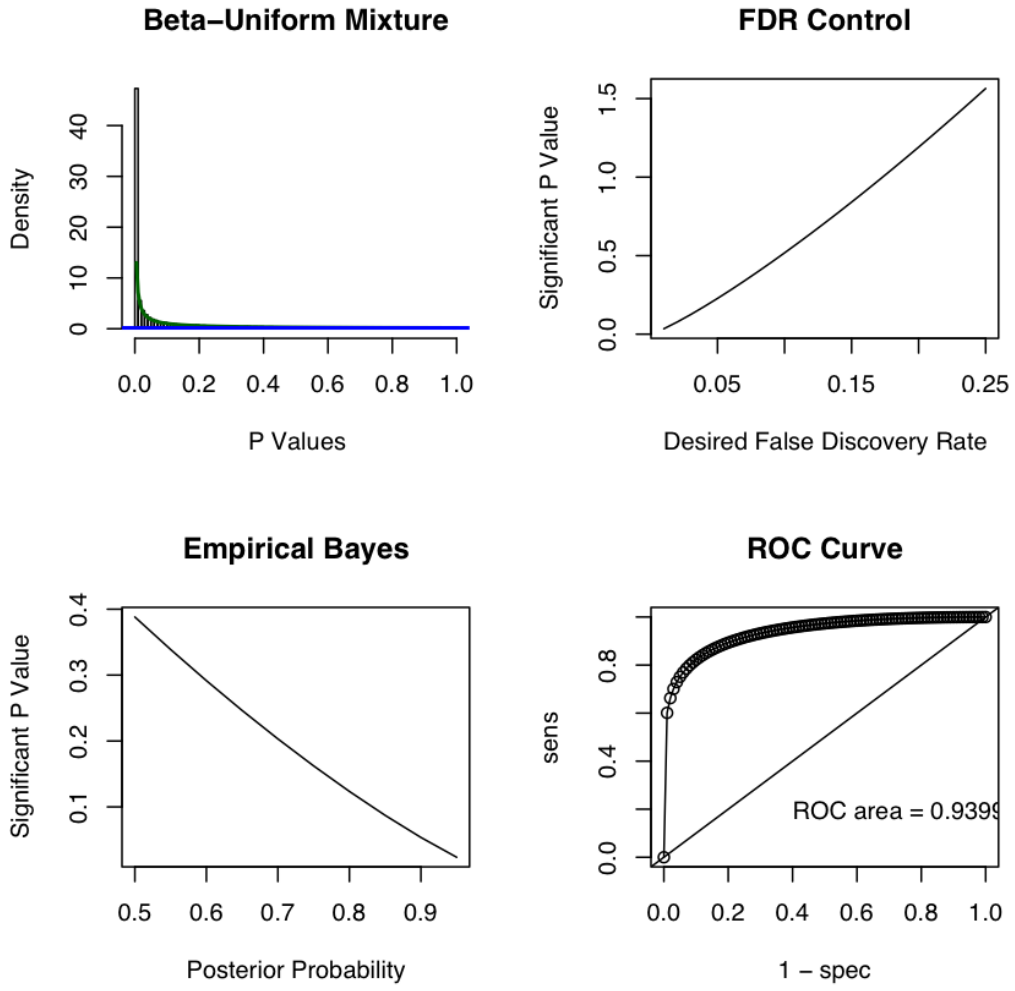
# How Many Batches? (Top ANOVAs)




The smallest ANOVA p-values show the same.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

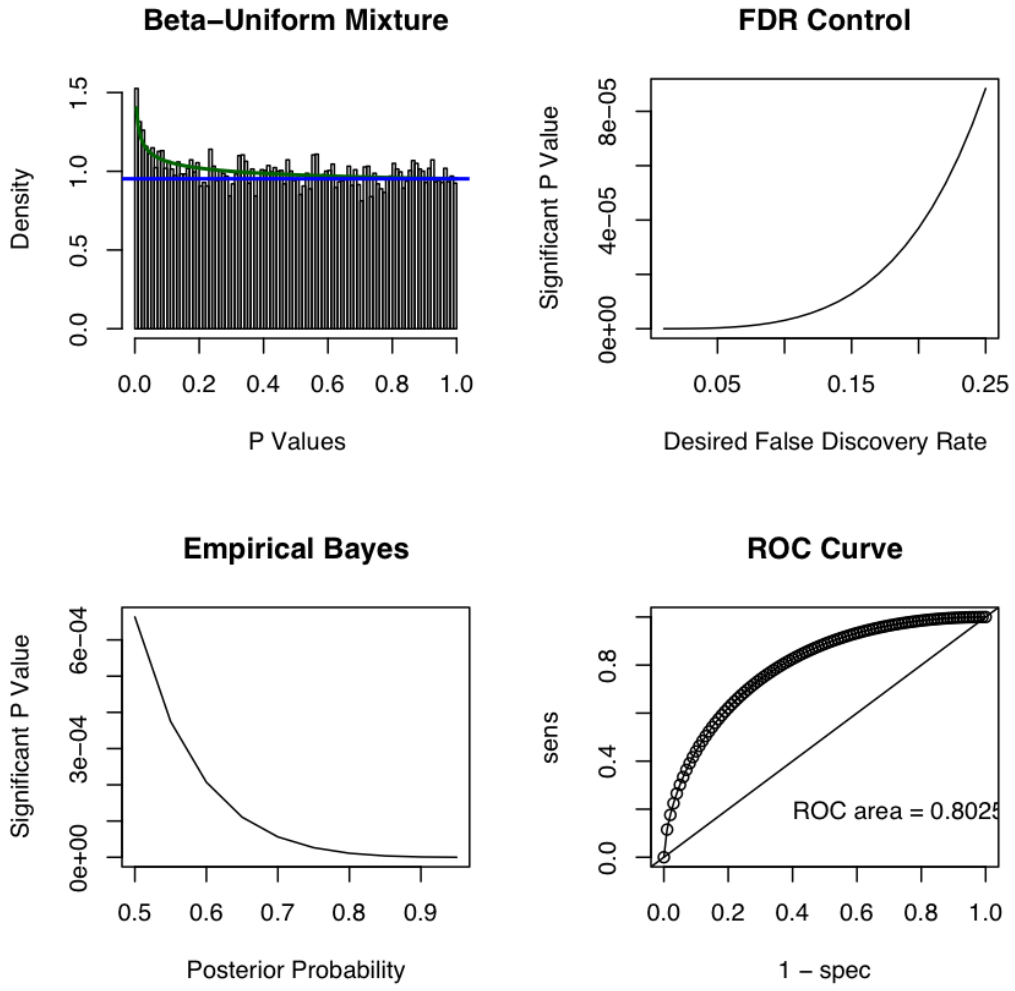
# Is There Division by Batch?




We think we caught most of it.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

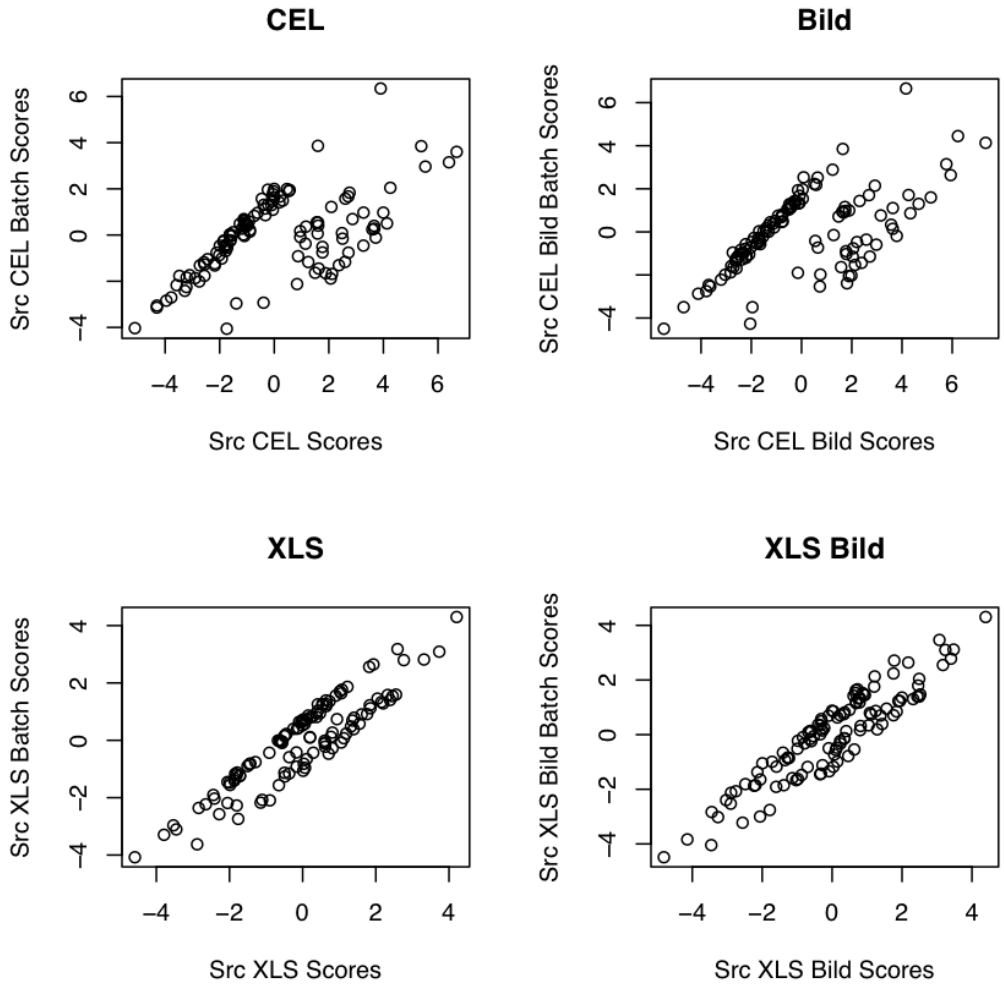
# Division by CR/NR After Batch?




Maybe some, but the FDR isn't small.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

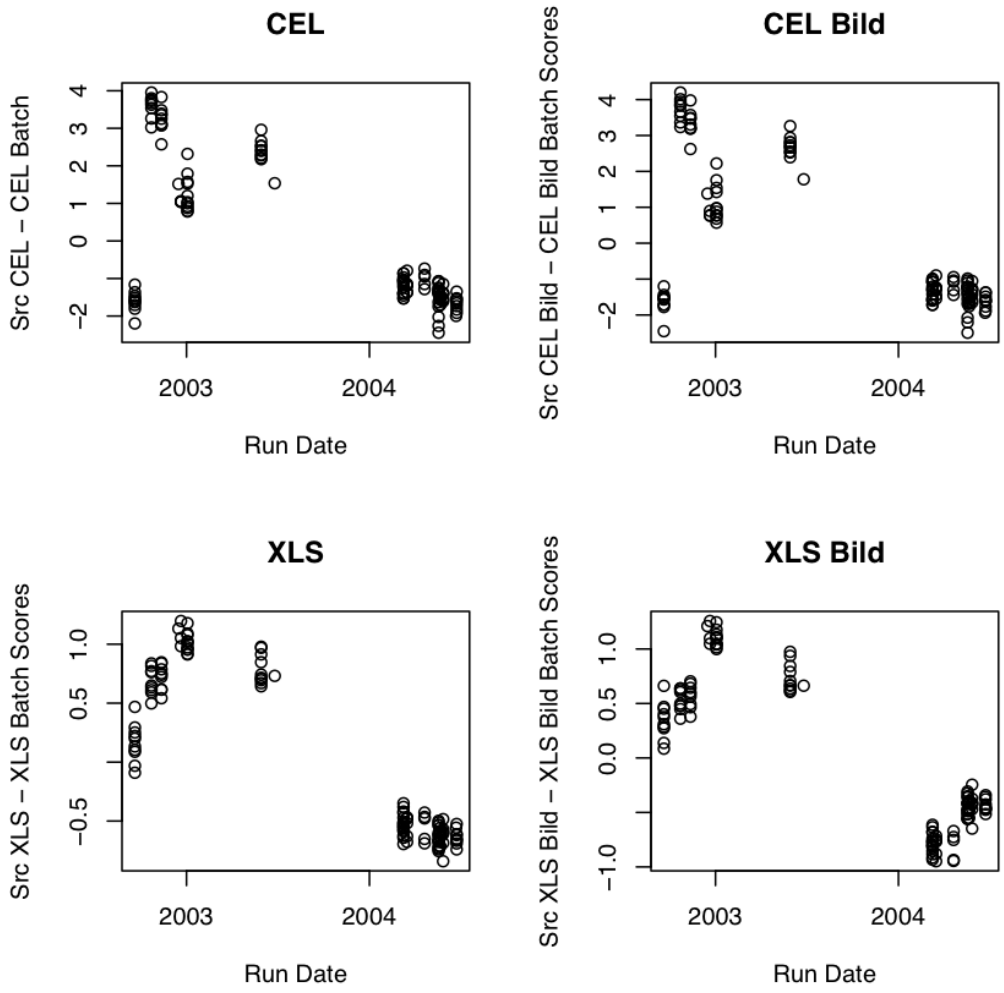
# Batches Affect Scores




Offsets are quite visible

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Early Batch Effects Are Larger




Remember, batches are confounded with Survival!

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# So, how did we get Run Date?

```
> celDatHeaders <- celFiles
> for (i1 in 1:length(celDatHeaders)) {
     temp <- read.celfile.header(file.path("DukeWe
         celFiles[i1]), info = "full")
     celDatHeaders[i1] <- temp$DatHeader
 }
```

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Eyeballing DatHeader Lines

```
> celDatHeaders[1]
872
 "[0..37764]  0074_1772_H133A_872:CLS=4733
  RWS=4733 XIN=3  YIN=3  VE=17        2.0
  09/20/02 11:43:50    \024  \024 HG-U133A.1sq
  \024  \024  \024  \024  \024  \024  \024  \024
  \024 6"
celDatHeaders[119]
M810
 "[0..29187]  Robyn 810 CR-2:CLS=5391 RWS=5391
  XIN=2  YIN=2  VE=30        2.0 04/20/04
  11:29:32 50101330  M10    \024  \024 HG-U133A.1s
  \024  \024  \024  \024  \024  \024  \024  \024
  \024 6"
```

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Extracting Dates

```
> tempDate1 <- strsplit(celDatHeaders, "2\\.0 ")
> tempDate2 <- unlist(lapply(tempDate1, function(x
    x[2]
 }))
> tempDate3 <- unlist(lapply(tempDate2, function(x
    substr(x, 1, 17)
 }))
> tempDate3[1]
                872
"09/20/02 11:43:50"
```

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Formatting Dates

```
> celRunDate <- as.Date(tempDate3,
    format = "%m/%d/%y %H:%M:%S")
> names(celRunDate) <- names(tempDate3)
> celRunDate <- celRunDate[rownames(clinicalInfo)]
> celRunDate[1:3]
      0.08          860          872
"2002-10-23" "2002-10-23" "2002-09-20"
```

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Tabulating Dates

```
> table(celRunDate)
celRunDate
2002-09-20 2002-10-23 2002-11-12 2002-12-16 2002-1
        10          9          9          1
2003-06-26 2004-03-09 2004-03-16 2004-04-20 2004-0
         1         16          6          5
2004-06-22 2004-06-23
         1          8
```

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Using Dates

```
> names(celRunDate[celRunDate < "2004-03-09"])
 [1] "0.08" "860"  "872"  "922" ...
[51] "3102" "3107" "3142" "3249"
> names(celRunDate[celRunDate >= "2004-03-09"])
 [1] "D1805" "D1837" "D1859" "D2098" ...
[64] "M6199" "M810"
> sum(celRunDate < "2004-03-09")
[1] 54
> table(celRunDate, clinicalInfo$Response)
celRunDate   CR NR
  2002-09-20 10  0
  2002-10-23  2  7
  2002-11-12  8  1
  2002-12-16  1  0
```

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# So, how did we Correct for Batch?

```
> runBatch <- rep(6, 119)
> runBatch[celRunDate == "2002-09-20"] <- 1
> runBatch[celRunDate == "2002-10-23"] <- 2
> runBatch[celRunDate == "2002-11-12"] <- 2
> runBatch[celRunDate == "2002-12-16"] <- 3
> runBatch[celRunDate == "2002-12-21"] <- 3
> runBatch[celRunDate == "2003-01-03"] <- 3
> runBatch[celRunDate == "2003-05-30"] <- 4
> runBatch[celRunDate == "2003-06-26"] <- 4
> runBatch[celRunDate == "2004-03-09"] <- 5
> runBatch[celRunDate == "2004-03-16"] <- 5
> runBatch[celRunDate == "2004-04-20"] <- 5
> runBatch[celRunDate == "2004-05-18"] <- 6
> runBatch[celRunDate == "2004-05-21"] <- 6
```

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

```
> runBatch[celRunDate == "2004-05-27"] <- 6
> runBatch[celRunDate == "2004-06-22"] <- 6
> runBatch[celRunDate == "2004-06-23"] <- 6
> runBatch <- as.factor(runBatch)
> names(runBatch) <- names(celRunDate)
```

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Modeling Residuals

Fit gene expression values as a function of runBatch, then fit residuals after correcting for batch.

```
> batchModelForm <- Y ~ runBatch
> batchModelLMAll <-
    MultiLinearModel(batchModelForm,
    data.frame(runBatch = runBatch),ovcaRMAFromCEL
> ovcaRMAFromCELResids <- ovcaRMAFromCEL -
     t(batchModelLMAll@predictions)
> responseModelForm <- Y ~ Response
> responseModelLMAll <-
     MultiLinearModel(responseModelForm,
     data.frame(Response = clinicalInfo$Response),
        ovcaRMAFromCELResids)
```

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Other Approaches

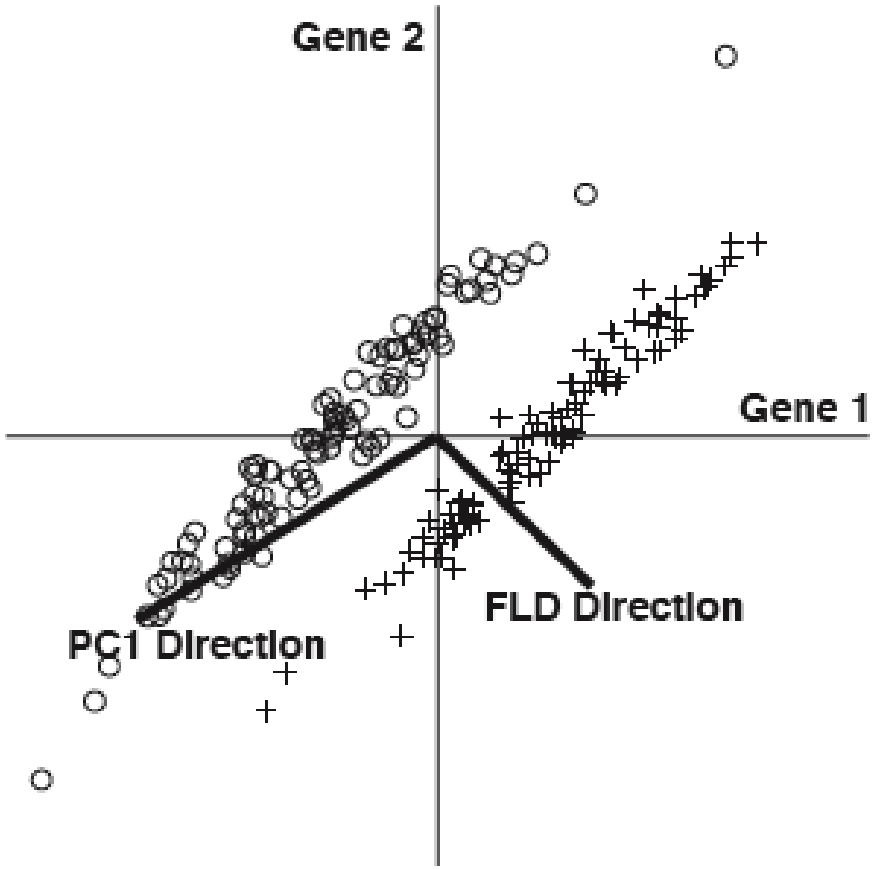
Principal Components/Singular Value Decompostion

Excellent for getting a high-level view of the structure in your data, and seeing if there are outliers or outlying clusters.

Hence, useful for revealing batches, but not so useful for removing them – it doesn't use labels at all.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

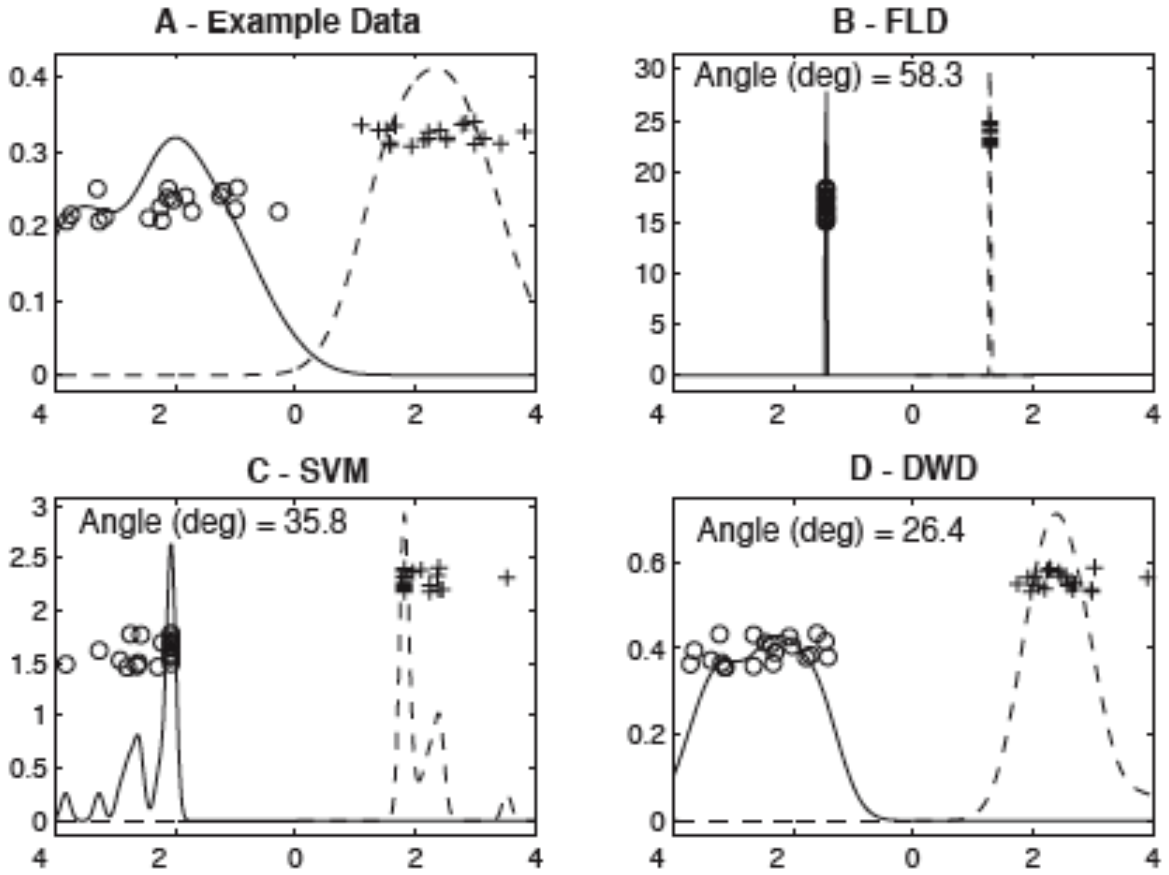
# Where SVD Breaks




Benito et al, Bioinf 2004, fig 1. The batch effect isn't clearly dominant. LDA is ok here.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Where LDA Breaks




Benito et al, Bioinf 2004, fig 2. Low-d structure in high-d data. LDA sidetracked.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Distance Weighted Discrimination (DWD)

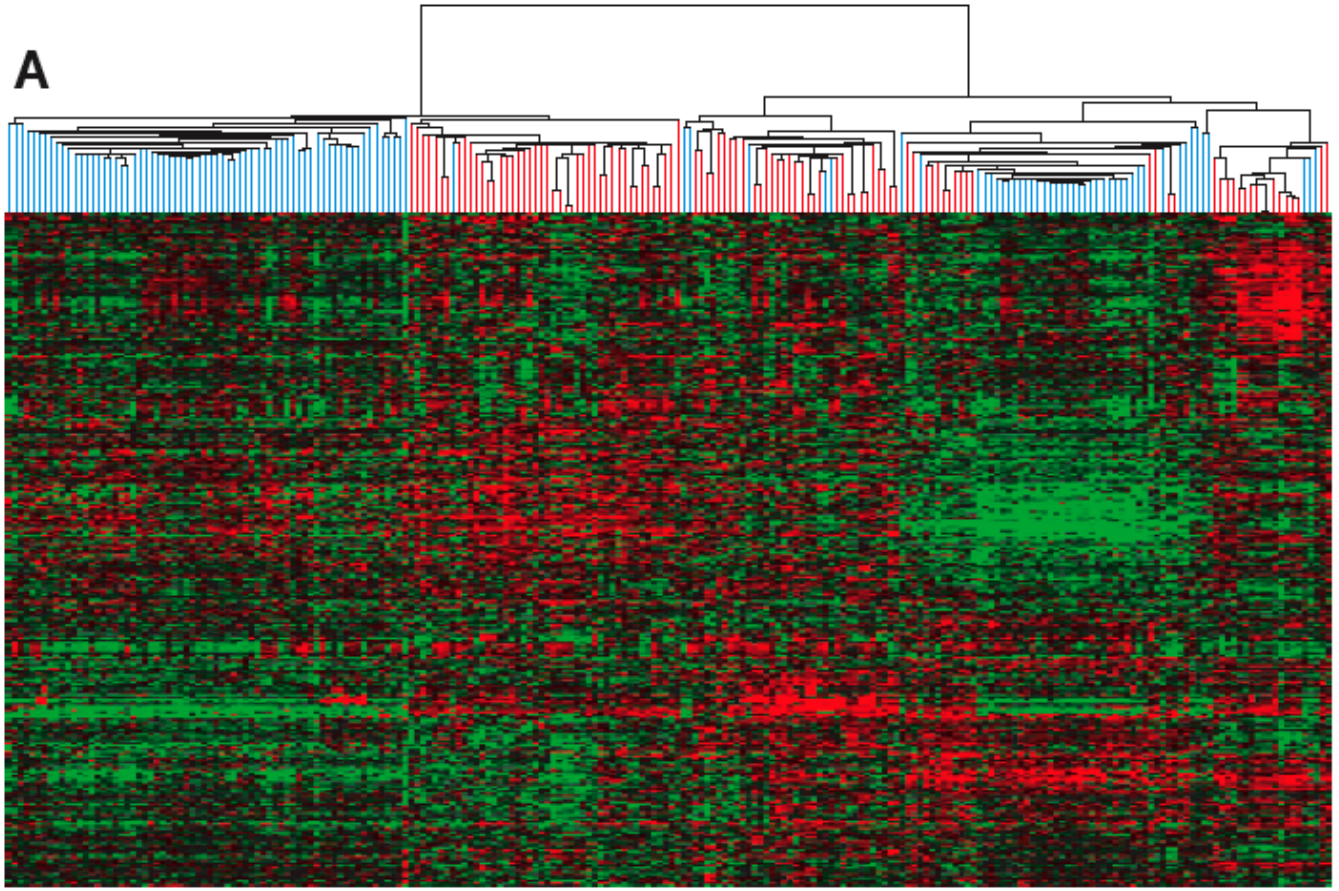
Similar to a Support Vector Machine (SVM) approach, but not solely driven by extrema (maximizing the minimum distance). Instead, maximize the sum of the inverse distances — lets all points contribute, but still stays robust to outliers.

Matlab implementation code exists, involves some sophisticated optimization algorithms.

Focuses on two groups, works better if groups are moderately large.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

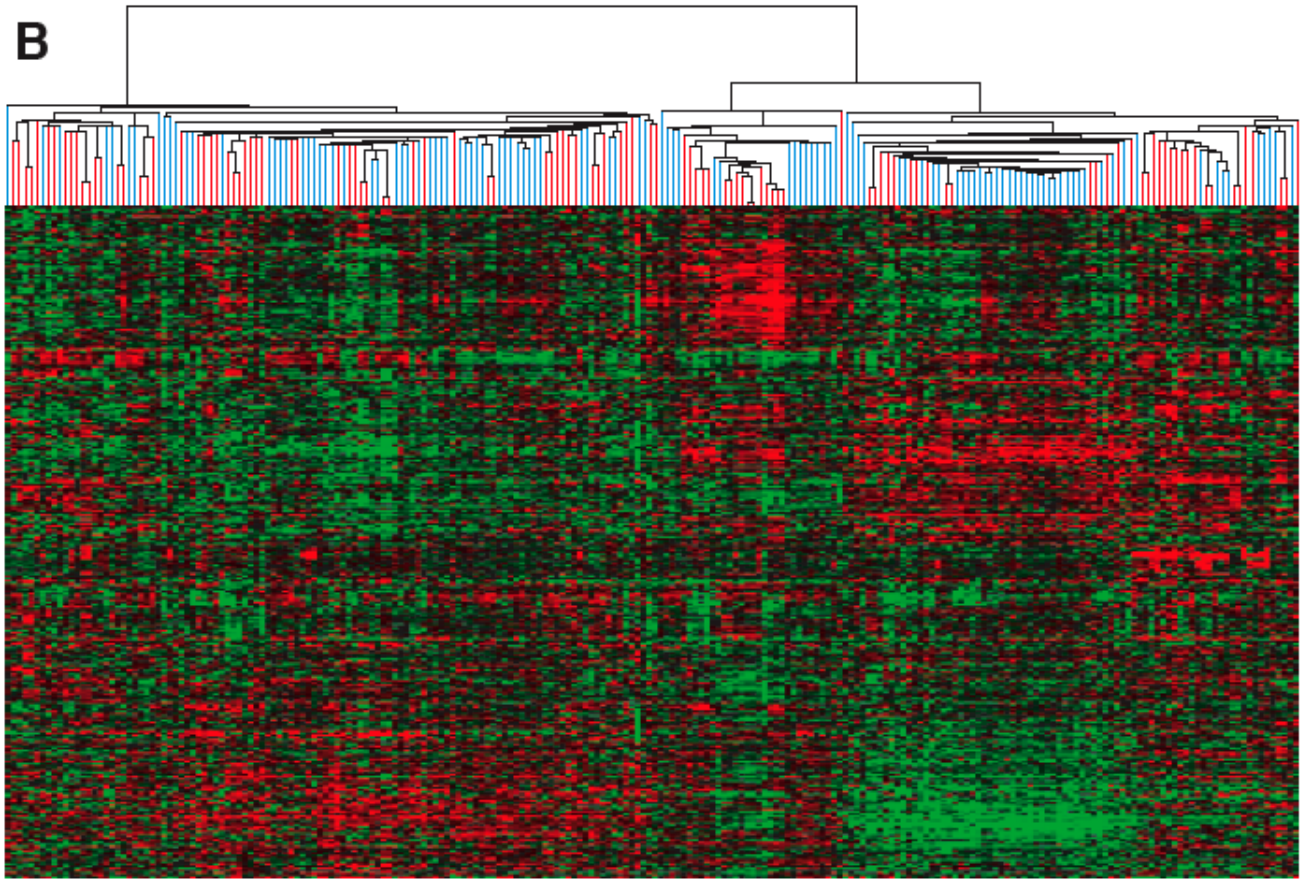
# Real Data pre DWD




Benito et al, Bioinf 2004, fig 6a. Datasets from two platforms (logratio), each median centered.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Real Data post DWD




Benito et al, Bioinf 2004, fig 6b. Datasets after DWD adjustment. More coherent.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# An Alternative – COMBAT

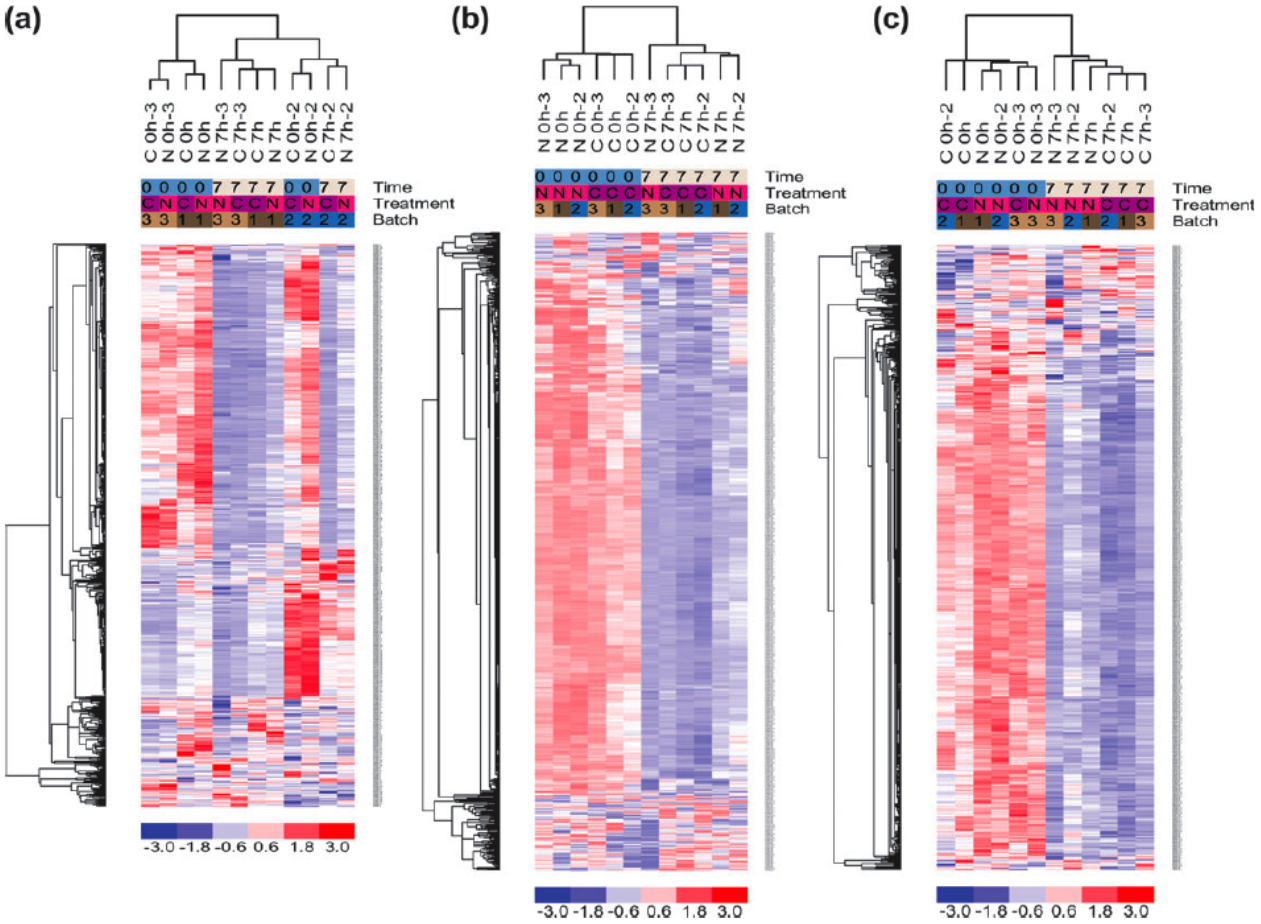
Starts with an adjustment similar to what we tried in the first place with the linear models, with two shifts.

First, it allows for shifts in scale as well as center.

Second, it borrows strength across genes and shrinks towards the central location and scale adjustments.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

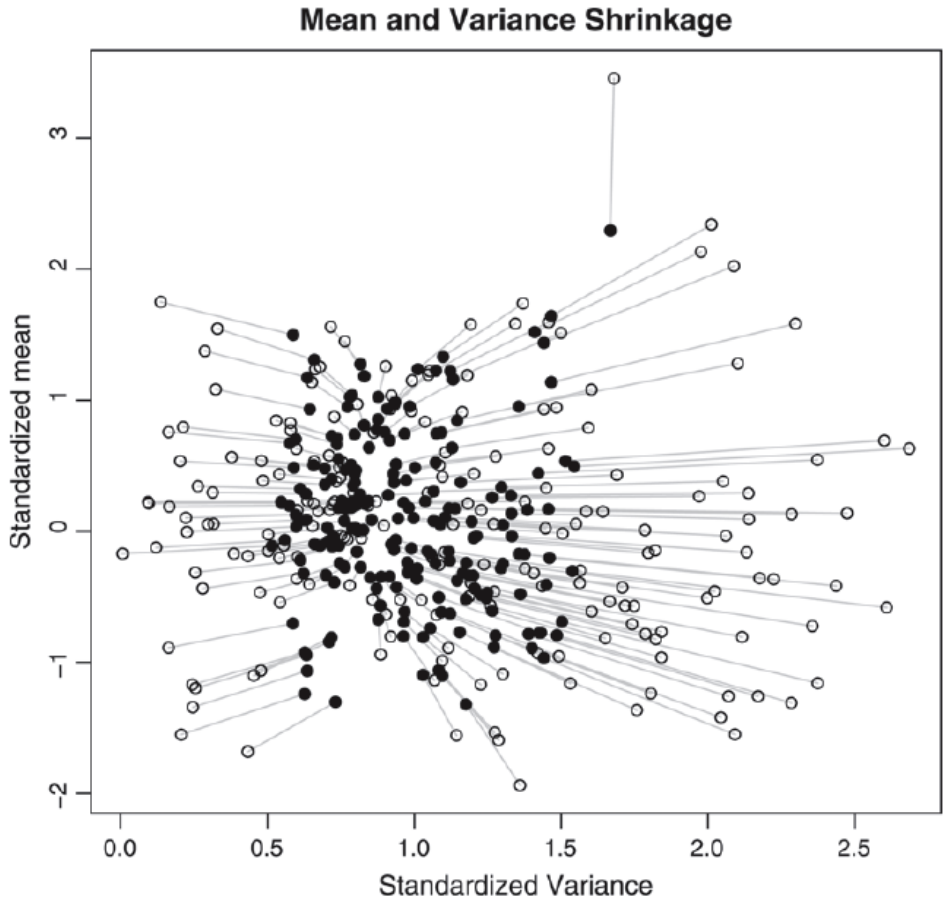
# A COMBATive Triptych




Bad, Good, Good

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

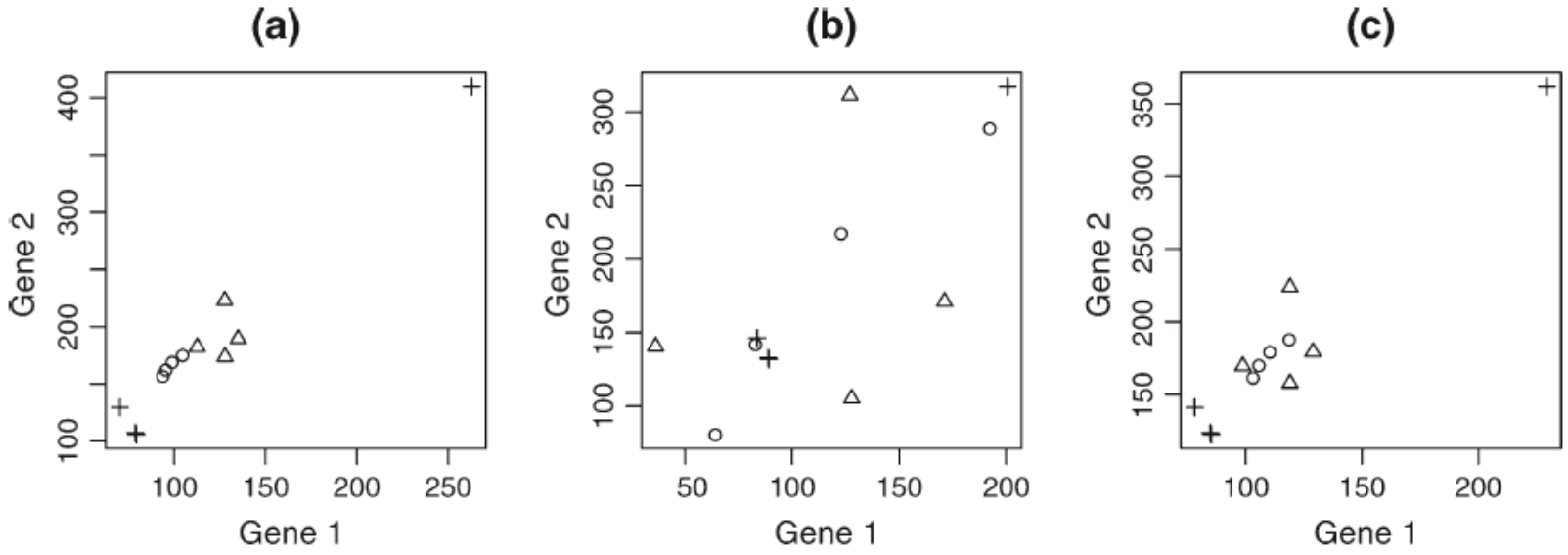
# A Shrinking Black Hole




Which values are “better”? Why?

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# The Value of Shrinkage




Borrowing can let you recognize (and downweight) outliers.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Implementing COMBAT

This is actually pretty easy, since a set of R scripts is available from Evan Johnson's web site at BYU:

`http://statistics.byu.edu/johnson/ComBat`

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Review: Leek et al

How pervasive are batch effects?

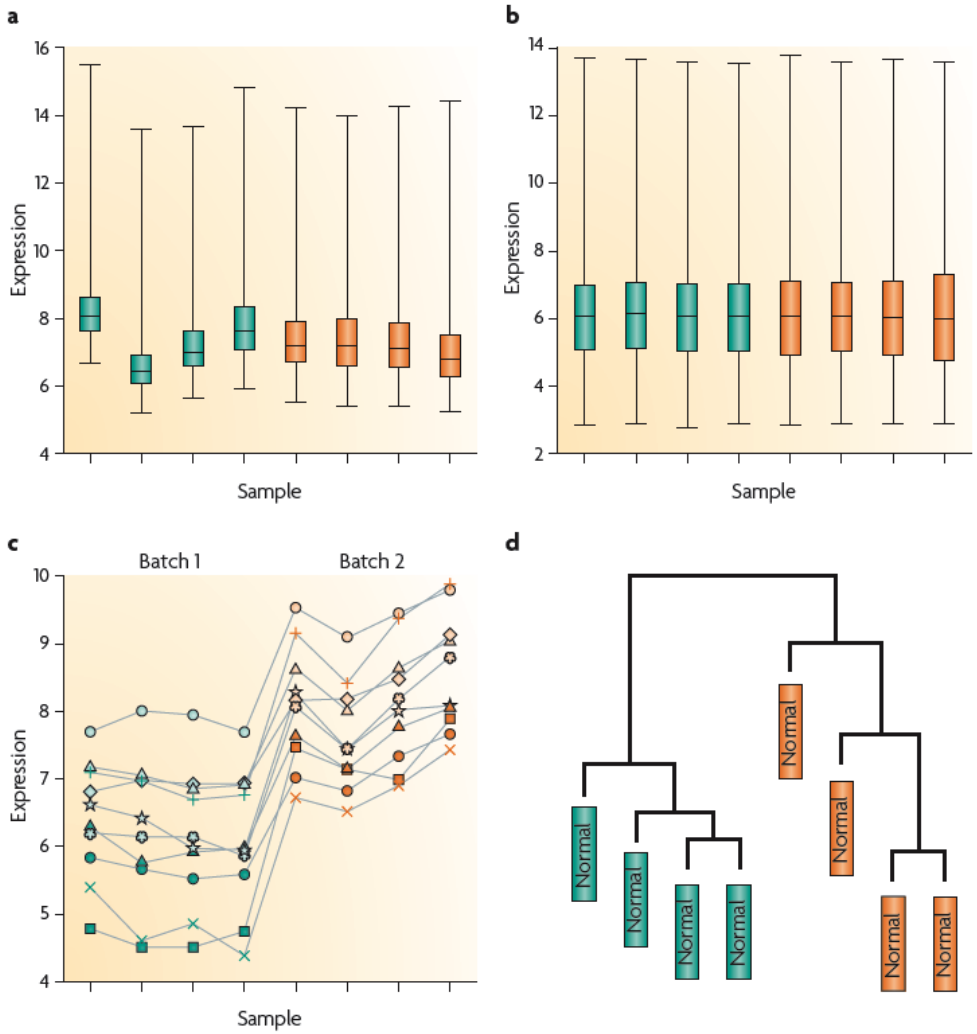
How big are they?

How can we fix them?

When can we fix them?

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Will Normalization Fix Things?




Leek et al Fig 1: Batches can survive normalization.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Are New Assays Immune?

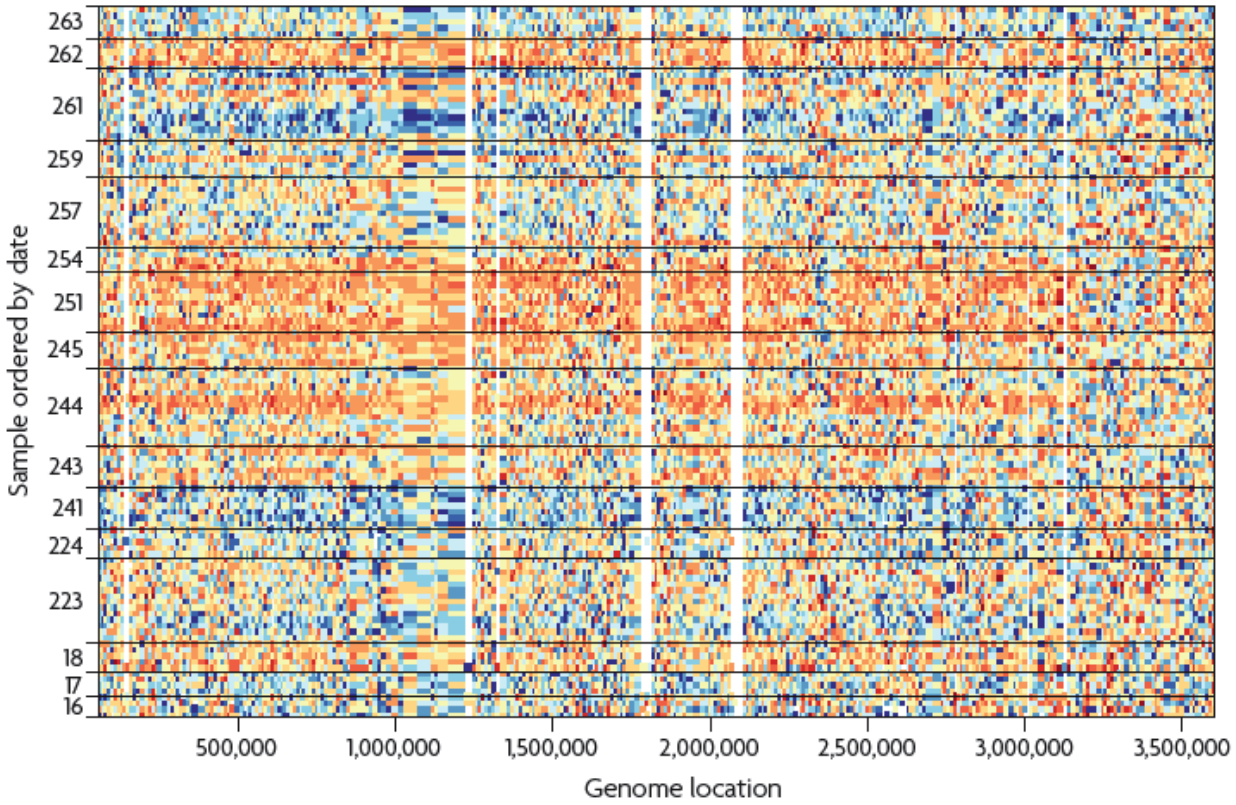



Figure 2 | **Batch effects for second-generation sequencing data from the 1000 Genomes Project.** Each row is a different HapMap sample processed in the same facility with the same platform. See Supplementary information S1 (box) for a description of the data represented here. The

Leek et al Fig 2: Next gen – sd +/- in orange/blue.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Are Pathways Immune?

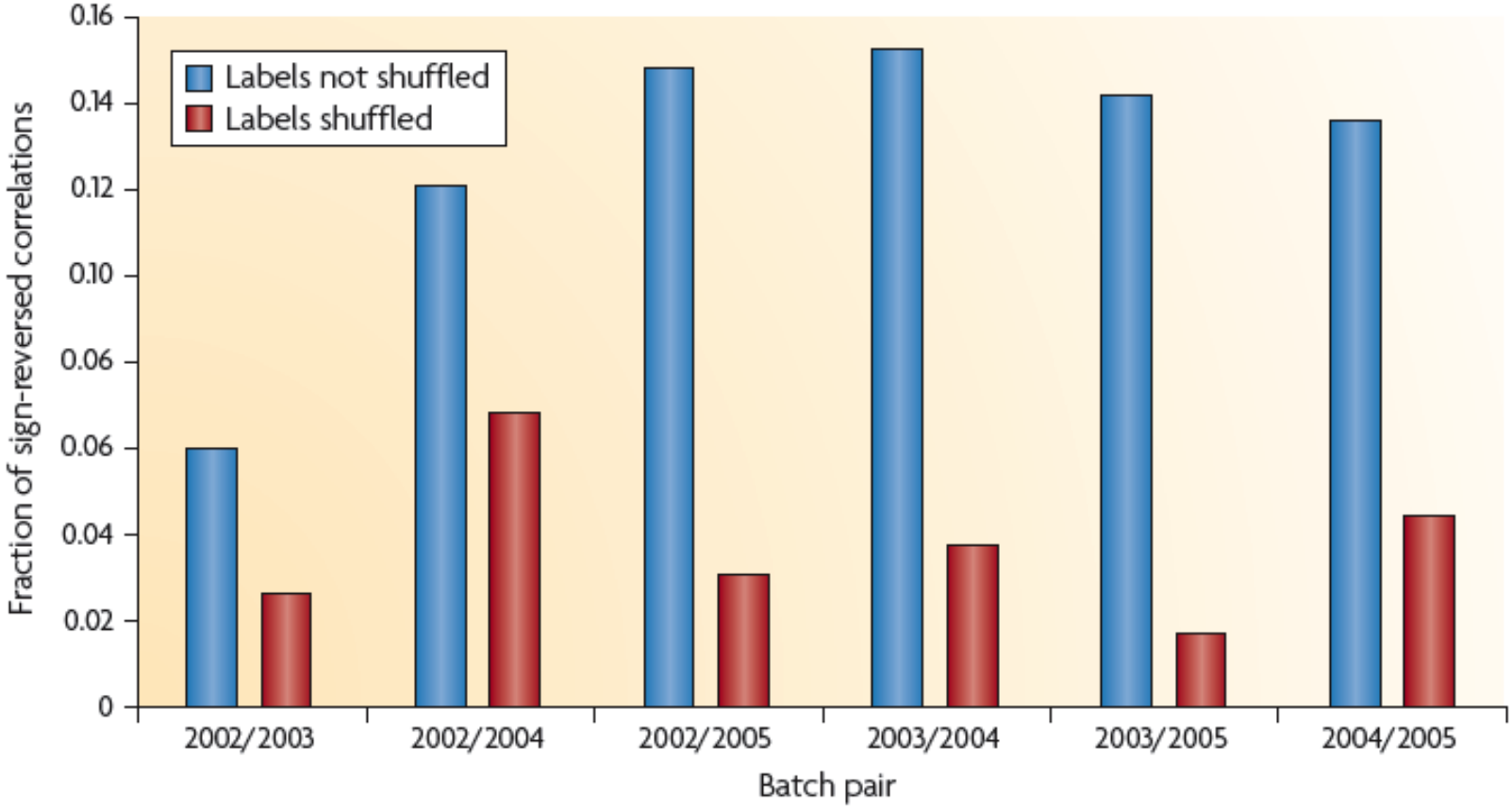



Figure 3 | **Batch effects also change the correlations between genes.** We normalized every gene

Leek et al Fig 3: Correlations can change sign

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# What Was Examined?

Table 1 | **Batch effects seen for a range of high-throughput technologies**

| Study description* | Known variable used as a surrogate | | | Principal components used as a surrogate | | | Association with outcome | Refs |
|---|---|---|---|---|---|---|---|---|
| | 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 | 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 | 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 | 3 |
| 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 | 16 |
| 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 | 17 |
| Data set 6: gene expression, Affymetrix ($N_p$ = 22,277) | Processing group | NA | 83.8 | 5 (0.369) | NA | 97.1 | NA | 18 |
| Data set 7: gene expression, Agilent ($N_p$ = 17,594) | Date | NA | 62.8 | 2 (0.248) | NA | 96.7 | NA | 18 |
| Data set 8: DNA methylation, Agilent ($N_p$ = 27,578) | Processing group | NA | 78.6 | 3 (0.381) | NA | 99.8 | NA | 18 |
| Data set 9: DNA sequencing, Solexa ($N_p$ = 2,886) | Date | 24.2 | 32.1 | 2 (0.846) | 2 (0.213) | 72.7 | 16.9 | 1000 Genomes Project |

Leek et al Table 1: Batches are everywhere

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# How Do We Fix Things?

Design

Linear Models

ComBat

SVA

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# Which do I Prefer?

I like them all, but at present I lean towards either linear model approaches or ComBat.

This is for reasons of relative simplicity and ease of implementation.

In general, I think the choice of a particular method of batch adjustment is an order of magnitude less important than recognizing that batches may be present in the first place, and coming up with reasonable ideas as to what they are.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# TCGA Batches

The TCGA samples are processed in batches.

This is largely by necessity, because not all of the samples are coming into the processing center at once, and they don't want to wait for all 500 before starting.

Batches typically involve about 30 samples.

The batches have not been assembled with balance in mind.

For some batches, mild correction will likely be necessary.

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

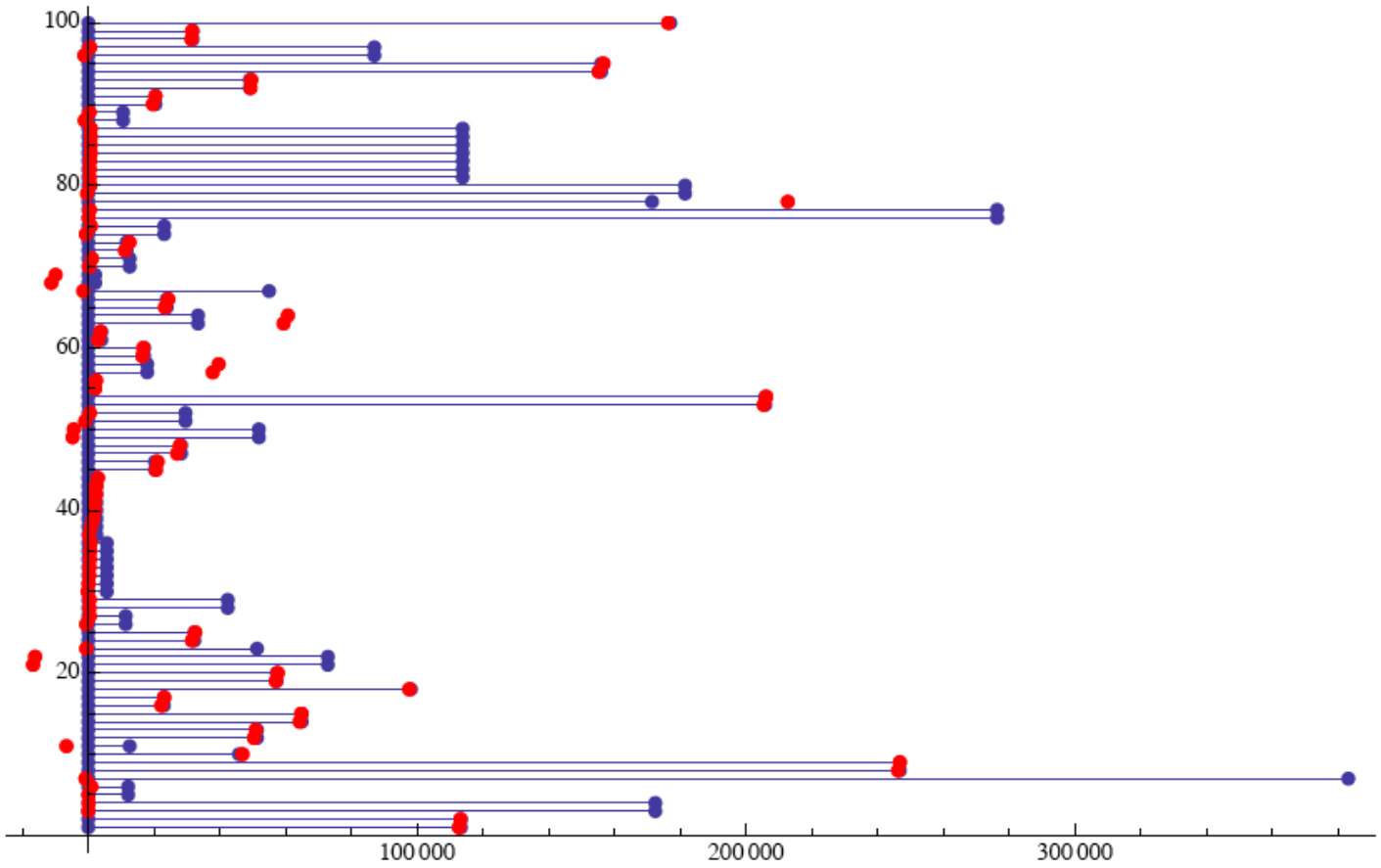
# TCGA Genes

"GBM Phase I + II.xls" (from the TCGA web site)

| Phase | Name | Chr | Start | Stop | locusLinkId | omimId | proteinAcc |
|---|---|---|---|---|---|---|---|
| Phase1 | ABI1 | chr10 | 27077005 | 27190298 | 10006 | 603050 | NP_005461 |
| Phase1 | ABL1 | chr9 | 132579028 | 132751391 | 25 | 189980 | NP_009297 |
| Phase1 | ADAM15 | chr1 | 153289987 | 153302102 | 8751 | 605548 | NP_003806 |

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

# TCGA Positions?




Shouldn't they line up?

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom