1 Introduction

The interchange between Baggerly, Coombes and Neeley (BCN) and Dressman, Potti and Nevins (DPN) in XXX is of broad interest to cancer bioinformaticians. We are indebted to Dressman and colleagues for voluntarily making available much of the data and analytic resources underlying the "Integrated genomic-based approach" paper of 2007. We are indebted to Baggerly and colleagues for their meticulous approach to re-assessing the paper in conjunction with the data.

To review, BCN use the archive at `http://data.cgt.duke.edu/platinum.php` along with the GEO submission of Bild et al. (XXX, GSE3156) to raise a number of concerns regarding the main paper on targeting treatment of ovarian cancer. I select those that are simplest to illustrate and interpret.

2 Identifier scrambling

First, the "corrected RMA" expression quantifications

`https://discovery.genome.duke.edu/express/resources/193/correctedplatinum_RMA.xls`,

retrieved online May 8 2009, appear to have **mislabeled columns**.

This was apparently discovered through an attempt to understand the difference between RMA preprocessed arrays, readily generated using the CEL files in

`https://discovery.genome.duke.edu/express/resources/1144/PlatinumJCO.zip`,

and the "corrected RMA" data, which consists of RMA quantifications that are additionally processed by "sparse factor regression" to remove artifacts.

Figure 1 shows how the mislabeling can be discovered and corrected; full details are provided in the original supplements of BCN's letter, but using a Bioconductor experimental data package, *dressCheck*, we can illustrate the problem readily.

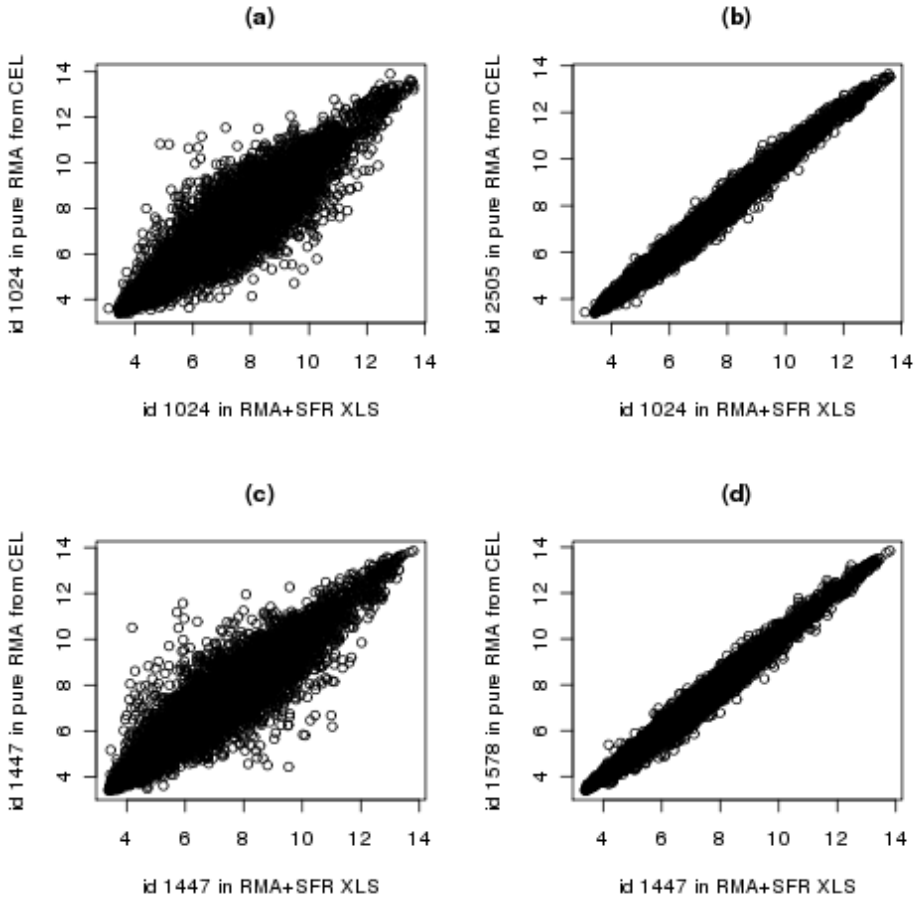
```
> library(dressCheck)
> if (!exists("c119")) data(c119) # pure RMA on CEL files, with trimming of sample na
> if (!exists("DrAsGiven")) data(DrAsGiven) # read of corrected platinum XLS to CSV
>     # some names are not preserved between two images
> setdiff(sampleNames(c119), sampleNames(DrAsGiven))

[1] ".08"  "3250"

>     # use the common ones
> okn = intersect(sampleNames(DrAsGiven), sampleNames(c119))
>     # the corrected XLS does not have all genes, so c119 needs trim
> c119r = c119[ featureNames(DrAsGiven), ]
>     # now demonstrate
> allc1 = sapply(okn, function(i) cor(exprs(DrAsGiven)[,okn[1]],
```

```
+    exprs(c119r[,i])))
> png(file="corChk.png")
> par(mfrow=c(2,2))
> remap1 = which.max(allc1)
> plot(exprs(DrAsGiven)[,okn[1]], exprs(c119r)[,okn[1]],
+  xlab=paste("id", okn[1], "in RMA+SFR XLS"),
+  ylab=paste("id", okn[1], "in pure RMA from CEL"),
+  main="(a)")
> plot(exprs(DrAsGiven)[,okn[1]], exprs(c119r)[,names(remap1)],
+  xlab=paste("id", okn[1], "in RMA+SFR XLS"),
+  ylab=paste("id", names(remap1), "in pure RMA from CEL"),
+  main="(b)")
> allc2 = sapply(okn, function(i) cor(exprs(DrAsGiven)[,okn[2]],
+    exprs(c119r[,i])))
> remap2 = which.max(allc2)
> plot(exprs(DrAsGiven)[,okn[2]], exprs(c119r)[,okn[2]],
+  xlab=paste("id", okn[2], "in RMA+SFR XLS"),
+  ylab=paste("id", okn[2], "in pure RMA from CEL"),
+  main="(c)")
> plot(exprs(DrAsGiven)[,okn[2]], exprs(c119r)[,names(remap2)],
+  xlab=paste("id", okn[2], "in RMA+SFR XLS"),
+  ylab=paste("id", names(remap2), "in pure RMA from CEL"),
+  main="(d)")
> dev.off()

pdf
  2
```



Panels (a) and (c) show the kind of relationship expected when two arrays on different samples are compared. Panels (b) and (d) shows the kind of relationship expected when slightly different preprocessing method are applied to the same sample. BCN did this exercise very systematically, concluding that the 'corrected RMA' arrays 1024 and 1447 in the XLS matche the CEL files (and thus the clinical data series for) 2505 and 1578 respectively. In their response, DPN acknowledge the identifier scrambling and remark that mislabeling only affected the publication of data on the web site, not the analyses underlying the paper.

The unscrambling using correlation to CEL succeeds for 116/119 samples. The unscrambled data are in ExpressionSet `corrp116`.

3 Persistence of batch effects through sparse factor regression corrections

Once the corrected quantifications are properly relabeled they can be associated with the clinical data appropriately. They can also be associated with processing information

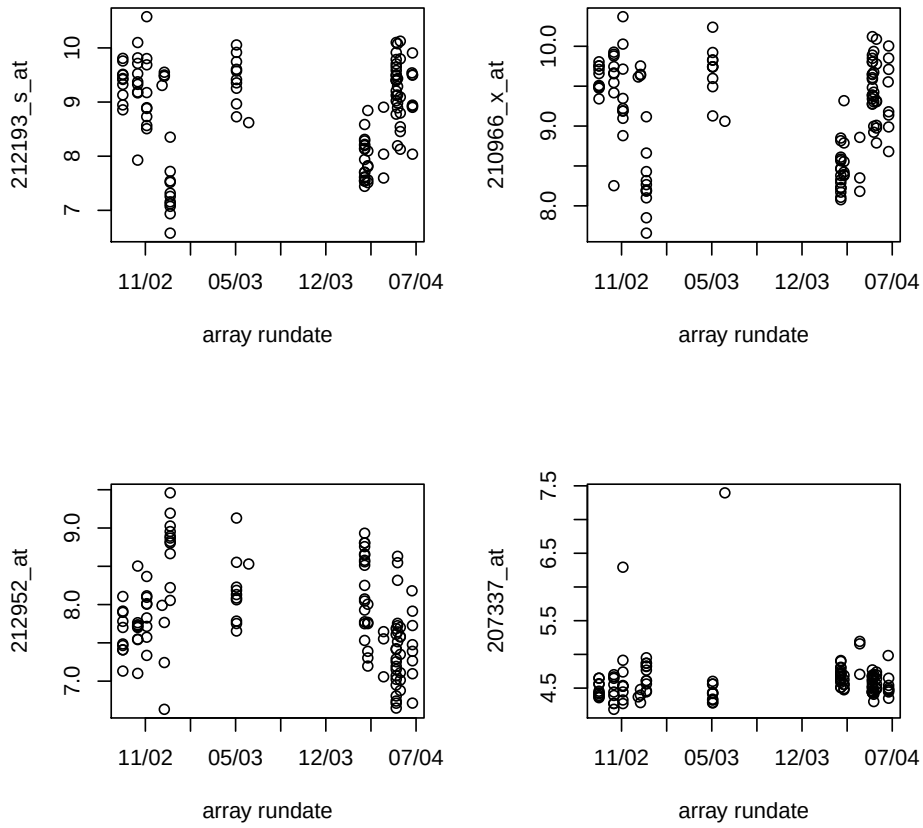
in the CEL files. The corrp116 ExpressionSet includes this information.

```
> library(chron)
> if (!exists("corrp116")) data(corrp116)
> dt = table(chron(corrp116$rundate))
> cdt = chron(as.numeric(names(dt)))
> names(dt) = cdt
> dt

<NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
  10    9    9    1    3   11   10    1   16    6    3   15    7    7    1    7
```

We will test for differential expression by batch.

```
> library(limma)
> corrp116$cdate = factor(chron(corrp116$rundate))
> des = model.matrix(~cdate, pData(corrp116))
> if (!exists("f1")) f1 = lmFit(corrp116, des)
> ef1 = eBayes(f1)
> options(digits = 4)
> tt = topTable(ef1, 2:16)[, -c(2:16)]
> bigtt = topTable(ef1, 2:16, n = 1000)[, -c(2:16)]
> mm = max(bigtt[, 5])
> tops = tt[, 1]
> par(mfrow = c(2, 2))
> plot(chron(corrp116$rundate), exprs(corrp116)[tops[1], ], xlab = "array rundate",
+     ylab = tops[1])
> plot(chron(corrp116$rundate), exprs(corrp116)[tops[2], ], xlab = "array rundate",
+     ylab = tops[2])
> plot(chron(corrp116$rundate), exprs(corrp116)[tops[3], ], xlab = "array rundate",
+     ylab = tops[3])
> plot(chron(corrp116$rundate), exprs(corrp116)[tops[4], ], xlab = "array rundate",
+     ylab = tops[4])
```



The maximum adjusted p-value (BH FDR) in the top 1000 genes is 0.00133, so there is statistical evidence for an association between mean expression and run date for many genes.

The display shows that this association can take various forms. For the top two panels, there is indication of a downward trend for all samples in late 2002, early 2003. The bottom left panel suggests an increasing trend in mean in the same period, and the bottom right panel shows that outliers can be identified in tests of batch effects.

Baggerly et al show that there is an association between survival time of sample and run date (and I do so as well in the short letter).

The basic upshot of this section: It is probably incorrect to dismiss the possibility of confounding of expression-survival associations with run batch. We don't know how to adjust for this in a way that is completely reliable. If we take the standard epidemiological approach of introducing a factor for batch in a survival regression model, we may be over-adjusting. But the retort in the DPN rebuttal to BCN, asserting that batch effects cannot be present because sparse factor regression corrects for them, is incorrect.

4 Non-reconstructibility of E2F3-survival relationship

The short letter is pretty clear on this.