Extracting Chip Run Dates

Keith A. Baggerly, Shannon Neeley and Kevin R. Coombes

October 9, 2007

1 Introduction

We have often found it worthwhile to look at chip run dates, as we have encountered batch effects in a variety of forms. Run date is a surrogate for many potential sources of batch-type variation. Here, we extract this information from the CEL files and check for potential problems.

2 Options and Libraries

```
> options(width = 80)

> library(affy)

> load(file.path("RDataObjects", "celFiles.Rda"))
> load(file.path("RDataObjects", "clinicalInfo.Rda"))
```

3 Grabbing the Run Dates

First, we extract the DatHeader lines.

```
> celDatHeaders <- celFiles
> for (i1 in 1:length(celDatHeaders)) {
+     temp <- read.celfile.header(file.path("DukeWebSite", "PlatinumJCO",
+         celFiles[i1]), info = "full")
+     celDatHeaders[i1] <- temp$DatHeader
+ }
> celDatHeaders[1:3]

 "[0..37764]  0074_1772_H133A_872:CLS=4733 RWS=4733 XIN=3  YIN=3  VE=17        2.0 09/20/02 11:43:50

 "[0..33251]  0074_1773_H133A_922:CLS=4733 RWS=4733 XIN=3  YIN=3  VE=17        2.0 09/20/02 11:54:47

"[0..43335]  0074_1774_H133A_1451:CLS=4733 RWS=4733 XIN=3  YIN=3  VE=17        2.0 09/20/02 12:06:27

> celDatHeaders[117:119]
```

```
    "[13..28971]  Robyn5775CR:CLS=5392 RWS=5392 XIN=2  YIN=2  VE=30        2.0 03/09/04 13:11:03 5010133

"[0..54480]  Robyn 16199 CR-2:CLS=5391 RWS=5391 XIN=2  YIN=2  VE=30        2.0 04/20/04 11:43:40 5010133

  "[0..29187]  Robyn 810 CR-2:CLS=5391 RWS=5391 XIN=2  YIN=2  VE=30        2.0 04/20/04 11:29:32 5010133
```

As we can see, we have a broad date range involved, going at least from 9/20/02 to 4/20/04. We now parse these strings to extract the dates, making use of the centrally placed "2.0".

```
> tempDate1 <- strsplit(celDatHeaders, "2\\.0 ")
> tempDate1[1]

$`872`
[1] "[0..37764]  0074_1772_H133A_872:CLS=4733 RWS=4733 XIN=3  YIN=3  VE=17        "
[2] "09/20/02 11:43:50    \024  \024 HG-U133A.1sq \024  \024  \024  \024  \024  \024  \024  \024  \024 6

> tempDate2 <- unlist(lapply(tempDate1, function(x) {
+     x[2]
+ }))
> tempDate2[1]

                                                                                                    872
"09/20/02 11:43:50    \024  \024 HG-U133A.1sq \024  \024  \024  \024  \024  \024  \024  \024  \024 6"

> tempDate3 <- unlist(lapply(tempDate2, function(x) {
+     substr(x, 1, 17)
+ }))
> tempDate3[1]

               872
"09/20/02 11:43:50"

> celRunDate <- as.Date(tempDate3, format = "%m/%d/%y %H:%M:%S")
> names(celRunDate) <- names(tempDate3)
> celRunDate[1:3]

         872          922         1451
"2002-09-20" "2002-09-20" "2002-09-20"

> celRunDate <- celRunDate[rownames(clinicalInfo)]
> celRunDate[1:3]

        0.08          860          872
"2002-10-23" "2002-10-23" "2002-09-20"

> rm(tempDate1, tempDate2, tempDate3)
```

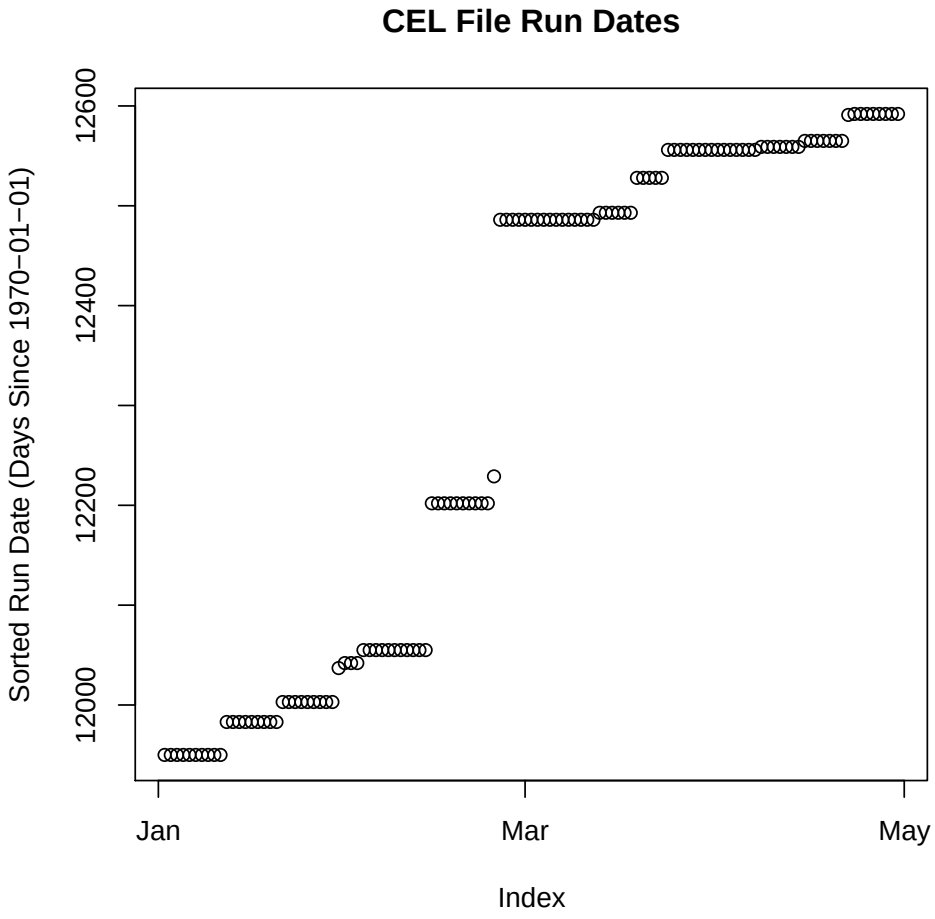
Let's see what the values are, and how they're distributed. First, we tabulate them.

```
> table(celRunDate)
```

```
celRunDate
2002-09-20 2002-10-23 2002-11-12 2002-12-16 2002-12-21 2003-01-03 2003-05-30
        10          9          9          1          3         11         10
2003-06-26 2004-03-09 2004-03-16 2004-04-20 2004-05-18 2004-05-21 2004-05-27
         1         16          6          5         15          7          7
2004-06-22 2004-06-23
         1          8
```

There are 16 distinct run dates, with a few gaps of a month or more and one major gap of more than eight months. Let's plot the distribution to get a better idea.

```
> plot(sort(celRunDate), xlab = "Index", ylab = "Sorted Run Date (Days Since 1970-01-01)",
+     main = "CEL File Run Dates")
```



The major gaps are more visible here.

At this point, the only other information that we are working with is file name. Let's look at the names before and after the divide.

```
> names(celRunDate[celRunDate < "2004-03-09"])

 [1] "0.08" "860"  "872"  "922"  "1024" "1447" "1451" "1504" "1526" "1552"
[11] "1578" "1590" "1615" "1623" "1665" "1674" "1675" "1774" "1784" "1834"
[21] "1846" "1877" "1913" "1929" "2046" "2063" "2064" "2075" "2198" "2204"
[31] "2324" "2419" "2422" "2424" "2465" "2476" "2479" "2505" "2542" "2573"
[41] "2673" "2739" "2802" "2849" "2895" "2967" "2981" "2999" "3018" "3090"
[51] "3102" "3107" "3142" "3249"

> names(celRunDate[celRunDate >= "2004-03-09"])

 [1] "D1805" "D1837" "D1859" "D2098" "D2208" "D2332" "D2342" "D2358" "D2421"
[10] "D2432" "D2433" "D2480" "D2557" "D2559" "D2560" "D2572" "D2575" "D2576"
[19] "D2581" "D2603" "D2611" "D2629" "D2640" "D2648" "D2668" "D2689" "D2691"
[28] "D2700" "D2726" "D2727" "D2733" "D2738" "D2749" "D2776" "D2792" "M1054"
[37] "M1055" "M120"  "M1241" "M1390" "M1503" "M1572" "M17"   "M1891" "M2070"
[46] "M2097" "M2184" "M2437" "M2515" "M2729" "M2807" "M3142" "M337"  "M3484"
[55] "M3514" "M359"  "M3627" "M4161" "M444"  "M485"  "M503"  "M5668" "M5775"
[64] "M6199" "M810"

> sum(celRunDate < "2004-03-09")

[1] 54
```

All of the samples run before the main gap have numerical ids. These 54 samples are the samples described by Berchuck et al. The more recent samples have "D" and "M" prefixes, presumably corresponding to whether the samples came from Duke or Moffitt. Let's look at the latter a bit more closely.

```
> table(celRunDate[celRunDate >= "2004-03-09"], substr(names(celRunDate[celRunDate >=
+     "2004-03-09"]), 1, 1))

             D  M
  2004-03-09  0 16
  2004-03-16  0  6
  2004-04-20  0  5
  2004-05-18 15  0
  2004-05-21  7  0
  2004-05-27  7  0
  2004-06-22  1  0
  2004-06-23  5  3
```

Most of the Moffitt samples were run, followed by most of the Duke samples, and then a mix from both on the last day.

4 Looking at Clinical Info by Run Date

Ideally, we would like the clinical factors of interest to be relatively balanced with respect to potential blocking strata, so as to avoid confounding.

4.1 Stage

```
> table(celRunDate, clinicalInfo$Stage)
```

```
celRunDate    2  3  4
  2002-09-20  0 10  0
  2002-10-23  0  7  2
  2002-11-12  0  9  0
  2002-12-16  0  1  0
  2002-12-21  0  3  0
  2003-01-03  0 10  1
  2003-05-30  0  8  2
  2003-06-26  0  0  1
  2004-03-09  1 12  3
  2004-03-16  0  5  1
  2004-04-20  0  3  1
  2004-05-18  0 14  1
  2004-05-21  0  6  1
  2004-05-27  0  5  2
  2004-06-22  0  0  1
  2004-06-23  0  5  3
```

No gross imbalances here; this is made easier by the fact that the vast majority of the cases are stage 3 of some type.

4.2 Grade

```
> table(celRunDate, clinicalInfo$Grade)
```

```
celRunDate       2/3  ?  1  2  3  4 UNK
  2002-09-20  0    0  0  0  5  5  0   0
  2002-10-23  0    0  0  0  4  5  0   0
  2002-11-12  0    0  0  1  3  4  1   0
  2002-12-16  0    0  0  0  1  0  0   0
  2002-12-21  0    0  1  0  1  1  0   0
  2003-01-03  0    0  0  0  7  4  0   0
  2003-05-30  0    0  0  0  9  1  0   0
  2003-06-26  0    0  0  0  1  0  0   0
  2004-03-09  0    2  0  2  2 10  0   0
  2004-03-16  0    0  0  0  1  5  0   0
  2004-04-20  0    0  0  0  0  5  0   0
  2004-05-18  1    0  0  1  8  5  0   0
  2004-05-21  0    0  0  0  7  0  0   0
  2004-05-27  0    0  0  0  2  5  0   0
  2004-06-22  0    0  0  0  1  0  0   0
  2004-06-23  0    0  0  0  2  5  0   1
```

As with stage, we do not see gross imbalances.

4.3 Debulk

```
> table(celRunDate, clinicalInfo$Debulk)
```

```
celRunDate    O  S
  2002-09-20  6  4
  2002-10-23  1  8
  2002-11-12  7  2
  2002-12-16  0  1
  2002-12-21  0  3
  2003-01-03  4  7
  2003-05-30  6  4
  2003-06-26  1  0
  2004-03-09 13  3
  2004-03-16  4  2
  2004-04-20  4  1
  2004-05-18  6  9
  2004-05-21  2  5
  2004-05-27  4  3
  2004-06-22  0  1
  2004-06-23  6  2
```

As above, we do not see gross imbalances.

4.4 Response

```
> 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
  2002-12-21  0  3
  2003-01-03  9  2
  2003-05-30  8  2
  2003-06-26  0  1
  2004-03-09 13  3
  2004-03-16  6  0
  2004-04-20  5  0
  2004-05-18  8  7
  2004-05-21  3  4
  2004-05-27  6  1
  2004-06-22  0  1
  2004-06-23  6  2
```

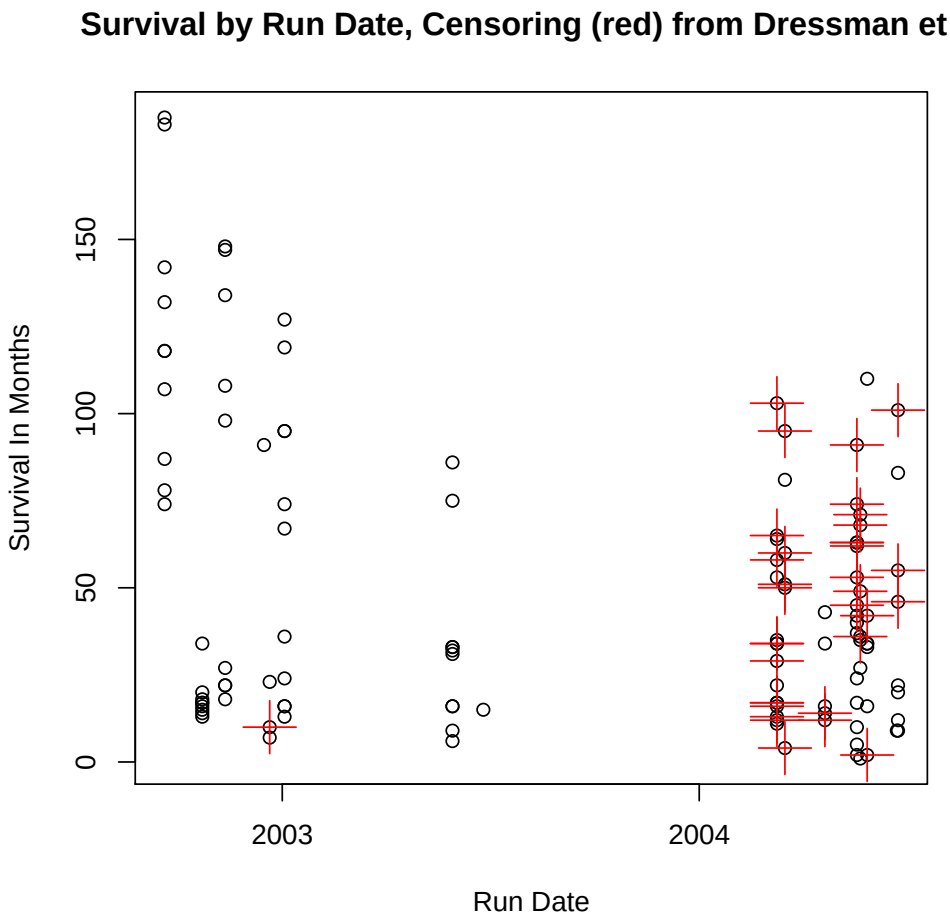
As above, we do not see gross deviations by date from the 85/34 ratio expected here. The 10/0 split on the 2002-09-20 has a one-sided p-value of 0.0346, and the 0/3 split on 2002-12-21 has a one-sided p-value of 0.0233 in the opposite direction. Given that we're looking at multiple dates, this is not too bad.

4.5 Survival and Censoring

For dealing with survival, plots are more effective than tables. In one plot, we superimpose the censoring information used in Dressman et al, and in the next we superimpose the censoring information used in Bild et al.

First, using the Dressman et al censoring.

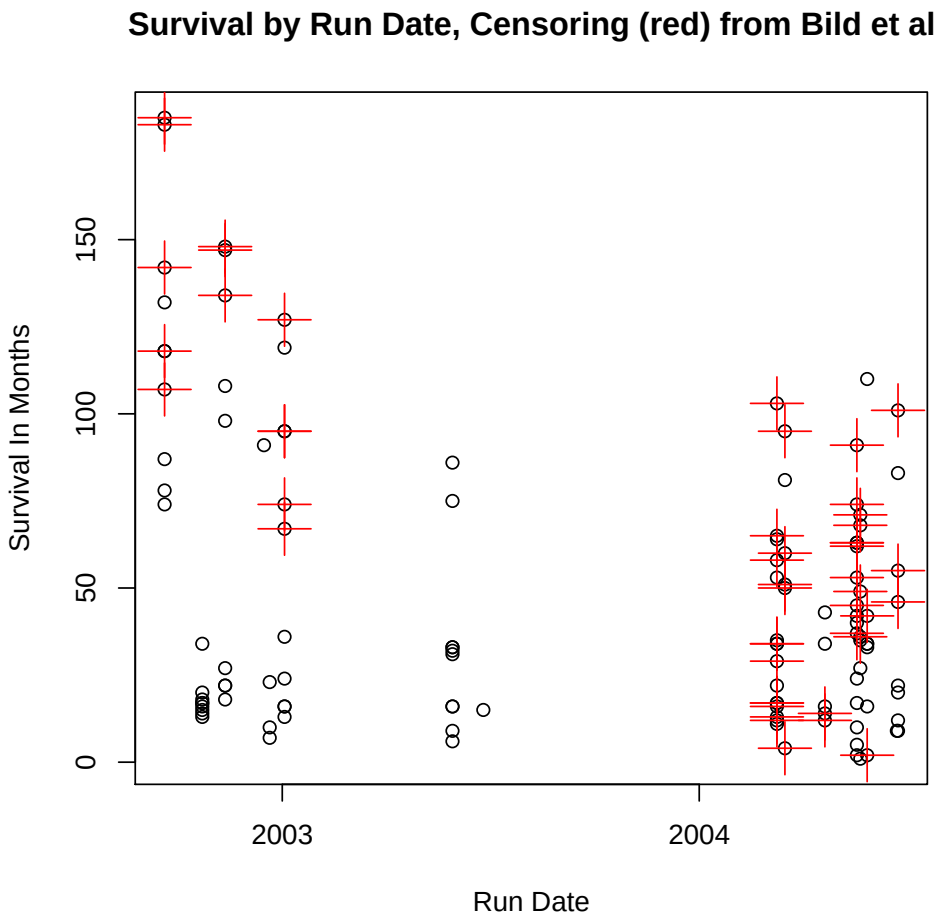
```
> plot(celRunDate, clinicalInfo$SurvMonths, xlab = "Run Date",
+     ylab = "Survival In Months", main = "Survival by Run Date, Censoring (red) from Dressman et al")
> points(celRunDate[clinicalInfo$Censoring == "Alive"], clinicalInfo$SurvMonths[clinicalInfo$Censoring =
+     "Alive"], pch = 3, cex = 3, col = "red")
```



The plot shows the clear time division between the Berchuck et al samples and the more recent ones. The almost complete lack of censored observations at the left is striking. The very first date blocks, as well, appear to exhibit some confounding with survival (all of the samples in the very first block have associated survival times in excess of 7 years).

Next, we apply the censoring from Bild et al.

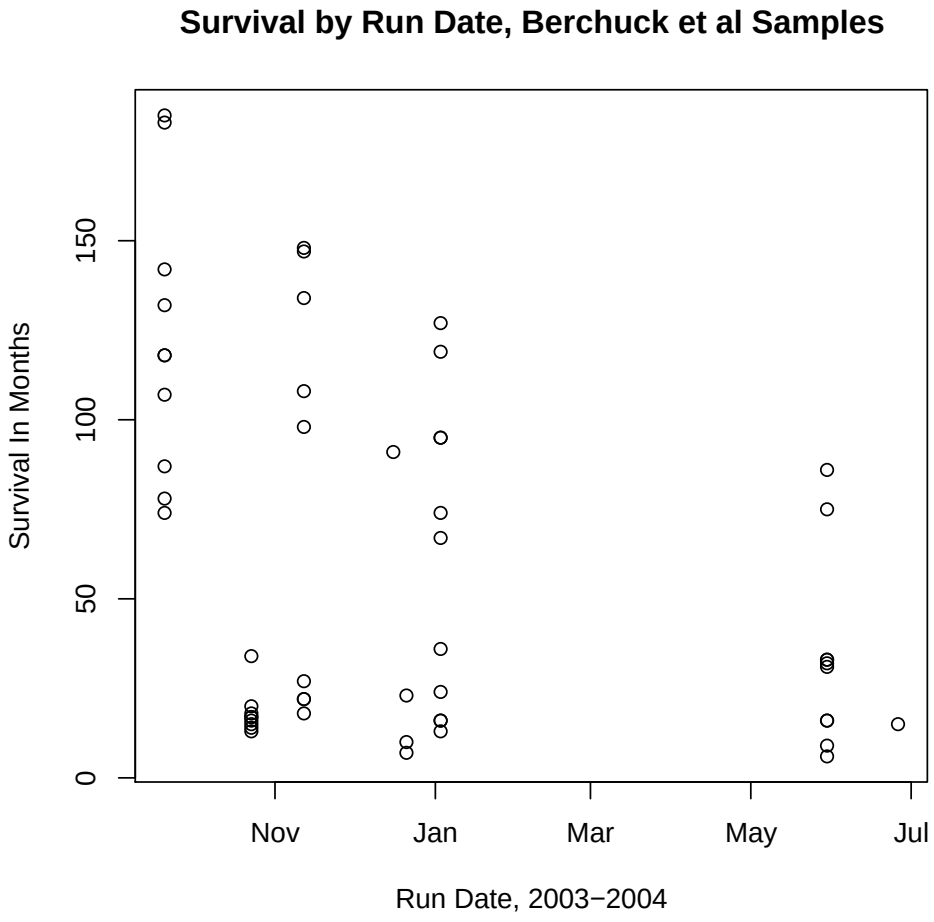
```
> plot(celRunDate, clinicalInfo$SurvMonths, xlab = "Run Date",
+     ylab = "Survival In Months", main = "Survival by Run Date, Censoring (red) from Bild et al")
> points(celRunDate[clinicalInfo$CensoringBild == "Alive"], clinicalInfo$SurvMonths[clinicalInfo$Censori
+     "Alive"], pch = 3, cex = 3, col = "red")
```



Adding the censoring information to the first group changes the picture quite a bit. As noted earlier, we suspect that this plot is more likely to be accurate than the other.

Finally, we zoom in on the data from Berchuck et al, to see the run date divisions more clearly.

```
> plot(celRunDate[celRunDate < "2004-03-09"], clinicalInfo$SurvMonths[celRunDate <
+     "2004-03-09"], xlab = "Run Date, 2003-2004", ylab = "Survival In Months",
+     main = "Survival by Run Date, Berchuck et al Samples")
```



Here it becomes apparent that the first two blocks are the major ones to worry about with respect to confounding with survival.

5 Summary

1. The CEL files can be separated into batches based on run date, with some large gaps in time between batches.
2. There are 54 CEL files that were run before 2004-03-09, all of which were described in the Berchuck paper. These samples have simple numerical IDs.
3. The samples run on or after 2004-03-09 have identifiers that either being with M (Moffitt) or D (Duke); the source is confounded with run date.
4. Most clinical data (stage, grade, debulking, response) do not appear to be confounded with run date.

5. Censoring of the survival based on the Dressman annotations is confounded with run date. However, using the alternative annotations in Bild, this confounding disappears.

6. Survival appears to be confounded with run date, especially in the samples that were processed earliest.

6 Appendix

6.1 Saves

```
> save(celRunDate, file = paste("RDataObjects", "celRunDate.Rda",
+     sep = .Platform$file.sep))
```

6.2 SessionInfo

```
> sessionInfo()

R version 2.5.1 (2007-06-27)
i386-pc-mingw32

locale:
LC_COLLATE=English_United States.1252;LC_CTYPE=English_United States.1252;LC_MONETARY=English_United Sta

attached base packages:
[1] "splines"   "tools"     "stats"     "graphics"  "grDevices" "utils"
[7] "datasets"  "methods"   "base"

other attached packages:
       survival  ClassDiscovery         cluster ClassComparison      PreProcess
         "2.32"         "2.5.0"        "1.11.7"         "2.5.0"         "2.5.0"
      oompaBase     geneplotter         lattice        annotate            affy
        "2.5.0"        "1.14.0"       "0.15-11"        "1.14.1"        "1.14.2"
         affyio         Biobase
        "1.4.1"        "1.14.1"
```