Building changAll.Rda

Keith A. Baggerly

July 18, 2009

Contents

List of Figures

List of Tables

1 Executive Summary

1.1 Introduction

In this report, we combine the Chang et al. [1] expression data from GEO with the other information supplied in their supplementary table and in their Table 1. In order to combine the data, we use the expression data given in their supplementary table for their key probesets to map the sample identifiers used in their clinical tables to the identifiers used at GEO.

1.2 Methods

We loaded three previously assembled Rda files: gse349, gse350, and changSuppAndTable. We then used the ordering of probeset values from the supplementary quantifications to suggest how the data from GEO should be ordered to match.

1.3 Results

We created a "changAll" matrix of the quantifications from GEO, using sample names from the clinical information. We created a "changAllInfo" data frame of sample information, including the mapping between GEO ids and sample names. We also created a vector of the "changKeyGenes". We stored these in RDataObjects as "changAll.Rda."

2 Options and Libraries

```
> options(width = 80)
```

3 Loading and Parsing Data

3.1 Earlier Rda Files

We begin by loading three Rda files assembled earlier: gse349, gse350, and changSuppAndTable.

```
> rdaList <- c("gse349", "gse350", "changSuppAndTable")
> for (rdaFile in rdaList) {
+     rdaFullFile <- file.path("RDataObjects", paste(rdaFile, "Rda",
+         sep = "."))
+     if (file.exists(rdaFullFile)) {
+         cat("loading ", rdaFullFile, " from cache\n")
+         load(rdaFullFile)
+     }
+     else {
+         cat("building ", rdaFullFile, " from raw data\n")
+         Stangle(file.path("RNowebSource", paste("buildRda", rdaFile,
+             "Rnw", sep = ".")))
+         source(paste("buildRda", rdaFile, "R", sep = "."))
+     }
+ }
```

```
loading  RDataObjects/gse349.Rda  from cache
loading  RDataObjects/gse350.Rda  from cache
loading  RDataObjects/changSuppAndTable.Rda  from cache
```

3.2 Bundling GEO Data

Next, we bundle the information from the two GEO files.

```
> if (all(rownames(gse349) == rownames(gse350))) {
+     geoAll <- cbind(gse349, gse350)
+     geoAllInfo <- rbind(gse349Info, gse350Info)
+ }
> geoAll[1:2, ]
```

```
                GSM4901 GSM4902 GSM4904 GSM4905 GSM4906  GSM4909 GSM4910
AFFX-MurIL2_at  54.1184 49.3722  49.123 75.3204 54.2316  52.3357 41.1346
AFFX-MurIL10_at 81.4273 65.2157  65.176 96.0592 50.1190 120.6330 99.8087
                GSM4911 GSM4912 GSM4913  GSM4916 GSM4918 GSM4922 GSM4924
AFFX-MurIL2_at  28.7636 43.5179 46.4989  37.4464 33.4237 35.5482 32.2541
AFFX-MurIL10_at 66.6838 73.8450 48.1602 108.1250 94.9995 50.2598 56.1386
                GSM4903 GSM4907 GSM4908 GSM4914 GSM4915 GSM4917 GSM4919
AFFX-MurIL2_at  38.0593 44.0123 41.8354 35.4666 43.0591 34.2387 28.6799
AFFX-MurIL10_at 59.5524 60.4303 61.2471 81.0051 57.9803 82.1848 48.3541
                 GSM4920 GSM4921 GSM4923
AFFX-MurIL2_at   30.1072 45.7519 47.4269
AFFX-MurIL10_at 109.1020 38.0342 53.7161
```

```
> geoAllInfo
```

```
                 dataFile sampleName    status
GSM4901 GSM4901-tbl-1.txt         44 Resistant
GSM4902 GSM4902-tbl-1.txt         51 Resistant
GSM4904 GSM4904-tbl-1.txt        113 Resistant
GSM4905 GSM4905-tbl-1.txt        118 Resistant
GSM4906 GSM4906-tbl-1.txt        136 Resistant
GSM4909 GSM4909-tbl-1.txt        356 Resistant
GSM4910 GSM4910-tbl-1.txt        358 Resistant
GSM4911 GSM4911-tbl-1.txt        359 Resistant
GSM4912 GSM4912-tbl-1.txt        370 Resistant
GSM4913 GSM4913-tbl-1.txt        377 Sensitive
GSM4916 GSM4916-tbl-1.txt        432 Resistant
GSM4918 GSM4918-tbl-1.txt        438 Resistant
GSM4922 GSM4922-tbl-1.txt        555 Resistant
GSM4924 GSM4924-tbl-1.txt        562 Resistant
GSM4903 GSM4903-tbl-1.txt         71 Sensitive
GSM4907 GSM4907-tbl-1.txt        142 Sensitive
GSM4908 GSM4908-tbl-1.txt        273 Sensitive
GSM4914 GSM4914-tbl-1.txt        413 Sensitive
GSM4915 GSM4915-tbl-1.txt        425 Sensitive
```

```
GSM4917 GSM4917-tbl-1.txt        437 Sensitive
GSM4919 GSM4919-tbl-1.txt        447 Sensitive
GSM4920 GSM4920-tbl-1.txt        458 Sensitive
GSM4921 GSM4921-tbl-1.txt        492 Sensitive
GSM4923 GSM4923-tbl-1.txt        558 Sensitive
```

Everything matches nicely.

4 Mapping Clinical Data to GEO

Chang et al. [1] name their samples N1-N24. The Chang et al. [1] arrays at GEO are named GSM4901-GSM4924. We want to match the simpler names with the array ids. We do this using the expression values given in the Chang et al. [1] supplementary table for the 92 “important”. We first try to match the expression values exactly.

```
> which(geoAll == changSuppQuants[1, 1])
```

```
integer(0)
```

Unfortunately, the numbers in the Lancet table do not exactly match the numbers from GEO. Some change has occurred, possibly in the version of dChip used or the size of the sample set used to define the models. We can still identify the samples, but the approach will be based upon high correlation rather than perfect identity. To do this, we compute pairwise correlations between the quantifications from the supplmentary table and the corresponding subset of the expression values from GEO.

```
> changKeyGenes <- rownames(changSuppQuants)
> changGEOCorrs <- cor(changSuppQuants, geoAll[changKeyGenes, ])
> changGEOCorrs[22:24, 1:3]
```

```
      GSM4901   GSM4902   GSM4904
N22 0.8610841 0.8759926 0.9947137
N23 0.8331194 0.6951595 0.7947931
N24 0.9438636 0.9660246 0.8269891
```

Looking at the subtable shown above, we see (for example) that N22 is very highly correlated with GSM4904. We can check to see how much the biggest correlations exceed the next biggest to see if there are “clear winners”. A plot of the top two correlations for each column from the supplementary table is shown in Figure 1. In each case, there is a clear winner.

Given that there is a clear mapping, we look at where the winners are. These are shown in Figure 2. The sensitive and resistant groups match with the exception of GSM4913. As noted by Coombes et al. [2], personal communication with Chang et al. [1] established that this sample was mistakenly labeled as “resistant” at GEO; it should have been labeled “sensitive”.

We now extract the matches, and use these to construct a quantification matrix with columns sorted N1-N24.

```
> bestPairs <- which(t(changGEOCorrs) > 0.988, arr.ind = TRUE)
> bestPairs[1:3, ]
> tempRowMaxes <- apply(changGEOCorrs, 1, max)
> tempRowSecondBiggest <- apply(changGEOCorrs, 1, function(x) {
+  -sort(-x)[2]})
> plot(sort(tempRowMaxes), ylim = c(0.8, 1),
+      ylab = "Corr with Chang Supp Data",
+      main = "Best and Next Best Corrs by Row, Sorted by Max")
> points(tempRowSecondBiggest[order(tempRowMaxes)], col = "red")
> min(tempRowMaxes)

[1] 0.9880765

> sum(changGEOCorrs >= min(tempRowMaxes))

[1] 24
```

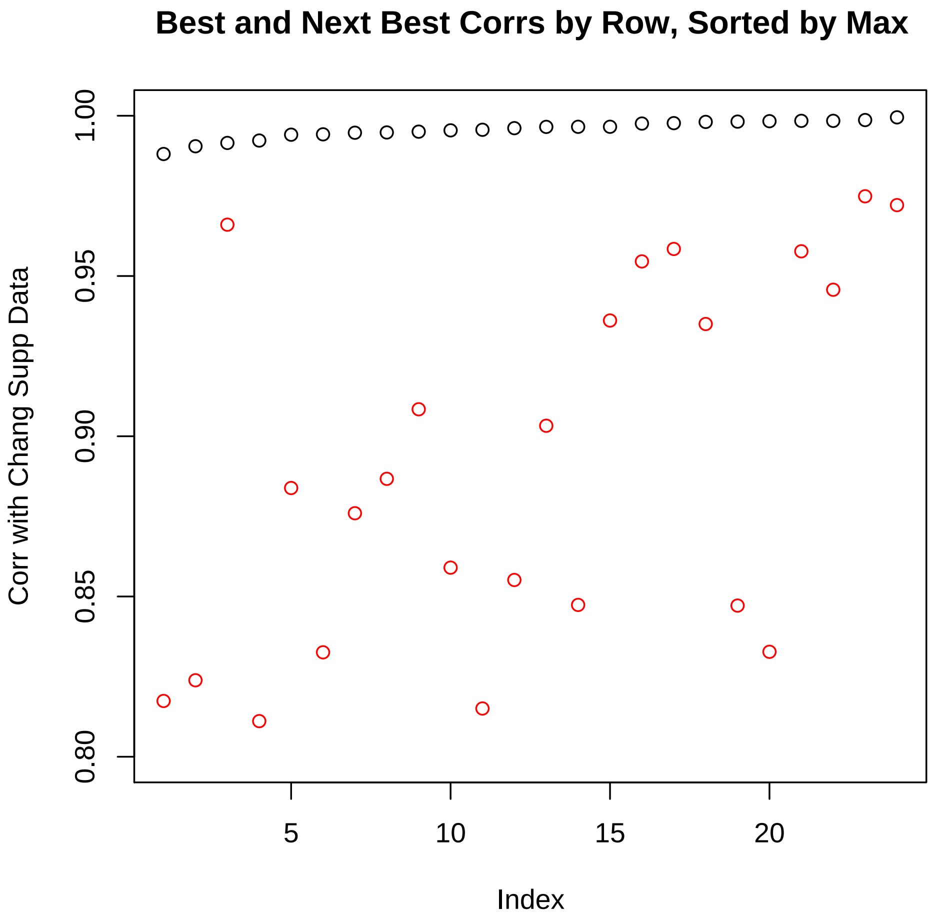

Figure 1: Plot of the best and second-best correlations in the GEO data for each column from the supplementary table. For each sample, there is a clear "winning" match.

```
> image(1:24, 1:24, changGEOCorrs < 0.988, axes = FALSE, xlab = "",
+     ylab = "", asp = 1, main = "Corrs > 0.988 Between Lancet and GEO",
+     ylim = c(24.5, 0.5))
> lines(c(0.5, 24.5), c(14.5, 14.5))
> lines(c(11.5, 11.5), c(0.5, 24.5))
> rect(0.5, 0.5, 24.5, 24.5)
> axis(1, at = 1:24, labels = rownames(changGEOCorrs), las = 2,
+     line = -0.5, tick = 0)
> axis(2, at = 1:24, labels = colnames(changGEOCorrs), las = 2,
+     line = -2.1, tick = 0)
> axis(3, at = c(6, 18), labels = c("Sensitive", "Resistant"),
+     las = 1, line = -1, tick = 0)
> mtext(text = c("GSE349 (Res)", "GSE350 (Sen)"), side = 4, at = c(7.5,
+     19.5), line = -1)
```

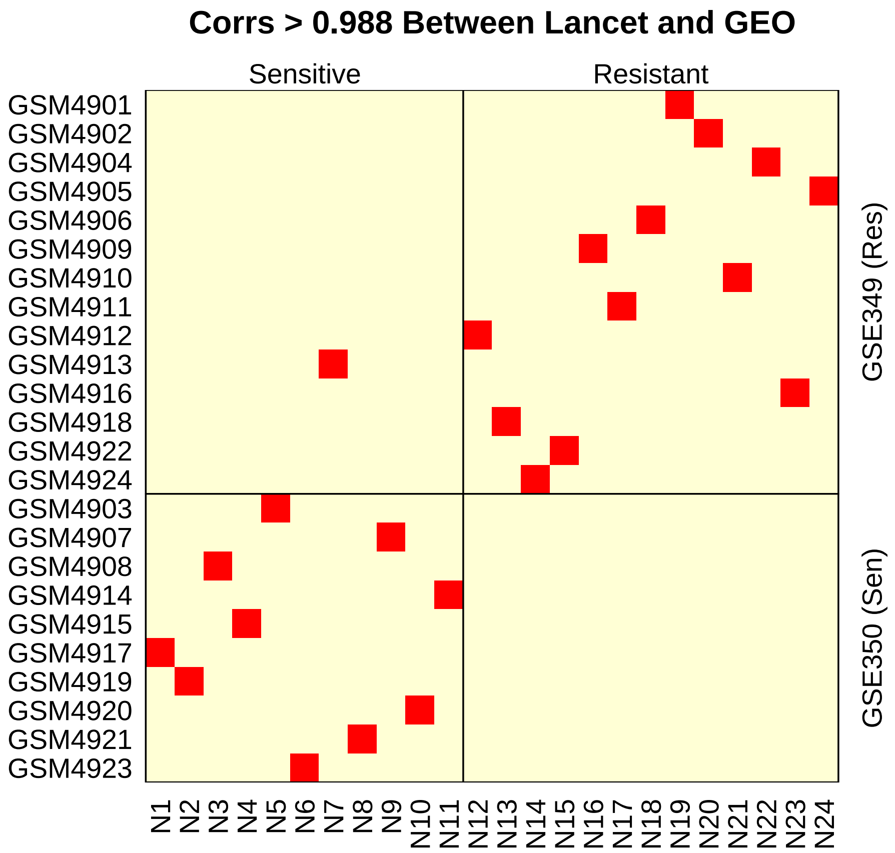

Figure 2: Pairs of samples showing the best correlations between the quantifications in the Lancet supplementary table and the quantifications at GEO. The sensitive and resistant groupings match with the exception of GSM4913, which was mistakenly labeled resistant at GEO.

```
        row col
GSM4917  20   1
GSM4919  21   2
GSM4908  17   3
```

```
> bestMapping <- rownames(bestPairs)
> names(bestMapping) <- colnames(changSuppQuants)
> bestMapping
```

```
       N1        N2        N3        N4        N5        N6        N7        N8 
"GSM4917" "GSM4919" "GSM4908" "GSM4915" "GSM4903" "GSM4923" "GSM4913" "GSM4921" 
       N9       N10       N11       N12       N13       N14       N15       N16 
"GSM4907" "GSM4920" "GSM4914" "GSM4912" "GSM4918" "GSM4924" "GSM4922" "GSM4909" 
      N17       N18       N19       N20       N21       N22       N23       N24 
"GSM4911" "GSM4906" "GSM4901" "GSM4902" "GSM4910" "GSM4904" "GSM4916" "GSM4905" 
```

```
> changAll <- geoAll[, bestMapping]
> colnames(changAll) <- colnames(changSuppQuants)
> geoAllInfoReorg <- geoAllInfo[bestMapping, ]
```

We can do a quick double-check to confirm that we have the ordering correct.

```
> changAllCorWChangSupp <- cor(changAll[changKeyGenes, ], changSuppQuants)
> sum(changAllCorWChangSupp > 0.988)
```

```
[1] 24
```

```
> sum(diag(changAllCorWChangSupp) > 0.988)
```

```
[1] 24
```

All of the high correlations are now on the main diagonal, where we want them to be.

5 Condense Sample Info

At this point, we have distinct sample specific information in three consistently ordered files: geoAllInfoReorg, changSuppClinical, and changTable1. We now examine the first few rows of each.

```
> geoAllInfoReorg[1:2, ]
```

```
                dataFile sampleName    status
GSM4917 GSM4917-tbl-1.txt        437 Sensitive
GSM4919 GSM4919-tbl-1.txt        447 Sensitive
```

```
> changSuppClinical[1:2, ]
```

```
   PercentResidualTumor    Status
N1                    1 Sensitive
N2                    1 Sensitive
```

```
> changTable1[1:2, ]
```

```
  Patient Age..years. Menopausal.status Ethnic.origin
1       1          37     Premenopausal      Hispanic
2       2          55    Postmenopausal      Hispanic
  Bidimensional.tumour.size..cm. Clinical.axillary.nodes
1                         10x10                      No
2                          10x8                     Yes
  Oestrogen..receptor.status Progesterone..receptor.status HER.2 Tumour.type
1                          -                             -     -         IMC
2                          -                             -     +         IDC
```

Now we combine most of the distinct information into a single data frame.

```
> changAllInfo <-
+   data.frame(row.names     = names(bestMapping),
+              geoID         = I(bestMapping),
+              geoTitle      = geoAllInfoReorg[,"sampleName"],
+              status        = geoAllInfoReorg[,"status"],
+              pctResidTumor = changSuppClinical[,"PercentResidualTumor"],
+              ageInYears    = changTable1[,"Age..years."],
+              er            = changTable1[,"Oestrogen..receptor.status"],
+              pr            = changTable1[,"Progesterone..receptor.status"],
+              her2          = changTable1[,"HER.2"],
+              tumorSizeInCm = changTable1[,"Bidimensional.tumour.size..cm."],
+              menopause     = changTable1[,"Menopausal.status"],
+              ethnicity     = changTable1[,"Ethnic.origin"],
+              tumorType     = changTable1[,"Tumour.type"])
```

Now we do a visual check.

```
> changAllInfo
```

```
      geoID geoTitle    status pctResidTumor ageInYears er pr her2
N1  GSM4917      437 Sensitive             1         37  -  -    -
N2  GSM4919      447 Sensitive             1         55  -  -    +
N3  GSM4908      273 Sensitive             6         41  +  +    -
N4  GSM4915      425 Sensitive             6         43  +  -    -
N5  GSM4903       71 Sensitive            13         50  -  -    -
N6  GSM4923      558 Sensitive            14         55  +  +    -
N7  GSM4913      377 Sensitive            16         42  +  +    -
N8  GSM4921      492 Sensitive            17         63  +  +    -
N9  GSM4907      142 Sensitive            18         50  +  +    -
N10 GSM4920      458 Sensitive            22         38  +  +    -
N11 GSM4914      413 Sensitive            25         58  +  +    -
N12 GSM4912      370 Resistant            36         62  +  -    -
N13 GSM4918      438 Resistant            38         40  +  +    -
N14 GSM4924      562 Resistant            39         36  +  +    -
N15 GSM4922      555 Resistant            44         56  +  -    -
N16 GSM4909      356 Resistant            45         38  +  -    -
N17 GSM4911      359 Resistant            47         54  +  +    +
N18 GSM4906      136 Resistant           60         52  +  +     -
N19 GSM4901       44 Resistant           64         57  -  -     -
N20 GSM4902       51 Resistant           65         52  -  -     -
N21 GSM4910      358 Resistant           70         44  -  -     -
N22 GSM4904      113 Resistant          100         41  +  +     -
N23 GSM4916      432 Resistant          100         38  +  +     -
N24 GSM4905      118 Resistant          131         54  +  +     -
    tumorSizeInCm      menopause ethnicity tumorType
N1          10x10  Premenopausal  Hispanic       IMC
N2           10x8 Postmenopausal  Hispanic       IDC
N3            6x5  Premenopausal     Black       IDC
N4          15x13  Premenopausal     Black       IMC
N5          20x23 Postmenopausal     Black       IDC
N6          11x11 Postmenopausal     Black       IDC
N7            7x9  Premenopausal     Black       IMC
N8            7x8 Postmenopausal     Black       IMC
N9           13x9 Postmenopausal     Black       IDC
N10           8x8  Premenopausal  Hispanic       IMC
N11           7x7 Postmenopausal  Hispanic       IMC
N12           4x4 Postmenopausal  Hispanic       IDC
N13       5.5x4.5  Premenopausal  Hispanic       IMC
N14           6x6  Premenopausal     Black       IDC
N15         5x5.5 Postmenopausal     Black       IMC
N16           6x6  Premenopausal     White       IDC
N17           5x6 Postmenopausal     White       IDC
N18         10x10 Postmenopausal     White       IDC
N19           8x8 Postmenopausal     White       IDC
N20         10x10 Postmenopausal     Black       IDC
N21         11x11  Premenopausal     Black       IDC
N22           6x5  Premenopausal     Black       IDC
N23           8x8  Premenopausal     White       IDC
N24           9x7 Postmenopausal     Black       IDC
```

Everything looks as expected.

6 Save Rda File

Finally, we save the reordered quantification matrix, the combined sample annotation and the list of key genes.

```
> save(changAll, changAllInfo, changKeyGenes, file = file.path("RDataObjects",
+     "changAll.Rda"))
```

7 Appendix

7.1 File Location

```
> getwd()
```

```
[1] "/Users/kabagg/ReproRsch/WebSite"
```

7.2 Saves

7.3 SessionInfo

```
> sessionInfo()

R version 2.8.1 (2008-12-22)
i386-apple-darwin8.11.1

locale:
en_US.UTF-8/en_US.UTF-8/C/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] XML_2.3-0
```

References

[1] Chang JC, Wooten EC, Tsimelzon A, et al.: Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet*, **362**:362-369, 2003.

[2] Coombes KR, Wang J, Baggerly KA: Microarrays: retracing steps. *Nat Med*, **13**:1276-7, 2007. Author reply, 1277-8.