Building gse349.Rda

Keith A. Baggerly

July 18, 2009

Contents

List of Figures

List of Tables

1 Executive Summary

1.1 Introduction

In this report, we assemble the quantification and annotation information available for GEO dataset GSE349, nominally corresponding to the samples from Chang et al. [1] that were resistant to docetaxel. As noted by Coombes et al. [2], personal communication with Chang et al. [1] revealed that sample GSM4913 should have been labeled "sensitive", not "resistant."

1.2 Methods

We acquired the MINiML files for GSE349 from `http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE349`. This gzipped tar file was uncompressed and the contents were stored in RawData/GEO/GSE349. We parsed the family.xml file to extract annotation information, and loaded the array quantifications from the individual tbl files.

1.3 Results

We created a "gse349" matrix of array quantifications and a "gse349Info" data frame of sample information. We stored these in RDataObjects as "gse349.Rda."

2 Options and Libraries

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

3 Loading and Parsing Data

3.1 Parse XML

We begin by exploring the family XML structure for GSE349 to see what annotation information we might want to include.

```
> gse349Dir <- file.path("RawData", "GEO", "GSE349")
> gse349Xml <- xmlTreeParse(file.path(gse349Dir, "GSE349_family.xml"))
> gse349Root <- xmlRoot(gse349Xml)
> xmlSize(gse349Root)

[1] 30

> table(xmlSApply(gse349Root, xmlName))

Contributor    Database    Platform      Sample      Series
         13           1           1          14           1

> xmlValue(gse349Root[[which(xmlSApply(gse349Root, xmlName) ==
+     "Series")]][["Summary"]])

[1] "These patients proved resistant to docetaxel treatment, exhibiting residual tumor of 25% or greate

> idxSample <- which(xmlSApply(gse349Root, xmlName) == "Sample")
> gse349Root[[idxSample[1]]]

<Sample iid="GSM4901">
 <Status database="GEO">
  <Submission-Date>2003-03-19</Submission-Date>
  <Release-Date>2003-06-26</Release-Date>
  <Last-Update-Date>2009-03-16</Last-Update-Date>
  <Comment>Raw data provided as supplementary file</Comment>
```

```
 </Status>
 <Title>44</Title>
 <Accession database="GEO">GSM4901</Accession>
 <Type>RNA</Type>
 <Channel-Count>1</Channel-Count>
 <Channel position="1">
  <Source>human breast cancer core biopsy</Source>
  <Organism>Homo sapiens</Organism>
  <Characteristics>none</Characteristics>
  <Molecule>total RNA</Molecule>
 </Channel>
 <Description>Antibody amplified expression data, normalized and modeled.</Description>
 <Data-Processing/>
 <Platform-Ref ref="GPL8300"/>
 <Contact-Ref ref="contrib1"/>
 <Supplementary-Data type="CEL">ftp://ftp.ncbi.nih.gov/pub/geo/DATA/supplementary/samples/GSM4nnn/GSM490
 <Data-Table>
  <Column position="1">
   <Name>ID_REF</Name>
  </Column>
  <Column position="2">
   <Name>VALUE</Name>
  </Column>
  <External-Data rows="12625">GSM4901-tbl-1.txt</External-Data>
 </Data-Table>
</Sample>
```

The series summary confirms that these are the resistant samples. For each sample, the three fields we want are "Title", which gives one of the names that Chang et al. [1] used for this sample, "Accession", which gives the corresponding GSM id, and "Data-Table/External-Data" which names the files to be loaded. Examining the subfields of the "Data-Table" tag tells us the structure of the files to be loaded.

```
> tempSampleNames <- sapply(idxSample, function(x) {
+     xmlValue(gse349Root[[x]][["Title"]])
+ })
> tempGSMID <- sapply(idxSample, function(x) {
+     xmlValue(gse349Root[[x]][["Accession"]])
+ })
> tempDataFile <- sapply(idxSample, function(x) {
+     xmlValue(gse349Root[[x]][["Data-Table"]][["External-Data"]])
+ })
> tempStatus <- rep("Resistant", length(tempGSMID))
> gse349Info <- data.frame(row.names = tempGSMID, dataFile = I(tempDataFile),
+     sampleName = I(tempSampleNames), status = I(tempStatus))
> rm(list = ls(pattern = "^temp"))
```

Finally, we correct the status entry for GSM4913.

```
> gse349Info["GSM4913", "status"] <- "Sensitive"
> gse349Info
```

```
                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
```

3.2 Loading Array Quantifications

Now that we have the annotation, we load the actual array quantifications. This takes about 8 seconds on my MacBook Pro laptop.

```
> tempTable <- read.table(file.path(gse349Dir, gse349Info[1, "dataFile"]),
+     row.names = 1)
> gse349 <- matrix(0, nrow = dim(tempTable)[1], ncol = dim(gse349Info)[1])
> rownames(gse349) <- rownames(tempTable)
> colnames(gse349) <- rownames(gse349Info)
> gse349[, 1] <- tempTable[, 1]
> for (tempSample in 2:dim(gse349Info)[1]) {
+     tempTable <- read.table(file.path(gse349Dir, gse349Info[tempSample,
+         "dataFile"]), row.names = 1)
+     if (all(rownames(tempTable) == rownames(gse349))) {
+         gse349[, tempSample] <- tempTable[, 1]
+     }
+ }
> rm(list = ls(pattern = "^temp"))
```

Now we check the first few values to make sure that everything loaded correctly.

```
> dim(gse349)

[1] 12625    14

> gse349[1:3, ]

                GSM4901  GSM4902  GSM4904  GSM4905  GSM4906   GSM4909 GSM4910
AFFX-MurIL2_at  54.11840 49.37220 49.12300 75.32040 54.23160  52.33570 41.1346
AFFX-MurIL10_at 81.42730 65.21570 65.17600 96.05920 50.11900 120.63300 99.8087
AFFX-MurIL4_at   5.89822  6.07494  5.89822  5.89822  8.19463   8.33974 16.0799
                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
AFFX-MurIL4_at  21.6423 30.9097 25.2147  20.1633 10.8711 27.7929 22.9879
```

```
> apply(gse349, 2, min)
```

```
GSM4901 GSM4902 GSM4904 GSM4905 GSM4906 GSM4909 GSM4910 GSM4911 GSM4912 GSM4913
5.89822 5.89822 5.89822 5.89822 5.89822 5.89822 5.89822 5.89822 5.89822 5.89822
GSM4916 GSM4918 GSM4922 GSM4924
5.89822 5.89822 5.89822 5.89822
```

There are no 0's, so all 14 samples loaded successfully. We note that one low value (5.89822) occurs multiple times. This is the smallest value reported for all of the columns, so we are seeing a truncation effect.

4 Save Rda File

Finally, we save the quantification matrix and the annotation information.

```
> save(gse349, gse349Info, file = file.path("RDataObjects", "gse349.Rda"))
```

5 Appendix

5.1 File Location

```
> getwd()
```

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

5.2 Saves

5.3 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.