Building gse350.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 GSE350, nominally corresponding to the samples from Chang et al. [1] that were sensitive to docetaxel. As noted by Coombes et al. [2], personal communication with Chang et al. [1] revealed that sample GSM4913 (in GSE349) should have been labeled "sensitive", not "resistant."

1.2 Methods

We acquired the MINiML files for GSE350 from `http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE350`. This gzipped tar file was uncompressed and the contents were stored in RawData/GEO/GSE350. 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 "gse350" matrix of array quantifications and a "gse350Info" data frame of sample information. We stored these in RDataObjects as "gse350.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 GSE350 to see what annotation information we might want to include.

```
> gse350Dir <- file.path("RawData", "GEO", "GSE350")
> gse350Xml <- xmlTreeParse(file.path(gse350Dir, "GSE350_family.xml"))
> gse350Root <- xmlRoot(gse350Xml)
> xmlSize(gse350Root)

[1] 26

> table(xmlSApply(gse350Root, xmlName))

Contributor    Database    Platform      Sample      Series
         13           1           1          10           1

> xmlValue(gse350Root[[which(xmlSApply(gse350Root, xmlName) ==
+     "Series")]][["Summary"]])

[1] "These patients were sensitive to docetaxel treatment, exhibiting less than 25% residual tumor.\nKey

> idxSample <- which(xmlSApply(gse350Root, xmlName) == "Sample")
> gse350Root[[idxSample[1]]]

<Sample iid="GSM4903">
 <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>71</Title>
 <Accession database="GEO">GSM4903</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>Ab amplified expression values, normalized and modelled.</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">GSM4903-tbl-1.txt</External-Data>
 </Data-Table>
</Sample>
```

The series summary confirms that these are sensitive 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(gse350Root[[x]][["Title"]])
+ })
> tempGSMID <- sapply(idxSample, function(x) {
+     xmlValue(gse350Root[[x]][["Accession"]])
+ })
> tempDataFile <- sapply(idxSample, function(x) {
+     xmlValue(gse350Root[[x]][["Data-Table"]][["External-Data"]])
+ })
> tempStatus <- rep("Sensitive", length(tempGSMID))
> gse350Info <- data.frame(row.names = tempGSMID, dataFile = I(tempDataFile),
+     sampleName = I(tempSampleNames), status = I(tempStatus))
> rm(list = ls(pattern = "^temp"))
```

Finally, we list the info for GSE350 for visual checking.

```
> gse350Info
```

```
                dataFile sampleName    status
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
```

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(gse350Dir, gse350Info[1, "dataFile"]),
+     row.names = 1)
> gse350 <- matrix(0, nrow = dim(tempTable)[1], ncol = dim(gse350Info)[1])
> rownames(gse350) <- rownames(tempTable)
> colnames(gse350) <- rownames(gse350Info)
> gse350[, 1] <- tempTable[, 1]
> for (tempSample in 2:dim(gse350Info)[1]) {
+     tempTable <- read.table(file.path(gse350Dir, gse350Info[tempSample,
+         "dataFile"]), row.names = 1)
+     if (all(rownames(tempTable) == rownames(gse350))) {
+         gse350[, tempSample] <- tempTable[, 1]
+     }
+ }
> rm(list = ls(pattern = "^temp"))
```

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

```
> dim(gse350)

[1] 12625    10

> gse350[1:3, ]

               GSM4903  GSM4907  GSM4908 GSM4914  GSM4915 GSM4917 GSM4919
AFFX-MurIL2_at  38.0593 44.01230 41.83540 35.4666 43.05910 34.2387 28.6799
AFFX-MurIL10_at 59.5524 60.43030 61.24710 81.0051 57.98030 82.1848 48.3541
AFFX-MurIL4_at  10.3508  5.89822  6.66312 13.3874  9.84302 22.4733 12.3610
                GSM4920 GSM4921  GSM4923
AFFX-MurIL2_at   30.1072 45.7519 47.42690
AFFX-MurIL10_at 109.1020 38.0342 53.71610
AFFX-MurIL4_at   14.6647 14.7236  5.91698

> apply(gse350, 2, min)
```

```
GSM4903 GSM4907 GSM4908 GSM4914 GSM4915 GSM4917 GSM4919 GSM4920 GSM4921 GSM4923
5.89822 5.89822 5.89822 5.89822 5.89822 5.89822 5.89822 5.89822 5.89822 5.89822
```

There are no 0's, so all 10 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(gse350, gse350Info, file = file.path("RDataObjects", "gse350.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.