Building changSuppAndTable.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 clinical and expression information provided by Chang et al. [1] in their supplementary information file and in their Table 1.

1.2 Methods

We acquired the supplmentary table for Chang et al. [1] from `http://image.thelancet.com/extras/01art11086webtable.pdf`. We rearranged the file into csv form and dropped extraneous header rows for easier loading. We stored this revised file in RawData/ChangLancet/01art11086webtable_rev.csv. We also acquired Table 1 of Chang et al. [1] for comparison and parsed the clinical information into a csv file for easier loading. We stored this file in RawData/ChangLancet/changTable1.csv.

1.3 Results

We created a "changSuppQuants" matrix of array quantifications, a "changSuppProbeAnnot" data frame of probe annotation for their 92 key probesets, and a "changSuppClinical" data frame of sample information. We also created a "changTable1" data frame of sample information. We stored these in RDataObjects as "changSuppAndTable.Rda."

2 Options and Libraries

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

3 Loading and Parsing Data

3.1 The Supplementary Table from Chang et al. [1]

Chang et al. [1] provided some clinical and quantification information for the samples they examined in a supplementary table. This table is available at `http://image.thelancet.com/extras/01art11086webtable.pdf`. In particular, this table lists (a) a sample identifier (N1-N24), (b) the percent residual disease, and (c) expression profile values of 92 important probesets for each of the 24 samples profiled.

```
> changSupp <- read.table(file.path("RawData", "ChangLancet", "01art11086webtable_rev.csv"),
+     sep = ",", strip.white = TRUE)
> dim(changSupp)

[1] 96 29

> changSupp[1:5, c(1:4, 6, 29)]

                      V1      V2         V3              V4     V6    V29
1          Sample number                                      N1    N24
2     Residual tumour, %                                       1    131
3 Sensitive or resistant                                       S      R
4              Probe set GenBank Locus Link Official Symbol
5              1008_f_at  U50648       5610            PRKR 247.45 638.04
```

The supplementary table has four header rows of explanatory information, but not all headers are relevant for all columns. Of the 29 columns, the first five give annotation information for the Affymetrix probesets described (Probe Set, GenBank, Locus Link, Official Symbol, and Gene Name). The last 24 give sample specific information (Sample Name, Percent Residual Tumor, Sensitive/Resistant status, and probeset expression values). We partition this information into tables for probe annotation, clinical information, and expression.

We begin with probe annotation. What we want here are the first five columns of information for the last 92 (probe specific) rows.

```
> changSuppProbeAnnot <-
+   data.frame(GenBank   = as.character(changSupp[5:96,2]),
+              LocusLink = as.numeric(  changSupp[5:96,3]),
+              Symbol    = as.character(changSupp[5:96,4]),
+              GeneName  = as.character(changSupp[5:96,5]),
+              row.names = as.character(changSupp[5:96,1]))
> dim(changSuppProbeAnnot)

[1] 92  4

> changSuppProbeAnnot[1:3, ]

          GenBank LocusLink Symbol
1008_f_at  U50648        58   PRKR
1199_at    D13748        25 EIF4A1
1250_at    U47077        57  PRKDC
                                                                      GeneName
1008_f_at protein kinase, interferon-inducible double stranded RNA dependent
1199_at               eukaryotic translation initiation factor 4A, isoform 1
1250_at                 protein kinase, DNA-activated, catalytic polypeptide
```

We turn next to the clinical information. What we want here are the first three rows of header information for the last 24 (sample specific) columns.

```
> changSuppClinical <-
+   data.frame(PercentResidualTumor =   as.numeric(  t(changSupp[2,6:29])),
+              Status               = I(as.character(t(changSupp[3,6:29]))),
+              row.names            =   as.character(t(changSupp[1,6:29])))
> dim(changSuppClinical)

[1] 24  2

> changSuppClinical[c(1:2,24),]

    PercentResidualTumor Status
N1                     1      S
N2                     1      S
N24                  131      R
```

To keep things clear, we expand "S" and "R" to "Sensitive" and "Resistant", respectively.

```
> changSuppClinical[changSuppClinical[, "Status"] == "S", "Status"] <- "Sensitive"
> changSuppClinical[changSuppClinical[, "Status"] == "R", "Status"] <- "Resistant"
> changSuppClinical

    PercentResidualTumor    Status
N1                     1 Sensitive
N2                     1 Sensitive
```

```
N3                    6 Sensitive
N4                    6 Sensitive
N5                   13 Sensitive
N6                   14 Sensitive
N7                   16 Sensitive
N8                   17 Sensitive
N9                   18 Sensitive
N10                  22 Sensitive
N11                  25 Sensitive
N12                  36 Resistant
N13                  38 Resistant
N14                  39 Resistant
N15                  44 Resistant
N16                  45 Resistant
N17                  47 Resistant
N18                  60 Resistant
N19                  64 Resistant
N20                  65 Resistant
N21                  70 Resistant
N22                 100 Resistant
N23                 100 Resistant
N24                 131 Resistant
```

Finally, we turn to the quantifications for the 92 key genes. Here, we want the last 92 rows (genes) and the last 24 columns (samples).

```
> changSuppQuants <- matrix(as.numeric(as.matrix(changSupp[5:96,
+     6:29])), nrow = 92, ncol = 24)
> rownames(changSuppQuants) <- rownames(changSuppProbeAnnot)
> colnames(changSuppQuants) <- rownames(changSuppClinical)
> dim(changSuppQuants)
```

```
[1] 92 24
```

```
> changSuppQuants[1:5, 1:4]
```

```
               N1     N2     N3     N4
1008_f_at  247.45 149.39 321.16 301.74
1199_at   1004.11 764.48 300.12 594.85
1250_at     17.96  44.70  75.46  11.23
1624_at     31.54  35.99  63.05  24.35
1635_at     65.33  65.08  15.21  99.41
```

3.2 Table 1 from Chang et al. [1]

Chang et al. [1] provide more extensive clinical information for their 24 samples in their Table 1.

```
> changTable1 <- read.table(file.path("RawData", "ChangLancet",
+     "changTable1.csv"), sep = ",", header = TRUE)
> dim(changTable1)
```

```
[1] 24 10
```

```
> 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
```

The patient id values here are 1 through 24, as opposed to N1 through N24 in the supplementary table, but we assume the direct one-to-one correspondence holds.

4 Save Rda File

Finally, we save the supplementary quantification matrix, probe annotation, and sample annotation and the table sample annotation information.

```
> save(changSuppProbeAnnot, changSuppClinical, changSuppQuants,
+     changTable1, file = file.path("RDataObjects", "changSuppAndTable.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.