Easy testing for differential expression

Florian Hahne, Wolfgang Huber

June 22, 2006

In this short exercise, we will explore the most basic approach to the selection of differentially expressed genes: first, an unspecific filtering step to remove probes for genes that appear to be always unexpressed, second, a probe-by-probe statistical test.

There are many variations and improvements to the procedure shown here, and you can learn more about these in the full differential expression lab.

First, we load the necessary libraries and data.

```
> library("Biobase")
> library("genefilter")
> library("ALL")
> data("ALL")
```

The ALL (acute lymphoblastic leukemea) data set is quite large, so we select the subset of B-cell ALLs whose molecular type is either *BCR/ABL* or *NEG*.

```
> s1 <- grep("^B", as.character(ALL$BT))
> s2 <- which(as.character(ALL$mol.biol) %in% c("BCR/ABL", "NEG"))
> ALLs <- ALL[, intersect(s1, s2)]
> table(ALLs$mol.biol)
```

```
ALL1/AF4   BCR/ABL E2A/PBX1       NEG   NUP-98  p15/p16
       0        37        0        42        0        0
```

First, we calculate the overall variability across arrays of each probeset, regardless of its sample label. For this, we can use the function *rowSds*, which calculates the standard deviation for each row. An alternative is to calculate the interquartile range (IQR), for this we could employ the *rowQ* function also from the genefilter package.

```
> sds = rowSds(exprs(ALLs))
> sh = shorth(sds)
> sh
```

```
[1] 0.2423124
```

We can plot the histogram of the distribution of sds, see Figure 1. The function *shorth* calculates the midpoint of the *shorth* (the shortest interval containing half of the data), and is in many cases a reasonable estimator of the "peak" of a distribution. Its value 0.242 is drawn as a vertical line in Figure 1.

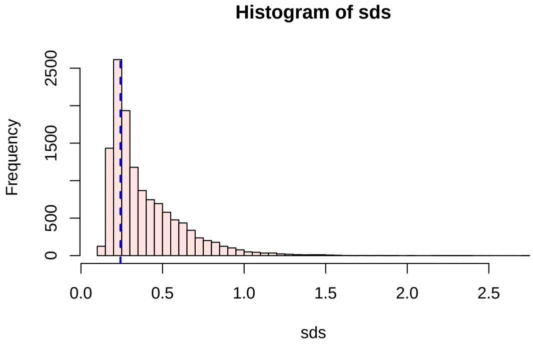

Figure 1: Histogram of `sds`.

```
> hist(sds, breaks = 50, col = "mistyrose")
> abline(v = sh, col = "blue", lwd = 2, lty = 2)
```

We will now discard all probe sets whose standard deviation is below the value of `sh`.

```
> ALLs <- ALLs[sds >= sh, ]
> dim(exprs(ALLs))
```

```
[1] 8812   79
```

Now let's perform a probe-by-probe t-test. The function *rowttests* can deal with `exprsSets`. It performs row-by-row tests for a significant difference in the location of two groups defined by a factor variable. In this case, we use the information about BCR/ABL mutation status in column `mol.biol` of `ALL`'s phenoData slot as grouping factor.

```
> tt <- rowttests(ALLs, "mol.biol")
> names(tt)
```

```
[1] "statistic" "dm"        "df"        "p.value"
```

Take a look at the histogram of resulting p-values (Figure 2):

```
> hist(tt$p.value, breaks = 50, col = "orange")
```

Now create a gene list containing the 20 highest-ranking genes with respect to t-test p-value

```
> g <- geneNames(ALLs[order(tt$p.value)])[1:20]
```

and print their gene symbols:

```
> library("hgu95av2")
> unlist(mget(g, hgu95av2SYMBOL))
```

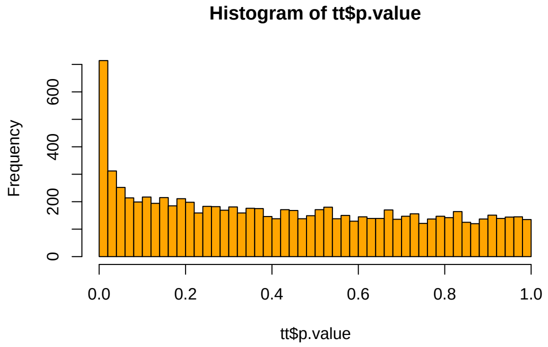

Figure 2: Histogram of *p*-values.

```
      1636_g_at        39730_at         1635_at         1674_at        40504_at
         "ABL1"          "ABL1"          "ABL1"          "YES1"          "PON2"
       37015_at        40202_at        32434_at        37027_at      39837_s_at
      "ALDH1A1"          "KLF9"        "MARCKS"         "AHNAK"        "ZNF467"
       41274_at      40167_s_at        37403_at      40480_s_at        41815_at
"DKFZp667G2110"          "WSB2"         "ANXA1"           "FYN"         "SYNE2"
       33774_at        36591_at        37363_at        39631_at        34472_at
        "CASP8"         "TUBA1"         "MTSS1"          "EMP2"          "FZD6"
```

The version number of R and packages loaded for generating this document are:

```
Version 2.3.1 Patched (2006-06-08 r38315)
powerpc-apple-darwin8.6.0

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

other attached packages:
  hgu95av2        ALL genefilter   survival    Biobase
  "1.12.0"    "1.2.1"   "1.10.1"     "2.26"   "1.10.0"
```