Solutions for chapter Gene Set Enrichment Analysis

Exercise 1

a There are 79 samples, and the table below shows how many of each type.

```
> table(ALLfilt_bcrneg$mol.biol)
BCR/ABL     NEG
     37      42
```

b 4502 distinct genes (as determined by Entrez Gene ID) have been selected.

Exercise 2

There is one gene set for each row of the incidence matrix, so there are 194 gene sets. We can find out how many gene sets have fewer than 10 genes by computing `rowSums(Am)`, so there are 76 gene sets. The largest number of gene sets a gene is in can be found by finding the largest of the column sums, which is 32 gene sets for `976_s_at`.

Exercise 3

There are 755 positive statistics, and 785 negative ones. There are 128 with p-values less than 0.01.

Exercise 4

You should notice that all of the points lie above the 45 degree line, indicating that they have higher y values than x values. Or, that the mean value in the NEG group is larger than the mean value in the BCR/ABL group.

Exercise 5

Although the mean plot seemed to suggest a strong separation between the two groups, we see from the heatmap that the distinction is not that clear.

The row in the heatmap that corresponds to the gene labeled `41214_at` indicates that the gene is on in some samples and off in others. It is on the Y chromosome, and hence we are seeing a pattern of expression that distinguishes the male samples from the females.

Exercise 6

```
> apply(pvals, 2, min)
Lower Upper
0.014 0.000
> rownames(pvals)[apply(pvals, 2, which.min)]
[1] "03010" "04510"
```

Exercise 7

To obtain the p-values from the permutation approach we must obtain the rowwise minima of the permutation p-values. The p-values for the parametric approach can be obtained by calling `pnorm`, and then taking the smaller of the observed p-value or one minus it.

```
> permpvs  = pmin(pvals[,1], pvals[,2])
> pvsparam = pnorm(tAadj)
> pvspara  = pmin(pvsparam, 1-pvsparam)
```

```
> plot(permpvs, pvspara, xlab="Permutation p-values",
       ylab="Parametric p-values")
```

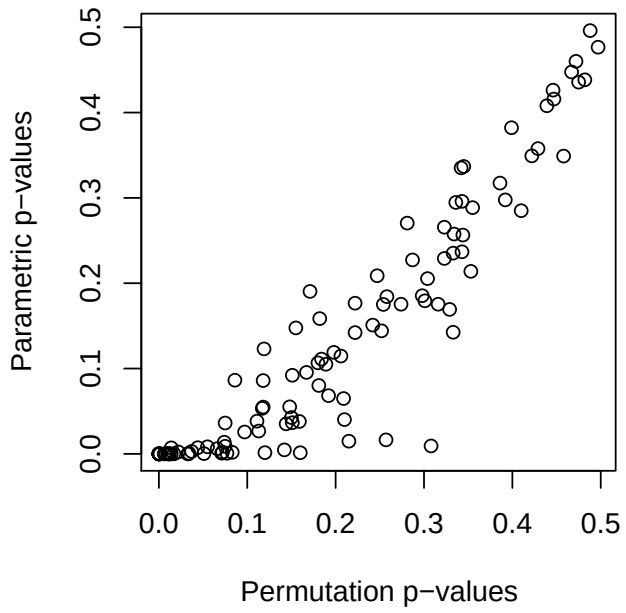

Figure 1. Scatter plot comparing the permutation p-values to those obtained from using a Normal approximation.

Exercise 8

It indicates that the gene is on the p arm of chromosome 17 in band 3, subband 3, subsubband 2.

Exercise 9

```
> ## depending on which annotation infrastructure we use
> ## hgu95av2MAP will either be an environment or an
> ## AnnDbBimap object
> fnames = featureNames(ALLfilt_bcrneg)
> if(is(hgu95av2MAP, "environment")){
      chrLocs = mget(fnames, hgu95av2MAP)
      mapping = names(chrLocs[sapply(chrLocs,
          function(x) !all(is.na(x)))])
  }else{
      mapping = toTable(hgu95av2MAP[fnames])$probe_id
  }
> psWithMAP = unique(mapping)
> nsF2 = ALLfilt_bcrneg[psWithMAP, ]
```

Exercise 10

The value returned by `MAPAmat` is a matrix where the rows are the chromosome bands and the columns are the genes. So there are 4495 genes and 1055 map positions.

```
> dim(chrMat)
[1] 1055 4495
```

Exercise 11

```
> chrMat = chrMat[rowSums(chrMat) >= 5, ]
> dim(chrMat)
[1]  519 4495
```

Exercise 12

```
> EGlist = mget(featureNames(nsF2),  hgu95av2ENTREZID)
> EGIDs = sapply(EGlist, "[", 1)
> idx = match(EGIDs, colnames(chrMat))
> chrMat = chrMat[, idx]
```

Now you can simply repeat the analysis using the new incidence matrix `chrMat`.

Exercise 13

```
> rowSums(Ams)[c("04510", "04512", "04514", "04940")]
04510 04512 04514 04940
   89    25    50    26
> Amx["04512", "04510"]
[1] 16
> Amx["04940", "04514"]
[1] 16
```

Exercise 14

Example code

```
> P04514 = Ams["04514",]
> P04940 = Ams["04940",]
> P04514.Only = ifelse(P04514 != 0 & P04940 == 0, 1, 0)
> P04940.Only = ifelse(P04514 == 0 & P04940 != 0, 1, 0)
> Both        = ifelse(P04514 != 0 & P04940 != 0, 1, 0)
> lm5 = lm(rttStat ~ P04514.Only + P04940.Only + Both)
> summary(lm5)
Call:
lm(formula = rttStat ~ P04514.Only + P04940.Only + Both)

Residuals:
   Min     1Q Median     3Q    Max
-4.206 -1.074 -0.162  0.886  7.172

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1081     0.0401    2.70   0.0071 **
P04514.Only   0.7493     0.2675    2.80   0.0052 **
P04940.Only   0.7017     0.4894    1.43   0.1518
Both          0.8387     0.3877    2.16   0.0307 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.54 on 1536 degrees of freedom
Multiple R-squared: 0.00919,        Adjusted R-squared: 0.00726
F-statistic: 4.75 on 3 and 1536 DF,  p-value: 0.00266
```

The answer is a bit less clear here. Genes only in `04514` give an extremely small p-value, whereas those in both and those only in `04940` have a lesser effect.