Exercise 1

- Make sure you have the latest version of R (3.0.2)
- Install Bioconductor default and the LMGene package
- Download the 12 .CEL files and read them into an AffyBatch object with ReadAffy
- Summarize the probe sets with rma. This also transforms and normalizes the arrays.
- Find the significant probe sets using both the gene specific and the posterior p-values and in both cases find the FDR-adjusted p-values.

```
library(affy)
rrdata <- ReadAffy()

> class(rrdata)
[1] "AffyBatch"
attr(,"package")
[1] "affy“

> dim(exprs(rrdata))
[1] 409600     12

> colnames(exprs(rrdata))
 [1] "LN0A.CEL" "LN0B.CEL" "LN1A.CEL" "LN1B.CEL" "LN2A.CEL" "LN2B.CEL"
 [7] "LN3A.CEL" "LN3B.CEL" "LN4A.CEL" "LN4B.CEL" "LN5A.CEL" "LN5B.CEL"
```

```
> eset <- rma(rrdata)
trying URL 'http://bioconductor.org/packages/2.1/…
Content type 'application/zip' length 1352776 bytes (1.3 Mb)
opened URL
downloaded 1.3 Mb

package 'hgu95av2cdf' successfully unpacked and MD5 sums checked

The downloaded packages are in
        C:\Documents and Settings\dmrocke\Local Settings…
updating HTML package descriptions
Background correcting
Normalizing
Calculating Expression

> class(eset)
[1] "ExpressionSet"
attr(,"package")
[1] "Biobase"
> dim(exprs(eset))
[1] 12625    12
```

```
> summary(exprs(eset))
    LN0A.CEL          LN0B.CEL          LN1A.CEL          LN1B.CEL
 Min.   : 2.713   Min.   : 2.585   Min.   : 2.611   Min.   : 2.636
 1st Qu.: 4.478   1st Qu.: 4.449   1st Qu.: 4.458   1st Qu.: 4.477
 Median : 6.080   Median : 6.072   Median : 6.070   Median : 6.078
 Mean   : 6.120   Mean   : 6.124   Mean   : 6.120   Mean   : 6.128
 3rd Qu.: 7.443   3rd Qu.: 7.473   3rd Qu.: 7.467   3rd Qu.: 7.467
 Max.   :12.042   Max.   :12.146   Max.   :12.122   Max.   :11.889
    LN2A.CEL          LN2B.CEL          LN3A.CEL          LN3B.CEL
 Min.   : 2.598   Min.   : 2.717   Min.   : 2.633   Min.   : 2.622
 1st Qu.: 4.444   1st Qu.: 4.469   1st Qu.: 4.425   1st Qu.: 4.428
 Median : 6.008   Median : 6.058   Median : 6.017   Median : 6.028
 Mean   : 6.109   Mean   : 6.125   Mean   : 6.116   Mean   : 6.117
 3rd Qu.: 7.426   3rd Qu.: 7.422   3rd Qu.: 7.444   3rd Qu.: 7.459
 Max.   :13.135   Max.   :13.110   Max.   :13.106   Max.   :13.138
    LN4A.CEL          LN4B.CEL          LN5A.CEL          LN5B.CEL
 Min.   : 2.742   Min.   : 2.634   Min.   : 2.615   Min.   : 2.590
 1st Qu.: 4.468   1st Qu.: 4.433   1st Qu.: 4.448   1st Qu.: 4.487
 Median : 6.074   Median : 6.050   Median : 6.053   Median : 6.068
 Mean   : 6.122   Mean   : 6.120   Mean   : 6.121   Mean   : 6.123
 3rd Qu.: 7.460   3rd Qu.: 7.478   3rd Qu.: 7.477   3rd Qu.: 7.457
 Max.   :12.033   Max.   :12.162   Max.   :11.925   Max.   :11.952
```

```
> dim(exprs(eset))
[1] 12625    12
> group <- as.factor(c(0,0,1,1,2,2,3,3,4,4,5,5))
> group
 [1] 0 0 1 1 2 2 3 3 4 4 5 5
Levels: 0 1 2 3 4 5
> anova(lm(exprs(eset)[942,] ~ group))
Analysis of Variance Table

Response: exprs(eset)[942, ]
          Df Sum Sq Mean Sq F value   Pr(>F)
group      5 3.7235  0.7447  10.726 0.005945 **
Residuals  6 0.4166  0.0694
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> colnames(exprs(eset))
 [1] "LN0A.CEL" "LN0B.CEL" "LN1A.CEL" "LN1B.CEL" "LN2A.CEL" "LN2B.CEL"
 [7] "LN3A.CEL" "LN3B.CEL" "LN4A.CEL" "LN4B.CEL" "LN5A.CEL" "LN5B.CEL"

> group <- factor(c(0,0,1,1,2,2,3,3,4,4,5,5))
> vlist <- list(group=group)
> vlist
$group
 [1] 0 0 1 1 2 2 3 3 4 4 5 5
Levels: 0 1 2 3 4 5

> eset.lmg <- neweS(exprs(eset),vlist)
```

```
> genediff.results <- genediff(eset.lmg)
> names(genediff.results)
[1] "Gene.Specific" "Posterior"
> hist(genediff.results$Gene.Specific)
> hist(genediff.results$Posterior)
> pv2 <- pvadjust(genediff.results)
> names(pv2)
[1] "Gene.Specific"     "Posterior"          "Gene.Specific.FDR"
[4] "Posterior.FDR"
> sum(pv2$Gene.Specific < .05)
[1] 2615
> sum(pv2$Posterior < .05)
[1] 3082
> sum(pv2$Gene.Specific.FDR < .05)
[1] 119
> sum(pv2$Posterior.FDR < .05)
[1] 1173
```

Using genediff results in two lists of 12625 p-values. One uses the standard 6df denominator and the other uses the moderated F-statistic with a denominator derived from an analysis of all of the MSE's from all the linear models.

```
> genediff.results <- genediff(eset.lmg)
> class(genediff.results)
[1] "list"
> length(genediff.results)
[1] 2
> names(genediff.results)
[1] "Gene.Specific" "Posterior"
> length(genediff.results$Posterior)
[1] 12625
> genediff.results$Posterior[1:5]
[1] 0.02939858 0.07524596 0.34409615 0.26903574 0.19198230
> sort(genediff.results$Posterior)[1:5]
[1] 5.939886e-08 2.059263e-07 2.257690e-07 2.429635e-07 2.750870e-07
> order(genediff.results$Posterior)[1:5]
[1]  4343  7278 12607  7030  8691
> genediff.results$Posterior[order(genediff.results$Posterior)[1:5]]
[1] 5.939886e-08 2.059263e-07 2.257690e-07 2.429635e-07 2.750870e-07
```

```
> featureNames(eset.lmg)[1:5]
[1] "100_g_at"  "1000_at"   "1001_at"   "1002_f_at" "1003_s_at"
> featureNames(eset.lmg)[order(genediff.results$Posterior)[1:5]]
[1] "34301_r_at"       "37208_at"         "AFFX-M27830_5_at" "36962_at"
[5] "38608_at"
> featureNames(eset.lmg)[order(genediff.results$Posterior)[1:10]]
 [1] "34301_r_at"       "37208_at"         "AFFX-M27830_5_at" "36962_at"
 [5] "38608_at"         "33646_g_at"       "2027_at"          "31957_r_at"
 [9] "34592_at"         "256_s_at"
```

Feature	Gene	Feature	Gene
34301_r_at	KRT17	33646_g_at	GM2A
37208_at	PSPH1	2027_at	S100A2
AFFX-M27830_5_at	Endogenous control	31957_r_at	RPLP1
36962_at	COPA	34592_at	RPS17
38608_at	LGALS7B	256_s_at	RPSA

For Example, RPSA
40S ribosomal protein SA

GO:0006412 : translation
993 Gene Products
Biological Process
IEA

GO:0015935 : small ribosomal subunit
89 Gene Products
Cellular Component
IEA

GO:0003735 : structural constituent of ribosome
474 Gene Products
Molecular Function
IEA

Exercise 2

- Using the AD data, we will try to improve the performance of our analysis.
- First, the scales of the analytes are arbitrary, and even if they were calibrated, the actual amounts don't matter, just the relative values of each analyte across samples.
- Separate out the first column, which is diagnosis, from the remaining 124 columns which carry assay data. We will have then a 105 by 124 matrix of values.
- Normalize each analyte to a common median of 50. To do this, write a small program to take a vector, determine the median M of the vector, and multiply each element of the vector by 50/M. The vector will then have median 50. Now use the apply() function to do this to each column of the data matrix.

- Take logs of the data. None of them will be negative, but we may need to go back later and use a started log instead.
- Now we want to normalize the samples (rows) so that they all have the same median. Find the mean of the medians of the samples (rows), and using a slight variant of the previous procedure, normalize the matrix so that the median across analytes is the same for each sample.
- Use the tools of LMGene to investigate which analytes are most related to the diagnosis categories.

```
mnorm <- function(vec1,val1)
{
  vec2 <- vec1*val1/median(vec1)
  return(vec2)
}
> source(mnorm.R

> ad.data <- read.csv("AD-Luminex.csv")
> names(ad.data)
  [1] "Diagnosis"                                  "ACE..CD143."
  [3] "ACTH"                                       "Adiponectin"
  [5] "Agouti.Related.Protein..AgRP."              "Alpha.1.Antitrypsin"
                    ........
> diag <- ad.data[,1]
> admat0 <- ad.data[,-1]
> dim(admat0)
[1] 104 124                  104 subjects (rows) and 124 analytes (cols)
> summary(apply(admat0,2,median))
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
    0.026     2.240    23.950   253.700    96.910 14450.000
> admat1 <- apply(admat0,2,mnorm,val1=50)
> dim(admat1)
[1] 104 124
> summary(apply(admat1,2,median))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     50      50      50      50      50      50
```

```
> summary(apply(admat1,1,median))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  38.45   46.14   50.06   50.28   53.18   64.29
> med.all <- mean(apply(admat1,1,median))
> admat2 <- apply(admat1,1,mnorm,val1=med.all)
> dim(admat2)
[1] 124 104                                          transposed, but we want
> summary(apply(admat2,1,median))                    it that way
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  46.06   49.65   50.33   50.30   51.10   53.55
> summary(apply(admat2,2,median))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  50.28   50.28   50.28   50.28   50.28   50.28
> admat.log <- log(admat2)
> vlist <- list(diagnosis=diag)
> library(LMGene)
> ad.eset <- neweS(admat.log,vlist)
> pv1 <- genediff(ad.eset)
Prior d.f. =  2.591863
Prior mean reciprocal precision =  0.1769471
> source("pvadjust.R")
> pv2 <- pvadjust(pv1)
```

```
> sum(pv2$Gene.Specific < .05)
[1] 51
> sum(pv2$Posterior < .05)
[1] 51
> sum(pv2$Gene.Specific.FDR < .1)
[1] 43
> sum(pv2$Posterior.FDR < .1)
[1] 45
> rownames(admat.log)[pv2$Gene.Specific.FDR < .1]
 [1] "Agouti.Related.Protein..AgRP."      "Alpha.1.Antitrypsin"
 [3] "Alpha.Fetoprotein"                  "ApoAI"
 [5] "ApoB"                               "ApoCIII"
 [7] "ApoE"                               "ApoH"
 [9] "ApoJ"                               "Apolipoprotein.A1"
[11] "Apolipoprotein.CIII"                "Apolipoprotein.H"
[13] "ASP"                                "Brain.Derived.Neurotrophic.Factor"
[15] "C.Reactive.Protein"                 "Calcitonin"
[17] "CD40.Ligand"                        "Cortisol"
[19] "EGF"                                "ENA.78"
[21] "Fatty.Acid.Binding.Protein"         "Fibrinogen"
[23] "FSH"                                "GLP.1.active"
[25] "GRO.alpha"                          "HGF"
[27] "IGF.1"                              "IL.18"
[29] "MCP.3"                              "MIF"
[31] "MIP.1.alpha"                        "MIP.1beta"
[33] "PAI.1"                              "PDGF"
[35] "PDGF.BB"                            "Prostate.Specific.Antigen..Free"
[37] "Prostatic.Acid.Phosphatase"         "Pulmonary.and.Activation.Regulated.Chemokine..PARC."
[39] "RANTES"                             "SGOT"
[41] "SHBG"                               "Stem.Cell.Factor"
[43] "Testosterone"
```

```
> sum(pv2$Gene.Specific < .05)
[1] 51
> sum(pv2$Posterior < .05)
[1] 51
> sum(pv2$Gene.Specific.FDR < .1)
[1] 43
> sum(pv2$Posterior.FDR < .1)
[1] 45
> rownames(admat.log)[pv2$Gene.Specific.FDR < .1]
 [1] "Agouti.Related.Protein..AgRP."      "Alpha.1.Antitrypsin"
 [3] "Alpha.Fetoprotein"                  "ApoAI"
 [5] "ApoB"                               "ApoCIII"
 [7] "ApoE"                               "ApoH"
 [9] "ApoJ"                               "Apolipoprotein.A1"
[11] "Apolipoprotein.CIII"                "Apolipoprotein.H"
[13] "ASP"                                "Brain.Derived.Neurotrophic.Factor"
[15] "C.Reactive.Protein"                 "Calcitonin"
[17] "CD40.Ligand"                        "Cortisol"
[19] "EGF"                                "ENA.78"
[21] "Fatty.Acid.Binding.Protein"         "Fibrinogen"
[23] "FSH"                                "GLP.1.active"
[25] "GRO.alpha"                          "HGF"
[27] "IGF.1"                              "IL.18"
[29] "MCP.3"                              "MIF"
[31] "MIP.1.alpha"                        "MIP.1beta"
[33] "PAI.1"                              "PDGF"
[35] "PDGF.BB"                            "Prostate.Specific.Antigen..Free"
[37] "Prostatic.Acid.Phosphatase"         "Pulmonary.and.Activation.Regulated.Chemokine..PARC."
[39] "RANTES"                             "SGOT"
[41] "SHBG"                               "Stem.Cell.Factor"
[43] "Testosterone"
```