We will now perform some statistical analyses to the normalized, adjusted data. Note that you can use either RUV2-adjusted or RUV4-adjusted data, whichever you think is better. However, in the lab note, I will be assuming that RUV2-adjusted data is used.

1. Read in the datasets and the workspace from Lab 3 using the following command:

```
#datasets
load(file= paste('Datasets_','YOURNAME','.Rdata',sep=""))
# read workspace from previous lab
filaneme1 <- paste('Lab3_','YOURNAME','.Rdata',sep="")
load(filename1)
```

2. During the last lab, we investigated the p-values associated with season of birth and AFB1 levels using adjusted and unadjusted (normalized) data. Now, we will apply the empirical Bayes (EB) method to the t-statistic associated with season of birth and AFB1 levels using adjusted data. We will also investigate the impact of applying EB method to probes with small variance.

```
#apply EB to the adjusted data
ruv2.adj.test <- variance_adjust(ruv2.adj)
# extract the FDR-adjusted p-value, after and before applying EB
pval.adj.aft <- ruv2.adj.test$p.ebayes.BH
pval.adj.bef <- ruv2.adj.test$p.BH
# calculate log SD for each probe and plot against change in pvalues for AFB1
log.sd <- log( apply(y.norm,1,sd))
delta.log10.pval <- log10(pval.adj.bef[2,])-log10(pval.adj.aft[2,])
plot(log.sd,delta.log10.pval,cex=0.1)
```

You should see a pattern where the change in log10 pvalues are negative for probes with small SD; this is an effect of applying EB method where evidence from probes with small variance will be down-weighted as it is more likely that this evidence is 'false'.

3. We can also perform regression analysis with omics data. Here, we will use only the probes with FDR-corrected p-value less than 0.20 after EB-adjustment. These probes will be used to predict the probability of high toxin levels (defined as AFB1 $\geq$ 50).

```
require(ruv)
#extract adjusted data
y.norm.adj <- y.norm - t(ruv2.adj$alpha) %*% t(ruv2.adj$W)
# probes for analysis with FDR < 0.20 for AFB1
y.select <- y.norm.adj[ (ruv2.adj.test$p.ebayes.BH[2,]) < 0.20,]
high.AFB1 <- info$AFB1>50

xvars <- t(y.select)
# perform logistic regression
model.logit <- glm(high.AFB1 ~ xvars, family='binomial')
summary(model.logit)

require(glmnet)
# find optimal lambda
find.lambda <- cv.glmnet(y=high.AFB1,x= xvars,nfolds=10)
# perform LASSO-logistic regression
model.logit.L1 <- glmnet(y=high.AFB1,x= xvars, family='binomial',alpha=1,
                    lambda=find.lambda$lambda.min)
```

```
# compare predicte probability
predict.prob <- predict(model.logit.L1,newx=xvars,type='response')
boxplot(split(predict.prob,high.AFB1))
```

4. Which probes have nonzero coefficients in the LASSO logistic regression? Can you find the UCSC gene names overlapping these probes? (HINT: you need to perform matching with the appropriate column in `anno` data frame. Once you located the implicated genes, you can learn about the genes by looking up from NCBI Gene Card (`http://www.ncbi.nlm.nih.gov/gene/`).

5. When you finished don't forget to save the commands you have been executing and the workspace containing all your works, but first you need to remove the three datasets `data,anno,info` so that the workspace size is not overwhelming. The following commands can be used:

```
# remove big datasets
rm(data);rm(anno);rm(info)
# save commands
filaneme1 <- paste('Lab4_','YOURNAME','.R',sep="")
savehistory(filename1)
# save workspace
filaneme2 <- paste('Lab4_','YOURNAME','.RData',sep="")
save.image(filename2)
```