Gene expression dataset for **VANGUARD** prospective study - **(ADJ vs. CONTRALATERAL)**

Li Shen, Jing Wang, and Kevin R. Coombes
Dept. of Bioinformatics and Computational Biology
The University of Texas M.D. Anderson Cancer Center (MDACC)

December 1, 2011

Contents

1 Executive Summary

1.1 Statistical Methods

In this report we perform feature-by-feature `Paired T-test` for each patients to compare the site effects between (ADJ vs. CONTRA). The distributions of p-values are shown in Figure 1. The beta-uniform mixture (BUM) model, described by Pounds and Morris (*Bioinformatics*, 2003; **19**:1236–42), was used to control false discovery rate (FDR).

1.2 Results

We performed the analyses described in the Statistical Methods section. The results are summarized in the following figures and tables:

1. Figure 1 shows the distributions of p-values from the feature-by-feature Paired t-tests.
2. Table 1 to summarize the numbers of significant features using different FDR cutoff values for **ADJ vs. CONTRA**.
3. Figure 2 shows a heatmap using the top 238 most differentially expressed genes selected by FDR 0.05.
4. Figure 3 shows a scatter plot of the first two principal components using the top 238 most differentially expressed genes selected by FDR 0.05.

List of Tables

List of Figures

2 Loading the Data

2.1 Setting working directory

We first set up directory we use for the analysis.

```
> ### working directory ###
> getwd()

[1] "/data/bioinfo2/Lung-HN/Wistuba-VANGUARD/Analysis"

>
```

2.2 R Libraries

We begin by loading all the libraries we will need for this analysis. A list of the current versions of the libraries used for the analysis can be found in the appendix.

```
> require(nlme)
> library(gplots)
> library(xtable)
> library(ClassComparison)
> library(ClassDiscovery)
> library(RColorBrewer)    # colorbrewer
> library(Hmisc)    # latex function (\usepackage{longtable})
```

2.3 Processed Data

We now load the processed gene expression data.

```
> ### load processed data ###
> load(file="gExpr-RMA-Aroma.RData")
>
```

2.3.1 Data Description

The processed data is stored in a data frame or matrix of the following dimensions:

```
> ### datasets ###
> normData <- as.matrix(normData)
> dim(normData) # normalized data

[1] 33252   391
```

```
>
> # identical(rownames(annot), rownames(normData))
> # out.1 <- data.frame(ProbeID=rownames(normData), normData, annot)
> # write.csv(out.1, file="VANGUARD-normData.csv", row.names = FALSE)
>
```

The columns of this data matrix represent samples and the rows represent the features. All relevant clinical covariates are stored in a data frame of the following dimensions:

```
> ### Sample Info ###
> dim(si)

[1] 391  61

>
```

The rows of the data frame of covariates (Group column in sampleInfo file) represent samples and each column represent a different covariate. Thus, the rows of the covariate matrix must match the columns of the data matrix. Before proceeding, we confirm this agreement.

```
> nrow(si) == ncol(normData)

[1] TRUE

> all(rownames(si)==colnames(normData))

[1] TRUE

>
```

We are interested in differential expression between the groups defined by the covariate column labeled 'FACTOR'. The folowing table lists the group sizes. (Note that, if there are more than two groups, then the default behavior is to compare the first group to everything else.

```
> ### factors in the model ###
> with(si, table(Batch))

Batch
  I  II
 71 320

> with(si, table(Case.ID=V_Case.ID.Inclusion_number.))

Case.ID
 1  3  6 10 16 18 20 23 30 31 35 38 40 41 44 46 47 48 50
25 23 24 18 15 23 23 17 23 24 17 23 24 18 24 17 18 17 18
```

```
> with(si, table(Time.point))

Time.point
  0  12  24  36
109 108 113  61

> with(si, table(Map))

Map
            ADJ      MC NON-ADJ
      1      62      60     268

> with(si, table(DetMap))

DetMap
            ADJ  CONTRA      MC NON-ADJ
      1      62     161      60     107

> with(si, table(Batch, DetMap))

     DetMap
Batch     ADJ CONTRA  MC NON-ADJ
   I    0  10     16  34      11
   II   1  52    145  26      96

>
> #with(si, table(Time.point, DetMap,
> #      Case.ID=V_Case.ID.Inclusion_number.))
```

2.4 Gene/Probe Annotations

Here we load the annotations for the probes on the ST 1.0 array.

```
> load(file=file.path("RNW", "HG-1ST-annot.Rda"))
>
> #annot <- read.csv(file.path("RNW",
> #                            "Human Gene 1.0 ST annotations for Li Shen.csv"),
> #                  header=TRUE, as.is=TRUE, na.strings=c("NA","Un", ""),
> #                  row.names=1)
> #all(rownames(normData) %in% rownames(annot))
> #annot <- annot[rownames(normData),]
> #annot$Symbol <- factor(annot$Symbol)
> #annot$UGCluster <- factor(annot$UGCluster)
> #annot$Chromosome <- factor(annot$Chromosome, levels=c(1:22, "X", "Y"))
```

```
> #annot$Cytoband <- factor(annot$Cytoband)
> #summary(annot)
> #
> #gene.label <- rep("NA", nrow(annot))
> #for (gene in 1:nrow(normData)) {
> #    gene.label[gene] <- ifelse(is.na(annot[gene, "Symbol"]), rownames(annot)[gene],
> #                           as.character(annot[gene, "Symbol"]))
> #}
> #
> #annot <- cbind(annot, gene.label)
> #save("annot", file=file.path("RNW", "HG-1ST-annot.Rda"))
>
```

2.5 Clinical Data

Now we clean up the Sample Information data.

```
> si$Batch <- factor(si$Batch)
> si$Gender <- factor(si$Gender)
> si$Diagnosis..Histology. <- factor(si$Diagnosis..Histology.)
> colnames(si)[19] <- "Histology"
> si$Off.study_Reason <- factor(si$Off.study_Reason)
> si$Differentiation <- factor(si$Differentiation)
> si$Leison.Site <- factor(si$Leison.Site)
> si$Anatomical_site <- factor(si$Anatomical_site)
> si$site.of.collection <- factor(si$site.of.collection)
> si$Contralateral <- factor(si$Contralateral)
> dmlev <- c("ADJ", "NON-ADJ", "MC", "CONTRA")
> si$DetMap <- factor(si$DetMap, levels=dmlev)
> mlev <- c("ADJ", "NON-ADJ", "MC")
> si$Map <- factor(si$Map, levels=mlev)
> #si$Time.point <- as.numeric(si$Time.point)
> si$Code.4.time.point <- factor(si$Code.4.time.point)
> si$Code.4.Site.of.collection <- factor(si$Code.4.Site.of.collection)
> si$pT <- factor(si$pT)
> si$pN <- factor(si$pN)
> si$Final.Pat.Stage <- factor(si$Final.Pat.Stage)
> si$EGFR.status <- factor(si$EGFR.status)
> si$KRAS.status <- factor(si$KRAS.status)
> si$V_Case.ID.Inclusion_number. <- factor(paste("P",
+                                                si$V_Case.ID.Inclusion_number.,
+                                                sep=''))
> colnames(si)[8] <- "Case"
```

```
> si$MRN..MDAH. <- factor(si$MRN..MDAH.)
> simplify <- c(2, 8, 13,  27, 31, 34, 12, 14:20, 22, 24, 43:47)
> rm(dmlev, mlev)
> summary(si[, simplify])

 Batch         Case          DetMap         Map        Time.point
 I : 71   P1     : 25   ADJ    : 62   ADJ    : 62   Min.   : 0.00
 II:320   P31    : 24   NON-ADJ:107   NON-ADJ:268   1st Qu.: 0.00
          P40    : 24   MC     : 60   MC     : 60   Median :12.00
          P44    : 24   CONTRA :161   NA's   :  1    Mean   :15.87
          P6     : 24   NA's   :  1                  3rd Qu.:24.00
          P18    : 23                                Max.   :36.00
          (Other):247
          Off.study_Reason Gender  DOB..DOBirth.       DOSurgery
                  : 17     F:134   Length:391         Length:391
 N                :332     M:257   Class :character   Class :character
 Y_died           : 25             Mode  :character   Mode  :character
 Y_recurrence lung: 17



 DOInclusion          Surgery                    Histology   Differentiation
 Length:391         Length:391         Adenocarcinoma:309   MOD     :180
 Class :character   Class :character   Squamous      : 82   MOD-POOR: 24
 Mode  :character   Mode  :character                       POOR    : 35
                                                          W       : 42
                                                          WELL    : 23
                                                          NA's    : 87

 Anatomical_site Contralateral    pT      pN      Final.Pat.Stage      EGFR.status
 LLL: 58                :  1    1A  :203   0:366   I  : 23         MUT del E19: 18
 LUL:102         CONTRA:161    1B  : 43   1: 25   IA :221         WT         : 71
 RLL: 80         IPSI  :169    2A  : 98           IB : 98         NA's       :302
 RML: 17         MC    : 60    2B  : 24           IIA: 49
 RUL:134                       NA's: 23


     KRAS.status
 MUT cod 12: 24
 WT        : 65
 NA's      :302

>
```

```
> tmp <- si[match(levels(as.factor(si$MRN)), si$MRN), c(8,13)]
> tmp2 <- read.csv(file.path("RNW",
+                            "Copy of VANGUARD_06012011_BRONCHIAL BRUSHES.csv"),
+                  header=TRUE, as.is=TRUE, na.strings=c("NA","Un", ""),
+                  row.names="MDAH")
> ci <- data.frame(tmp, Event=tmp2[match(tmp$MRN, rownames(tmp2)),"Event"])
> ci <- ci[order(ci$Case),]
> rownames(ci) <- ci$Case
> Event.col <- rep("Suspicion", nrow(ci))
> Event.col[which(ci$Event=="YES")] <- "Recurrence"
> Event.col[which(ci$Event=="NO")]  <- "No"
> ci <- cbind(ci, Event.col)
> ci

    Case    Event  Event.col
P1    P1 YES Recurrence
P10  P10 YES Recurrence
P16  P16 NO         No
P18  P18 NO (suspicious RML nodule and RUL ground glass)  Suspicion
P20  P20 NO (suspicious LUL, RUL nodules, bilateral ground glass)  Suspicion
P23  P23 NO         No
P3    P3 NO         No
P30  P30 NO         No
P31  P31 NO         No
P35  P35 YES Recurrence
P38  P38 NO         No
P40  P40 NO         No
P41  P41 NO (suspicious ground glass nodule)  Suspicion
P44  P44 NO         No
P46  P46 NO         No
P47  P47 NO         No
P48  P48 NO         No
P50  P50 NO         No
P6    P6 YES Recurrence
```

Three patients originally suspicious for relapse, developed recurrence by the end of the study.

3 Result from Mixed Effect Models

We used a mixed-effects model to understand the expression patterns of each gene. Here we load in the result from Dr. Coombes' analysis.

```
> data.dir <- file.path(osbase, 'bioinfo2','Lung-HN',
+                       'KRC-Analyses')
> #f <- "modlist.rda"
> #load(file.path(data.dir, f))
> #rm(f)
```

3.1 Batch Correction

In order to generate additional plots, we need to adjust for the batch effects that are imposed on almost all genes. The information that we need to make this adjustment is already contained in the (fixed-effects) coefficients in the statistical models that we computed for each gene. For example,

```
> #x <- modlist[[1]]
> #fixef(x)
```

The next block of code extracts all of the fixed-effects coefficients from the statistical models.

```
> f <- "fixcoef.rda"
> load(file.path(data.dir, f))
> rm(f)
```

Now we use the batch coefficients to adjust the data.

```
> adjData <- normData
> temp <- sweep(adjData[, si$Batch=="I"], 1, fixcoef$BatchII, "+")
> adjData[, si$Batch=="I"] <- temp
>
> # identical(rownames(annot), rownames(adjData))
> # out.2 <- data.frame(ProbeID=rownames(adjData), adjData, annot)
> # write.csv(out.2, file="VANGUARD-adjData.csv", row.names = FALSE)
```

4 Differential Expression - Paired t-tests

4.1 ADJ vs. CONTRA

Here we only focus adjacent versus CONTRA airways.

```
> ### adjusted for Batch ###
> dataset <- data.frame(adjData)
> my.si <- si
> nrow(my.si)

[1] 391

> ### get ADJ ###
> keep.adj <- "ADJ"
> keep.idx <- which(my.si$DetMap %in% keep.adj)
> si.adj   <- my.si[keep.idx,]
> data.adj <- dataset[,keep.idx]
> nrow(si.adj)

[1] 62

> ncol(data.adj)

[1] 62

> group.mean2=apply(data.adj,1, function (i)
+ {tapply(i,as.factor(si.adj$Case), mean) })
> mean.adj=t(group.mean2)
> colnames(mean.adj)

 [1] "P1"  "P10" "P16" "P18" "P20" "P23" "P3"  "P30" "P31" "P35" "P38" "P40" "P41" "P44"
[15] "P46" "P47" "P48" "P50" "P6"

> ### get CONTRA ###
> keep.non <- "CONTRA"
> keep.idx <- which(my.si$DetMap %in% keep.non)
> si.non   <- my.si[keep.idx,]
> data.non <- dataset[,keep.idx]
> nrow(si.non)

[1] 161

> ncol(data.non)
```

```
[1] 161
```

```
> group.mean2=apply(data.non,1, function (i)
+ {tapply(i,as.factor(si.non$Case), mean) })
> mean.non=t(group.mean2)
> colnames(mean.non)
```

```
 [1] "P1"  "P10" "P16" "P18" "P20" "P23" "P3"  "P30" "P31" "P35" "P38" "P40" "P41" "P44"
[15] "P46" "P47" "P48" "P50" "P6"
```

```
> ### check ###
> all(colnames(mean.adj)==colnames(mean.non))
```

```
[1] TRUE
```

```
> all(rownames(mean.adj)==rownames(annot))
```

```
[1] TRUE
```

We can use above data matrix to perform the analysis.

```
> ### data matrix - dat ###
> colnames(mean.adj) <- paste("ADJ", colnames(mean.adj), sep=".")
> colnames(mean.non) <- paste("CON", colnames(mean.non), sep=".")
>
>
> #dat <- cbind(mean.adj, mean.non)
> #dim(dat)
> #colnames(dat)
> #cla <- factor(rep(c('ADJ', 'CONTRA'), 19))
> #pairing <- rep(1:19, each=2)
>
> #dat <- matrix(0, nrow(mean.adj), 38)
> #dat.cn <- rep("names", 38)
> #for (i1 in 1:19) {
> #    dat[,(i1*2)-1] <- mean.adj[,i1]
> #    dat[,(i1*2)]   <- mean.non[,i1]
> #    dat.cn[(i1*2)-1] <- colnames(mean.adj)[i1]
> #    dat.cn[(i1*2)]   <- colnames(mean.non)[i1]
> #}
> #colnames(dat) <- dat.cn
>
```

The purpose of this analysis is to get a list of features that are differentially expressed between ADJ and CONTRA. We fit Paired t-tests for each patient and check the difference for **Site** effects.

```
> pt.res <- matrix(0, nrow(mean.adj), 2)
> colnames(pt.res) <- c("p.value", "t.stat")
> rownames(pt.res) <- rownames(mean.adj)
> for (i1 in 1:nrow(mean.adj)){
+     a <- mean.adj[i1,]
+     b <- mean.non[i1,]
+     mtpt <- t.test(a,b, paired=TRUE, alternative = "two.sided")
+     pt.res[i1, "p.value"] <- mtpt$p.value
+     pt.res[i1, "t.stat"]  <- mtpt$statistic
+ }
>
```

Because we are performing many separate analysis (one per each row or feature of the data matrix), we must adjust for multiple testing. We make this adjustment by fitting a beta-uniform mixture (BUM) model to the set of feature-by-feature p-values. This model was first described by Pounds and Morris in *Bioinformatics*, 2003 Jul 1; **19**(10): 1236–42. After fitting the BUM model, we plot a histogram of the p-values (Figure 1).

```
> bum <- Bum(pt.res[,"p.value"])
```

From Figure 1, we note that there are many features showing significant differences between ADJ and CONTRA.

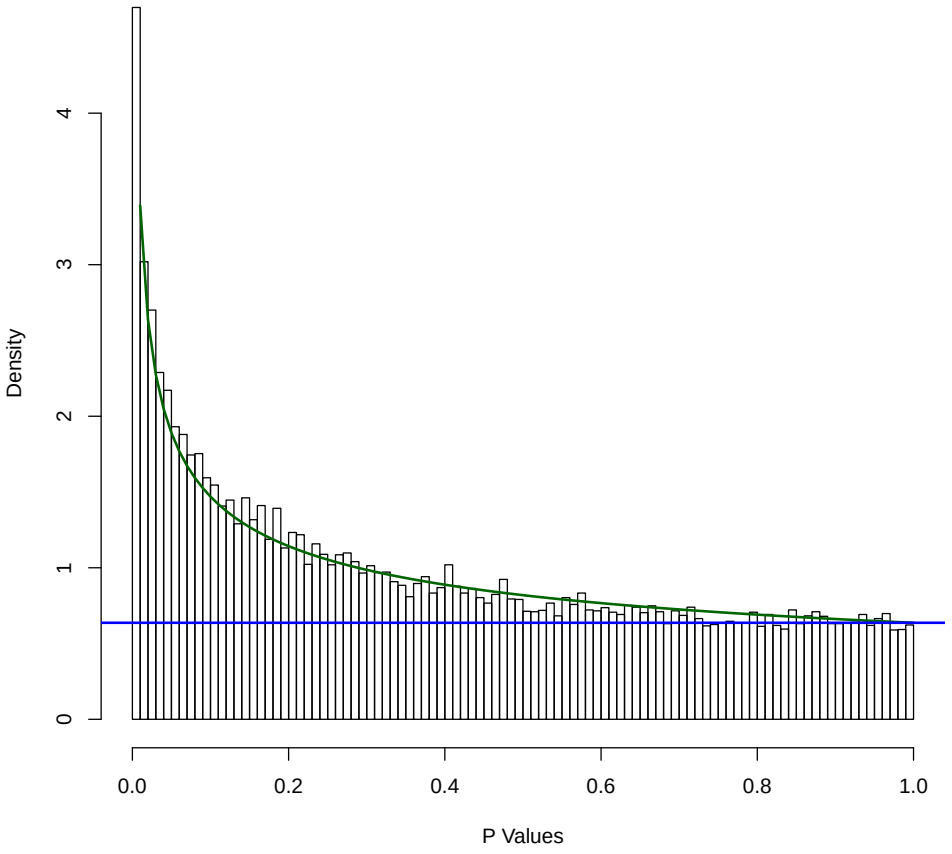

Figure 1: (**ADJ vs. CONTRA**) - Histogram showing the distributions of p-values from the feature-by-feature Paired t-tests. Superimposed curves represent the fit of a beta-uniform-mixture model.

5 FDR

We can use the BUM model to estimate the false discovery rate (FDR). Given a desired level for the FDR, we can compute a correponding cutoff on the p-values that achieves that FDR, and we can count the number of features that meet this selection criterion (Table 1).

```
> fdrs <- c(0.05, 0.10, 0.15, 0.20, 0.25, 0.30)
> counts <- sapply(fdrs, function(alpha) {
+   countSignificant(bum, alpha, by="FDR")
+ })
> cuts <- sapply(fdrs, function(alpha) {
+   cutoffSignificant(bum, alpha, by="FDR")
+ })
> selectors <- sapply(fdrs, function(alpha) {
+   selectSignificant(bum, alpha, by="FDR")
+ })
```

	FDR	Number Significant	P-value Cutoff
1	0.05	238	0.0009040
2	0.10	1074	0.0060996
3	0.15	2417	0.0186341
4	0.20	4311	0.0411553
5	0.25	6556	0.0760928
6	0.30	9163	0.1257278

Table 1: **ADJ vs. CONTRA** - Summary of significant features by different FDR cutoff values

Now we can generate a heatmap (Figure 2) to see the expression levels of the 238 most differentially expressed features selected at FDR 0.05 in difference of ADJ and CONTRA.

```
X11cairo
       2
```

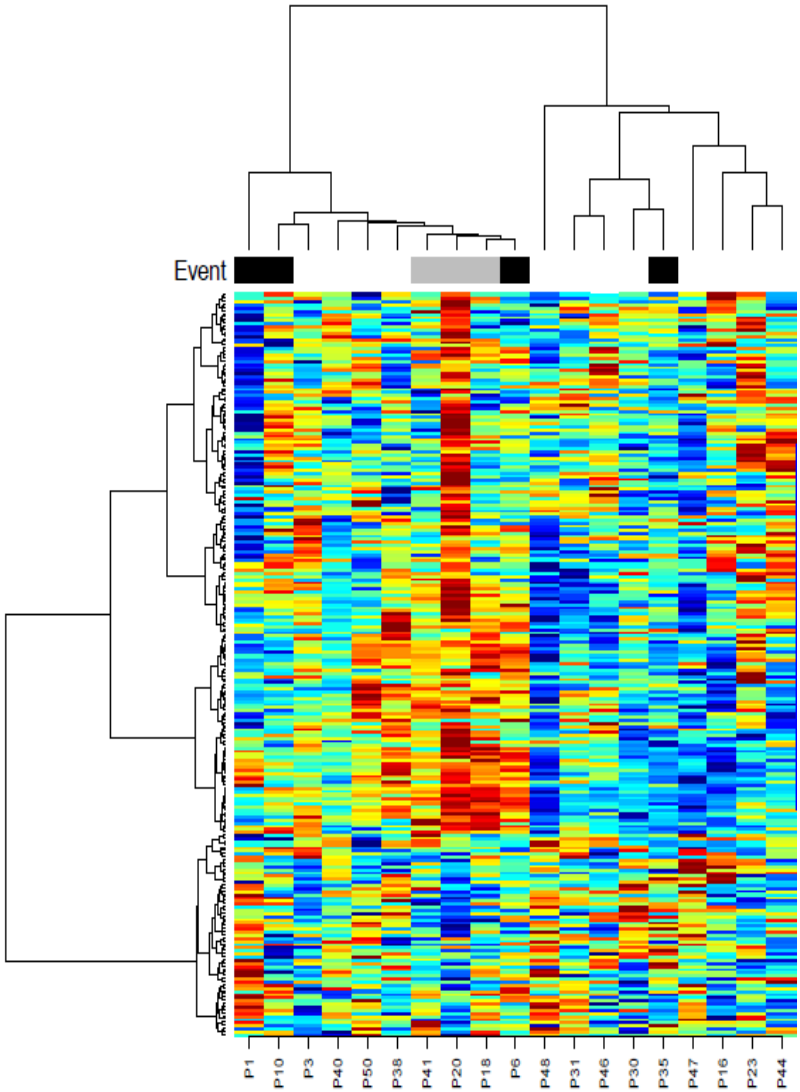

Figure 2: Heatmap of the top 238 gene selected by FDR 0.05. The expression values shown on the heatmap have been standardized and at ±2 standard deviations for display purpose (The scale of the values is indicated in the color key).

```
X11cairo
       2
```

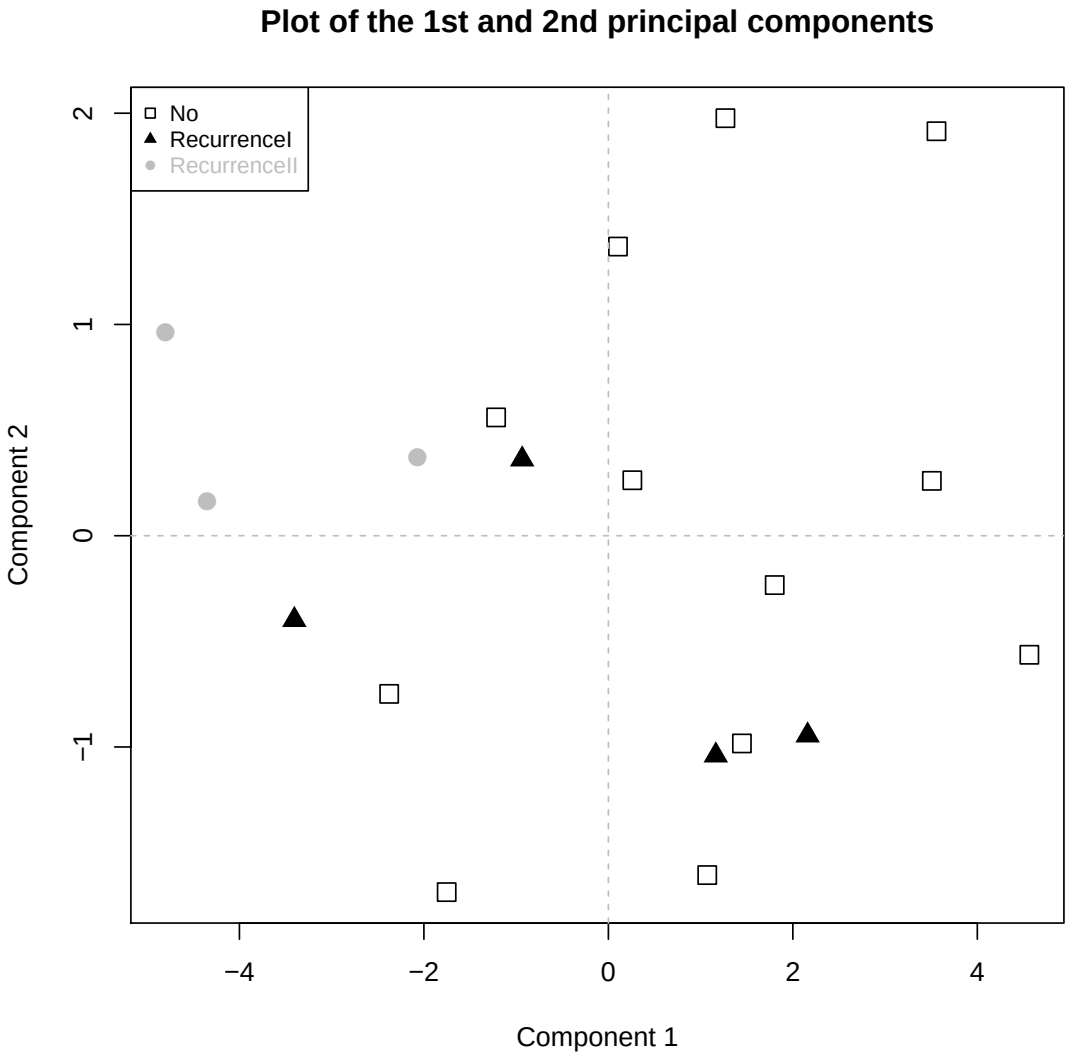

Figure 3: Scatter plot of the first two principal components using the top 238 most differentially expressed genes selected by FDR 0.05.

Appendices

A References

References

B Saving the Processed Data

The results of this analysis are saved in the following file.

```
> save.image("05-4_VANGUARD_ADJvsCONTRA-batchCorrection.RData")
```

C File Location

This analysis was run in the following directory:

```
> getwd()
```

```
[1] "/data/bioinfo2/Lung-HN/Wistuba-VANGUARD/Analysis"
```

D SessionInfo

This analysis was run in the following software environment:

```
> sessionInfo()
```

```
R version 2.14.0 (2011-10-31)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C               LC_TIME=en_US.UTF-8
 [4] LC_COLLATE=en_US.UTF-8     LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=C                 LC_NAME=C                  LC_ADDRESS=C
[10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] splines   grid      stats     graphics  grDevices utils     datasets  methods
[9] base

other attached packages:
 [1] Hmisc_3.9-0            survival_2.36-10       RColorBrewer_1.0-5
 [4] ClassDiscovery_2.10.2  mclust_3.4.10          cluster_1.14.1
 [7] ClassComparison_2.10.1 PreProcess_2.10.1      oompaBase_2.12.0
```

```
[10] Biobase_2.14.0         xtable_1.6-0           gplots_2.10.1
[13] KernSmooth_2.23-7      caTools_1.12           bitops_1.0-4.1
[16] gdata_2.8.2            gtools_2.6.2           nlme_3.1-102

loaded via a namespace (and not attached):
[1] lattice_0.20-0 tools_2.14.0
```