Lung Spore Analysis Report: Affymetrix Data Analysis

Lixia Diao, Jing Wang and Kevin R. Coombes

September 11, 2013

Contents

List of Figures

List of Tables

1 Executive Summary

1.1 Introduction

This report describes the analysis of a data set from Dr. Milind Suraokar, a member of the laboratory of PI Dr. Ignacio I. Wistuba. This dataset was acquired using Affymetrix HG-U133Plus2.0. There are 96 mesothelioma tissue arrays in total.

1.1.1 Aims/Objectives

In the series of the reports, we are interested in the probesets that show different between normal and tumor tissues. For the previous analysis we have done on the clean data, the separation of groups is very fragile corresponding to the selected length of genes or including number of samples. In this report, we find the groups defined most stably.

1.2 Methods

1.2.1 Description of the Data

The data are processed by Affymetrix HGU133plus2.0. There are 96 arrays in total. Among them, 41 are coded as matched tumor and normal. Another 14 arrays are no paired tumor tissues. Among those 41 matched arrays, Dr. Suraokar indicated that Normal 5 is actually tumor tissue, and Tumor 22, Tumor 4 are actually normal tissues. Normal 5 is a normal sample but we found it is tumor, so during our tumor cluster anaysis, we remove it. So, there are 53 tumor tissues in total. There are 38 paired tumor and normal samples. In the 38 paired samples, there are 29 epithelioid, 4 sarcomatoid and 5 biphasic samples. For the clinical related analysis, samples Tumor 6, Tumor 17,

Tumor 44, Tumor 54, Tumor 45 have their death seems surgery related, and Normal 5 is a normal sample but we found it is tumor, so we remove them and keep only 48 tumors in use.

1.2.2 Statistical Methods

We first remove the batch effect. Based on the data, we want to find the most robust subgroups of patients:

1. Compare tumor vs normal in paired data
2. Randomly select n (an uniform number from 500 to 2000) probes from the probesets that are selected at the FDR level of 0.01 (around 20000 probes in total), and randomly select 80% of samples. Do cluster.
3. Repeat 2 for 1000 times, record each time whether any two sample are grouped together if cut tree with 3 or 4 groups. Define the consensus matrix as the average time that any two samples are grouped together in each clustering.
4. Use the consensus matrix as the similarity matrix to define the final clustering group.
5. Use cophenetic correlation or other criteria to define the number of subgroups we would like to use
6. For the subgroups we defined, we test in KM to check the association with clinical variable and find the genes that are different between the groups

We also apply KM or Cox model for different clinical information corresponding to overall survival.

1. Age
2. Gender
3. Original Histology
4. T Stage
5. N Stage
6. Overall Stage
7. Chemo Treatment.

Variables are included in the Cox model for multivariate analysis. Akaike Information Criterion (AIC) to eliminate redundant variables from the model.

After the best survival model is selected, we add one more variable factor (the three sub-tumor groups) in the Cox model and perform the overall survival analysis.

1.3 Results

For univariate analysis, the three sub-tumor group is not significant associated with overall survival.

There are four variables remained for multivariate analysis after Akaike Information Criterion (AIC) to eliminate redundant variables from the model:

1. Gender
2. Original Histology
3. Combined N Stage
4. Chemo Treatment

After we include one more variable factor (the three sub-tumor groups) in the Cox model we selected in the previous step and perform the overall survival analysis, this new variable will not be excluded after AIC to eliminate redundant variables.

1.4 Conclusions

We define three surival related sub-tumor groups.

2 Loading the Data

2.1 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.

```
> library(affy)
> library(simpleaffy)
> library(geneplotter)
> library(xtable)
> library(ClassComparison)
> library(ClassDiscovery)
> library(limma)
> library(hgu133plus2.db)
> library(gplots)
> library(gdata)
> library(affyio)
> library(nlme)
> library(survival)
```

2.2 Load Data Object

We first load in the data object we have saved in the previous report.

```
> mainDirectory<-"//data/bioinfo/Private/LungSpore"
> DataDirectory<-"/data/bioinfo2/Lung-HN/Mesothelioma"
> setwd(file.path(mainDirectory, "ReportWithNewClin"))
> load(file.path(mainDirectory, "Report","processedDataAffy.RData"))
> my.fig <- "Report15-Affymetrix-Analsys-Remove-Batch-AllSample-Simulation-Edit-NewID"
```

We check the name and the clinical information consistency.

```
> AllData<-exprs(x.rma)
> identical(substr(colnames(AllData), 19,21), substr(si[,"core-id"], 8,10))

[1] TRUE

>
```

We then read in the sample ID.

```
> si2 <- read.xls("Affy-U133-samples_091111.xls")
> identical(si[,"core-id"], si2[,"core.id"])

[1] TRUE

> si <- cbind(si, si2)
> si3 <- read.xls("Numbered-DeID-samples.xls")
> identical(si[,"core-id"], si3[,"core.id"])

[1] TRUE

> si <- cbind(si, SepID = as.vector(si3[,"Sep.2013.ID"]))
> colnames(AllData) <- as.vector(si[,"SepID"])
>
```

We read in the run date of each sample first.

```
> celDatHeaders<-colnames(AllData)
> for (i1 in 1:length(colnames(AllData))) {
+ temp <- read.celfile.header(file.path(DataDirectory, "Affymetrix-mRNA", si[i1,"fnames"]), in
+ celDatHeaders[i1] <- temp$DatHeader
+ }
> #celDatHeaders[1:3]
> temp <- strsplit(celDatHeaders, "[[:space:]]+")
> celRunDates <- unlist(lapply(temp, function(x) {x[8]}))
> zed <- as.Date(celRunDates,format="%m/%d/%y")
> names(zed)<-si[,c("Sample.ID..IW.")]
> table(zed)
```

```
zed
2009-02-24 2009-02-26 2009-03-12 2009-03-13 2009-03-25 2009-04-07 2009-04-09 2009-04-14
        12         12         12         12         12         12         12         12
```

```
> si <- cbind(si, zed)
>
```

We create another date factor to separate the run date for months.

```
> dateFactor2<-rep("2009-02",length(zed))
> dateFactor2[grep("2009-03", zed)]<-"2009-03"
> dateFactor2[grep("2009-04", zed)]<-"2009-04"
> names(dateFactor2)<-si[,"Sample ID"]
> si <- cbind(si,  dateFactor2)
> rm(dateFactor2, zed, celDatHeaders, celRunDates)
```

A mesothelioma tumor can be of three types: epitheloid, sarcomatoid and biphasic (contains a mix of both sarcomatoid and epitheloid). PI submit one diagnosis: the "Path Report Diagnosis" (the original tumor diagnosis in the patient). We load in the corresponding clinical information.

```
> Histology <- read.xls("Meso-histology-May2010.xlsx")
> dim(Histology)
```

```
[1] 56  2
```

```
> identical(as.vector(Histology[,1]), as.vector(si[,"Sample ID"]))
```

```
[1] TRUE
```

```
> si <- cbind(si, Histology)
> rm(temp1, temp2,  temp)
```

3 Remove Batch Effect Using All Samples

Figure 1 shows the hierarchical clustering for all the samples and all the probesets.

```
> load("QCResult.RData")
> hc <- hclust(distanceMatrix(AllData,"euclidean"),"ward")
> branches <- cutree(hc, k=4)
> pp <- c(x.qc1@percent.present, x.qc2@percent.present)
> names(pp) <- sub(".present", "", names(pp))
> all(names(pp) ==rownames(si))
```

```
[1] TRUE
```

```
> sum(names(pp) ==rownames(si))

[1] 96

> Infor<-cbind(Primary=as.vector(si[,"Primary"]), dateFactor = as.character(si[,"zed"]),
+                 dateFactor2 = as.vector(si[,"dateFactor2"]), branches, pp)
> si <- cbind(si, branches)
> gd <- groupedData(pp ~ 1|branches, data=as.data.frame(Infor), order.groups=FALSE)
> gsummary(gd)

  Primary dateFactor dateFactor2 branches              pp
1   Tumor 2009-02-24     2009-03        1 37.4833104709648
2  Normal 2009-02-26     2009-02        2 42.1380887059899
3   Tumor 2009-04-09     2009-04        3 19.3488797439415
4  Normal 2009-04-07     2009-04        4 25.1742112482853

>
>
```

Figure 2 shows the box plot of the percent present calls for the four clusters in the Figure 5. We revmove the two big branches effect that are caused by quality.

```
> branchesCorrect <- cutree(hc, k=2)
> table(branchesCorrect, branches)

               branches
branchesCorrect  1  2  3  4
              1 43 25  0  0
              2  0  0 16 12

> cla <- as.factor(branchesCorrect)
> covars <- data.frame(mainBatch=cla)
> mlm <- MultiLinearModel(Y ~ mainBatch, covars, AllData)
> debatch <- AllData
> mm <- matrixMean(AllData)
> debatch <- sweep(AllData - t(mlm@predictions), 1, mm, "+")
> AllDataDebatch<-debatch
> Tumorsi <- si[which(si[,"Primary"] == "Tumor" & si[,"SepID"]!="Normal 5"),]
> #Tumorsi <- si[which(si[,"Primary"] == "Tumor" ),]
> TumorSampleDatadebatch<-debatch[, match(as.vector(Tumorsi[,"SepID"]), colnames(debatch))]
> identical(colnames(TumorSampleDatadebatch), as.vector(Tumorsi[,"SepID"]))

[1] TRUE
```

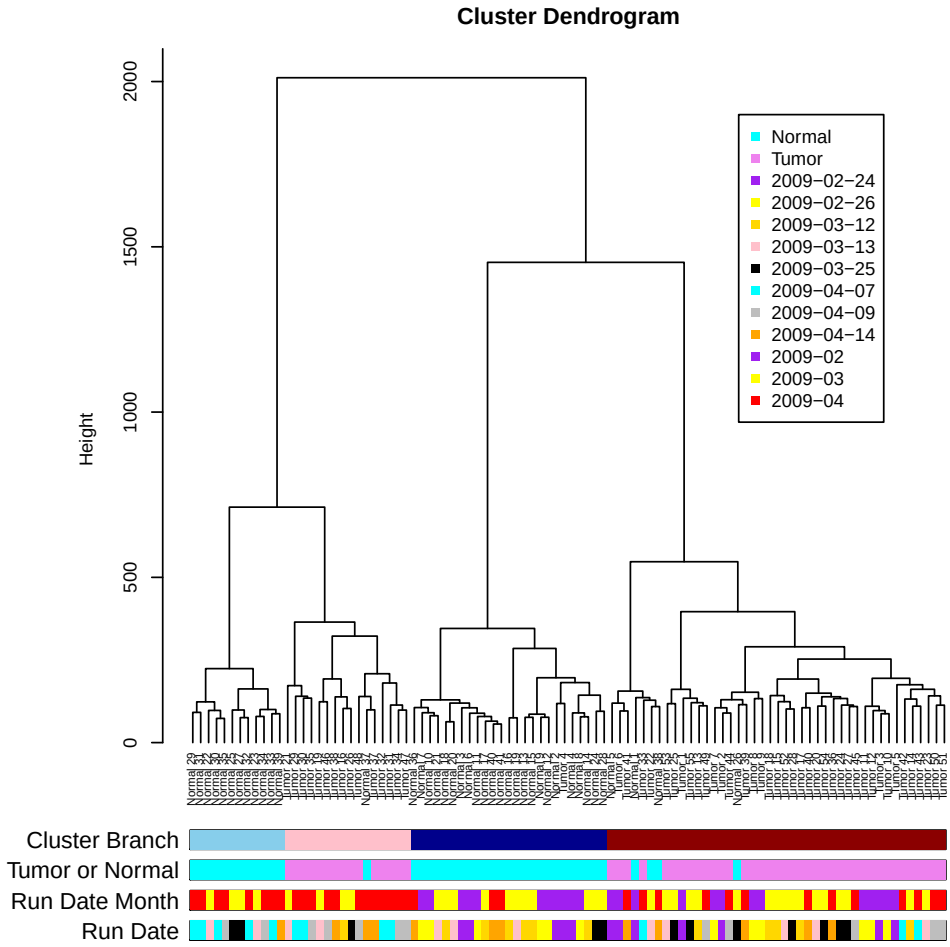

Figure 1: Hierarchical Clustering For All Samples, using euclidean and ward method.

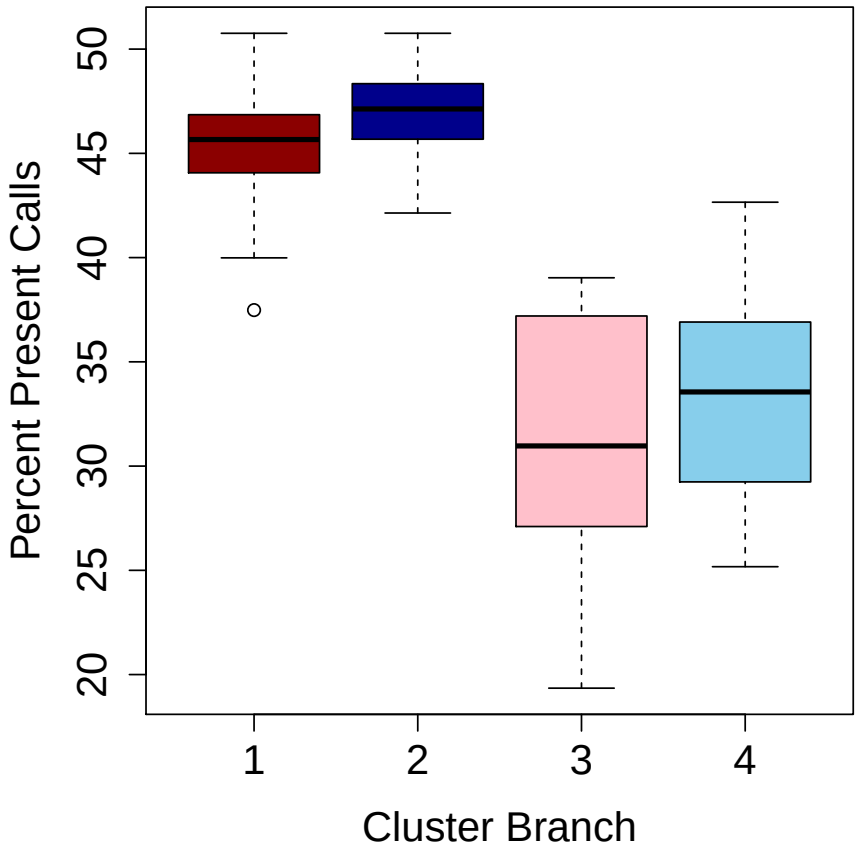

Figure 2: Box plot of Percent Present Calls for Four Clusters.

We then find the paired data.

The samples are either tumor or normal. Some of them are paired tumor and normal samples. There are 55 unique patients. Among them, there are 14 unmatched samples, which are all tumors.

```
> PatientID<-gsub("T","",gsub("N","",si[,"Sample ID"]))
> names(PatientID)<-si[,"Sample ID"]
> as.vector(si[match(names(table(gsub("T","",gsub("N","",si[,"Sample ID"])))[table(gsub("T","",

 [1] "Tumor 44" "Tumor 42" "Tumor 53" "Tumor 43" "Tumor 47" "Tumor 52" "Tumor 55"
 [8] "Tumor 45" "Tumor 49" "Tumor 54" "Tumor 50" "Tumor 51" "Tumor 46" "Tumor 48"

>
```

From PI's information, Normal 5 is actually tumor tissue, and Tumor 4, Tumor 22 are actually normal tissues, we exclude these three patients and those unmatched samples from the analysis.

```
> UnPairedSamples<-c(as.vector(si[match(names(table(gsub("T","",gsub("N","",si[,"Sample ID"])))
+         "Normal 5", "Tumor 5", "Tumor 22", "Normal 22", "Tumor 4", "Normal 4")
> Pairedsi<-si[-match(UnPairedSamples, si[,"SepID"]),]
> PairedPatientID<-gsub("Tumor","",gsub("Normal","",Pairedsi[,"SepID"]))
> length(unique(gsub("Tumor","",gsub("Normal","",Pairedsi[,"SepID"]))))

[1] 38

> table(Pairedsi[,"Primary"])

Normal  Tumor
    38     38
```

So we have 38 paired tumor and normal samples.

```
> PairedData<-AllData[, -match(UnPairedSamples, si[,"SepID"])]
> identical(colnames(PairedData), as.vector(Pairedsi[,"SepID"]))

[1] TRUE

> PairedDataDebatch <- debatch[, match(colnames(PairedData), colnames(debatch))]
> identical(colnames(PairedDataDebatch), as.vector(Pairedsi[,"SepID"]))

[1] TRUE
```

We save some information for further use.

```
> probes<-rownames(PairedData)
> acc <- unlist(mget(probes, hgu133plus2ACCNUM))
> chr <- mget(probes, hgu133plus2CHR)
> chr <- unlist(lapply(chr, function(x) x[1]))
> gene <- unlist(mget(probes, hgu133plus2GENENAME))
> sym <- unlist(mget(probes, hgu133plus2SYMBOL))
> uni <- mget(probes, hgu133plus2UNIGENE)
> uni <- unlist(lapply(uni, function(x) x[1]))
> annot <- data.frame(GenBank=acc, Symbol=sym, UniGene=uni,
+                     Chrom=chr, Description=gene)
> identical(rownames(annot), rownames(PairedData))

[1] TRUE

> save(AllDataDebatch, si, TumorSampleDatadebatch, Tumorsi, PairedDataDebatch, Pairedsi,annot,
+         file = "Report15-Affymetrix-Analsys-Remove-Batch-AllSample-Simulation-Edit-NewID-Deba
>
```

4 Paired T Test Comparing Tumor vs Normal Within Paired Samples

We apply paired t test to check the difference between tumor vs normal effect within paired samples.

```
>         identical(PairedPatientID[which(Pairedsi[,"Primary"] == "Tumor")],
+                 PairedPatientID[which(Pairedsi[,"Primary"] == "Normal")])

[1] TRUE

>         Paired.t.testp <- rep(1, dim(PairedDataDebatch)[1])
>         Paired.t.testt <- rep(1, dim(PairedDataDebatch)[1])
>         for(i1 in 1:length(Paired.t.testp))
+         {
+         Paired.t.test <- t.test(PairedDataDebatch[i1,which(Pairedsi[,"Primary"] == "Tumor")],
+                         PairedDataDebatch[i1, which(Pairedsi[,"Primary"] !="Tumor")], pair=Th
+         Paired.t.testp[i1]<-Paired.t.test$p.value
+         Paired.t.testt[i1]<-Paired.t.test$statistic
+
+         }
>
>
>

> action.bum <- Bum(Paired.t.testp)
>
```

```
   FDRs Significant Probesets Corresponding p-value
1 5e-10                  1017              2.37e-11
2 1e-09                  1157              5.42e-11
3 5e-09                  1575              3.70e-10
4 1e-08                  1761              8.46e-10
5 5e-08                  2306              5.77e-09
6 1e-02                 21267              1.22e-02
7 5e-02                 30459              8.57e-02
```

```
> TumorNormalMean<-apply(PairedDataDebatch, 1, function(x) {tapply(x,list(Pairedsi[,"Primary"])
> TumorNormalFold<-sign((TumorNormalMean["Tumor",]-TumorNormalMean["Normal",]))*2^(abs(TumorNor
> combined <- data.frame(Probe = rownames(PairedData), tstat=Paired.t.testt,
+         pvalue=Paired.t.testp, PairedDataTumor=TumorNormalMean["Tumor",],
+         PairedDataNormal=TumorNormalMean["Normal",], PairedDataFoldChange=TumorNormalFold ,
+         annot)
> write.csv(combined, file="Report15-Affymetrix-Analsys-Remove-Batch-AllSample-Simulation-Edit-
>
```

5 Robust Clustering Using Tumor Samples

In this section, we generate the robust hierarchical clustering for tumor samples and the probesets are selected for different comparing tumor vs normal in the paired data. We do this 1000 times. Each time, we randomly select n (an uniform number from 500 to 2000) probes from the probesets that are selected at the FDR level of 0.01 (around 20000 probes in total), and randomly select 80$Record each time whether any two sample are grouped together if cut tree with 3 or 4 groups. Define the consensus matrix as the average time that any two samples are grouped together in each clustering.

```
> ############  define data frame
> DataSource <- TumorSampleDatadebatch[selectSignificant(action.bum, 0.01, by='FDR'),]
> dim(DataSource)

[1] 21267    53

> totalrepnum <- 1000
> ConsensusMatrix3 <- matrix(0,dim(TumorSampleDatadebatch)[2], dim(TumorSampleDatadebatch)[2],
> rownames(ConsensusMatrix3) <- colnames(TumorSampleDatadebatch)
> colnames(ConsensusMatrix3) <- colnames(TumorSampleDatadebatch)
> ConsensusMatrix4 <- matrix(0,dim(TumorSampleDatadebatch)[2], dim(TumorSampleDatadebatch)[2],
> rownames(ConsensusMatrix4) <- colnames(TumorSampleDatadebatch)
> colnames(ConsensusMatrix4) <- colnames(TumorSampleDatadebatch)
> SampleMatrix <- matrix(0,dim(TumorSampleDatadebatch)[2], dim(TumorSampleDatadebatch)[2], byro
```

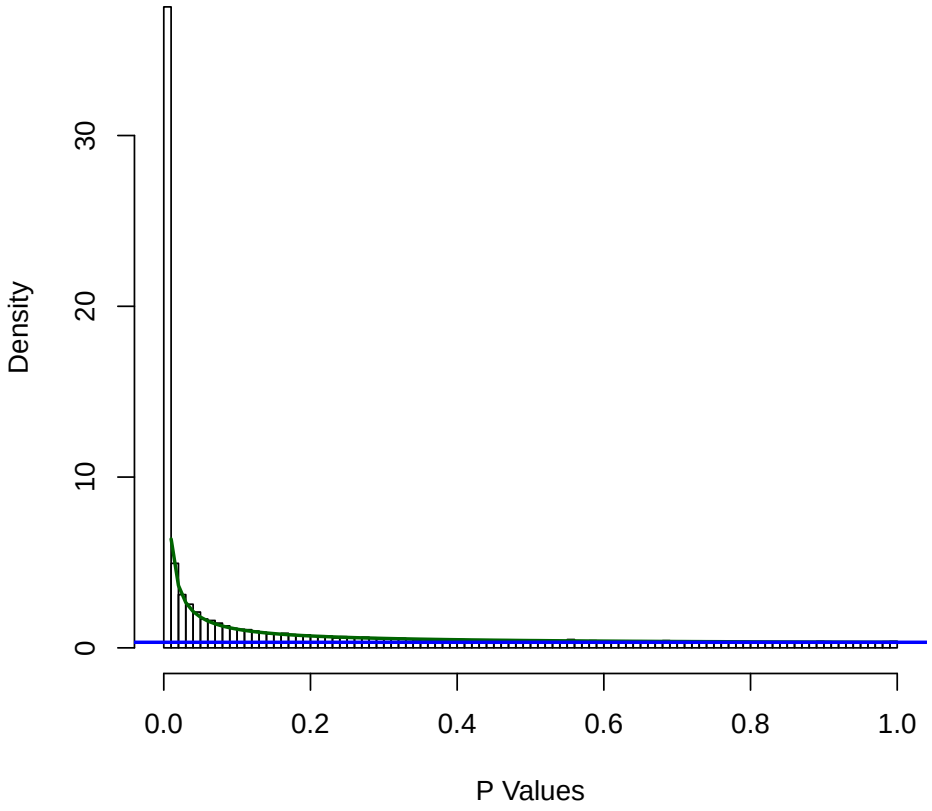

Figure 3: Results of the BUM analysis for difference between tumor vs normal using paired t test within pair samples. There is a peak at the left side. Some probesets are associated with the gene expression with some appropriate FDR level.

```
> rownames(SampleMatrix) <- colnames(TumorSampleDatadebatch)
> colnames(SampleMatrix) <- colnames(TumorSampleDatadebatch)
> for(repnum in 1:totalrepnum)
+ {
+                 selectlength <-  sample(c(500:2000), 1)
+                 selectcollength <- floor(0.8*dim(DataSource)[2])
+                 selectrow <- sample(c(1:dim(DataSource)[1]), selectlength)
+                 selectcol <- sample(c(1:dim(DataSource)[2]), selectcollength)
+                 selectuseddata <- DataSource[selectrow, selectcol]
+                 SampleMatrix[selectcol, selectcol] <- SampleMatrix[selectcol, selectcol] + 1
+
+                 hcTumorSamplesim <- hclust(distanceMatrix(selectuseddata,"euclidean"),"ward")
+                 hcacsim3 <- cutree(hcTumorSamplesim, k=3)
+                 hcacsim4 <- cutree(hcTumorSamplesim, k=4)
+                 usedmatrix3 <- as.matrix(hcacsim3, col=1)%*%t(as.matrix(1/hcacsim3)) == 1
+                 usedmatrix3[which(usedmatrix3)] <- 1
+                 ConsensusMatrix3[selectcol, selectcol] <- ConsensusMatrix3[selectcol, selecto
+                 usedmatrix4 <- as.matrix(hcacsim4, col=1)%*%t(as.matrix(1/hcacsim4)) == 1
+                 usedmatrix4[which(usedmatrix4)] <- 1
+                 ConsensusMatrix4[selectcol, selectcol] <- ConsensusMatrix4[selectcol, selecto
+ }
> Consensus3 <- ConsensusMatrix3/SampleMatrix
> Consensus4 <- ConsensusMatrix4/SampleMatrix
> hcTumorSample3 <- hclust(distanceMatrix((1-Consensus3),"euclidean"),"ward")
> hcac3 <- factor(cutree(hcTumorSample3, k=3))
> SeqLMR3<-rep("Group 2", length(hcac3))
> SeqLMR3[hcac3==unique(hcac3[order.dendrogram(as.dendrogram(hcTumorSample3))])[2]]<-"Group 3"
> SeqLMR3[hcac3==unique(hcac3[order.dendrogram(as.dendrogram(hcTumorSample3))])[3]]<-"Group 1"
> Infor<-cbind(Histology=as.vector(Tumorsi[,"Path.Report.Diagnosis"]), Seq = SeqLMR3)
>
>


> OutputInfor <- cbind(Tumorsi, Seq = SeqLMR3)[colInd,]
> write.csv(OutputInfor, file="Report15-Affymetrix-Analsys-Remove-Batch-AllSample-Simulation-Ed
```

6 ANOVA Comparing the Expression Levels of Probesets among Three Major Splits

We would like to check which probesets show significant different expression levels among these three splits.

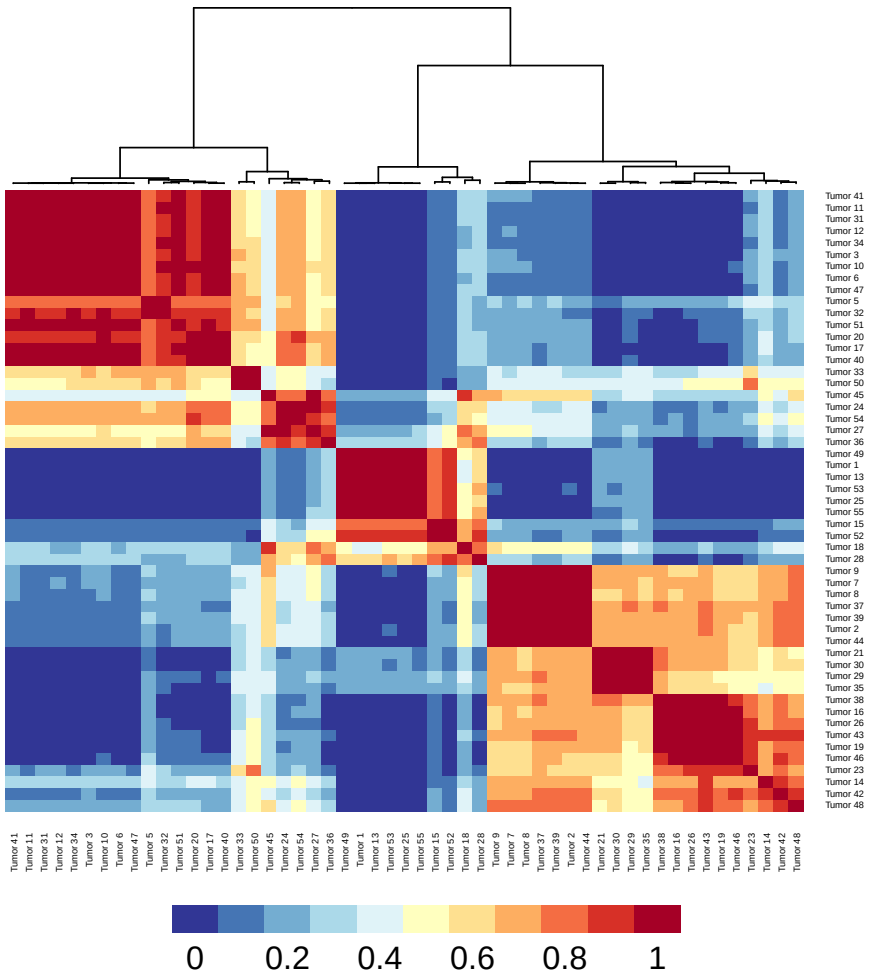

Figure 4: Robust Consensus For tumor samples and the probesets are selected as described in Method section, separation into three groups, using euclidean and ward method.

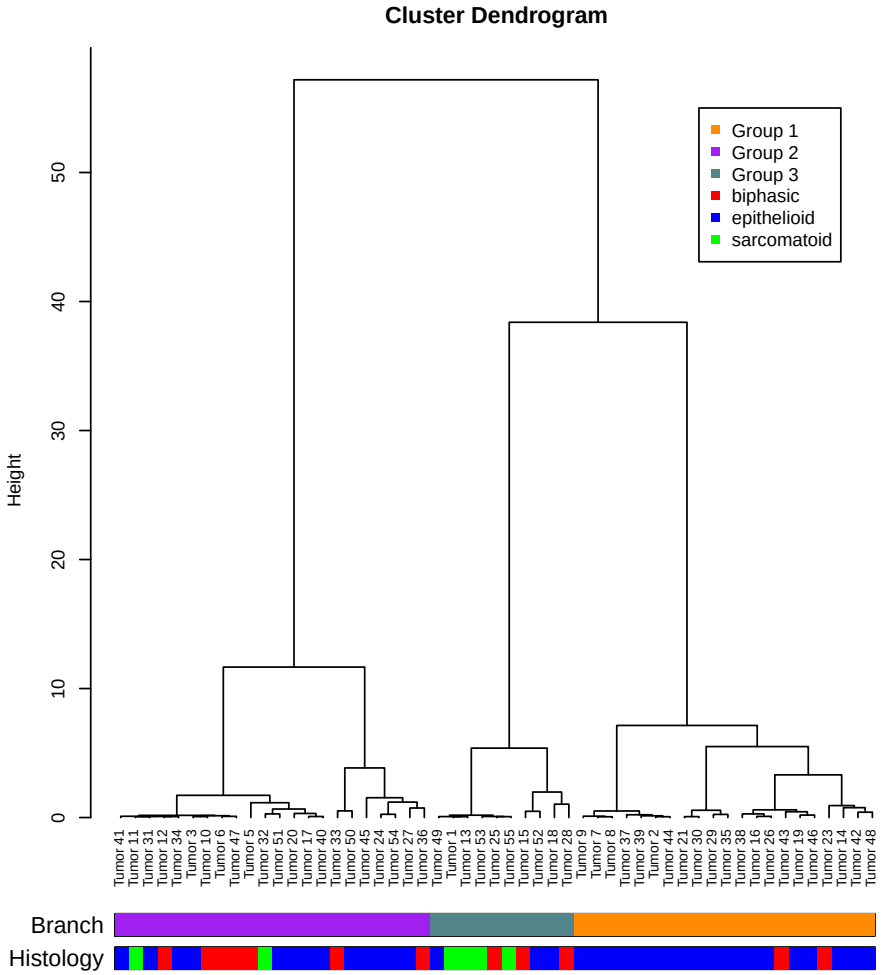

Figure 5: Robust Hierarchical Clustering For tumor samples and the probesets are selected as described in Method section, separation into three groups, using euclidean and ward method.

ANOVA is applied on a probe-by-probe basis using `MultiLinearModel`. The F-statistics and the p-values associated with the F-statistics comparing the linear model to the null model are used to evaluate the expressions levels of each probeset among the three splits.

```
> SeqLMR <- SeqLMR3
> #######  check the three branch information
> table(SeqLMR)

SeqLMR
Group 1 Group 2 Group 3
     21      22      10

> #######  fit in ANOVA model
> covars <- data.frame(SeqLMR=SeqLMR)
> ThreeBranchResult <- MultiLinearModel(Y ~ SeqLMR, covars, TumorSampleDatadebatch)
```

Because of the multiple testing involved in this approach, the individual p-values are not particularly meaningful. However, when we look across the entire set of tests, the distribution of the p-values (under the null hypothesis that no mRNAs provide useful information) should be uniform. If, on the other hand, some mRNAs provide useful information about predicting the response, we would expect an overabundance of small p-values. We can capture this situation by modeling the distribution of the p-values with a Beta-uniform Mixture (BUM). To identify significantly differentially expressed genes (associated with major split effect), we choose a cutoff for the single test p-values by controlling the false discovery rate (FDR), which is defined as the percentage of genes called significant that are expected to turn out false.

```
> ThreeBranchResult.bum <- Bum(ThreeBranchResult@p.values)
```

The following includes all the number of probesets selected at the different FDR levels comparing the probeset expression levels among all three major splits.

```
   FDRs Significant Probesets Corresponding p-value
1 5e-05                  1628             7.43e-06
2 1e-04                  2301             1.92e-05
3 5e-04                  4706             1.75e-04
4 1e-03                  6306             4.54e-04
5 5e-03                 12358             4.14e-03
6 1e-02                 16454             1.07e-02
7 5e-02                 30757             9.78e-02
```

We define the left group as the Group 2, corresponding to most of the biphasic, the right group as the Group 1, corresponding to epithelioid, and the middle group as the Group 3, corresponding to most of the sarcomatoid.

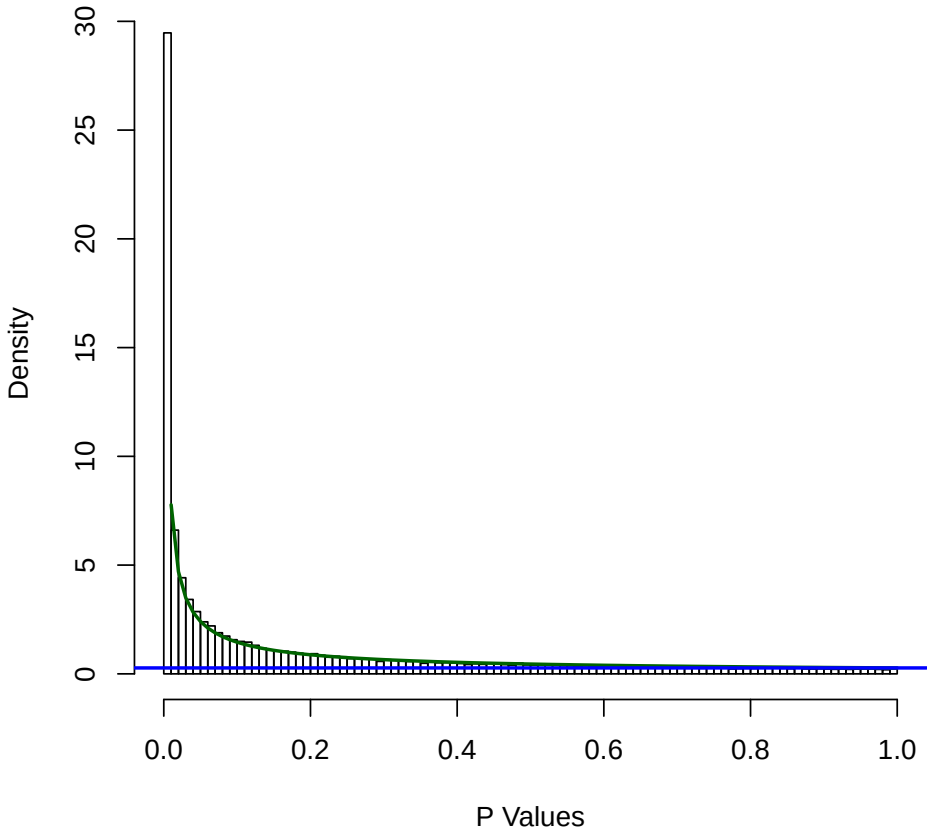

Figure 6: Results of a BUM analysis of the ANOVA comparing linear model to the null model to define the major three split of Figure5. The curve tends to be a peak at the left end, suggesting there is some differential expression present at an appropriate FDR level.

```
> ######  define the three groups created in the previous report
>
> ThreeGroupTvsNTestProbe<-SeqLMR
> Tumorsi <- cbind(Tumorsi,ThreeGroupTvsNTestProbe )
```

We save the result.

```
> ######  calculate the mean for each group
> ThreeBranchMeanTvsNTestProbe<-t(apply(TumorSampleDatadebatch, 1, function(x) {tapply(x, as.fa
> ######  save the result
>
> ThreeGroupFDR <- 0.0001
> ReportResult<-data.frame(Probes=rownames(TumorSampleDatadebatch), FStatistics= ThreeBranchRes
+         pValues=ThreeBranchResult@p.values, ThreeBranchMeanTvsNTestProbe,
+         ANOVAThreeGroup =selectSignificant(ThreeBranchResult.bum, ThreeGroupFDR, by='FDR'))
>
```

6.1 Define Probesets Distribution In Three Subgroups

Then, we need to define how the probes distributed in three groups.

We first define the order for the three groups.

```
> ##############  define the order of the probesets
> MeanOrder <- apply(ThreeBranchMeanTvsNTestProbe, 1, function(x){paste(order(x, decreasing=TRU
> table(MeanOrder)

MeanOrder
1-2-3 1-3-2 2-1-3 2-3-1 3-1-2 3-2-1
14145 14252  5726  4442  7377  8733

> table(MeanOrder[which(ReportResult[,"ANOVAThreeGroup"])])

1-2-3 1-3-2 2-1-3 2-3-1 3-1-2 3-2-1
  382   108   342   283   244   942

> ReportResult<- cbind(ReportResult, MeanOrder)
> ############## Tukey HSD is applied
>
> TUKEYResult <- matrix(0, dim(TumorSampleDatadebatch)[1], 6, byrow=TRUE)
> colnames(TUKEYResult) <- c("Group2vs1Diff", "Group2vs1PValue",
+                                 "Group3vs1Diff", "Group3vs1PValue",
+                                 "Group3vs2Diff", "Group3vs2PValue")
> for(i1 in 1:dim(TUKEYResult)[1])
+ {
```

```
+         tempdata <- as.vector(t(TumorSampleDatadebatch[i1,]))
+         temp <- TukeyHSD(aov(tempdata~as.factor(ThreeGroupTvsNTestProbe)))
+         TUKEYResult[i1, c("Group2vs1Diff", "Group2vs1PValue")] <- temp$'as.factor(ThreeGroupT
+         TUKEYResult[i1, c("Group3vs1Diff", "Group3vs1PValue")] <- temp$'as.factor(ThreeGroupT
+         TUKEYResult[i1, c("Group3vs2Diff", "Group3vs2PValue")] <- temp$'as.factor(ThreeGroupT
+ }
> ReportResult <- cbind(ReportResult, TUKEYResult)
>
>
>
```

We change the categories as whether the three tests are significant at the 0.05 p value cutoff.

```
> TUKEYResultIndi <- (TUKEYResult[, c("Group2vs1PValue",
+                                     "Group3vs1PValue",
+                                     "Group3vs2PValue")] <=0.05)*1
> IndicatorCombine <- cbind(MeanOrder, TUKEYResultIndi)
> IndicatorCombineFinal <- apply(IndicatorCombine, 1, function(x){paste(x, collapse = "-")})
> ################  check the categories for ANOVA at FDR 0.0001 level
> table(IndicatorCombineFinal[which(ReportResult[,"ANOVAThreeGroup"])])

1-2-3-0-1-1 1-2-3-1-1-0 1-2-3-1-1-1 1-3-2-1-0-0 1-3-2-1-0-1 1-3-2-1-1-0 2-1-3-0-1-1
         43         244          95           8          24          76          53
2-1-3-1-0-1 2-1-3-1-1-1 2-3-1-1-0-0 2-3-1-1-0-1 2-3-1-1-1-0 2-3-1-1-1-1 3-1-2-0-1-1
        192          97          16          95         167           5         187
3-1-2-1-0-1 3-1-2-1-1-1 3-2-1-0-1-1 3-2-1-1-1-0 3-2-1-1-1-1
         24          33         295         373         274

> ReportResult <- cbind(ReportResult, IndicatorCombineFinal)
>
```

We then generate the heatmap using the selected probesets (FDR levle of 1e-04). The clustering for samples does not change.

```
> #######   hierarchical clustering is applied on both directions
> alpha<-ThreeGroupFDR
> tempMatrix<-TumorSampleDatadebatch[selectSignificant(ThreeBranchResult.bum, alpha, by='FDR'),]
> agencl <- hclust(distanceMatrix(t(tempMatrix), "pearson"), "ward")
> asamcl <- hcTumorSample3
> #######   truncate the standardized gene expression for explore purpose
>
> ulim <- 2
> temp <- t(scale(t(tempMatrix)))
> temp[temp > ulim] <- ulim
```

```
> temp[temp < -ulim] <- -ulim
> tempIndicator <- IndicatorCombineFinal[selectSignificant(ThreeBranchResult.bum, alpha, by='FD
> dim(tempMatrix)

[1] 2301   53

>
```

We check how the groups corresponding to the four big groups defined by the clustering.

```
> x <- cutree(agencl, k=4)
> table(x)

x
  1   2   3   4
522 922 354 503

> ########   get the sequence of the cutting tree from left to right
> unique(x[order.dendrogram(as.dendrogram(agencl))])

[1] 1 2 3 4

> ##################  association with the indicator
>
> GeneInThreeGroupANOVAProbe <- x
> GeneInThreeGroupANOVAProbe[x==unique(x[order.dendrogram(as.dendrogram(agencl))])[1]]<-"Probeg
> GeneInThreeGroupANOVAProbe[x==unique(x[order.dendrogram(as.dendrogram(agencl))])[2]]<-"Probeg
> GeneInThreeGroupANOVAProbe[x==unique(x[order.dendrogram(as.dendrogram(agencl))])[3]]<-"Probeg
> GeneInThreeGroupANOVAProbe[x==unique(x[order.dendrogram(as.dendrogram(agencl))])[4]]<-"Probeg
> table(tempIndicator, GeneInThreeGroupANOVAProbe)

             GeneInThreeGroupANOVAProbe
tempIndicator Probegroup1 Probegroup2 Probegroup3 Probegroup4
  1-2-3-0-1-1           0           0          11          32
  1-2-3-1-1-0           0           0           0         244
  1-2-3-1-1-1           0           0           1          94
  1-3-2-1-0-0           0           0           0           8
  1-3-2-1-0-1           1           0           0          23
  1-3-2-1-1-0           0           0           0          76
  2-1-3-0-1-1           0           0          48           5
  2-1-3-1-0-1           0          19         173           0
  2-1-3-1-1-1           0           0          97           0
  2-3-1-1-0-0           0          16           0           0
  2-3-1-1-0-1           0          72          23           0
```

```
  2-3-1-1-1-0          0         166           1           0
  2-3-1-1-1-1          0           5           0           0
  3-1-2-0-1-1        184           2           0           1
  3-1-2-1-0-1          9           0           0          15
  3-1-2-1-1-1         27           1           0           5
  3-2-1-0-1-1        238          57           0           0
  3-2-1-1-1-0          4         369           0           0
  3-2-1-1-1-1         59         215           0           0
>
```

From the distribution, there are four groups. From lowest to highest, they represent:

1. High in Group 3
   - mostly, mean value of group 3 is higher than the other two
   - small part of the probesets, mean value of group 3 is higher than one of the other groups
   - at least one test is significant that Group 3 is higher than at least one of the other two
2. Minerhigh in Group 3
   - almost half of the probeset, mean value of group 3 is higher than one of the other two groups
   - some probesets, mean value of group 3 is higher than the other two
   - at least one test is significant that Group 3 is higher than at least one of the other two.
3. High in Group 2
   - mostly, mean value of group 2 is higher than the other two
   - only one probeset, mean value of group 2 is higher than 1, lower than group 3
   - at least one test is significant that Group 2 is higher than one of the other two
4. High in Group 1
   - mostly, mean value of group 1 is higher than the other two
   - small part of the probesets, mean value of group 1 is higher than one of the other groups
   - at least one test is significant that Group 1 is higher than at least one of the other two

In this way, we define them as the representation.

```
> GeneInThreeGroupANOVAProbe <- x
> GeneInThreeGroupANOVAProbe[x==unique(x[order.dendrogram(as.dendrogram(agencl))])[4]]<-"High G
> GeneInThreeGroupANOVAProbe[x==unique(x[order.dendrogram(as.dendrogram(agencl))])[1]]<-"High G
> GeneInThreeGroupANOVAProbe[x==unique(x[order.dendrogram(as.dendrogram(agencl))])[3]]<-"High G
> GeneInThreeGroupANOVAProbe[x==unique(x[order.dendrogram(as.dendrogram(agencl))])[2]]<-"Miner
> table(tempIndicator, GeneInThreeGroupANOVAProbe)
```

```
             GeneInThreeGroupANOVAProbe
tempIndicator High Group 1 High Group 2 High Group 3 Miner High Group 3
  1-2-3-0-1-1           32           11            0                   0
  1-2-3-1-1-0          244            0            0                   0
  1-2-3-1-1-1           94            1            0                   0
  1-3-2-1-0-0            8            0            0                   0
  1-3-2-1-0-1           23            0            1                   0
  1-3-2-1-1-0           76            0            0                   0
  2-1-3-0-1-1            5           48            0                   0
  2-1-3-1-0-1            0          173            0                  19
  2-1-3-1-1-1            0           97            0                   0
  2-3-1-1-0-0            0            0            0                  16
  2-3-1-1-0-1            0           23            0                  72
  2-3-1-1-1-0            0            1            0                 166
  2-3-1-1-1-1            0            0            0                   5
  3-1-2-0-1-1            1            0          184                   2
  3-1-2-1-0-1           15            0            9                   0
  3-1-2-1-1-1            5            0           27                   1
  3-2-1-0-1-1            0            0          238                  57
  3-2-1-1-1-0            0            0            4                 369
  3-2-1-1-1-1            0            0           59                 215

>
>
>

null device
          1


> ######  output the result
>
>
> tempresult <- ReportResult[selectSignificant(ThreeBranchResult.bum, alpha, by='FDR'),]
> tempresult <- cbind(tempresult, GeneInThreeGroupANOVAProbe)
> ReportResult <- cbind(ReportResult, annot)
> write.csv(ReportResult, file="Report15-Affymetrix-Analsys-Remove-Batch-AllSample-Simulation-
> tempresult <- cbind(tempresult, annot[selectSignificant(ThreeBranchResult.bum, alpha, by='FDR
> write.csv(tempresult, file="Report15-Affymetrix-Analsys-Remove-Batch-AllSample-Simulation-Edi
>
```

We check the assiciation with other clinical variables.

```
> ############  histology
> table(Group = Tumorsi[,"ThreeGroupTvsNTestProbe"], histology = as.vector(Tumorsi[,"Path.Repo
```

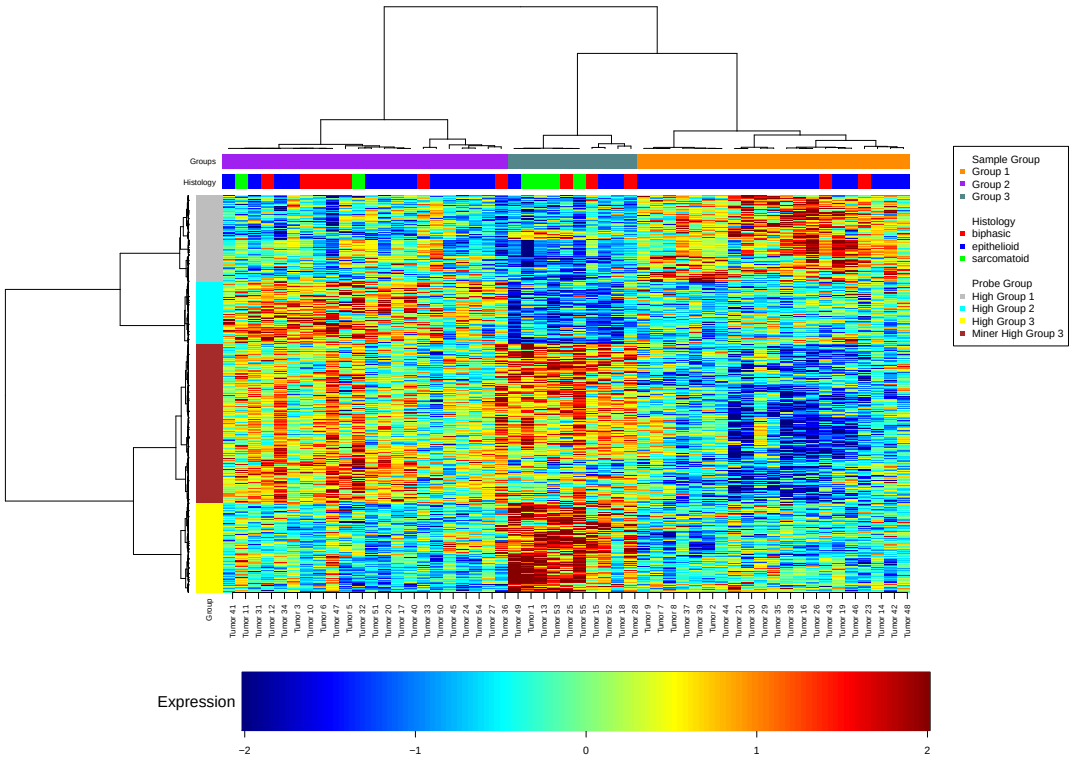

Figure 7: Heatmap for tumor samples, probesets comparing three major splits using ANOVA at the FDR level of 1e-04. For display purposes, we truncate the standardized gene expression values at ± 2 standard deviations. The sample clustering is the robust clusters defined by our algorithm.

```
         histology
Group     biphasic epithelioid sarcomatoid
  Group 1        2          19           0
  Group 2        7          13           2
  Group 3        3           3           4
```

```
> fisher.test(table(Group = Tumorsi[,"ThreeGroupTvsNTestProbe"], histology = as.vector(Tumorsi
```

```
        Fisher's Exact Test for Count Data

data:
p-value = 0.002433
alternative hypothesis: two.sided
```

```
> ############  with the samples that will be removed from analysis
>
> link <- c( "Tumor 6","Tumor 17", "Tumor 44", "Tumor 45", "Tumor 54")
> table(Tumorsi[match(link, Tumorsi[,"SepID"]),"ThreeGroupTvsNTestProbe"])
```

```
Group 1 Group 2 Group 3
      1       4       0
```

```
>
```

7 Overall Survival Analysis

7.1 Load in Clinical Information

We then load in the survival data file.

```
> SurvivalData<-read.xls("Meso-Clinical-Data-June20-2012.xlsx", sheet = 2)
> SurvivalData1<-read.xls("Meso-Clinical-Data-June20-2012.xlsx", sheet = 1)
> identical(as.vector(SurvivalData[,"Spore.ID"]), as.vector(SurvivalData1[,"Spore.ID"]))
```

```
[1] TRUE
```

```
> SurvivalData <- cbind(SurvivalData, LFUS= SurvivalData1[,"LFUS"], Gender = SurvivalData1[,"Ge
```

We clean the tumor samples for further analysis. In the survival analysis, we remove this sample. Samples Tumor 6, Tumor 17, Tumor 44, Tumor 45, Tumor 54 have their death seems surgery related, so we also remove them from further analysis.

```
> RemoveID <- c( "Tumor 6","Tumor 17", "Tumor 44", "Tumor 45", "Tumor 54")
> Removelink <- match(RemoveID, Tumorsi[,"SepID"])
> TumorSurvivalsi <- Tumorsi[-Removelink,]
> dim(TumorSurvivalsi)
```

```
[1] 48 17
```

We would like to use the first "Date.Death" column as the end time point of overall survival analysis. There is 4 patient has empty Date of death. We will use the last follow up date to replace.

```
> ######  look at the patient with no date of death
>
> SurvivalData[SurvivalData[, "Date.Death"]=="", "vital.status"]

[1] A A A A
Levels: A D

> ######  create vector to replace the patient
>
> temp <- ifelse(as.character(SurvivalData[, "Date.Death"])=="",
+         as.character(gsub(" another tumor", "", SurvivalData[, "LFUS"])),
+         as.character(SurvivalData[, "Date.Death"])  )
> temp2 <- temp
> temp2[is.na(as.Date(temp))] <- as.character(as.Date(temp[is.na(as.Date(temp))], "%m/%d/%Y"))
> ######  replace the information with the patient using LFUS
>
> SurvivalData<-data.frame(DeathorLFUS=temp2, SurvivalData)
> ######  we calculate the overall survival time as Death or LFUS minus Date Surgery
>
> OverallTime<- (as.Date(as.character(as.vector(SurvivalData[,"DeathorLFUS"]))) -  as.Date(as.c
> SurvivalData<-data.frame(OverallTime=as.numeric(OverallTime), SurvivalData)
> ######  we cdefine the census of overall survival
>
> table(SurvivalData[,"vital.status"])

 A  D
 4 47

> cen.status <- ifelse(as.character(SurvivalData[,"vital.status"])=="D",
+                 as.character(1), as.character(0) )
> SurvivalData<-data.frame(VitalStatusCen=cen.status, SurvivalData)
> ###### derive the age at the date of surgery  ###############
>
> PatientAge<-(as.Date(as.character(as.vector(SurvivalData[,"Date.Surgery"]))) -
+         as.Date(as.character(as.vector(SurvivalData[,"DOB"]))))/365.25
> ###### summary the age at the date of surgery  ###############
>
> summary(as.numeric(PatientAge))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  40.38   58.85   63.33   63.36   68.41   81.41
```

```
> SurvivalData <-data.frame(PatientAge=as.numeric(PatientAge), SurvivalData)
> rm(PatientAge)
> rm(cen.status)
> rm(OverallTime)
>
```

There is one patient Tumor 47, should be NEO-Adj true based on clinical PI. We change it.

We then create another survival data frame which matches to the tumor samples that will be used in further survival analysis.

```
> MatchID<-match(gsub("T", "", gsub("N", "", TumorSurvivalsi[, "Sample ID"])), gsub("T", "", gs
> #####   check the sample ID not matched to the survival data
>
> sum(is.na(MatchID))

[1] 0

> #####   create the Survival Data Matrix with the matched tumor samples
>
> TumorSampleSurvivalData<-SurvivalData[MatchID,]
> #####   check the consistency between sample ID
>
> identical(as.vector(gsub("T", "", gsub("N", "", TumorSurvivalsi[, "Sample ID"]))),
+                         as.vector(gsub("T", "", gsub("N", "", TumorSampleSurvivalData[,"Spore

[1] TRUE

> identical(as.vector(Tumorsi[, "SepID"]), colnames(TumorSampleDatadebatch))

[1] TRUE

> rm(MatchID)
>
```

We check all the tumor samples.

```
> MatchID<-match(gsub("T", "", gsub("N", "", Tumorsi[, "Sample ID"])), gsub("T", "", gsub("N",
> #####   check the sample ID not matched to the survival data
>
> sum(is.na(MatchID))

[1] 4
```

```
> Tumorsi[is.na(MatchID), "SepID"]

[1] Tumor 6  Tumor 54 Tumor 44 Tumor 45
96 Levels: Normal 1 Normal 10 Normal 11 Normal 12 Normal 13 Normal 14 ... Tumor 9

> #####   create the Survival Data Matrix with the matched tumor samples
>
> TumorSampleAllSurvivalData<-SurvivalData[MatchID,]
> #####   check the consistency between sample ID
>
> #identical(as.vector(gsub("T", "", gsub("N", "", Tumorsi[, "Sample ID"]))),
> #                        as.vector(gsub("T", "", gsub("N", "", TumorSampleAllSurvivalData[,"S
>
>
>
> identical(as.vector(Tumorsi[, "SepID"]), colnames(TumorSampleDatadebatch))

[1] TRUE

> rm(MatchID)
> save(Tumorsi, TumorSampleAllSurvivalData,
+          file="Report15-Affymetrix-Analsys-Remove-Batch-AllSample-Simulation-Edit-NewID-All-S
> ##############  some of the tumor samples we do not have
> ############## clinical information
> table(Group =Tumorsi[,"ThreeGroupTvsNTestProbe"] , NEO=  as.vector(TumorSampleAllSurvivalData

         NEO
Group     No Yes
  Group 1 19   1
  Group 2 11   8
  Group 3  7   3

> fisher.test(table(Group =Tumorsi[,"ThreeGroupTvsNTestProbe"] , NEO=  as.vector(TumorSampleAll

        Fisher's Exact Test for Count Data

data:
p-value = 0.01889
alternative hypothesis: two.sided
```

We save the object for further analysis.

```
> save(TumorSurvivalsi, TumorSampleSurvivalData,
+          file="Report15-Affymetrix-Analsys-Remove-Batch-AllSample-Simulation-Edit-NewID-Survi
>
>
```

7.2 Overall Survival Analysis For Tumor Samples

There are three subgroups defined among all the tumor samples. We would like to check whether the overall survival is different among the three groups.

```
> #######  find the time and census status
>
> Time.dfs <- as.numeric(TumorSampleSurvivalData[,"OverallTime"])
> cen.status <- as.numeric(as.vector(TumorSampleSurvivalData[,"VitalStatusCen"]))
> #######  fit into KM model
> groupused <- TumorSurvivalsi[,"ThreeGroupTvsNTestProbe"]
> OverallThreeGroup<- survfit(Surv(Time.dfs, cen.status) ~ as.factor(groupused),  na.action=na.
> OverallThreeGroup.res <- survdiff(Surv(Time.dfs , cen.status) ~ as.factor(groupused), na.acti
> Factorlevel<-groupused
> pValue3All <- 1-pchisq(OverallThreeGroup.res$chisq, df=length(levels(factor(Factorlevel)))-1)
> #######  show the KM result comparing the three groups
>
> OverallThreeGroup.res

Call:
survdiff(formula = Surv(Time.dfs, cen.status) ~ as.factor(groupused),
    na.action = na.exclude)

                              N Observed Expected (O-E)^2/E (O-E)^2/V
as.factor(groupused)=Group 1 20       17    22.22     1.225      2.52
as.factor(groupused)=Group 2 18       18    15.92     0.272      0.43
as.factor(groupused)=Group 3 10        9     5.86     1.678      1.97

 Chisq= 3.3  on 2 degrees of freedom, p= 0.197

> OverallThreeGroup

Call: survfit(formula = Surv(Time.dfs, cen.status) ~ as.factor(groupused),
    na.action = na.exclude)

                              records n.max n.start events median 0.95LCL 0.95UCL
as.factor(groupused)=Group 1      20    20      20     17  1.562   0.810    3.03
as.factor(groupused)=Group 2      18    18      18     18  0.820   0.454    3.03
as.factor(groupused)=Group 3      10    10      10      9  0.571   0.320      NA

> summary(coxph(Surv(Time.dfs, cen.status) ~ as.factor(groupused),  na.action=na.exclude))

Call:
coxph(formula = Surv(Time.dfs, cen.status) ~ as.factor(groupused),
```

```
    na.action = na.exclude)

  n= 48, number of events= 44

                              coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(groupused)Group 2 0.3991    1.4906   0.3399 1.174   0.2403
as.factor(groupused)Group 3 0.7175    2.0493   0.4181 1.716   0.0861 .
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

                            exp(coef) exp(-coef) lower .95 upper .95
as.factor(groupused)Group 2     1.491     0.6709    0.7656     2.902
as.factor(groupused)Group 3     2.049     0.4880    0.9031     4.650

Concordance= 0.596  (se = 0.046 )
Rsquare= 0.063   (max possible= 0.997 )
Likelihood ratio test= 3.13  on 2 df,   p=0.209
Wald test            = 3.19  on 2 df,   p=0.2034
Score (logrank) test = 3.27  on 2 df,   p=0.1947
```

7.3 Overall Survival Analysis for Histology

There are three histology groups defined among all the tumor samples. We would like to check whether the overall survival is different among the three histology groups.

```
> #######  fit into KM model
>
> UsedFactor <- TumorSurvivalsi[, "Path.Report.Diagnosis"]
> OverallHistGroup<- survfit(Surv(Time.dfs, cen.status) ~ as.factor(UsedFactor),  na.action=na.
> OverallHistGroup.res <- survdiff(Surv(Time.dfs , cen.status) ~ as.factor(UsedFactor), na.acti
> Factorlevel<-TumorSurvivalsi[, "Path.Report.Diagnosis"]
> pValue3All <- 1-pchisq(OverallHistGroup.res$chisq, df=length(levels(factor(Factorlevel)))-1)
> #######  show the KM result comparing the three groups
>
> OverallHistGroup.res

Call:
survdiff(formula = Surv(Time.dfs, cen.status) ~ as.factor(UsedFactor),
    na.action = na.exclude)

                                  N Observed Expected (O-E)^2/E (O-E)^2/V
as.factor(UsedFactor)=biphasic    11       11     7.09  2.161367  2.606452
```

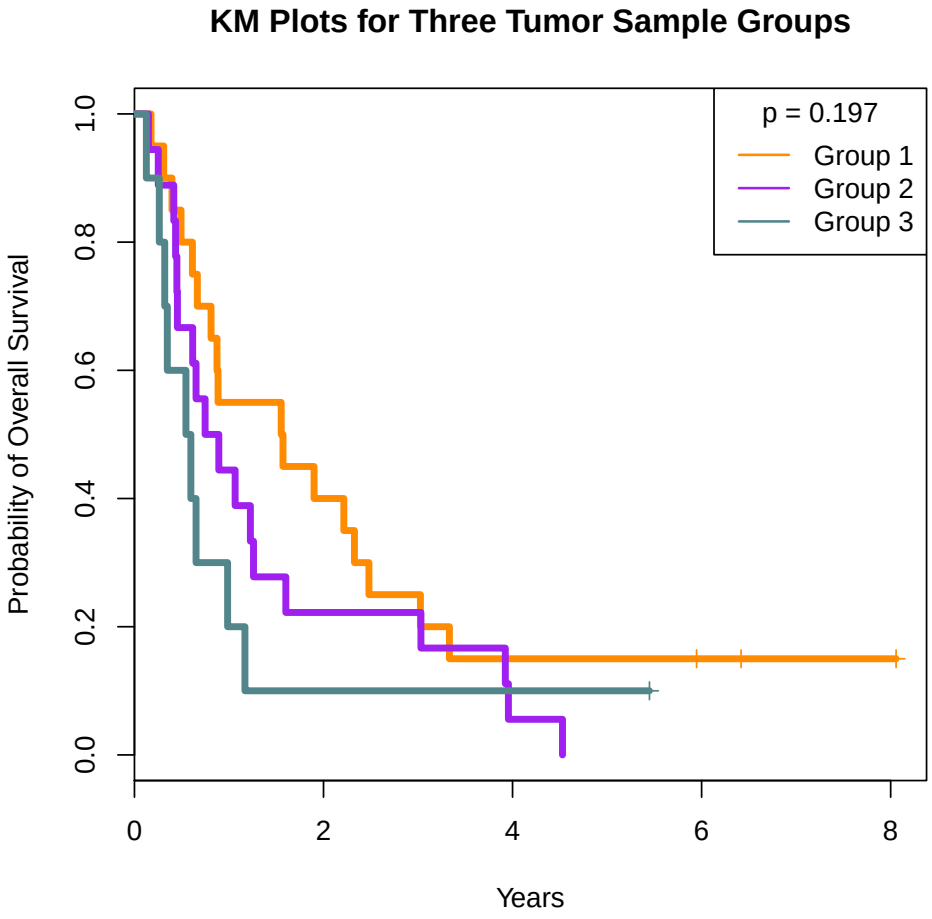

Figure 8: KM for comparing the three groups defined in Figure5, excluding sample Normal 5, Tumor 6, Tumor 17, Tumor 44, Tumor 45, Tumor 54.

```
as.factor(UsedFactor)=epithelioid 31       28    31.86  0.468750  1.720596
as.factor(UsedFactor)=sarcomatoid  6        5     5.05  0.000472  0.000541

 Chisq= 2.7  on 2 degrees of freedom, p= 0.264
```

```
> OverallHistGroup
```

```
Call: survfit(formula = Surv(Time.dfs, cen.status) ~ as.factor(UsedFactor),
    na.action = na.exclude)

                                 records n.max n.start events median 0.95LCL 0.95UCL
as.factor(UsedFactor)=biphasic         11    11      11     11  0.449   0.263      NA
as.factor(UsedFactor)=epithelioid      31    31      31     28  1.065   0.665    2.33
as.factor(UsedFactor)=sarcomatoid       6     6       6      5  0.634   0.597      NA
```

```
> summary(coxph(Surv(Time.dfs, cen.status) ~ as.factor(UsedFactor),  na.action=na.exclude))
```

```
Call:
coxph(formula = Surv(Time.dfs, cen.status) ~ as.factor(UsedFactor),
    na.action = na.exclude)

  n= 48, number of events= 44

                                    coef exp(coef) se(coef)      z Pr(>|z|)
as.factor(UsedFactor)epithelioid -0.5753    0.5625   0.3584 -1.605    0.108
as.factor(UsedFactor)sarcomatoid -0.4522    0.6362   0.5439 -0.831    0.406

                                 exp(coef) exp(-coef) lower .95 upper .95
as.factor(UsedFactor)epithelioid    0.5625      1.778    0.2786     1.136
as.factor(UsedFactor)sarcomatoid    0.6362      1.572    0.2191     1.847

Concordance= 0.578  (se = 0.04 )
Rsquare= 0.048   (max possible= 0.997 )
Likelihood ratio test= 2.36  on 2 df,   p=0.3076
Wald test            = 2.58  on 2 df,   p=0.2749
Score (logrank) test = 2.65  on 2 df,   p=0.2658
```

7.4 Overall Survival Analysis for Epithelial in Group 2 and 3

There are only 2 epithelial samples in grouthree. We would like to check whether the overall survival is different among the other two groups groups.

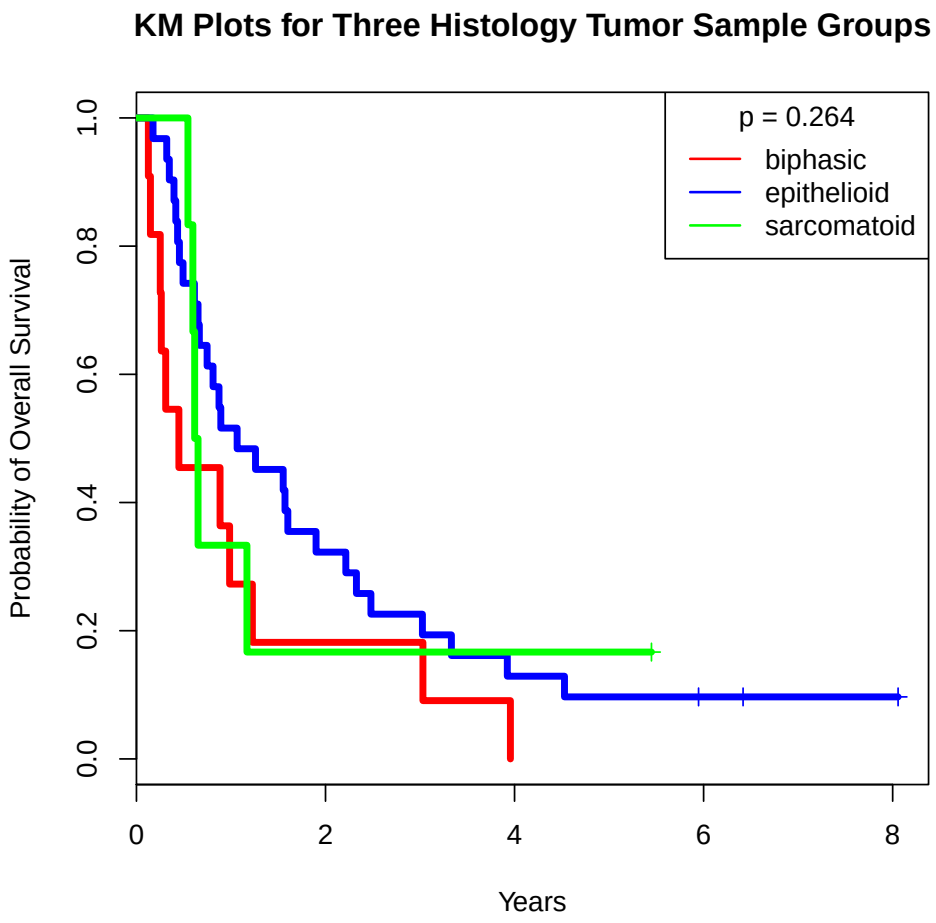

Figure 9: KM for comparing the three groups defined by histology, excluding samples Normal 5, Tumor 6, Tumor 17, Tumor 44, Tumor 45, Tumor 54.

```
> #######  find the index of epithelioid
>
> EpiIndex <- which(TumorSurvivalsi[, "Path.Report.Diagnosis"] == "epithelioid" & groupused !=
> length(EpiIndex)

[1] 28

> #######  find the time and census status
>
> Time.dfs <- as.numeric(TumorSampleSurvivalData[EpiIndex,"OverallTime"])
> cen.status <- as.numeric(as.vector(TumorSampleSurvivalData[EpiIndex,"VitalStatusCen"]))
> #######  fit into KM model
>
> OverallThreeGroupEpi<- survfit(Surv(Time.dfs, cen.status) ~ as.factor(as.vector(groupused[Epi
> OverallThreeGroup.resEpi <- survdiff(Surv(Time.dfs , cen.status) ~ as.factor(as.vector(groupu
> Factorlevel<-as.vector(groupused[EpiIndex])
> table(as.vector(groupused[EpiIndex]))

Group 1 Group 2
     18      10

> pValue3All <- 1-pchisq(OverallThreeGroup.resEpi$chisq, df=length(levels(factor(Factorlevel))
> #######  show the KM result comparing the three groups
>
> OverallThreeGroup.resEpi

Call:
survdiff(formula = Surv(Time.dfs, cen.status) ~ as.factor(as.vector(groupused[EpiIndex])),
    na.action = na.exclude)

                                                 N Observed Expected (O-E)^2/E
as.factor(as.vector(groupused[EpiIndex]))=Group 1 18       15    17.49     0.354
as.factor(as.vector(groupused[EpiIndex]))=Group 2 10       10     7.51     0.825
                                                 (O-E)^2/V
as.factor(as.vector(groupused[EpiIndex]))=Group 1       1.2
as.factor(as.vector(groupused[EpiIndex]))=Group 2       1.2

 Chisq= 1.2  on 1 degrees of freedom, p= 0.273

> OverallThreeGroupEpi

Call: survfit(formula = Surv(Time.dfs, cen.status) ~ as.factor(as.vector(groupused[EpiIndex])),
    na.action = na.exclude)
```

```
                                                 records n.max n.start events median
as.factor(as.vector(groupused[EpiIndex]))=Group 1      18    18      18     15  1.736
as.factor(as.vector(groupused[EpiIndex]))=Group 2      10    10      10     10  0.979
                                                 0.95LCL 0.95UCL
as.factor(as.vector(groupused[EpiIndex]))=Group 1   0.810    3.33
as.factor(as.vector(groupused[EpiIndex]))=Group 2   0.454      NA
```

```
> summary(coxph(Surv(Time.dfs, cen.status) ~ as.factor(as.vector(groupused[EpiIndex])),  na.act
```

```
Call:
coxph(formula = Surv(Time.dfs, cen.status) ~ as.factor(as.vector(groupused[EpiIndex])),
    na.action = na.exclude)

  n= 28, number of events= 25

                                                   coef exp(coef) se(coef)     z
as.factor(as.vector(groupused[EpiIndex]))Group 2 0.4482    1.5655   0.4124 1.087
                                                 Pr(>|z|)
as.factor(as.vector(groupused[EpiIndex]))Group 2    0.277

                                                 exp(coef) exp(-coef) lower .95
as.factor(as.vector(groupused[EpiIndex]))Group 2     1.566     0.6388    0.6976
                                                 upper .95
as.factor(as.vector(groupused[EpiIndex]))Group 2     3.513

Concordance= 0.551  (se = 0.055 )
Rsquare= 0.04   (max possible= 0.991 )
Likelihood ratio test= 1.14  on 1 df,   p=0.2857
Wald test            = 1.18  on 1 df,   p=0.2771
Score (logrank) test = 1.2  on 1 df,   p=0.2733
```

7.5 Overall Survival Analysis Removing NEO-adj Treated Patients

We remove the Neo-adj patients and compare the three groups.

```
> #######  find the index of epithelioid
>
> EpiIndex <- which(TumorSampleSurvivalData[,"Neoadjuvant.chemo"] == "No")
> length(EpiIndex)

[1] 37

> #######  find the time and census status
>
```

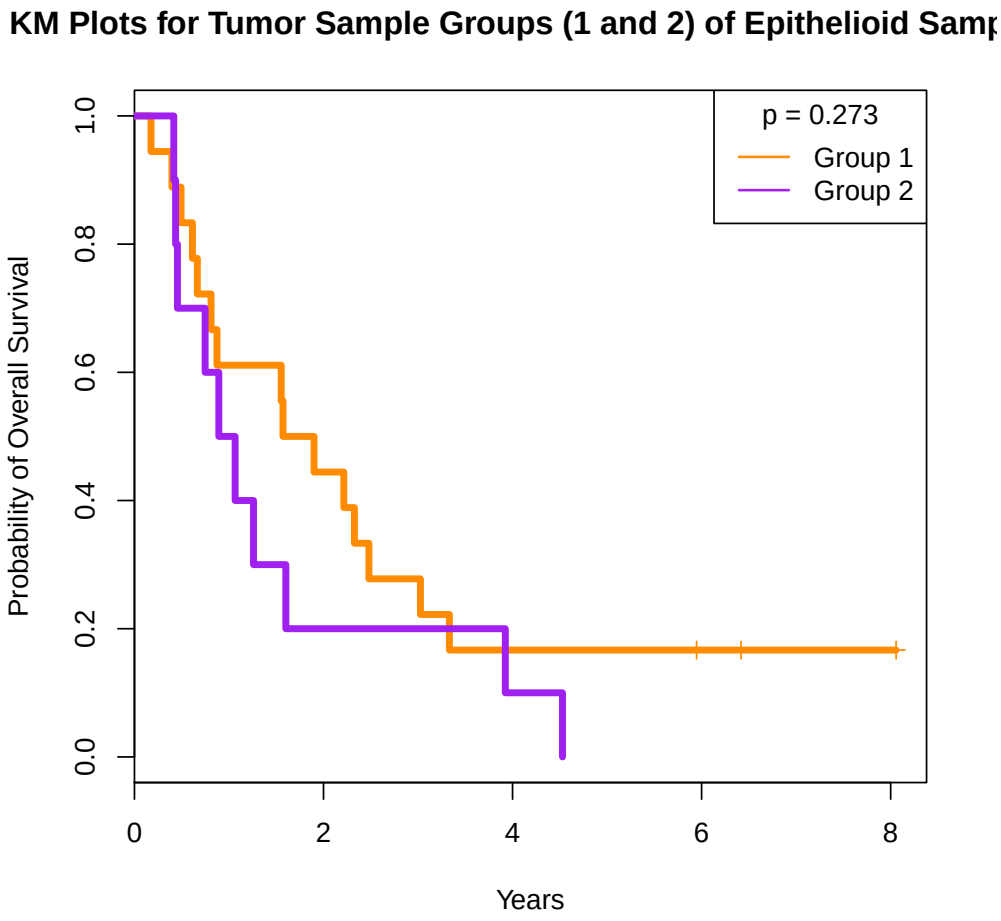

Figure 10: KM for comparing the groups 1 and 2 defined in Figure5 with only epithelioid samples, excluding samples Normal 5, Tumor 6, Tumor 17, Tumor 44, Tumor 45, Tumor 54.

```
> Time.dfs <- as.numeric(TumorSampleSurvivalData[EpiIndex,"OverallTime"])
> cen.status <- as.numeric(as.vector(TumorSampleSurvivalData[EpiIndex,"VitalStatusCen"]))
> #######  fit into KM model
>
> OverallThreeGroupEpi<- survfit(Surv(Time.dfs, cen.status) ~ as.factor(groupused[EpiIndex]),
> OverallThreeGroup.resEpi <- survdiff(Surv(Time.dfs , cen.status) ~ as.factor(groupused[EpiInd
> Factorlevel<-groupused[EpiIndex]
> table(groupused[EpiIndex])

Group 1 Group 2 Group 3
     19      11       7

> pValue3All <- 1-pchisq(OverallThreeGroup.resEpi$chisq, df=length(levels(factor(Factorlevel)))
> #######  show the KM result comparing the three groups
>
> OverallThreeGroup.resEpi

Call:
survdiff(formula = Surv(Time.dfs, cen.status) ~ as.factor(groupused[EpiIndex]),
    na.action = na.exclude)

                                      N Observed Expected (O-E)^2/E (O-E)^2/V
as.factor(groupused[EpiIndex])=Group 1 19       16    18.29    0.2861    0.6481
as.factor(groupused[EpiIndex])=Group 2 11       11    10.33    0.0435    0.0641
as.factor(groupused[EpiIndex])=Group 3  7        6     4.38    0.5962    0.7000

 Chisq= 0.9  on 2 degrees of freedom, p= 0.625

> OverallThreeGroupEpi

Call: survfit(formula = Surv(Time.dfs, cen.status) ~ as.factor(groupused[EpiIndex]),
    na.action = na.exclude)

                                     records n.max n.start events median 0.95LCL
as.factor(groupused[EpiIndex])=Group 1      19    19      19     16  1.572   0.873
as.factor(groupused[EpiIndex])=Group 2      11    11      11     11  1.227   0.616
as.factor(groupused[EpiIndex])=Group 3       7     7       7      6  0.652   0.545
                                     0.95UCL
as.factor(groupused[EpiIndex])=Group 1    3.33
as.factor(groupused[EpiIndex])=Group 2      NA
as.factor(groupused[EpiIndex])=Group 3      NA

> summary(coxph(Surv(Time.dfs, cen.status) ~ as.factor(groupused[EpiIndex]),  na.action=na.excl
```

```
Call:
coxph(formula = Surv(Time.dfs, cen.status) ~ as.factor(groupused[EpiIndex]),
    na.action = na.exclude)

  n= 37, number of events= 33

                                       coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(groupused[EpiIndex])Group 2 0.1979    1.2188   0.3935 0.503    0.615
as.factor(groupused[EpiIndex])Group 3 0.4555    1.5770   0.4838 0.942    0.346

                                      exp(coef) exp(-coef) lower .95 upper .95
as.factor(groupused[EpiIndex])Group 2     1.219     0.8205    0.5636     2.636
as.factor(groupused[EpiIndex])Group 3     1.577     0.6341    0.6110     4.071

Concordance= 0.558  (se = 0.052 )
Rsquare= 0.024   (max possible= 0.994 )
Likelihood ratio test= 0.89  on 2 df,   p=0.6411
Wald test            = 0.93  on 2 df,   p=0.6287
Score (logrank) test = 0.94  on 2 df,   p=0.6252
```

8 Survival Analysis With Other Clinical Information

In this section, we will check the survival analysis with other clinical information. We first calculate the summary of each category in the patients.

```
> #######  find the time and census status
>
> Time.dfs <- TumorSampleSurvivalData[,"OverallTime"]
> cen.status <- as.numeric(as.vector(TumorSampleSurvivalData[,"VitalStatusCen"]))
> ###### derive the age at the date of surgery  ###############
>
> PatientAge<-TumorSampleSurvivalData[,"PatientAge"]
> ###### summary the age at the date of surgery  ###############
>
> summary(PatientAge)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  40.38   58.68   63.19   63.01   67.69   81.41

> ####### gender ##############
>
> PatientGender<-TumorSampleSurvivalData[,"Gender"]
> table(PatientGender)
```

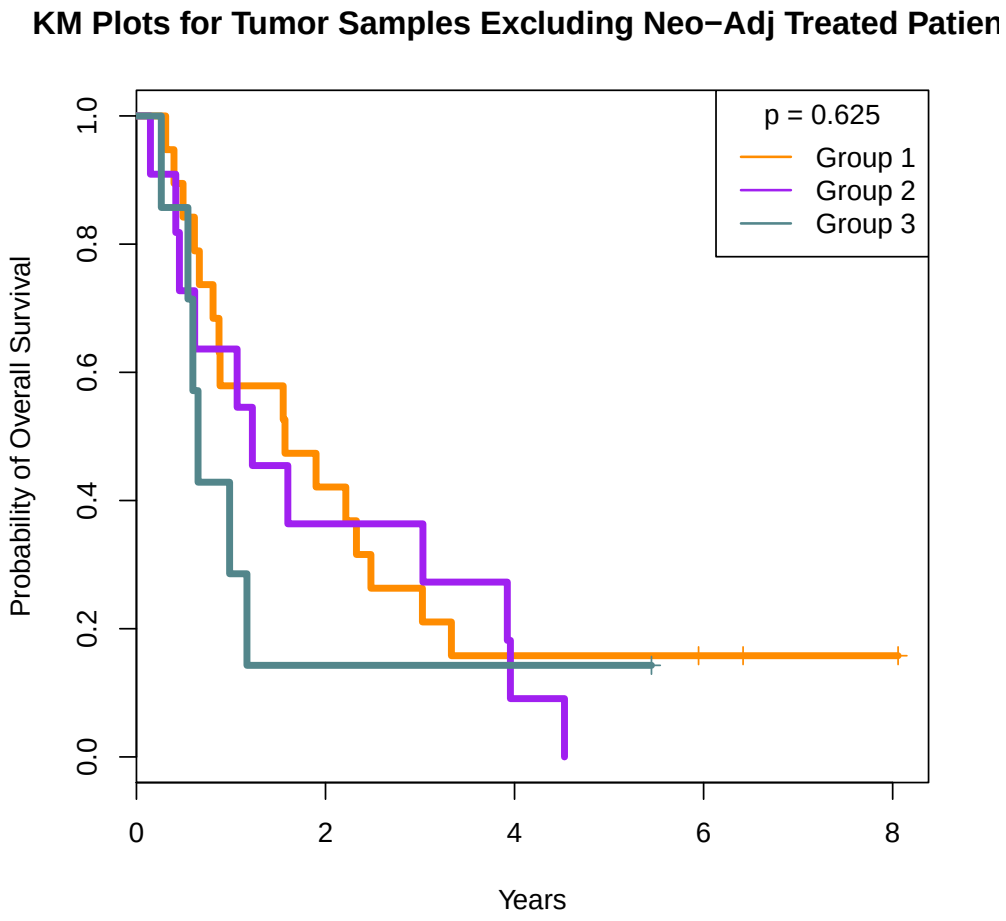

Figure 11: KM for comparing the three groups defined in Figure5, excluding sample Normal 5, Tumor 6, Tumor 17, Tumor 44, Tumor 45, Tumor 54 and samples that are NEo-adj treated.

```
PatientGender
Female   Male
     8     40

> table(PatientGender, cen.status)

             cen.status
PatientGender  0  1
       Female  4  4
       Male    0 40

> table(PatientGender)/length(PatientGender)

PatientGender
   Female      Male
0.1666667 0.8333333

> #######original histology #############
>
> PatientOriginalDiag <- TumorSurvivalsi[, "Path.Report.Diagnosis"]
> table(PatientOriginalDiag )

PatientOriginalDiag
   biphasic epithelioid sarcomatoid
         11          31           6

> table(PatientOriginalDiag , cen.status)

                  cen.status
PatientOriginalDiag  0  1
        biphasic     0 11
        epithelioid  3 28
        sarcomatoid  1  5

> table(PatientOriginalDiag )/length(PatientOriginalDiag )

PatientOriginalDiag
   biphasic epithelioid sarcomatoid
  0.2291667   0.6458333   0.1250000

> ####### T Stage #############
>
> PatientTStage <- as.vector(TumorSampleSurvivalData[,"Pathologic.T.stage"])
> table(PatientTStage)
```

```
PatientTStage
T1 T2 T3 T4
 2  5 36  5
```

```
> table(PatientTStage, cen.status)
```

```
             cen.status
PatientTStage  0  1
           T1  1  1
           T2  0  5
           T3  3 33
           T4  0  5
```

```
> table(PatientTStage)/length(PatientTStage)
```

```
PatientTStage
        T1         T2         T3         T4
0.04166667 0.10416667 0.75000000 0.10416667
```

```
> ##      only two patient at stage 1, so I combine stage 1 and 2
>
> PatientTStageCombine <- PatientTStage
> PatientTStageCombine[PatientTStage == "T1" | PatientTStage == "T2"] <- "T1 or T2"
> table(PatientTStageCombine)
```

```
PatientTStageCombine
T1 or T2       T3       T4
       7       36        5
```

```
> table(PatientTStageCombine, cen.status)
```

```
                   cen.status
PatientTStageCombine  0  1
            T1 or T2  1  6
            T3        3 33
            T4        0  5
```

```
> table(PatientTStageCombine)/length(PatientTStageCombine)
```

```
PatientTStageCombine
 T1 or T2        T3        T4
0.1458333 0.7500000 0.1041667
```

```
> ####### N Stage ##############
>
> PatientNStage <- as.vector(TumorSampleSurvivalData[,"Path.N.stage"])
> table(PatientNStage)
```

```
PatientNStage
N0 N1 N2 N3
23  7 16  2

> table(PatientNStage, cen.status)

             cen.status
PatientNStage  0  1
           N0  3 20
           N1  1  6
           N2  0 16
           N3  0  2

> table(PatientNStage)/length(PatientNStage)

PatientNStage
        N0         N1         N2         N3
0.47916667 0.14583333 0.33333333 0.04166667

> ##      only two patient at stage 3, so I combine stage 2 and 3
>
> PatientNStageCombine <- PatientNStage
> PatientNStageCombine[PatientNStage == "N2" | PatientNStage == "N3"] <- "N2 or N3"
> table(PatientNStageCombine)

PatientNStageCombine
      N0       N1 N2 or N3
      23        7       18

> table(PatientNStageCombine, cen.status)

                    cen.status
PatientNStageCombine  0  1
            N0        3 20
            N1        1  6
            N2 or N3  0 18

> table(PatientNStageCombine)/length(PatientNStageCombine)

PatientNStageCombine
       N0        N1  N2 or N3
0.4791667 0.1458333 0.3750000
```

```
> #######   Overall.Pathologic.Stage #################
>
>
>
> PatientOverStage <- as.vector(TumorSampleSurvivalData[,"Overall.Pathologic.Stage"])
> table(PatientOverStage)

PatientOverStage
  I  II III  IV
  2   2  37   7

> table(PatientOverStage, cen.status)

               cen.status
PatientOverStage  0  1
             I    1  1
             II   0  2
             III  3 34
             IV   0  7

> table(PatientOverStage)/length(PatientOverStage)

PatientOverStage
         I         II        III         IV
0.04166667 0.04166667 0.77083333 0.14583333

> ##      combine stage I and II and III
>
>
> PatientOverStageCombine <- PatientOverStage
> PatientOverStageCombine[PatientOverStage == "I" | PatientOverStage == "II" | PatientOverStage
> table(PatientOverStageCombine)

PatientOverStageCombine
I to III       IV
      41        7

> table(PatientOverStageCombine, cen.status)

                     cen.status
PatientOverStageCombine  0  1
              I to III  4 37
              IV        0  7

> table(PatientOverStageCombine)/length(PatientOverStageCombine)
```

```
PatientOverStageCombine
 I to III        IV
0.8541667 0.1458333

> #######   chemo treatment or not #################
>
>
>
> PatientTreat <- as.vector(TumorSampleSurvivalData[,"Neoadjuvant.chemo"])
> table(PatientTreat)

PatientTreat
 No Yes
 37  11

> table(PatientTreat, cen.status)

            cen.status
PatientTreat  0  1
         No   4 33
         Yes  0 11

> table(PatientTreat)/length(PatientTreat)

PatientTreat
       No       Yes
0.7708333 0.2291667

>
>
>
```

We would like to check each clinical features independently first to check if any of the variables predict overall survival.

8.1 Function

In order to minimize the amount of code we need to rewrite, we use the function below to generate Kaplan-Meier plots.

```
> makeKMplot <- function(Time, Status, Clinical,
+                 main="Overall Survival")
+ {
+   degf <- length(levels(Clinical)) - 1
```

```
+   sf <- survfit(Surv(Time, Status) ~ Clinical)
+   sdv <- survdiff(Surv(Time, Status) ~ Clinical)
+
+
+   colset <- c("red", "blue", "green", "orange", "cyan", "magenta")
+   plot(sf, col=colset, xlab="Years", ylab="Probability of Overall Survival",main=main)
+
+   legend("topright", levels(Clinical), col=colset, lwd=2,
+         title = ifelse((1-pchisq(sdv$chisq, degf) > 0.0001), paste("P value =", round(1-pchis
+                 "P value < 0.0001 "))
+
+ }
>
>
```

8.2 Age

We calculated the age associated with the survival difference.

```
>         model0 <- coxph(Surv(Time.dfs, cen.status) ~ PatientAge, na.action=na.exclude)
>         summary(model0)

Call:
coxph(formula = Surv(Time.dfs, cen.status) ~ PatientAge, na.action = na.exclude)

  n= 48, number of events= 44

              coef exp(coef) se(coef)     z Pr(>|z|)
PatientAge 0.04725   1.04838  0.02329 2.029   0.0425 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

           exp(coef) exp(-coef) lower .95 upper .95
PatientAge     1.048     0.9539     1.002     1.097

Concordance= 0.624  (se = 0.05 )
Rsquare= 0.086   (max possible= 0.997 )
Likelihood ratio test= 4.32  on 1 df,   p=0.03766
Wald test            = 4.12  on 1 df,   p=0.04248
Score (logrank) test = 4.15  on 1 df,   p=0.04165

>         model<-survfit(coxph(Surv(Time.dfs, cen.status) ~ PatientAge, na.action=na.exclude))
>         model
```

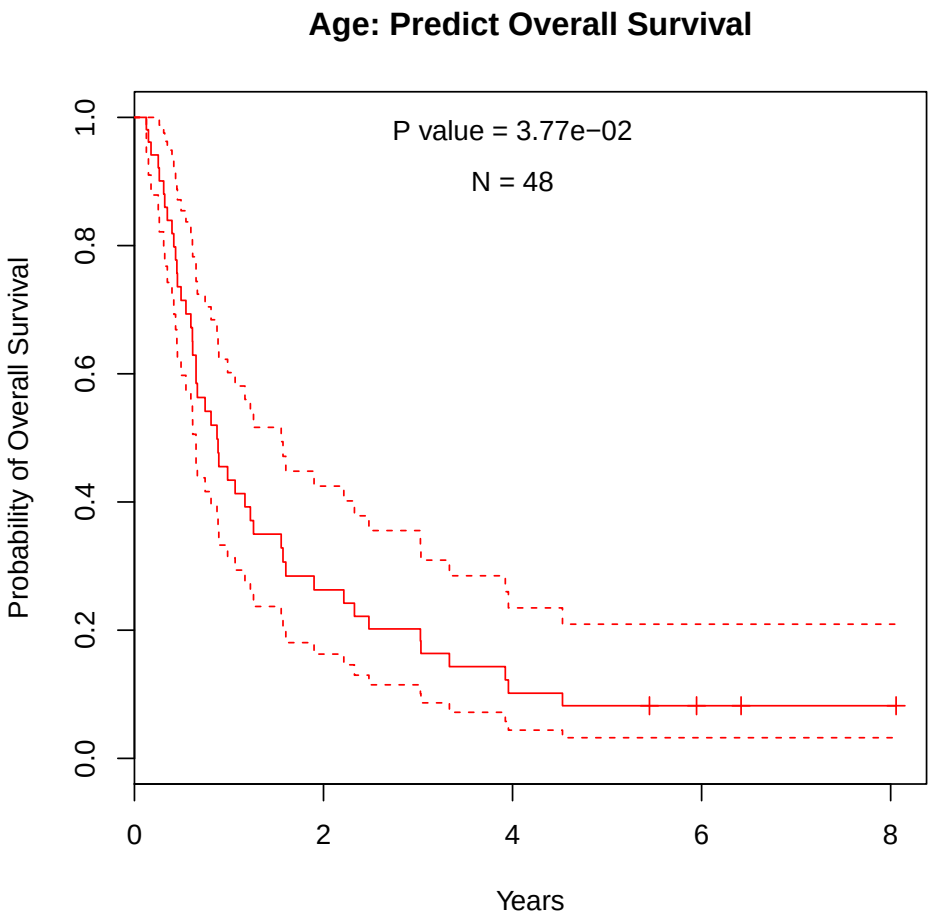

Figure 12: KM for overall survival predicted by age.

```
Call: survfit(formula = coxph(Surv(Time.dfs, cen.status) ~ PatientAge,
    na.action = na.exclude))

records   n.max n.start  events  median 0.95LCL 0.95UCL
 48.000  48.000  48.000  44.000   0.873   0.652   1.552

>
```

8.3 Gender

We compared the gender checking the survival difference.

```
>         model <- coxph(Surv(Time.dfs, cen.status) ~ as.factor(PatientGender), na.action=na.ex
>         summary(model)

Call:
coxph(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientGender),
    na.action = na.exclude)

  n= 48, number of events= 44

                               coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(PatientGender)Male 1.5625    4.7708   0.5433 2.876  0.00403 **
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

                             exp(coef) exp(-coef) lower .95 upper .95
as.factor(PatientGender)Male     4.771     0.2096     1.645     13.84

Concordance= 0.582  (se = 0.035 )
Rsquare= 0.221   (max possible= 0.997 )
Likelihood ratio test= 11.99  on 1 df,   p=0.0005346
Wald test            = 8.27  on 1 df,   p=0.004026
Score (logrank) test = 9.7  on 1 df,   p=0.001841

>         model <- survfit(Surv(Time.dfs, cen.status) ~ as.factor(PatientGender), na.action=na.
>         model

Call: survfit(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientGender),
    na.action = na.exclude)

                               records n.max n.start events median 0.95LCL 0.95UCL
as.factor(PatientGender)=Female       8     8       8      4  2.480   1.552      NA
as.factor(PatientGender)=Male        40    40      40     40  0.706   0.597    1.23

>         model2 <-survdiff(Surv(Time.dfs, cen.status) ~ as.factor(PatientGender), na.action=na
>         model2

Call:
survdiff(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientGender),
    na.action = na.exclude)

                                N Observed Expected (O-E)^2/E (O-E)^2/V
as.factor(PatientGender)=Female  8        4       13      6.21       9.7
as.factor(PatientGender)=Male   40       40       31      2.60       9.7

 Chisq= 9.7  on 1 degrees of freedom, p= 0.00184
```

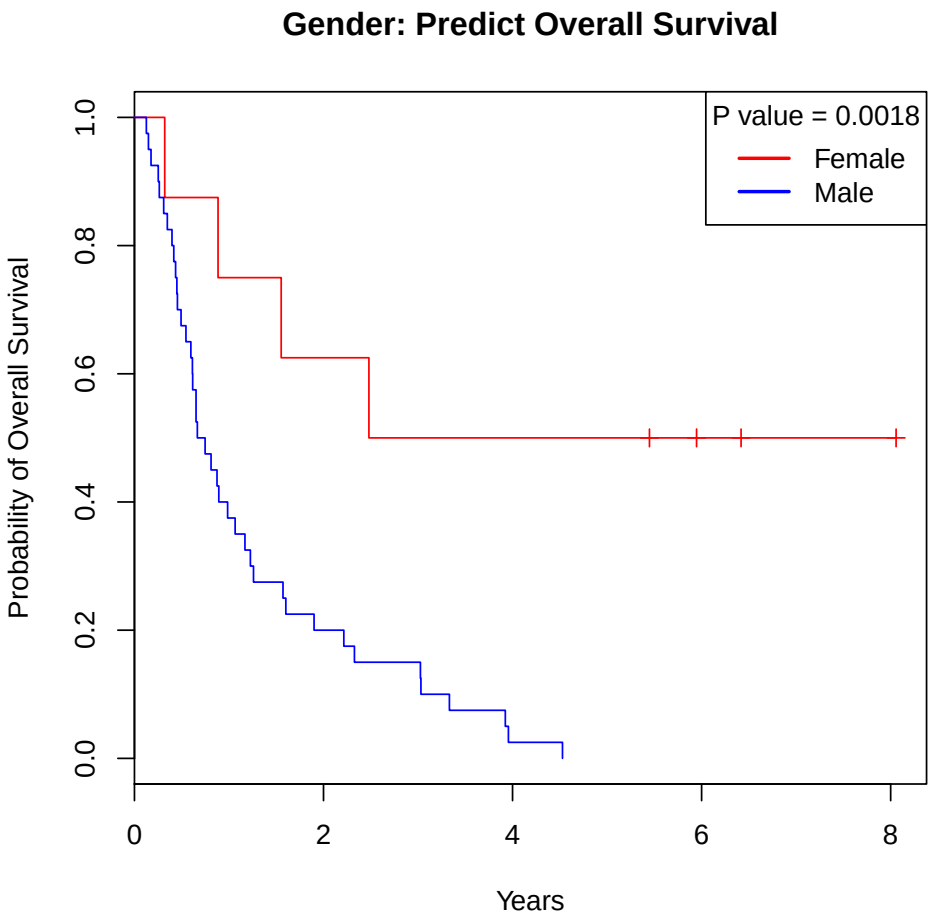

Figure 13: KM for overall survival predicted by gender.

```
>
```

8.4 Original Histology

We compared the original histology checking the survival difference.

```
>         model <- coxph(Surv(Time.dfs, cen.status) ~ as.factor(PatientOriginalDiag), na.actio
>         summary(model)

Call:
coxph(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientOriginalDiag),
```

```
    na.action = na.exclude)

  n= 48, number of events= 44

                                          coef exp(coef) se(coef)      z Pr(>|z|)
as.factor(PatientOriginalDiag)epithelioid -0.5753    0.5625   0.3584 -1.605    0.108
as.factor(PatientOriginalDiag)sarcomatoid -0.4522    0.6362   0.5439 -0.831    0.406

                                         exp(coef) exp(-coef) lower .95 upper .95
as.factor(PatientOriginalDiag)epithelioid    0.5625      1.778    0.2786     1.136
as.factor(PatientOriginalDiag)sarcomatoid    0.6362      1.572    0.2191     1.847

Concordance= 0.578  (se = 0.04 )
Rsquare= 0.048   (max possible= 0.997 )
Likelihood ratio test= 2.36  on 2 df,   p=0.3076
Wald test            = 2.58  on 2 df,   p=0.2749
Score (logrank) test = 2.65  on 2 df,   p=0.2658

>         model <- survfit(Surv(Time.dfs, cen.status) ~ as.factor(PatientOriginalDiag), na.acti
>         model

Call: survfit(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientOriginalDiag),
    na.action = na.exclude)

                                         records n.max n.start events median 0.95LCL
as.factor(PatientOriginalDiag)=biphasic         11    11      11     11  0.449   0.263
as.factor(PatientOriginalDiag)=epithelioid      31    31      31     28  1.065   0.665
as.factor(PatientOriginalDiag)=sarcomatoid        6     6       6      5  0.634   0.597
                                         0.95UCL
as.factor(PatientOriginalDiag)=biphasic         NA
as.factor(PatientOriginalDiag)=epithelioid    2.33
as.factor(PatientOriginalDiag)=sarcomatoid      NA

>         model2 <-survdiff(Surv(Time.dfs, cen.status) ~ as.factor(PatientOriginalDiag), na.act
>         model2

Call:
survdiff(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientOriginalDiag),
    na.action = na.exclude)

                                          N Observed Expected (O-E)^2/E (O-E)^2/V
as.factor(PatientOriginalDiag)=biphasic    11       11     7.09  2.161367  2.606452
as.factor(PatientOriginalDiag)=epithelioid 31       28    31.86  0.468750  1.720596
```

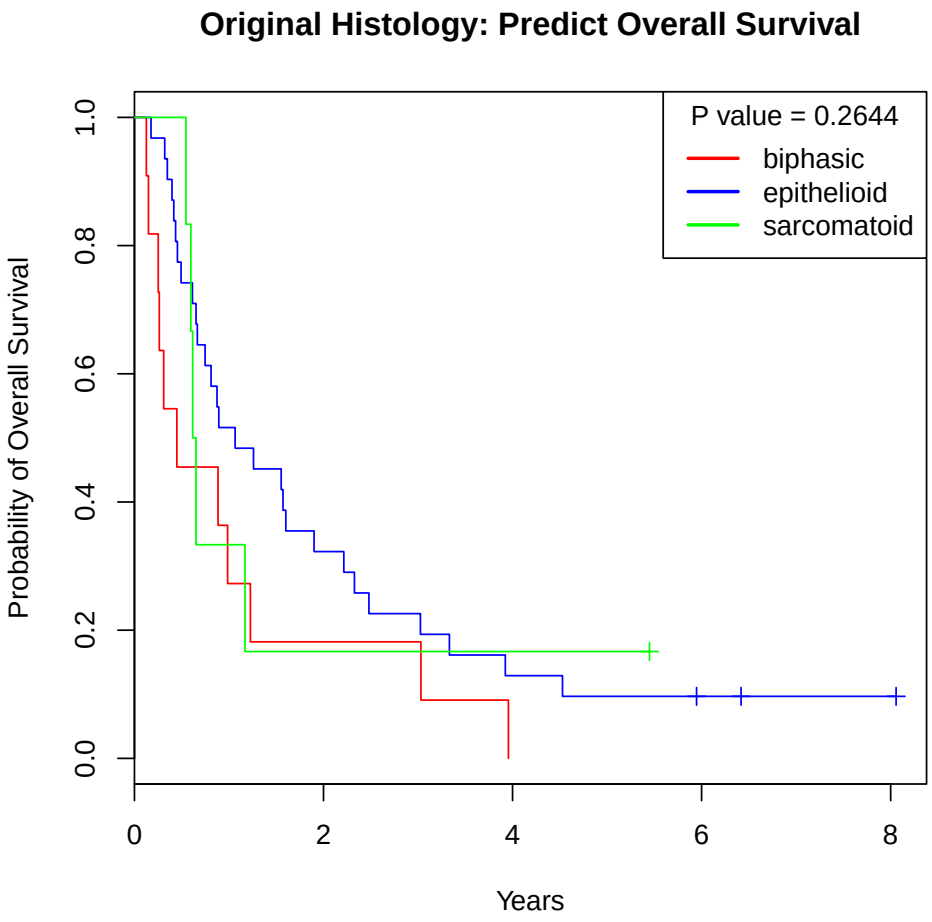

Figure 14: KM for overall survival predicted by original histology.

```
as.factor(PatientOriginalDiag)=sarcomatoid  6        5     5.05  0.000472  0.000541

 Chisq= 2.7  on 2 degrees of freedom, p= 0.264
```

8.5 T Stage

We compared the T stage checking the survival difference.

```
>         model <- coxph(Surv(Time.dfs, cen.status) ~ as.factor(PatientTStage), na.action=na.ex
>         summary(model)
```

```
Call:
coxph(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientTStage),
    na.action = na.exclude)

  n= 48, number of events= 44

                             coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(PatientTStage)T2  2.883    17.876    1.126 2.560   0.0105 *
as.factor(PatientTStage)T3  1.424     4.153    1.018 1.398   0.1621
as.factor(PatientTStage)T4  1.983     7.266    1.111 1.785   0.0743 .
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

                           exp(coef) exp(-coef) lower .95 upper .95
as.factor(PatientTStage)T2    17.876    0.05594    1.9664    162.51
as.factor(PatientTStage)T3     4.153    0.24081    0.5642     30.57
as.factor(PatientTStage)T4     7.266    0.13762    0.8229     64.17

Concordance= 0.603  (se = 0.036 )
Rsquare= 0.195   (max possible= 0.997 )
Likelihood ratio test= 10.43  on 3 df,   p=0.01526
Wald test            = 10.73  on 3 df,   p=0.01329
Score (logrank) test = 12.77  on 3 df,   p=0.005157

>         model <- survfit(Surv(Time.dfs, cen.status) ~ as.factor(PatientTStage), na.action=na.
>         model

Call: survfit(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientTStage),
    na.action = na.exclude)

                            records n.max n.start events median 0.95LCL 0.95UCL
as.factor(PatientTStage)=T1       2     2       2      1  3.031   3.031      NA
as.factor(PatientTStage)=T2       5     5       5      5  0.397   0.309      NA
as.factor(PatientTStage)=T3      36    36      36     33  0.939   0.652     1.9
as.factor(PatientTStage)=T4       5     5       5      5  0.616   0.545      NA

>         model2 <-survdiff(Surv(Time.dfs, cen.status) ~ as.factor(PatientTStage), na.action=na
>         model2

Call:
survdiff(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientTStage),
    na.action = na.exclude)
```

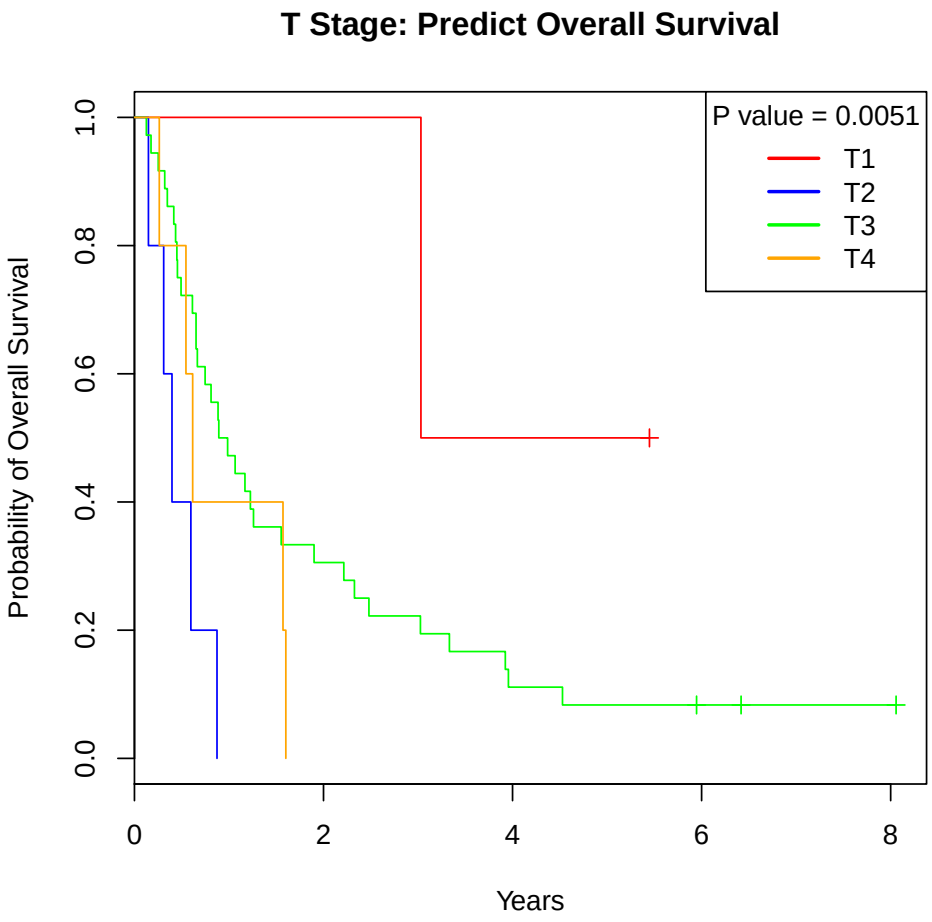

Figure 15: KM for overall survival predicted by T Stage.

```
                              N Observed Expected (O-E)^2/E (O-E)^2/V
as.factor(PatientTStage)=T1   2        1     4.11     2.357     2.680
as.factor(PatientTStage)=T2   5        5     1.46     8.568     9.229
as.factor(PatientTStage)=T3  36       33    35.14     0.130     0.650
as.factor(PatientTStage)=T4   5        5     3.29     0.891     0.991

 Chisq= 12.8  on 3 degrees of freedom, p= 0.00507

>
```

We also consider the combined T stage because of the small number of T 1 stage patients

```
>         model <- coxph(Surv(Time.dfs, cen.status) ~ as.factor(PatientTStageCombine), na.actio
>         summary(model)

Call:
coxph(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientTStageCombine),
    na.action = na.exclude)

  n= 48, number of events= 44

                                     coef exp(coef) se(coef)      z Pr(>|z|)
as.factor(PatientTStageCombine)T3 -0.1327    0.8757   0.4455 -0.298    0.766
as.factor(PatientTStageCombine)T4  0.3625    1.4370   0.6175  0.587    0.557

                                  exp(coef) exp(-coef) lower .95 upper .95
as.factor(PatientTStageCombine)T3    0.8757     1.1419    0.3657     2.097
as.factor(PatientTStageCombine)T4    1.4370     0.6959    0.4284     4.820

Concordance= 0.54  (se = 0.036 )
Rsquare= 0.02   (max possible= 0.997 )
Likelihood ratio test= 0.95  on 2 df,   p=0.6216
Wald test            = 1.05  on 2 df,   p=0.5909
Score (logrank) test = 1.07  on 2 df,   p=0.5852

>         model <- survfit(Surv(Time.dfs, cen.status) ~ as.factor(PatientTStageCombine), na.act
>         model

Call: survfit(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientTStageCombine),
    na.action = na.exclude)

                                        records n.max n.start events median 0.95LCL
as.factor(PatientTStageCombine)=T1 or T2       7     7       7      6  0.597   0.309
as.factor(PatientTStageCombine)=T3            36    36      36     33  0.939   0.652
as.factor(PatientTStageCombine)=T4             5     5       5      5  0.616   0.545
                                        0.95UCL
as.factor(PatientTStageCombine)=T1 or T2      NA
as.factor(PatientTStageCombine)=T3           1.9
as.factor(PatientTStageCombine)=T4            NA

>         model2 <-survdiff(Surv(Time.dfs, cen.status) ~ as.factor(PatientTStageCombine), na.ac
>         model2

Call:
survdiff(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientTStageCombine),
```

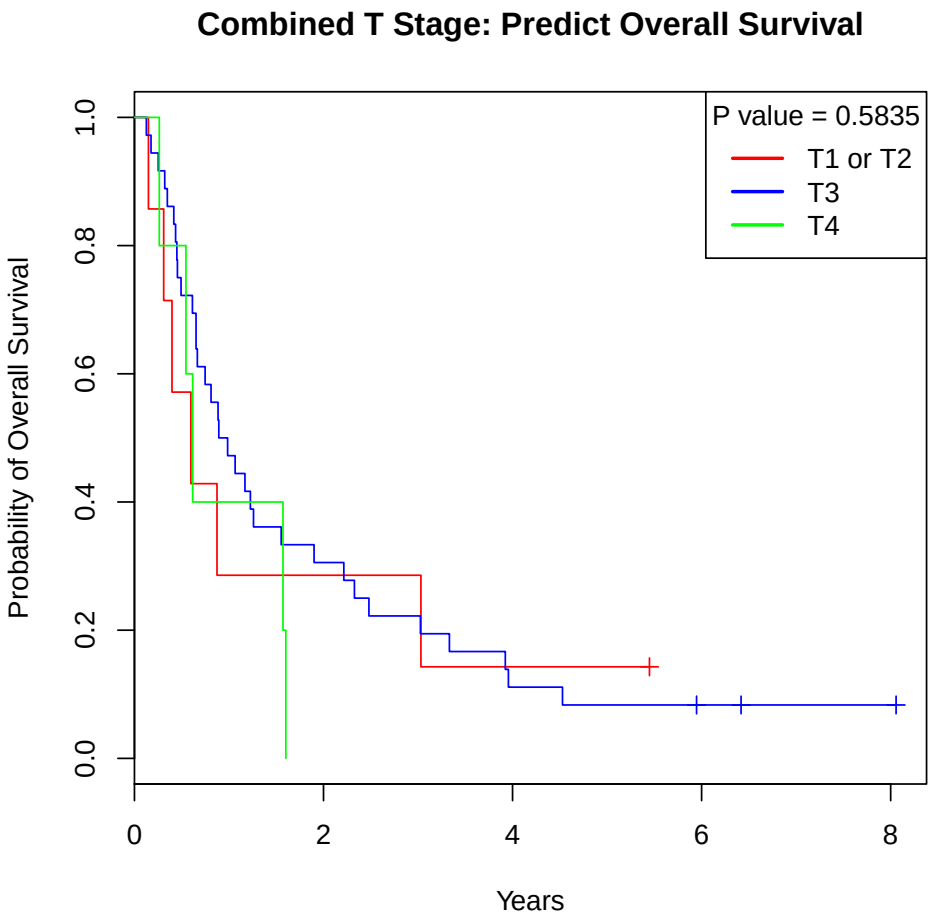

Figure 16: KM for overall survival predicted by combined T Stage.

```
    na.action = na.exclude)

                                          N Observed Expected (O-E)^2/E (O-E)^2/V
as.factor(PatientTStageCombine)=T1 or T2  7        6     5.58    0.0323    0.0374
as.factor(PatientTStageCombine)=T3       36       33    35.14    0.1299    0.6497
as.factor(PatientTStageCombine)=T4        5        5     3.29    0.8913    0.9909

 Chisq= 1.1  on 2 degrees of freedom, p= 0.583

>
```

8.6 N Stage

We compared the N stage checking the survival difference.

```
>         model <- coxph(Surv(Time.dfs, cen.status) ~ as.factor(PatientNStage), na.action=na.ex
>         summary(model)

Call:
coxph(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientNStage),
    na.action = na.exclude)

  n= 48, number of events= 44

                             coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(PatientNStage)N1 0.2220    1.2485   0.4696 0.473  0.63648
as.factor(PatientNStage)N2 1.0446    2.8424   0.3662 2.853  0.00434 **
as.factor(PatientNStage)N3 1.6313    5.1103   0.7759 2.102  0.03552 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

                           exp(coef) exp(-coef) lower .95 upper .95
as.factor(PatientNStage)N1     1.249     0.8009    0.4973     3.134
as.factor(PatientNStage)N2     2.842     0.3518    1.3867     5.826
as.factor(PatientNStage)N3     5.110     0.1957    1.1168    23.383

Concordance= 0.618  (se = 0.045 )
Rsquare= 0.184   (max possible= 0.997 )
Likelihood ratio test= 9.76  on 3 df,   p=0.02075
Wald test            = 10.36  on 3 df,   p=0.01576
Score (logrank) test = 11.38  on 3 df,   p=0.009861

>         model <- survfit(Surv(Time.dfs, cen.status) ~ as.factor(PatientNStage), na.action=na.
>         model

Call: survfit(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientNStage),
    na.action = na.exclude)

                            records n.max n.start events median 0.95LCL 0.95UCL
as.factor(PatientNStage)=N0      23    23      23     20  1.227   0.652    3.92
as.factor(PatientNStage)=N1       7     7       7      6  1.259   0.416      NA
as.factor(PatientNStage)=N2      16    16      16     16  0.615   0.435    1.55
as.factor(PatientNStage)=N3       2     2       2      2  0.461   0.175      NA

>         model2 <-survdiff(Surv(Time.dfs, cen.status) ~ as.factor(PatientNStage), na.action=na
>         model2
```

```
Call:
survdiff(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientNStage),
    na.action = na.exclude)

                              N Observed Expected (O-E)^2/E (O-E)^2/V
as.factor(PatientNStage)=N0 23       20   27.481     2.037     5.855
as.factor(PatientNStage)=N1  7        6    6.896     0.116     0.139
as.factor(PatientNStage)=N2 16       16    8.956     5.540     7.536
as.factor(PatientNStage)=N3  2        2    0.667     2.663     2.760

 Chisq= 11.4  on 3 degrees of freedom, p= 0.00959

>
```

We also consider the combined N stage because of the small number of N 3 stage patients

```
>         model <- coxph(Surv(Time.dfs, cen.status) ~ as.factor(PatientNStageCombine), na.actio
>         summary(model)

Call:
coxph(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientNStageCombine),
    na.action = na.exclude)

  n= 48, number of events= 44

                                       coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(PatientNStageCombine)N1       0.2209    1.2472   0.4696 0.470  0.63806
as.factor(PatientNStageCombine)N2 or N3 1.0885    2.9698   0.3576 3.044  0.00234 **
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

                                      exp(coef) exp(-coef) lower .95 upper .95
as.factor(PatientNStageCombine)N1         1.247     0.8018    0.4968     3.131
as.factor(PatientNStageCombine)N2 or N3     2.970     0.3367    1.4733     5.986

Concordance= 0.613  (se = 0.045 )
Rsquare= 0.175   (max possible= 0.997 )
Likelihood ratio test= 9.24  on 2 df,   p=0.009846
Wald test            = 9.62  on 2 df,   p=0.008156
Score (logrank) test = 10.38  on 2 df,   p=0.005561

>         model <- survfit(Surv(Time.dfs, cen.status) ~ as.factor(PatientNStageCombine), na.act
>         model
```

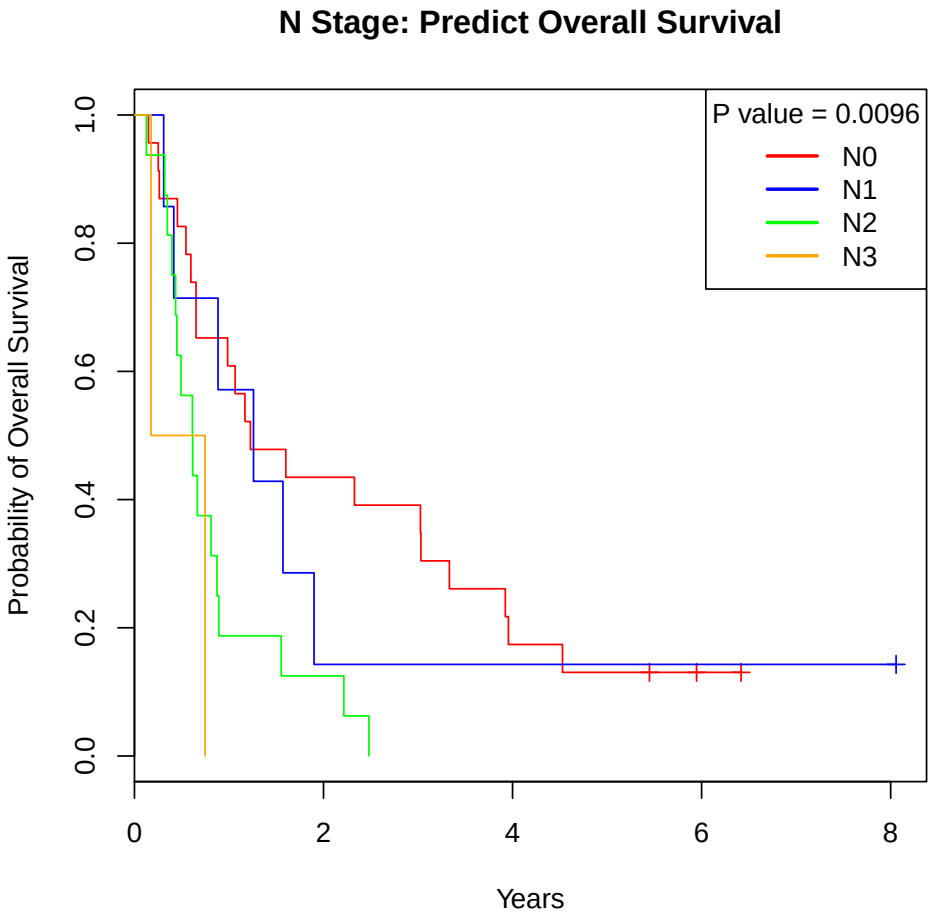

Figure 17: KM for overall survival predicted by N Stage.

```
Call: survfit(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientNStageCombine),
    na.action = na.exclude)

                                       records n.max n.start events median 0.95LCL
as.factor(PatientNStageCombine)=N0          23    23      23     20  1.227   0.652
as.factor(PatientNStageCombine)=N1           7     7       7      6  1.259   0.416
as.factor(PatientNStageCombine)=N2 or N3    18    18      18     18  0.615   0.435
                                       0.95UCL
as.factor(PatientNStageCombine)=N0       3.923
as.factor(PatientNStageCombine)=N1          NA
as.factor(PatientNStageCombine)=N2 or N3 0.893
```

```
> model2 <-survdiff(Surv(Time.dfs, cen.status) ~ as.factor(PatientNStageCombine), na.ac
> model2
```

```
Call:
survdiff(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientNStageCombine),
    na.action = na.exclude)

                                        N Observed Expected (O-E)^2/E (O-E)^2/V
as.factor(PatientNStageCombine)=N0     23       20    27.48     2.037     5.855
as.factor(PatientNStageCombine)=N1      7        6     6.90     0.116     0.139
as.factor(PatientNStageCombine)=N2 or N3 18     18     9.62     7.292    10.264

 Chisq= 10.4  on 2 degrees of freedom, p= 0.00542

>
```

8.7 Overall Stage

We compared the overall stage checking the survival difference.

```
> model <- coxph(Surv(Time.dfs, cen.status) ~ as.factor(PatientOverStage), na.action=na
> summary(model)
```

```
Call:
coxph(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientOverStage),
    na.action = na.exclude)

  n= 48, number of events= 44

                                coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(PatientOverStage)II  3.147    23.256    1.260 2.496   0.0125 *
```

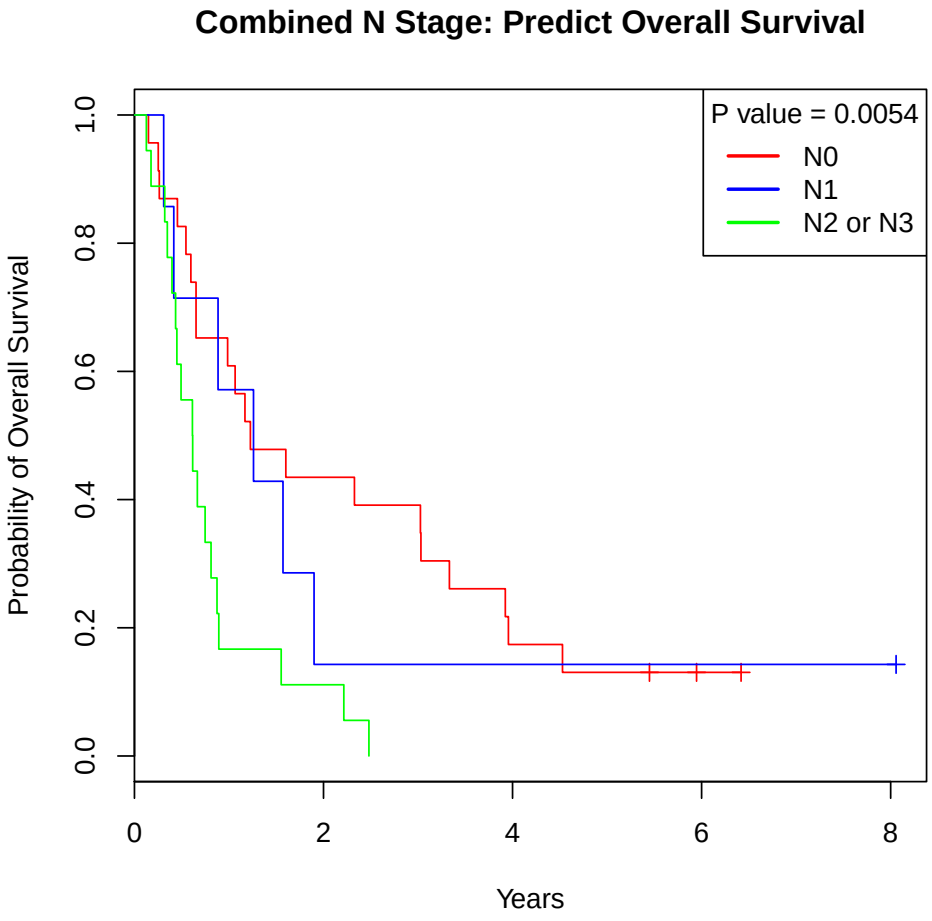

Figure 18: KM for overall survival predicted by combined N Stage.

```
as.factor(PatientOverStage)III  1.433     4.193    1.018 1.408   0.1591
as.factor(PatientOverStage)IV   2.124     8.367    1.086 1.957   0.0504 .
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

                              exp(coef) exp(-coef) lower .95 upper .95
as.factor(PatientOverStage)II     23.256     0.0430    1.9663    275.06
as.factor(PatientOverStage)III     4.193     0.2385    0.5701     30.84
as.factor(PatientOverStage)IV      8.367     0.1195    0.9965     70.25

Concordance= 0.594  (se = 0.035 )
Rsquare= 0.17   (max possible= 0.997 )
Likelihood ratio test= 8.94  on 3 df,   p=0.03013
Wald test            = 9.01  on 3 df,   p=0.02916
Score (logrank) test = 10.85  on 3 df,   p=0.01256

>         model <- survfit(Surv(Time.dfs, cen.status) ~ as.factor(PatientOverStage), na.action=
>         model

Call: survfit(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientOverStage),
    na.action = na.exclude)

                               records n.max n.start events median 0.95LCL 0.95UCL
as.factor(PatientOverStage)=I        2     2       2      1  3.031   3.031      NA
as.factor(PatientOverStage)=II       2     2       2      2  0.372   0.148      NA
as.factor(PatientOverStage)=III     37    37      37     34  0.893   0.652     1.9
as.factor(PatientOverStage)=IV       7     7       7      7  0.616   0.263      NA

>         model2 <-survdiff(Surv(Time.dfs, cen.status) ~ as.factor(PatientOverStage), na.actio
>         model2

Call:
survdiff(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientOverStage),
    na.action = na.exclude)

                                N Observed Expected (O-E)^2/E (O-E)^2/V
as.factor(PatientOverStage)=I   2        1    4.114    2.3572     2.680
as.factor(PatientOverStage)=II  2        2    0.442    5.4838     5.647
as.factor(PatientOverStage)=III 37      34   35.488    0.0624     0.325
as.factor(PatientOverStage)=IV  7        7    3.955    2.3440     2.670

 Chisq= 10.9  on 3 degrees of freedom, p= 0.0125

>
```

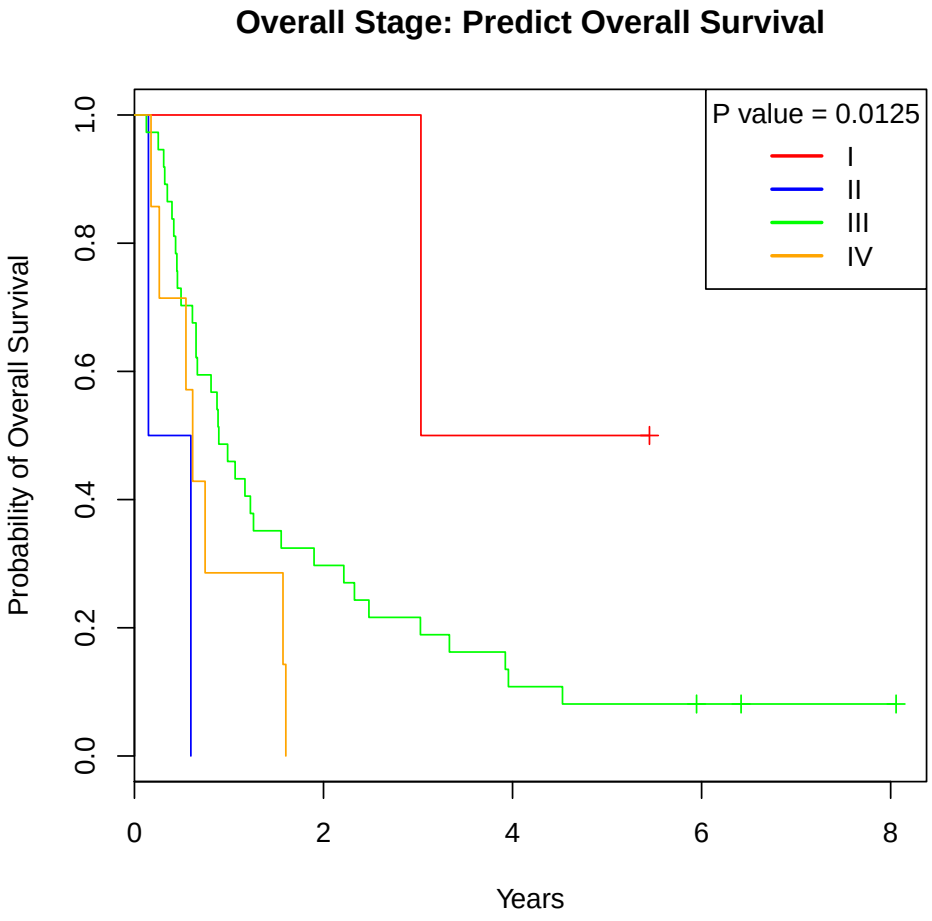

Figure 19: KM for overall survival predicted by Overall Stage.

We also consider the combined Overall stage because of the small number of overall 1 and 2 stage patients

```
>         model <- coxph(Surv(Time.dfs, cen.status) ~ as.factor(PatientOverStageCombine), na.ac
>         summary(model)

Call:
coxph(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientOverStageCombine),
    na.action = na.exclude)

  n= 48, number of events= 44

                                       coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(PatientOverStageCombine)IV 0.6809    1.9757   0.4258 1.599     0.11

                                     exp(coef) exp(-coef) lower .95 upper .95
as.factor(PatientOverStageCombine)IV     1.976     0.5061    0.8576     4.552

Concordance= 0.537  (se = 0.028 )
Rsquare= 0.045   (max possible= 0.997 )
Likelihood ratio test= 2.23  on 1 df,   p=0.1354
Wald test            = 2.56  on 1 df,   p=0.1098
Score (logrank) test = 2.66  on 1 df,   p=0.1031

>         model <- survfit(Surv(Time.dfs, cen.status) ~ as.factor(PatientOverStageCombine), na.
>         model

Call: survfit(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientOverStageCombine),
    na.action = na.exclude)

                                         records n.max n.start events median 0.95LCL
as.factor(PatientOverStageCombine)=I to III      41    41      41     37  0.893   0.652
as.factor(PatientOverStageCombine)=IV             7     7       7      7  0.616   0.263
                                         0.95UCL
as.factor(PatientOverStageCombine)=I to III     1.9
as.factor(PatientOverStageCombine)=IV            NA

>         model2 <-survdiff(Surv(Time.dfs, cen.status) ~ as.factor(PatientOverStageCombine), na
>         model2

Call:
survdiff(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientOverStageCombine),
    na.action = na.exclude)
```

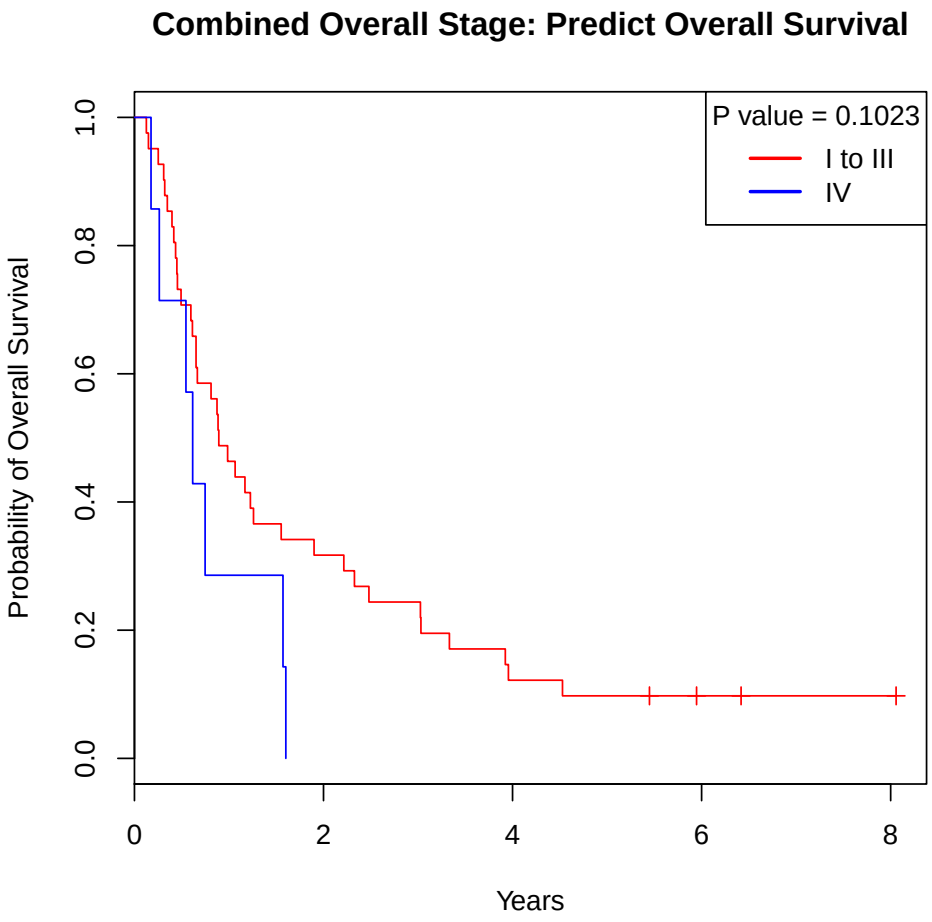

Figure 20: KM for overall survival predicted by combined Overall Stage.

```
                                          N Observed Expected (O-E)^2/E (O-E)^2/V
as.factor(PatientOverStageCombine)=I to III 41       37    40.04     0.232      2.67
as.factor(PatientOverStageCombine)=IV        7        7     3.96     2.344      2.67

 Chisq= 2.7  on 1 degrees of freedom, p= 0.102

>
```

8.8 Chemo Treatment

We compared the chemo treatment checking the survival difference.

```
>         model <- coxph(Surv(Time.dfs, cen.status) ~ as.factor(PatientTreat), na.action=na.exc
>         summary(model)

Call:
coxph(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientTreat),
    na.action = na.exclude)

  n= 48, number of events= 44

                             coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(PatientTreat)Yes 1.3787    3.9697   0.3859 3.573 0.000353 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

                           exp(coef) exp(-coef) lower .95 upper .95
as.factor(PatientTreat)Yes      3.97     0.2519     1.863     8.457

Concordance= 0.612  (se = 0.03 )
Rsquare= 0.202   (max possible= 0.997 )
Likelihood ratio test= 10.81  on 1 df,   p=0.00101
Wald test            = 12.77  on 1 df,   p=0.000353
Score (logrank) test = 14.75  on 1 df,   p=0.0001226

>         model <- survfit(Surv(Time.dfs, cen.status) ~ as.factor(PatientTreat), na.action=na.e
>         model

Call: survfit(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientTreat),
    na.action = na.exclude)

                           records n.max n.start events median 0.95LCL 0.95UCL
as.factor(PatientTreat)=No      37    37      37     33  1.169    0.81    2.33
as.factor(PatientTreat)=Yes     11    11      11     11  0.435    0.32      NA

>         model2 <-survdiff(Surv(Time.dfs, cen.status) ~ as.factor(PatientTreat), na.action=na.
>         model2

Call:
survdiff(formula = Surv(Time.dfs, cen.status) ~ as.factor(PatientTreat),
    na.action = na.exclude)

                            N Observed Expected (O-E)^2/E (O-E)^2/V
as.factor(PatientTreat)=No  37       33    40.03      1.24      14.7
as.factor(PatientTreat)=Yes 11       11     3.97     12.47      14.7

 Chisq= 14.7  on 1 degrees of freedom, p= 0.000126
```

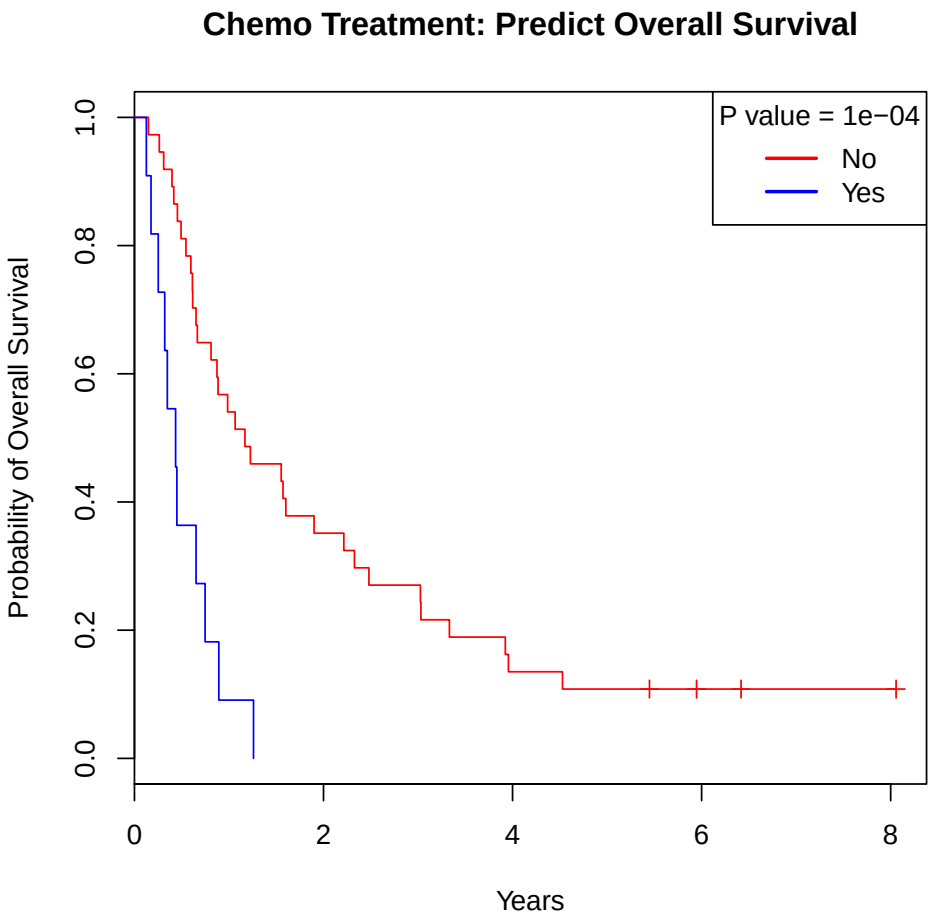

Figure 21: KM for overall survival predicted by chemo treatment.

>

8.9 Overall Survival Analysis: Multivariate Analysis for Different Clinical Variables

Here we perform a multivariate analysis to identify the selected clinical variables (significant from univariate model and have been reduced at smaller levels) that best explain overall survival. We exclude shaving diagnosis histology because it is not significant at the univariate level.

```
> ###############   create the data
> dataset <- data.frame(Time.dfs, cen.status, PatientAge, PatientGender,
```

```
+                 PatientOriginalDiag, PatientTStageCombine, PatientNStageCombine, PatientOverS
> ################   fit in the model
> mod0 <- coxph(Surv(Time.dfs, cen.status) ~ ., data = na.omit(dataset))
>
```

Now we use the Akaike Information Criterion (AIC) to eliminate redundant variables from the model.

```
>  mod1 <- step(mod0)

Start:  AIC=258.57
Surv(Time.dfs, cen.status) ~ PatientAge + PatientGender + PatientOriginalDiag +
    PatientTStageCombine + PatientNStageCombine + PatientOverStageCombine +
    PatientTreat

                          Df    AIC
- PatientTStageCombine     2 254.96
- PatientOverStageCombine  1 256.62
- PatientAge               1 257.71
- PatientOriginalDiag      2 257.82
<none>                       258.57
- PatientTreat             1 258.85
- PatientGender            1 265.01
- PatientNStageCombine     2 266.39

Step:  AIC=254.96
Surv(Time.dfs, cen.status) ~ PatientAge + PatientGender + PatientOriginalDiag +
    PatientNStageCombine + PatientOverStageCombine + PatientTreat

                          Df    AIC
- PatientOverStageCombine  1 253.12
- PatientAge               1 254.25
- PatientTreat             1 254.94
<none>                       254.96
- PatientOriginalDiag      2 255.22
- PatientGender            1 262.60
- PatientNStageCombine     2 262.90

Step:  AIC=253.13
Surv(Time.dfs, cen.status) ~ PatientAge + PatientGender + PatientOriginalDiag +
    PatientNStageCombine + PatientTreat

                      Df    AIC
```

```
- PatientAge            1 252.61
- PatientTreat          1 253.05
<none>                    253.12
- PatientOriginalDiag   2 253.32
- PatientGender         1 261.90
- PatientNStageCombine  2 262.13

Step:  AIC=252.61
Surv(Time.dfs, cen.status) ~ PatientGender + PatientOriginalDiag +
    PatientNStageCombine + PatientTreat

                       Df    AIC
<none>                    252.61
- PatientTreat          1 253.91
- PatientOriginalDiag   2 254.97
- PatientNStageCombine  2 260.68
- PatientGender         1 262.37

>
```

Here is the Final model:

```
> summary(mod1)

Call:
coxph(formula = Surv(Time.dfs, cen.status) ~ PatientGender +
    PatientOriginalDiag + PatientNStageCombine + PatientTreat,
    data = na.omit(dataset))

  n= 48, number of events= 44

                                   coef exp(coef) se(coef)      z Pr(>|z|)
PatientGenderMale               1.69855   5.46603  0.57978  2.930 0.003394 **
PatientOriginalDiagepithelioid -0.92876   0.39504  0.40654 -2.285 0.022340 *
PatientOriginalDiagsarcomatoid  0.04285   1.04378  0.57411  0.075 0.940501
PatientNStageCombineN1          0.81833   2.26672  0.54586  1.499 0.133832
PatientNStageCombineN2 or N3    1.55030   4.71291  0.44434  3.489 0.000485 ***
PatientTreatYes                 0.76645   2.15212  0.41034  1.868 0.061784 .
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

                               exp(coef) exp(-coef) lower .95 upper .95
PatientGenderMale                  5.466     0.1829    1.7545   17.0287
```

```
PatientOriginalDiagepithelioid     0.395     2.5314    0.1781    0.8764
PatientOriginalDiagsarcomatoid     1.044     0.9581    0.3388    3.2158
PatientNStageCombineN1             2.267     0.4412    0.7776    6.6074
PatientNStageCombineN2 or N3       4.713     0.2122    1.9727   11.2593
PatientTreatYes                    2.152     0.4647    0.9629    4.8101

Concordance= 0.741  (se = 0.05 )
Rsquare= 0.511   (max possible= 0.997 )
Likelihood ratio test= 34.38  on 6 df,   p=5.685e-06
Wald test            = 29.78  on 6 df,   p=4.336e-05
Score (logrank) test = 34.74  on 6 df,   p=4.845e-06
```

```
> anova(mod1)

Analysis of Deviance Table
 Cox model: response is Surv(Time.dfs, cen.status)
Terms added sequentially (first to last)

                      loglik   Chisq Df Pr(>|Chi|)
NULL                 -137.50
PatientGender        -131.50 11.9909  1  0.0005346 ***
PatientOriginalDiag  -130.84  1.3117  2  0.5190077
PatientNStageCombine -121.95 17.7846  2  0.0001374 ***
PatientTreat         -120.31  3.2912  1  0.0696534 .
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

In the final model, gender, original diagnosis, combined N stage and chemo treatment are included.

8.10 Overall Survival Analysis: Multivariate Analysis with sub-tumor Groups and Clinical Variables

In the previous section, we have selected the final model including the clinical variables for the survival analysis. In this section, we include the three sub-tumor groups we have defined in the earlier report and check the model.

```
> ###############   create the data
> dataset <- data.frame(Time.dfs, cen.status,  PatientGender,PatientOriginalDiag,
+                 PatientNStageCombine, PatientTreat, groupused)
> ###############   fit in the model
> mod0 <- coxph(Surv(Time.dfs, cen.status) ~ ., data = na.omit(dataset))
> summary(mod0)
```

```
Call:
coxph(formula = Surv(Time.dfs, cen.status) ~ ., data = na.omit(dataset))

  n= 48, number of events= 44

                                  coef exp(coef) se(coef)      z Pr(>|z|)
PatientGenderMale               2.1814    8.8587   0.6350  3.435 0.000592 ***
PatientOriginalDiagepithelioid -1.1243    0.3249   0.4876 -2.306 0.021134 *
PatientOriginalDiagsarcomatoid -0.5639    0.5690   0.6239 -0.904 0.366111
PatientNStageCombineN1          0.9290    2.5320   0.6310  1.472 0.140942
PatientNStageCombineN2 or N3    1.6623    5.2715   0.5500  3.022 0.002507 **
PatientTreatYes                 1.2098    3.3527   0.6579  1.839 0.065927 .
groupusedGroup 2               -0.6455    0.5244   0.6252 -1.032 0.301864
groupusedGroup 3                0.9981    2.7130   0.7115  1.403 0.160706
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

                               exp(coef) exp(-coef) lower .95 upper .95
PatientGenderMale                 8.8587     0.1129    2.5517   30.7545
PatientOriginalDiagepithelioid    0.3249     3.0780    0.1249    0.8449
PatientOriginalDiagsarcomatoid    0.5690     1.7575    0.1675    1.9328
PatientNStageCombineN1            2.5320     0.3949    0.7351    8.7210
PatientNStageCombineN2 or N3      5.2715     0.1897    1.7939   15.4911
PatientTreatYes                   3.3527     0.2983    0.9235   12.1725
groupusedGroup 2                  0.5244     1.9069    0.1540    1.7858
groupusedGroup 3                  2.7130     0.3686    0.6727   10.9422

Concordance= 0.771  (se = 0.05 )
Rsquare= 0.597   (max possible= 0.997 )
Likelihood ratio test= 43.57  on 8 df,   p=6.868e-07
Wald test            = 35.92  on 8 df,   p=1.817e-05
Score (logrank) test = 43.59  on 8 df,   p=6.793e-07

> anova(mod0)

Analysis of Deviance Table
 Cox model: response is Surv(Time.dfs, cen.status)
Terms added sequentially (first to last)

                     loglik   Chisq Df Pr(>|Chi|)
NULL                -137.50
PatientGender       -131.50 11.9909  1  0.0005346 ***
PatientOriginalDiag -130.84  1.3117  2  0.5190077
```

```
PatientNStageCombine -121.95 17.7846  2  0.0001374 ***
PatientTreat         -120.31  3.2912  1  0.0696534 .
groupused            -115.71  9.1888  2  0.0101083 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

>
```

We use the Akaike Information Criterion (AIC) to eliminate redundant variables from the model. The same model stays.

```
>  mod1 <- step(mod0)

Start:  AIC=247.42
Surv(Time.dfs, cen.status) ~ PatientGender + PatientOriginalDiag +
    PatientNStageCombine + PatientTreat + groupused

                       Df    AIC
<none>                    247.43
- PatientTreat          1 248.64
- PatientOriginalDiag   2 248.69
- groupused             2 252.61
- PatientNStageCombine  2 253.64
- PatientGender         1 262.05

>
```

9 Appendix

This analysis was run in the following directory:

```
> getwd()

[1] "/data/bioinfo/Private/LungSpore/ReportWithNewClin"
```

This analysis was run in the following software environment:

```
> sessionInfo()

R version 2.15.1 (2012-06-22)
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] grid      splines   parallel  stats     graphics  grDevices utils     datasets
 [9] methods   base

other attached packages:
 [1] RColorBrewer_1.0-5     survival_2.37-4        nlme_3.1-108
 [4] affyio_1.24.0          gplots_2.11.0          MASS_7.3-23
 [7] KernSmooth_2.23-10     caTools_1.14           gdata_2.12.0.2
[10] gtools_2.7.1           hgu133plus2.db_2.7.1   org.Hs.eg.db_2.7.1
[13] RSQLite_0.11.3         DBI_0.2-6              limma_3.14.0
[16] ClassDiscovery_2.10.2  mclust_4.1             cluster_1.14.4
[19] ClassComparison_2.10.1 PreProcess_2.10.1      oompaBase_2.12.0
[22] xtable_1.7-1           geneplotter_1.34.0     lattice_0.20-15
[25] annotate_1.34.1        AnnotationDbi_1.22.5   simpleaffy_2.32.0
[28] gcrma_2.28.0           BiocInstaller_1.4.9    genefilter_1.38.0
[31] affy_1.34.0            Biobase_2.16.0         BiocGenerics_0.6.0

loaded via a namespace (and not attached):
[1] Biostrings_2.24.1     bitops_1.0-5          IRanges_1.18.0
[4] preprocessCore_1.18.0 stats4_2.15.1         tools_2.15.1
[7] XML_3.96-1.1          zlibbioc_1.2.0
```