Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival

Contents

Introduction

This work arises with the aim of developing scripts in R with the assistance of Bioconductor packages for the analysis of gene expression data. The data being analyzed come from a study related to the gene expression profile of tobacco consumption and its role in the development of lung adenocarcinoma (Landi et al. 2008).

Lung adenocarcinoma is a common form of cancer in the lungs, found in smokers, although it can also develop in people who have never smoked. Typically, it starts in the outer tissues of the lung, remaining undetectable for long periods of time. It presents as one of the most common and deadly types of cancer. However, the molecular changes induced by tobacco consumption that lead to cancer development are not yet fully understood (Subramanian and Govindan 2007). Therefore, studying and understanding these mechanisms is important as it will enable the identification of genes that may be associated with cancer development and, consequently, identify potential targets for future treatment.

The study on which this work is based included 105 individuals with lung adenocarcinoma aged between 44 and 72 years, belonging to the Environment And Genetics in Lung cancer Etiology (EAGLE), another study focused on lung cancer in the Lombardy region of Italy. Gene expression data were obtained using HG-U133A microarrays from Affimetrix, comprising 135 samples of normal tissue (**NT**) and adenocarcinoma (**T**), from individuals who are current smokers (**C**), former smokers (**F**), and never smokers (**N**). After preprocessing by the authors, the final dataset is already normalized, totaling 107 samples, with 58 corresponding to tumor tissues and 49 to non-tumor tissues, from 20 never smokers, 26 former smokers, and 28 current smokers. The dataset used in this work is available in the GEO database of NCBI under the accession code GDS3257.

Loading and checking the data

Required packages:

```r
library(Biobase)
library(GEOquery)
library(hgu133a.db)
library(genefilter)
library(limma)
library(caret)
library(gtools)
library(GOstats)
library(gplots)
library(MLInterfaces)
```

Download and loading of the dataset:

```r
gds3257 <- getGEO('GDS3257', destdir = ".")
```

Transform the dataset into an expressionset:

```r
eset <- GDS2eSet(gds3257)
eset
```

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 22283 features, 107 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: GSM254629 GSM254648 ... GSM254685 (107 total)
##   varLabels: sample tissue ... description (7 total)
##   varMetadata: labelDescription
## featureData
##   featureNames: 1007_s_at 1053_at ... AFFX-TrpnX-M_at (22283 total)
##   fvarLabels: ID Gene title ... GO:Component ID (21 total)
##   fvarMetadata: Column labelDescription
## experimentData: use ’experimentData(object)’
##   pubMedIds: 18297132
## Annotation:
```

Data

The expressionset contains 107 samples and a total of 22283 features, corresponding to the 22283 probes present in the microarray.

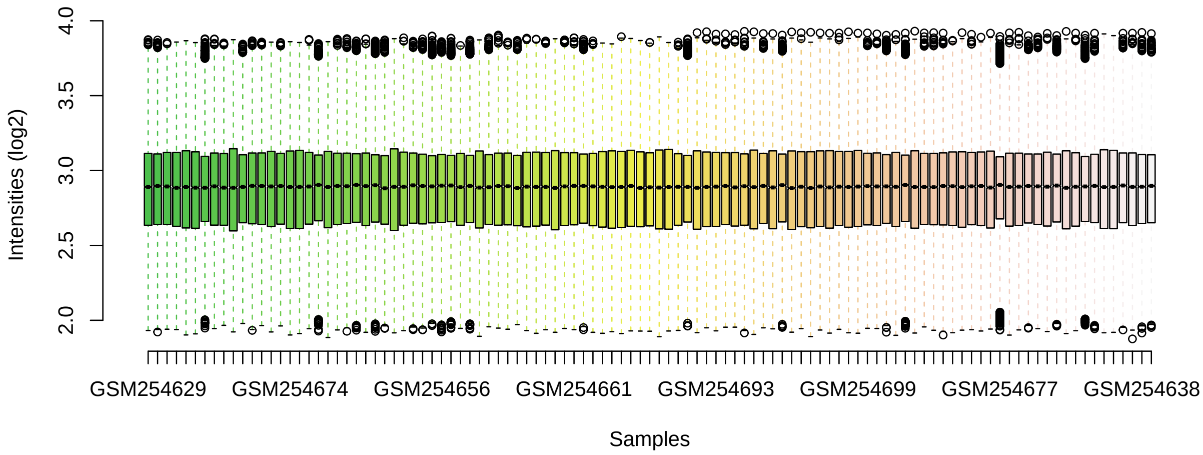
Example of expression data for the first 10 genes and the first 6 samples:

```
dados = exprs(eset)
dados[1:10,1:6]
```

```
##           GSM254629 GSM254648 GSM254694 GSM254701 GSM254728 GSM254726
## 1007_s_at    10.989    10.692    10.898    11.749    10.903    10.769
## 1053_at       6.826     6.910     6.803     6.818     6.838     6.740
## 117_at        7.776     7.684     7.885     7.938     8.010     8.159
## 121_at        9.855    10.132     9.841     9.900     9.872     9.791
## 1255_g_at     4.824     4.985     4.877     4.709     4.789     4.609
## 1294_at       9.104     8.996     9.262     9.686     9.032     8.865
## 1316_at       6.193     6.314     6.257     6.189     6.310     6.217
## 1320_at       6.119     6.021     6.076     6.006     6.017     5.862
## 1405_i_at     7.753     8.141     7.441     7.860     8.234     8.488
## 1431_at       4.968     5.137     4.915     4.875     4.864     4.863
```

Making a boxplot with the base 2 logarithm of the expression data for all samples allows us to easily visualize the scale and distribution of the data. These show similar shape and positioning among samples, as they have a nearly homogeneous distribution along the horizontal line, suggesting that the expression data are already normalized, as expected and mentioned by the authors of the article.

```
col = terrain.colors(length(pData(eset)$individual), alpha = 0.7)
boxplot(log2(dados), col = col, whiskcol = col, pch = 21, cex = 0.8, xlab = 'Samples', ylab = 'Intensit
```



Metadata

The metadata contains 6 fields, including: *tissue*, indicating whether the sample's lung tissue is tumoral or normal, *individual*, indicating whether the individual is currently smoking, has smoked before, or has never smoked, *disease.state*, indicating the state of the tumor in the patient, and *gender*, indicating the patient's gender, among other descriptive fields.

```
vars = pData(eset)
names(vars)
```

```
## [1] "sample"        "tissue"        "individual"    "disease.state" "gender"        "other"
## [7] "description"
```

```
levels(vars$tissue); levels(vars$individual); levels(vars$disease.state); levels(vars$gender)
```

```
## [1] "normal" "tumor"
```

```
## [1] "current smoker" "former smoker"  "never smoker"
```

```
## [1] "stage I"   "stage II"  "stage III" "stage IV"
```

```
## [1] "female" "male"
```

The following table shows the number of samples corresponding to the *tissue* and *individual* variables, making it easier to understand how many of the samples (107), i.e., their frequency, belong to each respective variable.

```
addmargins(table(vars$tissue,vars$individual))
```

```
##
##          current smoker former smoker never smoker Sum
##   normal             16            18           15  49
##   tumor              24            18           16  58
##   Sum                40            36           31 107
```

Preprocessing

Check for NAs in the expressionset:

```
sum(which(is.na(dados)))
```

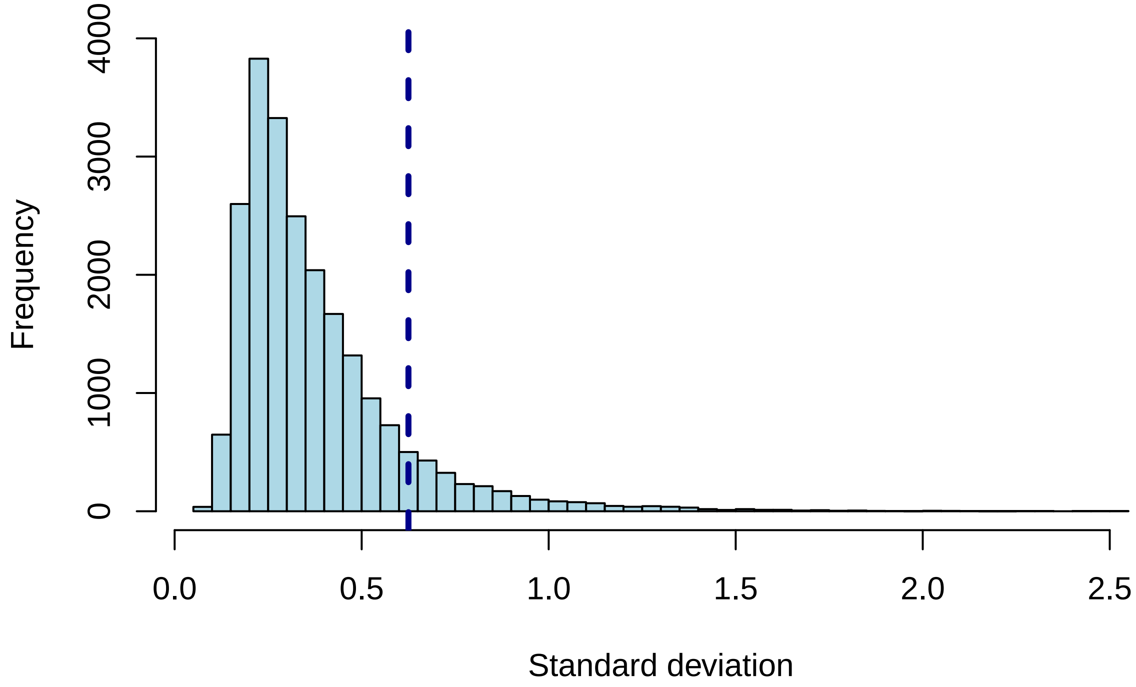
```
## [1] 0
```

As verified, the expressionset does not have any missing data, which aligns with the previous findings, as the dataset is already normalized. The 22283 features will be used in subsequent analyses, an approach also chosen by the authors (Landi et al. 2008). For predictive and clustering analysis, it was decided to filter the data, reducing the number of features to make the process less computationally intensive. Genes were filtered based on whether the standard deviation of the expression values is greater than twice the median of the standard deviations of all genes:

```
# Data filtered by standard deviation
sds = rowSds(dados)
m = median(sds)
eset.f = eset[sds > 2*m]
```

Visually, this corresponds to selecting all genes whose standard deviation of expression values lies beyond the dark blue line in the histogram below:

```
hist(sds,breaks = 50,ylim = c(0,4000),col = "lightblue",xlab = 'Standard deviation',ylab = 'Frequency',m
abline(v = m*2, col = "darkblue",lwd = 3,lty = 2)
```



Differential expression and enrichment analysis

The following statistical analyses were performed using the *limma* package (Ritchie et al. 2015), which is a library used for gene expression analysis in data and microarrays, allowing the creation of linear models using a Bayesian extension, suitable for this type of studies. The genes of interest identified by these statistical analyses were subjected to an enrichment analysis to determine if there is statistically significant "enrichment" in the genes from any biologically coherent set of genes. This latter analysis was conducted using the *GOstats* package (Falcon and Gentleman 2007), which allows testing gene lists for association with Gene Ontology (GO) terms.

In an attempt to find molecular alterations associated with tobacco consumption, a differential expression analysis was conducted between current smokers and never smokers (**C**/**N**) and former smokers and never smokers (**F**/**N**), for all tumor samples, for early tumor stages (stages I and II), and for all normal tissue samples. Priority was given to the analysis of tumor tissue samples in early stages, as the advanced state of the tumor may cause potential gene expression changes. A significance criterion was set at p-value < 0.01 and Fold Change > 1.5.

Never smokers vs. current smokers (Tumor tissue)

Creation of the design for the linear model, defining *never smoker* as the reference:

```r
eset.tumor = eset[, eset$tissue == "tumor"]
indivs1 = relevel(eset.tumor$individual, 'never smoker')
design1 = model.matrix(~ indivs1)
head(design1)
```

```
##   (Intercept) indivs1current smoker indivs1former smoker
## 1           1                     0                    0
## 2           1                     0                    0
## 3           1                     0                    0
## 4           1                     0                    0
## 5           1                     0                    0
## 6           1                     0                    0
```

Creation of linear regression models and performing statistical tests:

```r
fit1 = lmFit(eset.tumor, design1)
fit.bayes1 = eBayes(fit1)
diff1 = topTable(fit.bayes1, coef = 2, 1000, genelist = fit1$genes$NAME)
diff1[1:10,]
```

```
##                logFC AveExpr      t   P.Value adj.P.Val     B
## 203560_at     1.5397   7.377  6.726 8.396e-09 0.0001616 9.702
## 204822_at     1.4822   7.009  6.584 1.451e-08 0.0001616 9.209
## 219787_s_at   1.3950   6.848  5.815 2.732e-07 0.0016185 6.561
## 201088_at     0.8156  10.325  5.799 2.905e-07 0.0016185 6.505
## 204277_s_at  -0.4069   6.588 -5.682 4.506e-07 0.0018951 6.109
## 202558_s_at   0.9419   7.537  5.649 5.103e-07 0.0018951 5.996
## 207828_s_at   1.2021   8.557  5.530 7.956e-07 0.0024235 5.595
## 201761_at     0.9958   9.720  5.483 9.501e-07 0.0024235 5.434
## 218053_at     0.6483   9.747  5.465 1.014e-06 0.0024235 5.375
## 204887_s_at   0.7671   6.953  5.446 1.088e-06 0.0024235 5.312
```

According to the previous tests, we will now check for the most differentially expressed genes using the significance criterion defined earlier (p-value < 0.01 and Fold Change > 1.5):

```
treshold = foldchange2logratio (1.5)
upregulated1 = diff1[which(diff1$logFC > treshold & diff1$adj.P.Val < 0.01),]
downregulated1 = diff1[which(diff1$logFC < -treshold & diff1$adj.P.Val < 0.01),]
unlist(mget(rownames(upregulated1),hgu133aSYMBOL))
```

```
##   203560_at   204822_at 219787_s_at   201088_at 202558_s_at 207828_s_at   201761_at   218053_at
##       "GGH"       "TTK"      "ECT2"     "KPNA2"    "HSPA13"     "CENPF"    "MTHFD2"   "PRPF40A"
## 204887_s_at 203362_s_at 218875_s_at 204092_s_at 212020_s_at 222077_s_at 201897_s_at 209172_s_at
##      "PLK4"    "MAD2L1"     "FBXO5"     "AURKA"     "MKI67"   "RACGAP1"     "CKS1B"     "CENPF"
## 201291_s_at   209642_at 208808_s_at   218039_at   201636_at   203418_at   218542_at   201292_at
##     "TOP2A"      "BUB1"     "HMGB2"    "NUSAP1"      "FXR1"     "CCNA2"     "CEP55"     "TOP2A"
## 213911_s_at   204127_at 219918_s_at 209434_s_at   204641_at   209096_at   204146_at 212295_s_at
##     "H2AZ1"      "RFC3"      "ASPM"      "PPAT"      "NEK2"    "UBE2V2"  "RAD51AP1"    "SLC7A1"
## 208079_s_at
##     "AURKA"
```

```
unlist(mget(rownames(downregulated1),hgu133aSYMBOL))
```

```
##   212256_at   206170_at   201286_at   201525_at   220622_at   204276_at 200810_s_at 200696_s_at
##   "GALNT10"     "ADRB2"      "SDC1"      "APOD"    "LRRC31"       "TK2"     "CIRBP"       "GSN"
```

Found that there are 30 overexpressed genes (in 33 probes) and 8 underexpressed genes (in 8 probes) in the tumor tissue of smoking patients compared to never smokers. Next, an enrichment analysis will be performed to try to determine the most likely biological class to which this set of genes belongs.

For all enrichment analyses, a significance criterion of p-value < 0.025 was considered. The universe of genes corresponds to all genes present in the expressionset.

Enrichment analysis for the overexpressed genes:

```
entrezUniverse = unlist(mget(featureNames(eset), hgu133aENTREZID))
selectedEntrezIds1 = unlist(mget(rownames(upregulated1), hgu133aENTREZID))
params1 = new("GOHyperGParams", geneIds = selectedEntrezIds1, universeGeneIds = entrezUniverse,
             annotation = "hgu133a.db", ontology = "BP", pvalueCutoff = 0.025, testDirection = "over")
hgOver1 = hyperGTest(params1)
summary(hgOver1)[1:15,]
```

```
##        GOBPID    Pvalue OddsRatio ExpCount Count Size                                   Term
## 1  GO:0051301 4.208e-15     27.65   1.2289    16  478                          cell division
## 2  GO:0022402 9.027e-15     21.70   2.2547    19  877                     cell cycle process
## 3  GO:0051783 1.128e-13     56.55   0.2879    10  112         regulation of nuclear division
## 4  GO:0000280 3.596e-13     27.85   0.8330    13  324                       nuclear division
## 5  GO:0051726 7.260e-13     18.11   2.0593    17  801               regulation of cell cycle
## 6  GO:0007049 1.057e-12     15.86   3.4013    20 1323                             cell cycle
## 7  GO:0007088 1.089e-12     58.95   0.2391     9   93 regulation of mitotic nuclear division
## 8  GO:0048285 1.379e-12     24.88   0.9255    13  360                      organelle fission
## 9  GO:0000278 1.713e-12     18.19   1.8099    16  704                     mitotic cell cycle
## 10 GO:0010564 3.365e-12     20.57   1.2572    14  489       regulation of cell cycle process
## 11 GO:0007059 1.280e-11     27.73   0.6402    11  249                 chromosome segregation
## 12 GO:1903047 2.716e-11     17.41   1.4680    14  571             mitotic cell cycle process
## 13 GO:0098813 3.895e-11     30.13   0.5142    10  200         nuclear chromosome segregation
## 14 GO:0000819 1.149e-10     33.74   0.3985     9  155           sister chromatid segregation
## 15 GO:0140014 1.616e-10     25.83   0.5939    10  231               mitotic nuclear division
```

The previous result suggests a statistically significant "enrichment" in genes whose biological processes are related to cell division/cycle, mitotic process, and chromosome segregation, among others. Considering that one of the most evident characteristics in cancer development involves rapid and uncontrolled cell proliferation, these results are in line with expectations. In fact, among the identified genes are "TTK" (which encodes a protein essential for chromosomal alignment during mitosis (Mills et al. 1992)), "ECT2" (thought to play an important role in cytokinesis regulation (Tatsumoto et al. 1999)), and "CENPF" (essential for kinetochore function and chromosomal segregation during mitosis (Liao et al. 1995)), among others, reinforcing the obtained result.

Enrichment analysis for the underexpressed genes:

```
selectedEntrezIds2 = unlist(mget(rownames(downregulated1), hgu133aENTREZID))
params2 = new("GOHyperGParams", geneIds = selectedEntrezIds2, universeGeneIds = entrezUniverse,
              annotation = "hgu133a.db", ontology = "BP", pvalueCutoff = 0.025, testDirection = "over")
hgOver2 = hyperGTest(params2)
summary(hgOver2)[c(1:6,8:14),] #linha 7 demasiado longa
```

```
##        GOBPID    Pvalue OddsRatio  ExpCount Count Size
## 1  GO:0009409 0.0002501     119.2 0.0245951     2   41
## 2  GO:0034443 0.0005999       Inf 0.0005999     1    1
## 3  GO:0044858 0.0005999       Inf 0.0005999     1    1
## 4  GO:0046092 0.0005999       Inf 0.0005999     1    1
## 5  GO:0060588 0.0005999       Inf 0.0005999     1    1
## 6  GO:0061884 0.0005999       Inf 0.0005999     1    1
## 8  GO:0098816 0.0005999       Inf 0.0005999     1    1
## 9  GO:1903906 0.0005999       Inf 0.0005999     1    1
## 10 GO:0002032 0.0011995    1943.5 0.0011998     1    2
## 11 GO:0034439 0.0011995    1943.5 0.0011998     1    2
## 12 GO:0034442 0.0011995    1943.5 0.0011998     1    2
## 13 GO:0042160 0.0011995    1943.5 0.0011998     1    2
## 14 GO:0042161 0.0011995    1943.5 0.0011998     1    2
##                                                                            Term
## 1                                                              response to cold
## 2                                  negative regulation of lipoprotein oxidation
## 3                                             plasma membrane raft polarization
## 4                                               deoxycytidine metabolic process
## 5                            negative regulation of lipoprotein lipid oxidation
## 6                          regulation of mini excitatory postsynaptic potential
## 8                                        mini excitatory postsynaptic potential
## 9                               regulation of plasma membrane raft polarization
## 10 desensitization of G protein-coupled receptor signaling pathway by arrestin
## 11                                                  lipoprotein lipid oxidation
## 12                                           regulation of lipoprotein oxidation
## 13                                                     lipoprotein modification
## 14                                                        lipoprotein oxidation
```

The enrichment analysis for underexpressed genes suggests a statistically significant "enrichment" in genes whose biological processes are related to the cellular defense response and immunology. Among the identified genes are "SDC1" (which encodes a transmembrane protein responsible for mediating cell binding and signaling (Lories et al. 1992)) and "CIRBP" (cold-inducible, encoding a protein with a protective role in response to genotoxic stress by stabilizing transcripts involved in cell survival (Nishiyama et al. 1997)), among others. These functions reinforce the result obtained in the enrichment analysis. The fact that the “SDC1” gene is underexpressed and related to cell binding also somewhat aligns with expectations, as another characteristic in cancers is metastasis, where cells can break away from the primary tumor site and travel through the bloodstream, spreading the tumor to other organs.

Never smokers vs. former smokers (Tumor tissue)

The design used in this case is the same as in the previous case. Creation of linear regression models and performing statistical tests:

```
fit2 = lmFit(eset.tumor, design1)
fit.bayes2 = eBayes(fit2)
diff2 = topTable(fit.bayes2, coef = 3, 1000, genelist = fit2$genes$NAME)
diff2[1:10,]
```

```
##                logFC AveExpr      t   P.Value adj.P.Val      B
## 218222_x_at  0.3235   7.851  4.785 1.214e-05    0.1670 2.6970
## 38671_at    -0.8171   9.479 -4.706 1.608e-05    0.1670 2.4705
## 202134_s_at  0.3912   8.473  4.611 2.248e-05    0.1670 2.2005
## 204862_s_at -0.7199   8.753 -4.415 4.459e-05    0.1968 1.6484
## 209856_x_at  0.3121   8.339  4.275 7.213e-05    0.1968 1.2607
## 213440_at   -0.4512   6.998 -4.232 8.344e-05    0.1968 1.1432
## 221728_x_at -2.4508   7.781 -4.221 8.659e-05    0.1968 1.1134
## 214218_s_at -2.2003   6.034 -4.197 9.376e-05    0.1968 1.0492
## 200810_s_at -0.6214   9.844 -4.182 9.870e-05    0.1968 1.0078
## 211089_s_at  0.3824   7.030  4.103 1.286e-04    0.1968 0.7947
```

According to the previous tests, check for the most differentially expressed genes using the significance criterion defined earlier:

```
upregulated2 = diff2[which(diff2$logFC > treshold & diff2$adj.P.Val < 0.01),]
downregulated2 = diff2[which(diff2$logFC < -treshold & diff2$adj.P.Val < 0.01),]
unlist(mget(rownames(upregulated2),hgu133aSYMBOL))
```

```
## NULL
```

```
unlist(mget(rownames(downregulated2),hgu133aSYMBOL))
```

```
## NULL
```

We can thus verify that there are no overexpressed or underexpressed genes in tissues from former smokers compared to tissues from never smokers.

Never smokers vs. current smokers (Tumor tissue - stages I and II)

Creation of the design for the linear model, defining *never smoker* as the reference:

```
eset.stumor = eset[, eset$tissue == "tumor" & (eset$disease.state == 'stage I' | eset$disease.state ==
indivs2 = relevel(eset.stumor$individual, 'never smoker')
design2 = model.matrix(~ indivs2)
head(design2)
```

```
##   (Intercept) indivs2current smoker indivs2former smoker
## 1           1                     0                    0
## 2           1                     0                    0
## 3           1                     0                    0
## 4           1                     0                    0
## 5           1                     0                    0
## 6           1                     0                    0
```

Creation of linear regression models and performing statistical tests:

```
fit3 = lmFit(eset.stumor, design2)
fit.bayes3 = eBayes(fit3)
diff3 = topTable(fit.bayes3, coef = 2, 1000, genelist = fit3$genes$NAME)
head(diff3)
```

```
##              logFC AveExpr      t   P.Value adj.P.Val     B
## 206170_at -1.0444   8.265 -6.648 4.188e-08 0.0009333 8.228
## 203560_at  1.7140   7.452  6.329 1.218e-07 0.0013566 7.280
## 209667_at -0.7054   8.689 -5.776 7.764e-07 0.0037178 5.627
## 204822_at  1.5852   7.058  5.730 9.038e-07 0.0037178 5.491
## 208760_at -0.9523   8.180 -5.727 9.129e-07 0.0037178 5.482
## 201088_at  0.8608  10.360  5.699 1.003e-06 0.0037178 5.397
```

According to the previous tests, check for the most differentially expressed genes using the significance criterion defined earlier:

```
upregulated3 = diff3[which(diff3$logFC > treshold & diff3$adj.P.Val < 0.01),]
downregulated3 = diff3[which(diff3$logFC < -treshold & diff3$adj.P.Val < 0.01),]
unlist(mget(rownames(upregulated3),hgu133aSYMBOL))
```

```
##   203560_at   204822_at   201088_at 207828_s_at 219787_s_at   218542_at   201761_at   201292_at
##       "GGH"       "TTK"     "KPNA2"     "CENPF"      "ECT2"     "CEP55"    "MTHFD2"     "TOP2A"
## 204887_s_at 201291_s_at
##      "PLK4"     "TOP2A"
```

```
unlist(mget(rownames(downregulated3),hgu133aSYMBOL))
```

```
##   206170_at   209667_at   208760_at   201286_at   220622_at   213244_at 208704_x_at 204519_s_at
##     "ADRB2"      "CES2"     "UBE2I"      "SDC1"    "LRRC31"    "SCAMP4"     "APLP2"      "PLLP"
## 200696_s_at
##       "GSN"
```

We found that there are 9 overexpressed genes (in 10 probes) and 9 underexpressed genes (in 9 probes) in tumor tissue of smoking patients compared to never smokers. Next, an enrichment analysis will be performed to try to determine the most likely biological class to which this set of genes belongs.

Enrichment analysis for the overexpressed genes:

```
selectedEntrezIds3 = unlist(mget(rownames(upregulated3), hgu133aENTREZID))
params3 = new("GOHyperGParams", geneIds = selectedEntrezIds3, universeGeneIds = entrezUniverse,
              annotation = "hgu133a.db", ontology = "BP", pvalueCutoff = 0.025, testDirection = "over")
hgOver3 = hyperGTest(params3)
summary(hgOver3)[1:15,]
```

```
##        GOBPID    Pvalue OddsRatio ExpCount Count Size                                             Te
## 1  GO:0098813 9.865e-06     46.79  0.15425     4  200                   nuclear chromosome segregati
## 2  GO:0022402 1.222e-05     24.77  0.67641     6  877                               cell cycle proce
## 3  GO:0051301 1.242e-05     29.56  0.36867     5  478                                    cell divisi
## 4  GO:0051304 1.842e-05     83.99  0.05553     3   72                            chromosome separati
## 5  GO:2001251 1.842e-05     83.99  0.05553     3   72   negative regulation of chromosome organizati
## 6  GO:0007059 2.344e-05     37.27  0.19205     4  249                           chromosome segregati
## 7  GO:0051307 6.310e-05    237.67  0.01234     2   16                    meiotic chromosome separati
## 8  GO:0000278 8.099e-05     19.60  0.54298     5  704                               mitotic cell cyc
## 9  GO:0007143 8.981e-05    195.68  0.01465     2   19                  female meiotic nuclear divisi
## 10 GO:0006760 9.975e-05    184.79  0.01543     2   20 folic acid-containing compound metabolic proce
## 11 GO:0000910 1.106e-04     45.05  0.10104     3  131                                      cytokines
## 12 GO:0007049 1.302e-04     15.71  1.02040     6 1323                                       cell cyc
## 13 GO:0033044 1.685e-04     38.89  0.11646     3  151            regulation of chromosome organizati
## 14 GO:0032506 1.702e-04    138.52  0.02005     2   26                              cytokinetic proce
## 15 GO:0042558 1.702e-04    138.52  0.02005     2   26  pteridine-containing compound metabolic proce
```

The previous result suggests a statistically significant "enrichment" in genes whose biological processes are related to cell division/cycle, mitotic process, and chromosome regulation/segregation, among others. These results are indeed similar to those obtained when considering all tumor stages. Also here, the previously mentioned genes are among the most overexpressed, whose function relates to the obtained result. However, when considering only the early tumor stages (stages I and II), the number of overexpressed genes decreases, which may be related to the fact that the advanced state of the tumor (not considered here) could cause potential changes in gene expression.

Enrichment analysis for the underexpressed genes:

```
selectedEntrezIds4 = unlist(mget(rownames(downregulated3), hgu133aENTREZID))
params4 = new("GOHyperGParams", geneIds = selectedEntrezIds4, universeGeneIds = entrezUniverse,
              annotation = "hgu133a.db", ontology = "BP", pvalueCutoff = 0.025, testDirection = "over")
hgOver4 = hyperGTest(params4)
summary(hgOver4)[1:15,] #linhas demasiado longas
```

```
##        GOBPID    Pvalue OddsRatio  ExpCount Count Size
## 1  GO:0044858 0.0006856       Inf 0.0006856     1    1
## 2  GO:0061884 0.0006856       Inf 0.0006856     1    1
## 3  GO:0061885 0.0006856       Inf 0.0006856     1    1
## 4  GO:0098816 0.0006856       Inf 0.0006856     1    1
## 5  GO:1903182 0.0006856       Inf 0.0006856     1    1
## 6  GO:1903755 0.0006856       Inf 0.0006856     1    1
## 7  GO:1903906 0.0006856       Inf 0.0006856     1    1
## 8  GO:0002032 0.0013707   1665.71 0.0013712     1    2
## 9  GO:0044855 0.0013707   1665.71 0.0013712     1    2
## 10 GO:0044856 0.0013707   1665.71 0.0013712     1    2
## 11 GO:0048627 0.0013707   1665.71 0.0013712     1    2
## 12 GO:1900756 0.0013707   1665.71 0.0013712     1    2
## 13 GO:1903921 0.0013707   1665.71 0.0013712     1    2
## 14 GO:1903923 0.0013707   1665.71 0.0013712     1    2
## 15 GO:0043244 0.0020471     38.54 0.0699289     2  102
##                                                                        Term
## 1                                          plasma membrane raft polarization
## 2                        regulation of mini excitatory postsynaptic potential
## 3               positive regulation of mini excitatory postsynaptic potential
## 4                                       mini excitatory postsynaptic potential
## 5                                      regulation of SUMO transferase activity
## 6                             positive regulation of SUMO transferase activity
## 7                            regulation of plasma membrane raft polarization
## 8  desensitization of G protein-coupled receptor signaling pathway by arrestin
## 9                                          plasma membrane raft distribution
## 10                                         plasma membrane raft localization
## 11                                                      myoblast development
## 12                                     protein processing in phagocytic vesicle
## 13                       regulation of protein processing in phagocytic vesicle
## 14              positive regulation of protein processing in phagocytic vesicle
## 15                       regulation of protein-containing complex disassembly
```

The enrichment analysis for underexpressed genes suggests a statistically significant "enrichment" in genes whose biological processes are related to distribution/localization/polarization of rafts in the plasma membrane, regulation of macromitophagy (a specialized form of autophagy by which mitochondria are selectively degraded and recycled), and organization of the extracellular matrix. Since the extracellular matrix is composed of a set of molecules that provide structural and biochemical support to surrounding cells, some of the identified genes may be related to tumor metastasis. Another common characteristic of cancer cells is increased resistance to mitochondrial apoptosis, so it would be expected that genes positively regulating this process would be underexpressed. Among the most underexpressed genes are "ADRB2" (involved in cell adhesion process (Boulay et al. 2012)) and "SDC1" (involved with the extracellular matrix and cell adhesion process (Lories et al. 1992)), among others.

Never smokers vs. former smokers (Tumor tissue - stages I and II)

The design used in this case is the same as in the previous case. Creation of linear regression models and performing statistical tests:

```
fit4 = lmFit(eset.stumor, design2)
fit.bayes4 = eBayes(fit4)
diff4 = topTable(fit.bayes4, coef = 3, 1000, genelist = fit4$genes$NAME)
head(diff4)
```

```
##               logFC AveExpr      t   P.Value adj.P.Val     B
## 206170_at   -0.8622   8.265 -5.053 8.515e-06   0.07561 3.068
## 209667_at   -0.6662   8.689 -5.023 9.415e-06   0.07561 2.987
## 208760_at   -0.8991   8.180 -4.979 1.088e-05   0.07561 2.870
## 213509_x_at -0.5529   9.332 -4.911 1.357e-05   0.07561 2.691
## 204519_s_at -0.9484   8.133 -4.793 1.991e-05   0.08875 2.380
## 212326_at   -0.6100   7.631 -4.660 3.061e-05   0.11366 2.030
```

According to the previous tests, check for the most differently expressed genes using the significance criterion defined earlier:

```
upregulated4 = diff4[which(diff4$logFC > treshold & diff4$adj.P.Val < 0.01),]
downregulated4 = diff4[which(diff4$logFC < -treshold & diff4$adj.P.Val < 0.01),]
unlist(mget(rownames(upregulated4),hgu133aSYMBOL))
```

```
## NULL
```

```
unlist(mget(rownames(downregulated4),hgu133aSYMBOL))
```

```
## NULL
```

We can verify that there are no overexpressed or underexpressed genes in tissues from former smokers compared to tissues from never smokers.

Never smokers vs. current smokers (Normal tissue)

Creation of the design for the linear model, defining *never smoker* as the reference:

```
eset.normal = eset[, eset$tissue == "normal"]
indivs3 = relevel(eset.normal$individual, 'never smoker')
design3 = model.matrix(~ indivs3)
head(design3)
```

```
##   (Intercept) indivs3current smoker indivs3former smoker
## 1           1                     0                    0
## 2           1                     0                    0
## 3           1                     0                    0
## 4           1                     0                    0
## 5           1                     0                    0
## 6           1                     0                    0
```

Creation of linear regression models and performing statistical tests:

```
fit5 = lmFit(eset.normal, design3)
fit.bayes5 = eBayes(fit5)
diff5 = topTable(fit.bayes5, coef = 2, 1000, genelist = fit5$genes$NAME)
head(diff5)
```

```
##                logFC AveExpr      t   P.Value adj.P.Val      B
## 202437_s_at  2.1955   7.866  8.956 5.858e-12 1.305e-07 16.130
## 202436_s_at  1.8094   9.424  8.165 9.411e-11 1.049e-06 13.697
## 205576_at    1.3752   6.518  7.070 4.721e-09 3.507e-05 10.221
## 202435_s_at  1.4417   8.559  6.887 9.134e-09 5.089e-05  9.631
## 220911_s_at -0.4782   9.089 -6.536 3.224e-08 1.437e-04  8.500
## 211276_at   -0.9230   6.290 -5.888 3.291e-07 1.167e-03  6.409
```

According to the previous tests, check for the most differently expressed genes using the significance criterion defined earlier:

```r
upregulated5 = diff5[which(diff5$logFC > treshold & diff5$adj.P.Val < 0.01),]
downregulated5 = diff5[which(diff5$logFC < -treshold & diff5$adj.P.Val < 0.01),]
unlist(mget(rownames(upregulated5),hgu133aSYMBOL))
```

```
## 202437_s_at 202436_s_at   205576_at 202435_s_at 221266_s_at 220625_s_at   204580_at 206700_s_at
##    "CYP1B1"    "CYP1B1"  "SERPIND1"    "CYP1B1"   "DCSTAMP"      "ELF5"     "MMP12"     "KDM5D"
##   219890_at
##    "CLEC5A"
```

```r
unlist(mget(rownames(downregulated5),hgu133aSYMBOL))
```

```
##   211276_at   205433_at   213071_at 204428_s_at 205109_s_at   202746_at   204589_at 208096_s_at
##    "TCEAL2"      "BCHE"       "DPT"      "LCAT"   "ARHGEF4"     "ITM2A"     "NUAK1"   "COL21A1"
##   213456_at   202908_at 203349_s_at
##   "SOSTDC1"      "WFS1"      "ETV5"
```

We found that there are 7 overexpressed genes (in 9 probes) and 11 underexpressed genes (in 11 probes) in normal tissue of smoking patients compared to never smokers. Next, an enrichment analysis will be performed to try to determine the most likely biological class to which this set of genes belongs.

Enrichment analysis for the overexpressed genes:

```r
selectedEntrezIds5 = unlist(mget(rownames(upregulated5), hgu133aENTREZID))
params5 = new("GOHyperGParams", geneIds = selectedEntrezIds5, universeGeneIds = entrezUniverse,
              annotation = "hgu133a.db", ontology = "BP", pvalueCutoff = 0.025, testDirection = "over")
hgOver5 = hyperGTest(params5)
summary(hgOver5)[c(1:12,14:15),] #linha 13 demasiado longa
```

```
##        GOBPID    Pvalue OddsRatio  ExpCount Count Size
## 1  GO:1904905 0.0005999       Inf 0.0005999     1    1
## 2  GO:0002930 0.0011995    1943.5 0.0011998     1    2
## 3  GO:0060309 0.0017987     971.7 0.0017996     1    3
## 4  GO:0060435 0.0017987     971.7 0.0017996     1    3
## 5  GO:0034239 0.0023977     647.7 0.0023995     1    4
## 6  GO:0034241 0.0023977     647.7 0.0023995     1    4
## 7  GO:0034721 0.0023977     647.7 0.0023995     1    4
## 8  GO:0071603 0.0023977     647.7 0.0023995     1    4
## 9  GO:0034238 0.0029963     485.8 0.0029994     1    5
## 10 GO:0072675 0.0029963     485.8 0.0029994     1    5
## 11 GO:0002457 0.0035947     388.6 0.0035993     1    6
## 12 GO:0060054 0.0035947     388.6 0.0035993     1    6
## 14 GO:0072674 0.0035947     388.6 0.0035993     1    6
## 15 GO:0090674 0.0035947     388.6 0.0035993     1    6
##                                                                               Term
## 1         negative regulation of endothelial cell-matrix adhesion via fibronectin
## 2                                                  trabecular meshwork development
## 3                                                        elastin catabolic process
## 4                                                           bronchiole development
## 5                                                  regulation of macrophage fusion
## 6                                         positive regulation of macrophage fusion
## 7                            histone H3-K4 demethylation, trimethyl-H3-K4-specific
## 8                                                   endothelial cell-cell adhesion
## 9                                                               macrophage fusion
## 10                                                              osteoclast fusion
## 11                                     T cell antigen processing and presentation
## 12 positive regulation of epithelial cell proliferation involved in wound healing
## 14                                        multinuclear osteoclast differentiation
## 15                               endothelial cell-matrix adhesion via fibronectin
```

The previous result suggests a statistically significant “enrichment” in genes whose biological processes are related to cellular defense and immune response. Among the most overexpressed genes, as expected, is “CYP1B1,” a gene known to be induced by tobacco consumption (Lampe et al. 2004), and whose encoded enzyme metabolizes procarcinogens, chemicals that become carcinogens after metabolism.

Enrichment analysis for the underexpressed genes:

```
selectedEntrezIds6 = unlist(mget(rownames(downregulated5), hgu133aENTREZID))
params6 = new("GOHyperGParams", geneIds = selectedEntrezIds6, universeGeneIds = entrezUniverse,
              annotation = "hgu133a.db", ontology = "BP", pvalueCutoff = 0.025, testDirection = "over")
hgOver6 = hyperGTest(params6)
summary(hgOver6)[c(1:4,6:15),-4]
```

```
##        GOBPID    Pvalue OddsRatio Count Size
## 1  GO:0046448 0.0007713       Inf     1    1
## 2  GO:0050783 0.0007713       Inf     1    1
## 3  GO:0060648 0.0015420    1457.4     1    2
## 4  GO:1903892 0.0015420    1457.4     1    2
## 6  GO:0006581 0.0023122     728.6     1    3
## 7  GO:0014016 0.0023122     728.6     1    3
## 8  GO:0042078 0.0023122     728.6     1    3
## 9  GO:0048133 0.0023122     728.6     1    3
## 10 GO:0090107 0.0023122     728.6     1    3
## 11 GO:0098722 0.0023122     728.6     1    3
## 12 GO:0098728 0.0023122     728.6     1    3
## 13 GO:0019695 0.0030819     485.7     1    4
## 14 GO:1903891 0.0030819     485.7     1    4
## 15 GO:2000015 0.0030819     485.7     1    4
##                                                              Term
## 1                              tropane alkaloid metabolic process
## 2                                       cocaine metabolic process
## 3                                 mammary gland bud morphogenesis
## 4  negative regulation of ATF6-mediated unfolded protein response
## 6                                 acetylcholine catabolic process
## 7                                      neuroblast differentiation
## 8                                    germ-line stem cell division
## 9                    male germ-line stem cell asymmetric division
## 10       regulation of high-density lipoprotein particle assembly
## 11                                  asymmetric stem cell division
## 12                         germline stem cell asymmetric division
## 13                                      choline metabolic process
## 14          regulation of ATF6-mediated unfolded protein response
## 15                 regulation of determination of dorsal identity
```

The enrichment analysis for underexpressed genes suggests a statistically significant “enrichment” in genes whose biological processes are related to metabolic processes and, interestingly, with morphogenesis/development of the mammary gland. Although this study is on lung adenocarcinoma, it is still interesting to note that in normal tissue of smokers, there is differential expression in genes related to the morphogenesis/development of the mammary gland, as tobacco consumption is thought to be related to the onset of breast cancer (Catsburg et al. 2014). However, this may just be a coincidence. Among the most underexpressed genes are "DPT," "ARHGEF4," and "NUAK1," all related to the process of cell adhesion (Superti-Furga et al. 1993; Thiesen et al. 2000; Hou et al. 2011).

Never smokers vs. former smokers (Normal tissue)

The design used in this case is the same as in the previous case. Creation of linear regression models and performing statistical tests:

```
fit6 = lmFit(eset.normal, design3)
fit.bayes6 = eBayes(fit6)
diff6 = topTable(fit.bayes6, coef = 3, 1000, genelist = fit6$genes$NAME)
head(diff6)
```

```
##                logFC AveExpr      t   P.Value adj.P.Val     B
## 206700_s_at  1.6509   8.795  6.499 3.679e-08 0.0008198 4.921
## 214218_s_at -2.7029   5.673 -5.965 2.503e-07 0.0024288 3.798
## 201909_at    3.2166  10.443  5.818 4.226e-07 0.0024288 3.486
## 205000_at    2.9676   7.875  5.749 5.390e-07 0.0024288 3.340
## 221728_x_at -3.1062   7.386 -5.746 5.450e-07 0.0024288 3.334
## 203992_s_at -0.5797   7.818 -4.891 1.086e-05 0.0397467 1.506
```

According to the previous tests, check for the most differently expressed genes using the significance criterion defined earlier:

```
upregulated6 = diff6[which(diff6$logFC > treshold & diff6$adj.P.Val < 0.01),]
downregulated6 = diff6[which(diff6$logFC < -treshold & diff6$adj.P.Val < 0.01),]
unlist(mget(rownames(upregulated6),hgu133aSYMBOL))
```

```
## 206700_s_at   201909_at   205000_at
##     "KDM5D"    "RPS4Y1"     "DDX3Y"
```

```
unlist(mget(rownames(downregulated6),hgu133aSYMBOL))
```

```
## 214218_s_at 221728_x_at
##      "XIST"      "XIST"
```

We found that there are 3 overexpressed genes (in 3 probes) and 1 underexpressed gene (in 2 probes) in normal tissue of former smoking patients compared to never smokers. Next, an enrichment analysis will be performed to try to determine the most likely biological class to which this set of genes belongs.

Enrichment analysis for the overexpressed genes:

```
selectedEntrezIds7 = unlist(mget(rownames(upregulated6), hgu133aENTREZID))
params7 = new("GOHyperGParams", geneIds = selectedEntrezIds7, universeGeneIds = entrezUniverse,
              annotation = "hgu133a.db", ontology = "BP", pvalueCutoff = 0.025, testDirection = "over")
hgOver7 = hyperGTest(params7)
summary(hgOver7)[1:15,]
```

```
##          GOBPID   Pvalue OddsRatio ExpCount Count Size
## 1    GO:0034721 0.001028   1943.83 0.001028     1    4
## 2    GO:0002457 0.001542   1166.10 0.001543     1    6
## 3    GO:0034720 0.001799    971.67 0.001800     1    7
## 4    GO:0070076 0.006414    242.54 0.006427     1   25
## 5    GO:0016577 0.006926    223.85 0.006941     1   27
## 6    GO:0060765 0.006926    223.85 0.006941     1   27
## 7    GO:0006482 0.007438    207.82 0.007456     1   29
## 8    GO:0008214 0.007438    207.82 0.007456     1   29
## 9    GO:0030521 0.010760    141.77 0.010798     1   42
## 10   GO:0070988 0.015093    100.07 0.015168     1   59
## 11   GO:0033143 0.017636     85.28 0.017739     1   69
## 12   GO:0002456 0.024227     61.55 0.024424     1   95
## 13   GO:0019882 0.024480     60.90 0.024681     1   96
## NA         <NA>       NA        NA       NA    NA   NA
## NA.1       <NA>       NA        NA       NA    NA   NA
##                                                                        Term
## 1                     histone H3-K4 demethylation, trimethyl-H3-K4-specific
## 2                                T cell antigen processing and presentation
## 3                                               histone H3-K4 demethylation
## 4                                              histone lysine demethylation
## 5                                                     histone demethylation
## 6                         regulation of androgen receptor signaling pathway
## 7                                                     protein demethylation
## 8                                                      protein dealkylation
## 9                                       androgen receptor signaling pathway
## 10                                                            demethylation
## 11   regulation of intracellular steroid hormone receptor signaling pathway
## 12                                                 T cell mediated immunity
## 13                                      antigen processing and presentation
## NA                                                                     <NA>
## NA.1                                                                   <NA>
```

The previous result suggests a statistically significant "enrichment" in genes whose biological processes are related to cellular defense and immune response ("KDM5D" (Rezvani and Barrett 2008),"DDX3Y" (Rosinski et al. 2008)), peptide and histone biosynthesis and processing ("RPS4Y1", (Andrés et al. 2008)).

Enrichment analysis for the underexpressed genes:

```
selectedEntrezIds8 = unlist(mget(rownames(downregulated6), hgu133aENTREZID))
params8 = new("GOHyperGParams", geneIds = selectedEntrezIds8, universeGeneIds = entrezUniverse,
              annotation = "hgu133a.db", ontology = "BP", pvalueCutoff = 0.025, testDirection = "over")
hgOver8 = hyperGTest(params8)
```

As we can see, this analysis cannot be performed because the gene being tested, "XIST," does not have any corresponding GO term. This gene is involved in the inactivation of one copy of the X chromosome in female mammals.

Former smokers vs. current smokers (Tumor tissue - stages I and II)

To investigate whether the pattern observed between current smokers/never smokers is also observed for former smokers, a gene expression analysis was performed between current smokers and former smokers in early-stage tumor tissues.

Creation of the design for the linear model (*current smoker* as the reference):

```
design4 = model.matrix(~ eset.stumor$individual)
head(design4)
```

```
##   (Intercept) eset.stumor$individualformer smoker eset.stumor$individualnever smoker
## 1           1                                   0                                  1
## 2           1                                   0                                  1
## 3           1                                   0                                  1
## 4           1                                   0                                  1
## 5           1                                   0                                  1
## 6           1                                   0                                  1
```

Creation of linear regression models and performing statistical tests:

```
fit7 = lmFit(eset.stumor, design4)
fit.bayes7 = eBayes(fit7)
diff7 = topTable(fit.bayes7, coef = 2, 1000, genelist = fit7$genes$NAME)
head(diff7)
```

```
##                logFC AveExpr      t   P.Value adj.P.Val       B
## 207788_s_at  0.4580   6.928  4.569 4.104e-05    0.6568 -0.9912
## 212846_at   -0.5195   8.854 -4.148 1.553e-04    0.6568 -1.5285
## 212729_at    0.4508   7.562  4.115 1.717e-04    0.6568 -1.5695
## 203560_at   -1.0166   7.452 -4.081 1.912e-04    0.6568 -1.6136
## 214326_x_at  0.6461   7.294  4.073 1.957e-04    0.6568 -1.6232
## 204395_s_at  0.2559   8.719  4.047 2.120e-04    0.6568 -1.6560
```

According to the previous tests, check for the most differently expressed genes using the significance criterion defined earlier:

```
upregulated7 = diff7[which(diff7$logFC > treshold & diff7$adj.P.Val < 0.01),]
downregulated7 = diff7[which(diff7$logFC < -treshold & diff7$adj.P.Val < 0.01),]
unlist(mget(rownames(upregulated7),hgu133aSYMBOL))
```

```
## NULL
```

```
unlist(mget(rownames(downregulated7),hgu133aSYMBOL))
```

```
## NULL
```

We found that there are no differentially expressed genes between former smokers and current smokers. This may indicate that although the patients had quit smoking some time ago, the gene-level changes remained.

Summary table of differential expression analysis

The following table summarizes the number of genes identified during the previous differential expression analyses:

Tissue	Tumor		Tumor (Stages I e II)		Normal	
Condition	**24C vs 16N**	**18F vs 16N**	**20C vs 10N**	**13F vs 10N**	**16C vs 15N**	**18F vs 15N**
Over Genes	30	0	9	0	7	3
Over Probes	33	0	10	0	9	3
Und Genes	8	0	9	0	11	1
Und Probes	8	0	9	0	11	2

Table 1: Number of probes and genes (overexpressed and underexpressed) differentiating current smokers (**C**) from never smokers (**N**) and former smokers (**F**) from never smokers in all tumor samples, tumor samples in stage I or II, and normal tissue samples. The significance criterion for analysis was set as p-value < 0.01 and Fold Change > 1.5.

Normal tissue vs. tumor tissue

A differential expression analysis was also conducted between samples from normal tissue vs. tumor tissue in order to obtain an overall view of gene expression in the two tissues regardless of the patient's profile.

Creation of the design for the linear model (*normal* as the reference):

```
design5 = model.matrix(~ eset$tissue)
head(design5)
```

```
##   (Intercept) eset$tissuetumor
## 1           1                1
## 2           1                1
## 3           1                1
## 4           1                1
## 5           1                1
## 6           1                1
```

Creation of linear regression models and performing statistical tests:

```
fit8 = lmFit(eset, design5)
fit.bayes8 = eBayes(fit8)
diff8 = topTable(fit.bayes8, coef = 2, 1000, genelist = fit8$genes$NAME)
head(diff8)
```

```
##              logFC AveExpr      t   P.Value adj.P.Val     B
## 209074_s_at -3.372   8.995 -24.01 3.231e-45 7.200e-41 92.44
## 209555_s_at -2.417   8.261 -22.42 1.588e-42 1.769e-38 86.33
## 204396_s_at -2.388   8.677 -22.19 3.992e-42 2.965e-38 85.42
## 206209_s_at -2.554   8.082 -22.00 8.567e-42 4.773e-38 84.66
## 204271_s_at -2.227   8.784 -21.94 1.096e-41 4.886e-38 84.42
## 204677_at   -2.620   8.004 -21.82 1.771e-41 6.577e-38 83.94
```

According to the previous tests, check for the most differently expressed genes using the significance criterion defined earlier:

```
upregulated8 = diff8[which(diff8$logFC > treshold & diff8$adj.P.Val < 0.01),]
downregulated8 = diff8[which(diff8$logFC < -treshold & diff8$adj.P.Val < 0.01),]
length(unlist(mget(rownames(upregulated8),hgu133aSYMBOL))) #número sondas
```

```
## [1] 284
```

```
length(unique(unlist(mget(rownames(upregulated8),hgu133aSYMBOL)))) #número genes
```

```
## [1] 230
```

```
length(unlist(mget(rownames(downregulated8),hgu133aSYMBOL)))
```

```
## [1] 636
```

```
length(unique(unlist(mget(rownames(downregulated8),hgu133aSYMBOL))))
```

```
## [1] 488
```

Due to the high number of genes in the result, we chose to present the total number of genes instead of the gene list itself. We found that, for the defined significance criterion, there are 230 overexpressed genes (in 284 probes) and 488 underexpressed genes (in 636 probes) in tumor tissue compared to normal tissue. Next, an enrichment analysis will be performed to try to determine the most likely biological class in which this set of genes fits.

An enrichment analysis for overexpressed genes:

```
selectedEntrezIds9 = unlist(mget(rownames(upregulated8), hgu133aENTREZID))
params9 = new("GOHyperGParams", geneIds = selectedEntrezIds9, universeGeneIds = entrezUniverse,
              annotation = "hgu133a.db", ontology = "BP", pvalueCutoff = 0.025, testDirection = "over")
hgOver9 = hyperGTest(params9)
summary(hgOver9)[1:15,]
```

```
##        GOBPID    Pvalue OddsRatio ExpCount Count Size
## 1  GO:0009132 5.094e-07     7.192    1.943    12  104
## 2  GO:0006165 5.385e-07     8.060    1.607    11   86
## 3  GO:0046939 6.819e-07     7.850    1.644    11   88
## 4  GO:2001251 8.220e-07     8.831    1.345    10   72
## 5  GO:0033044 8.583e-07     5.668    2.821    14  151
## 6  GO:0055086 8.900e-07     3.394    8.332    25  446
## 7  GO:0007059 9.540e-07     4.371    4.652    18  249
## 8  GO:0000819 1.178e-06     5.505    2.896    14  155
## 9  GO:0009123 1.313e-06     9.819    1.102     9   59
## 10 GO:0022402 2.180e-06     2.582   16.384    37  877
## 11 GO:1903047 2.687e-06     2.960   10.667    28  571
## 12 GO:0046031 2.787e-06     7.598    1.532    10   82
## 13 GO:0098813 5.045e-06     4.500    3.736    15  200
## 14 GO:0009165 5.362e-06     4.475    3.755    15  201
## 15 GO:1901293 5.698e-06     4.451    3.774    15  202
##                                                       Term
## 1                 nucleoside diphosphate metabolic process
## 2                   nucleoside diphosphate phosphorylation
## 3                               nucleotide phosphorylation
## 4           negative regulation of chromosome organization
## 5                    regulation of chromosome organization
## 6  nucleobase-containing small molecule metabolic process
## 7                                   chromosome segregation
## 8                             sister chromatid segregation
## 9               nucleoside monophosphate metabolic process
## 10                                    cell cycle process
## 11                            mitotic cell cycle process
## 12                                 ADP metabolic process
## 13                        nuclear chromosome segregation
## 14                       nucleotide biosynthetic process
## 15             nucleoside phosphate biosynthetic process
```

The previous result indicates a statistically significant "enrichment" in genes whose biological processes are related to cell division, mitotic cycle, and sugar catabolic processes. This aligns with the expected outcomes as mentioned earlier. The overexpression of genes related to sugar catabolism is consistent with the fact that cancer cells alter their metabolism to achieve faster proliferation.

An enrichment analysis for underexpressed genes:

```
selectedEntrezIds10 = unlist(mget(rownames(downregulated8), hgu133aENTREZID))
params10 = new("GOHyperGParams", geneIds = selectedEntrezIds10, universeGeneIds = entrezUniverse,
              annotation = "hgu133a.db", ontology = "BP", pvalueCutoff = 0.025, testDirection = "over")
hgOver10 = hyperGTest(params10)
summary(hgOver10)[1:15,]
```

```
##        GOBPID    Pvalue OddsRatio ExpCount Count Size                               Term
## 1  GO:0001944 2.688e-21     4.070    24.88    80  627            vasculature development
## 2  GO:0072359 4.443e-21     3.453    36.82   100  928     circulatory system development
## 3  GO:0001568 1.317e-20     4.067    23.85    77  601           blood vessel development
## 4  GO:0048514 4.009e-19     4.087    21.35    70  538         blood vessel morphogenesis
## 5  GO:0048856 2.069e-18     2.298   173.23   264 4366   anatomical structure development
## 6  GO:0009653 6.957e-17     2.426    83.84   157 2113 anatomical structure morphogenesis
## 7  GO:0001525 8.704e-17     4.086    18.01    60  454                       angiogenesis
## 8  GO:0035239 1.264e-16     3.350    28.45    78  717                 tube morphogenesis
## 9  GO:0048731 2.053e-16     2.198   146.93   230 3703                 system development
## 10 GO:0032501 2.659e-16     2.198   213.47   299 5380   multicellular organismal process
## 11 GO:0007275 3.681e-16     2.173   157.68   241 3974 multicellular organism development
## 12 GO:0007155 1.024e-15     2.730    47.30   105 1192                      cell adhesion
## 13 GO:0022610 1.534e-15     2.711    47.57   105 1199                biological adhesion
## 14 GO:0032502 1.847e-15     2.127   189.10   272 4766              developmental process
## 15 GO:0035295 2.002e-15     2.987    34.96    86  881                   tube development
```

A hierarchical clustering analysis was performed on the previously filtered expression set data to understand if there is a clear clustering among samples based on expression values. The distance between clusters was calculated using the average linkage method, and the distance matrix was calculated using the one minus the Pearson correlation formula, which was also chosen by the authors of this study.

Hierarchical clustering with distance matrix calculation (using filtered data):

```
corPDist = as.dist(1 - cor(exprs(eset.f), method = "pearson"))
cl.hier = hclust (corPDist, method = "average")
```

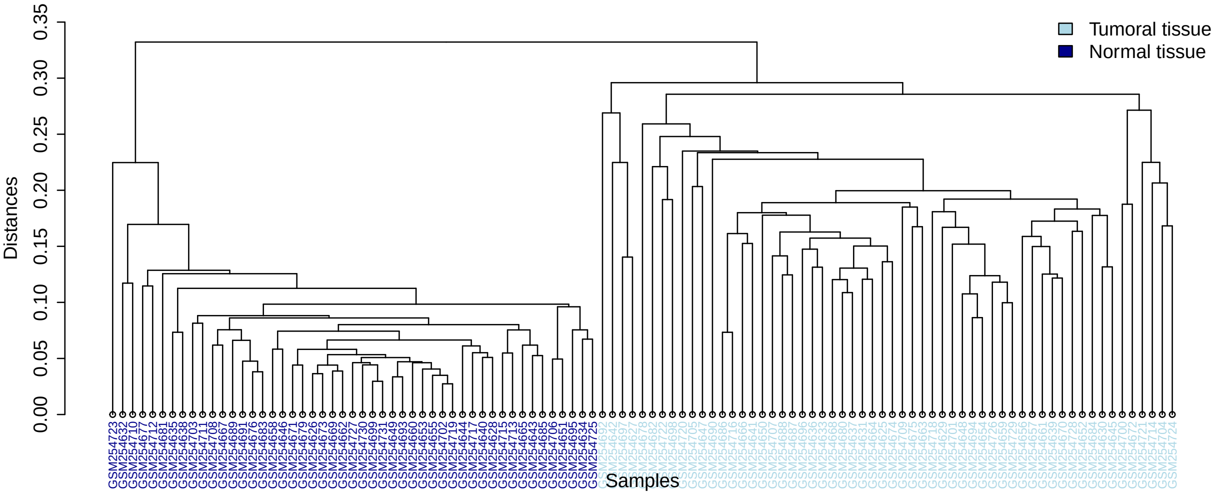
In order to apply colors to the clustering result for better visualization, the following auxiliary function was created:

```
clusMember = cutree(cl.hier, 2)
labelCol = c('lightblue', 'darkblue')

colLab <- function(n) {
  if (is.leaf(n)) {
    a <- attributes(n)
    labCol <- labelCol[clusMember[which(names(clusMember) == a$label)]]
    attr(n, "nodePar") <- c(a$nodePar, lab.col = labCol)
  }
  n
}
```

The result of the hierarchical clustering analysis can then be visualized by creating a plot, with colors indicating the tissue type associated with each sample (tumor or normal tissue):

```
clusDendro = dendrapply(as.dendrogram(cl.hier), colLab)
par(cex = 0.6); plot(clusDendro, axes = F, ylim = c(0, 0.35))
par(cex = 0.9); title(xlab="Samples", ylab="Distances", main = NULL); axis(2)
legend('topright', c('Tumoral tissue', 'Normal tissue'), fill = c('lightblue', 'darkblue'), bty = "n")
```



Analyzing the graph, it is observed the formation of two evident clusters, with samples from tumoral tissue and normal tissue in different branches of the tree. This means that genes from tumoral tissue generally present a closer level of expression among themselves than in relation to genes from normal tissue. This is consistent with expectations, as adenocarcinoma presents its own characteristics that depend on the overexpression of some genes (related to cell multiplication, mitosis, etc.) and underexpression of others (related to cell adhesion, mitophagy, etc.), thus showing differences in expression between the two tissues.

Next, a clustering analysis was performed for both samples, similar to what was done, and for probes, that is, applying clustering to both columns and rows. Thus, two auxiliary functions were defined, one for calculating the distance matrix and another for clustering, necessary for creating the heatmap.

Auxiliary function for calculating the distance matrix:

```r
dist.fun = function(x) {
  return (as.dist (1 - cor(t (x), method = "pearson")))
}
```

Auxiliary function for performing hierarchical clustering:

```r
clust.fun = function (x) {
  return (hclust (x, method = "average"))
}
```

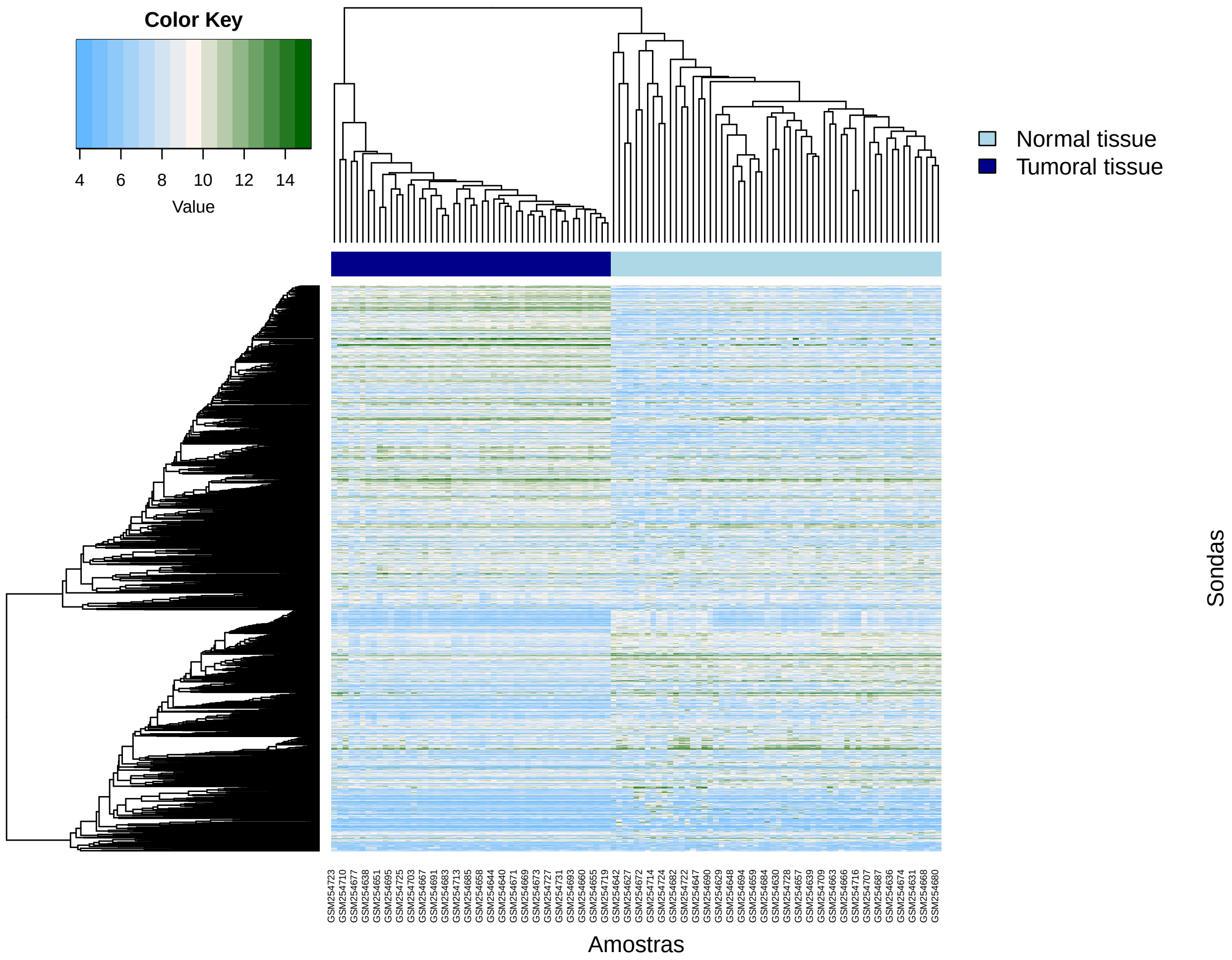
Another function was also defined in order to be applied in the heatmap to associate a color with each type of tissue, namely dark blue for samples of tumor tissue and light blue for samples of normal tissue.

```r
color.map.tissue <- function(tissue) { if (tissue == "tumor") "lightblue" else "darkblue" }
tissuecolors <- unlist(lapply(eset.f$tissue, color.map.tissue))
```

Defining all these functions beforehand, the heatmap can then be created:

```r
heatmap.2(exprs(eset.f), col = colorRampPalette(c("steelblue1", "seashell1", "darkgreen")), scale = "non
          ColSideColors = tissuecolors, key = TRUE, symkey = FALSE, density.info = "none", trace = "none
          cexRow = 0.5, distfun = dist.fun, hclustfun = clust.fun, labRow = F, margins = c(6,16),
          ylab = "Sondas", xlab = "Amostras")

legend('topright', c("Normal tissue","Tumoral tissue"), bty = "n", fill = c("lightblue", "darkblue"))
```



According to the heatmap result, we observe a clear separation of gene expression between samples from tumoral tissue and normal tissue. Furthermore, within the same type of tissue, the formation of two gene groups is observed, belonging to different branches of the clustering tree, with one group being more underexpressed than the other. In tumoral tissue, we observe a greater underexpression of genes compared to genes present in normal tissue, mainly in the group of genes from the upper cluster. On the other hand, there are also more underexpressed genes in tumoral tissue compared to normal tissue, although fewer, a result that is consistent with the gene expression analysis between the two tissues previously performed (227 overexpressed genes versus 481 underexpressed genes in tumoral tissue). From the previous analyses, we know that the overexpressed genes in tumoral tissue are involved in the processes of cell division, mitosis, and sugar catabolism, and the underexpressed genes are involved in the mechanism of angiogenesis.

Dimensionality Reduction

```r
smokers.data <- dados
smokers.data.t = t(smokers.data)
pca_smokers = prcomp(smokers.data.t, scale = T)
summary(pca_smokers)
```

```
## Importance of components:
```

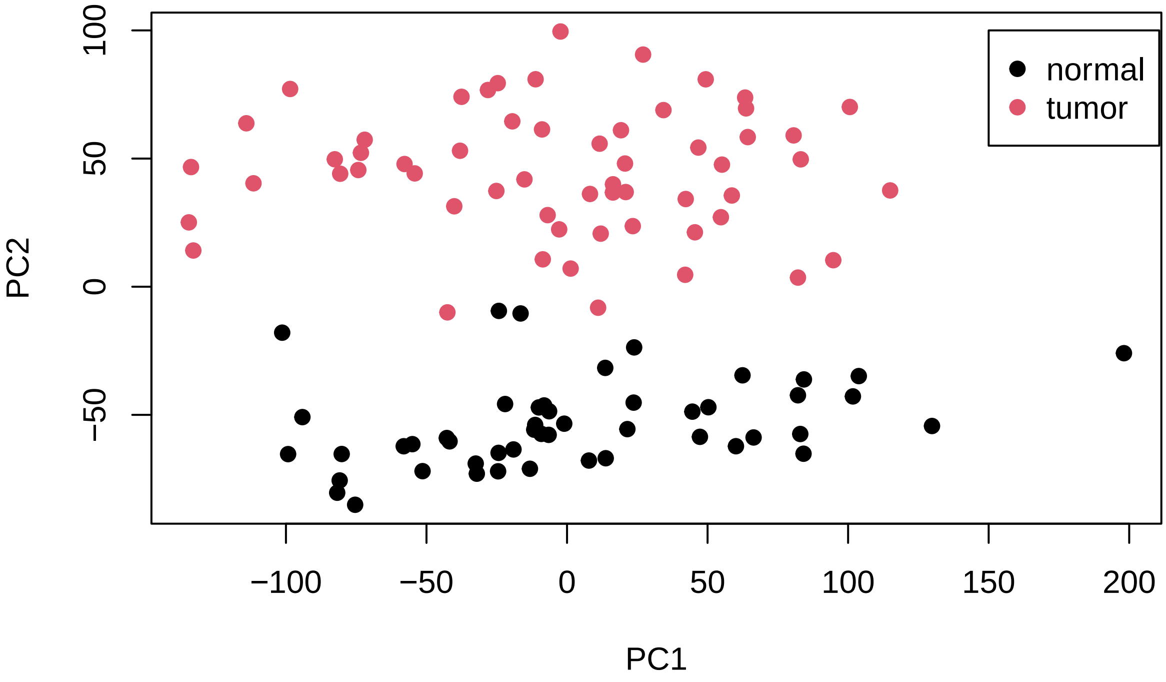
```
##                           PC1    PC2     PC3     PC4     PC5     PC6    PC7     PC8    PC9    PC10
## Standard deviation     63.043 53.708 38.4296 27.2183 25.9261 25.0874 22.625 20.4172 20.052 19.2285 18
## Proportion of Variance  0.178  0.129  0.0663  0.0333  0.0302  0.0282  0.023  0.0187  0.018  0.0166  0
## Cumulative Proportion   0.178  0.308  0.3741  0.4073  0.4375  0.4657  0.489  0.5074  0.525  0.5421  0
##                           PC12    PC13    PC14    PC15    PC16    PC17    PC18   PC19     PC20     PC
## Standard deviation     18.3353 17.4113 16.7838 16.3013 15.7862 15.4407 15.2707 14.928 14.76016 14.404
## Proportion of Variance  0.0151  0.0136  0.0126  0.0119  0.0112  0.0107  0.0105  0.010  0.00978  0.009
## Cumulative Proportion   0.5730  0.5866  0.5992  0.6112  0.6224  0.6331  0.6435  0.654  0.66330  0.672
##                           PC22     PC23     PC24     PC25     PC26     PC27     PC28     PC29     PC3
## Standard deviation     14.3935 13.85536 13.66842 13.41952 13.38401 12.87284 12.73229 12.64261 12.5470
## Proportion of Variance  0.0093  0.00862  0.00838  0.00808  0.00804  0.00744  0.00728  0.00717  0.0070
## Cumulative Proportion   0.6819  0.69053  0.69891  0.70699  0.71503  0.72247  0.72974  0.73692  0.7439
##                            PC31     PC32    PC33     PC34     PC35     PC36     PC37     PC38     PC3
## Standard deviation     12.30044 12.20286 12.1257 11.96917 11.95820 11.76832 11.70047 11.35632 11.2540
## Proportion of Variance  0.00679  0.00668  0.0066  0.00643  0.00642  0.00622  0.00614  0.00579  0.0056
## Cumulative Proportion   0.75077  0.75745  0.7641  0.77048  0.77690  0.78311  0.78926  0.79504  0.8007
##                            PC40     PC41     PC42     PC43     PC44     PC45     PC46     PC47     PC
## Standard deviation     11.19825 10.99659 10.92653 10.68174 10.54417 10.46952 10.39434 10.27464 10.201
## Proportion of Variance  0.00563  0.00543  0.00536  0.00512  0.00499  0.00492  0.00485  0.00474  0.004
## Cumulative Proportion   0.80636  0.81178  0.81714  0.82226  0.82725  0.83217  0.83702  0.84176  0.840
##                            PC49     PC50    PC51    PC52   PC53    PC54    PC55    PC56   PC57    PC5
## Standard deviation     10.06615 10.02075 9.82722 9.75863 9.6694 9.51124 9.46228 9.42546 9.1964 9.1931
## Proportion of Variance  0.00455  0.00451 0.00433 0.00427 0.0042 0.00406 0.00402 0.00399 0.0038 0.0037
## Cumulative Proportion   0.85097  0.85548 0.85981 0.86409 0.8683 0.87234 0.87636 0.88035 0.8841 0.8879
##                           PC59    PC60    PC61    PC62    PC63    PC64    PC65   PC66    PC67    PC6
## Standard deviation     9.10020 8.93242 8.86447 8.75604 8.65114 8.63886 8.52662 8.4402 8.35168 8.29904
## Proportion of Variance 0.00372 0.00358 0.00353 0.00344 0.00336 0.00335 0.00326 0.0032 0.00313 0.00309
## Cumulative Proportion  0.89165 0.89523 0.89876 0.90220 0.90556 0.90891 0.91217 0.9154 0.91850 0.92159
##                           PC69    PC70    PC71    PC72    PC73    PC74    PC75    PC76    PC77    PC7
## Standard deviation     8.22298 8.11884 8.05660 8.02171 7.94224 7.88415 7.80752 7.77609 7.52409 7.4923
## Proportion of Variance 0.00303 0.00296 0.00291 0.00289 0.00283 0.00279 0.00274 0.00271 0.00254 0.0025
## Cumulative Proportion  0.92462 0.92758 0.93050 0.93338 0.93621 0.93900 0.94174 0.94445 0.94699 0.9495
##                           PC79    PC80    PC81    PC82    PC83    PC84    PC85    PC86    PC87    PC8
## Standard deviation     7.38843 7.35039 7.18097 7.09737 7.01644 6.95287 6.94525 6.87523 6.72198 6.6600
## Proportion of Variance 0.00245 0.00242 0.00231 0.00226 0.00221 0.00217 0.00216 0.00212 0.00203 0.0019
## Cumulative Proportion  0.95196 0.95439 0.95670 0.95896 0.96117 0.96334 0.96551 0.96763 0.96965 0.9716
##                           PC89    PC90    PC91    PC92    PC93    PC94    PC95    PC96    PC97    PC9
## Standard deviation     6.61106 6.48725 6.43249 6.32341 6.24887 6.22209 6.14351 6.04679 5.98444 5.9103
## Proportion of Variance 0.00196 0.00189 0.00186 0.00179 0.00175 0.00174 0.00169 0.00164 0.00161 0.0015
## Cumulative Proportion  0.97361 0.97550 0.97735 0.97915 0.98090 0.98264 0.98433 0.98597 0.98758 0.9891
##                           PC99   PC100   PC101   PC102   PC103   PC104   PC105   PC106    PC107
## Standard deviation     5.86415 5.75653 5.67841 5.61318 5.46871 5.41061 5.15408 4.98341 1.96e-13
## Proportion of Variance 0.00154 0.00149 0.00145 0.00141 0.00134 0.00131 0.00119 0.00111 0.00e+00
## Cumulative Proportion  0.99069 0.99218 0.99362 0.99504 0.99638 0.99769 0.99889 1.00000 1.00e+00
```

```
min(which(summary(pca_smokers)$importance[c("Cumulative Proportion"),] > 0.9))
```

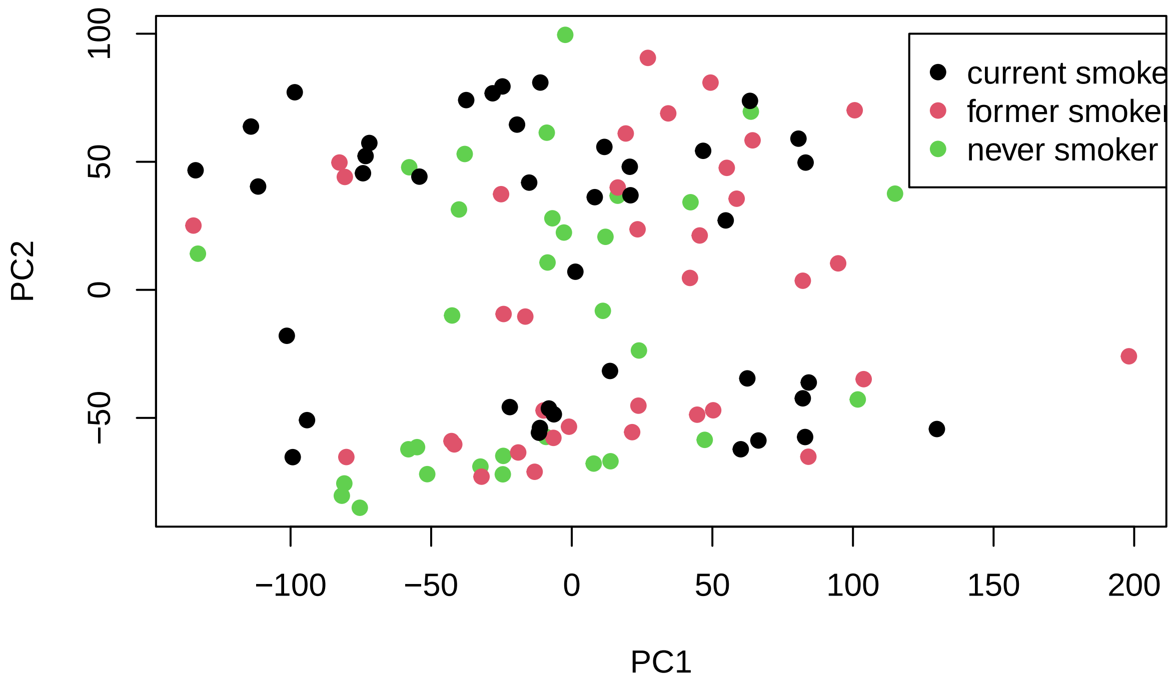
```
## [1] 62
```

It was possible to verify how many essential genes exist to explain the variability of the data. In this study, only 62 genes explain 90% of the data variability. PC1 + PC2 explain only about 30%.

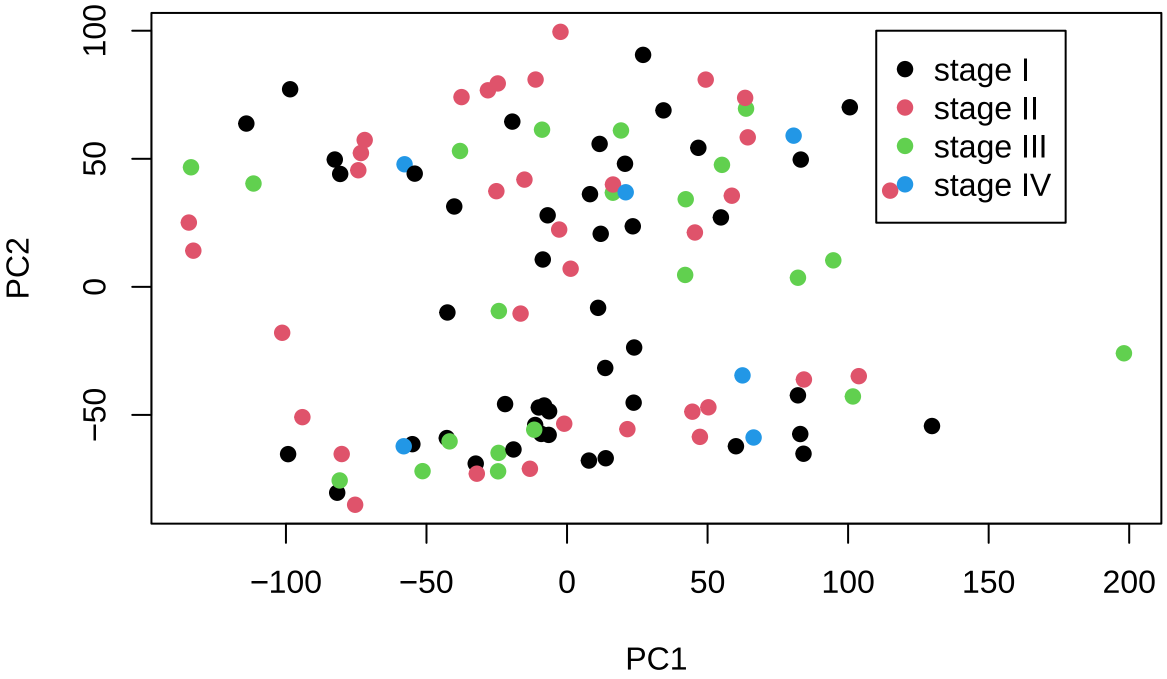
```
plot(pca_smokers$x, col = vars$tissue , pch = 19)
legend(150,100,legend=levels(vars$tissue), col = 1:6, pch=19)
```



```
plot(pca_smokers$x, col = vars$individual , pch = 19)
legend(120,100,legend=levels(vars$individual), col = 1:6, pch=19)
```



```
plot(pca_smokers$x, col = vars$disease.state , pch = 19)
legend(110,100,legend=levels(vars$disease.state), col = 1:6, pch=19)
```



From these plot scores it is possible to verify that for the tissue the samples are not correlated as they are distant from each other, for the rest there is a slight correlation between the samples. The expression data for the tissues are different.

Predictive Analysis

A predictive analysis was conducted to predict both the type of tissue (tumor/normal) and the individual's profile (current/former or never smoker). For this purpose, the *MLInterfaces* package (Carey et al. 2016) was used, which provides a standard way of parameterization and presentation of results for learning algorithms. In the analysis, the machine learning methods of k-nearest neighbors, regression trees, and support vector machines (SVMs) were used, using the cross-validation method that allows the use of all available data. The number of iterations for cross-validation was set to 10.

K-nearest Neighbors

Tissue Type Prediction

Model construction and results:

```
knnResult.tissue.cv <- MLearn(tissue ~ ., eset.f, knnI(k = 1), xvalSpec("LOG", 10, balKfold.xvspec(10))
addmargins(confuMat(knnResult.tissue.cv))
```

```
##         predicted
## given    normal tumor Sum
##   normal     48     1  49
##   tumor       0    58  58
##   Sum        48    59 107
```

Model accuracy for tissue type prediction:

```
precision(confuMat(knnResult.tissue.cv))
```

```
## normal  tumor
## 0.9796 1.0000
```

Prediction of Individual Profile

Construction of the model and results:

```
knnResult.indivs.cv <- MLearn(individual ~ ., eset.f, knnI(k = 1), xvalSpec("LOG", 10, balKfold.xvspec(
addmargins(confuMat(knnResult.indivs.cv))
```

```
##                  predicted
## given             current smoker former smoker never smoker Sum
##   current smoker             19            11           10  40
##   former smoker               9            10           17  36
##   never smoker                4             9           18  31
##   Sum                        32            30           45 107
```

Model Accuracy for Predicting Individual Profile:

Accuracy of the model for predicting individual profile:

```
precision(confuMat(knnResult.indivs.cv))
```

```
## current smoker  former smoker   never smoker
##         0.4750         0.2778         0.5806
```

Regression Trees

Tissue Type Prediction

Model construction and results:

```
treeResult.tissue.cv <- MLearn(tissue ~ ., eset.f, rpartI, xvalSpec("LOG", 10, balKfold.xvspec(10)))
addmargins(confuMat(treeResult.tissue.cv))
```

```
##         predicted
## given    normal tumor Sum
##   normal     45     4  49
##   tumor       3    55  58
##   Sum        48    59 107
```

Model Accuracy for Tissue Type Prediction:

```
precision(confuMat(treeResult.tissue.cv))
```

```
## normal  tumor
## 0.9184 0.9483
```

Individual profile prediction

Model construction and results:

```
treeResult.indivs.cv <- MLearn(individual ~ ., eset.f, rpartI, xvalSpec("LOG", 10, balKfold.xvspec(10))
addmargins(confuMat(treeResult.indivs.cv))
```

```
##                 predicted
## given            current smoker former smoker never smoker Sum
##   current smoker             18            13            9  40
##   former smoker              11            16            9  36
##   never smoker                5             9           17  31
##   Sum                        34            38           35 107
```

Model accuracy for predicting individual profile:

```
precision(confuMat(treeResult.indivs.cv))
```

```
## current smoker  former smoker   never smoker
##         0.4500         0.4444         0.5484
```

Support Vector Machines (SVMs)

Fabric type forecast

Model construction and results:

```
svmResult.tissue.cv <- MLearn(tissue ~ ., eset.f, svmI , xvalSpec("LOG", 10, balKfold.xvspec(10)))
addmargins(confuMat(svmResult.tissue.cv))
```

```
##         predicted
## given    normal tumor Sum
##   normal     48     1  49
##   tumor       0    58  58
##   Sum        48    59 107
```

Model accuracy for tissue type prediction:

```
precision(confuMat(svmResult.tissue.cv))
```

```
## normal  tumor
## 0.9796 1.0000
```

Individual profile prediction

Model construction and results:

```
svmResult.indivs.cv <- MLearn(individual ~ ., eset.f, svmI , xvalSpec("LOG", 10, balKfold.xvspec(10)))
addmargins(confuMat(svmResult.indivs.cv))
```

```
##                  predicted
## given             current smoker former smoker never smoker Sum
##   current smoker             35             3            2  40
##   former smoker              17            11            8  36
##   never smoker                6             9           16  31
##   Sum                        58            23           26 107
```

Model accuracy for predicting individual profile:

```
precision(confuMat(svmResult.indivs.cv))
```

```
## current smoker  former smoker   never smoker
##         0.8750         0.3056         0.5161
```

In general, the 3 types of models used, k-nearest neighbors, regression trees and SVMs, present a good level of accuracy in predicting the type of tissue associated with each sample, with the k-nearest neighbors method being better. close with an accuracy level of 97.96% for normal tissue samples and 100% for tumor tissue samples. As we saw in the clustering analysis, there is a clear separation between the two tissue types, which could make the prediction process easier, as we see here.

On the other hand, the level of precision drops considerably when it comes to predicting the profile of the individual associated with each sample, with the precision levels of the three models for this situation being around or below 50% (with the exception of the precision of the model of SVM for current smokers, with 75%). In this case we have one more variable to predict than in the previous case, which makes the prediction process even more complicated. Furthermore, this difference in the prediction between tissues and the individual's profile is due to the fact that the difference in expression is evident between cancerous and normal cells (genes related to cell cycle, mitosis, immunology, etc.) and this difference is not so evident when what is at stake is whether or not the individual consumes tobacco.

Conclusions

Differential expression analysis in 107 samples of tumor tissue and normal tissue from current, former and never-smoking patients demonstrated that there are changes at the genetic level caused by tobacco consumption. In fact, in samples from smokers there was a higher expression of genes related to the cell division/cycle process, mitosis process and chromosome segregation in the analysis with all tumor stages or just the first stages (e.g. "TTK" , "ECT2", "CENPF"). This result is consistent with the fact that cancer cells present a high level of cellular proliferation and is in line with the result obtained by the authors of this study:

> "We found that smoking induces deregulation of this very mitotic process (. . .) comprises genes that regulate the mitotic spindle formation (. . .) such as CENPF. (. . .) TTK (linked to cell mitosis through EGFR,a critical drug target for lung adenocarcinoma."

Samples from smokers showed lower expression of genes related to the cellular defense response process and immunology in tumor tissue (e.g. "SDC1" and "CIRBP"). In normal tissue this trend is reversed, with this type of genes being overexpressed, including the "CYP1B1" gene, which encodes an enzyme capable of metabolizing procarcinogens. This result is also in line with the results obtained by the authors:

> "In the non-tumor tissue, current smoking strongly altered immune response genes, consistent with the defense mechanisms of the lung tissue against the acute toxic effects of smoking. (...) Our results are consistent with some previous findings, such as smoking-related alteration of CYP1B1"

An analysis of genetic expression between former smokers and current smokers demonstrated that, for the defined significance criterion, there are no differentially expressed genes, which may indicate that although the patients had already stopped smoking some time ago, changes in the expression level of genes remained.

Another analysis of differential expression between samples from normal tissue and tumor tissue demonstrated that there are more underexpressed genes and fewer overexpressed genes in tumor tissue, a result also evident by the hierarchical clustering analysis, with a clear separation between samples from tumor and normal tissue. . From the enrichment analysis it appears that there is a statistically significant "enrichment" in genes whose biological processes are related to cell division, mitotic cycle and sugar catabolic processes, which is in line with what is expected since cancer cells evolve in order to alter their metabolism in order to achieve faster proliferation.

Regarding underexpressed genes, there is a statistically significant "enrichment" in genes whose biological processes are related to the development of the cardiovascular/circulatory system, morphogenesis and development of blood vessels. As mentioned, although it may seem contradictory in relation to what is known about cancer, particularly with regard to the increase in the angiogenesis mechanism, it is known that in some types of cancer the angiogenesis mechanism is not that relevant and is, in in many cases, reduced, so lung adenocarcinoma could be one of these cases.

After carrying out a predictive analysis using the k-nearest neighbors, regression trees and SVMs methods, it is concluded that the methods present very good accuracy when it comes to predicting the tissue associated with each sample, with the k-neighbors method being better at predicting closest with an accuracy level of 97.96% for normal tissue samples and 100% for tumor tissue samples. However, for predicting the profile of the individual associated with each sample, this is not the case, with the accuracy levels of the three models for this situation being around or below 50% (with the exception of the accuracy of the SVM model for current smokers, with 75%), which could be due to the existence of a clear separation between the genes expressed in the two types of tissue, as verified in the clustering analysis, which will not be so evident when it comes to distinguishing between the profiles of individuals .

References

Andrés, Olga, Thomas Kellermann, Francesc López-Giráldez, Julio Rozas, Xavier Domingo-Roura, and Montserrat Bosch. 2008. "RPS4Y Gene Family Evolution in Primates." *BMC Evolutionary Biology* 8 (1): 1.

Boulay, Gaylor, Nicolas Malaquin, Ingrid Loison, Bénédicte Foveau, Capucine Van Rechem, Brian R Rood, Albin Pourtier, and Dominique Leprince. 2012. "Loss of Hypermethylated in Cancer 1 (HIC1) in Breast Cancer Cells Contributes to Stress-Induced Migration and Invasion Through β-2 Adrenergic Receptor (ADRB2) Misregulation." *Journal of Biological Chemistry* 287 (8): 5379–89.

Carey, Vince, Robert Gentleman, Jess Mar, contributions from Jason Vertrees, and Laurent Gatto. 2016. *MLInterfaces: Uniform Interfaces to r Machine Learning Procedures for Data in Bioconductor Containers.*

Catsburg, Chelsea, Victoria A Kirsh, Colin L Soskolne, Nancy Kreiger, and Thomas E Rohan. 2014. "Active Cigarette Smoking and the Risk of Breast Cancer: A Cohort Study." *Cancer Epidemiology* 38 (4): 376–81.

Falcon, Seth, and Robert Gentleman. 2007. "Using GOstats to Test Gene Lists for GO Term Association." *Bioinformatics* 23 (2): 257–58.

Hou, X, JE Liu, W Liu, CY Liu, ZY Liu, and ZY Sun. 2011. "A New Role of NUAK1: Directly Phosphorylating P53 and Regulating Cell Proliferation." *Oncogene* 30 (26): 2933–42.

Lampe, Johanna W, Sergey B Stepaniants, Mao Mao, Jerald P Radich, Hongyue Dai, Peter S Linsley, Stephen H Friend, and John D Potter. 2004. "Signatures of Environmental Exposures Using Peripheral

Leukocyte Gene Expression: Tobacco Smoke." *Cancer Epidemiology Biomarkers & Prevention* 13 (3): 445–53.

Landi, Maria Teresa, Tatiana Dracheva, Melissa Rotunno, Jonine D Figueroa, Huaitian Liu, Abhijit Dasgupta, Felecia E Mann, et al. 2008. "Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival." *PloS One* 3 (2): e1651.

Liao, H, RJ Winkfein, G Mack, JB Rattner, and TJ Yen. 1995. "CENP-f Is a Protein of the Nuclear Matrix That Assembles onto Kinetochores at Late G2 and Is Rapidly Degraded After Mitosis." *The Journal of Cell Biology* 130 (3): 507–18.

Lories, Veerle, Jean-Jacques Cassiman, Herman Van den Berghe, and Guido David. 1992. "Differential Expression of Cell Surface Heparan Sulfate Proteoglycans in Human Mammary Epithelial Cells and Lung Fibroblasts." *Journal of Biological Chemistry* 267 (2): 1116–22.

Mills, Gordon Brent, Rosemarie Schmandt, Martha McGill, Antonella Amendola, Mary Hill, Kathleen Jacobs, Christopher May, Anna-Marie Rodricks, Susan Campbell, and David Hogg. 1992. "Expression of TTK, a Novel Human Protein Kinase, Is Associated with Cell Proliferation." *Journal of Biological Chemistry* 267 (22): 16000–16006.

Nishiyama, Hiroyuki, Hiroaki Higashitsuji, Hiromichi Yokoi, Katsuhiko Itoh, Shozo Danno, Tadashi Matsuda, and Jun Fujita. 1997. "Cloning and Characterization of Human CIRP (Cold-Inducible RNA-Binding Protein) cDNA and Chromosomal Assignment of the Gene." *Gene* 204 (1): 115–20.

Rezvani, Katayoun, and A John Barrett. 2008. "Characterizing and Optimizing Immune Responses to Leukaemia Antigens After Allogeneic Stem Cell Transplantation." *Best Practice & Research Clinical Haematology* 21 (3): 437–53.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015. "Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research*, gkv007.

Rosinski, Kellie V, Nobuharu Fujii, Jeffrey K Mito, Kevin KW Koo, Suzanne M Xuereb, Olga Sala-Torra, James S Gibbs, et al. 2008. "DDX3Y Encodes a Class i MHC–Restricted HY Antigen That Is Expressed in Leukemic Stem Cells." *Blood* 111 (9): 4817–26.

Subramanian, Janakiraman, and Ramaswamy Govindan. 2007. "Lung Cancer in Never Smokers: A Review." *Journal of Clinical Oncology* 25 (5): 561–70.

Superti-Furga, A, M Rocchi, BW Schäfer, and R Gitzelmann. 1993. "Complementary DNA Sequence and Chromosomal Mapping of a Human Proteoglycan-Binding Cell-Adhesion Protein (Dermatopontin)." *Genomics* 17 (2): 463–67.

Tatsumoto, Takashi, Xiaozhen Xie, Rayah Blumenthal, Isamu Okamoto, and Toru Miki. 1999. "Human ECT2 Is an Exchange Factor for Rho GTPases, Phosphorylated in G2/m Phases, and Involved in Cytokinesis." *The Journal of Cell Biology* 147 (5): 921–28.

Thiesen, Signe, S Kübart, H-H Ropers, and HG Nothwang. 2000. "Isolation of Two Novel Human RhoGEFs, ARHGEF3 and ARHGEF4, in 3p13-21 and 2q22." *Biochemical and Biophysical Research Communications* 273 (1): 364–69.