# Reproducible Research in High-Throughput Biology: A Case Study

Paolo Sonego[1]

December 2, 2012

[1]paolo.sonego@gmail.com

**Aim of this Document:** Using the open source R, CRAN and Bioconductor, we want to verify the reproducibility of the paper "Reversal of gene expression changes in the colorectal normal-adenoma pathway by NS398 selective COX2 inhibitor, Galamb, O. and at. 2010 on the basis of the information and procedure depicted in the original manuscript.

**Findings:** We decided to replicate the Class Prediction Analysis depicted in the manuscript. The authors taking advantage of the nearest shrunken centroid method (PAM) [2] wanted to identify the gene expressions patterns (*signatures*) of the contrasts *adenoma vs normal biopsy samples* and *colorectal cancer (CRC) vs normal biopsy samples*. On the basis of the information contained in the manuscript's Material and Methods, we were not capable to reproduce the *exact* original results.

**Conclusions:** This document combines the Literate Programming paradigm and the power of R and Bioconductor in providing an example of Reproducible Research in the analysis of high-throughput biological data.

## Introduction

The original aim of the paper [1] were to analyse the gene expression modulating effect of NS398 selective COX2 inhibitor on the HT29 colon adenocarcinoma cell line and to correlate this effect to the modulation in gene expression observed during normal-adenoma and normal-CRC transition when biopsy samples were analysed.

## 1 Description of the data analysis

Fifty-three samples of total RNA extracted from cells obtained using biopsy in patients having colon adenoma, colorectal cancer, bowel disease and healthy control were hybridized to Affymetrix HG U133 Plus 2.0 oligonucleotide arrays and are available from the Gene Expression Omnibus database (series accession numbers: *GSE183*) or from the ArrayExpress database (series accession numbers: *E-GEOD-4183*).

We focused our attention on the classification tasks and, following the original Material and Methods from [1], we used the nearest shrunken centroid method (PAM) [2] in order to identify the Gene Expressions patterns (signatures) of the samples in the two contrast *adenoma vs normal biopsy samples* and *colorectal cancer (CRC) vs normal biopsy samples*.

# 2 Loading microarray data into Bioconductor

The data used in this example can be downloaded from the ArrayExpress database and imported into R using the `ArrayExpress` package by typing:

```
library("ArrayExpress")
EGEOD4183.affybatch <- ArrayExpress(accession = "E-GEOD-4183", save = TRUE)
fns <- list.celfiles(path = "../../DATA/", full.names = TRUE)
EGEOD4183.affybatch <- read.affyBatch(fns)
```

A brief description of the `EGEOD4183.affybatch` object can be obtained by using the `print(EGEOD4183.affy` command:

```
AffyBatch object
size of arrays=1164x1164 features (67 kb)
cdf=HG-U133_Plus_2 (54675 affyids)
number of samples=53
number of genes=54675
annotation=hgu133plus2
notes=E-GEOD-4183
E-GEOD-4183
NA
c("unknown_experiment_design_type", "transcription profiling by array")
```

If the Affymetrix microarray data sets have been downloaded into a single directory, then the `.cel` files can be loaded into R using the `ReadAffy` command.

# 3 Pre-processing

## 3.1 Data Quality Assessment

The powerful and chip technology agnostic `ArrayQualityMetrics` package can be used to check the quality of the data, a mandatory step in every data analysis. The following code create a directory `aQM` depicting all kind of Quality Control Metrics and a table highlighting bad quality experiments.

```
library("arrayQualityMetrics")
arrayQualityMetrics(expressionset = as(EGEOD4183.affybatch, "ExpressionSet"),
    outdir = "aQM", force = TRUE)
```

## 3.2 Data Normalisation

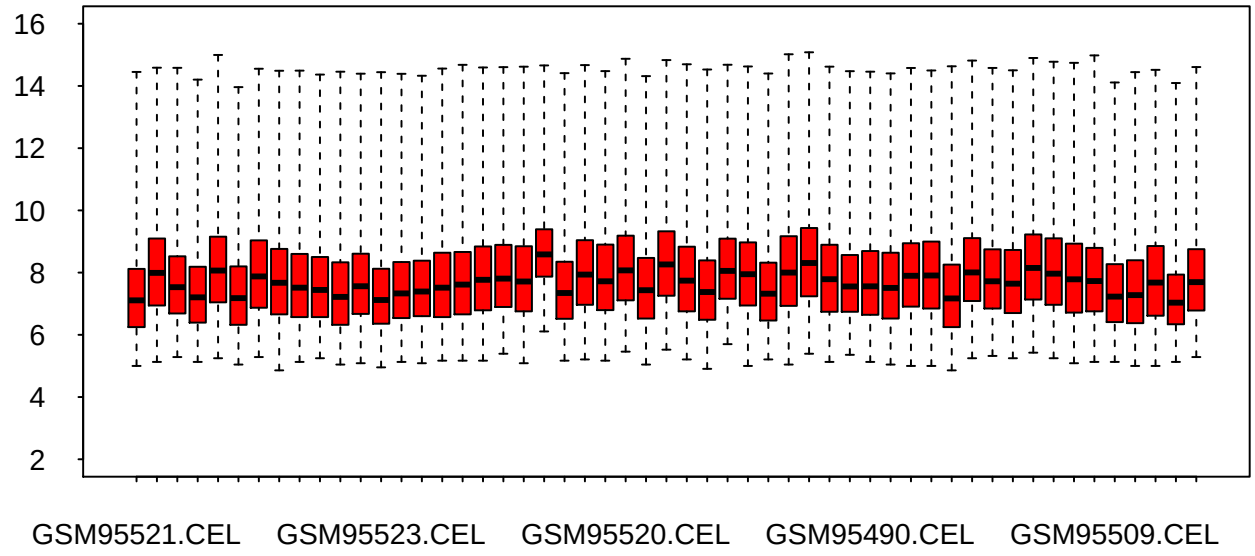
Following the Material and Methods of the original paper we used the GeneChip RMA (GCRMA) with quantile normalisation.

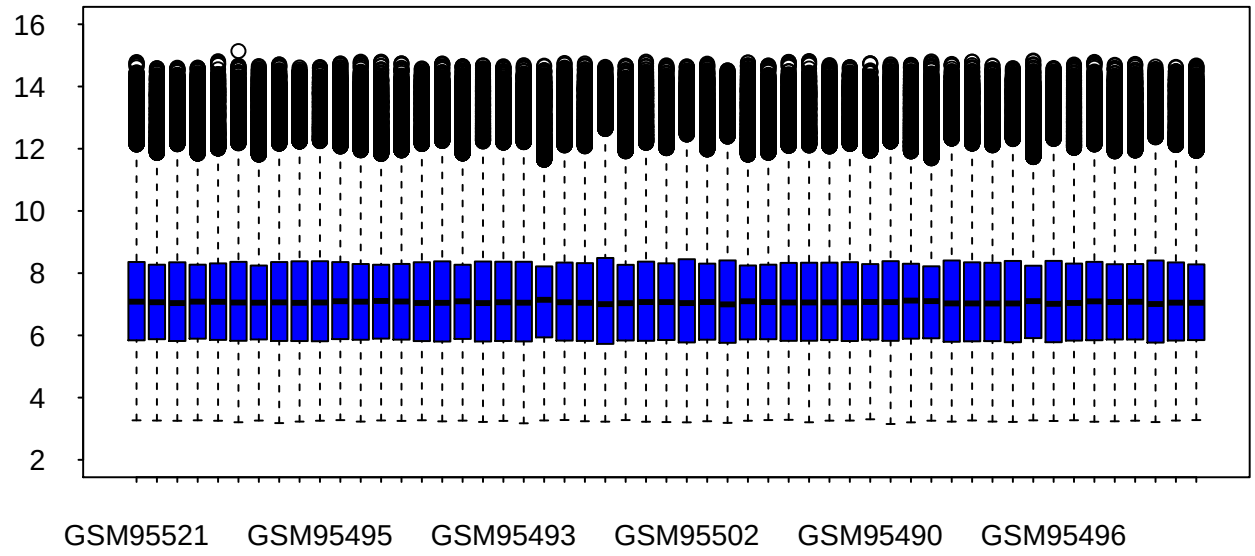
```
library("gcrma")
eset.rma <- rma(EGEOD4183.affybatch)

## Background correcting
## Normalizing
## Calculating Expression
```

The aim of the normalisation is to make the distribution of probe intensities for each array in a set of arrays the same. We illustrate its effect by studying boxplots of the raw data against their normalised counterparts.

```
library("affyPLM")
par(mar = c(3, 3, 2, 1), mfrow = c(2, 1), las = 1, tck = -0.01, cex.axis = 0.9)
# Raw data intensities
boxplot(EGEOD4183.affybatch, col = "red", main = "", ylim = c(2, 16))
# Normalised intensities
boxplot(exprs(eset.rma), col = "blue", ylim = c(2, 16))
```



# 4 Classification using the nearest shrunken centroid method

Gene Expressions patterns (*signatures*) of the contrast between adenoma and healthy biopsy samples were determined using the nearest shrunken centroid classification algorithm (PAM).

Using the **pamr** package between adenoma and healthy biopsy samples, 20 classifiers were identified (sensitivity:100%, specificity: 100%). See Table 1 and Fig 1 for the identified classifiers.

```
x.avsn <- exprs(eset.rma)[, c(grep("adenoma", pData(eset.rma)$Description),
    grep("healthy", pData(eset.rma)$Description))]
mydata.avsn <- list(x = x.avsn, y = factor(c(rep("adenoma", length(grep("adenoma",
    pData(eset.rma)$Description))), rep("normal", length(grep("healthy",
    pData(eset.rma)$Description))))), genenames = featureNames(eset.rma),
    geneid = featureNames(eset.rma))
library("pamr")
set.seed(123)  # for reproducibility
mytrain.avsn <- pamr.train(mydata.avsn)

## 1234567891011121314151617181920212223242526272829 30

mycv.avsn <- pamr.cv(mytrain.avsn, mydata.avsn)

## 12Fold 1 :1234567891011121314151617181920212223242526272829 30
## Fold 2 :1234567891011121314151617181920212223242526272829 30
## Fold 3 :1234567891011121314151617181920212223242526272829 30
## Fold 4 :1234567891011121314151617181920212223242526272829 30
## Fold 5 :1234567891011121314151617181920212223242526272829 30
## Fold 6 :1234567891011121314151617181920212223242526272829 30
## Fold 7 :1234567891011121314151617181920212223242526272829 30
## Fold 8 :1234567891011121314151617181920212223242526272829 30

pamr.confusion(mycv.avsn, threshold = 6.8)

##         adenoma normal Class Error rate
## adenoma      15      0                0
## normal        0      8                0
## Overall error rate= 0

avsn.signature <- data.frame(pamr.listgenes(mytrain.avsn, mydata.avsn,
    threshold = 6.8, genenames = T))

##       id          name        adenoma-score normal-score
##  [1,] 237530_at   237530_at   -0.75         1.4062
##  [2,] 212942_s_at 212942_s_at 0.7479        -1.4022
##  [3,] 227475_at   227475_at   0.3957        -0.742
##  [4,] 212531_at   212531_at   0.3871        -0.7259
##  [5,] 204719_at   204719_at   -0.3207       0.6013
##  [6,] 230204_at   230204_at   -0.1871       0.3509
##  [7,] 219727_at   219727_at   0.1587        -0.2975
##  [8,] 1552296_at  1552296_at  -0.1268       0.2378
##  [9,] 201563_at   201563_at   0.1161        -0.2177
```

```
## [10,] 203256_at   203256_at   0.114         -0.2137
## [11,] 240157_at   240157_at   -0.1043       0.1955
## [12,] 203510_at   203510_at   0.0765        -0.1435
## [13,] 207504_at   207504_at   -0.0689       0.1291
## [14,] 240389_at   240389_at   -0.0374       0.0701
## [15,] 203962_s_at 203962_s_at 0.0208        -0.039
## [16,] 224412_s_at 224412_s_at -0.0132       0.0247
## [17,] 204470_at   204470_at   0.0128        -0.0239
## [18,] 217996_at   217996_at   0.012         -0.0226
## [19,] 207850_at   207850_at   0.0101        -0.019
## [20,] 221019_s_at 221019_s_at -0.0013       0.0024

library("hgu133plus2.db")  # chip annotations
mp.entrez <- mappedkeys(hgu133plus2ENTREZID)
mp.entrez.lst <- as.list(hgu133plus2ENTREZID[mp.entrez])
mp.symbol <- mappedkeys(hgu133plus2SYMBOL)
mp.symbol.lst <- as.list(hgu133plus2SYMBOL[mp.symbol])
avsn.df <- data.frame(Affymetrix_id = as.character(avsn.signature$id),
    ENTREZID = unlist(mp.entrez.lst)[as.character(avsn.signature$id)],
    SYMBOL = unlist(mp.symbol.lst)[as.character(avsn.signature$id)],
    avsn.signature[, 3:4])
write.table(avsn.df, file = "adenomavsnormal_signature_pam.txt", sep = "\t",
    quote = F, row.names = F)
```

Table 1: Classificatory genes identified by pam - adenoma vs normal

| Affymetrix_id | ENTREZID | SYMBOL | adenoma.score | normal.score |
|---|---|---|---|---|
| 237530_at | | | -0.75 | 1.4062 |
| 212942_s_at | 57214 | KIAA1199 | 0.7479 | -1.4022 |
| 227475_at | 94234 | FOXQ1 | 0.3957 | -0.742 |
| 212531_at | 3934 | LCN2 | 0.3871 | -0.7259 |
| 204719_at | 10351 | ABCA8 | -0.3207 | 0.6013 |
| 230204_at | 1404 | HAPLN1 | -0.1871 | 0.3509 |
| 219727_at | 50506 | DUOX2 | 0.1587 | -0.2975 |
| 1552296_at | 266675 | BEST4 | -0.1268 | 0.2378 |
| 201563_at | 6652 | SORD | 0.1161 | -0.2177 |
| 203256_at | 1001 | CDH3 | 0.114 | -0.2137 |
| 240157_at | | | -0.1043 | 0.1955 |
| 203510_at | 4233 | MET | 0.0765 | -0.1435 |
| 207504_at | 766 | CA7 | -0.0689 | 0.1291 |
| 240389_at | 140803 | TRPM6 | -0.0374 | 0.0701 |
| 203962_s_at | 10529 | NEBL | 0.0208 | -0.039 |
| 224412_s_at | 140803 | TRPM6 | -0.0132 | 0.0247 |
| 204470_at | 2919 | CXCL1 | 0.0128 | -0.0239 |
| 217996_at | 22822 | PHLDA1 | 0.012 | -0.0226 |
| 207850_at | 2921 | CXCL3 | 0.0101 | -0.019 |
| 221019_s_at | 81035 | COLEC12 | -0.0013 | 0.0024 |

```
pamr.plotcen(mytrain.avsn, mydata.avsn, threshold = 6.8)
```

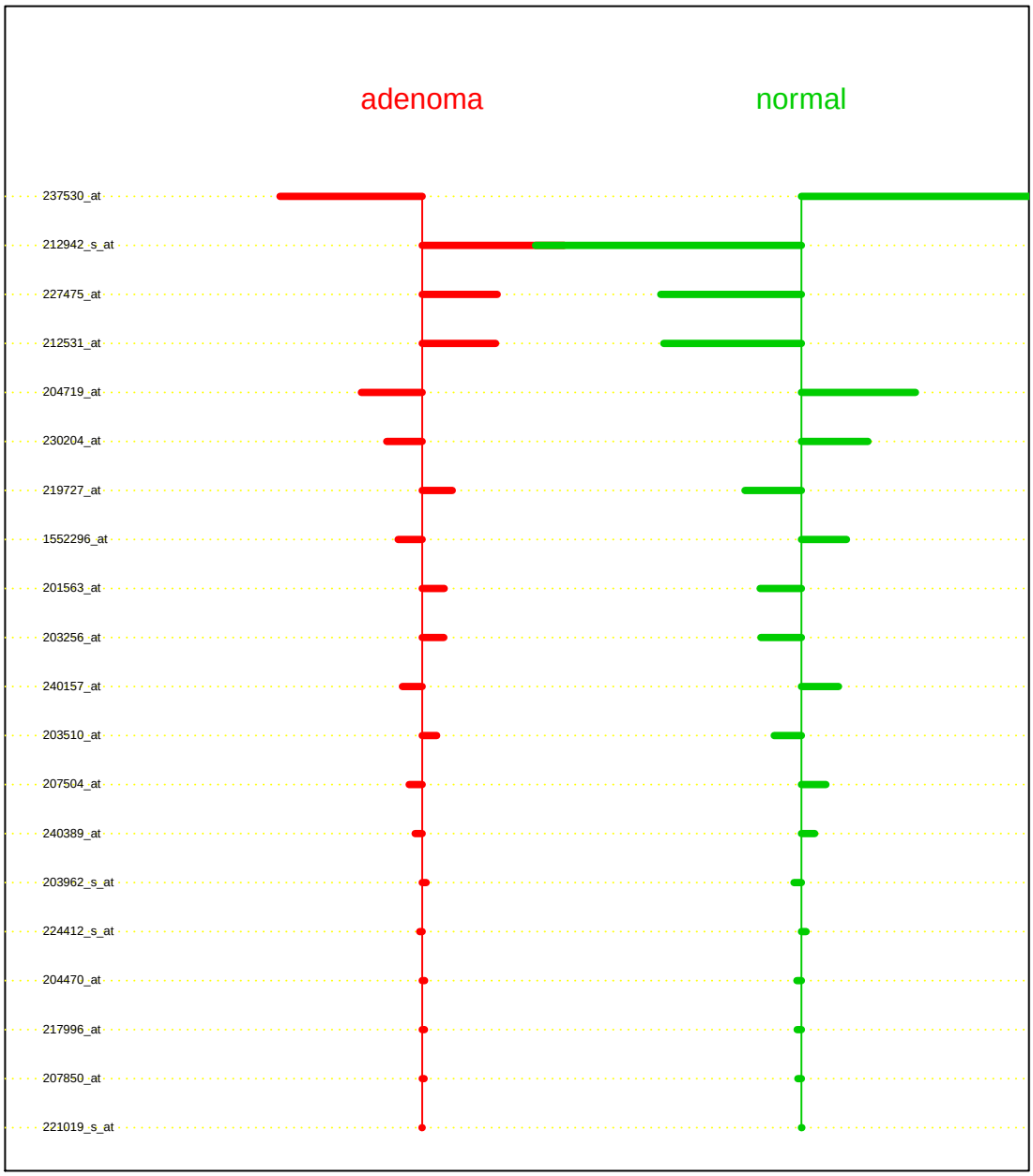



Figure 1: Centroids of the 20 features in the contras: adenoma vs normal

Normal and CRC biopsy samples could be distinguished using 15 discriminatory genes (sensitivity:86%, specificity: 100%). See Table 2 and Fig 2 for the identified classifiers.

```
x.crcvsn <- exprs(eset.rma)[, c(grep("colorectal cancer", pData(eset.rma)$Description), g
mydata.crcvsn <- list(x = x.crcvsn, y = factor(c(rep("CRC", length(grep("colorectal cance
set.seed(123)
mytrain.crcvsn <- pamr.train(mydata.crcvsn)

## 12345678910111213141516171819202122324252627282930

mycv.crcvsn <- pamr.cv(mytrain.crcvsn, mydata.crcvsn)

## 12Fold 1 :12345678910111213141516171819202122324252627282930
## Fold 2 :12345678910111213141516171819202122324252627282930
## Fold 3 :12345678910111213141516171819202122324252627282930
## Fold 4 :12345678910111213141516171819202122324252627282930
## Fold 5 :12345678910111213141516171819202122324252627282930
## Fold 6 :12345678910111213141516171819202122324252627282930
## Fold 7 :12345678910111213141516171819202122324252627282930
## Fold 8 :12345678910111213141516171819202122324252627282930

pamr.confusion(mycv.crcvsn, threshold = 4.45)

##        CRC normal Class Error rate
## CRC     13      2           0.1333
## normal   0      8           0.0000
## Overall error rate= 0.086

crcvsn.signature <- data.frame(pamr.listgenes(mytrain.crcvsn, mydata.crcvsn, threshold =

##       id          name        CRC-score normal-score
##  [1,] 202112_at   202112_at   0.3549    -0.6655
##  [2,] 211959_at   211959_at   0.1558    -0.2921
##  [3,] 211980_at   211980_at   0.144     -0.2699
##  [4,] 211981_at   211981_at   0.1405    -0.2635
##  [5,] 212531_at   212531_at   0.1195    -0.224
##  [6,] 206336_at   206336_at   0.1003    -0.188
##  [7,] 209395_at   209395_at   0.053     -0.0994
##  [8,] 209081_s_at 209081_s_at 0.0507    -0.095
##  [9,] 204470_at   204470_at   0.0443    -0.0831
## [10,] 202917_s_at 202917_s_at 0.0391    -0.0733
## [11,] 218468_s_at 218468_s_at 0.02      -0.0376
## [12,] 228863_at   228863_at   0.0181    -0.0339
## [13,] 212950_at   212950_at   0.0166    -0.0311
## [14,] 202291_s_at 202291_s_at 0.0089    -0.0168
## [15,] 204464_s_at 204464_s_at 0.005     -0.0094

crcvsn.df <- data.frame(Affymetrix_id = as.character(crcvsn.signature$id), ENTREZID = unl
write.table(crcvsn.df, file = "crcvsnormal_signature_pam.txt", sep = "\t", quote = F, row
```

```
pamr.plotcen(mytrain.crcvsn, mydata.crcvsn, threshold = 4.45)
```

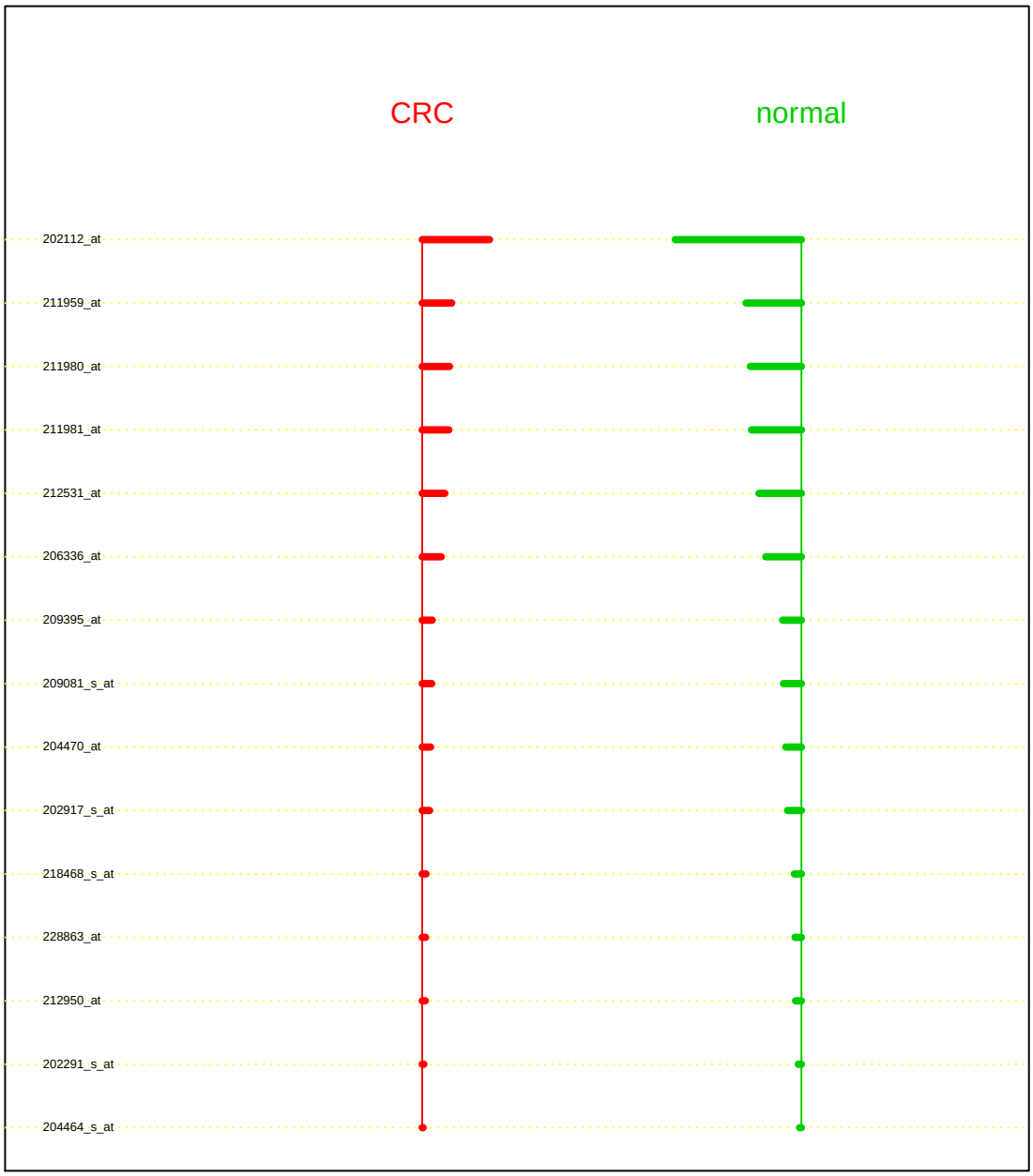



Figure 2: Centroids of the 15 features in the contras: normal vs colorectal cancer

Table 2: Classificatory genes identified by pam - normal vs colorectal cancer

| Affymetrix_id | ENTREZID | SYMBOL | CRC.score | normal.score |
|---|---|---|---|---|
| 202112_at | 7450 | VWF | 0.3549 | -0.6655 |
| 211959_at | 3488 | IGFBP5 | 0.1558 | -0.2921 |
| 211980_at | 1282 | COL4A1 | 0.144 | -0.2699 |
| 211981_at | 1282 | COL4A1 | 0.1405 | -0.2635 |
| 212531_at | 3934 | LCN2 | 0.1195 | -0.224 |
| 206336_at | 6372 | CXCL6 | 0.1003 | -0.188 |
| 209395_at | 1116 | CHI3L1 | 0.053 | -0.0994 |
| 209081_s_at | 80781 | COL18A1 | 0.0507 | -0.095 |
| 204470_at | 2919 | CXCL1 | 0.0443 | -0.0831 |
| 202917_s_at | 6279 | S100A8 | 0.0391 | -0.0733 |
| 218468_s_at | 26585 | GREM1 | 0.02 | -0.0376 |
| 228863_at | 27253 | PCDH17 | 0.0181 | -0.0339 |
| 212950_at | 221395 | GPR116 | 0.0166 | -0.0311 |
| 202291_s_at | 4256 | MGP | 0.0089 | -0.0168 |
| 204464_s_at | 1909 | EDNRA | 0.005 | -0.0094 |

# 5 Conclusion and Discussion

Following the incomplete information depicted in the paper we were capable of reproducing only part of the original results. We think that the main reasons of this outcome are:

- No source code for the analyses was included
- The versions of the different packages were not specified
- The exact parameters selected for the analyses were not reported

This case study has achieved two aims:

- Depicting a basic pipeline for the analysis of high-throughput biological data following the principle of the Reproducible Research
- Showing a basic selection of common mistakes that can make the replication of a published work difficult

# R Version information

- R version 2.15.2 (2012-10-26), `x86_64-apple-darwin9.8.0`
- Locale: `en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, graphics, grDevices, methods, splines, stats, utils
- Other packages: affy 1.36.0, affyPLM 1.34.0, AnnotationDbi 1.20.2, ArrayExpress 1.18.0, arrayQualityMetrics 3.14.0, Biobase 2.18.0, BiocGenerics 0.4.0, cluster 1.14.3, codetools 0.2-8, colorout 0.9-9, dataframe 2.5, DBI 0.2-5, gcrma 2.30.0, hgu133plus2.db 2.8.0, hgu133plus2cdf 2.11.0, impute 1.32.0, knitr 0.8, matrixStats 0.6.2, org.Hs.eg.db 2.8.0, pamr 1.54, preprocessCore 1.20.0, RSQLite 0.11.2, samr 2.0, survival 2.36-14, xtable 1.7-0
- Loaded via a namespace (and not attached): affyio 1.26.0, annotate 1.36.0, beadarray 2.8.1, BeadDataPackR 1.10.0, BiocInstaller 1.8.3, Biostrings 2.26.2, Cairo 1.5-2, colorspace 1.2-0, digest 0.5.2, evaluate 0.4.2, formatR 0.6, genefilter 1.40.0, grid 2.15.2, Hmisc 3.10-1, hwriter 1.3, IRanges 1.16.4, lattice 0.20-10, latticeExtra 0.6-24, limma 3.14.1, parallel 2.15.2, plyr 1.7.1, R.methodsS3 1.4.2, RColorBrewer 1.0-5, reshape2 1.2.1, setRNG 2011.11-2, stats4 2.15.2, stringr 0.6.1, SVGAnnotation 0.93-1, tools 2.15.2, vsn 3.26.0, XML 3.95-0.1, zlibbioc 1.4.0

# References


[1] O. Galamb, S. Spisak, F. Sipos, K. Toth, N. Solymosi, B. Wichmann, T. Krenacs, G. Valcz, Z. Tulassay, and B. Molnar. Reversal of gene expression changes in the colorectal normal-adenoma pathway by ns398 selective cox2 inhibitor. *British journal of cancer*, 102(4):765–773, 2010.

[2] T. Hastie, R. Tibshirani, B. Narasimhan, and G. Chu. *pamr: Pam: prediction analysis for microarrays*, 2011. R package version 1.54.