Selected GEO Data for "A novel approach to gene ordering using beta distributions on Montecarlo p-values reveals an expression pattern in colorectal cancer"

Riffo-Campos, A. Ayala, G. Domingo, J.

March 13, 2021

1 GEO Colorectal Cancer datasets

The datasets were obtained at GEO website (https://www.ncbi.nlm.nih.gov/gds/), filtering the search by:
'Expression profiling by array' AND 'Colorectal cancer' AND 'Homo sapiens'

finding 218 results (until May, 2019). Of these, the datasets repeated, those that did not include controls, xenografts, animal models, cell culture and others were discarted.

The datasets not included, preceded by the reason for exclusion, were:

- Aging-associated, 3 patients: GSE80981
- All are normal samples: GSE97689, GSE42690, GSE4107, GSE56789, GSE20931
- Cells culture: GSE109592, GSE41469, GSE64595, GSE76583, GSE33112, GSE93254, GSE43576, GSE63916, GSE86598, GSE8565, GSE70468, GSE75867, GSE70018, GSE86597, GSE79038
- CRC subtypes: GSE54483, GSE42559
- Circulating Tumor Cells: GSE31023
- Colon normal children: GSE37267
- Colon polyp: GSE81804
- Colonospheres: GSE38706
- Crohn's disease: GSE75459, GSE17594
- Curative large bowel resection for CRC: GSE23194
- Duodenal cancer: GSE111156
- Ex vivo platform: GSE56386
- F. nucleatum infection: GSE122183
- Fetal and adult liver samples: GSE61276

- Fibroblast: GSE51257, GSE46824, GSE93253
- Hepatocellular carcinoma: GSE112790
- Hereditary: GSE23011
- Human cancers processed in different ways: GSE27175
- LncRNA: GSE52413, GSE126092, GSE109454, GSE110715
- Lung cancer: GSE43580, GSE32863
- Lymph node: GSE63596
- Mice model: GSE63544, GSE106535
- NanoString nCounter: GSE68306, GSE130557, GSE101479, GSE101481, GSE86561, GSE78248
- Not include normal tissue: GSE27157, GSE30540, GSE10961, GSE3493, GSE131027, GSE101472, GSE94104, GSE97781, GSE116305, GSE96528, GSE90814, GSE101896, GSE85043, GSE81653, GSE58394, GSE79959, GSE45404, GSE86559, GSE86557, GSE71222, GSE66726, GSE79460, GSE69182, GSE63624, GSE69657, GSE64256, GSE54986, GSE40367, GSE62080, GSE33193, GSE42284, GSE35452, GSE19862, GSE19860, GSE34489, GSE27854, GSE37892, GSE36335, GSE30378, GSE14095, GSE26906, GSE4459, GSE4526, GSE16534, GSE5851, GSE4554, GSE27544, GSE14010, GSE103479, GSE75117
- Organoid: GSE75916, GSE79461, GSE57965, GSE28907, GSE74843, GSE13471, GSE45270, GSE83513, GSE56448
- Patients with Ulcerative Colitis: GSE37283
- Prostate cancer: GSE101607
- Skin biopsy of the infant: GSE54162
- Stool samples: GSE99573
- SuperSeries: GSE101651, GSE126095, GSE93255, GSE110225, GSE114061, GSE79794, GSE86599, GSE86566, GSE79462, GSE64258, GSE41015, GSE62322, GSE47076, GSE52847, GSE44073, GSE37182, GSE27913, GSE38940, GSE33114, GSE24551, GSE23768
- Supplementary raw data files not provided: GSE3964, GSE5364
- Test of the technique: GSE73883
- Ulcerative colitis: GSE3629
- Xenografts: GSE47087, GSE76402, GSE70915, GSE36006, GSE113965, GSE73906, GSE103340, GSE35144

In addition, the datasets of other platforms were also excluded, namely:

- Datasets from ABI: GSE25071 and GSE15781
- Datasets from Illumina: GSE106582, GSE79793, GSE83889, GSE74602, GSE47063, GSE31279, GSE38939, GSE37182 (GSE37175 (normal) and GSE37178 (tumor))

- Datasets from Agilent: GSE104645, GSE89076, GSE71187, GSE70880, GSE75970, GSE89287, GSE90524, GSE77199, GSE50746, GSE28000, GSE68204, GSE35982 and GSE35602
- Datasets from Phalanx Biotech Group: GSE41011
- Datasets from spotted oligonucleotide: GSE12032

2 Download and preprocessing of the data

The datasets included in the analysis were separated by the biological designed in paired i. e. case/normal tissue from the same patient, and not-paired. In any case, only Affymetrix microarray datasets with samples obtained directly from patients were included. The platforms included were: Affymetrix Human Genome U133A Array, Affymetrix Human Genome U133 Plus 2.0 Array, Affymetrix HT HG-U133+ PM Array and Affymetrix Human Genome U219 Array.

2.1 Paired colorectal cancer datasets

2.1.1 Affymetrix array: GSE32323 dataset

The GSE32323 dataset was published by Khamas et al in 2012 (PMID: 22399497). The dataset include 17 colorectal cancer patients with samples from cancer and non-cancerous tissues and 10 samples from CRC cell lines. The cell lines samples were excluded in this analysis.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133plus2.db)
library(hgu133plus2cdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE32323 = ReadAffy(celfile.path = "GSE32323_RAW")
raw_GSE32323
dim(exprs(raw_GSE32323))

# Quality controls

table(probes(raw_GSE32323,"pm") >= probes(raw_GSE32323,"mm"))

affy::hist(raw_GSE32323)
affy::boxplot(raw_GSE32323)

# arrayQualityMetrics(expressionset = raw_GSE32323,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE32323 = affy::rma(raw_GSE32323)
dim(exprs(RMA_raw_GSE32323))
class(RMA_raw_GSE32323)
```

```
# write.csv(colnames(RMA_raw_GSE32323), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE32323_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE32323)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Patient_ID = metadata_fen$patient_ID
Type = metadata_fen$type
Pair = metadata_fen$pair
Stage = metadata_fen$stage
RMA_raw_GSE32323_CRC = RMA_raw_GSE32323
pData(RMA_raw_GSE32323_CRC) = data.frame(Series_sample_id,
                        Patient_ID, Type, Pair, Stage)
colnames(RMA_raw_GSE32323_CRC) = colnames(RMA_raw_GSE32323)
head(colnames(RMA_raw_GSE32323_CRC))
head(row.names(RMA_raw_GSE32323_CRC))
head(pData(RMA_raw_GSE32323_CRC))
GSE32323 = RMA_raw_GSE32323_CRC

# Plots
graphics::boxplot(exprs(GSE32323))
affy::hist(exprs(GSE32323))

# Save
# save(GSE32323, file ="GSE32323.rda")
```

2.1.2 Affymetrix array: GSE8671 dataset

The GSE8671 dataset was published by Sabates-Bellver et al in 2007 (PMID: 18171984) and conrespond to paired samples of colorectal adenomas and normal mucosas from 32 patients (64 samples).

```
library(affy)
library(org.Hs.eg.db)
library(hgu133plus2.db)
library(hgu133plus2cdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE8671 = ReadAffy(celfile.path = "GSE8671_RAW")
raw_GSE8671
dim(exprs(raw_GSE8671))

# Quality controls

table(probes(raw_GSE8671,"pm") >= probes(raw_GSE8671,"mm"))
```

```
affy::hist(raw_GSE8671)
affy::boxplot(raw_GSE8671)

# arrayQualityMetrics(expressionset = raw_GSE8671,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE8671 = affy::rma(raw_GSE8671)
dim(exprs(RMA_raw_GSE8671))
class(RMA_raw_GSE8671)
# write.csv(colnames(RMA_raw_GSE8671), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE8671_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE8671)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Patient_ID = metadata_fen$patient_ID
Type = metadata_fen$type
Pair = metadata_fen$pair
Sample_ID = metadata_fen$sample_ID
Location = metadata_fen$Location
RMA_raw_GSE8671_CRC = RMA_raw_GSE8671
pData(RMA_raw_GSE8671_CRC) = data.frame(Series_sample_id,
                        Patient_ID, Type, Pair, Sample_ID, Location)
colnames(RMA_raw_GSE8671_CRC) = colnames(RMA_raw_GSE8671)
head(colnames(RMA_raw_GSE8671_CRC))
head(row.names(RMA_raw_GSE8671_CRC))
head(pData(RMA_raw_GSE8671_CRC))
GSE8671 = RMA_raw_GSE8671_CRC

# Plots
graphics::boxplot(exprs(GSE8671))
affy::hist(exprs(GSE8671))

# Save
# save(GSE8671, file ="GSE8671.rda")
```

2.1.3 Affymetrix array: GSE15960 dataset

The GSE15960 was published by Galamb et al in 2010 (PMID:20087348) with 6 laser microdissected human colon epithelial cells, including adenoma, colorectal cancer (tumor) and normal samples. Only colorectal cancer (tumor) and normal samples were included.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133plus2.db)
library(hgu133plus2cdf)
```

```
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE15960 = ReadAffy(celfile.path = "GSE15960_RAW")
raw_GSE15960
dim(exprs(raw_GSE15960))

# Quality controls

table(probes(raw_GSE15960,"pm") >= probes(raw_GSE15960,"mm"))

affy::hist(raw_GSE15960)
affy::boxplot(raw_GSE15960)

# arrayQualityMetrics(expressionset = raw_GSE15960,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE15960 = affy::rma(raw_GSE15960)
dim(exprs(RMA_raw_GSE15960))
class(RMA_raw_GSE15960)

# write.csv(colnames(RMA_raw_GSE15960), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE15960_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE15960)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE15960_CRC = RMA_raw_GSE15960
pData(RMA_raw_GSE15960_CRC) = data.frame(Series_sample_id,
                           Type, Pair)
colnames(RMA_raw_GSE15960_CRC) = colnames(RMA_raw_GSE15960)
head(colnames(RMA_raw_GSE15960_CRC))
head(row.names(RMA_raw_GSE15960_CRC))
head(pData(RMA_raw_GSE15960_CRC))
GSE15960 = RMA_raw_GSE15960_CRC

# Plots
graphics::boxplot(exprs(GSE15960))
affy::hist(exprs(GSE15960))

# Save
# save(GSE15960, file ="GSE15960.rda")
```

2.1.4 Affymetrix array: GSE110224 dataset

The GSE110224 was published by Vlachavas et al in 2019 (PMID: 30809322) and includes 17 patients with histologically confirmed colorectal adenocarcinomas and adjacent normal samples.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133plus2.db)
library(hgu133plus2cdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE110224 = ReadAffy(celfile.path = "GSE110224_RAW")
raw_GSE110224
dim(exprs(raw_GSE110224))

# Quality controls

table(probes(raw_GSE110224,"pm") >= probes(raw_GSE110224,"mm"))

affy::hist(raw_GSE110224)
affy::boxplot(raw_GSE110224)

# arrayQualityMetrics(expressionset = raw_GSE110224,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE110224 = affy::rma(raw_GSE110224)
dim(exprs(RMA_raw_GSE110224))
class(RMA_raw_GSE110224)

# write.csv(colnames(RMA_raw_GSE110224), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE110224_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE110224)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE110224_CRC = RMA_raw_GSE110224
pData(RMA_raw_GSE110224_CRC) = data.frame(Series_sample_id,
                            Type, Pair)
colnames(RMA_raw_GSE110224_CRC) = colnames(RMA_raw_GSE110224)
head(colnames(RMA_raw_GSE110224_CRC))
head(row.names(RMA_raw_GSE110224_CRC))
head(pData(RMA_raw_GSE110224_CRC))
```

```
GSE110224 = RMA_raw_GSE110224_CRC

# Plots
graphics::boxplot(exprs(GSE110224))
affy::hist(exprs(GSE110224))

# Save
# save(GSE110224, file ="GSE110224.rda")
```

2.1.5 Affymetrix array: GSE110223 dataset

The GSE110223 also was published by Vlachavas et al in 2019 (PMID: 30809322), but these dataset includes other 13 patients with histologically confirmed colorectal adenocarcinomas and adjacent normal samples, using U133A array.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133a.db)
library(hgu133acdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE110223 = ReadAffy(celfile.path = "GSE110223_RAW")
raw_GSE110223
dim(exprs(raw_GSE110223))

# Quality controls

table(probes(raw_GSE110223,"pm") >= probes(raw_GSE110223,"mm"))

affy::hist(raw_GSE110223)
affy::boxplot(raw_GSE110223)

# arrayQualityMetrics(expressionset = raw_GSE110223,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE110223 = affy::rma(raw_GSE110223)
dim(exprs(RMA_raw_GSE110223))
class(RMA_raw_GSE110223)

# write.csv(colnames(RMA_raw_GSE110223), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE110223_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE110223)
head(metadata_fen)
```

```
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE110223_CRC = RMA_raw_GSE110223
pData(RMA_raw_GSE110223_CRC) = data.frame(Series_sample_id,
                            Type, Pair)
colnames(RMA_raw_GSE110223_CRC) = colnames(RMA_raw_GSE110223)
head(colnames(RMA_raw_GSE110223_CRC))
head(row.names(RMA_raw_GSE110223_CRC))
head(pData(RMA_raw_GSE110223_CRC))
GSE110223 = RMA_raw_GSE110223_CRC

# Plots
graphics::boxplot(exprs(GSE110223))
affy::hist(exprs(GSE110223))

# Save
# save(GSE110223, file ="GSE110223.rda")
```

2.1.6 Affymetrix array: GSE44861 dataset

The GSE44861 was published by Ryan et al in 2014 (PMID: 23982929) and includes 111 colon tissues samples from tumors and adjacent noncancerous tissues. Of these, 47 patients with paired samples (94 samples in total) were selected.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133a.db)
library(hthgu133acdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE44861 = ReadAffy(celfile.path = "GSE44861_RAW")
raw_GSE44861
dim(exprs(raw_GSE44861))

# Quality controls

table(probes(raw_GSE44861,"pm") >= probes(raw_GSE44861,"mm"))

affy::hist(raw_GSE44861)
affy::boxplot(raw_GSE44861)

# arrayQualityMetrics(expressionset = raw_GSE44861,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA
```

```
RMA_raw_GSE44861 = affy::rma(raw_GSE44861)
dim(exprs(RMA_raw_GSE44861))
class(RMA_raw_GSE44861)

# write.csv(colnames(RMA_raw_GSE44861), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE44861_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE44861)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Patient_ID = metadata_fen$patient_ID
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE44861_CRC = RMA_raw_GSE44861
pData(RMA_raw_GSE44861_CRC) = data.frame(Series_sample_id,
                          Patient_ID, Type, Pair)
colnames(RMA_raw_GSE44861_CRC) = colnames(RMA_raw_GSE44861)
head(colnames(RMA_raw_GSE44861_CRC))
head(row.names(RMA_raw_GSE44861_CRC))
head(pData(RMA_raw_GSE44861_CRC))
GSE44861 = RMA_raw_GSE44861_CRC

# Plots
graphics::boxplot(exprs(GSE44861))
affy::hist(exprs(GSE44861))

# Save
# save(GSE44861, file ="GSE44861.rda")
```

2.1.7 Affymetrix array: GSE44076 dataset

The GSE44076 dataset was published by Cordero et al in 2014 (PMID: 25253512). The dataset include 246 samples: 50 mucosa samples from healthy donnor, 98 patients with colorectal tumor and 98 adjacent normal tissue. We analyze only the paired samples.

```
library(affy)
library(org.Hs.eg.db)
library(hgu219.db)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE44076 = ReadAffy(celfile.path = "GSE44076_RAW")
raw_GSE44076

dim(exprs(raw_GSE44076))
```

```r
# Quality controls

table(probes(raw_GSE44076,"pm") >= probes(raw_GSE44076,"mm"))

affy::hist(raw_GSE44076)
affy::boxplot(raw_GSE44076)

# arrayQualityMetrics(expressionset = raw_GSE44076,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE44076 = affy::rma(raw_GSE44076)
dim(exprs(RMA_raw_GSE44076))
class(RMA_raw_GSE44076)
# write.csv(colnames(RMA_raw_GSE44076), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE44076_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE44076)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Patient_ID = metadata_fen$patient_ID
Type = metadata_fen$type
Pair = metadata_fen$pair
Stage = metadata_fen$stage
Location = metadata_fen$location
Gender = metadata_fen$gender
Age = metadata_fen$age
RMA_raw_GSE44076_CRC = RMA_raw_GSE44076
pData(RMA_raw_GSE44076_CRC) = data.frame(Series_sample_id,
                         Patient_ID, Type, Pair, Stage, Location, Gender, Age)
colnames(RMA_raw_GSE44076_CRC) = colnames(RMA_raw_GSE44076)
head(colnames(RMA_raw_GSE44076_CRC))
head(row.names(RMA_raw_GSE44076_CRC))
head(pData(RMA_raw_GSE44076_CRC))
GSE44076 = RMA_raw_GSE44076_CRC

# Plots
graphics::boxplot(exprs(GSE44076))
affy::hist(exprs(GSE44076))

# Save
# save(GSE44076, file ="GSE44076.rda")
```

2.1.8 Affymetrix array: GSE18462 dataset

The GSE18462 dataset was published by Wang and Tsai 2009 (not paper associated) and includes 8 samples: paired normal colon and primary colorectal carcinoma; between primary colorectal carcinoma and liver metas-

tasis colorectal carcinoma. Only 4 samples; 2 tumor and 2 normal paired were analyzed. This data set was discarded, too few samples.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133plus2.db)
library(hgu133plus2cdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE18462 = ReadAffy(celfile.path = "GSE18462_RAW")
raw_GSE18462
dim(exprs(raw_GSE18462))

# Quality controls

table(probes(raw_GSE18462,"pm") >= probes(raw_GSE18462,"mm"))

affy::hist(raw_GSE18462)
affy::boxplot(raw_GSE18462)

# arrayQualityMetrics(expressionset = raw_GSE18462,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE18462 = affy::rma(raw_GSE18462)
dim(exprs(RMA_raw_GSE18462))
class(RMA_raw_GSE18462)

# write.csv(colnames(RMA_raw_GSE18462), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE18462_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE18462)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE18462_CRC = RMA_raw_GSE18462
pData(RMA_raw_GSE18462_CRC) = data.frame(Series_sample_id,
                           Type, Pair)
colnames(RMA_raw_GSE18462_CRC) = colnames(RMA_raw_GSE18462)
head(colnames(RMA_raw_GSE18462_CRC))
head(row.names(RMA_raw_GSE18462_CRC))
head(pData(RMA_raw_GSE18462_CRC))
GSE18462 = RMA_raw_GSE18462_CRC
```

```
# Plots
graphics::boxplot(exprs(GSE18462))
affy::hist(exprs(GSE18462))

# Save
# save(GSE18462, file ="Normalized_annonated_CRC/GSE18462.rda")
```

2.1.9 Affymetrix array: GSE41328 dataset

The GSE41328 dataset was published by Lin et al in 2012 (PMID: 17160039). The dataset include five colorectal adenocarcinomas and matched normal colonic tissues. The 10 samples are analyzed by 2 laboratories (20 data), so we only include lab1 samples in the analysis.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133plus2.db)
library(hgu133plus2cdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE41328 = ReadAffy(celfile.path = "GSE41328_RAW")
raw_GSE41328
dim(exprs(raw_GSE41328))

# Quality controls

table(probes(raw_GSE41328,"pm") >= probes(raw_GSE41328,"mm"))

affy::hist(raw_GSE41328)
affy::boxplot(raw_GSE41328)

# arrayQualityMetrics(expressionset = raw_GSE41328,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE41328 = affy::rma(raw_GSE41328)
dim(exprs(RMA_raw_GSE41328))
class(RMA_raw_GSE41328)

# write.csv(colnames(RMA_raw_GSE41328), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE41328_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE41328)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
```

```
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE41328_CRC = RMA_raw_GSE41328
pData(RMA_raw_GSE41328_CRC) = data.frame(Series_sample_id,
                            Type, Pair)
colnames(RMA_raw_GSE41328_CRC) = colnames(RMA_raw_GSE41328)
head(colnames(RMA_raw_GSE41328_CRC))
head(row.names(RMA_raw_GSE41328_CRC))
head(pData(RMA_raw_GSE41328_CRC))
GSE41328 = RMA_raw_GSE41328_CRC

# Plots
graphics::boxplot(exprs(GSE41328))
affy::hist(exprs(GSE41328))

# Save
# save(GSE41328, file ="GSE41328.rda")
```

2.1.10 Affymetrix array: GSE18105 dataset

The GSE18105 dataset was published by Matsuyama et al in 2010 (PMID: 20162577) and includes 111 samples (77 for LCM samples, and 17 pairs for homogenized samples from CRC tumor and adjacent tissues). Only the 17 pairs were included.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133plus2.db)
library(hgu133plus2cdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE18105 = ReadAffy(celfile.path = "GSE18105_RAW")
raw_GSE18105
dim(exprs(raw_GSE18105))

# Quality controls

table(probes(raw_GSE18105,"pm") >= probes(raw_GSE18105,"mm"))

affy::hist(raw_GSE18105)
affy::boxplot(raw_GSE18105)

# arrayQualityMetrics(expressionset = raw_GSE18105,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE18105 = affy::rma(raw_GSE18105)
dim(exprs(RMA_raw_GSE18105))
```

```
class(RMA_raw_GSE18105)

# write.csv(colnames(RMA_raw_GSE18105), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE18105_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE18105)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Patient_ID = metadata_fen$patient_ID
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE18105_CRC = RMA_raw_GSE18105
pData(RMA_raw_GSE18105_CRC) = data.frame(Series_sample_id,
                           Patient_ID, Type, Pair)
colnames(RMA_raw_GSE18105_CRC) = colnames(RMA_raw_GSE18105)
head(colnames(RMA_raw_GSE18105_CRC))
head(row.names(RMA_raw_GSE18105_CRC))
head(pData(RMA_raw_GSE18105_CRC))
GSE18105 = RMA_raw_GSE18105_CRC

# Plots
graphics::boxplot(exprs(GSE18105))
affy::hist(exprs(GSE18105))

# Save
# save(GSE18105, file ="GSE18105.rda")
```

2.2 Non-paired Colorectal cancer datasets

2.2.1 Affymetrix array: GSE21510 dataset

The GSE21510 dataset was published by Tsukamoto et al, 2011 (PMID: 21270110) and include 104 samples from laser-capture microdissection (LCM) and 44 homogenized normal tissues, of colorectal cancer patients. In the paper dataset analysis, the data were normalized separately for LCM dataset and homogenized tissue dataset using log2-transformed values by RMA. Howeber, is it correct to normalize controls on one side and cases on the other?

```
library(affy)
library(org.Hs.eg.db)
library(hgu133plus2.db)
library(hgu133plus2cdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data
```

```
raw_GSE21510 = ReadAffy(celfile.path = "GSE21510_RAW")
raw_GSE21510
dim(exprs(raw_GSE21510))

# Quality controls

table(probes(raw_GSE21510,"pm") >= probes(raw_GSE21510,"mm"))

affy::hist(raw_GSE21510)
affy::boxplot(raw_GSE21510)

# arrayQualityMetrics(expressionset = raw_GSE21510,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE21510 = affy::rma(raw_GSE21510)
dim(exprs(RMA_raw_GSE21510))
class(RMA_raw_GSE21510)

# write.csv(colnames(RMA_raw_GSE21510), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE21510_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE21510)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Patient_ID = metadata_fen$patient_ID
Type = metadata_fen$type
Pair = metadata_fen$pair
Stage = metadata_fen$stage

RMA_raw_GSE21510_CRC = RMA_raw_GSE21510
pData(RMA_raw_GSE21510_CRC) = data.frame(Series_sample_id,
                          Patient_ID, Type, Pair, Stage)
colnames(RMA_raw_GSE21510_CRC) = colnames(RMA_raw_GSE21510)
head(colnames(RMA_raw_GSE21510_CRC))
head(row.names(RMA_raw_GSE21510_CRC))
head(pData(RMA_raw_GSE21510_CRC))
GSE21510 = RMA_raw_GSE21510_CRC

# Plots

graphics::boxplot(exprs(GSE21510))
affy::hist(exprs(GSE21510))

# Save
# save(GSE21510, file ="GSE21510.rda")
```

2.2.2 Affymetrix array: GSE24514 dataset

The GSE24514 dataset was published by Alhopuro et al in 2012 (PMID: 21544814), including 34 microsatellite instability (MSI) colorectal cancers and 15 normal colonic mucosas.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133a.db)
library(hgu133acdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE24514 = ReadAffy(celfile.path = "GSE24514_RAW")
raw_GSE24514
dim(exprs(raw_GSE24514))

# Quality controls

table(probes(raw_GSE24514,"pm") >= probes(raw_GSE24514,"mm"))

affy::hist(raw_GSE24514)
affy::boxplot(raw_GSE24514)

# arrayQualityMetrics(expressionset = raw_GSE24514,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE24514 = affy::rma(raw_GSE24514)
dim(exprs(RMA_raw_GSE24514))
class(RMA_raw_GSE24514)

# write.csv(colnames(RMA_raw_GSE24514), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE24514_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE24514)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Patient_ID = metadata_fen$patient_ID
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE24514_CRC = RMA_raw_GSE24514
pData(RMA_raw_GSE24514_CRC) = data.frame(Series_sample_id,
                           Patient_ID, Type, Pair)
colnames(RMA_raw_GSE24514_CRC) = colnames(RMA_raw_GSE24514)
head(colnames(RMA_raw_GSE24514_CRC))
head(row.names(RMA_raw_GSE24514_CRC))
```

```
head(pData(RMA_raw_GSE24514_CRC))
GSE24514 = RMA_raw_GSE24514_CRC

# Plots
graphics::boxplot(exprs(GSE24514))
affy::hist(exprs(GSE24514))

# Save
# save(GSE24514, file ="GSE24514.rda")
```

2.2.3 Affymetrix array: GSE4183 dataset

The GSE4183 dataset was published by Galamb in 2008 (PMID:18776587). The dataset includes colonic biopsies of 15 patients with CRC (cancer), 15 with adenoma (tumor), 15 with inflammatory bowel diseases (IBD) and 8 healthy normal controls. The inflammatory bowel diseases samples were excluded for the analysis.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133plus2.db)
library(hgu133plus2cdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE4183 = ReadAffy(celfile.path = "GSE4183_RAW")
raw_GSE4183
dim(exprs(raw_GSE4183))

# Quality controls

table(probes(raw_GSE4183,"pm") >= probes(raw_GSE4183,"mm"))

affy::hist(raw_GSE4183)
affy::boxplot(raw_GSE4183)

# arrayQualityMetrics(expressionset = raw_GSE4183,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE4183 = affy::rma(raw_GSE4183)
dim(exprs(RMA_raw_GSE4183))
class(RMA_raw_GSE4183)

# write.csv(colnames(RMA_raw_GSE4183), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE4183_RAW/metadata.csv",
                        header = TRUE,sep=",")
```

```
rownames(metadata_fen) = colnames(RMA_raw_GSE4183)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Patient_ID = metadata_fen$patient_ID
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE4183_CRC = RMA_raw_GSE4183
pData(RMA_raw_GSE4183_CRC) = data.frame(Series_sample_id,
                          Patient_ID, Type, Pair)
colnames(RMA_raw_GSE4183_CRC) = colnames(RMA_raw_GSE4183)
head(colnames(RMA_raw_GSE4183_CRC))
head(row.names(RMA_raw_GSE4183_CRC))
head(pData(RMA_raw_GSE4183_CRC))
GSE4183 = RMA_raw_GSE4183_CRC

# Plots
graphics::boxplot(exprs(GSE4183))
affy::hist(exprs(GSE4183))

# Save
# save(GSE4183, file ="GSE4183.rda")
```

2.2.4 Affymetrix array: GSE9348 dataset

The GSE9348 was published by Hong in 2010 (PMID: 20143136) and includes tumors from age- and ethnicity-matched of 70 patients and biopsies from 12 healthy controls.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133plus2.db)
library(hgu133plus2cdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE9348 = ReadAffy(celfile.path = "GSE9348_RAW")
raw_GSE9348
dim(exprs(raw_GSE9348))

# Quality controls

table(probes(raw_GSE9348,"pm") >= probes(raw_GSE9348,"mm"))

affy::hist(raw_GSE9348)
affy::boxplot(raw_GSE9348)

# arrayQualityMetrics(expressionset = raw_GSE9348,
# do.logtransform = TRUE) ## Transforma a log2
```

```
# Normalization with RMA

RMA_raw_GSE9348 = affy::rma(raw_GSE9348)
dim(exprs(RMA_raw_GSE9348))
class(RMA_raw_GSE9348)

# write.csv(colnames(RMA_raw_GSE9348), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE9348_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE9348)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Patient_ID = metadata_fen$patient_ID
Type = metadata_fen$type
Pair = metadata_fen$pair
Sex = metadata_fen$Sex
Age = metadata_fen$age

RMA_raw_GSE9348_CRC = RMA_raw_GSE9348
pData(RMA_raw_GSE9348_CRC) = data.frame(Series_sample_id,
                          Patient_ID, Type, Pair, Sex, Age)
colnames(RMA_raw_GSE9348_CRC) = colnames(RMA_raw_GSE9348)
head(colnames(RMA_raw_GSE9348_CRC))
head(row.names(RMA_raw_GSE9348_CRC))
head(pData(RMA_raw_GSE9348_CRC))
GSE9348 = RMA_raw_GSE9348_CRC

# Plots
graphics::boxplot(exprs(GSE9348))
affy::hist(exprs(GSE9348))

# Save
# save(GSE9348, file ="GSE9348.rda")
```

2.2.5 Affymetrix array: GSE20916 dataset

The GSE20916 dataset was published by Skrzypczak et al, 2010 (PMID: 20957034). This dataset includes 145 samples: 36 adenocarcinoma, 45 adenoma, 5 colonic crypt epithelial cells adenoma, 5 colonic crypt epithelial cells carcinoma, 5 colonic crypt epithelial cells distant normal colon, 5 colonic crypt epithelial cells normal colon, 5 mucosa adenoma, 5 mucosa carsinoma, 5 mucosa distant normal colon, 5 mucosa normal colon and 24 normal colon. Here, these samples were grouped in 44 normal and 101 tumor.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133plus2.db)
library(hgu133plus2cdf)
```

```r
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE20916 = ReadAffy(celfile.path = "GSE20916_RAW")
raw_GSE20916
dim(exprs(raw_GSE20916))

# Quality controls

table(probes(raw_GSE20916,"pm") >= probes(raw_GSE20916,"mm"))

affy::hist(raw_GSE20916)
affy::boxplot(raw_GSE20916)

# arrayQualityMetrics(expressionset = raw_GSE20916,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE20916 = affy::rma(raw_GSE20916)
dim(exprs(RMA_raw_GSE20916))
class(RMA_raw_GSE20916)

# write.csv(colnames(RMA_raw_GSE20916), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE20916_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE20916)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Patient_ID = metadata_fen$patient_ID
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE20916_CRC = RMA_raw_GSE20916
pData(RMA_raw_GSE20916_CRC) = data.frame(Series_sample_id,
                           Patient_ID, Type, Pair)
colnames(RMA_raw_GSE20916_CRC) = colnames(RMA_raw_GSE20916)
head(colnames(RMA_raw_GSE20916_CRC))
head(row.names(RMA_raw_GSE20916_CRC))
head(pData(RMA_raw_GSE20916_CRC))
GSE20916 = RMA_raw_GSE20916_CRC

# Plots
graphics::boxplot(exprs(GSE20916))
affy::hist(exprs(GSE20916))

# Save
```

```
# save(GSE20916, file ="GSE20916.rda")
```

2.2.6 Affymetrix array: GSE23878 dataset

The GSE23878 dataset was published by Uddin et al. in 2011 (PMID: 21281787) and include 35 colorectal cancer samples versus 24 normal samples.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133plus2.db)
library(hgu133plus2cdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE23878 = ReadAffy(celfile.path = "GSE23878_RAW")
raw_GSE23878
dim(exprs(raw_GSE23878))

# Quality controls

table(probes(raw_GSE23878,"pm") >= probes(raw_GSE23878,"mm"))

affy::hist(raw_GSE23878)
affy::boxplot(raw_GSE23878)

# arrayQualityMetrics(expressionset = raw_GSE23878,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE23878 = affy::rma(raw_GSE23878)
dim(exprs(RMA_raw_GSE23878))
class(RMA_raw_GSE23878)

# write.csv(colnames(RMA_raw_GSE23878), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE23878_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE23878)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Patient_ID = metadata_fen$patient_ID
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE23878_CRC = RMA_raw_GSE23878
pData(RMA_raw_GSE23878_CRC) = data.frame(Series_sample_id,
                           Patient_ID, Type, Pair)
```

```
colnames(RMA_raw_GSE23878_CRC) = colnames(RMA_raw_GSE23878)
head(colnames(RMA_raw_GSE23878_CRC))
head(row.names(RMA_raw_GSE23878_CRC))
head(pData(RMA_raw_GSE23878_CRC))
GSE23878 = RMA_raw_GSE23878_CRC

# Plots
graphics::boxplot(exprs(GSE23878))
affy::hist(exprs(GSE23878))

# Save
# save(GSE23878, file ="GSE23878.rda")
```

2.2.7 Affymetrix array: GSE33113 dataset

The GSE33113 dataset was published by de Sousa E Melo F et al. in 2011 (PMID:22056143) and include 90 colorectal cancer stage II samples versus 6 normal samples.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133plus2.db)
library(hgu133plus2cdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE33113 = ReadAffy(celfile.path = "GSE33113_RAW")
raw_GSE33113
dim(exprs(raw_GSE33113))

# Quality controls

table(probes(raw_GSE33113,"pm") >= probes(raw_GSE33113,"mm"))

affy::hist(raw_GSE33113)
affy::boxplot(raw_GSE33113)

# arrayQualityMetrics(expressionset = raw_GSE33113,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE33113 = affy::rma(raw_GSE33113)
dim(exprs(RMA_raw_GSE33113))
class(RMA_raw_GSE33113)

# write.csv(colnames(RMA_raw_GSE33113), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE33113_RAW/metadata.csv",
```

```
                          header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE33113)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE33113_CRC = RMA_raw_GSE33113
pData(RMA_raw_GSE33113_CRC) = data.frame(Series_sample_id,
                           Type, Pair)
colnames(RMA_raw_GSE33113_CRC) = colnames(RMA_raw_GSE33113)
head(colnames(RMA_raw_GSE33113_CRC))
head(row.names(RMA_raw_GSE33113_CRC))
head(pData(RMA_raw_GSE33113_CRC))
GSE33113 = RMA_raw_GSE33113_CRC

# Plots
graphics::boxplot(exprs(GSE33113))
affy::hist(exprs(GSE33113))

# Save
# save(GSE33113, file ="GSE33113.rda")
```

2.2.8 Affymetrix array: GSE37364 dataset

The GSE37364 was published by Valcz et al in 2014 (PMID: 25405986) and includes 29 Adenoma, 14 Adenocarcinoma, 13 Dukes and 38 normal samples. Of these, only Adenocarcinoma and normal samples were analysed.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133plus2.db)
library(hgu133plus2cdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE37364 = ReadAffy(celfile.path = "GSE37364_RAW")
raw_GSE37364
dim(exprs(raw_GSE37364))

# Quality controls

table(probes(raw_GSE37364,"pm") >= probes(raw_GSE37364,"mm"))

affy::hist(raw_GSE37364)
affy::boxplot(raw_GSE37364)

# arrayQualityMetrics(expressionset = raw_GSE37364,
# do.logtransform = TRUE) ## Transforma a log2
```

```
# Normalization with RMA

RMA_raw_GSE37364 = affy::rma(raw_GSE37364)
dim(exprs(RMA_raw_GSE37364))
class(RMA_raw_GSE37364)

# write.csv(colnames(RMA_raw_GSE37364), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE37364_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE37364)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Patient_ID = metadata_fen$patient_ID
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE37364_CRC = RMA_raw_GSE37364
pData(RMA_raw_GSE37364_CRC) = data.frame(Series_sample_id,
                           Patient_ID, Type, Pair)
colnames(RMA_raw_GSE37364_CRC) = colnames(RMA_raw_GSE37364)
head(colnames(RMA_raw_GSE37364_CRC))
head(row.names(RMA_raw_GSE37364_CRC))
head(pData(RMA_raw_GSE37364_CRC))
GSE37364 = RMA_raw_GSE37364_CRC

# Plots
graphics::boxplot(exprs(GSE37364))
affy::hist(exprs(GSE37364))

# Save
# save(GSE37364, file ="GSE37364.rda")
```

2.2.9 Affymetrix array: GSE77953 dataset

The GSE77953 dataset was published by Qu et al in 2016 (PMID: 27270421) and includes 58 laser capture micro-dissected tissues, 17 Adenoma, 17 Carcinoma, 11 Metastasis, 7 normal crypt and 6 normal surface. We includes only the carcinoma and normal samples.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133a.db)
library(hgu133acdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE77953 = ReadAffy(celfile.path = "GSE77953_RAW")
```

```
raw_GSE77953
dim(exprs(raw_GSE77953))

# Quality controls

table(probes(raw_GSE77953,"pm") >= probes(raw_GSE77953,"mm"))

affy::hist(raw_GSE77953)
affy::boxplot(raw_GSE77953)

# arrayQualityMetrics(expressionset = raw_GSE77953,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE77953 = affy::rma(raw_GSE77953)
dim(exprs(RMA_raw_GSE77953))
class(RMA_raw_GSE77953)

# write.csv(colnames(RMA_raw_GSE77953), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE77953_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE77953)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE77953_CRC = RMA_raw_GSE77953
pData(RMA_raw_GSE77953_CRC) = data.frame(Series_sample_id,
                           Type, Pair)
colnames(RMA_raw_GSE77953_CRC) = colnames(RMA_raw_GSE77953)
head(colnames(RMA_raw_GSE77953_CRC))
head(row.names(RMA_raw_GSE77953_CRC))
head(pData(RMA_raw_GSE77953_CRC))
GSE77953 = RMA_raw_GSE77953_CRC

# Plots
graphics::boxplot(exprs(GSE77953))
affy::hist(exprs(GSE77953))

# Save
# save(GSE77953, file ="GSE77953.rda")
```

2.2.10 Affymetrix array: GSE49355 dataset

The GSE49355 dataset was published by Del Rio et. al. in 2013 (PMID: 24023955, PMID: 17327601) and includes 19 liver metastasis samples, 18 normal colon samples and 20 primary tumour, of these, 15 patient have

paired samples. We include all normal and tumor samples for this analysis.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133a.db)
library(hgu133acdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE49355 = ReadAffy(celfile.path = "GSE49355_RAW")
raw_GSE49355
dim(exprs(raw_GSE49355))

# Quality controls

table(probes(raw_GSE49355,"pm") >= probes(raw_GSE49355,"mm"))

affy::hist(raw_GSE49355)
affy::boxplot(raw_GSE49355)

# arrayQualityMetrics(expressionset = raw_GSE49355,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE49355 = affy::rma(raw_GSE49355)
dim(exprs(RMA_raw_GSE49355))
class(RMA_raw_GSE49355)

# write.csv(colnames(RMA_raw_GSE49355), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE49355_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE49355)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Type = metadata_fen$type
Pair = metadata_fen$pair
Sex = metadata_fen$Sex
Age = metadata_fen$Age

RMA_raw_GSE49355_CRC = RMA_raw_GSE49355
pData(RMA_raw_GSE49355_CRC) = data.frame(Series_sample_id,
                           Type, Pair, Sex, Age)
colnames(RMA_raw_GSE49355_CRC) = colnames(RMA_raw_GSE49355)
head(colnames(RMA_raw_GSE49355_CRC))
head(row.names(RMA_raw_GSE49355_CRC))
head(pData(RMA_raw_GSE49355_CRC))
GSE49355 = RMA_raw_GSE49355_CRC
```

```
# Plots
graphics::boxplot(exprs(GSE49355))
affy::hist(exprs(GSE49355))

# Save
# save(GSE49355, file ="GSE49355.rda")
```

2.2.11 Affymetrix array: GSE41258 dataset

The GSE41258 dataset was published by Sheffer et. al. in 2012 (PMID: 19359472) and includes 12 cell lines samples, 67 liver metastasis samples, 2 microadenoma, 54 normal colon, 13 normal liver, 7 normal lung, 49 polyp samples, 186 primary tumor. We include the 54 normal colon and the 186 primary tumor samples.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133a.db)
library(hgu133acdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE41258 = ReadAffy(celfile.path = "GSE41258_RAW")
raw_GSE41258
dim(exprs(raw_GSE41258))

# Quality controls

table(probes(raw_GSE41258,"pm") >= probes(raw_GSE41258,"mm"))

affy::hist(raw_GSE41258)
affy::boxplot(raw_GSE41258)

# arrayQualityMetrics(expressionset = raw_GSE41258,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE41258 = affy::rma(raw_GSE41258)
dim(exprs(RMA_raw_GSE41258))
class(RMA_raw_GSE41258)

# write.csv(colnames(RMA_raw_GSE41258), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE41258_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE41258)
head(metadata_fen)
sapply(metadata_fen,class)
```

```
Series_sample_id = metadata_fen$series_sample_id
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE41258_CRC = RMA_raw_GSE41258
pData(RMA_raw_GSE41258_CRC) = data.frame(Series_sample_id,
                            Type, Pair)
colnames(RMA_raw_GSE41258_CRC) = colnames(RMA_raw_GSE41258)
head(colnames(RMA_raw_GSE41258_CRC))
head(row.names(RMA_raw_GSE41258_CRC))
head(pData(RMA_raw_GSE41258_CRC))
GSE41258 = RMA_raw_GSE41258_CRC

# Plots
graphics::boxplot(exprs(GSE41258))
affy::hist(exprs(GSE41258))

# Save
# save(GSE41258, file ="GSE41258.rda")
```

2.2.12 Affymetrix array: GSE19249 dataset

The GSE19249 dataset was published by Abdueva et al in 2010 (PMID: 20522636). The dataset include 53 samples: 16 FF colon, 7 FFPE colon, 6 FF lung, 8 FFPE lung, 8 FF kidney, and 8 FFPE kidney specimens. We analyzed only the 23 (8 normal and 15 tumor) sample for colon.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133a2.db)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE19249 = ReadAffy(celfile.path = "GSE19249_RAW")
raw_GSE19249
dim(exprs(raw_GSE19249))

# Quality controls

table(probes(raw_GSE19249,"pm") >= probes(raw_GSE19249,"mm"))

affy::hist(raw_GSE19249)
affy::boxplot(raw_GSE19249)

# arrayQualityMetrics(expressionset = raw_GSE19249,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE19249 = affy::rma(raw_GSE19249)
dim(exprs(RMA_raw_GSE19249))
```

```
class(RMA_raw_GSE19249)
# write.csv(colnames(RMA_raw_GSE19249), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE19249_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE19249)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE19249_CRC = RMA_raw_GSE19249
pData(RMA_raw_GSE19249_CRC) = data.frame(Series_sample_id,
                        Type, Pair)
colnames(RMA_raw_GSE19249_CRC) = colnames(RMA_raw_GSE19249)
head(colnames(RMA_raw_GSE19249_CRC))
head(row.names(RMA_raw_GSE19249_CRC))
head(pData(RMA_raw_GSE19249_CRC))
GSE19249 = RMA_raw_GSE19249_CRC

# Plots
graphics::boxplot(exprs(GSE19249))
affy::hist(exprs(GSE19249))

# Save
# save(GSE19249, file ="GSE19249.rda")
```

2.2.13 Affymetrix array: GSE22242 dataset

The GSE22242 dataset was published by Tang et al in 2010 (PMID: 20878084) and includes 3 CRC, 1 adenoma and 1 normal samples. This data set was discarded, too few samples.

```
library(affy)
library(org.Hs.eg.db)
library(hgu133plus2.db)
library(hgu133plus2cdf)
library(arrayQualityMetrics)
library(geneplotter)

# Load the raw data

raw_GSE22242 = ReadAffy(celfile.path = "GSE22242_RAW")
raw_GSE22242
dim(exprs(raw_GSE22242))

# Quality controls

table(probes(raw_GSE22242,"pm") >= probes(raw_GSE22242,"mm"))
```

```r
affy::hist(raw_GSE22242)
affy::boxplot(raw_GSE22242)

# arrayQualityMetrics(expressionset = raw_GSE22242,
# do.logtransform = TRUE) ## Transforma a log2

# Normalization with RMA

RMA_raw_GSE22242 = affy::rma(raw_GSE22242)
dim(exprs(RMA_raw_GSE22242))
class(RMA_raw_GSE22242)

# write.csv(colnames(RMA_raw_GSE22242), file = "metadata.csv")

# Including the metadata

metadata_fen = read.csv(file = "GSE22242_RAW/metadata.csv",
                        header = TRUE,sep=",")
rownames(metadata_fen) = colnames(RMA_raw_GSE22242)
head(metadata_fen)
sapply(metadata_fen,class)
Series_sample_id = metadata_fen$series_sample_id
Type = metadata_fen$type
Pair = metadata_fen$pair

RMA_raw_GSE22242_CRC = RMA_raw_GSE22242
pData(RMA_raw_GSE22242_CRC) = data.frame(Series_sample_id,
                          Type, Pair)
colnames(RMA_raw_GSE22242_CRC) = colnames(RMA_raw_GSE22242)
head(colnames(RMA_raw_GSE22242_CRC))
head(row.names(RMA_raw_GSE22242_CRC))
head(pData(RMA_raw_GSE22242_CRC))
GSE22242 = RMA_raw_GSE22242_CRC

# Plots
graphics::boxplot(exprs(GSE22242))
affy::hist(exprs(GSE22242))

# Save
# save(GSE22242, file ="GSE22242.rda")
```