Correlation analysis between Expression (microarrays & RNA-seq) and methylation data in a set of cancer colon cell lines data. (1): Creating matched datasets

Alex Sánchez-Pla.
Statistics department. UB
& Statistics and Bioinformatics Unit (UEB). VHIR.

January 15, 2016

Contents


```
[1] "Package knitr already installed"
[1] "Package gdata already installed"
[1] "Package VennDiagram already installed"
[1] "Package Biobase already installed"
[1] "Package annotate already installed"
[1] "Package hgu133plus2.db already installed"
```

1 Introduction

The goal of this study is to check the correlation between expression and methylation values in a set of cell lines that have been analyzed to look for biomarkers

for sensitivity to some drugs used in colon cancer treatment. Expression data have been obtained from microarrays and RNA-seq. Methylation has been measured on Illumina methylation arrays.

2 Data for the analysis

The data correspond to expression and methylation data from a series of colon cancer cell lines characterized by their different sensitivities to drugs.

Data have been generated and preprocessed separately

- Expression microarrays have been normalized using the RMA algorithm. Probesets corresponding to duplicate identifiers have been removed (the probe with the highest variance is retained in each case). Data have been batch-centered to remove batch effect due to the place from where samples were processed.
- Methylation values have been normalized using standard approaches for this type of data. Data from several CPG sites corresponding to the same gene have been averaged. Sites non-associated with a gene have been removed.
- RNA-seq data have been preprocessed using standard approaches and turned into normalized counts using the RPKM algorithm. Only counts that could be assigned to genes have been retained. Genes with an excess of zero counts have been removed.

Preprocessed data have been stored as binary files for its further re-use.

```
> load(file.path(resultsDir,"expresMicroarraysNewFiltered.Rda"))
> load(file.path(resultsDir,"methylationDataNewAgregated.Rda"))
> load(file.path(resultsDir,"RNAseqDataNew.Rda"))
> # which(rownames(expres2)=="ZBTB18")
> # which(rownames(numDataMethByMean)=="ZNF8")
> # which(rownames(dataRNAseq)=="ZBTB18")
> # which(rownames(dataRNAseqA)=="ZBTB18")
```

2.1 Genes in common between the three datasets

2.1.1 Genes in common between the three datasets

The dimensions of the datasets availabe are not the same (although there are two methylation datasets –aggreggated by their mean and by thir highest variance– thir dimensions and names are indeed the same.

```
> class (expres2); dim(expres2)
```

```
[1] "matrix"
[1] 19991    42

> colnames(expres2) <- toupper(colnames(expres2))
> class (numDataMethByMean); dim(numDataMethByMean)

[1] "data.frame"
[1] 14476    46

> class (numDataMethByVar); dim(numDataMethByVar)

[1] "matrix"
[1] 14476    46

> class (dataRNAseqA); dim(dataRNAseqA)

[1] "data.frame"
[1] 17408    59

> numDataMethByMean <- as.matrix(numDataMethByMean)
> numDataMethByMean <-numDataMethByMean[order(rownames(numDataMethByMean)),]
> numDataMethByVar <-numDataMethByVar[order(rownames(numDataMethByVar)),]
> sum(rownames(numDataMethByMean)!=rownames(numDataMethByVar))

[1] 0

> sum(colnames(numDataMethByMean)!=colnames(numDataMethByVar))

[1] 0

> dataRNAseqA <-as.matrix(dataRNAseqA)
```

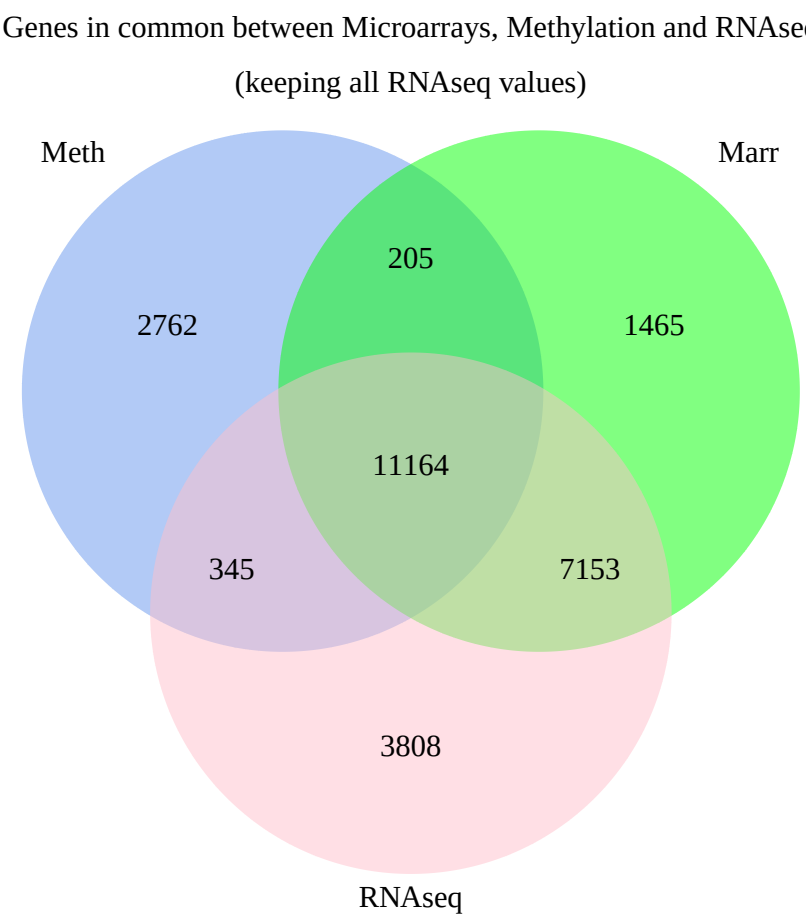
This means that we may expect to have different gene names and different sample names between them.

```
> marrSymbols <- rownames(expres2)
> methSymbols <- rownames(numDataMethByMean)
> RNAseqSymbols <- rownames(dataRNAseq)
> RNAseqSymbolsA <- rownames(dataRNAseqA)
```

If we consider all genes available in the RNAseq files we get:

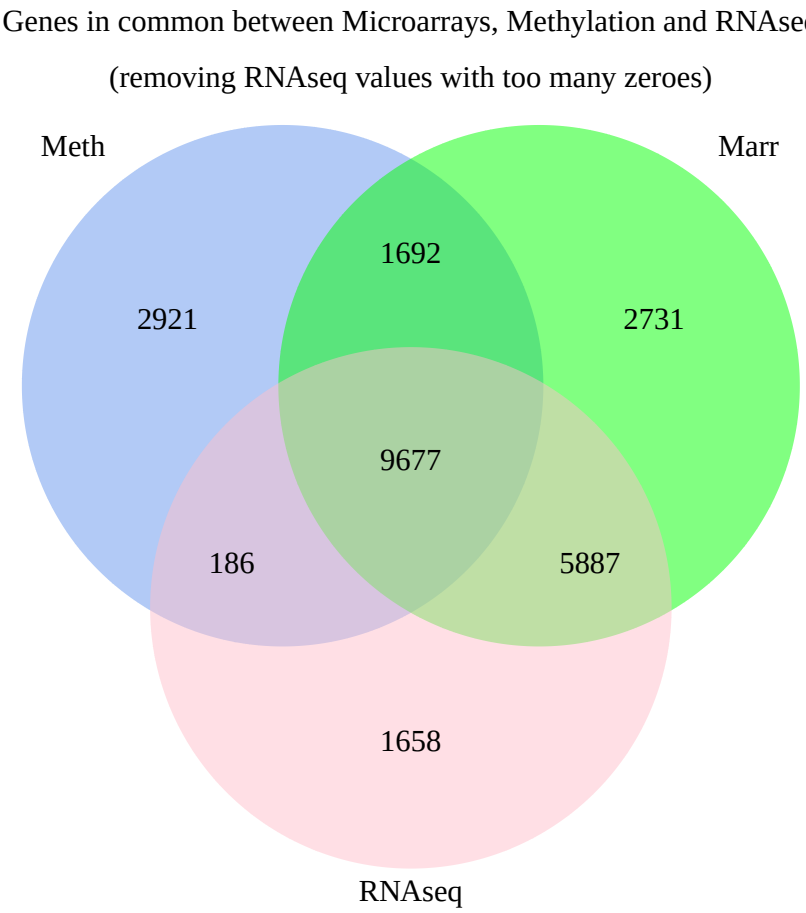
```
> par(mfrow=c(2,1))
> require(VennDiagram)
> vd<- venn.diagram(list(Meth=methSymbols, Marr=marrSymbols, RNAseq=RNAseqSymbols),
+                   filename=NULL,
+                   col = "transparent", fill = c("cornflowerblue", "green", "pink"),
```

```
+                   main ="Genes in common between Microarrays, Methylation and RNAseq",
+                   sub="(keeping all RNAseq values)")
> grid.draw(vd)
```



Ignoring those with too many zeroes the result is:

```
> #dev.new()
> vd1<- venn.diagram(list(Meth=methSymbols, Marr=marrSymbols, RNAseq=RNAseqSymbolsA),
+                   filename=NULL,
+                   col = "transparent", fill = c("cornflowerblue", "green", "pink"),
+                   main ="Genes in common between Microarrays, Methylation and RNAseq",
+                   sub="(removing RNAseq values with too many zeroes)")
> grid.draw(vd1)
```

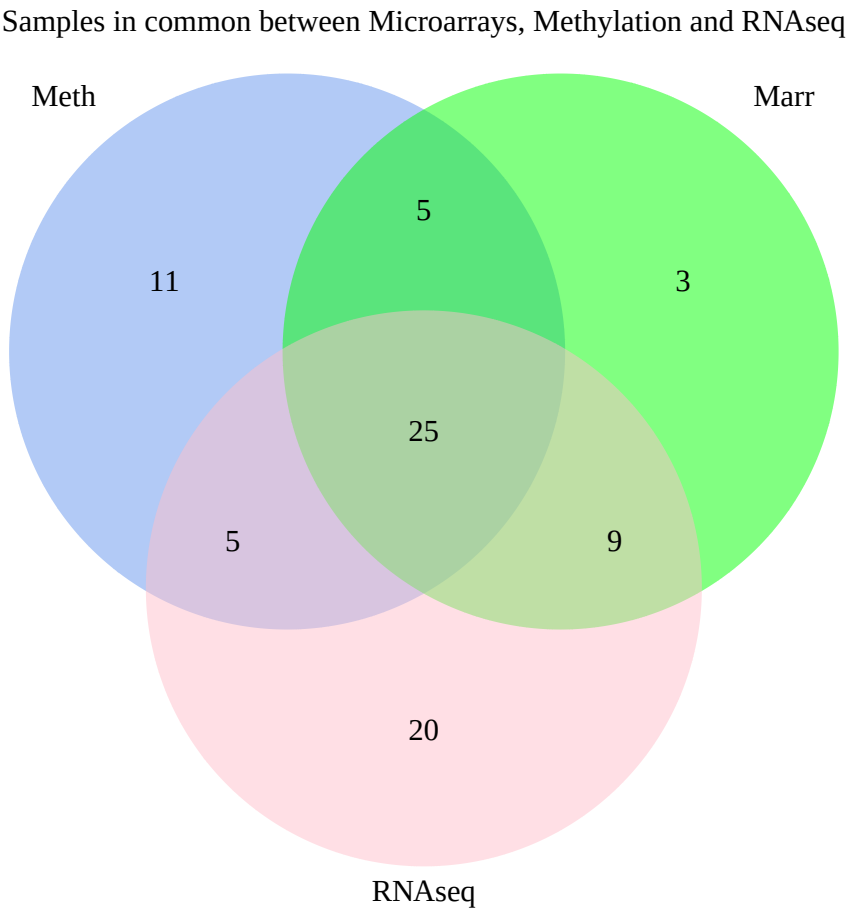


2.1.2 Samples in common between the three datasets

Methylation and gene expression has been measured on many common cell lines, although some are not the same

```
> namesMarr<-marrNames <- toupper(colnames(expres2))
> namesMeth<-methNames <- toupper(colnames(numDataMethByMean))
> namesRNAseq<-RNAseqNames <- toupper(colnames(dataRNAseqA))
```

```
> #dev.new()
> vd2<- venn.diagram(list(Meth=methNames, Marr=marrNames, RNAseq=RNAseqNames),
+                    filename=NULL,
+                    col = "transparent", fill = c("cornflowerblue", "green", "pink"),
+                    main ="Samples in common between Microarrays, Methylation and RNAseq")
> grid.draw(vd2)
```



Before we can investigate the existing correlation between each dataset some work has to be done to match the datasets on a "per-gene" and "per-sample" basis.

```
> sampleNames<-cbind(c(sort(marrNames), sort(methNames), sort(RNAseqNames)),
+                    c(rep("marr", length(marrNames)), rep("meth", length(methNames)),rep("R
> table(sampleNames[,1], sampleNames[,2])


         marr meth RNAseq
  ALA       0    1      0
  C125.PM   0    0      1
  C135      0    0      1
  C70       0    0      1
  CACO2     1    1      1
  CCK81     0    0      1
  CO115     1    1      0
```

```
COLO201    1    1      1
COLO205    1    1      1
COLO320    1    1      1
CX1        0    0      1
DIFI       1    0      1
DLD1       1    1      1
FET        0    1      0
GEO        0    0      1
GP2D       0    0      1
GP5D       0    1      1
HCA7       0    1      1
HCC2998    1    1      1
HCT116     1    1      1
HCT15      1    1      1
HCT8       1    0      1
HDC108     0    1      0
HDC111     0    1      0
HDC114     0    1      0
HDC15      0    1      0
HDC54      0    0      1
HDC57      0    0      1
HDC75      0    1      0
HDC87      0    1      0
HDC9       0    1      0
HDC90      0    0      1
HRA19      0    0      1
HT115      0    0      1
HT29       1    1      1
HT55       0    0      1
IS1        1    1      1
IS2        1    1      1
IS3        0    1      1
KM12       1    1      1
LIM1215    1    1      1
LIM1863    1    0      1
LIM1899    1    0      1
LIM2099    0    0      1
LIM2405    1    1      1
LIM2537    1    0      1
LIM2550    1    0      1
LIM2551    1    0      1
LOVO       1    1      0
LS1034     0    1      0
LS174T     1    1      0
LS411      1    0      0
```

```
LS513      0    1      1
NCIH747    0    0      1
NOMET      0    1      0
RKO        1    1      1
RW2982     1    1      1
RW7213     1    1      1
SKCO1      1    1      1
SNU175     0    0      1
SNUC2B     0    0      1
SW1112     0    0      1
SW1116     1    1      1
SW1222     1    0      1
SW1417     1    0      0
SW403      1    1      1
SW48       1    1      0
SW480      1    0      1
SW620      1    1      1
SW837      1    1      1
SW948      1    1      1
T84        1    1      1
TC71       1    1      0
V9P        0    1      1
VACO10     0    0      1
VACO432    1    0      0
VACO4S     0    0      1
VACO5      1    1      1
```

3 Subsetting datasets to match rows and columns

3.1 Matching columns (cell lines)

As we have seen there is not a one-to-one correspondence between the names of cell lines for which there is methylation or expression (microarray or RNAseq) values.

1. RNAseq and Microarray data:

```
> namesMarr # colnames(dataMeth)

 [1] "CACO2"   "CO115"   "COLO201" "COLO205" "COLO320"
 [6] "DIFI"    "DLD1"    "HCC2998" "HCT116"  "HCT15"
[11] "HCT8"    "HT29"    "IS1"     "IS2"     "KM12"
[16] "LIM1215" "LIM1863" "LIM1899" "LIM2405" "LIM2537"
[21] "LIM2550" "LIM2551" "LOVO"    "LS174T"  "LS411"
```

```
[26] "RKO"     "RW2982"  "RW7213"  "SKCO1"   "SW48"
[31] "SW1116"  "SW1222"  "SW1417"  "SW403"   "SW480"
[36] "SW620"   "SW837"   "SW948"   "T84"     "TC71"
[41] "VACO432" "VACO5"

> namesRNAseq # colnames(dataRNAseq)

 [1] "C125.PM" "C135"    "C70"     "CACO2"   "CCK81"
 [6] "COLO201" "COLO205" "COLO320" "CX1"     "DIFI"
[11] "DLD1"    "GEO"     "GP2D"    "GP5D"    "HCA7"
[16] "HCC2998" "HCT116"  "HCT15"   "HCT8"    "HDC54"
[21] "HDC57"   "HDC90"   "HRA19"   "HT115"   "HT29"
[26] "HT55"    "IS1"     "IS2"     "IS3"     "KM12"
[31] "LIM1215" "LIM1863" "LIM1899" "LIM2099" "LIM2405"
[36] "LIM2537" "LIM2550" "LIM2551" "LS513"   "NCIH747"
[41] "RKO"     "RW2982"  "RW7213"  "SKCO1"   "SNU175"
[46] "SNUC2B"  "SW1112"  "SW1116"  "SW1222"  "SW403"
[51] "SW480"   "SW620"   "SW837"   "SW948"   "T84"
[56] "V9P"     "VACO10"  "VACO4S"  "VACO5"

> common0<- intersect(namesMarr, namesRNAseq); length(common0)

[1] 34
```

2. RNAseq and Methylation data:

```
> namesMeth # colnames(dataMeth)

 [1] "ALA"     "CACO2"   "CO115"   "COLO201" "COLO205"
 [6] "COLO320" "DLD1"    "GP5D"    "HCA7"    "HCC2998"
[11] "HCT116"  "HCT15"   "HDC108"  "HDC9"    "HT29"
[16] "IS1"     "IS2"     "IS3"     "KM12"    "LIM1215"
[21] "LIM2405" "LOVO"    "LS1034"  "LS174T"  "LS513"
[26] "RKO"     "RW2982"  "RW7213"  "SKCO1"   "SW1116"
[31] "SW403"   "SW48"    "SW620"   "SW837"   "SW948"
[36] "T84"     "TC71"    "V9P"     "VACO5"   "FET"
[41] "HDC111"  "HDC114"  "HDC75"   "HDC87"   "HDC15"
[46] "NOMET"

> namesRNAseq # colnames(dataRNAseq)

 [1] "C125.PM" "C135"    "C70"     "CACO2"   "CCK81"
 [6] "COLO201" "COLO205" "COLO320" "CX1"     "DIFI"
[11] "DLD1"    "GEO"     "GP2D"    "GP5D"    "HCA7"
```

```
[16] "HCC2998" "HCT116"  "HCT15"   "HCT8"    "HDC54"
[21] "HDC57"   "HDC90"   "HRA19"   "HT115"   "HT29"
[26] "HT55"    "IS1"     "IS2"     "IS3"     "KM12"
[31] "LIM1215" "LIM1863" "LIM1899" "LIM2099" "LIM2405"
[36] "LIM2537" "LIM2550" "LIM2551" "LS513"   "NCIH747"
[41] "RKO"     "RW2982"  "RW7213"  "SKCO1"   "SNU175"
[46] "SNUC2B"  "SW1112"  "SW1116"  "SW1222"  "SW403"
[51] "SW480"   "SW620"   "SW837"   "SW948"   "T84"
[56] "V9P"     "VACO10"  "VACO4S"  "VACO5"

> common1<- intersect(namesMeth, namesRNAseq); length(common1)

[1] 30
```

3. Microarray and methylation data:

```
> common2<- intersect(namesMeth, namesMarr); length(common2)

[1] 30
```

4. The three of them Microarray, RNAseq and Methylation data:

```
> common3<- intersect(namesRNAseq, intersect(namesMeth, namesMarr)); length(common3)

[1] 25
```

There are only 25 lines for which there are microarrays, RNAseq and expression values **so only these will be used in a first analysis**.

Two types of subsets will be made.

1. One based on the intersection of the three datasets, intended to do a joint analysis of their relations

```
> dataMarrC <- expres2[,common3];dim(dataMarrC); colnames(dataMarrC)

[1] 19991    25
 [1] "CACO2"   "COLO201" "COLO205" "COLO320" "DLD1"
 [6] "HCC2998" "HCT116"  "HCT15"   "HT29"    "IS1"
[11] "IS2"     "KM12"    "LIM1215" "LIM2405" "RKO"
[16] "RW2982"  "RW7213"  "SKCO1"   "SW1116"  "SW403"
[21] "SW620"   "SW837"   "SW948"   "T84"     "VACO5"

> dataMethC <- numDataMethByMean[,common3];dim(dataMethC); colnames(dataMethC)
```

```
[1] 14476    25
 [1] "CACO2"   "COLO201" "COLO205" "COLO320" "DLD1"
 [6] "HCC2998" "HCT116"  "HCT15"   "HT29"    "IS1"
[11] "IS2"     "KM12"    "LIM1215" "LIM2405" "RKO"
[16] "RW2982"  "RW7213"  "SKCO1"   "SW1116"  "SW403"
[21] "SW620"   "SW837"   "SW948"   "T84"     "VACO5"

> dataMethVarC  <- numDataMethByVar[,common3] ;dim(dataMethVarC); colnames(dataMethVarC

[1] 14476    25
 [1] "CACO2"   "COLO201" "COLO205" "COLO320" "DLD1"
 [6] "HCC2998" "HCT116"  "HCT15"   "HT29"    "IS1"
[11] "IS2"     "KM12"    "LIM1215" "LIM2405" "RKO"
[16] "RW2982"  "RW7213"  "SKCO1"   "SW1116"  "SW403"
[21] "SW620"   "SW837"   "SW948"   "T84"     "VACO5"

> dataRNAseqC <- dataRNAseqA[,common3];dim(dataRNAseqC); colnames(dataRNAseqC)

[1] 17408    25
 [1] "CACO2"   "COLO201" "COLO205" "COLO320" "DLD1"
 [6] "HCC2998" "HCT116"  "HCT15"   "HT29"    "IS1"
[11] "IS2"     "KM12"    "LIM1215" "LIM2405" "RKO"
[16] "RW2982"  "RW7213"  "SKCO1"   "SW1116"  "SW403"
[21] "SW620"   "SW837"   "SW948"   "T84"     "VACO5"
```

2. Two based on pairwise intersections, that is the common microarray and methylation values by one side and the common RNAseq and methylation values by the other side.

```
> dataMarrC0<- expres2[,common0];dim(dataMarrC0); colnames(dataMarrC0)

[1] 19991    34
 [1] "CACO2"   "COLO201" "COLO205" "COLO320" "DIFI"
 [6] "DLD1"    "HCC2998" "HCT116"  "HCT15"   "HCT8"
[11] "HT29"    "IS1"     "IS2"     "KM12"    "LIM1215"
[16] "LIM1863" "LIM1899" "LIM2405" "LIM2537" "LIM2550"
[21] "LIM2551" "RKO"     "RW2982"  "RW7213"  "SKCO1"
[26] "SW1116"  "SW1222"  "SW403"   "SW480"   "SW620"
[31] "SW837"   "SW948"   "T84"     "VACO5"

> dataRNAseqC0 <- dataRNAseqA[,common0];dim(dataRNAseqC0); colnames(dataRNAseqC0)

[1] 17408    34
 [1] "CACO2"   "COLO201" "COLO205" "COLO320" "DIFI"
```

```
 [6] "DLD1"    "HCC2998" "HCT116"  "HCT15"   "HCT8"
[11] "HT29"    "IS1"     "IS2"     "KM12"    "LIM1215"
[16] "LIM1863" "LIM1899" "LIM2405" "LIM2537" "LIM2550"
[21] "LIM2551" "RKO"     "RW2982"  "RW7213"  "SKCO1"
[26] "SW1116"  "SW1222"  "SW403"   "SW480"   "SW620"
[31] "SW837"   "SW948"   "T84"     "VACO5"

> sum(colnames(dataMarrC0)==colnames(dataRNAseqC0))

[1] 34

> dataRNAseqC1 <- dataRNAseqA[,common1];dim(dataRNAseqC1); colnames(dataRNAseqC1)

[1] 17408    30
 [1] "CACO2"   "COLO201" "COLO205" "COLO320" "DLD1"
 [6] "GP5D"    "HCA7"    "HCC2998" "HCT116"  "HCT15"
[11] "HT29"    "IS1"     "IS2"     "IS3"     "KM12"
[16] "LIM1215" "LIM2405" "LS513"   "RKO"     "RW2982"
[21] "RW7213"  "SKCO1"   "SW1116"  "SW403"   "SW620"
[26] "SW837"   "SW948"   "T84"     "V9P"     "VACO5"

> dataMethC1 <- numDataMethByMean[,common1];dim(dataMethC1); colnames(dataMethC1)

[1] 14476    30
 [1] "CACO2"   "COLO201" "COLO205" "COLO320" "DLD1"
 [6] "GP5D"    "HCA7"    "HCC2998" "HCT116"  "HCT15"
[11] "HT29"    "IS1"     "IS2"     "IS3"     "KM12"
[16] "LIM1215" "LIM2405" "LS513"   "RKO"     "RW2982"
[21] "RW7213"  "SKCO1"   "SW1116"  "SW403"   "SW620"
[26] "SW837"   "SW948"   "T84"     "V9P"     "VACO5"

> dataMethVarC1 <- numDataMethByVar[,common1];dim(dataMethVarC1); colnames(dataMethVarC

[1] 14476    30
 [1] "CACO2"   "COLO201" "COLO205" "COLO320" "DLD1"
 [6] "GP5D"    "HCA7"    "HCC2998" "HCT116"  "HCT15"
[11] "HT29"    "IS1"     "IS2"     "IS3"     "KM12"
[16] "LIM1215" "LIM2405" "LS513"   "RKO"     "RW2982"
[21] "RW7213"  "SKCO1"   "SW1116"  "SW403"   "SW620"
[26] "SW837"   "SW948"   "T84"     "V9P"     "VACO5"

> sum(colnames(dataMethC1)==colnames(dataRNAseqC1))

[1] 30
```

```
> sum(colnames(dataMethVarC1)==colnames(dataRNAseqC1))

[1] 30

> dataMarrC2<- expres2[,common2];dim(dataMarrC2); colnames(dataMarrC2)

[1] 19991    30
 [1] "CACO2"   "CO115"   "COLO201" "COLO205" "COLO320"
 [6] "DLD1"    "HCC2998" "HCT116"  "HCT15"   "HT29"
[11] "IS1"     "IS2"     "KM12"    "LIM1215" "LIM2405"
[16] "LOVO"    "LS174T"  "RKO"     "RW2982"  "RW7213"
[21] "SKCO1"   "SW1116"  "SW403"   "SW48"    "SW620"
[26] "SW837"   "SW948"   "T84"     "TC71"    "VACO5"

> dataMethC2 <- numDataMethByMean[,common2];dim(dataMethC2); colnames(dataMethC2)

[1] 14476    30
 [1] "CACO2"   "CO115"   "COLO201" "COLO205" "COLO320"
 [6] "DLD1"    "HCC2998" "HCT116"  "HCT15"   "HT29"
[11] "IS1"     "IS2"     "KM12"    "LIM1215" "LIM2405"
[16] "LOVO"    "LS174T"  "RKO"     "RW2982"  "RW7213"
[21] "SKCO1"   "SW1116"  "SW403"   "SW48"    "SW620"
[26] "SW837"   "SW948"   "T84"     "TC71"    "VACO5"

> dataMethVarC2 <- numDataMethByVar[,common2];dim(dataMethVarC2); colnames(dataMethVarC

[1] 14476    30
 [1] "CACO2"   "CO115"   "COLO201" "COLO205" "COLO320"
 [6] "DLD1"    "HCC2998" "HCT116"  "HCT15"   "HT29"
[11] "IS1"     "IS2"     "KM12"    "LIM1215" "LIM2405"
[16] "LOVO"    "LS174T"  "RKO"     "RW2982"  "RW7213"
[21] "SKCO1"   "SW1116"  "SW403"   "SW48"    "SW620"
[26] "SW837"   "SW948"   "T84"     "TC71"    "VACO5"

> sum(colnames(dataMethC2)==colnames(dataMarrC2))

[1] 30

> sum(colnames(dataMethVarC2)==colnames(dataMarrC2))

[1] 30
```

3.2 Matching rows (genes)

It is clear that if we wish to compute correlations between different sets of measurements done on genes we need to have a *common set of genes*. This can be done, in a similar way as with the samples, by taking common genes between two each datasets or by taking genes in common to the three of them.

If we consider "new" datasets processed by us:

```
> commongenes0 <- intersect (rownames(dataMarrC), rownames(dataRNAseqC))
> cat("Genes in common between Microarrays and RNAseq:", length(commongenes0), "\n")

Genes in common between Microarrays and RNAseq: 15564

> commongenes1 <- intersect (rownames(dataMethC1), rownames(dataRNAseqC1))
> cat("Genes in common between RNAseq and Methylation:", length(commongenes1), "\n")

Genes in common between RNAseq and Methylation: 9863

> commongenes2 <- intersect (rownames(dataMethC2), rownames(dataMarrC2))
> cat("Genes in common between Microarrays and Methylation:", length(commongenes2), "\n")

Genes in common between Microarrays and Methylation: 11369

> commongenes3 <- intersect (rownames(dataMethC),
+                            intersect(rownames(dataMarrC),rownames(dataRNAseqC)))
> cat("Genes in common between RNAseq and Microarrays and Methylation:", length(commongenes3

Genes in common between RNAseq and Microarrays and Methylation: 9677
```

There are 9677 genes in common between methylation , microarrays and RNAseq data, so these will be the basis for the search for correlation between expression (RNAseq vs microarrays) and between expression and methylation.

```
> dataMarrCR <- dataMarrC[commongenes3,];dim(dataMarrCR);

[1] 9677   25

> dataMethCR <- dataMethC[commongenes3,];dim(dataMethCR);

[1] 9677   25

> dataMethVarCR <- dataMethVarC[commongenes3,];dim(dataMethVarCR);

[1] 9677   25

> dataRNAseqCR <- dataRNAseqC[commongenes3,];dim(dataRNAseqCR);

[1] 9677   25
```

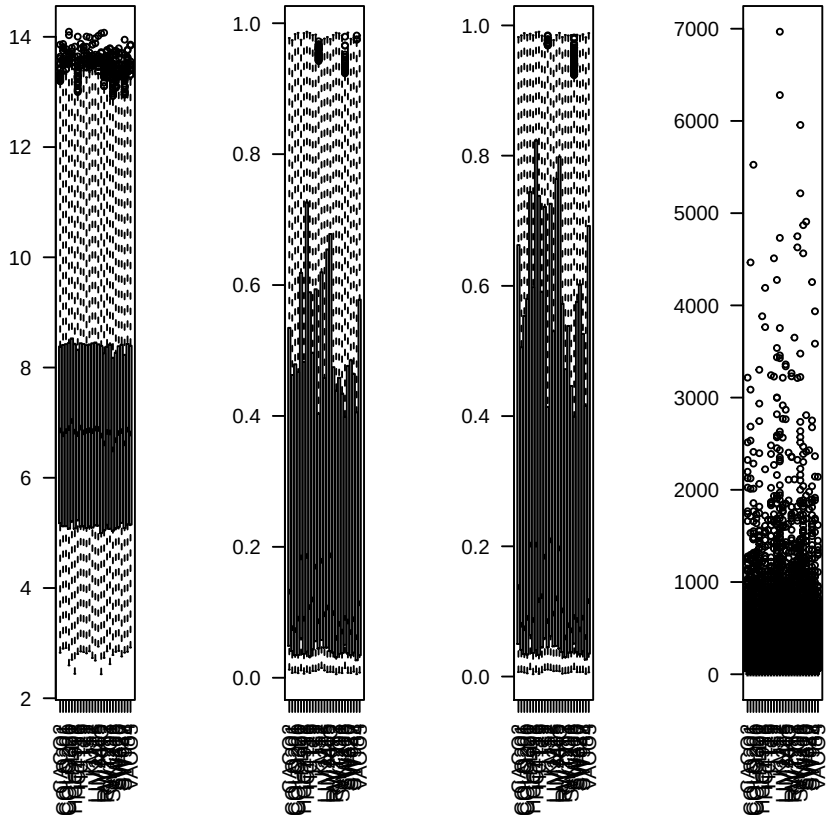
We will similarly prepare datasets with pairwise common sets of genes

```
> dataMarrCR0 <- dataMarrC0[commongenes0,];dim(dataMarrCR0);

[1] 15564    34

> dataRNAseqCR0 <- dataRNAseqC0[commongenes0,];dim(dataRNAseqCR0);

[1] 15564    34

> dataRNAseqCR1 <- dataRNAseqC1[commongenes1,];dim(dataRNAseqCR1);

[1] 9863   30

> dataMethCR1 <- dataMethC1[commongenes1,];dim(dataMethCR1);

[1] 9863   30

> dataMethVarCR1 <- dataMethVarC1[commongenes1,];dim(dataMethVarCR1);

[1] 9863   30

> dataMarrCR2 <- dataMarrC2[commongenes2,];dim(dataMarrCR2);

[1] 11369    30

> dataMethCR2 <- dataMethC2[commongenes2,];dim(dataMethCR2);

[1] 11369    30

> dataMethVarCR2 <- dataMethVarC2[commongenes2,];dim(dataMethVarCR2);

[1] 11369    30
```

3.3 Centering and scaling expression values

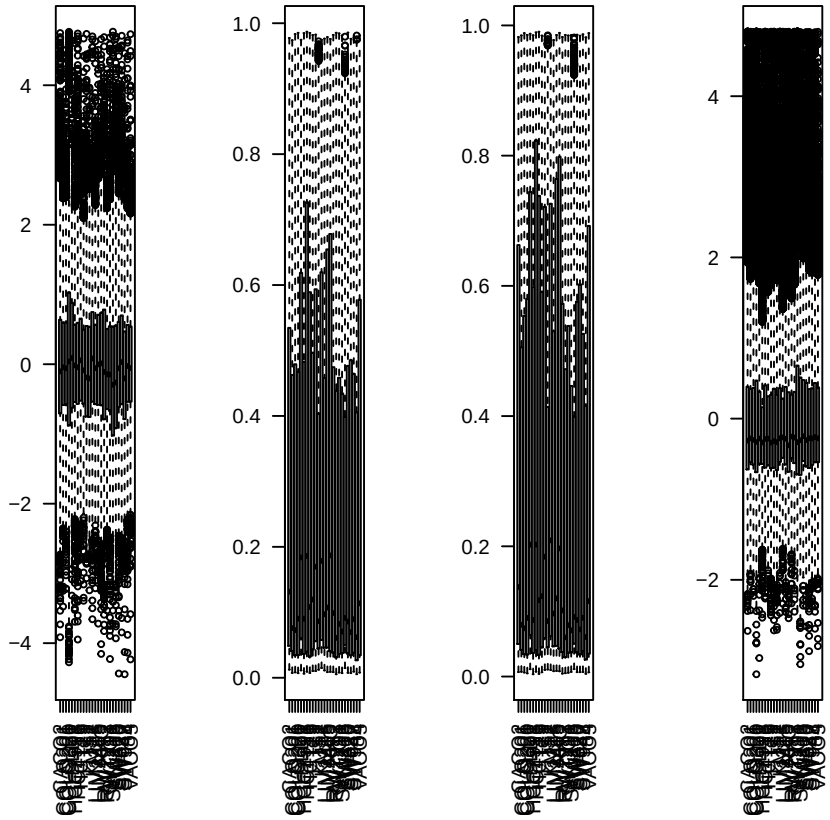
Normalized microarray and RNAseq expression values vary in a wide scale, whereas methylation values are percentages ranging between 0 and 1 and In order to facilitate joint analyses expression values will be centered and scaled and then combined into a commom data matrix again.

```
> par(mfrow=c(1,4))
> boxplot(dataMarrCR, las=2)
> boxplot(dataMethCR, las=2)
> boxplot(dataMethVarCR, las=2)
> boxplot(dataRNAseqCR, las=2)
```




```
> # Notice that scaling is applied gene-wise
> dataMarrCRCS <- t(apply(dataMarrCR,1,scale))
> colnames(dataMarrCRCS) <- colnames(dataMarrCR)
> dataRNAseqCRCS<- t(apply(dataRNAseqCR,1,scale))
> colnames(dataRNAseqCRCS) <- colnames(dataRNAseqCR)
> dataRNAseqCR0CS <- t(apply(dataRNAseqCR0,1,scale))
> colnames(dataRNAseqCR0CS) <- colnames(dataRNAseqCR0)
> dataRNAseqCR1CS  <- t(apply(dataRNAseqCR1,1,scale))
> colnames(dataRNAseqCR1CS) <- colnames(dataRNAseqCR1)
```

```
> par(mfrow=c(1,4))
> boxplot(dataMarrCRCS, las=2)
> boxplot(dataMethCR, las=2)
> boxplot(dataMethVarCR, las=2)
> boxplot(dataRNAseqCRCS, las=2)
```



This transformation sets all values in the same scale (mean=0, SD=1) which may be better for correlation analysis.

```
> cat("Microarray Data (RMA)\n")

Microarray Data (RMA)

> X<-dataMarrCR; Xcs <- dataMarrCRCS; Y <- scale(t(X))
> apply(X[1:3,], 1,function (x) {return (unlist(list(mean=mean(x), sd=sd(x))))})

      A1BG    A2M  A2ML1
mean 5.120 4.3218 3.8204
sd   0.158 0.2154 0.5447

> apply(Xcs[1:3,], 1,function (x) {return (unlist(list(mean=mean(x), sd=sd(x))))})

                      A1BG                   A2M
mean 0.0000000000000004518 -0.000000000000001568
```

```
sd   1.000000000000000000  1.000000000000000000
                      A2ML1
mean -0.000000000000002498
sd    1.000000000000000000

> apply(Y[1:3,], 1,function (x) {return (unlist(list(mean=mean(x), sd=sd(x))))})

         CACO2 COLO201 COLO205
mean -0.003146 0.03171 0.03204
sd    1.070961 0.89820 0.90548

> X<-dataRNAseqCR; Xcs <- dataRNAseqCRCS
> apply(X[1:3,], 1,function (x) {return (unlist(list(mean=mean(x), sd=sd(x))))})

        A1BG     A2M  A2ML1
mean 0.08488 0.03278 0.8814
sd   0.10984 0.04971 4.2020

> apply(Xcs[1:3,], 1,function (x) {return (unlist(list(mean=mean(x), sd=sd(x))))})

                      A1BG                       A2M
mean 0.0000000000000002938 -0.0000000000000000004172
sd   0.9999999999999977796  1.0000000000000000000000
                     A2ML1
mean 0.0000000000000000612
sd   0.9999999999999977796
```

3.4 Storing intermediate values

```
> showV <- function(x){
+    cat (substitute(x),": ",dim(x),"\n")
+ }
```

The result of the preprocessing steps done has outputted the following data tables:

1. Data matched at gene and sample level between the three data types: RNA-seq –unscaled or scaled row wise– Microarrays and Methylation – aggregated by mean or by variance.

```
> showV(dataRNAseqCR); showV(dataRNAseqCRCS); showV(dataMarrCR);  showV(dataMethCR); sh

dataRNAseqCR :  9677 25
dataRNAseqCRCS :  9677 25
dataMarrCR :  9677 25
```

```
dataMethCR :  9677 25
dataMethVarCR :  9677 25
```

2. Data matched at gene and sample level between every two data types: microarrays and RNA-seq, methylation and microarrays and methylation and RNA-seq.

```
> showV(dataMarrCR0); showV(dataRNAseqCR0)

dataMarrCR0 :  15564 34
dataRNAseqCR0 :  15564 34

> showV(dataRNAseqCR1) ; showV(dataMethCR1); showV(dataMethVarCR1)

dataRNAseqCR1 :  9863 30
dataMethCR1 :  9863 30
dataMethVarCR1 :  9863 30

> showV(dataMarrCR2); showV(dataMethCR2);showV(dataMethVarCR2);

dataMarrCR2 :  11369 30
dataMethCR2 :  11369 30
dataMethVarCR2 :  11369 30
```

These data will be used for correlation analyses to be performed in the next sections so they can be saved for further analyses.

```
> save(dataRNAseqCR, dataRNAseqCRCS, dataMarrCR,  dataMarrCR, dataMethCR, dataMethVarCR,
+      file=file.path(combinedDir, "matchedMarrRNAseqMethData.Rda"))
> save(dataMarrCR0, dataRNAseqCR0, dataRNAseqCR0CS,
+      dataRNAseqCR1,  dataRNAseqCR1CS, dataMethCR1, dataMethVarCR1,
+      dataMarrCR2,  dataMethCR2, dataMethVarCR2,
+      file=file.path(combinedDir, "matchedMarrRNAseqMethDataByPAIRS.Rda"))
```