Correlation analysis between Expression (microarrays & RNA-seq) and methylation data in a set of cancer colon cell lines data (2) Computing correlations between datasets

Alex Sánchez-Pla. Statistics department. UB
Statistics and Bioinformatics Unit. VHIR

January 15, 2016

Contents


```
> setwd(workingDir)
> #Sys.setenv(TEXINPUTS=getwd(),
> #           BIBINPUTS=getwd(),
> #           BSTINPUTS=getwd())
```

1 Introduction

```
> load(file=file.path(combinedDir, "matchedMarrRNAseqMethData.Rda"))
> load(file=file.path(combinedDir, "matchedMarrRNAseqMethDataByPAIRS.Rda"))
```

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

In a previous step data have been prepared so that genes and samples match between those datasets that we wish to study.

Once we have matched the data we can proceed to compute correlations between datasets either to check consistency between datasets (microarrays and RNAseq data, methylation aggregated by means or by vars) or between either of these and methylation values.

1.1 Data for correlation analysis

The data that have been prepared are the following:

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)

dataRNAseqCR :  9677 25

> showV(dataRNAseqCRCS)

dataRNAseqCRCS :  9677 25

> showV(dataMarrCR)

dataMarrCR :  9677 25

> showV(dataMethCR)

dataMethCR :  9677 25

> showV(dataMethVarCR)

dataMethVarCR :  9677 25

> # which(rownames(dataMarrCR)=="ZBTB18")
> # which(rownames(dataMethCR)=="ZBTB18")
> # which(rownames(dataRNAseqCR)=="ZBTB18")
```

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
```

2 Correlation Analysis

Once we have the matched datasets there are two main questions to investigate:

- How consistent are related datasets that is: how good is the correlation between methylation values aggregated differently or how good is the correlation between RNAseq and Microarray values?[1]
- How good is the correlation between each expression value (RNAseq or microarrays) and methylation?

This suggests a series of possible correlation analyses.

1. Correlation analysis between "possibly redundant" datasets (or should we study their concordance?).
   (a) Methylation aggregated by averaging CPG sites vs Methylation aggregated by taking the most variable CPG sites
   (b) RNA-seq and microarrays.
2. Correlation analysis for data matched at gene and sample level between **the three data types**: methylation, microarrays and RNA-seq.

[1] RNAseq and RNAseq centered and scaled are identical in terms of computing correlations, so RNA-seq centered and scale will be omitted from the analyses

(a) RNA-seq and Microarrays.

(b) RNA-seq and Methylation.

(c) Microarrays and Methylation.

3. Correlation analysis for data matched at gene and sample level between **every two data types**: microarrays and RNA-seq, methylation and microarrays and methylation and RNA-seq.

   (a) RNA-seq and Microarrays.

   (b) Microarrays and Methylation.

   (c) RNA-seq and Methylation.

Depending of the results of these analyses redundant data will be removed to avoid results that are also the same except for differences attributable to noise.

2.1 Analysis methods

For the correlation analysis between genes we have used several standard association measures, such as Pearson and Spearman Correlation Coefficients. For each of them We have computed p-values for the significance test $H_0 : \rho = 0$ and we have adjusted p-values using the Banjamini-Hochberg mehod.

We have also computed a newer measure known as "Distance Correlation" [**?**] that uses the distances between observations as part of its calculation. Distance correlation varies between 0 and 1 and it can be computed on a gene per gene basis ("univariately") or between matrices, in which case it becomes a multivariate coefficient.

Last we have computed a purely multivariate correlation coefficent, the "RV coefficient" ([**?**, **?**]), which, similarly to the matrix Distance correlation can be seen as a measure of "global" similarity between the datasets. The closer to 1, in the scale 0-1 the greater the correlation between the two datasets.

```
> source(file.path(codeDir, "correlationFunctions.R"))
```

3 Data Analysis

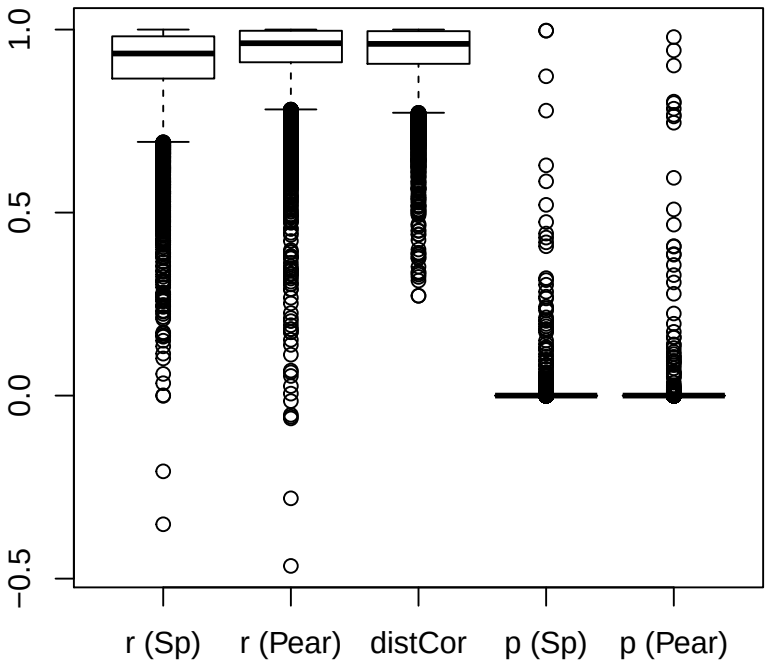
3.1 Relation between "possibly redundant" datasets

We have prepared several variations of the same dataset because it is not clear which is the most appropriate for the analysis. If the values are very very similar we can omit one of them.

3.1.1 Relation between methylation values aggregated differently

```
> MethCorr1 <- matAllCorrs (dataMethCR, dataMethVarCR);
> head(MethCorr1)

      r (Sp) r (Pear) distCor p (Sp) p (Pear) adj.Spear.Pval adj.Pear.Pval
A1BG  1.0000   1.0000  1.0000      0        0              0             0
A2M   1.0000   1.0000  1.0000      0        0              0             0
AAAS  0.9754   0.9773  0.9814      0        0              0             0
AADAC 1.0000   1.0000  1.0000      0        0              0             0
AAK1  1.0000   1.0000  1.0000      0        0              0             0
AANAT 0.9992   0.9987  0.9982      0        0              0             0

> tail(MethCorr1)

           r (Sp) r (Pear) distCor p (Sp)          p (Pear) adj.Spear.Pval    adj.Pear.Pval
SARM1   0.1146154 0.779185  0.7957 0.5854 0.000004445704285         0.5857 0.00000463939182
PDCD5   0.1015385 0.068352  0.3307 0.6291 0.745452074153160         0.6294 0.74606885113043
CXCL16  0.0592308 0.005415  0.2729 0.7785 0.979505650670827         0.7788 0.97950565067083
TRIM45  0.0338462 0.478627  0.4688 0.8724 0.015504595476246         0.8726 0.01559322078816
PDLIM1  0.0007692 0.988610  0.9800 0.9971 0.000000000000000         0.9971 0.00000000000000
TMCO1  -0.0007692 0.944772  0.9316 0.9971 0.000000000001247         0.9971 0.00000000000208
```

```
> boxplot(as.data.frame(MethCorr1[,1:5]), main ="Correlations between methylations aggregate
```




```
> multivCorr(dataMethCR, dataMethVarCR)
```

It can be seen that the correlation between the aggregated values is almost perfect in both univariate and multivariate measures**so the analyses will be performed based on methylation data aggregated with the mean**.

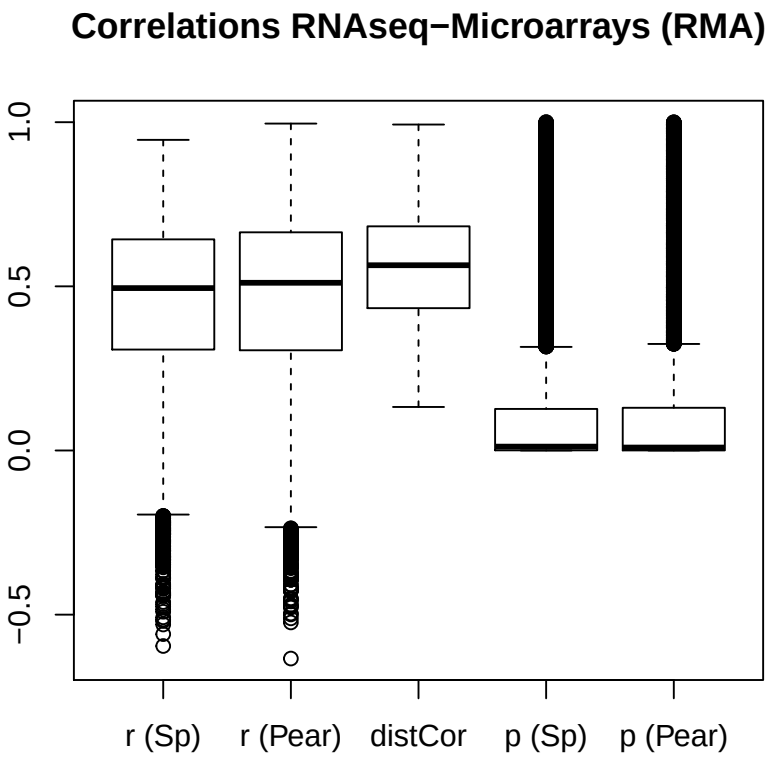
3.2 Relation between different expression measures

Microarrays and RNAseq are both used to quantify gene expression. In this case we expect that they yield similar values but there is no warranty that this is the case because of the different technological approaches used.

```
> MarrRNaseqCorr1 <- matAllCorrs (dataMarrCR, dataRNAseqCR)
> head(MarrRNaseqCorr1)

        r (Sp) r (Pear) distCor            p (Sp)         p (Pear) adj.Spear.Pval
```

```
CLDN1    0.9462   0.8506  0.8832 0.0000000000009379 0.00000007251256 0.000000009076
WWTR1    0.9392   0.7574  0.8504 0.0000000000036424 0.00001168876427 0.000000017624
TRAF3IP2 0.9200   0.8131  0.8500 0.0000000000781566 0.00000078041374 0.000000172492
PIWIL1   0.9192   0.9520  0.9447 0.0000000000869138 0.00000000000026 0.000000172492
ITPR3    0.9169   0.8478  0.8617 0.0000000001187861 0.00000008857403 0.000000172492
SLC26A11 0.9169   0.8021  0.8476 0.0000000001187861 0.00000142085839 0.000000172492
            adj.Pear.Pval
CLDN1    0.00000208220785
WWTR1    0.00010522062500
TRAF3IP2 0.00001304328795
PIWIL1   0.00000000004839
ITPR3    0.00000243503088
SLC26A11 0.00002079642568
```

```
> boxplot(as.data.frame(MarrRNaseqCorr1[,1:5]), main ="Correlations RNAseq-Microarrays (RMA)
```



We can make a global assessment of correlation using the `Distcor` correlation distance and the `RV` coefficient.

```
> multivCorr(dataMarrCR, dataRNAseqCR)
```

It can be seen that against what one would expect both distance correlation and the RV coefficient the correlation between microarrays and RNAseq is not very high.

Simple Correlation values are written into a text file.

```
> write.csv2(MarrRNaseqCorr1, file=file.path(resultsDir, "Correlations-MArr-RNAseq.csv"))
```

3.3 Relation between Expression and Methylation values

3.3.1 RNAseq vs methylation. 3 way matched

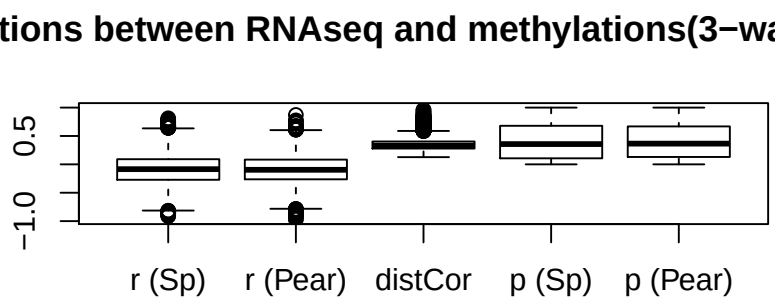
```
> dataMethCRM<- dataMethCR
> means <- apply(dataMethCRM, 1, mean)
> withMissings <- dataMethCRM[which(is.na(means)),]
> head(withMissings)

     CACO2 COLO201 COLO205 COLO320 DLD1 HCC2998 HCT116 HCT15 HT29 IS1 IS2 KM12 LIM1215 LIM24
     RW2982 RW7213 SKCO1 SW1116 SW403 SW620 SW837 SW948 T84 VACO5

> RNAseqMethCorr <- matAllCorrs (dataMethCR, dataRNAseqCR)
> head(RNAseqMethCorr)

        r (Sp) r (Pear) distCor          p (Sp)     p (Pear) adj.Spear.Pval adj.Pear.Pval
PYCARD -0.9138  -0.6961  0.7715 0.0000000001777 0.0001112136    0.000001720     0.0135066
GPX2   -0.8938  -0.8276  0.8332 0.0000000017723 0.0000003323    0.000007236     0.0001786
CTSS   -0.8915  -0.7003  0.7289 0.0000000022433 0.0000971352    0.000007236     0.0120510
SGK2   -0.8646  -0.7153  0.8114 0.0000000250416 0.0000585012    0.000060582     0.0088456
DAPP1  -0.8566  -0.8224  0.8537 0.0000000463929 0.0000004564    0.000082238     0.0002325
ANXA4  -0.8554  -0.6739  0.7497 0.0000000509899 0.0002213555    0.000082238     0.0216369

> sum(RNAseqMethCorr[,"adj.Spear.Pval"] < 0.05)

[1] 411

> write.csv2(RNAseqMethCorr, file=file.path(resultsDir, "Correlations-RNAseq-Meth (3-way-Mat
```

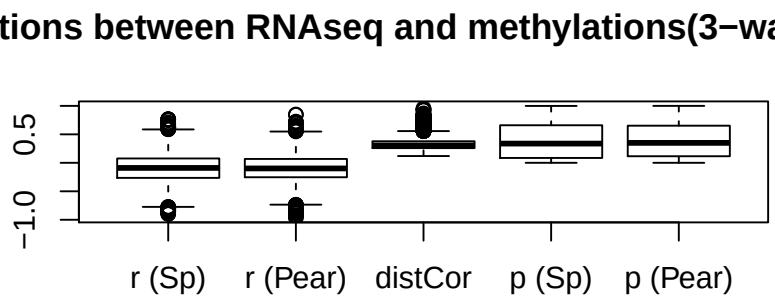
```
> multivCorr(dataMethCR, dataRNAseqCR)
```

3.3.2 RNAseq vs methylation. 2 way matched

```
> RNAseqMethCorr1 <- matAllCorrs (dataMethCR1, dataRNAseqCR1)
> head(RNAseqMethCorr1)

        r (Sp) r (Pear) distCor           p (Sp)      p (Pear) adj.Spear.Pval adj.Pear.Pval
PYCARD -0.8999  -0.6828  0.7456 0.00000000001336 0.00003225877   0.0000001178    0.00338477
GPX2   -0.8954  -0.7891  0.7946 0.00000000002388 0.00000021972   0.0000001178    0.00009030
CTSS   -0.8643  -0.6940  0.7360 0.00000000075274 0.00002108479   0.0000024748    0.00263240
DAPP1  -0.8579  -0.8110  0.8434 0.00000000137512 0.00000005493   0.0000032763    0.00002851
SGK2   -0.8558  -0.6860  0.8044 0.00000000166091 0.00002855236   0.0000032763    0.00320014
ANXA4  -0.8523  -0.6600  0.7449 0.00000000228365 0.00007250839   0.0000037539    0.00644280

> sum(RNAseqMethCorr1[,"adj.Spear.Pval"] < 0.05)

[1] 677

> write.csv2(RNAseqMethCorr1, file=file.path(resultsDir, "Correlations-RNAseq-Meth (2-way-Ma
```

```
> opt<-par(mfrow=c(2,1))
> boxplot(as.data.frame(RNAseqMethCorr[,1:5]), main ="Correlations between RNAseq and methyl
> boxplot(as.data.frame(RNAseqMethCorr1[,1:5]), main ="Correlations between RNAseq and methy
```






```
> par(opt)
```

```
> multivCorr(dataMethCR1, dataRNAseqCR1)
```

3.3.3 Microarray (RMA) vs methylation. 3 way matched

```
> MarrMethCorr <- matAllCorrs (as.matrix(dataMethCR), dataMarrCR)
> head(MarrMethCorr)

        r (Sp) r (Pear) distCor         p (Sp)       p (Pear) adj.Spear.Pval adj.Pear.Pval
EVPL   -0.8969  -0.8982  0.8845 0.000000001284 0.000000001119     0.00001242  0.0000009842
BST2   -0.8692  -0.8761  0.8730 0.000000017207 0.000000009594     0.00005696  0.0000054611
RIPK3  -0.8669  -0.8203  0.8352 0.000000020794 0.000000517899     0.00005696  0.0000864088
```

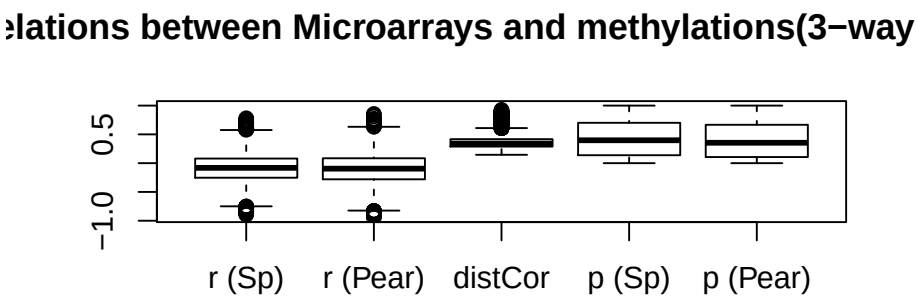
```
PYCARD -0.8654  -0.8908  0.8759 0.000000023546 0.000000002424     0.00005696  0.0000019548
PVRL4  -0.8454  -0.7395  0.7617 0.000000104474 0.000023986733     0.00020220  0.0015790450
FADS1  -0.8308  -0.8153  0.7983 0.000000273636 0.000000691586     0.00044133  0.0001079433

> sum(MarrMethCorr[,"adj.Spear.Pval"] < 0.05)

[1] 232

> write.csv2(MarrMethCorr, file=file.path(resultsDir, "Correlations-Microarrays(RMA)-Meth (3
```

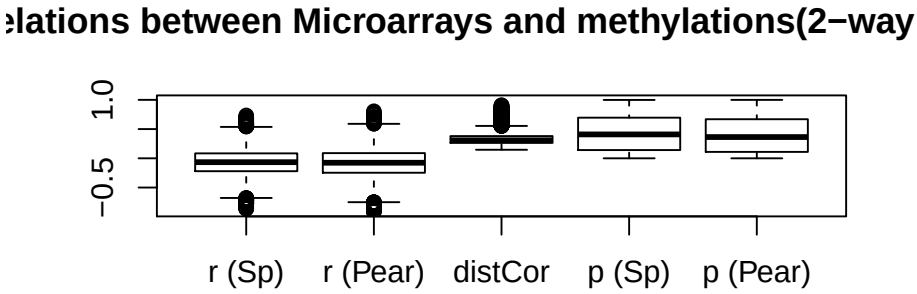
```
> multivCorr(as.matrix(dataMethCR), dataMarrCR)
```

3.3.4 Microarrays vs methylation. 2 way matched

```
> MarrMethCorr1 <- matAllCorrs (dataMethCR2, dataMarrCR2)
> head(MarrMethCorr1)

        r (Sp) r (Pear) distCor          p (Sp)         p (Pear) adj.Spear.Pval adj.Pear.Pva
BST2   -0.8652  -0.8930  0.8945 0.0000000006905 0.00000000003242    0.000005805 0.0000000737
RIPK3  -0.8581  -0.8040  0.8361 0.0000000013553 0.00000008724365    0.000005805 0.0000198374
EVPL   -0.8567  -0.8813  0.8587 0.0000000015318 0.00000000012898    0.000005805 0.0000001659
PYCARD -0.8469  -0.8692  0.8474 0.0000000036262 0.00000000046556    0.000009582 0.0000004410
ZNF420 -0.8452  -0.8051  0.8332 0.0000000042142 0.00000008095809    0.000009582 0.0000191655
TNS4   -0.8287  -0.8230  0.8253 0.0000000155721 0.00000002366899    0.000029506 0.0000072727

> sum(MarrMethCorr1[,"adj.Spear.Pval"] < 0.05)

[1] 279

> write.csv2(MarrMethCorr1, file=file.path(resultsDir, "Correlations-Microarrays(RMA)-Meth (
```

```
> opt<-par(mfrow=c(2,1))
> boxplot(as.data.frame(MarrMethCorr[,1:5]), main ="Correlations between Microarrays and met
> boxplot(as.data.frame(MarrMethCorr1[,1:5]), main ="Correlations between Microarrays and me
```






```
> par(opt)
```

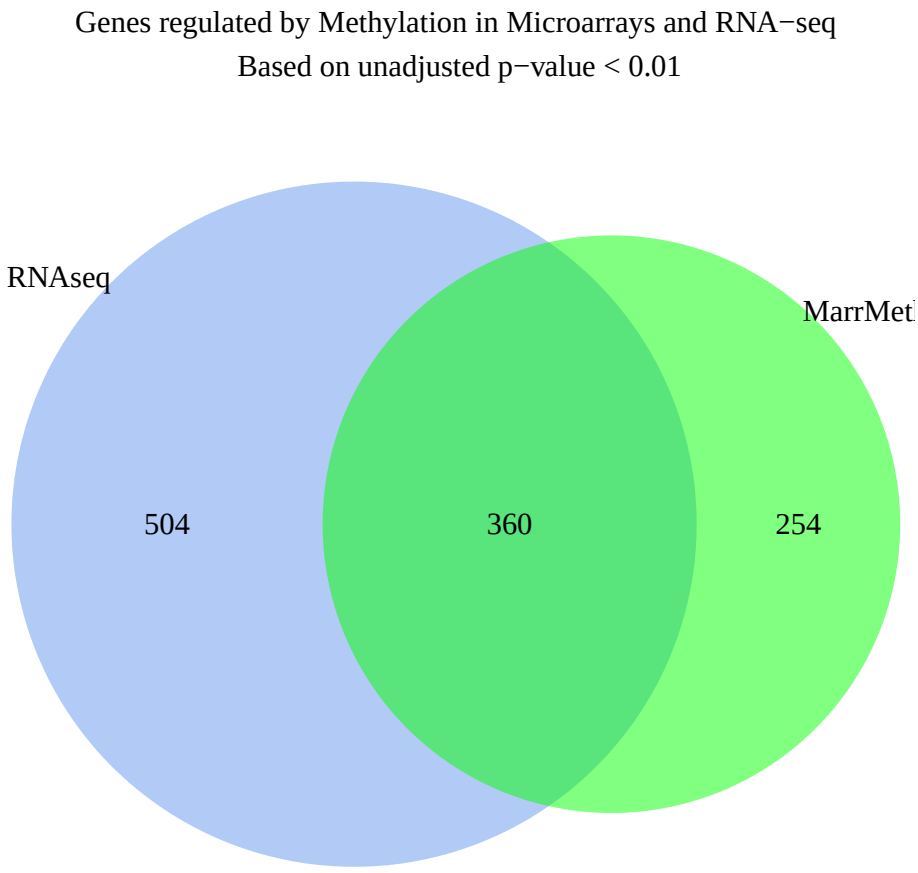
```
> multivCorr(dataMethCR1, dataRNAseqCR1)
```

3.4 Relating Microarray–RNAseq–Methylation correlations

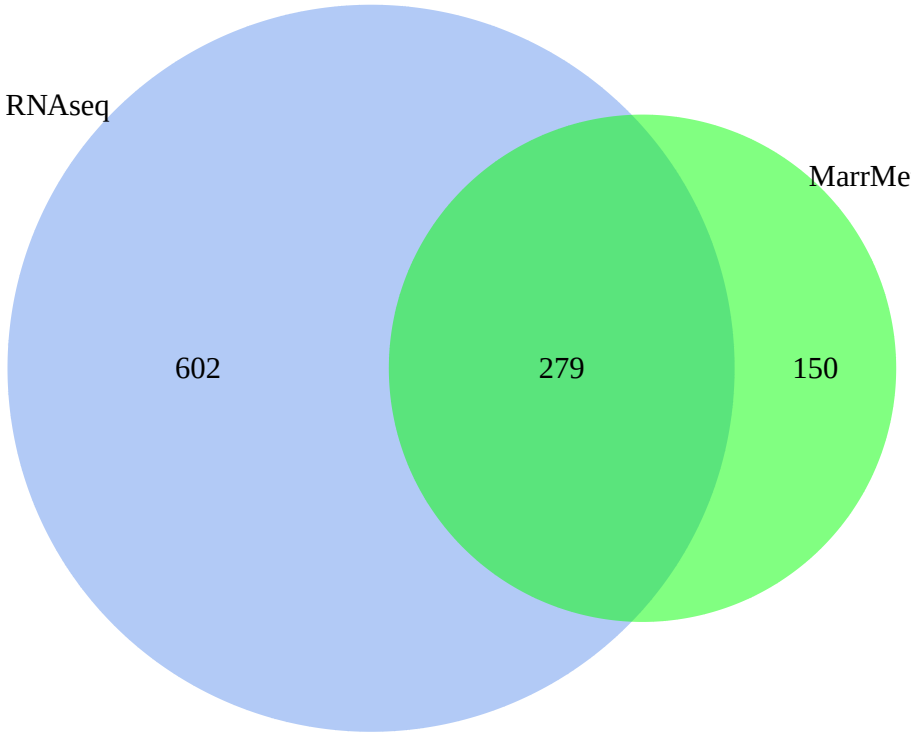
A way to decide which genes can be called "regulated by methylation" consists of selecting those genes that are significantly and negatively correlated with methylation, that is selecting those genes with a negative correlation coefficient and a p-value of less than 0.05. Depending on if we rely on adjusted or unadjusetd p-values the number of candidate genes will change.

```
> cond01 <- (RNAseqMethCorr1[,"r (Sp)"]< 0) & (RNAseqMethCorr1[,"p (Sp)"] < 0.01)
> sigRNAseqMeth0 <- RNAseqMethCorr1 [cond01, ]
> sigRNAseqMethGenes0<-rownames(sigRNAseqMeth0)
> length(sigRNAseqMethGenes0)

[1] 864

> cond02 <-(RNAseqMethCorr1[,"r (Sp)"]< 0) & (RNAseqMethCorr1[,"adj.Spear.Pval"] < 0.10)
> sigRNAseqMeth <- RNAseqMethCorr1 [cond02,]
> sigRNAseqMethGenes<-rownames(sigRNAseqMeth)
> length(sigRNAseqMethGenes)

[1] 881

> cond11 <- (MarrMethCorr1[,"r (Sp)"] < 0) & MarrMethCorr1[,"p (Sp)"] < 0.01
> sigMarrMeth0 <- MarrMethCorr1 [cond11,]
> sigMarrMethGenes0 <-rownames(sigMarrMeth0)
> length(sigMarrMethGenes0)

[1] 614

> cond12 <- (MarrMethCorr1[,"r (Sp)"] < 0) & (MarrMethCorr1[,"adj.Spear.Pval"] < 0.10)
> sigMarrMeth <- MarrMethCorr1 [cond12,]
> sigMarrMethGenes<-rownames(sigMarrMeth)
> length(sigMarrMethGenes)

[1] 429
```

```
> require(VennDiagram)
> vd01<- venn.diagram(list(RNAseq=sigRNAseqMethGenes0, MarrMeth=sigMarrMethGenes0),filename=
> grid.draw(vd01)
```




```
> vd02<- venn.diagram(list(RNAseq=sigRNAseqMethGenes, MarrMeth=sigMarrMethGenes),
+                     filename=NULL, col = "transparent",
+                     fill = c("cornflowerblue", "green"),
+                     main ="Genes regulated by Methylation in Microarrays and RNA-seq\n Based
> grid.draw(vd02)
```

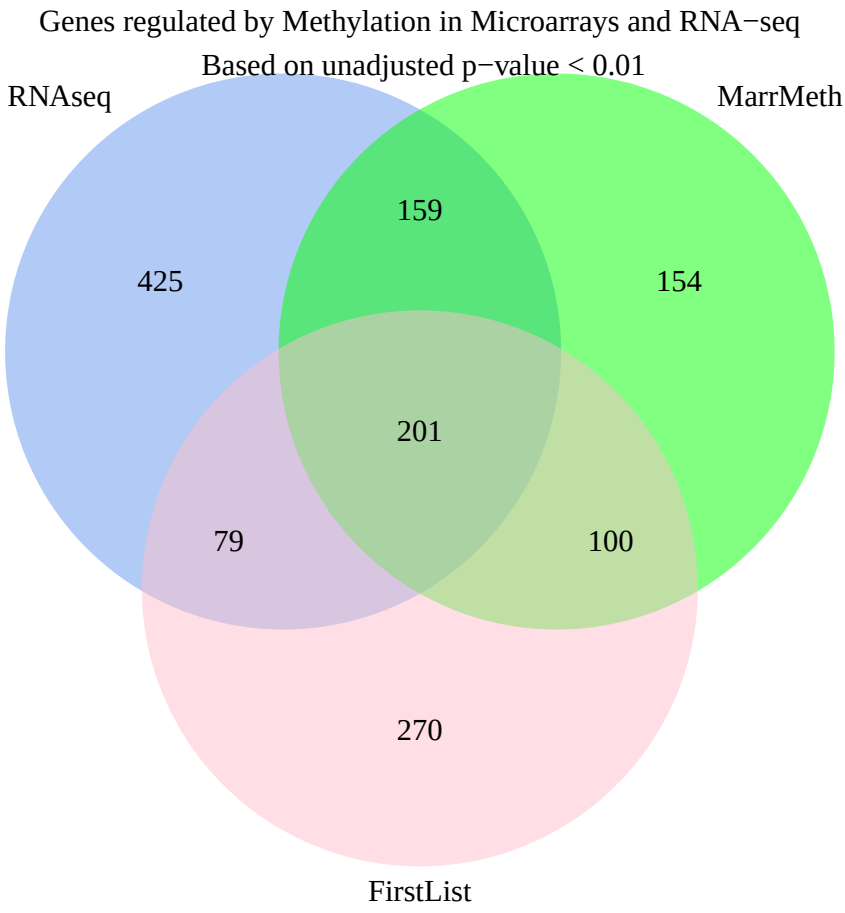


A previous analysis had been performed and a first list had been obtained

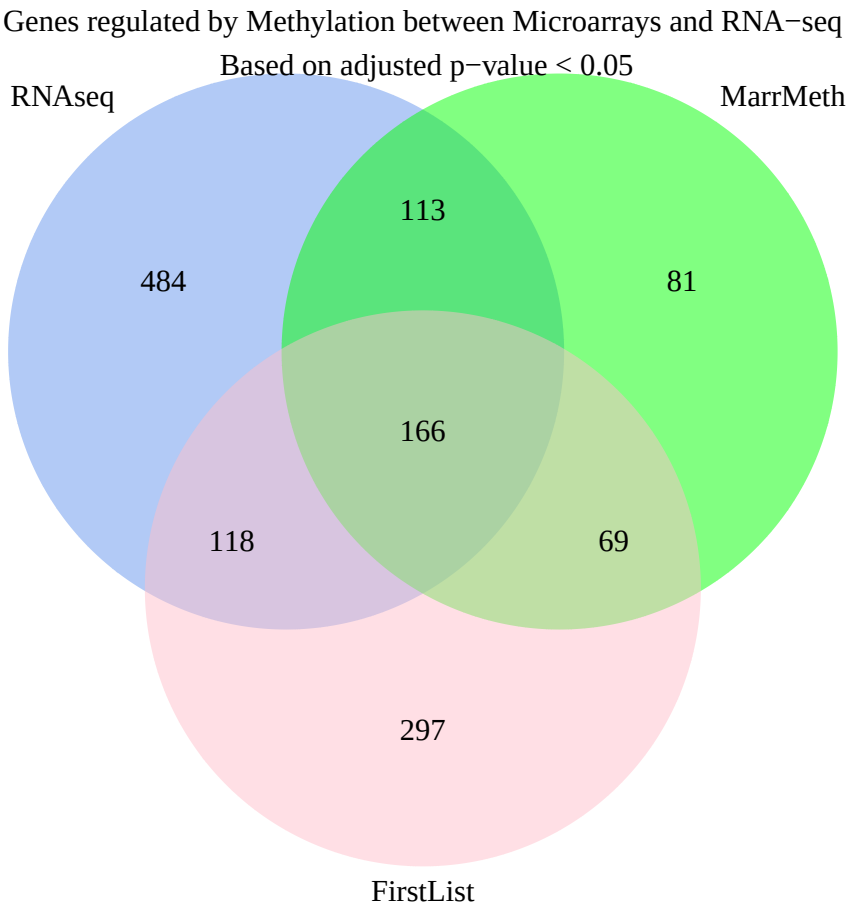
```
> firstList <- read.csv(file.path(combinedDir,"GRbyM-FirstList.txt"), header=TRUE)
> #firstList <- read.xls(file.path(combinedDir,"GRbyM-FirstList.xls"), sheet=1, header=TRUE,
> firstList<- as.character (firstList[,1])
> firstGenes <-toupper(firstList)
```

Common genes between these and the new lists can be seen as follows:

```
> require(VennDiagram)
> vd11<- venn.diagram(list(RNAseq=sigRNAseqMethGenes0,
+                          MarrMeth=sigMarrMethGenes0, FirstList=firstGenes),
+                     filename=NULL, col = "transparent",
+                     fill = c("cornflowerblue", "green", "pink"),
+                     main ="Genes regulated by Methylation in Microarrays and RNA-seq\n Bas
> grid.draw(vd11)
```




```
> vd12<- venn.diagram(list(RNAseq=sigRNAseqMethGenes, MarrMeth=sigMarrMethGenes, FirstList=f
+                     filename=NULL, col = "transparent",
+                     fill = c("cornflowerblue", "green", "pink"),
+                     main ="Genes regulated by Methylation between Microarrays and RNA-seq\
> grid.draw(vd12)
```


Genes regulated by Methylation between Microarrays and RNA−seq
Based on adjusted p−value < 0.05
RNAseq
MarrMeth
FirstList
484
113
81
166
118
69
297