Working notes 

introduction to the R package `mgmtstp27`
(document in preparation!)

BADY P.

11 mai 2015

Résumé

This document contains the description and the use of some functions propsoed in the `R` packages `mgmtstp27`. Additionally, it provides information related to the effect of normalization, bacth, etc ... of HM-450K Infinium platform on the prediction of the DNA methylation status of the MGMT promoter (Bady et al. 2012).

Table des matières

1 Motivations

This document present a set of function used to predict the DNA methylation of the MGMT promoter from the model based on Infinium HM-450K platforms (DNA methylation) proposed in [2]. This model is usable with the Infinium HM-27K platform.

2 Data

Two datasets are used to illustrate the package `mgmtstp27`. The first dataset come from TCGA project (The Cancer Genome Atlas Research Network 2008, http ://cancergenome.nih.gov/) where the DNA methylation was evaluated by platform Infinium HM-450K and HM-27K. The second dataset was used as training dataset (M-GBM) in [2].The data come from "raw" normalisation procedure orresponding to the method initially used to preprocess the data from HM-27K platform.The function `preprocessRaw` from `R` package `minfi` can be used to perform this preprocessing means converting the Red and Green channel into unmethylated and methylated signal.
The two datasets are available in the package mgmtstp27 and they can be loaded as follow :

```
require(mgmtstp27)
data(NCHgbm450)
colnames(NCHgbm450)
```

```
 [1] "Code"              "Age"               "Sex"               "OS"
 [5] "Status"            "PrGBM"             "TMZ_RT"            "NTB"
 [9] "PatientID"         "MGMTmsp"           "IDH1status"        "CIMP"
[13] "ExpressionSubtype" "Trial"             "STP27link"         "STP27response"
[17] "STP27class"        "cg00618725"        "cg01341123"        "cg02022136"
[21] "cg02330106"        "cg02802904"        "cg02941816"        "cg05068430"
[25] "cg12434587"        "cg12575438"        "cg12981137"        "cg14194875"
[29] "cg16215402"        "cg18026026"        "cg19706602"        "cg23998405"
[33] "cg25946389"        "cg26201213"        "cg26950715"
```

```
data(TCGAgbm27)
colnames(TCGAgbm27)
```

```
[1] "bcr_patient_barcode" "STP27response"       "STP27class"
[4] "cg12434587"          "cg12981137"
```

3 Probability that MGMT pomoter is methylated

The function `MGMTpredict` provides prediction of DNA methylation status of MGMT promoter as described in [2]. The model and data are contains in an in-

ternal object `glm` called `MGMTSTP27`. An additional numerical vector called `perf` containing performance information and optimal cut-off (see [2]) was associated with this object.The model is described below :

```
mgmtstp27::MGMTSTP27
```

```
Call:  glm(formula = y ~ cg12434587 + cg12981137, family = binomial,
    data = tmp)

Coefficients:
(Intercept)   cg12434587   cg12981137
     4.3215       0.5271       0.9265

Degrees of Freedom: 67 Total (i.e. Null);  65 Residual
Null Deviance:              94.03
Residual Deviance: 30.14        AIC: 36.14
```

```
names(mgmtstp27::MGMTSTP27)
```

```
 [1] "coefficients"      "residuals"         "fitted.values"     "effects"
 [5] "R"                 "rank"              "qr"                "family"
 [9] "linear.predictors" "deviance"          "aic"               "null.deviance"
[13] "iter"              "weights"           "prior.weights"     "df.residual"
[17] "df.null"           "y"                 "converged"         "boundary"
[21] "model"             "call"              "formula"           "terms"
[25] "data"              "offset"            "control"           "method"
[29] "contrasts"         "xlevels"           "anova"             "perf1"
[33] "perf2"
```

```
summary(mgmtstp27::MGMTSTP27)
```

```
Call:
glm(formula = y ~ cg12434587 + cg12981137, family = binomial,
    data = tmp)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0674  -0.2682  -0.1469   0.2098   2.2753

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.3215     1.2200   3.542 0.000397
cg12434587    0.5271     0.3021   1.745 0.080988
cg12981137    0.9265     0.3018   3.069 0.002145

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 94.033  on 67  degrees of freedom
Residual deviance: 30.143  on 65  degrees of freedom
AIC: 36.143

Number of Fisher Scoring iterations: 6
```

```
mgmtstp27::MGMTSTP27$perf1
```

```
        cut    sens      spec       pvp      pvn      prev
1 0.3582476 0.96875 0.8888889 0.8857143 0.969697 0.4705882
```

The prediction can be simply obtained as follow :

```
prednewnch <- MGMTpredict(NCHgbm450)
prednewtcga <- MGMTpredict(TCGAgbm27)
```

4 DNA methylation state of MGMT promoter

To validate the prediction computed by the package `mgmtstp27`, we compare the results from [2] and the output from our function `MGMTpredict`. The predicted DNA methylated states of the MGMT promoter are exactly the same for the two datasets (training and TCGA datasets, see below).

```
table(prednewtcga$state,TCGAgbm27$STP27class,useNA="always")
```

```
        M   U <NA>
 M    120   0    0
 U      0 121    0
 <NA>   0   0    0
```

```
table(prednewnch$state,NCHgbm450$STP27class,useNA="always")
```

```
       M  U <NA>
 M    35  0    0
 U     0 33    0
 <NA>  0  0    0
```

The two following figures confirm that the outputs (probabilities) from the function `MGMTpredict` correspond exactly to the values from [2].

```
plot(prednewnch$pred,NCHgbm450$STP27response,xlab="proba from MGMTpredict",
     ylab="proba from table S3",panel.first=c(grid()),pch=19)
abline(0,1,col="red",lwd=2)
```

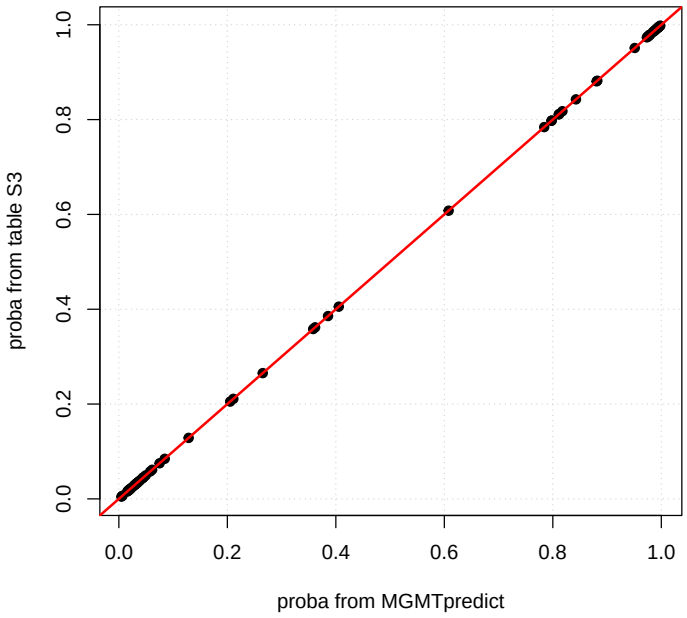

FIGURE 1 – Comparison of the prediction from the table S3 ([2]) and the outputs from the function `MGMTpredict`.

.

```
plot(prednewtcga$pred,TCGAgbm27$STP27response,xlab="proba from MGMTpredict",
     ylab="proba from table S3",panel.first=c(grid()),pch=19)
abline(0,1,col="red",lwd=2)
```

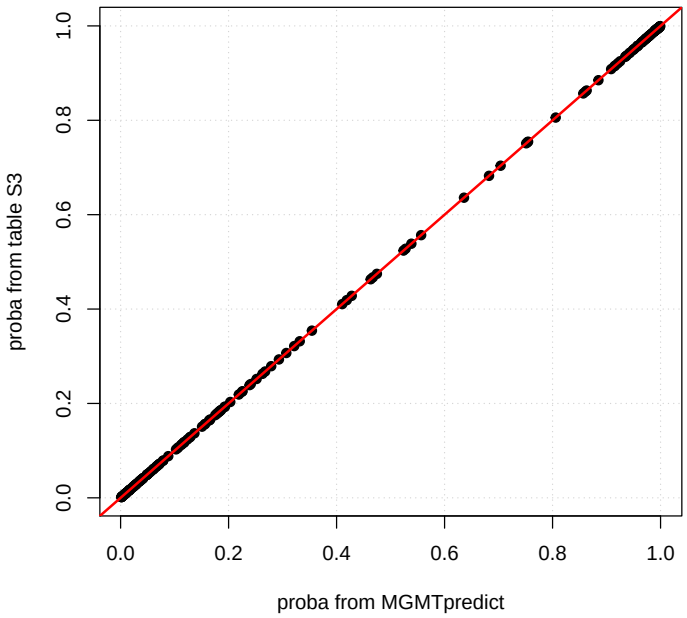

FIGURE 2 – Comparison of the prediction from the table S5 ([2]) and the outputs from the function `MGMTpredict`.

.

5 Quality control for dataset prediction

The graphical tools proposed in this section postulate that the new population is comparable to the training datasets (giloma grade IV populations). Consequently, it could be not necessary relevant to use them to investigate the quality of prediction for non-GBM populations. For NCH population, we obtained the exact results of ([2]).

```
MGMTqc.pop(prednewnch)
```

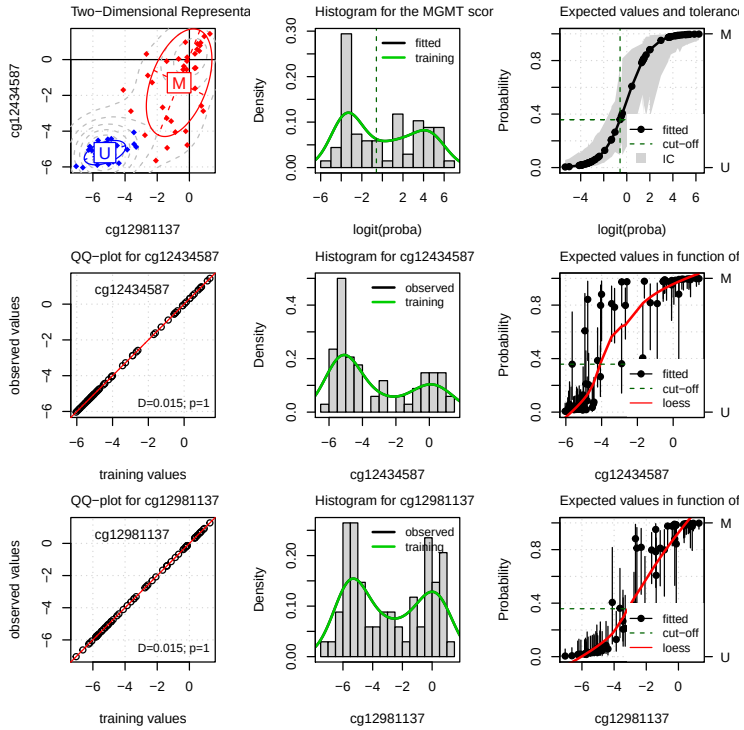

FIGURE 3 – Graphical quality control for prediction from NCH datasets

.

In TCGA population, We observe that the M-values distribution of cg12434587 and cg12981137 are comparable to the training distributions (see below).

```
MGMTqc.pop(prednewtcga)
```

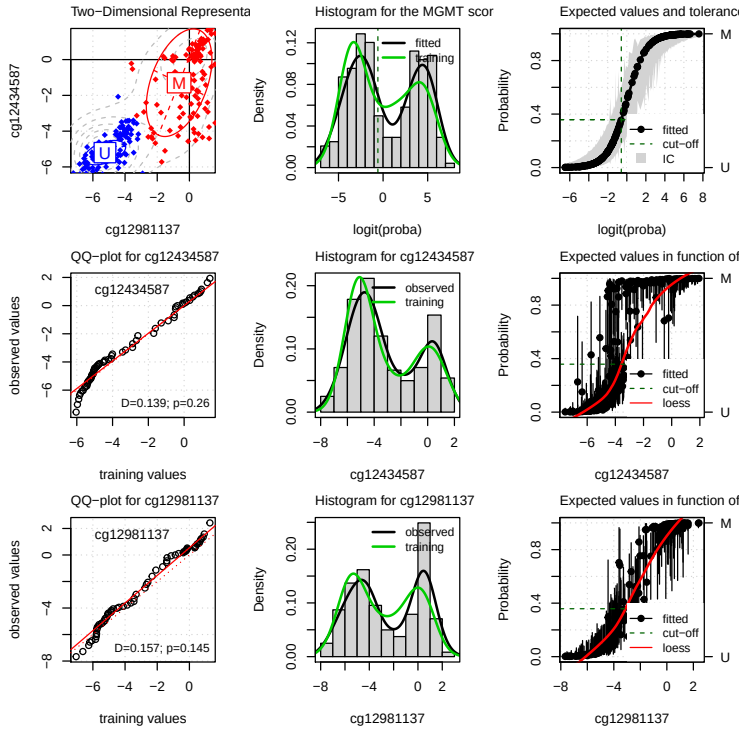

FIGURE 4 – Graphical quality control for population prediction from TCGA datasets

.

6 Quality control for single sample prediction

For the single sample prediction the graphical control quality is simplified and We only consider the Dimension plot, the representation of the prediction with its tolerance interval and the density plot of the MGMT score (logit-transformed probability) and probe distribution. the code is given below :

```
pred1 <- MGMTpredict(NCHgbm450[1,])
pred1
```

```
     sample cg12434587 cg12981137      pred      lower    upper state
1076   1076  -4.725846  -3.437487 0.2051936 0.07826202 0.439771     U
```

```
MGMTqc.single(pred1)
```

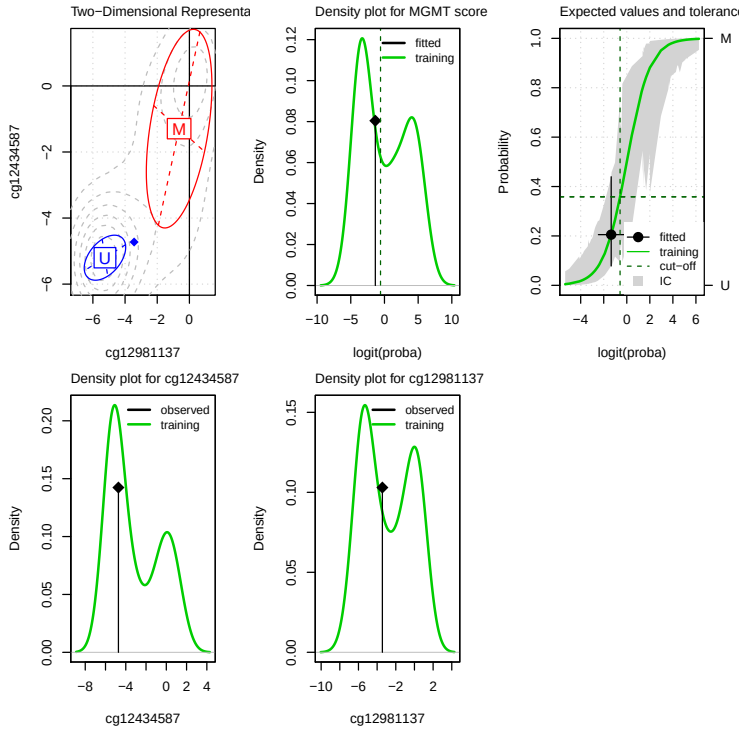

FIGURE 5 – Graphical quality control for single sample prediction the first sample from TCGA datasets

.

The function can be used on multi-sample datset to investigate the quality control of a given sample (14th in the example below).

```
MGMTqc.single(prednewtcga,nsample=14)
```

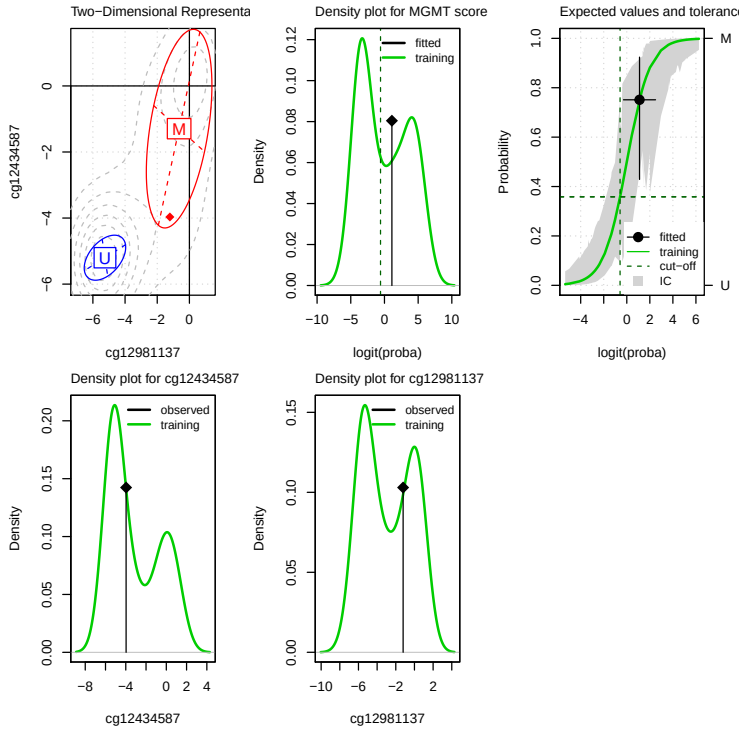

FIGURE 6 – Graphical quality control for single sample prediction for the 14th samples from TCGA datasets.

.

7 Uncertainty and reliability of the model

An essential issue for any model is the estimation uncertainty related to the prediction.A model is simplification of the real systems and its utility depends in part on the accuracy and reliability of its outputs.

7.1 Extreme values

The comportment of the model was evaluated in presence of extreme M-values for the two Infinium predictors `cg12434587` and `cg12981137`. An illustration of these results was given in the following figure.

```
library(lattice)
library(latticeExtra)
grid1 <- expand.grid("cg12434587"=seq(-15,5,len=100),"cg12981137"=seq(-15,5,len=100))
ted <- cbind(grid1,MGMTpredict(grid1))
## graph
cut1 <- MGMTSTP27$perf$cut
funcol <- function(...) grey.colors(...,start=0.9,end=0.3)
margin <- 0.05
grid1 <- seq(-15,10,by=5)
lvp <- levelplot(pred~cg12434587+cg12981137, data=ted,xlim=c(-10,5),ylim=c(-10,5),
                 col.regions = funcol,panel=function(...){
                   panel.levelplot(...);
                   panel.abline(h=grid1,col="white",lty=3);
                   panel.abline(v=grid1,col="white",lty=3);
                   panel.abline(h=0,col="white",lwd=2);
                   panel.abline(v=0,col="white",lwd=2);
                   panel.abline(a=0,b=1,col="white",lwd=2,lty=2);
                 })
cvp <- contourplot(pred~cg12434587+cg12981137, data=ted,
                   xlim=c(-10,5),ylim=c(-10,5),at=c(0.2,0.4,0.6),labels=TRUE,lty=2)
cvpbis <- contourplot(pred~cg12434587+cg12981137, data=ted,
                      xlim=c(-10,5),ylim=c(-10,5),at=c(cut1),col="red3",labels=FALSE)
xyp <- xyplot(cg12981137~cg12434587, data=MGMTpredict(NCHgbm450),xlim=c(-10,5),
              ylim=c(-10,5),groups=state,col="black",fill=c("darkgray","white"),
              pch=c(24,25),cex=1.25)
lvp <- lvp+as.layer(xyp, axes = NULL)+as.layer(cvp,axes=NULL)+as.layer(cvpbis,axes=NULL)
print(lvp)
```

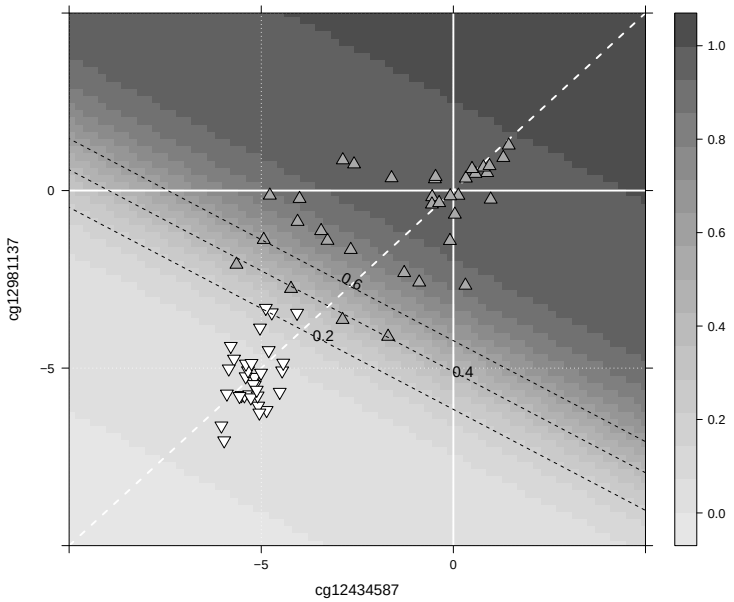

FIGURE 7 – Representation of model sensitivity to extreme values for the two Infinium probes `cg12434587` and `cg12981137`. The output related to training dataset are sumperimpose on the graphic. The dotted line give probability limits and the red line identified the cut-off given in [2]

.

7.2 Confidence and tolerance intervals

Computation of confidence and tolerance intervals in the function MGMTpredict from the package mgmtstp27 is based on the error propagation principle to obtain limits contained in the range [0 ;1]. The computtaion is provided below :

$$IC_{1-\alpha} = g^{-1}(\hat{y} \pm z_{\alpha}\sigma_{\hat{y}})$$

Where $\hat{y}$, $\sigma_{\hat{y}}$ and z_{α} correspond to the estimated values, standard deviation associated with estimated values and the theoretical values from Normal distribution (α= error type I). The Estimation of the variance of the estimated values for confidence interval is given below :

$$\sigma_{\hat{y}} = \sigma X^t (X^t X)^{-1} X$$

The value σ corresponds to the dispersion term. This value is postulated equal to 1 in logistique regression by default.An estimation of σ can be computed as follow :

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

where the term $n - p - 1$ corresponds to the degree of freedom of the model.

8 Acknowledgments

The results published here are in part based upon data generated by The Cancer TCGA Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at (http ://cancergenome.nih.gov). The dbGaP accession number to the specific version of the TCGA data set is phs000178.v8.p7.

Références

[1] Martin J Aryee, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. Minfi : A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays. *Bioinformatics*, 2014.

[2] Pierre Bady, Davide Sciuscio, Annie-Claire Diserens, Jocelyne Bloch, Martin J. van den Bent, Christine Marosi, Pierre-Yves Dietrich, Michael Weller, Luigi Mariani, Frank L. Heppner, David R. McDonald, Denis Lacombe, Roger Stupp, Mauro Delorenzi, and Monika E. Hegi. Mgmt methylation analysis of glioblastoma on the infinium methylation beadchip identifies two distinct cpg regions associated with gene silencing and outcome, yielding a prediction model for comparisons across datasets, tumor grades, and cimp-status. *Acta Neuropathologica*, 124(4) :547–560, 2012. Times Cited : 6.

[3] Sean Davis, Pan Du, Sven Bilke, Tim Triche, Jr., and Moiz Bootwalla. *methylumi : Handle Illumina methylation data*, 2014. R package version 2.10.0.

9 Appendix

9.1 Import raw HM-27K data (format .IDAT)

To import raw data (format **.IDAT**), the function contains in R package `minfi` ([1]) don't work with HM-27K. However, it's possible to import data with functions from R package `methylumi` ([3]).

```
require(methylumi)
rgset0<- methylumIDAT(barcode=as.character(File.Name),idatPath=datadir)
# no normalization for HM-27k,
# see help "For HumanMethylation27 data, the function does nothing"
norm27k <- normalizeMethyLumiSet(rgset0)
u27k <- unmethylated(norm27k)
m27k <- methylated(norm27k)
mvalue0 <- log2((m27k+1)/(u27k+1))
```

9.2 Session

```
print(sessionInfo(),locale=FALSE)
```

```
R version 3.1.2 (2014-10-31)
Platform: x86_64-w64-mingw32/x64 (64-bit)

attached base packages:
[1] stats4    parallel  stats     graphics  grDevices utils     datasets  methods
[9] base

other attached packages:
 [1] latticeExtra_0.6-26  RColorBrewer_1.1-2   mgmtstp27_0.6
 [4] MASS_7.3-40          methylumi_2.12.0     matrixStats_0.14.0
 [7] ggplot2_1.0.1        reshape2_1.4.1       scales_0.2.4
[10] ade4_1.7-2           lumi_2.18.0          minfi_1.12.0
[13] bumphunter_1.6.0     locfit_1.5-9.1       iterators_1.0.7
[16] foreach_1.4.2        Biostrings_2.34.1    XVector_0.6.0
[19] GenomicRanges_1.18.4 GenomeInfoDb_1.2.5   IRanges_2.0.1
[22] S4Vectors_0.4.0      lattice_0.20-31      Biobase_2.26.0
[25] BiocGenerics_0.12.1

loaded via a namespace (and not attached):
 [1] affy_1.44.0             affyio_1.34.0           annotate_1.44.0
 [4] AnnotationDbi_1.28.2    base64_1.1              base64enc_0.1-2
 [7] BatchJobs_1.6           BBmisc_1.9              beanplot_1.2
[10] BiocInstaller_1.16.4    BiocParallel_1.0.3      biomaRt_2.22.0
[13] bitops_1.0-6            brew_1.0-6              checkmate_1.5.2
[16] codetools_0.2-11        colorspace_1.2-6        DBI_0.3.1
[19] digest_0.6.8            doRNG_1.6               fail_1.2
[22] genefilter_1.48.1       GenomicAlignments_1.2.2 GenomicFeatures_1.18.7
[25] grid_3.1.2              gtable_0.1.2            illuminaio_0.8.0
[28] KernSmooth_2.23-14      limma_3.22.7            Matrix_1.2-0
[31] mclust_5.0.0            mgcv_1.8-6              multtest_2.22.0
[34] munsell_0.4.2           nleqslv_2.7             nlme_3.1-120
[37] nor1mix_1.2-0           pkgmaker_0.22           plyr_1.8.1
[40] preprocessCore_1.28.0   proto_0.3-10            quadprog_1.5-5
[43] Rcpp_0.11.5             RCurl_1.95-4.6          registry_0.2
[46] reshape_0.8.5           rngtools_1.2.4          Rsamtools_1.18.3
[49] RSQLite_1.0.0           rtracklayer_1.26.3      sendmailR_1.2-1
[52] siggenes_1.40.0         splines_3.1.2           stringr_0.6.2
[55] survival_2.38-1         tools_3.1.2             XML_3.98-1.1
[58] xtable_1.7-4            zlibbioc_1.12.0
```