Working notes 

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

BADY P.

2 octobre 2014

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 **2**

2 Data **2**

3 Probability that MGMT pomoter is methylated **2**

4 DNA methylation state of MGMT promoter **3**

5 Quality control for population prediction (in development) **5**

6 Acknowledgments **7**

7 Appendix **8**
7.1 Import HM-27K data in IDAT format 8
7.2 Session 8

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 internal 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 elow :

```
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"             "perf"
```

```
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$perf
```

```
        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)
```

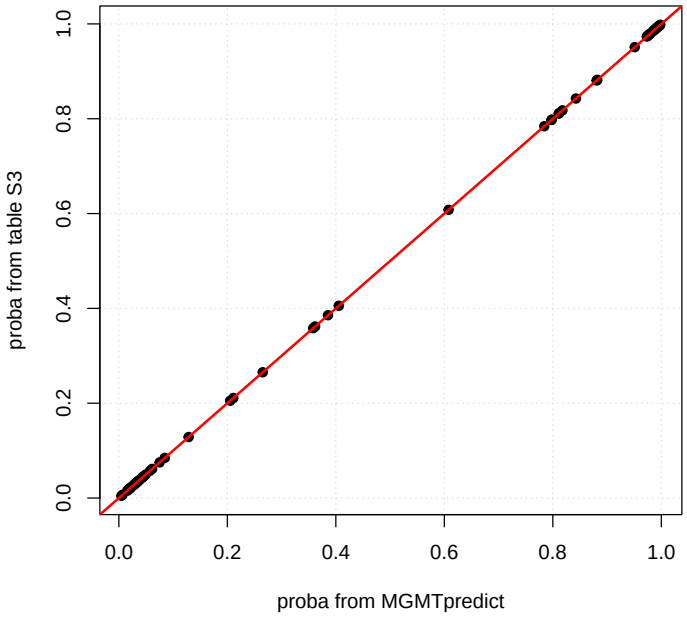

FIGURE 1 – Comparison of the prediction from the table S3 ([2]) and the outputs from the function `MGMTpredict`.

.

```
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)
```

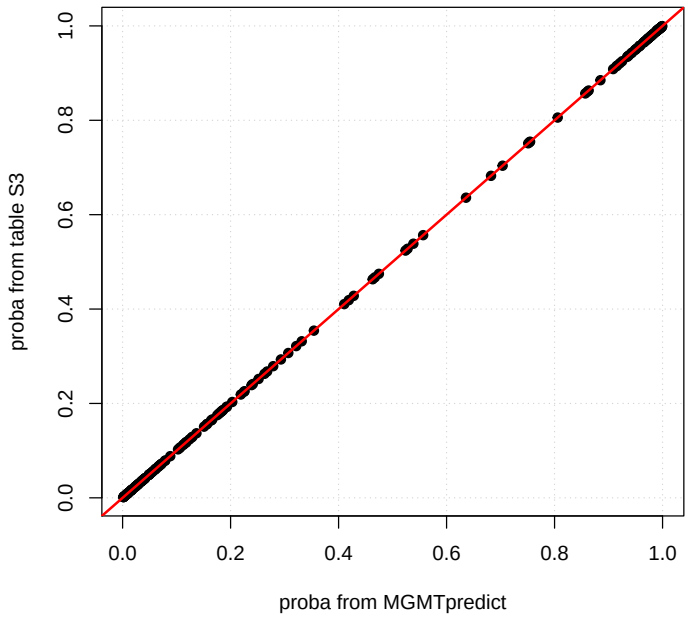

FIGURE 2 – Comparison of the prediction from the table S5 ([2]) and the outputs from the function `MGMTpredict`.

.

5 Quality control for population prediction (in development)

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 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(prednewnch)
```

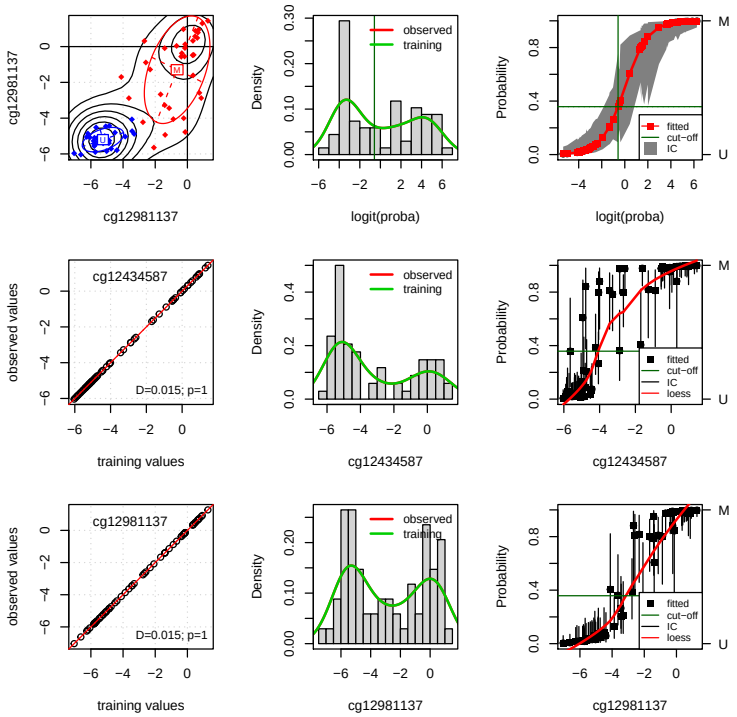

FIGURE 3 – Graphical quality control for prediction from NCH datasets

.

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

```
MGMTqc(prednewtcga)
```

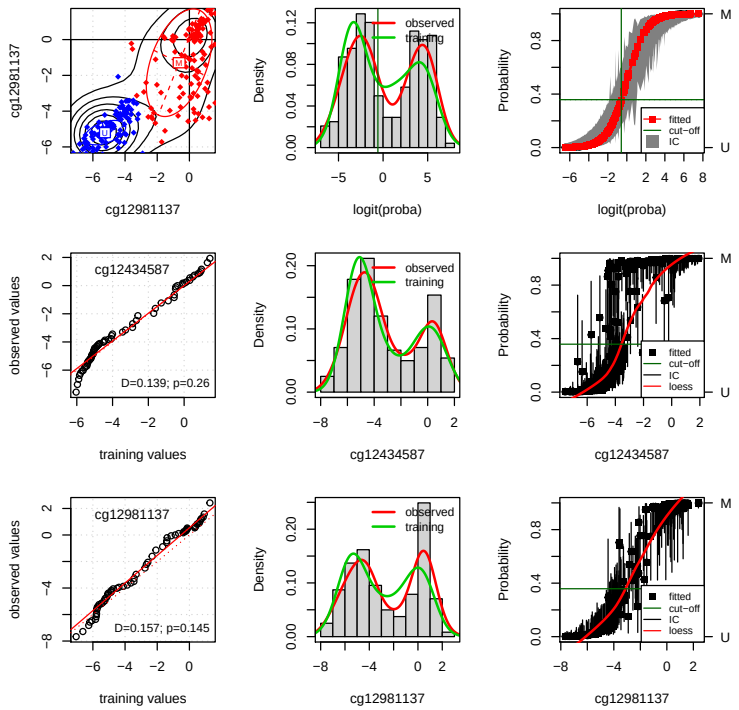

FIGURE 4 – Graphical quality control for prediction from TCGA datasets

.

6 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.

7 Appendix

7.1 Import HM-27K data in IDAT format

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))
```

7.2 Session

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

```
R version 3.1.1 (2014-07-10)
Platform: x86_64-w64-mingw32/x64 (64-bit)

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

other attached packages:
 [1] mgmtstp27_0.1        MASS_7.3-35          methylumi_2.10.0
 [4] matrixStats_0.10.0   ggplot2_1.0.0        reshape2_1.4
 [7] scales_0.2.4         ade4_1.6-2           lumi_2.16.0
[10] minfi_1.10.2         bumphunter_1.4.2     locfit_1.5-9.1
[13] iterators_1.0.7      foreach_1.4.2        Biostrings_2.32.1
[16] XVector_0.4.0        GenomicRanges_1.16.4 GenomeInfoDb_1.0.2
[19] IRanges_1.22.10      lattice_0.20-29      Biobase_2.24.0
[22] BiocGenerics_0.10.0

loaded via a namespace (and not attached):
 [1] affy_1.42.3             affyio_1.32.0           annotate_1.42.1
 [4] AnnotationDbi_1.26.0    base64_1.1              base64enc_0.1-2
 [7] BatchJobs_1.4           BBmisc_1.7              beanplot_1.2
[10] BiocInstaller_1.14.2    BiocParallel_0.6.1      biomaRt_2.20.0
[13] bitops_1.0-6            brew_1.0-6              BSgenome_1.32.0
[16] checkmate_1.4           codetools_0.2-9         colorspace_1.2-4
[19] DBI_0.3.1               digest_0.6.4            doRNG_1.6
[22] fail_1.2                genefilter_1.46.1       GenomicAlignments_1.0.6
[25] GenomicFeatures_1.16.2  grid_3.1.1              gtable_0.1.2
[28] illuminaio_0.6.1        KernSmooth_2.23-13      limma_3.20.9
[31] Matrix_1.1-4            mclust_4.4              mgcv_1.8-3
[34] multtest_2.20.0         munsell_0.4.2           nleqslv_2.5
[37] nlme_3.1-117            nor1mix_1.2-0           pkgmaker_0.22
[40] plyr_1.8.1              preprocessCore_1.26.1   proto_0.3-10
[43] R.methodsS3_1.6.1       RColorBrewer_1.0-5      Rcpp_0.11.3
[46] RCurl_1.95-4.3          registry_0.2            reshape_0.8.5
[49] rngtools_1.2.4          Rsamtools_1.16.1        RSQLite_0.11.4
[52] rtracklayer_1.24.2      sendmailR_1.2-1         siggenes_1.38.0
```

```
[55] splines_3.1.1           stats4_3.1.1            stringr_0.6.2
[58] survival_2.37-7         tools_3.1.1             XML_3.98-1.1
[61] xtable_1.7-4            zlibbioc_1.10.0
```