Working notes:

Effect of normalization on the prediction of DNA methylation status of MGMT promoter: example with HM-450K Infinium data from TCGA.

Contributors:

Pierre Bady

Monika Hegi

Date: 2012-08-07

Revision: 2012-08-10

Abstract:

This document contains information related to the effect of normalization of HM-450K Infinium platform on the prediction of the DNA methylation status of the MGMT promoter (Bady et al. 2012).

Keywords: MGMT, DNA methylation, HM-450K, Infinium, TCGA

Contents

Motivations

In this document, we propose to evaluate the effect of the normalization of the data from Infinium HM-450K platform (DNA methylation) on the prediction of the DNA methylation of the MGMT promoter from the model proposed in Bady et al. (2012).

Material and method

Biological Data

Dataset came from TCGA project (The Cancer Genome Atlas Research Network 2008, http://cancergenome.nih.gov/). The DNA methylation was evaluated by platform Infinium HM-450K. The first dataset come from the older version dated to 2012-05-25 where the level 1 (see TCGA documentation) directly contained the preprocessed information from 74 samples (e.g. unmethylated and methylated intensities). However, little information was provided to describe the normalization/preprocessing used to prepare this dataset. A second version (2012-07-31) of this data set in raw format (the information of the two colors is separated in two different files) were used to determine the normalization used in the initial dataset and to compare the preprocessing methods. The dataset used in Bady et al. (2012) as training dataset (M-GBM), was analyzed in a similar way.

Preprocessing and Normalization

For the initial dataset, we have some doubts on the method used to preprocess the data. Concerning the new dataset (updated version), we used two different methods available in Genome Studio:

- "raw" version corresponding to the method initially used to prepare the data from HM-27K platform. Preprocessing means converting the Red and Green channel into unmethylated and methylated signal.
- The second method corresponds to a new method proposed by Illumina to preprocess the HM-450K data. The procedure includes background correction and normalization using a sample as reference (the second by default, see documentation of R package minfi, Kasper & Martin 2012).

The functions from the R package **minfi**[1] (Kasper & Martin 2012) were used to perform both these normalizations. In this study, we didn't take into account the chemistry effect because the two probes considered in our model came from the chemistry I only. The R package **lumi** provided additional functions for normalization (Du & Lin 2008, not used here).

Statistical analyses

Analyses and Graphical representations were performed using R-2.15.1 (R Development Core Team 2012) and the R package **minfi** (Kasper & Martin 2012) and **methylumi** (Davis et al. 2012).

[1] The package **minfi** is used by the package **methylumi** (Davis et al. 2012) for normalization in the function **normalizeMethyLumiSet**.

Results

Data preparation

The importation and preparation of the three datasets were relatively facilitated by the use of the function from R package **minfi**. The functions **preprocessRaw** and **preprocessIllumina** provided the two new datasets from the last update (see R code below). The dataset used in the table S4 (Bady et al. 2012, R object called *predTCGA450K*) was built manually because the old structure of the level 1 data was not compatible with the functions of R packages **minfi** or **methylumi**.

```
#--------------------------
# data importation
#--------------------------
library(minfi)
library(IlluminaHumanMethylation450kmanifest)

# data importation
datadir <- paste(getwd(),"/JHU_USC__HumanMethylation450/Level_1/",sep="")
list.files(datadir)
infofile0 <- read.table("file_manifest.txt",h=TRUE,sep="\t")
infofile1 <- infofile0[infofile0$Level==1,]
rgset0 <- read.450k.exp(datadir)

# preprocessing
rawdata0 <- preprocessRaw(rgset0)
normdata0 <- preprocessIllumina(rgset0)

# meylation and unmethylation data
rawunmeth0 <- getUnmeth(rawdata0)
rawmeth0 <- getMeth(rawdata0)
normunmeth0 <- getUnmeth(normdata0)
normmeth0 <- getMeth(normdata0)

# table containing the probes used in the model
load("promoterprobes.rda")
rawunmeth1 <- rawunmeth0[promoterprobes,]
rawmeth1 <- rawmeth0[promoterprobes,]
normunmeth1 <- normunmeth0[promoterprobes,]
normmeth1 <- normmeth0[promoterprobes,]
mvalueraw1 <- log2((rawmeth1+1)/(rawunmeth1+1))
mvaluenorm1 <- log2((normmeth1+1)/(normunmeth1+1))

# initial dataset (74 samples used in the table S4 in Bady et al. 2012)
load("/export/scratch/data/monikaproject/TCGA6/DNAmethylation/450k/predTCGA450.rda"
)
```

Comparison of the three datasets (PP-27K, PP-450K, TCGA-450K)

In this section, we only kept the samples common to the three datasets. Consequently, we had three measures by probes for a given sample:

- The dataset called PP-27K corresponds to the raw dataset from TCGA (update 2012-07-31) after classical preprocessing/normalization (that correspond to the normalization initially used for HM-27K platform). It contained 124 samples before matching step.
- The dataset called PP-450K corresponds to the raw dataset from TCGA (update 2012-07-31) after "new" Illumina preprocessing. It contained 124 samples before matching step.
- TCGA-450K corresponds to the dataset (update 2012-05-25) used for the prediction in the table S4 (Bady et al. 2012). It contained 74 samples.

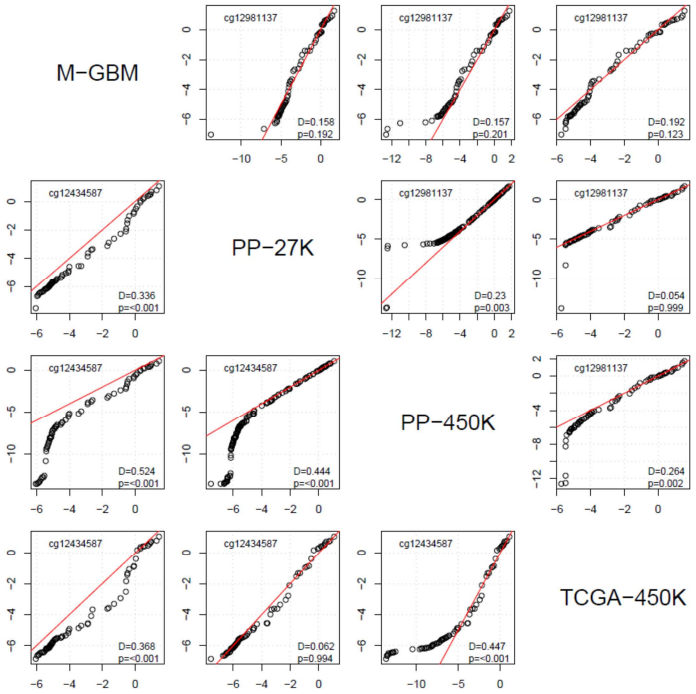

Figure 1. Comparison of M-value distributions between the three "unmatched" datasets and the training dataset (M-GBM). The M-values of the probes cg12434587 and cg12434587 used in MGMT-STP27 were compared by quantile-quantile representation (QQ-plot). The red line corresponds to the line y=x. The terms 'D' and 'p' refer to the comparison of the distribution by the Kolmogorov-Smirnov test. The platform Illumina used is indicated for each dataset. When the p-value is inferior to 0.05, the two distributions are considered as significantly different (for better resolution see PDF file).

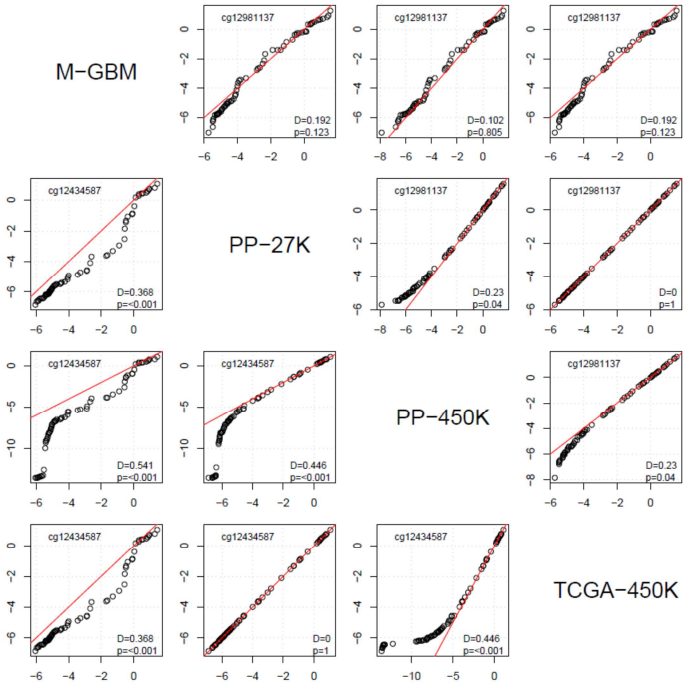

Figure 2. Comparison of M-value distributions between the three "matched" datasets and the training dataset (M-GBM). The M-values of the probes cg12434587 and cg12434587 used in MGMT-STP27 were compared by quantile-quantile representation (QQ-plot). The red line corresponds to the line y=x. The terms 'D' and 'p' refer to the comparison of distribution by the Kolmogorov-Smirnov test. The platform Illumina used is indicated for each dataset. When the p-value is inferior to 0.05, the two distributions are considered as significantly different (for better resolution see PDF file). (for better resolution see PDF file).

After matching based on the sample names, the three datasets contained 74 samples. The analyses in Figure 2 and Figure 3 show that the initial dataset (TCGA-450K) is exactly similar to the dataset normalized by the "raw" preprocessing (PP-27K). The procedure used to preprocess the initial dataset is certainly the same and corresponds to the procedure used to prepare the HM-27K datasets. Consequently, the prediction proposed in the table S4 (Bady et al. 2012) is the same.

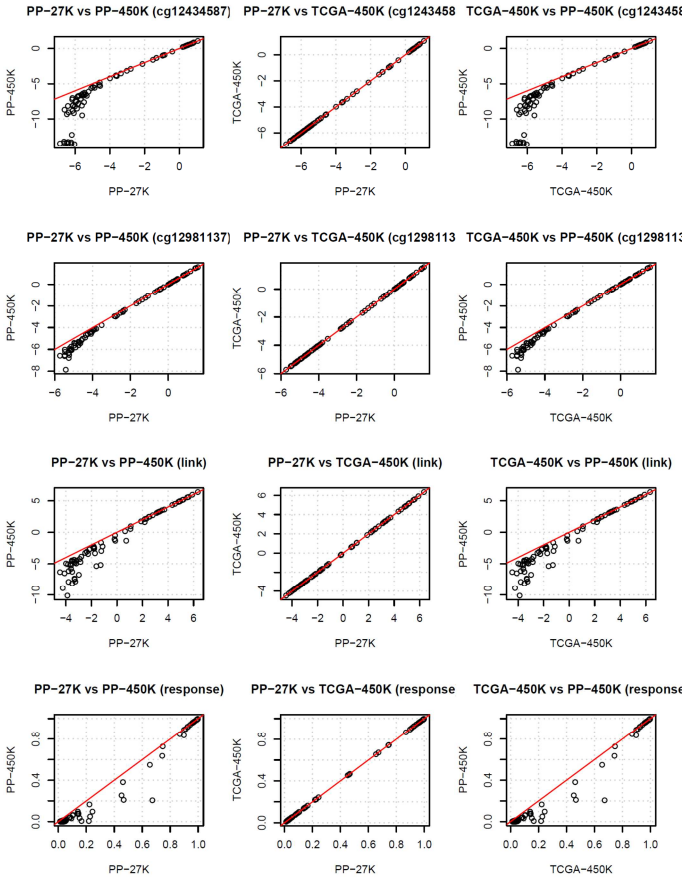

Figure 3. Comparisons of the values of the both probes (cg12434587 and cg12434587) and predictions (link and response values) between the three matched dataset from TCGA (for better resolution see PDF file).

The correlation between PP-27K and TCGA-450K datasets is perfect. The discrepancies between training (M-GBM) and PP-450K datasets were excessively increased by the normalization proposed in Genome Studio. The highest deviations between PP-27K and PP-450K were observed for the probe cg12434587 and they were mainly observed for the low M-values (Figure 3). The evaluation of the concordance between predicted statuses is provided below:

```
R> predraw2 <- predict(step27k,dfraw2,type="response")
R> mgmtraw2 <- ifelse(predraw2>=step27k$perf$cut,"M","U")
R> prednorm2 <- predict(step27k,dfnorm2,type="response")
R> mgmtnorm2 <- ifelse(prednorm2>=step27k$perf$cut,"M","U")
R> predini2 <- predict(step27k,predTCGA2,type="response")
R> mgmtini2 <- ifelse(predini2>=step27k$perf$cut,"M","U")
R>
R> table(mgmtraw2,mgmtini2)
        mgmtini2
mgmtraw2  M  U
       M 32  0
       U  0 42
```

```
R> table(mgmtraw2,mgmtnorm2)
        mgmtnorm2
mgmtraw2  M  U
       M 29  3
       U  0 42
R> table(mgmtnorm2,mgmtini2)
         mgmtini2
mgmtnorm2  M  U
        M 29  0
        U  3 42
R>
```

We observe that the initial dataset was in perfect concordance with the dataset normalized by “raw” preprocessing. When the dataset was normalized by new Illumina procedure, we observe that **three** samples were not correctly classified.

Normalization effect for the training dataset (M-GBM, Bady et al. 2012)

As previously, three datasets were considered in these analyses:

- **M-PP-27K** corresponds to the raw dataset after classical preprocessing/normalization (that correspond to the normalization initially used for HM-27K platform).
- **M-PP-450K** corresponds to the raw dataset after "new" Illumina preprocessing
- **M-GBM-450K** corresponds to the training dataset used to perform the model proposed in Bady et al. (2012)

As expected, the results observed were very similar to the ones presented in the previous section. We observed deviations between PP-27K (identical to M-GBM dataset) and PP-450K for the both probes and they mainly observed the low M-values (Figure 4).

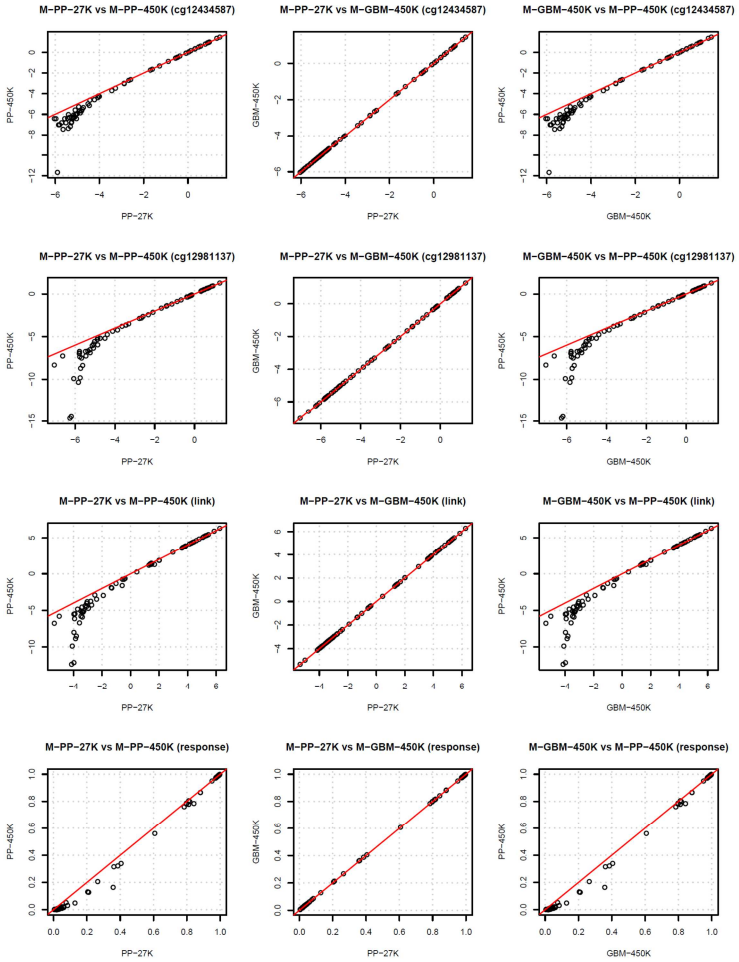

Figure 4. Comparisons of the values of the both probes (cg12434587 and cg12434587) and predictions (link and response values) between the three matched datasets from M-GBM data used as training dataset in Bady et al. 2012 (for better resolution see PDF file).

Four samples were not correctly classified. The evaluation of the concordance between the predicted statuses is provided below:

```
R> # comparison prediction
R> predraw2 <- predict(step27k,dfraw2,type="response")
R> mgmtraw2 <- ifelse(predraw2>=step27k$perf$cut,"M","U")
R> prednorm2 <- predict(step27k,dfnorm2,type="response")
R> mgmtnorm2 <- ifelse(prednorm2>=step27k$perf$cut,"M","U")
R> predini2 <- predict(step27k,predGBM2,type="response")
R> mgmtini2 <- ifelse(predini2>=step27k$perf$cut,"M","U")
R>
R> table(mgmtraw2,mgmtini2)
        mgmtini2
mgmtraw2  M  U
       M 35  0
       U  0 33
R> table(mgmtraw2,mgmtnorm2)
        mgmtnorm2
mgmtraw2  M  U
       M 31  4
       U  0 33
R> table(mgmtnorm2,mgmtini2)
         mgmtini2
mgmtnorm2  M  U
        M 31  0
        U  4 33
```

```
R>
```

Conclusion

Comment and instructions (as they come) related to the choice of normalization for using the model predicting the DNA methylation of the MGMT promoter, are given below:

- Original data from TCGA (update 2012-05-25) was preprocessed as HM-27K data (e.g. the function **rawpreprocessRaw** from R package **minfi**).
- The normalization can affect the prediction of the DNA methylation (it's not really a surprise).
- The generalization of the model can be affected by the new normalization proposed by Illumina (**preprocessIllumina**). The reference samples used during the normalization procedure were fixed within each dataset and they were not the same among the datasets.
- The model in Bady et al. (2012) requires to use the initial preprocessing (normalization) proposed initialy by Illumina that corresponds to the function **preprocessRaw** from R package **minfi**.

- The predictions proposed in the table S4 for the dataset based on HM-450K of the paper are consistent with our previous comments/recommendations (see above).
- There are no problem/bias induced by chemistry type, because the two probes used in the model come from the chemistry I as in the HM-27K platform

References

- Bady P., Sciuscio D., Diserens A.-C., Bloch J., van den Bent M.J., Marosi C., Dietrich P.-Y., Weller M., Mariani L., Heppner F.L., Macdonald D.R., Lacombe D., Stupp R., Delorenzi M. and Hegi M.E. (2012) 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 CIMPstatus, Acta Neuropathologica.
- Du, P., Kibbe, W.A. and Lin, S.M., (2008) 'lumi: a pipeline for processing Illumina microarray', Bioinformatics 24(13):1547-1548
- Kasper Daniel Hansen and Martin Aryee (2012) minfi: Analyze Illumina's 450k methylation arrays. R package version 1.2.0.
- Sean Davis, Pan Du, Sven Bilke, Tim Triche, Jr. and Moiz Bootwalla (2012). methylumi: Handle Illumina methylation data. R package version 2.2.0.
- R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
- The Cancer Genome Atlas (TCGA) Research Network (2008) The Cancer Genome Atlas (TCGA) Research Network, Comprehensive genomic characterization defines human glioblastoma genes and core pathways, Nature 455, pp. 1061–1068.