BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 12 2011, pages 1693–1695
doi:10.1093/bioinformatics/btr187

Genome analysis

Advance Access publication April 14, 2011

# SMETHILLIUM: spatial normalization METHod for ILLumina InfinIUM HumanMethylation BeadChip

Camille Sabbah[1,2,3,†], Gildas Mazo[1,2,3,†], Caroline Paccard[1,2,3], Fabien Reyal[1,4,5] and Philippe Hupé[1,2,3,4,*]

[1]Institut Curie, 26 rue d'Ulm, F-75248, [2]INSERM, U900, Paris F-75248, [3]Mines ParisTech, Fontainebleau F-77300, [4]CNRS UMR144, 26 rue d'Ulm, F-75248 and [5]Institut Curie, Department of Surgery, 26 rue d'Ulm, Paris, F-75248

Associate Editor: Alfonso valencia

## ABSTRACT

**Summary:** DNA methylation is a major epigenetic modification in human cells. Illumina HumanMethylation27 BeadChip makes it possible to quantify the methylation state of 27 578 loci spanning 14 495 genes. We developed a non-parametric normalization method to correct the spatial background noise in order to improve the signal-to-noise ratio. The prediction performance of the proposed method was assessed on three fully methylated samples and three fully unmethylated DNA samples. We demonstrate that the spatial normalization outperforms BeadStudio to predict the methylation state of a given locus.

**Availability and Implementation:** A R script and the data are available at the following address: http://bioinfo.curie.fr/projects/smethillium.

**Contact:** smethillium@curie.fr

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on September 15, 2010; revised on March 11, 2011; accepted on April 5, 2011

## 1 INTRODUCTION

DNA methylation is a major epigenetic modification that regulates transcription by gene silencing and have a role in the protection of chromosomal integrity. Methylation occurs in cytosines which precede guanines in dinucleotide called CpGs. Illumina HumanMethylation27 BeadChip (Illumina, 2008; Weisenberger *et al.*, 2008) is a valuable technology for genome-wide screen of DNA methylation. As any microarray technology, BeadChip suffers from spatial artifact that requires a correction in order to improve the signal-to-noise ratio. To our knowledge, the methylumi R package (Davis and Bilke, 2010) is the only one to deal with Illumina methylation BeadChip but does not address the spatial correction. The proposed method SMETHILLIUM consists of a spatial correction of the background noise, a probe summarization and a confidence assessment. The performance of the method is evaluated on six real samples with known methylation states.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

## 2 METHOD

### 2.1 BeadChip design

Let us note $(B^M_{i,j})_{j=1,\ldots,n^M_i}$ the set of beads for the $n^M_i$ methylated probes, and $(B^U_{i,j})_{j=1,\ldots,n^U_i}$ the set of beads for the $n^U_i$ unmethylated probes for a locus $i$ ($i=1,\ldots,N$ where $N$ is the total number of loci). Whatever the methylation state and for any locus $i$, the intensity signal for both types of probes is measured by the same single colour channel using the single-base extension process (Weisenberger *et al.*, 2008). However, as purines and pyrimidines are labelled, respectively, with a red and green dyes, beads can be attributed to a red set $R$, or a green set $G$. For all probes, both red and green signal intensities are quantified for any $B^M_{i,j}$ and $B^U_{i,j}$. For a given channel, the raw intensities are noted $I^M_{i,j}$ and $I^U_{i,j}$.

### 2.2 Spatial normalization

On all Illumina methylation BeadChip, a number of $n_0$ Negative Control Beads (NCB) are calibrated to not hybridize with any DNA template. We thus consider that the intensities measured on the NCB correspond to spatial noise. More precisely, we assume that the red (respectively green) intensities $I^R_{0,i}$, $i=1,\ldots,n_0$ (respectively $I^G_{0,i}$, $i=1,\ldots,n_0$) measured on these probes are related to their locations $\mathrm{BC}^R_i$, $i=1,\ldots,n_0$ (respectively $BC^G_i$, $i=1,\ldots,n_0$) on the chip *via* an unknown function $N_R(\cdot)$ (respectively $N_G(\cdot)$). In the following, we only consider the spatial noise in the red channel since the procedure for the green channel is the same.

As local polynomial estimation techniques apply at the boundaries (Fan and Gijbels, 1996), we estimate the function $N_R(\cdot)$ non-parametrically by a local polynomial estimator $\widehat{N}_R(\cdot,s)$ constructed upon the 2D *LOESS* model. The parameter $s$ in $(0,1]$ is called the *span*. A small $s$ value will reduce the bias of $\widehat{N}_R(\cdot,s)$ and increase variance while a large $s$ value will reduce variance and increase the bias. Choosing $s$ is an issue and we propose an adaptive method for the choice of an optimal data-driven $s$ based on cross-validation approach. Namely, for each colour channel, and for each $s$ on an arithmetic grid $\mathcal{S}$, we compute the average prediction error (PE) of $N_R(\cdot)$ by $\widehat{N}_R(\cdot,s)$ as follows:

$$\mathrm{PE}(s)=\frac{1}{n_0}\sum_{i=1}^{n_0}\left(\widehat{N}_{R,-i}(\mathrm{BC}^R_i,s)-I^R_{0,i}\right)^2,$$

where $\widehat{N}_{R,-i}(\cdot,s)$ is the leave-one-out version of the *LOESS* estimator $\widehat{N}_R(\cdot,s)$. We then choose the span $s^*_R$ which satisfies:

$$s^*_R=\underset{s\in\mathcal{S}}{\operatorname{argmin}}\ PE(s).$$

The $s^*_R$ which minimizes the average PE is afterward used in the estimation of $N_R(\cdot)$. Our local polynomial estimator of $N_R(\cdot)$ is then $\widehat{N}_R(\cdot)=\widehat{N}_R(\cdot,s^*_R)$. Note that the resulting estimator achieves the global optimal rate of Stone (1984).

Once the *span* selected, the *LOESS* model allows the extrapolation of the background noise for all the beads in the set $R$ (respectively $G$) taking

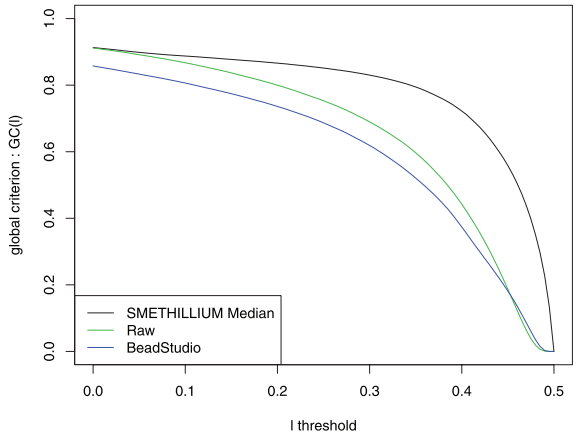


**Fig. 1.** Global performance criterion for the six pooled samples.

into account their location. For the set $R$ of all red beads, $\widehat{N}_R(B^S_{i,j})$ is the extrapolated value for a given bead where $S$ is either $M$ or $U$. The $\widehat{I}^S_{i,j}$ intensity corrected from the red background correction is computed as follows:

$$\widehat{I}^S_{i,j} = \begin{cases} I^S_{i,j} - \widehat{N}_R(B^S_{i,j}) + 1 \text{ if } I^S_{i,j} - \widehat{N}_R(B^S_{i,j}) > \delta, \\ ed(I^S_{i,j}, \widehat{N}_R(B^S_{i,j})) + 1 \text{ otherwise,} \end{cases}$$

with $ed(i,b) = \delta \exp[1-(b+\delta)/i]$ where $i$ is the raw signal intensity and $b$ is its estimated background. We choose $\delta = 1$ as recommended by Edwards (2003). A similar procedure is applied to correct the green background noise.

## 2.3 Probe summarization

For each locus $i$, the method computes the methylated, unmethylated and beta values using median (another method of summarization based on the mean is available and described in the Supplementary Material) as follows:

$$\widetilde{M}_i = \mathrm{Med}^{n_i^M}_{j=1}\left(\widehat{I}^M_{i,j}\right),\ \widetilde{U}_i = \mathrm{Med}^{n_i^U}_{j=1}\left(\widehat{I}^U_{i,j}\right),\ \widetilde{\beta}_i = \frac{\widetilde{M}_i}{\widetilde{M}_i + \widetilde{U}_i},$$

where $\mathrm{Med}^n_{j=1}(X_j)$ is the median of $(X_1, \ldots, X_n)$.

Depending on the beta value, a methylation state is assigned according to the following decision rule for each locus $i$:

$$S_i = \begin{cases} M \text{ if } \widetilde{\beta}_i \geq 0.5 + l, \\ U \text{ if } \widetilde{\beta}_i \leq 0.5 - l, \\ H \text{ if } \widetilde{\beta}_i \in (0.5 - l, 0.5 + l), \end{cases}$$

where $H$ represents a hemi-methylation state and $0 < l \leq 0.5$.

## 2.4 Confidence for methylation state

In order to define a confidence value for the assigned methylation state, a non-parametric test is proposed. In the following, we only consider the red beads since the test procedure for the green beads is the same. Namely, we use the Wilcoxon–Mann–Whitney statistic test. For any given locus $i$, let $\mathrm{med}^U_i$ and $\mathrm{med}^M_i$ be the theoretical values of the median of the normalized intensities. Let $\mathrm{med}_0$ be the theoretical value of the median of the normalized NCB intensities. The test we propose works as follows:

$$\begin{array}{l} {}^iH_0 : \max\left(\mathrm{med}^U_i, \mathrm{med}^M_i\right) = \mathrm{med}_0 \\ {}^iH_1 : \max\left(\mathrm{med}^U_i, \mathrm{med}^M_i\right) < \mathrm{med}_0. \end{array} \text{ versus}$$

Typically, we expect that ${}^iH_0$ is rejected for all loci since at least one type of probes must be significantly different from the NCB. Therefore, loci showing systematically high $P$-values over a set of experiments will be unreliable, possibly for technical reasons.

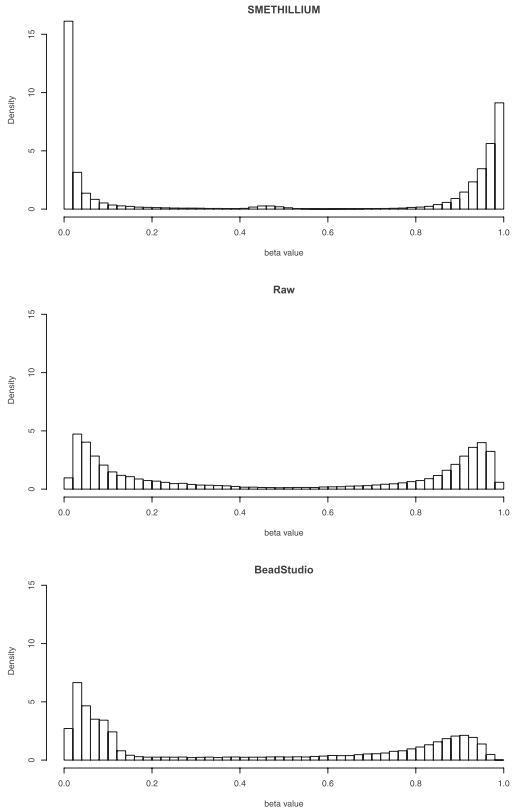


**Fig. 2.** Histograms for the beta values obtained with SMETHILLIUM, raw data and BeadStudio for the six pooled samples.

## 2.5 ROC and performance prediction

In order to validate the efficiency of the method, DNA extracted from a T24 bladder cancer cell line was used. A first group of DNA samples (called *met*1.1, *met*3.1 and *met*3.2) was treated with Methylase SssI such that all CpGs loci are methylated. A second group of DNA samples (called *RCA*, *Zeb*3.1 and *Zeb*3.2) was amplified by PCR for the *RCA* sample and by the Zebularine methylation inhibitor for the *Zeb*3.1 and the *Zeb*3.2 samples such that all CpGs loci are unmethylated after treatment. Therefore, both groups of samples are representative of a fully methylated DNA and a fully unmethylated DNA. Then, both groups of samples were individually hybridized on Illumina HumanMethylation27 BeadChip and separately normalized according to the proposed method. Raw data are available from our web site. The pooled beta values of the six samples were considered to assess the prediction performance. For each sample and each locus, the methylation state was assigned as described above. For each $l$ in $[0, 0.5]$, the proportions of correctly classified loci $p_M(l)$ and $p_U(l)$ are computed as follows: let $(\beta^{\mathrm{met}}_i)_{i=1,\ldots,N}$ and $(\beta^{\mathrm{RCA}}_i)_{i=1,\ldots,N}$ be, respectively, the beta values for the methylated and the unmethylated samples. Then,

$$p_M(l) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\beta^{\mathrm{met}}_i \geq 0.5 + l\right),$$

$$p_U(l) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\beta^{\mathrm{RCA}}_i \leq 0.5 - l\right),$$

where $\mathbb{I}(\cdot)$ is the indicator function. The global performance criterion is defined by averaging both previous proportions:

$$\mathrm{GC}(l) = \frac{1}{2}\left(p_M(l) + p_U(l)\right).$$

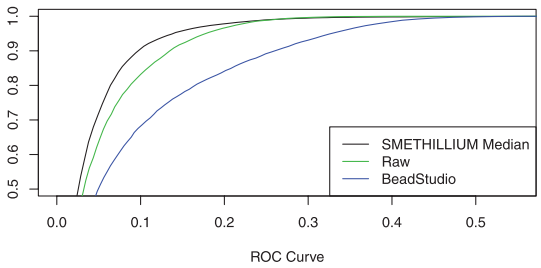


**Fig. 3.** ROC curves for the pooled six samples.

Figure 1 shows the global performance criterion on the beta values calculated with the SMETHILLIUM procedure, the BeadStudio procedure and also the raw data (beta values computed without any normalization and with intensities summarised using the mean). Figure 3 illustrates the receiver operating characteristic (ROC) curves obtained by the same three methods.

### 2.6 Results

Figures 1, 2 and 3 shows that the spatial normalization allows a better separation between the methylation state than the Illumina's proprietary software BeadStudio (Illumina, 2008). This is confirmed by the global criterion in Figure 1. Our method allows a better concentration of the beta values around their expected value allowing the possibility to better identify intermediate state such as hemi-methylation. For a $l$ threshold of 0.35, the global performance criterion is 80% after SMETHILLIUM normalization, 53% with BeadStudio and 60% without normalization (raw).

## 3 CONCLUSION

We developed a non-parametric spatial normalization method that improves the signal-to-noise ratio allowing a more reliable prediction of the methylation state as compared with BeadStudio. The method is implemented in R and publicly available. We recommend the user to use in-house experiments in order to define the most optimal $l$ threshold for their data.

## ACKNOWLEDGEMENT

The authors are very grateful to Pierre Gestraud and Jonas Mandel for their helpful comments, suggestions and observations.

*Funding*: P.H. is member of the team *Systems Biology of Cancer* and F.R. is member of the team *Oncologie Molculaire*, *équipes labellisées par la Ligue Nationale Contre le Cancer*. The project was supported by the Institut Curie translational department.

*Conflict of Interest*: none declared.

## REFERENCES

Davis,S. and Bilke,S. (2010) An introduction to the methylumi package. *Bioconductor package*.

Edwards,D. (2003) Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics*, **19**, 825–833.

Fan,J. and Gijbels,I. (1996) *Local Polynomial Modelling and Its Applications.* Chapman and Hall.

Illumina (2008) *Beadstudio Methylation Module v3.2, User Guide*. Illumina Inc., San Diego, CA, USA.

Stone,C.J. (1984) An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, **12**, 1285–1297.

Weisenberger,D.J. *et al.* (2008) Comprehensive DNA methylation analysis on the illumina infinium assay platform. *Illumina, Inc*.