Journal of Bioinformatics and Computational Biology
Vol. 12, No. 6 (2014) 1442005 (17 pages)
© Imperial College Press
DOI: 10.1142/S0219720014420050

Imperial College Press
www.icpress.co.uk

AKSmooth: Enhancing low-coverage bisulfite sequencing data via kernel-based smoothing

Junfang Chen[*,†,‡], Pavlo Lutsik[†], Ruslan Akulenko[*],
Jörn Walter[†,§,††] and Volkhard Helms[*,**,††]

[*]*Center for Bioinformatics, Saarland University*
Saarbrücken 66123, Germany

[†]*Department of Genetics, Saarland University*
Saarbrücken 66123, Germany
[‡]*s9juchen@stud.uni-saarland.de*
[§]*j.walter@mx.uni-saarland.de*
[**]*volkhard.helms@bioinformatik.uni-saarland.de*

Received 14 July 2014
Revised 12 September 2014
Accepted 12 September 2014
Published 26 December 2014

Whole-genome bisulfite sequencing (WGBS) is an approach of growing importance. It is the only approach that provides a comprehensive picture of the genome-wide DNA methylation profile. However, obtaining a sufficient amount of genome and read coverage typically requires high sequencing costs. Bioinformatics tools can reduce this cost burden by improving the quality of sequencing data. We have developed a statistical method Ajusted Local Kernel Smoother (AKSmooth) that can accurately and efficiently reconstruct the single CpG methylation estimate across the entire methylome using low-coverage bisulfite sequencing (Bi-Seq) data. We demonstrate the AKSmooth performance on the low-coverage ($\sim 4\times$) DNA methylation profiles of three human colon cancer samples and matched controls. Under the best set of parameters, AKSmooth-curated data showed high concordance with the gold standard high-coverage sample (Pearson 0.90), outperforming the popular analogous method. In addition, AKSmooth showed computational efficiency with runtime benchmark over 4.5 times better than the reference tool. To summarize, AKSmooth is a simple and efficient tool that can provide an accurate human colon methylome estimation profile from low-coverage WGBS data. The proposed method is implemented in R and is available at https://github.com/Junfang/AKSmooth.

Keywords: DNA methylation; whole-genome bisulfite sequencing; read coverage.

1. Introduction

DNA methylation is a biochemical modification of cytosine or adenine nucleotide bases. In mammals, DNA methylation occurs primarily in symmetrical CpG dinucleotide context[1,2] and 60% to 90% of all CpGs are methylated.[3,4] CpG islands are

[††]Corresponding authors.

genomic regions with a high CpG density that are usually hypomethylated. During the last decades, it has become clear that DNA methylation plays an essential role in down-regulating tumor-suppressor genes,[5] X-chromosome inactivation,[6] genomic imprinting,[7] aging[8] and the development of cancer.[9–11]

Previous studies found that the methylation levels of consecutive CpGs are correlated with each other. It has been reported that CpGs within a rather small genomic region are either methylated or unmethylated in a coordinated manner,[12,13] so a single-CpG methylation can be used to estimate the methylation status of its surrounding regions. More precisely in terms of genomic distances, a strong correlation in normal tissues for methylation of isolated CpGs in the range of 1 kb and 2 kb was detected[14] but the correlation decreased markedly when genomic distances were larger than 2 kb. Therefore, the underlying spatial dependence could be used to predict methylation levels of neighboring CpG sites that are not covered by reads and avoid missing methylation loci.[15] Furthermore, even the methylation patterns can be predicted based on the specific genomic sequence. Das and colleagues[16] tested the effect of window size on the methylation dependency with respect to the methylated sequences using support vector machines. This showed that methylation status is dependent within an 800 bp region centered on a CpG dinucleotide, relying on specific sequence features. Several bioinformatics tools have been developed which make use of this prior information, such as BSmooth,[17] BiSeq,[15] and CHARM.[18]

Over the past years a number of next generation sequencing (NGS) technologies have been developed to detect DNA methylation. Bisulfite sequencing (Bi-Seq) is still regarded as the gold standard to study whole-genome DNA methylation at single nucleotide resolution.[19] For this, the genomic DNA is specifically treated with sodium bisulfite, which leads to the selective deamination of cytosine residues to uracils and leaves 5-methylcytosine residues unchanged. Uracils are recognized as thymines by the sequencer. The main advantage of Bi-Seq compared with enzyme-based and enrichment-based methods is the high resolution information on the DNA methylation status.[20,21]

Reduced Representation Bi-Seq (RRBS),[22] Bisulfite Padlock Probes (BSPP),[23] and whole-genome bisulfite sequencing (WGBS)[24,25] are currently the most frequently used Bi-Seq-based methods to detect DNA methylation levels at single base-pair resolution. RRBS and BSPP involve much lower costs and have a higher read coverage since they sequence only a selected region of the genome. In contrast, WGBS needs more sample input and is usually very costly. However, WGBS is becoming increasingly popular and favorable due to its superior genome coverage (more than 90% of ~ 28 million CpGs in the human genome are covered) and the comprehensiveness of the methylome. Yet, the assessment of the genomic distribution of tissue-specific DNA methylation profiles is still limited in accuracy and completeness.[26]

The downside of WGBS is that at least $30\times$ coverage[27] is typically required in order to obtain a high-quality DNA methylation profile. Thus, it is very costly for routine studies of many samples and especially for a large methylome (e.g. human). If read coverage is not sufficient, potentially methylated CpGs in the genome may be missed[20]: Particularly CpGs located in genomic regions such as CpG island shores, genomic blocks or generic 2 kb regions. In order to avoid losing this information, sequencing machines with higher read coverage output, enabling input of more DNA samples, skilled personnel, and ideal experimental protocols are needed. All these findings and problems create a strong motivation to explore new strategies. To date, several tools have been developed for the alignment of Bi-Seq reads and methylation calling, such as BS-Seeker2,[28] BSMAP,[29] Bismark,[30] Bowtie 2,[31] and MethylCoder.[32] However, few bioinformatics tools or statistical methods exist that specifically deal with low-coverage Bi-Seq data.

It has been shown that results of comparable accuracy can be obtained even when using low-coverage data from modestly powerful sequencers.[17] BSmooth is the first tool that predicts methylation profiles accurately and precisely using low-coverage WGBS data and detects differentially methylated regions (DMRs) taking biological variability into account.[17] By applying BSmooth on low-coverage WGBS data, the authors compared the smoothed result to the matched high-coverage capture data with coverage $\geq 30\times$. When comparing the single-CpG methylation values from smoothed low-coverage WGBS data to that of the high-coverage data, a striking quantitative agreement was reported.

The BiSeq tool[15] applied a weighted local likelihood with a triangular kernel to smooth raw methylation data sample-wise. A small bandwidth (80 bp) within each CpG cluster is chosen for smoothing, and good results for detecting DMRs were obtained. However, the method is tailored for targeted Bi-Seq data (RRBS). A novel strategy (CHARM)[18] was described for array-based DNA methylation analysis of genome-weighted averaging from nearby genomic locations and the benefit of smoothing was demonstrated. But one should be aware of the fact that there is a relatively large difference between microarray-based and bisulfite sequencing-based measurements at the individual CpG level.

The aim of the present approach AKSmooth is to reconstruct the single CpG methylation profile of the full human methylome using low-coverage Bi-Seq data. The remainder of this work is organized as follows. Section 2 introduces the preparation of the experimental data. Then, we describe why and how the method is developed as well as the validation approaches. In Sec. 3, we apply AKSmooth to predict methylation data and compare the results with those obtained from BSmooth. In the last section, we conclude with a brief discussion of our results.

2. Materials and Methods

2.1. *DNA methylation data*

Two published Bi-Seq read data sets were used: One low-coverage WGBS raw data set and one capture bisulfite sequencing (Capture Bi-Seq) raw data set.[33] They were deposited in NCBI SRA with the accession number SRA036589. The low-coverage WGBS raw data were obtained from three colon cancer samples and their matched normal mucosa. Shotgun bisulfite whole-genome sequencing was performed on these samples using the ABI SOLiD platform, which produced 50 bp single-end reads. The data was collected in the laboratory of AP Feinberg.[33] The capture Bi-Seq raw data are considered as the gold standard, and will be used here to identify the accuracy of the methylation values estimated by statistical models. The data was prepared using a BSPP capture protocol on the same six samples as WGBS raw data, which were then sequenced on an Illumina GA II instrument yielding up to 80 base-pair single-end reads in approximately 40,000 capture regions. The capture regions can vary slightly for each sample.

Mapping color space Bi-Seq data from SOLiD requires a special algorithm for the bisulfite C/T conversion transforming two adjacent colors into 16 possible combinations. The low-coverage WGBS reads were mapped using the Merman aligner.[17] The alignment of the low-coverage WGBS reads was already done.[33] We used the tool Bismark[30] to perform the alignment of capture Bi-Seq reads. After mapping all the reads and extracting the DNA methylation data, we had to perform an additional step to match the genomic coordinates of the captured regional methylated CpG sites and low-coverage whole methylome. In doing so, it is possible to compare the smoothed low-coverage methylation data and high-coverage captured data. Therefore, according to the genomic information from both data sets, we could generate the matched genomic coordinate and matched high-coverage data set. In the end, the validation set was prepared (with $\geq 30\times$ coverage) from the matched high-coverage data set. The CpG sites on the sex chromosome were excluded since their proportion after matching is too small and also because the methylation behavior of sex chromosomes is quite different from other chromosomes.[34]

2.2. *Ajusted local kernel smoother (AKSmooth)*

The conventional wisdom is that more than 30× coverage is required to achieve an accurate DNA methylation profile in a genome region of interest or the whole genome. Capture Bi-Seq data in this work has a coverage of 30× on average. Thus it can serve as the gold standard. The WGBS data could only generate an average coverage of 4×, which provides only an unreliable estimate of the DNA methylation profile. Further analysis of DNA methylation based on this unprocessed information will be problematic due to the unreliable estimate. For this reason, statistical methods can be employed to improve the precision of the methylation estimates from the low-coverage Bi-Seq data.[17]

2.2.1. *Core idea*

When looking at low-coverage WGBS data, we need to pay attention to two major characteristics of the data. First, methylation levels of isolated CpGs within 1000 bps are often spatially dependent on each other.[14] As the average physical distance between two adjacent neighboring CpG dinucleotides lies in the range from tens to hundreds of base pairs, the methylation of CpG may be correlated with up to a hundred downstream and upstream CpGs. The second significant feature is the coverage of CpGs. For the data analyzed here, we observed that some of the CpGs have high-coverage but most of the CpGs possess only little coverage. Generally speaking, the larger the coverage of a CpG, the more reliable the methylation estimation. With these two important characteristics in mind, a novel smoothing method is developed below.

For a single sample, we denote the observed value y_i as the methylation ratio for the ith CpG in a given chromosome. The methylation value $\hat{f}_h(t)$ of the tth target CpG can be estimated using the following smoothing model, which is based on the Nadaraya–Watson estimator[35,36]:

$$\hat{f}_h(t) = \frac{\sum_i^N K_h(t,i) C_t(i) y_i}{\sum_i^N K_h(t,i) C_t(i)}, \tag{1}$$

where

$$K_h(t,i) = K\left(\frac{|i-t|}{h}\right), \tag{2}$$

$$C_t(i) = \begin{cases} g_t & \text{if } i = t; \\ 1 & \text{if } i \neq t. \end{cases} \tag{3}$$

We assume $\hat{f}_h(t)$ to be a smoothing function that can predict the methylation value of the tth target CpG site within a predefined bandwidth h, which is computed explicitly by the right hand side (RHS) expression. K is the kernel function that assigns different weights to the neighbors. The numerator of the RHS contains the summation of every weighted neighboring CpG value within a certain bandwidth h. In particular, all the neighboring CpGs receive the same coverage weight 1, whereby the target point is multiplied on purpose by $C_t(i)$. $C_t(i)$ is a coverage weighting function indicating the corresponding coverage of the ith CpG site. If $i = t$, $C_t(i)$ is defined by g_t, where g_t is the coverage of the tth CpG site. Otherwise, $C_t(i)$ is 1. The denominator sums all the weights and similarly the target weight should be multiplied by $C_t(i)$.

In this expression, local bi-weighted averaging is performed whereby the methylated points closer to the target point get higher weights. The attractive point is that the coverage information of each individual CpG (termed coverage weight) can also be accommodated. Higher coverage of a CpG will provide a relatively trustworthy methylation estimate.

2.2.2. *Choice of bandwidth*

In our work, the WGBS data for a single sample is univariate and assumed to have an evenly spaced design. Thus, a globally constant bandwidth is a favorable choice. Taking into account the methylation correlation of the neighboring CpGs and the smoothing window sizes suggested in the literature,[14,15,17] we have tested seven reasonable bandwidths for our experiment: 5, 10, 15, 20, 30, 50, and 70 CpGs.

2.2.3. *Choice of kernel*

The choice of the kernel weighting function plays an important role in defining the concept of AKSmooth. But it turned out that this choice has a weaker influence on the bias-variance tradeoff than the bandwidth. The aim of the kernel function is to estimate the target value from a limited set of neighboring values. The following three kernel functions were tested: The Gaussian kernel, the Epanechnikov, and tricubic kernel. This means that the kernel function

$$K_h(t,i) = D\left(\frac{|i-t|}{h}\right), \tag{4}$$

is either a standard Gaussian function

$$D(\mu) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{1}{2}\mu^2} \tag{5}$$

or the Epanechnikov kernel

$$D(\mu) = \begin{cases} \frac{3}{4}(1-\mu^2) & \text{if } |\mu| \le 1; \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

or the tricubic kernel

$$D(\mu) = \begin{cases} \frac{70}{81}(1-|\mu|^3)^3 & \text{if } |\mu| \le 1; \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

All three kernels yield larger weights to observations that are close to the target position. h determines the bandwidth of the neighborhood at the target CpG site t. The Gaussian kernel is a popular non-compact kernel with the standard deviation playing the role of the bandwidth. The Epanechnikov kernel and tricubic kernel are both commonly used compact kernels.

2.2.4. *AKSmooth algorithm*

The algorithm is described as follows:

AKSmooth Algorithm

Input:

(1) N: the total number of CpGs within one chromosome.
(2) y_i: the methylation ratio at the ith CpG site.
(3) $C_t(i)$: the coverage at the tth target CpG site if $i = t$, else $C_t(i) = 1$.
(4) K: the kernel function (Gaussian, tricubic or Epanechnikov).
(5) h: half the bandwidth.

Initialize:

u, l, and $\hat{f}_h$ are initialized as the zero vectors of length N.

```
Repeat:
For $t = 1$ to $N$:
  For $i = \max(t - h, 1)$ to $\min(t + h, N)$:
(1) Compute
      $u[t] = u[t] + K_h(t, i)C_t(i)y_i$,
      $l[t] = l[t] + K_h(t, i)C_t(i)$.
(2) Set
      $\hat{f}_h(t) = \frac{u[t]}{l[t]}$.
  Output:
    A vector $(\hat{f}_h(1), \ldots, \hat{f}_h(N))$ of length $N$ with the estimated methylation values.
```

2.3. *Model evaluation*

The Pearson correlation coefficient is used to assess the agreement between smoothed methylation data and the gold standard data. Assuming that two random variables F and Y are defined by a set of observations with f_i and y_i, where $i = 1, 2, \ldots, n$.

$$r = \frac{\sum_{i=1}^{n}((f_i - \bar{f})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{n}(f_i - \bar{f})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{8}$$

where $\bar{f}$ and $\bar{y}$ are the sample means of F and Y.

The concordance statistics is another well-known test for assessing the difference between the estimated and actual methylation profiles.[20] First, the difference between the methylation values in the smoothed data and high-coverage data is calculated for all n CpG sites present in both data sets. Then one can compute the percentage of sites whose discrepancy is smaller than 0.1 or 0.25.

The concordance statistics is mathematically defined as follows:

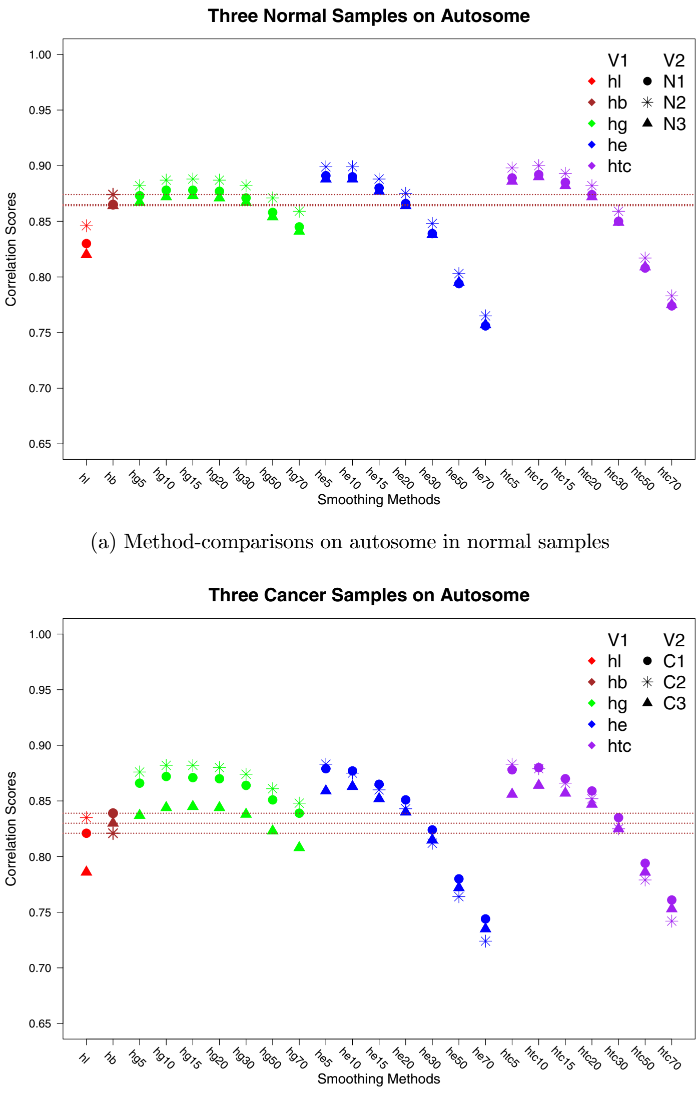
$$\phi_c = \frac{\sum_{i=1}^{n} \mathbb{1}_{|f_i - y_i|} \leq c}{n}, \tag{9}$$

where c is the cutoff 0.1 or 0.25.

3. Results

3.1. *Correlation measure*

We validated the smoothing methods on the basis of the global and regional methylation smoothing profiles, namely the autosomal and chromosomal methylation patterns. Using three different kernel functions and seven different window sizes, we



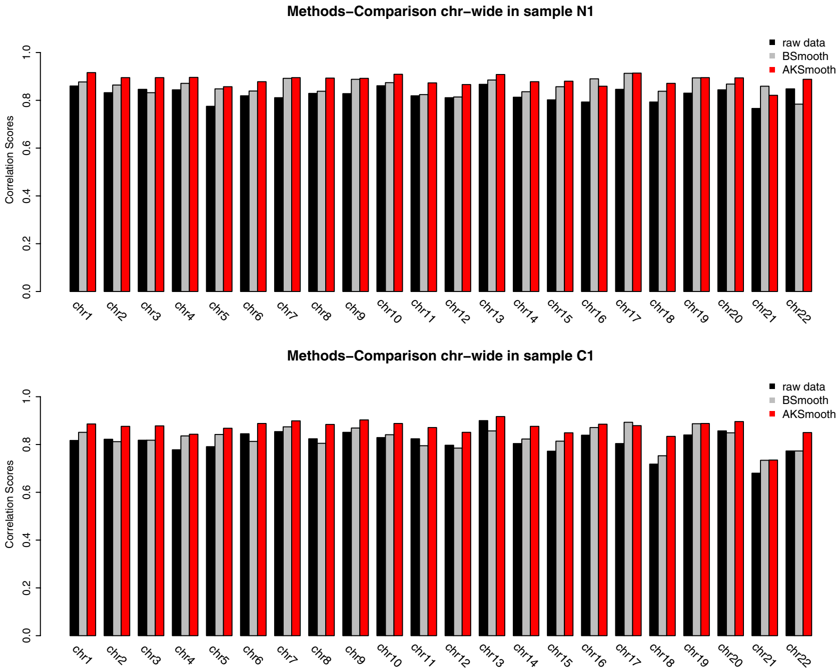
(a) Method-comparisons on autosome in normal samples

(b) Method-comparisons on autosome in cancer samples

Fig. 1. Method-comparisons on autosome with coverage $\geq 30\times$. Different symbols listed in the legend item V2 denote comparisons of different normal/cancer samples: Solid circles refer to normal/control sample N1/C1, the star symbols are used for sample N2/C2 and sample N3/C3 is marked as filled triangle. Different colors in the legend item V1 are used for different smoothing method-comparisons. The red symbol on the leftmost side is the correlation score between the gold standard data and the raw data of three normal/cancer samples. BSmoothed data and the gold standard data are compared using brown colors and the resulting scores are also marked by straight brown lines. Green, blue, and purple are the colors for three different kernel weighting functions of AKSmooth method: Gaussian, Epanechnikov, and tricubic kernels, respectively. Seven different window sizes are tested for each type of kernel function: 5, 10, 15, 20, 30, 50, and 70 CpGs. hl: high versus low-coverage data, hb: high versus BSmooth, hg: high versus AKSmooth with the Gaussian kernel, he: high versus AKSmooth with the Epanechnikov kernel, htc: high versus AKSmooth with the tricubic kernel.

measured the Pearson correlation coefficient between the gold standard data (high-coverage matched data) and raw unsmoothed data (low-coverage matched data), between the gold standard data and BSmoothed data with a recommended window size, and between the gold standard data and AKSmoothed data.

The comparison results for the autosomal methylation estimated profile are shown in Fig. 1. According to Fig. 1(a), AKSmooth performs better than BSmooth with all three kernels when a rather small window size is used. The highest correlation score of BSmooth is 0.87, while the four largest values obtained with AKSmooth reach 0.9. The best weighting functions are the Epanechnikov and tricubic kernels, where the top bandwidths are 5 and 10 CpGs. Interestingly, correlation values of AKSmooth with these three kernels show a similar tendency. They increase slightly at first, then fall gradually when the bandwidth becomes larger. The results obtained for cancer samples are shown in Fig. 1(b). We observed that BSmooth is not able to estimate the methylation in cancer sample C2 at all. The Gaussian kernel (bandwidth 5, 10 CpGs) followed by the Epanechnikov and the tricubic kernel with the same window size 5 CpGs provide the highest correlation coefficients (0.88), whereas the maximum value from BSmooth is 0.84.


Methods−Comparison chr−wide in sample N1

Methods−Comparison chr−wide in sample C1

Fig. 2. Chromosome-wide method-comparisons in samples N1 and C1. Three different comparisons are represented using three different colors. The comparison between the raw data and the high-coverage data is shown using black color. The gray color is the comparison between BSmoothed data and the high-coverage data. The red color shows the comparison between AKSmoothed data and the high-coverage data.

It turned out that AKSmooth with appropriate parameters is capable to reliably reconstruct the low-coverage methylation data on an autosome-wide scale for both normal and cancer samples, BSmooth also gives a rather good prediction but of slightly less accuracy than AKSmooth.

In Fig. 2, AKSmooth using the Epanechnikov kernel with a bandwidth of 5 CpGs performs better than BSmooth on most chromosomes in both normal (20 out of 22) and cancer samples (21 out of 22). The only exceptions to this are chromosomes 16 and 21 in normal sample and chromosome 17 in the cancer case. In general, the correlation scores obtained in the cancer sample are smaller than those of the control sample.

Therefore, we conclude that AKSmooth predicts methylation levels either globally or locally more accurately than BSmooth across the three pairs of normal/cancerous colon samples.

3.2. *Concordance statistic*

The concordance statistics with the cutoff 0.25 was evaluated between smoothed methylation estimates and the gold standard data. As shown in Fig. 3, AKSmooth

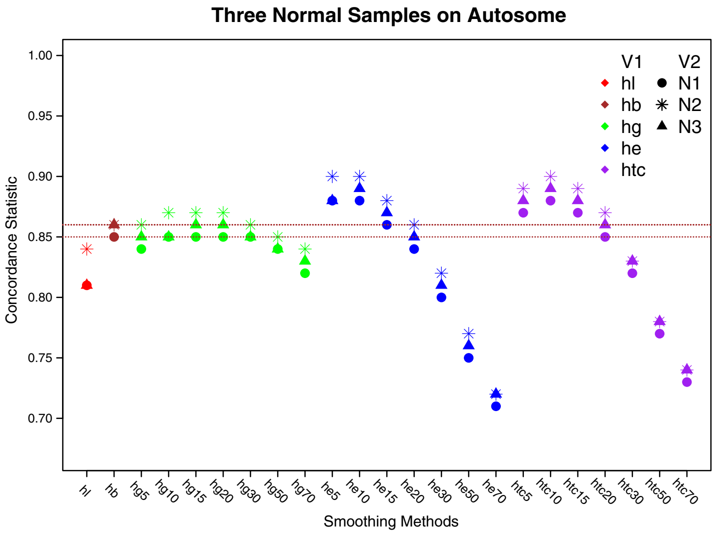


Fig. 3. Concordance statistics on autosome in normal samples. The symbols in the legend item V2 are used for the different normal samples: Solid circles refer to normal sample N1, the stars are used for normal sample N2, and normal sample N3 are labeled as filled triangles. Different colors in the legend item V1 are representing different smoothing method-comparisons. The red symbol on the leftmost side denotes the concordance scores between the gold standard data and the raw data. BSmoothed data and the gold standard data are compared using brown colors and the resulting scores are also marked by straight brown lines. Green, blue, and red colors stand for three different kernel weighting functions of the AKSmooth method, namely the Gaussian, Epanechnikov, and tricubic kernels, respectively. Seven different window sizes are tested for each type of kernel function: 5, 10, 15, 20, 30, 50, and 70 CpGs. hl: high versus low-coverage data, hb: high versus BSmooth, hg: high versus AKSmooth with the Gaussian kernel, he: high versus AKSmooth with the Epanechnikov kernel, htc: high versus AKSmooth with the tricubic kernel.

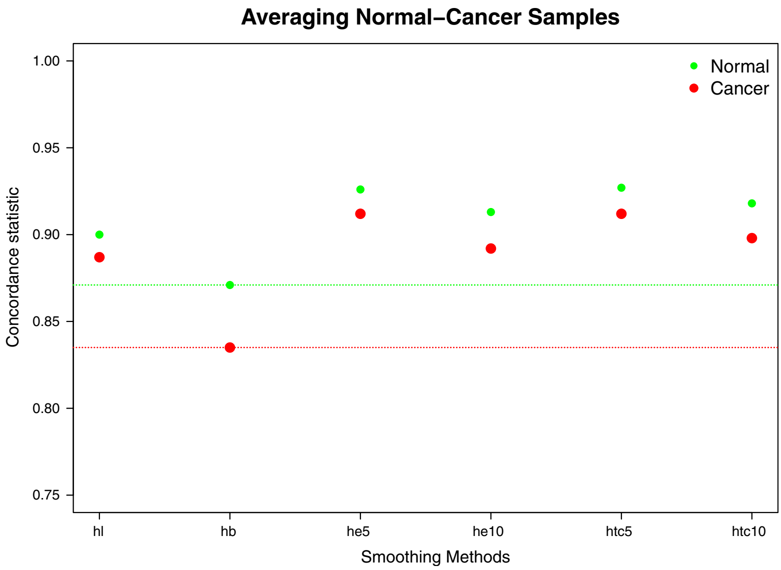


Fig. 4. Concordance test of averaging normal-cancer samples. hl: high versus low-coverage data, hb: high versus BSmooth, he: high versus AKSmooth with the Epanechnikov kernel, htc: high versus AKSmooth with the tricubic kernel. Here AKSmooth either used a bandwidth 5 or 10 CpGs. The high coverage data with coverage $\geq 30\times$ was used.

using the Epanechnikov and tricubic kernels with a small bandwidth reached a higher concordance, whereas the Gaussian kernel was only slightly better than BSmooth. Notably, about 90% of the AKSmoothed data from normal sample N2 was consistent with the true data of $\geq 30\times$ coverage, whereas BSmooth approached 86%. According to this concordance rate, the optimal parameters of AKSmooth were the Epanechnikov and tricubic kernels with rather small window sizes ranging from 5 to 15 CpGs. This agrees with the outcome from the correlation analysis. Similar results were obtained for the cancer case as well as the regional smoothing.

Furthermore, we averaged across three normal samples and the matched cancer samples after smoothing. Figure 4 demonstrated a remarkable concordance (92.7%) between AKSmoothed estimates using the tricubic kernel with a window size of 5 CpGs and the matched high coverage data in normal samples, where the result for BSmooth was 87.1%, for the raw data 90.0%. In the cancer case, we found that the difference between AKSmoothed data and the raw data was not large. The tricubic kernel with a bandwidth of 5 CpGs was 91.2% concordant with the gold standard, whereby the raw data shows 88.7% and BSmooth gave only 83.5% of the concordance. This implies that averaging the biological replicates would improve the concordance. The overall concordance rate in cancer is lower than that in the control sample which may be due to the higher variability in cancer cells.[33,37]

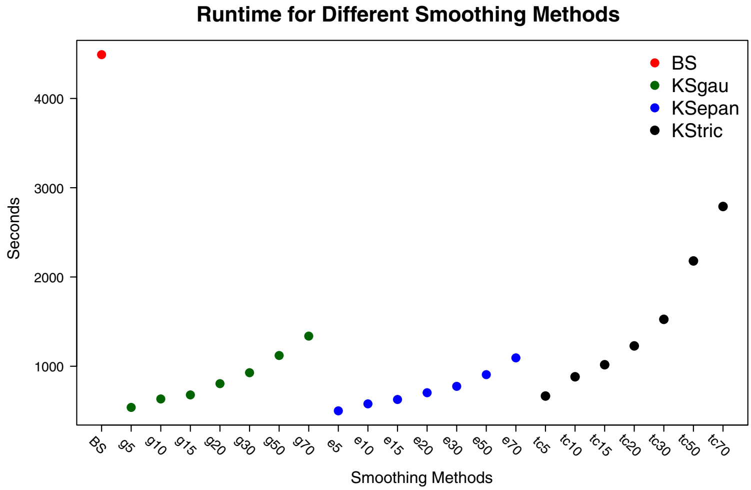


Fig. 5. Runtime performance of smoothing methods. BS: BSmooth, KSgau: AKSmooth with the Gaussian kernel, KSepan: AKSmooth with the Epanechnikov kernel, KStrc: AKSmooth with the tricubic kernel.

3.3. *Computation runtime*

In the end, we evaluated the runtime performance of BSmooth and AKSmooth. The test was executed on a single core of a server with (Intel (R) Xeon(R) CPU E5-4620 @ 2.20 GHz, 32 cores and 128 GB RAM). The results are displayed in Fig. 5. Both smoothing methods are implemented in R, but due to its more efficient algorithm, AKSmooth is approximately 4 to 5 folds faster than BSmooth on the whole methylome data set.

4. Discussion

We developed a new statistical method that turned out to predict a fairly accurate whole human colon methylome profile using low-coverage WGBS data. BSmooth employs a model that is known as locally weighted polynomial regression, whereas AKSmooth is a modified version of the Nadaraya–Watson estimator. On the one hand, the bandwidth plays a big role in estimating single CpG methylation levels. Similar to BSmooth, AKSmooth also takes the spatial dependence of neighboring CpGs into consideration. The methylated points closer to the target point have higher kernel weights. BSmooth uses a window size of up to 70 CpGs but at least 2 kb wide. This is useful when regional methylation level prediction is implemented, such as in CpG clusters. AKSmooth considers smaller window sizes, in which the methylation depends more strongly on the neighboring CpGs.[14] On the other hand, AKSmooth is able to accommodate the coverage information in a way that CpGs with higher coverage ($\geq 30\times$) from low-coverage data show relatively closer methylation values to those of the high-coverage data.

When compared to the performance of BSmooth, AKSmooth worked better in terms of correlation coefficients, concordance test, and runtime. First, correlation coefficients were computed for method comparison. AKSmooth using the tricubic kernel with a bandwidth of 10 CpGs and the Epanechnikov kernel (both 5 and 10 CpGs) showed the largest correlations (0.900, 0.899) on autosomes in normal sample N2, whereas the results of BSmooth and the raw data were 0.874 and 0.846, respectively. In the matched cancerous sample, the best scores (0.883, 0.883) were achieved by the Epanechnikov and tricubic kernel with the same bandwidth of 5 CpGs, whereas BSmooth and the raw data obtained 0.821 and 0.835, respectively.

We observed that AKSmooth using the Epanechnikov and tricubic kernel with window sizes ranging from 5 to 10 CpGs gave very good results in both matched normal and cancerous samples when compared to the high-coverage data. Moreover, the Gaussian kernel with the same bandwidth made the results better than those from BSmooth only on certain occasions, suggesting that a non-compact kernel is not as favorable as the compact ones for a Gaussian might also include more noisy data. On the other hand, the window size apparently plays a much more vital role in estimating methylation profiles than the kernel function. With increasing window size the correlation coefficients experienced a significant drop. The reason is likely that methylation levels of consecutive CpGs are correlated with only a limited number of neighbors. Besides, we found that AKSmooth with window sizes from 5 up to even 20 CpGs yielded a good outcome and outperformed BSmooth. This range of bandwidths is meaningful since it is reasonable to infer that the methylation levels of neighboring CpGs are correlated in a similar manner across the whole chromosome and the entire genome.

Second, the concordance statistics was used to test the performance of smoothing methods on the autosomes of the normal sample N2 and the cancer sample C2. The best smoothing results using normal sample N2 were provided by AKSmooth with the Epanechnikov kernel (5, 10 CpGs) and the tricubic kernel (10 CpGs). All of the resulting concordances were 90%, which was 4% and 6% better than with BSmooth (86%) and for the raw low-coverage data (84%). In the matched cancer case, the best concordances (87%, 87%) were obtained by the Epanechnikov kernel (5 CpGs) and the tricubic kernel (5 CpGs). In contrast, the best agreement obtained by BSmooth and the raw data were 77% and 82%, respectively. The findings from the concordance test are very similar to those of the correlation measure. Additionally, we also took the biological variability into account, which means that after sample-wise smoothing the autosomal methylation estimates across three normal samples and the matched cancer samples were averaged, respectively. The normal sample demonstrated a high concordance (92.7%) computed by AKSmooth with the tricubic kernel and a window width 5 CpGs, whereas BSmooth and the raw data showed concordance rates of 87.1% and 90.0%, respectively. In cancer, the concordance by the tricubic kernel with a bandwidth of 5 CpGs was 91.2%, while the results from BSmooth and the raw data were 83.5% and 88.7%, respectively. Overall, the concordance ratio in cancer was lower than in the control sample, which may be due to the higher variability in cancerous cells. Averaging the biological replicates could globally improve the concordance and is therefore recommended.

Lastly, we evaluated the runtime performance of BSmooth and AKSmooth methods. It turned out that AKSmooth is 4.5 times more efficient than BSmooth in algorithmic design.

In summary, our results showed that kernel-based smoothing is able to reconstruct the DNA methylome based on low-coverage Bi-Seq data with high accuracy, allowing to improve the sequencing data of suboptimal quality and optimize the sequencing costs. We implemented our method as an open-source R-package AKSmooth, which is freely available at https://github.com/Junfang/AKSmooth.

Acknowledgment

We would like to thank Dr. Ye Liang for the valuable comments on the manuscript.

Supplementary Information

Supplementary materials are available.

References

1. Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V, Bird AP, Jaenisch R, Non-cpg methylation is prevalent in embryonic stem cells and may be mediated by dna methyltransferase 3a, *Proc Nat Acad Sci* **97**(10):5237–5242, 2000.
2. Ziller MJ, Müller F, Liao J, Zhang Y, Gu H, Bock C, Boyle P, Epstein CB, Bernstein BE, Lengauer T *et al.*, Genomic distribution and inter-sample variation of non-cpg methylation across human cell types, *PLoS Genetics* **7**(12):e1002389, 2011.
3. Ehrlich M, Gama-Sosa MA, Huang LH, Midgett RM, Kuo KC, McCune RA, Gehrke C, Amount and distribution of 5-methylcytosine in human dna from different types of tissues or cells, *Nucleic Acids Res* **10**(8):2709–2721, 1982.
4. Tucker KL, Methylated cytosine and the brain: A new base for neuroscience, *Neuron* **30**(3):649–652, 2001.
5. Jones PA, Laird PW, Cancer-epigenetics comes of age, *Nat Genetics* **21**(2):163–167, 1999.
6. Panning B, Jaenisch R, RNA and the epigenetic regulation of x chromosome inactivation, *Cell* **93**(3):305–308, 1998.
7. Li E, Beard C, Jaenisch R, Role for DNA methylation in genomic imprinting, *Nature* **366**(6453):362–365, 1993.
8. Richardson B, Impact of aging on DNA methylation, *Ageing Res Rev* **2**(3):245–261, 2003.
9. Costello JF, Frühwald MC, Smiraglia DJ, Rush LJ, Robertson GP, Gao X, Wright FA, Feramisco JD, Peltomäki P, Lang JC *et al.*, Aberrant cpg-island methylation has non-random and tumour-type–specific patterns, *Nat Genetics* **24**(2):132–138, 2000.
10. Davis CD, Uthus EO, DNA methylation, cancer susceptibility, and nutrient interactions, *Exp Biol Med* **229**(10):988–995, 2004.
11. Esteller M, Cancer epigenomics: DNA methylomes and histone-modification maps, *Nat Rev Genetics* **8**(4):286–298, 2007.
12. Hatada I, Hayashizaki Y, Hirotsune S, Komatsubara H, Mukai T, A genomic scanning method for higher organisms using restriction sites as landmarks, *Proc Nat Acad Sci* **88**(21):9523–9527, 1991.

13. Smiraglia DJ, Kazhiyur-Mannar R, Oakes CC, Wu YZ, Liang P, Ansari T, Su J, Rush LJ, Smith LT, Yu L *et al.*, Restriction landmark genomic scanning (rlgs) spot identification by second generation virtual rlgs in multiple genomes with multiple enzyme combinations, *BMC Genomics* **8**(1):446, 2007.
14. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA *et al.*, DNA methylation profiling of human chromosomes 6, 20 and 22, *Nat Genetics* **38**(12):1378–1385, 2006.
15. Hebestreit K, Dugas M, Klein HU, Detection of significantly differentially methylated regions in targeted bisulfite sequencing data, *Bioinformatics* **29**(13):1647–1653, 2013.
16. Das R, Dimitrova N, Xuan Z, Rollins RA, Haghighi F, Edwards JR, Ju J, Bestor TH, Zhang MQ, Computational prediction of methylation status in human genomic sequences, *Proc Nat Acad Sci* **103**(28):10713–10716, 2006.
17. Hansen KD, Langmead B, Irizarry RA *et al.*, Bsmooth: From whole genome bisulfite sequencing reads to differentially methylated regions, *Genome Biol* **13**(10):R83, 2012.
18. Irizarry RA, Ladd-Acosta C, Carvalho B, Wu H, Brandenburg SA, Jeddeloh JA, Wen B, Feinberg AP, Comprehensive high-throughput arrays for relative methylation (charm), *Genome Res* **18**(5):780–790, 2008.
19. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL, A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands, *Proc Nat Acad Sci* **89**(5):1827–1831, 1992.
20. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao Y *et al.*, Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications, *Nat Biotechnol* **28**(10):1097–1105, 2010.
21. Sandoval J, Esteller M, Cancer epigenomics: Beyond genomics, *Curr Opin Genetics Development* **22**(1):50–55, 2012.
22. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB *et al.*, Genome-scale DNA methylation maps of pluripotent and differentiated cells, *Nature* **454**(7205):766–770, 2008.
23. Diep D, Plongthongkum N, Gore A, Fung HL, Shoemaker R, Zhang K, Library-free methylation sequencing with bisulfite padlock probes, *Nat Meth* **9**(3):270–272, 2012.
24. Laird PW, Principles and challenges of genome-wide DNA methylation analysis, *Nat Rev Genetics* **11**(3):191–203, 2010.
25. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM *et al.*, Human DNA methylomes at base resolution show widespread epigenomic differences, *Nature* **462**(7271):315–322, 2009.
26. Bock C, Analysing and interpreting DNA methylation data, *Nat Rev Genetics* **13** (10):705–719, 2012.
27. Stevens M, Cheng JB, Li D, Xie M, Hong C, Maire CL, Ligon KL, Hirst M, Marra MA, Costello JF *et al.*, Estimating absolute methylation levels at single-cpg resolution from methylation enrichment and restriction enzyme sequencing methods, *Genome Res* **23** (9):1541–1553, 2013.
28. Guo W, Fiziev P, Yan W, Cokus S, Sun X, Zhang MQ, Chen PY, Pellegrini M, Bs-seeker2: A versatile aligning pipeline for bisulfite sequencing data, *BMC Genomics* **14**(1):774, 2013.
29. Xi Y, Li W, Bsmap: Whole genome bisulfite sequence mapping program, *BMC Bioinformatics* **10**(1):232, 2009.
30. Krueger F, Andrews SR, Bismark: A flexible aligner and methylation caller for bisulfite-seq applications, *Bioinformatics* **27**(11):1571–1572, 2011.

31. Langmead B, Salzberg SL, Fast gapped-read alignment with bowtie 2, *Nat Meth* **9**(4):357–359, 2012.
32. Pedersen B, Hsieh TF, Ibarra C, Fischer RL, Methylcoder: Software pipeline for bisulfite-treated sequences, *Bioinformatics* **27**(17):2435–2436, 2011.
33. Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D *et al.*, Increased methylation variation in epigenetic domains across cancer types, *Nat Genetics* **43**(8):768–775, 2011.
34. Bernardino J, Lombard M, Niveleau A, Dutrillaux B, Common methylation characteristics of sex chromosomes in somatic and germ cells from mouse, lemur and human, *Chromosome Res* **8**(6):513–525, 2000.
35. Nadaraya EA, On estimating regression, *Theory Probab & Appl* **9**(1):141–142, 1964.
36. Watson GS, Smooth regression analysis, *Sankhyā: The Indian J Stat Series A* **26**(4):359–372, 1964.
37. Schoofs T, Rohde C, Hebestreit K, Klein HU, Göllner S, Schulze I, Lerdrup M, Dietrich N, Agrawal-Singh S, Witten A *et al.*, DNA methylation changes are a late event in acute promyelocytic leukemia and coincide with loss of transcription factor binding, *Blood* **121**(1):178–187, 2013.