BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 13 2013, pages 1647–1653
doi:10.1093/bioinformatics/btt263

Gene expression

Advance Access publication May 8, 2013

Detection of significantly differentially methylated regions in targeted bisulfite sequencing data

Katja Hebestreit*, Martin Dugas and Hans-Ulrich Klein

Institute of Medical Informatics, University of Münster, Albert-Schweitzer-Campus 1, 48149 Münster, Germany

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Bisulfite sequencing is currently the gold standard to obtain genome-wide DNA methylation profiles in eukaryotes. In contrast to the rapid development of appropriate pre-processing and alignment software, methods for analyzing the resulting methylation profiles are relatively limited so far. For instance, an appropriate pipeline to detect DNA methylation differences between cancer and control samples is still required.

Results: We propose an algorithm that detects significantly differentially methylated regions in data obtained by targeted bisulfite sequencing approaches, such as reduced representation bisulfite sequencing. In a first step, this approach tests all target regions for methylation differences by taking spatial dependence into account. A false discovery rate procedure controls the expected proportion of incorrectly rejected regions. In a second step, the significant target regions are trimmed to the actually differentially methylated regions. This hierarchical procedure detects differentially methylated regions with increased power compared with existing methods.

Availability: R/Bioconductor package BiSeq.

Contact: katja.hebestreit@uni-muenster.de

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

Received on January 29, 2013; revised on April 10, 2013; accepted on May 2, 2013

1 INTRODUCTION

DNA methylation is an epigenetic modification regulating gene transcription and is known to direct development and differentiation (Gopalakrishnan *et al*., 2008). A CG dinucleotide (CpG site) is called methylated if a methyl group is attached to the cytosine (C). DNA treatment with sodium bisulfite specifically introduces conversion of unmethylated cytosine to uracil (then read as thymine by DNA polymerase), whereas methylcytosine remains unmodified (Krueger *et al*., 2012). These changes are mapped by next-generation sequencing. To save per-sample costs, bisulfite sequencing (BS) can be combined with enrichment strategies to target bisulfite sequencing to a specific fraction of the genome (Bock, 2012). Along with DNA fragment capture, it is possible to use restriction enzymes as for reduced representation bisulfite sequencing (RRBS), which measures genome-wide DNA methylation in CpG-rich regions (Meissner *et al*., 2005). Msp1 (cleaves at CCGG) digested fragments are size-selected to obtain fragments with short distance between Msp1 sites. Bisulfite-treated DNA fragments are then sequenced and mapped via conversion-aware aligners, such as Bismark (Krueger and Andrews, 2011), RRBSMAP (Xi *et al*., 2012), BS Seeker (Chen *et al*., 2010) or PASS-bs (Campagna *et al*., 2012). The aligners usually return a list of CpG sites with the number of cytosines and thymines among all reads aligned to the cytosine in the genomic DNA sequence, which is then used for subsequent analyses.

Aberrant DNA methylation patterns are a characteristic feature of cancer (Das and Singal, 2004). Often, studies focus on promoter regions or gene bodies or—more generally—genomic regions that are differentially methylated between cancer and normal specimens. In this context, a differentially methylated region (DMR) denotes a genomic region of adjacent CpG sites that are differentially methylated. Because of lower costs of RRBS or other targeted bisulfite sequencing approaches increasingly larger number of samples are measured, which enables a comparison of DNA methylation and the detection of DMRs between groups of samples (Schoofs *et al*., 2013).

In the past 3 years, several software tools for descriptive BS data analysis have been published. MethVisual (Zackay and Steinhoff, 2010) is an R/Bioconductor (Gentleman *et al*., 2004; R Core Team, 2012) package for visualization and exploratory statistical analysis of BS data. BiQ Analyzer HT (Lutsik *et al*., 2011) allows a locus-specific analysis and visualization. SAAP-RRBS (Sun *et al*., 2012) provides methylation summary statistics and annotation of CpG sites. Nevertheless, few tools to detect DMRs exist. The most common approach is to perform Fisher's exact test CpG-wise or region-wise (Challen *et al*., 2012; Gu *et al*., 2010; Li *et al*., 2010). Recently, BSmooth was published, a pipeline to detect differentially methylated regions in whole-genome BS data (Hansen *et al*., 2012). BSmooth basically relies on smoothing the methylation values sample-wise and then testing for group differences via CpG-wise *t*-tests. DMRs are defined as adjacent CpG sites with absolute *t* statistics above a defined threshold.

To assess the goodness of the results, a suitable error measure is needed. For example, it would be convenient to control the false discovery rate (FDR) on genomic regions, i.e. the expected proportion of regions rejected erroneously out of all regions rejected, testing the null hypothesis that the regions are not differentially methylated. However, the desire for a convenient analysis of BS data and the establishing of an error measure pose challenges specific to the analysis of BS data: first, the coverage, i.e. the number of reads spanning a CpG site varies widely between different CpG sites and different samples. As a consequence, each sample has an individual profile of covered

*To whom correspondence should be addressed.

© The Author 2013. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

CpG sites, and missing values occur frequently. Second, the specific probability distribution of methylation levels requires specific statistical models. Third, the methylation of neighbored CpG sites is spatially correlated and, hence, are the CpG-wise test statistics and *P*-values.

The first and to our knowledge only approach accounting for spatial dependence in multiple hypothesis testing is methylKit, an R package for the analysis of RRBS data and its variants (Akalin *et al.*, 2012). methylKit models the methylation per CpG site within a logistic regression. A sliding linear model (SLIM) method is used to determine q values from *P*-values to correct for multiple hypothesis testing (Wang *et al.*, 2011).

We propose BiSeq, a DMR detecting approach that enables testing for DMRs within target regions and controlling a given FDR. This approach is tailored to data received by all kinds of targeted bisulfite sequencing approaches.

The article is organized as follows. In Section 2, we give a rough overview of the step-by-step approach. The individual steps are presented in detail in Section 3. In Section 4, we apply the approach to simulated data and compare its results with those from BSmooth and methylKit. Furthermore, we apply BiSeq to a published leukemia dataset. In Section 5, we briefly discuss our results.

2 APPROACH

The analysis is confined to the target region. For RRBS, these are genomic regions with a high spatial density of covered CpG sites. These regions are called CpG clusters. The approach has two aims: first, the detection of CpG clusters with at least one differentially methylated CpG site. Second, the trimming of the differentially methylated CpG clusters to the actually differentially methylated CpG sites. In detail, BiSeq is a five-step approach:

(1) Define CpG clusters
(2) Smooth methylation data within CpG clusters
(3) Model and test group effect within CpG clusters
(4) Apply hierarchical testing procedure:
    (a) Test CpG clusters for differential methylation and control weighted FDR on clusters
    (b) Trim rejected CpG clusters and control FDR on single CpG sites
(5) Define DMR boundaries

3 METHODS

3.1 Data

We used RRBS data of bone marrow specimens of 18 patients with acute promyelocytic leukemia at diagnosis that was recently published by Schoofs *et al.* (2013). Additionally, there are 16 control samples composed of four samples of healthy CD34+cells, four samples of promyelocytes (generated *in vitro* from the CD34+cells) and eight remission bone marrow samples of matched patient samples. After MspI digestion, size selection (40–220 bp) and bisulfite conversion, the RRBS libraries were sequenced on an Illumina HiScanSQ instrument. Using Bismark version 0.5, sequencing reads were mapped to hg19 genome and methylation calls were extracted (Krueger and Andrews, 2011). On average, 1.17×10^7 reads were uniquely mapped per sample yielding coverage of almost 9.3×10^6 CpG sites of which 9.4×10^5 CpG sites were covered in all samples.

To obtain a dataset with known DMRs, we took 12 control samples (four remission, four CD34+and four promyelocyte) from the data described earlier in the text. Half of the samples (two remission, two CD34+and two promyelocyte) were kept as control samples. Within the other half, we simulated DMRs by placing methylation differences of various intensities and lengths, so that these could be considered as cancer samples. CpG island positions were downloaded from UCSC database (Goldman *et al.*, 2012) within which 5000 DMRs were placed. Because of the fact that DMRs reported in literature usually fall within widths of a few hundred to a few thousand base pairs (Bock, 2012), we sampled the DMR lengths l from a truncated Gaussian: $l \sim \mathcal{N}(100, 150^2)$, with $10 < l < 1000$. Each DMR was placed into a CpG island that was at least as long as the respective DMR with probabilities proportional to the CpG island length. Overlapping of simulated DMRs was allowed. Differences d were sampled from a truncated Gamma: $d \sim \Gamma(1.9, 0.08)$, with $0 < d < 1$, so that the resulting differences were positive. For a detailed description of how simulated DMRs were incorporated into the data, see the Supplementary Data. In Supplementary Figures S1 and S2, the distributions of the lengths and differences of the resulting 4 034 DMRs are shown. Within the 12 samples, almost 8×10^6 CpG sites were covered of which 1.9×10^6 CpG sites were covered in all samples.

3.2 CpG clusters

The RRBS method provides methylation information of CpG-rich regions primarily. For this reason and because of the spatial correlation of the methylation of nearby CpG sites, we constrain the analysis on CpG sites within *CpG clusters*. These CpG clusters are regions detected as follows: first, we define *frequently covered CpG sites* as CpG sites that are covered in the majority of the samples. Second, we search for regions within which the frequently covered CpG sites are close to each other. Third, we retain only regions with a minimum number of frequently covered CpG sites. In both, the acute promyelocytic leukemia data as well as in the simulated data, CpG sites covered in at least 75% of samples were defined as frequently covered CpG sites. A maximum distance $d_{\max}$ of 100 bp to their nearest neighbor within a CpG cluster was accepted. Only CpG clusters with at least 20 frequently covered CpG sites were used for the analysis. Note that the frequently covered CpG sites are considered to define the CpG cluster boundaries only. For subsequent analysis, all methylation data within these CpG clusters are used. The CpG cluster detection is applicable for other targeted BS data in the same manner. Alternatively, rather than CpG clusters, the target regions may be used, e.g. if a DNA capture method followed by BS was used.

3.3 Smoothing

Methylation levels are strongly spatially correlated (Eckhardt *et al.*, 2006), that is why local smoothing of the data is appropriate (Hansen *et al.*, 2012; Jaffe *et al.*, 2012). Hansen *et al.* (2012) showed that smoothing of raw methylation data can reduce the required sequencing coverage. DMR detection approaches without smoothing often discard CpG sites of low coverage from further analyses. Thus, the underlying spatial correlation allows using information from neighboring CpG sites, on the one hand, reducing the variance of the methylation levels especially for lowly covered CpG sites and, on the other hand, predicting methylation levels at CpG-sites missing any sequence reads to avoid missing values.

Within each CpG cluster and for each sample, a smoothing function is modeled. For a given position x, the weighted local likelihood for methylation level y at position x

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

$$L(y|m,n,w) = \prod_{i=1}^{k} B(m_i|n_i,y)^{w_i}$$

is maximized, with B the binomial probability function, n_i the number of reads and m_i the number of methylated reads at x_i, $i = 1, \ldots, k$. Weights w_i are calculated using a triangular kernel with bandwidth h:

$$w_i = K(x_i) = \left(1 - \frac{|x - x_i|}{h}\right)\mathbf{1}_{\left\{\left(\frac{|x-x_i|}{h}\right)\leq 1\right\}}.$$

This smoothing approach ensures that CpG sites close to position x and CpG sites with a high coverage have high influence on the estimation of the methylation level at x. The resulting methylation levels represent the relative methylation between 0 and 1.

At each CpG site within a CpG cluster, the raw methylation data are smoothed with a bandwidth of $h = 80$ bp.

3.4 Model and test group effect

Testing of group effects on methylation levels is usually done via t-tests, Wilcoxon rank-sum tests or linear regression (Bock, 2012; Hansen *et al.*, 2012; Wang *et al.*, 2012). Nevertheless, as the methylation values $y \in (0, 1)$, the beta distribution is the appropriate distribution for modeling y (Ferrari and Cribari-Neto, 2004; Kuan *et al.*, 2010). Let $\mu = p/(p+q)$ and $\phi = p + q$, with $p > 0$, $q > 0$. The density of y can be written as

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},$$

with Gamma function Γ. The mean and variance of y are $\mathrm{E}(y) = \mu$ and $\mathrm{Var}(y) = \mu(1-\mu)/(1+\phi)$, respectively, and ϕ can be interpreted as a precision parameter. Thus, we model the mean of the methylation level y_i at position i within a beta regression.

$$g(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j,$$

where $\beta = (\beta_1, \ldots, \beta_k)^{\mathrm{T}}$ is a vector of the regression parameters and $x_{i1}, \ldots, x_{ik}$ are observations on k covariates. The link function $g(.)$ is strictly monotonic, twice differentiable and maps $(0, 1)$ into $\mathbb{R}$. The most common link for methylation levels is the logit function $g(\mu) = \log(\mu/(1-\mu))$. Nevertheless, we prefer to use the probit function $g(\mu) = \Phi^{-1}(\mu)$, where $\Phi^{-1}(.)$ is the inverse of the standard cumulative normal distribution function, as this link is more moderate in the transformation of extreme values close to 0 or 1. The maximum likelihood estimators of β and ϕ do not have a closed form and need to be obtained using numerical methods, for details see Ferrari and Cribari-Neto (2004).

The methylation variation of cancer samples is often higher than for normal samples (Hansen *et al.*, 2011; Schoofs *et al.*, 2013). An extension of the beta regression is the variable dispersion beta regression model. In this case, the biological variability and, hence, the precision parameter ϕ may be additionally modeled within a linear model with the group (e.g. cancer/normal) as independent variable (Cribari-Neto and Zeileis, 2010; Simas *et al.*, 2010).

We fit the beta regression model to the smoothed methylation levels at each CpG site and test for a group effect using the Wald test.

3.5 Hierarchical testing

As the number of tested CpG sites usually exceeds 1 million, the problem of multiple testing plays an important role in the detection of differentially methylated CpG sites. Bock (2012) points out that after multiple testing correction, only the strongest differences tend to remain significant. We suggest applying a hierarchical testing procedure by Benjamini and Heller (2007) on target regions. They introduced a hierarchical testing procedure that first tests regions and then tests the locations within the rejected regions. As region units are tested first rather than locations, the number of hypothesis tests is reduced, which increases statistical power.

3.6 Test CpG clusters

The aim is to detect CpG clusters containing at least one differentially methylated location and to control a size-weighted FDR on clusters. The weighted FDR (WFDR), proposed by Benjamini and Hochberg (1997), ensures that the chance of rejecting a cluster increases when it is larger, but on the other hand, that also the weight of the error increases if a large cluster is falsely rejected. To control the WFDR, the weighted Benjamini–Hochberg procedure (Benjamini and Hochberg, 1997) is applied on cluster P-values. For CpG site i in CpG cluster c, the P-value p_{ic} arising from the Wald test is transformed to the normally distributed z-score: $z_{ic} = \Phi^{-1}(1 - p_{ic})$. As cluster test statistic, the standardized z-score average $\bar{Z}_c/\hat{\sigma}_{\bar{Z}_c}$ of the cluster locations $i = 1, \ldots, m_c$ is used, with $\bar{Z}_c$ the z-score average and $\hat{\sigma}_{\bar{Z}_c}$ the estimated standard deviation of $\bar{Z}_c$. The standard deviation of $\bar{Z}_c$ is estimated as $\hat{\sigma}_{\bar{Z}_c} = \frac{1}{m_c}\sqrt{m_c + 2\sum_{i=1}^{m_c}\sum_{j=1}^{i-1}\hat{\rho}^c_{i,j}}$, with $\hat{\rho}^c_{i,j}$ the estimated correlation between two locations i and j. The correlation between two locations in a cluster is estimated as $\hat{\rho}^c_{i,j} = 1 - \hat{\gamma}(s_{ic} - s_{jc})$, with $\hat{\gamma}(s_{ic} - s_{jc})$ the estimated semivariogram corresponding to the distance between locations i and j within cluster c. The variogram for distance h is estimated robustly by $2\hat{\gamma}(h) = [\mathrm{median}(Z_{ic} - Z_{jc})^2 : s_{ic} - s_{jc} = h]/0.455$, see also Cressie, 1993. The weighted Benjamini–Hochberg procedure at level q is applied on the P-values $\tilde{\Phi}(\bar{Z}_c/\hat{\sigma}_{\bar{Z}_c})$ of all CpG clusters, with $\tilde{\Phi}$ the right tail probability of the standard normal distribution. The cluster P-values are ordered so that $p_{(1)} \leq \cdots \leq p_{(c)} \leq \cdots \leq p_{(m)}$. The CpG clusters corresponding to the smallest k P-values are rejected, with $k = \max\{c : p_{(c)} \leq (\sum_{b=1}^{c} w_{(b)}/m)q\}$, with $w_{(b)}$ being the weight associated with the size of cluster b and $\sum_{b=1}^{m} w_{(b)} = m$, the number of clusters. Benjamini and Hochberg, 1997 showed that, for independent cluster test statistics, $\mathrm{WFDR} \leq \left(\sum_{b\in I_0} w_b/m\right)q$, with I_0 the subset of indices corresponding to clusters without DMRs (true null clusters). If the joint distribution of the cluster test statistics is positive regression-dependent (PRDS) on the subset of true null clusters then $\mathrm{WFDR}_{\mathrm{clust}} \leq q$ (Benjamini and Yekutieli, 2001). In particular, the PRDS property is satisfied if the cluster test statistics are Gaussian, non-negatively correlated and the testing hypotheses are one-sided.

To ensure that the variance of the z-scores is Gaussian with variance 1, we recommend to estimate the variogram under the null hypothesis, e.g. for re-sampled data.

3.7 Trim significant CpG clusters

The aim is to remove the not differentially methylated CpG sites within the rejected CpG clusters and to control a location-wise FDR. For each of the resulting m_2 CpG sites within the rejected CpG clusters its conditional P-value $\hat{p}_{ic}$ on being in a rejected CpG cluster is estimated, see Benjamini and Heller, 2007. A multiple testing two-stage procedure at level q_2, as suggested by Benjamini *et al.* (2006), is applied on these conditional P-values that first estimates the number of null hypotheses and then uses it to enhance the power. The P-values are ordered so that $\hat{p}_{(1)} \leq \cdots \leq \hat{p}_{(m_2)}$. The number of null hypotheses $\hat{m}_{02}$ is estimated by $\hat{m}_{02} = m_2 - k_1$, with $k_1 = \max\{b : \hat{p}_{(b)} \leq (b/m_2)q_2'\}$ and $q_2' = q_2/(1+q_2)$. Then, $k_2 = \max\{b : \hat{p}_{(b)} \leq (b/\hat{m}_{02})q_2'\}$ is the number of rejected CpG sites. All locations with $\hat{p}_{ic} > (k_2/\hat{m}_{02})q_2'$ are trimmed. Benjamini and Heller (2007) showed that this two-stage procedure controls the location-wise FDR asymptotically at level q_2.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

3.8 Definition of DMR boundaries

The result of the cluster testing and trimming procedure is a list of differentially methylated CpG sites from which we expect that they form DMRs. Thus, we define DMRs as regions of adjacent rejected CpG sites within one CpG cluster. DMRs are divided if the methylation difference switches from positive to negative, or vice versa. This way we ensure that within a DMR, all CpG sites are hypermethylated or hypomethylated, respectively.

4 RESULTS

4.1 Simulation data

We applied the algorithm to our simulated dataset described in Section 3.1, obtaining 28 471 CpG clusters within which we smoothed the methylation levels for 1 710 164 CpG sites. Figure 1A shows the effect of smoothing for a selected sample and a selected genomic region. In Figure 1B, the smoothed methylation profiles for all samples are shown for the same region.

DMRs were identified with different choices of q and q_2. Table 1 shows the chosen q and q_2 together with the achieved cluster-wise WFDRs and location-wise FDRs. For $q = 0.05$ and $q = 0.1$, respectively, the WFDR could be controlled with $\mathrm{WFDR} < \left(\sum_{b \in I_0} w_b/m\right) q$. In contrast, the location-wise FDRs could not be controlled. However, this result was expected as the DMRs were expanded spatially by the smoothing step, as seen in Figure 1B. This explanation is supported by the fact that the location-wise FDRs could be controlled for pruned DMRs (Table 1, last column). Pruned DMRs were obtained as follows: after trimming and testing, the DMRs were defined as described in Section 3.8. Each DMR >80 bp, which is corresponding to the width of the smoothing window, was pruned by 40 bp (half the bandwidth) at each side. Smaller DMRs remained unchanged. Thus, we omitted CpG sites that were differentially methylated because of the smoothing. For the remaining CpG sites, we achieved control of the location-wise FDR. Results for other choices of parameters for CpG cluster definition ($d_{\max}$) and for smoothing (h) are shown in Supplementary Tables S1 and S2.

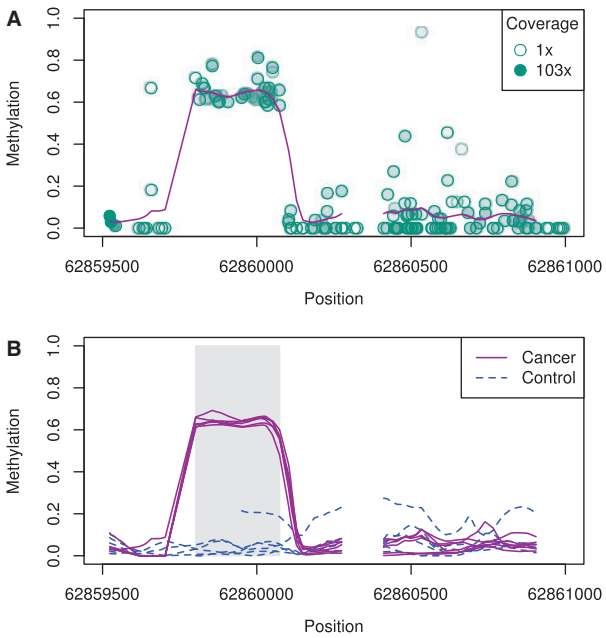

Fig. 1. Methylation values for a selected region. (**A**) Shown are the ratios of methylated and total reads (circles) and the smoothed methylation profile (line) for a simulated cancer sample. The grade of transparency of the circles represent the coverage of the respective CpG site. (**B**) Shown are the smoothed methylation profiles for all samples. The shaded box represents a simulated DMR. One of the control samples had an extreme low coverage that is why smoothing was impossible for some genomic regions

The probability to detect a differentially methylated CpG site (power) is depicted in Figure 2 depending on its simulated difference and on the fraction of differentially methylated CpG sites in its CpG cluster. As a standardized averaged z-score is used as cluster test statistic, the fraction of high z-scores affects the probability to reject a cluster. Figure 2 suggests that a CpG site with a methylation difference of at least 0.2 located in a CpG cluster with >30% differentially methylated CpG sites was detected with a probability of >90%. This implies that not only the methylation difference but also the choice of the CpG clusters is crucial for the power.

To illustrate the performance of our approach in comparison with others, we also applied BSmooth and methylKit to the simulated data. For BSmooth (version 0.4.3), we set the parameters for smoothing similar to the BiSeq parameter choices: the minimum number of methylation loci in a smoothing window (ns) was set to 20, the minimum bandwidth (h) was set to 80 and the maximum gap between two methylation loci, before the smoothing is broken across the gap (maxGap), was set to 100 bp. For CpG sites where at least two cancer samples and at least two normal samples had a minimum coverage of two reads (as recommended in the package vignette) were tested for differential methylation (3 187 862 CpG sites). Despite of comparable parameter choices, the number of tested CpG sites is higher than for BiSeq. This is because the given bandwidth is a minimum bandwidth and is enlarged until at least 20 CpG sites are included within the smoothing window. A consequence is that the degree of smoothing is different from sample to sample for the same genomic region and also within the genome of one sample. Furthermore, BSmooth adopts local

Table 1. Cluster-wise and location-wise FDRs for BiSeq

q	$\sum_{b \in I_0} \frac{w_b}{m} q$	q_2	WFDR	FDR	
			Cluster	Location	Location (pruned)
0.05	0.043	0.05	0.009	0.177	0.027
		0.1		0.189	0.032
0.1	0.086	0.05	0.015	0.184	0.028
		0.1		0.197	0.033

Note: Shown are the choices of q with the resulting upper bound for independent cluster test statistics and q_2 together with the observed cluster-wise WFDR, location-wise FDR and the location-wise FDR in pruned DMRs.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

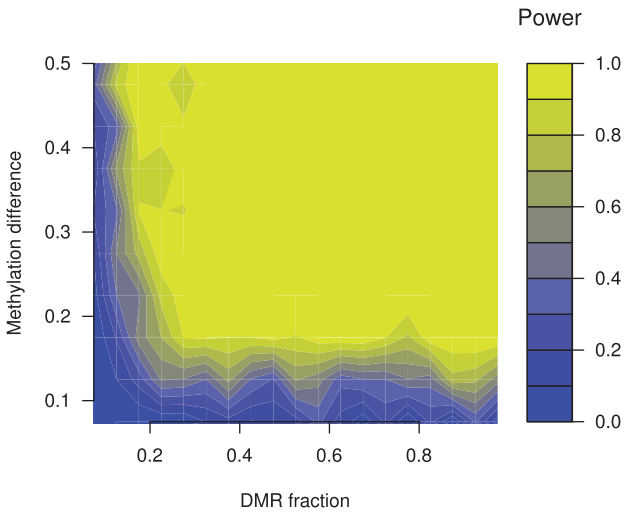

Fig. 2. Power to detect differentially methylated CpG sites. Shown is the influence of the CpG methylation difference and DMR fraction of the respective CpG clusters on the probability to detect a differentially methylated CpG site. The higher the simulated methylation difference the higher was the probability to reject the CpG site (*y*-axis). The higher the fraction of differentially methylated CpG sites in CpG clusters, the higher was the probability to reject the clusters and, hence, to reject the CpG sites

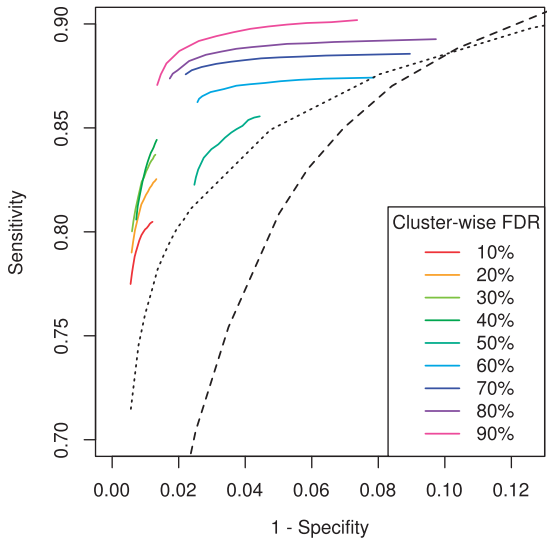

Fig. 3. ROC curves for differentially methylated CpG sites detected from BiSeq, BSmooth and methylKit. Shown are ROC curves for different *t*-statistic thresholds (BSmooth; dashed line), different q-value thresholds (methylKit; dotted line) and for different choices of q and q_2 (BiSeq; colored lines). Each of the ROC curves for BiSeq arises from a chosen q with different choices of q_2 (from 0.01 to 0.99)

logistic regression for smoothing. One main drawback of the regression approach is extrapolation: if methylation is predicted outside of a covered genomic region, the last observed slope is continued, often resulting in extrapolated methylation values of 0 and 1, see Supplementary Figure S3 for an example. The variance was estimated for the control group, as suggested in the package vignette. CpG sites with associated absolute *t*-statistics beyond a certain cut-off were defined as differentially methylated CpG sites.

methylKit (version 0.5.5) was applied to locations that are covered in at least four samples per group (2 903 270 CpG sites). It is necessary to set such a threshold because there is no smoothing step; hence, many CpG sites are covered in a fraction of the samples only. CpG sites with associated q values below a certain cut-off were defined as differentially methylated CpG sites.

Figure 3 shows ROC curves for all three approaches. As two FDR levels have to be chosen for BiSeq, we plotted several lines each showing the results for fixed q and varying q_2. In summary, BiSeq achieved a higher sensitivity than BSmooth and methylKit for false positive rates <8%.

4.2 Acute promyelocytic leukemia dataset

To analyze the acute promyelocytic leukemia data, we first detected CpG clusters with the same parameters as used for simulation data. We identified 30 205 CpG clusters with a total width of 1.4×10^7 bp (median: 373 bp, minimum: 42 bp, maximum: 4270 bp). We modeled the methylation levels applying a beta regression with group (cancer/control) and gender as explanatory variables. The precision parameter was modeled within an additional model with group as explanatory variable, as we expect higher methylation variances in the leukemia samples (Schoofs *et al.*, 2013). Testing the CpG clusters at a WFDR of 0.1 leads to the rejection of 7760 clusters. Cluster trimming at an FDR of 0.05 revealed 306 807 differentially methylated CpG sites, forming 9442 DMRs. The median of the DMR widths was 168 bp (minimum: 1 bp, maximum: 2401 bp). The vast majority (95.6%) of the differentially methylated CpG sites were hypermethylated in leukemia samples, see Supplementary Figure S4 for the frequency distribution. Plotting the chromosomal positions of the differentially methylated and of all tested CpG sites revealed that for some chromosomes, the distribution of differentially methylated CpG sites was shifted toward chromosomal ends in relation to the distribution of the tested CpG sites (Schoofs *et al.*, 2013). Supplementary Figure S5 depicts the distributions in chromosome 5. Moreover, it might be of interest whether certain genomic regions were located in DMRs more (or less) frequently than statistically expected. Briefly, we determined the centers of regions of interest, e.g. gene promoters, and then calculated the expected number of region centers within each DMR, assuming a uniform distribution of the centers in the CpG clusters. The number of expected region centers was then subtracted from the number of region centers actually observed in the DMRs. This analysis yielded that gene bodies were significantly overrepresented in hypermethylated DMRs, and promoters were underrepresented in both, hyper- and hypomethylated DMRs (Supplementary Fig. S6).

5 DISCUSSION

We proposed a DMR detection approach that tests regions for differential methylation with subsequent localization of DMRs within rejected regions. This hierarchical procedure is more powerful than BSmooth and methylKit.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

One strength of the proposed algorithm is that a region-wise FDR can be controlled. Nevertheless, we could ask whether smoothing of methylation data and focusing on regions is an appropriate strategy considering the fact that methylation of a single CpG site can prevent transcription factor binding (Gaston and Fried, 1995). To investigate whether focusing on region-wise methylation changes is meaningful, we analyzed the spatial clustering behavior of differentially methylated CpG sites in the leukemia dataset. We tested CpG sites for differential methylation in leukemia compared with controls without smoothing beforehand. This way we ensured that we did not add additional spatial correlation. We compared Ripley's K functions for both, tested CpG sites and differentially methylated CpG-sites and could show that the clustering of differentially methylated CpG sites is more pronounced than the clustering of tested CpG sites (which is due to the targeted sequencing). Differentially methylated CpG sites tend to form clusters of sizes up to 1000 bp (see Supplementary Fig. S7). The fact that entire regions are particularly effected by differential methylation in leukemia (and probably in other cancer types as well) justifies smoothing and detection of differentially methylated regions.

An advantage over other approaches is that the methylation levels are modeled within a regression model allowing adding further independent variables and confounders to the model. For instance, Boks *et al*., 2009 found associations of DNA methylation with gender and age; thus, it might be important to take these confounders into account. Neither BSmooth nor the methylKit package allows adding further variables. By modeling the methylation levels within a beta regression, we make sure that basic properties of methylation levels are respected, i.e. the distribution between 0 and 1 and a smaller variance near the boundaries. Another advantage of the beta regression is the option to consider higher variances in a group of samples (commonly the cancer samples) via a variable precision parameter. A drawback is the long computing time.

Most of the parameters that have to be chosen are related to CpG cluster detection and thus influence the extent of the genome that is to be analyzed. Basically, the grade of the minimal spatial density of covered CpG sites that form CpG clusters together with the width of the smoothing window should be adjusted to the expected DMR lengths. Even if many CpG sites are covered over a wide genomic region (for example, in whole-genome bisulfite sequencing data), it is advisable not to choose too large CpG clusters with many CpG sites, when the expected DMR widths are much smaller, as this would decrease the power markedly (Fig. 2). Likewise, a wide smoothing window results in a weaker sensitivity to small DMRs. Moreover, the smoothing bandwidth should be adjusted to the choice of CpG clusters.

Benjamini and Heller (2007) showed that the cluster testing procedure described earlier in the text controls the cluster-wise WFDR under independence or PRDS assumption, respectively. The procedure ensures that the cluster test statistics are Gaussian, and that the testing hypotheses are one-sided. Finally, to satisfy the PRDS property, the cluster test statistics have to be non-negatively correlated under the null hypothesis. We assume that this property is met in many cases for methylation data. Furthermore, it is important to note that the location-wise FDR is controlled asymptotically only, and DMR boundaries are slightly shifted by the smoothing step. However, we think that a targeted bisulfite sequencing approach primarily requires an analysis of the target units, e.g. CpG islands or gene bodies, rather than individual CpG sites. Besides, differential methylation typically is reported in terms of regions. This indicates a region-wise testing approach in the first place, followed by a CpG-wise detection tolerating more errors. This hierarchical testing procedure achieves a higher power than BSmooth and methylKit.

The approach is tailored to targeted BS data. The direct use on whole-genome BS data would be feasible if the analysis was restricted on discrete genomic regions, e.g. on genes.

We think that the two-stage testing procedure is applicable for other types of next-generation sequencing data as well. The histone modification H3K4me3 is known to occur in promoter regions primarily (Hebenstreit *et al*., 2011). Thus, promoters could be tested for differential histone modification between groups of samples in a first step and in bins (sub-regions) within rejected promoters in a second step. This is a point for further research.

ACKNOWLEDGEMENT

The authors thank Till Schoofs, Christian Rohde and Carsten Müller-Tidow (Department of Medicine A - Hematology, Oncology and Pneumology, University of Münster, Germany) for providing the data and discussing biological results.

Conflict of Interest: none declared.

REFERENCES

Akalin,A. *et al.* (2012) methylKit: a comprehensive r package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*., **13**, R87.

Benjamini,Y. and Heller,R. (2007) False discovery rates for spatial signals. *J. Am. Stat. Assoc*., **102**, 1272–1281.

Benjamini,Y. and Hochberg,Y. (1997) Multiple hypotheses testing with weights. *Scand. J. Stat*., **24**, 407–418.

Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat*., **29**, 1165–1188.

Benjamini,Y. *et al.* (2006) Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, **93**, 491–507.

Bock,C. (2012) Analysing and interpreting DNA methylation data. *Nat. Rev. Genet*., **13**, 705–719.

Boks,M.P. *et al.* (2009) The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PLoS One*, **4**, e6767.

Campagna,D. *et al.* (2012) Pass-bis: a bisulfite aligner suitable for whole methylome analysis of illumina and solid reads. *Bioinformatics*., **29**, 268–270.

Challen,G.A. *et al.* (2012) Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat. Genet*., **44**, 23–31.

Chen,P.Y. *et al.* (2010) Bs seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics*, **11**, 203.

Cressie,N.A.C. (1993) *Statistics for Spatial Data*. Wiley, New York.

Cribari-Neto,F. and Zeileis,A. (2010) Beta regression in r. *J. Stat. Softw*., **34**, 1–24.

Das,P.M. and Singal,R. (2004) DNA methylation and cancer. *J. Clin. Oncol*., **22**, 4632–4642.

Eckhardt,F. *et al.* (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet*., **38**, 1378–1385.

Ferrari,S.L. and Cribari-Neto,F. (2004) Beta regression for modelling rates and proportions. *J. Appl. Stat*., **31**, 799–815.

Gaston,K. and Fried,M. (1995) CpG methylation and the binding of YY1 and ETS proteins to the surf-1/surf-2 bidirectional promoter. *Gene*, **157**, 257–259.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.

Goldman,M. *et al.* (2012) The UCSC cancer genomics browser: update 2013. *Nucleic Acids Res*., **41**, D949–954.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Gopalakrishnan,S. *et al.* (2008) DNA methylation in development and human disease. *Mutat. Res.*, **647**, 30–38.
Gu,H. *et al.* (2010) Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat. Methods*, **7**, 133–136.
Hansen,K.D. *et al.* (2011) Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.*, **43**, 768–775.
Hansen,K.D. *et al.* (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.*, **13**, R83.
Hebenstreit,D. *et al.* (2011) EpiChIP: gene-by-gene quantification of epigenetic modification levels. *Nucleic Acids Res.*, **39**, e27.
Jaffe,A.E. *et al.* (2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.*, **41**, 200–209.
Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, **27**, 1571–1572.
Krueger,F. *et al.* (2012) DNA methylome analysis using short bisulfite sequencing data. *Nat Methods*, **9**, 145–151.
Kuan,P.F. *et al.* (2010) A statistical framework for illumina DNA methylation arrays. *Bioinformatics*, **26**, 2849–2855.
Li,Y. *et al.* (2010) The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol.*, **8**, e1000533.
Lutsik,P. *et al.* (2011) BiQ Analyzer HT: locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing. *Nucleic Acids Res.*, **39**, W551–W556.
Meissner,A. *et al.* (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.*, **33**, 5868–5877.
R Core Team. (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
Schoofs,T. *et al.* (2013) DNA methylation changes are a late event in acute promyelocytic leukemia and coincide with loss of transcription factor binding. *Blood*, **121**, 178–187.
Simas,A.B. *et al.* (2010) Improved estimators for a general class of beta regression models. *Comput. Stat. Data Anal.*, **54**, 348–366.
Sun,Z. *et al.* (2012) *SAAP-RRBS*: streamlined analysis and annotation pipeline for reduced representation bisulfite sequencing. *Bioinformatics*, **28**, 2180–2181.
Wang,D. *et al.* (2012) Ima: An R package for high-throughput analysis of illuminas 450 k infinium methylation data. *Bioinformatics*, **28**, 729–730.
Wang,H.Q. *et al.* (2011) Slim: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics*, **27**, 225–231.
Xi,Y. *et al.* (2012) RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing. *Bioinformatics*, **28**, 430–432.
Zackay,A. and Steinhoff,C. (2010) Methvisual - visualization and exploratory statistical analysis of DNA methylation profiles from bisulfite sequencing. *BMC Res. Notes*, **3**, 337.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016