An Optimization Approach to Detect Differentially Methylated Regions from Whole Genome Bisulfite Sequencing Data

Nina Hesse[1], **Christopher Schröder**[2], **and Sven Rahmann**[1,2]

[1]**Computer Science XI, TU Dortmund, Dortmund, Germany**
[2]**Genome Informatics, Human Genetics, Faculty of Medicine, University of Duisburg-Essen, Essen, Germany**

ABSTRACT

Whole genome bisulfite sequencing (WGBS) is the current method of choice to obtain the methylation status of each single CpG dinucleotide in a genome. The typical analysis asks for regions that are differentially methylated (DMRs) between samples of two classes, such as different cell types. However, even with current low sequencing costs, many studies need to cope with few samples and medium coverage to stay within budget. We present a method to conservatively estimate the methylation difference between the two classes. Starting from a Bayesian paradigm, we formulate an optimization problem related to LASSO approaches. We present a dynamic programming approach to efficiently compute the optimal solution and its implementation diffmer. We discuss the dependency of the resulting DMRs on the free parameters of our approach and compare the results to those obtained by other DMR discovery tools (BSmooth and RADMeth). We showcase that our method discovers DMRs that are missed by the other tools.

Keywords: differential methylation, whole genome bisulfite sequencing, optimization, LASSO, dynamic programming

INTRODUCTION

DNA methylation, a chemical modification that affects 70% to 80% of the approximately 30 million cytosines in CpG dinucleotides in the human genome, is an epigenetic process playing a key role in X-chromosome inactivation (Gartler and Goldman, 2005), genomic imprinting (Li et al., 1993), stem cell differentiation and gene regulation (Bock, 2012). Aberrant methylation is associated with various diseases, such as obesity (Xu et al., 2013) or different types of cancer (Hansen et al., 2011).

It is therefore of considerable interest to compare the methylation state of every single CpG dinucleotide in the human genome in two groups of samples, a case group and a control group, and to find genomic regions where the methylation differs significantly. Case vs. control group might be two different cell types, or tumor samples vs. normal (blood) samples. A *differentially methylated CpG* (DMC) is either hypo- or hyper-methylated in the case group in comparison to the control group, meaning that it has a significantly lower resp. higher methylation level. Multiple adjacent DMCs form a *differentially methylated region* (DMR).

The current gold standard method to obtain the genome-wide methylation status at single basepair resolution is whole genome bisulfite sequencing (WGBS): DNA is treated with sodium hydrogen sulfite (bisulfite), transforming unmethylated cytosine into uracil that is sequenced as thymine, but leaving methylated cytosine unchanged. After C/T-tolerant read mapping and comparison against the reference sequence, the methylation status of every CpG can be inferred.

The methylation status of the same CpG in different samples may differ due to biological variation, and the (average) methylation level of a particular CpG in a group of samples may be estimated as the ratio of reads showing `C` and reads showing `C` or `T` at the cytosine in question. Of course, the precision of such an estimate depends on the read coverage at that site and on the number of samples, both of which may be low in experiments with a limited budget. Therefore, reliable detection of DMRs is a challenge with currently available WGBS data.

Related work Several groups have addressed the problem of detecting DMCs and DMRs. A basic method to detect DMCs between two samples is Fisher's exact test (Lister et al., 2009). As this method is only suited to compare two samples at a time, it does not account for biological variation.

At low coverage, it suffers from low statistical power.

BSmooth (Hansen et al., 2012) uses smoothing to "borrow" information from neighboring CpGs and thereby increases its power to detect DMRs. The approach is based on the assumption that the true methylation level varies smoothly across the genome. First, the approach uses local-likelihood smoothing to estimate the methylation level of each single CpG per sample and then uses the averages of the smoothed values per class to compute estimates of the mean differences and standard errors per CpG site to form a test statistic similar to the one used in a t-test.

Bisulfighter (Saito et al., 2014) is based on a hidden Markov model (HMM) whose states represent hyper-methylation, hypo-methylation and no differential methylation. The transition probabilities are inferred based on the probability distribution of the distances between neighboring CpGs. This enables Bisulfighter to detect short as well as long DMRs. Bisulfighter pools all samples of a group into one aggregate sample and therefore does not account explicitly for biological variability.

Recently, Bayesian methods have been explored, where prior assumptions about methylation rates can be encoded in the prior distribution. Posterior estimates of methylation levels and differences can then be computed from the (beta-distributed) prior and the (binomial-distributed) likelihood function. The beta-binomial approach is flexible and computationally efficient. One may for example compute the (parametric) posterior distribution of each class methylation level, or the maximum a-posteriori estimate for the class methylation difference. Three recent tools, RADMeth (Dolzhenko and Smith, 2014), MOABS (Sun et al., 2014) and BEAT (Akman et al., 2014), make use of the Bayesian beta-binomial approach.

Our approach Our work follows the Bayesian paradigm as well, but (a) with more detailed prior assumptions (i.e., we do not use the beta-binomial model), and (b) using approximations that cast the maximum a-posteriori-estimation computational as a strictly convex optimization problem related to LASSO (least absolute shrinkage and selection operator) approaches (Tibshirani, 1996; Tibshirani and Wang, 2007). We estimate the (unknown) true group-specific methylation rate at each CpG by using both the available (small sample and low coverage) data and certain smoothness assumptions. Our target function that can (in principle) be optimized using any toolbox for convex optimization, or, as we show here, be discretized and solved efficiently by dynamic programming. We are not aware that a similar solution approach has been proposed before. Our DMC estimation procedure outputs conservative (shrunken) estimates of CpG methylation differences between the two classes. DMR calling then outputs regions of consecutive DMCs according to the user's parameters.

To validate our results, we compare called DMCs and DMRs with those obtained by other tools, namely BSmooth and RADMeth, and showcase some of the differences. In this methodological article, we do not report on specific genes or promoters or attempt to give biological interpretations.

DMC DETECTION AS AN OPTIMIZATION PROBLEM

Input, Output and Notation

We assume that we have processed the raw sequence reads into the following format.

There are m samples, indexed by $t = 1, \ldots, m$, belonging to two classes $k \in \{\oplus, \ominus\}$. We write $K(t)$ for the class of sample t. There are n CpGs (e.g., on a chromosome) indexed by $i = 1, \ldots, n$. The data then consists of two matrices $C = (C_i^{(t)})$ representing the number of methylated cytosines at CpG i in sample t and $N = (N_i^{(t)})$ representing the sum of the numbers of methylated and unmethylated cytosines. We also assume that the distance d_i between consecutive CpGs $i-1$ and i is known.

We will estimate the true methylation level $\mu_i^{(t)}$ for each CpG i and sample t under assumptions that we list below. The true methylation level may differ from the observed $C_i^{(t)}/N_i^{(t)}$ due to sampling variance, especially at low coverage. We also assume the existence of a class methylation level $\mu_{i,k}$ for each CpG-position i and class $k \in \{\oplus, \ominus\}$. Differences between $\mu_i^{(t)}$ and $\mu_{i,K(t)}$ can be interpreted as biological variation within a class.

We say that CpG i is a *differentially methylated CpG* (DMC) between the two classes if $|\mu_{i,\oplus} - \mu_{i,\ominus}|$ is sufficiently different from zero. A *differentially methylated region* (DMR) is an interval of contiguous DMCs with a given minimum length.

The goal is to estimate the unknown class methylation levels $\mu_{i,\oplus}$ and $\mu_{i,\ominus}$ for all i from the given data under certain assumptions for regularization, which we discuss next.

Assumptions

The following assumptions describe a null model for typical data without any exceptional features. Deviations from this null model (that we will be able to detect) may either indicate problems with the data or the object of interest: differential methylation.

1. $\mu_i^{(t)} \approx C_i^{(t)}/N_i^{(t)}$: The sample methylation level $\mu_i^{(t)}$ is close to the observed ratio $C_i^{(t)}/N_i^{(t)}$. The degree of assumed closeness depends on the read coverage at CpG i: lower coverage allows larger variation. More precisely, given $N_i^{(t)}$ and $\mu_i^{(t)}$, the random variable $C_i^{(t)}$ has a binomial distribution with these parameters.

2. $\mu_i^{(t)} \approx \mu_{i,K(t)}$: The sample methylation level $\mu_i^{(t)}$ is close to the class methylation level $\mu_{i,k}$ for the sample's class $k = K(t)$. The degree depends on the (biological) within-class variability between the samples.

3. $\mu_i^{(t)} \approx \mu_{i-1}^{(t)}$ and $\mu_{i,k} \approx \mu_{i-1,k}$: Adjacent CpGs in the same sample and in the same class are expected to have a similar methylation level, depending on their spatial distance.

4. $\mu_{i,\ominus} \approx \mu_{i,\oplus}$: For most CpGs, we do not expect differential methylation. Therefore the methylation difference between the classes is (close to) zero.

Formalization of Assumptions

We now formalize the notion of $\approx$ in all of the above assumptions. We write μ for the collection of all $(m+2)n$ variables $\mu_i^{(t)}$ and $\mu_{i,k}$ for all CpG sites i, samples t and classes k. Our approach for estimating μ is based on (approximate) maximum-a-posteriori estimation from the data. We write $f(\mu \mid C;N)$ for the a-posteriori density.

Assumption 1. above is formalized by a likelihood function $p(C \mid \mu;N)$, while assumptions 2.–4. describe prior expectations about typical methylation levels and are formalized by a prior density $\pi(\mu)$. Note that $(N_i^{(t)})$ is assumed to be fixed and not part of the random model.

By Bayes' Theorem, the a-posterior density is proportional to the product of the likelihood and the prior:

$$f(\mu \mid C;N) \propto p(C \mid \mu;N) \cdot \pi(\mu)$$

Maximizing f is thus equivalent to minimizing $-\log p(C \mid \mu;N) - \log \pi(\mu)$.

Likelihood function

Consider the random number of observed methylated cytosines $C_i^{(t)}$ at site i in sample t when $\mu_i^{(t)}$ is known and $N_i^{(t)}$ is fixed: $C_i^{(t)}$ has a binomial distribution $\mathscr{B}_{N_i^{(t)},\mu_i^{(t)}}$ with expectation $N_i^{(t)}\mu_i^{(t)}$ and variance $(\sigma_i^{(t)})^2 := N_i^{(t)}\mu_i^{(t)}(1-\mu_i^{(t)})$. For sufficiently large $N_i^{(t)}$, it can be approximated by a normal distribution with the same moments; its density is an approximation to $p(C \mid \mu;N)$.

$$p(C \mid \mu;N) \approx \prod_i \prod_t \frac{1}{\sqrt{2\pi(\sigma_i^{(t)})^2}} \cdot \exp\left(-\frac{(C_i^{(t)} - N_i^{(t)}\mu_i^{(t)})^2}{2\,(\sigma_i^{(t)})^2}\right). \tag{1}$$

The key point here is that the density is concentrated around its mean and decreases quadratically in the exponent. The variance $(\sigma_i^{(t)})^2$ unfortunately depends on the unknown parameter $\mu_i^{(t)}$, which, however, should be close to $C_i^{(t)}/N_i^{(t)}$. For the purpose of estimating the variance only, we approximate it by a regularized estimation using one pseudocount: $\mu_i^{(t)} \approx (C_i^{(t)}+1)/(N_i^{(t)}+2)$, i.e.,

$$(\sigma_i^{(t)})^2 \approx N_i^{(t)} \cdot \frac{C_i^{(t)}+1}{N_i^{(t)}+2} \cdot \left(1 - \frac{C_i^{(t)}+1}{N_i^{(t)}+2}\right). \tag{2}$$

This gives us an approximation of the variance that is independent of the unknown parameter $\mu_i^{(t)}$. Then the factor $\sqrt{2\pi(\sigma_i^{(t)})^2}$ does not depend on $\mu_i^{(t)}$ anymore and is a constant. We thus obtain

$$-\log(p(C \mid \mu;N)) \approx -\sum_i \sum_t \frac{N_i^{(t)}(N_i^{(t)}+2)^2}{2(C_i^{(t)}+1)(N_i^{(t)}-C_i^{(t)}+1)} \left(\frac{C_i^{(t)}}{N_i^{(t)}} - \mu_i^{(t)}\right)^2 + \text{const}. \tag{3}$$

Prior density

We model the unknown prior density $\pi(\mu)$ with three factors corresponding to assumptions 2. to 4.:

$$\pi(\mu) = \pi_{\text{class}}(\mu) \cdot \pi_{\text{space}}(\mu) \cdot \pi_{\text{diff}}(\mu) \tag{4}$$

The factor $\pi_{\text{class}}(\mu)$ makes a statement about the within-class variability of the methylation level at each site. We assume that most differences $|\mu_i^{(t)} - \mu_{i,K(t)}|$ are small and that the probability density for larger differences decreases exponentially with the difference. We therefore assume

$$\pi_{\text{class}}(\mu) \propto \prod_i \prod_t \exp\left(-\alpha_{\text{class};i,K(t)} \cdot |\mu_i^{(t)} - \mu_{i,K(t)}|\right). \tag{5}$$

A large parameter value of $\alpha_{\text{class};i,k}$ states that the difference is concentrated close to zero, so there is little variation in class k at CpG i. We discuss below how appropriate parameters can be chosen.

A similar assumption is made for the closeness of methylation levels at adjacent CpG sites $i-1$ and i. We could model the closeness in each individual sample or in each class. To keep the model simple and because each sample methylation level is already tied to its class methylation level by (5), we only model the closeness of class methylation levels between adjacent CpGs, thus avoiding redundancy in the model.

$$\pi_{\text{space}}(\mu) \propto \prod_i \prod_{k\in\{\oplus,\ominus\}} \exp\left(-\alpha_{\text{space};i,k} \cdot |\mu_{i,k} - \mu_{i-1,k}|\right), \tag{6}$$

where the parameter $\alpha_{\text{space};i,k}$ depends on the basepair distance between CpG sites i and $i-1$.

Finally, we make a similar distributional assumption for the difference between the two class methylation levels $|\mu_{i,\oplus} - \mu_{i,\ominus}|$. Most CpG sites are not differentially methylated and large differences are (exponentially) rare:

$$\pi_{\text{diff}}(\mu) \propto \prod_i \exp\left(-\alpha_{\text{diff};i} \cdot |\mu_{i,\oplus} - \mu_{i,\ominus}|\right) \tag{7}$$

Objective function

With the above considerations, the optimization problem is to minimize the objective function

$$\begin{aligned} -\log(f(\mu \mid C;N)) = &\sum_i \sum_t \frac{N_i^{(t)}(N_i^{(t)}+2)^2}{2(C_i^{(t)}+1)(N_i^{(t)}-C_i^{(t)}+1)} (\mu_i^{(t)} - C_i^{(t)}/N_i^{(t)})^2 \\ &+ \sum_i \sum_t \alpha_{\text{class};i,K(t)} |\mu_i^{(t)} - \mu_{i,K(t)}| \\ &+ \sum_{i\geq 2} \sum_{k\in\{\oplus,\ominus\}} \alpha_{\text{space};i,k} |\mu_{i,k} - \mu_{i-1,k}| \\ &+ \sum_i \alpha_{\text{diff};i} |\mu_{i,\oplus} - \mu_{i,\ominus}| \end{aligned} \tag{8}$$

for given weights $\alpha = (\alpha_{\text{class};i,k}, \alpha_{\text{space};i,k}, \alpha_{\text{diff};i})$ and given observed data C and N.

The objective function is (strictly) convex in the unknowns μ. We note that closeness to the observed data is modeled by a quadratic term, whereas the prior assumptions are encoded in ℓ_1-terms; thus the optimization problem is related to the LASSO (least absolute shrinkage and selection operator; Tibshirani (1996); Tibshirani and Wang (2007)). The main difference to the standard LASSO approaches is that we have several ℓ_1 terms with different weights in the objective function. This prohibits the direct application of existing LASSO algorithms to solve the problem.

Choosing weights

To choose meaningful weights $\alpha = (\alpha_{\text{class};i,k}, \alpha_{\text{space};i,k}, \alpha_{\text{diff};i})$, one has to know properties of the within-class variation of methylation at each site, of the "horizontal" correlation of methylation levels, and of the frequency of differentially methylated sites, respectively. This is unrealistic in most scenarios. However, we may be able to robustly estimate moments of the distributions of the differences by examining reliable parts of the given data.

We first reduce complexity by supposing that the weights α_{class} and α_{diff} are independent of class or position (i.e., they are constants independent of i and k) and that $\alpha_{\text{space};i}$ depends only on the basepair distance d_i to the previous CpG site (and not otherwise on i).

To estimate α_{class}, we can assess properties of typical values of $|\mu_i^{(t)} - \mu_{i,K(t)}|$ by tabulating these values at high-coverage CpG sites. Here it is justified to approximate $\mu_i^{(t)}$ by $C_i^{(t)}/N_i^{(t)}$ and $\mu_{i,K(t)}$ by

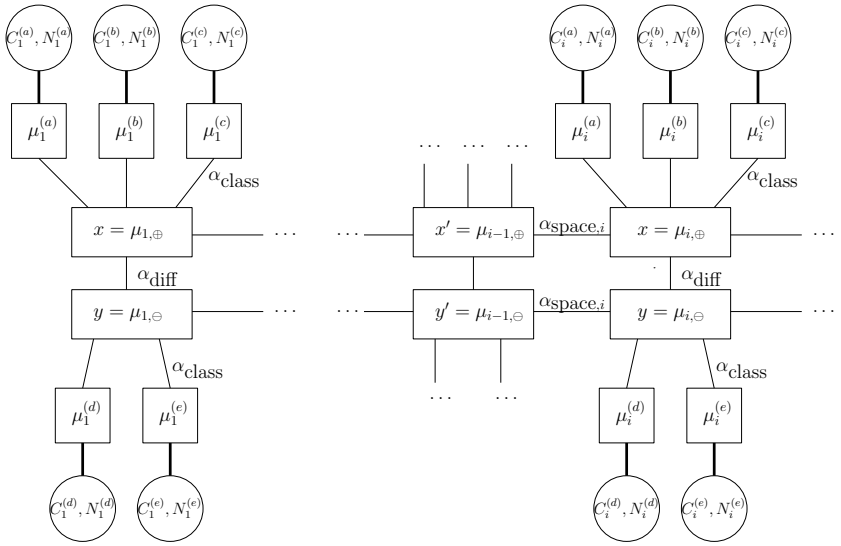

Figure 1. Visualization of parameters, data and terms in the target function. Square nodes correspond to unknown parameters to be estimated. Round nodes correspond to observed data. Edges correspond to differences between parameters or a parameter and data the objective function (thin: absolute difference, bold: quadratic difference). Variables x, y correspond to assumptions about (discretized) methylation levels during the dynamic programming solution.

either their mean or median. Plotting the empirical quantiles of the differences against the theoretical ones given by (5) in a Q-Q plot, we can check the model assumptions and estimate a reasonable range for α_{class}. By equating the theoretical mean or median of the distribution (5) with the empirical one, we obtain a point estimate for α_{class}. Similarly, we obtain diagnostic Q-Q plots and estimates for α_{diff}. We present an example in the Supplement.

As $\alpha_{\text{space};i}$ depends on spatial distances between CpGs, we compute an individual distribution for each possible (small) distance. For all adjacent CpGs with high overall coverage and genomic distance d between 2 and 25 bp, we compute the methylation level difference $|\mu_{i,k} - \mu_{i-1,k}|$ by taking the estimated levels as proxies for the true ones. This yields a median value for each distance $d = 2, \ldots, 25$. Linear regression yields a predictive function $L(d)$ of the median for larger distances (up to a distance where no predictions make sense). Assuming that $L(d_i)$ is the median of $|\mu_{i,k} - \mu_{i-1,k}|$ differences at CpGs with distance $d = d_i$, we obtain $\alpha_{\text{space};i} = \ln(2)/L(d_i)$ by equating medians.

With the weights estimated from data in this way, they may be influenced true differentially methylated CpGs or by outliers; both effects are not desired for the null model. Therefore these weights only represent a first estimate and should be fine-tuned in applications by systematically changing them and observing the output size (number of DMRs, number of CpGs in DMRs).

SOLUTION BY A DYNAMIC PROGRAMMING APPROACH

The optimization problem given by Eq. (8) has $(m+2)n$ variables $\mu_i^{(t)}$ and $\mu_{i,\oplus}, \mu_{i,\ominus}$, where n is the number of CpGs on a chromosome and m is the number of samples. A realistic setting is that n is of the order of 2.5 million and $m = 6$, yielding 20 million variables. This will often be a too large problem for an all-purpose quadratic programming solver.

While it is possible to subdivide the problem into smaller ones (e.g., a large distance between two adjacent CpGs means that their methylation levels are essentially uncorrelated, so one does not need to model them as connected) and use standard (commercial) solvers to obtain an exact solution, we here propose a different approach that exploits the specific structure of the problem and is fast in practice. It comes at the expense of discretizing a part of the possible solution space, i.e., the class methylation levels are only considered as a discrete set $\mu_{i,k} \in \{0, \varepsilon, 2\varepsilon, \ldots, 1\} =: M$ with $1/\varepsilon \in \mathbb{N}$. The running time then depends on $|M| = 1 + 1/\varepsilon$. In practice, it is sufficient to use $\varepsilon = 0.05$, as one is usually not interested in a more exact determination of the methylation level. We propose a dynamic programming approach to solve the discretized optimization problem.

Let $V_j(x, y)$ be the optimal value of the objective function up to CpG j under the condition that $\mu_{j,\oplus} = x \in M$ and $\mu_{j,\ominus} = y \in M$. In other words, define

$$V_j(x,y) := \min_{\mu:\ \substack{\mu_{j,\oplus}=x,\\ \mu_{j,\ominus}=y}} \sum_{i=1}^{j} \Bigg[\alpha_{\text{diff}} \left|\mu_{i,\oplus} - \mu_{i,\ominus}\right| + [\![i \geq 2]\!]\, \alpha_{\text{space};i} \left(\left|\mu_{i,\oplus} - \mu_{i-1,\oplus}\right| + \left|\mu_{i,\ominus} - \mu_{i-1,\ominus}\right| \right) + \sum_t \left(\frac{N_i^{(t)} (N_i^{(t)}+2)^2}{2(C_i^{(t)}+1)(N_i^{(t)} - C_i^{(t)} + 1)} \left(\mu_i^{(t)} - C_i^{(t)}/N_i^{(t)}\right)^2 + \alpha_{\text{class}} \left|\mu_i^{(t)} - \mu_{i,K(t)}\right| \right) \Bigg]. \quad (9)$$

For $j = 1$, computing the $|M|^2$ values $V_1(x,y)$ for each $x \in M$, $y \in M$ requires only solving simple "local" one-dimensional problems by finding optimal values $\mu_1^{(t)}$ for all samples t; see Figure 1 (left). Writing $z = \mu_1^{(t)}$, each problem has the form

$$\text{minimize } F(z) := w(z-p)^2 + v|z-q|, \quad (10)$$

where $p := C_1^{(t)}/N_1^{(t)}$ is the data, $q := \mu_{1,K(t)}$ is the class methylation rate (the given x or y, depending on the class of sample t) and w, v are the appropriate weights from (9). Such a one-dimensional strictly convex problem (with $w > 0, v > 0$) can be explicitly solved. Let $F^* := \min_z F(z)$ and $z^* := \operatorname{argmin}_z F(z)$. Then

$$F^* = F(z^*), \qquad z^* = \begin{cases} p & \text{if } q = p, \\ \max\{p - v/(2w), q\} & \text{if } q < p, \\ \min\{p + v/(2w), q\} & \text{if } q > p. \end{cases} \quad (11)$$

The optimal value is then $V_1(x,y) = \alpha_{\text{diff}}|x-y| + \sum_t F_1^{*(t)}$, where $F_1^{*(t)}$ is the optimal solution (11) of (10) for sample t for the given x, y at the first CpG. Note that for $j = 1$, the "connection term" weighted with $\alpha_{\text{space};i}$ does not exist.

For $j \geq 2$, we claim that

$$V_j(x,y) = \min_{x',y'} \left[V_{j-1}(x',y') + \alpha_{\text{diff}}|x-y| + \alpha_{\text{space};i}\left(|x'-x| + |y'-y|\right) + \sum_t F_j^{*(t)} \right], \quad (12)$$

where $F_j^{*(t)}$ is the optimal solution of the local problem (10) for sample t at CpG j with known class methylation rate $q \in \{x, y\}$; see Figure 1 (right).

The proof that (12) is correct follows the usual inductive argument as for the Viterbi algorithm for HMMs: Assume that $V_j(x,y)$ has a smaller value, say u, than the right-hand side of (12), and consider the class methylation rates x', y' at CpG $j-1$ that result in the optimal u. Since all the terms in (12) are constant or optimal, the only possible explanation is that we already missed a smaller value for $V_{j-1}(x',y')$. Descending the chain down to V_1 leads to a contradiction.

Therefore we compute a sequence of $|M| \times |M|$ matrices V_j, $j = 1, \ldots, n$, where each V_j depends only on the previous matrix. To obtain the final result, we see that $V := \min_{x,y} V_n(x,y)$ is the overall optimal value. To reconstruct the optimal sequence of class methylation levels, we additionally store traceback pointers, i.e., the values of x', y' minimizing (12) for each (j,x,y) in $O(n|M|^2)$ space.

To analyze the running time, we note that there are $n|M|^2$ values to compute. According to (12), each value $V_j(x,y)$ can be computed in $O(m + |M|^2)$ time, leading to a running time of $O(mn|M|^2 + n|M|^4)$. (With current small sample sizes $m = 6$, and the proposed $|M| = 21$ with $\varepsilon = 0.05$, the time is linear in the number n of CpGs. In fact, for $n = 30 \cdot 10^6$, we obtain $mn|M|^2 + n|M|^4 \approx 5.9 \cdot 10^{12}$.)

The space requirement of $O(n|M|^2)$ may be reduced to $O(|M|^2 + n)$ by a divide-and-conquer technique following the lines of Hirschberg's technique for pairwise sequence alignment (Hirschberg, 1975). In practice, the $\alpha_{\text{space};i}$ term vanishes for large distances between CpGs, and one can start a new independent problem.

DMR Calling

After computing class methylation values for every single CpG site, we call DMCs and DMRs as a function of the following user-defined parameters: the minimally required methylation difference Δ for a CpG to be a DMC, the minimum size S (in terms of number of CpGs) of a DMR, the minimally required mean methylation difference $\overline{\Delta}$ of each size-S window in a DMR, and optionally the maximally allowed number s of consecutive non-DMCs in a DMR ("skip" parameter, $s \ll S$),

We first create a binary sequence $\delta = (\delta_i) \in \{0,1\}^n$ of DMC indicators by comparing each $|\mu_{i,\oplus} - \mu_{i,\ominus}|$ with threshold Δ. Short runs of zeros (non-DMCs) of length $\leq s$ are filled in with 1s (the

default is $s = 0$, so no relabeling of non-DMCs as DMCs occurs). Next we determine the start indices of size-S windows with all ones whose mean methylation difference at at least $\overline{\Delta}$. The resulting runs of ones are the resulting DMRs.

As a post-processing step, DMRs called by different parameter sets can be joined, e.g., one might use $(\Delta, S, \overline{\Delta}, s) = (0.2, 5, 0.3, 1)$ to find good candidates of intermediate length and additionally $(0.7, 1, 0.0, 0)$ for clearly differential single CpGs.

Software

The optimization approach described above has been implemented in the `diffmer` (DIFFerentially MEthylated Regions) software[1]. Installation and usage instructions are on the project overview page. The software is written in Python (version 3.4 or newer is required) and makes use of just-in-time compilation with numba and LLVM for performance-critical code. Running `diffmer` consists of two steps:

1. Estimating class methylation levels (`diffmer estimate`) requires tabular count data in BED format (CpG position, followed by absolute number of methylated and total reads at each CpG in each sample), sample-to-class assignment (either on the command line or via a file) and setting appropriate weights (defaults are provided); see the repository README for details. This step produces tab-separated output of estimated class methylation levels.

2. Calling differentially methylated CpGs and DMRs (`diffmer dmrs`) uses the estimates from the first step and requires user-provided thresholds $(\Delta, S, \overline{\Delta}, s)$; defaults are provided. Called DMRs are output in BED format. For each DMR, we report the genomic interval, the number of CpGs in the interval, the mean absolute methylation difference, the kind of DMR (hyper- or hypo-methylation of the case class vs. the control class (`+1` or `-1`; or `0` for mixed), and finally the sequence of signed class methylation differences for each CpG in the region.

EVALUATION

We compare our approach to the existing popular approaches BSmooth and RADMeth on two real WGBS datasets, a public one on the social status of macaques, and an unpublished one on different types of uveal melanoma that is being studied at University Hospital Essen.

Datasets

The "macaques dataset" of Tung et al. (2012) consists of six DNA samples from female rhesus macaques, three individuals of low social rank and three individuals of high rank. The study states a relationship between DNA-methylation and social rank. Each sample consists of roughly 27 million CpGs at an average coverage between 11× and 14×. The data can be obtained with accession number GSE34128 from Gene Expression Omnibus[2].

The "uveal melanoma dataset" (provided by Michael Zeschnigk, Ludger Klein-Hitpass and Bernhard Horsthemke, University Hospital Essen) compares two different types of human uveal melanoma tissue. Each type is represented by two samples with roughly 28 million CpG-sites at about 35× coverage. The data is currently not publicly available.

Dependence of the Solution on Parameters

The number of discovered DMRs (and DMCs within DMRs) is influenced by two distinct types of parameters: the α weights for estimation, and the thresholds $(\Delta, S, \overline{\Delta}, s)$ for the caller. We estimated the α weights as follows.

	estimated weights			used weights		
dataset	α_{class}	α_{diff}	$\alpha_{\text{space};i}$	α_{class}	α_{diff}	$\alpha_{\text{space};i}$
macaque	34.83	33.50	$\frac{\ln(2)}{0.00180 \cdot d_i + 0.054}$	34.83	16.75	$\frac{3\ln(2)}{0.00180 \cdot d_i + 0.054}$
melanoma	17.82	11.09	$\frac{\ln(2)}{0.00224 \cdot d_i + 0.033}$	17.82	11.09	$\frac{3\ln(2)}{0.00224 \cdot d_i + 0.033}$

The weights α_{diff} and $\alpha_{\text{space};i}$ have the biggest impact on the result: A larger value of $\alpha_{\text{space};i}$ yields smoother class methylation values, while α_{diff} controls the number of resulting DMCs and therefore the number and size of DMRs and detection sensitivity. With the parameters as stated above, we observed little smoothing across adjacent CpGs. Therefore we scaled $\alpha_{\text{space};i}$ with a factor of 3. On the macaque data we found few and small DMRs with the estimated α_{diff}, so we scaled it by 0.5.

[1] https://bitbucket.org/genomeinformatics/diffmer/

[2] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE34128

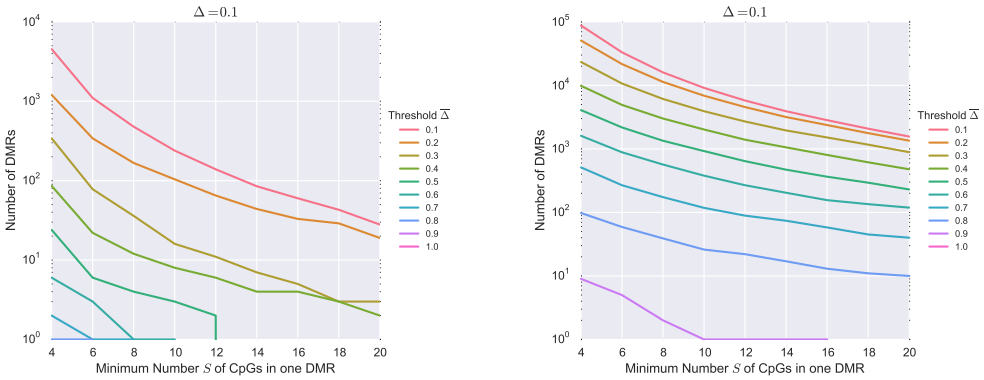

Figure 2. Total number of DMRs (y-axis: log-scale) as a function of the threshold $\overline{\Delta}$ (mean methylation difference; color) of the required minimum size S of a DMR (x-axis). Left: macaques. Right: uveal melanoma.

Figure 3. Venn diagrams comparing the number of DMCs in DMRs detected by diffmer, BSmooth and RADMeth. Left: macaques; right: uveal melanoma. Parameters were chosen to obtain an approximately equal number of DMCs in DMRs for each method.

With the weights fixed in this way, we set $\Delta := 0.1$ and varied the thresholds S and $\overline{\Delta}$ (minimum DMR size and mean methylation difference). The resulting number of discovered DMRs is shown in Figure 2. The total size of detected DMRs in both datasets differs strongly for the same parameters: There are much less differential methylated regions detected in the macaque dataset than in the uveal melanoma dataset, although the overall number of CpGs in the genome is comparable. Because the weights suggest that intra-class variability is considerably lower, we may speculate that there is indeed much less differential methylation present in the macaque dataset. On the other hand, the macaque dataset also has a smaller coverage, which produces more conservative estimates of methylation difference by shrinkage. It becomes clear that the optimal choice of thresholds depends on the organism, cell type and study question. From a practical perspective, one might want to verify a small number of most promising DMRs at interesting genomic locations (see below).

While the results do not strongly depend on the discretization granularity ε, the running time does. We measured the wall clock time for the estimation step on a modern)Intel® Core™ i7-4790 processor at 3.6 GHz CPU clock with 16 GB RAM, using a single thread. On the macaque dataset (6 samples), for $\varepsilon = 0.2, 0.1, 0.05, 0.02$, estimation took 8, 27, 240 and 7430 minutes, respectively; on the uveal melanoma dataset (4 samples), 8, 27, 235 and 7400 minutes (approx. 5 days). Single-thread times for RADMeth for macaque and uveal melanoma are 780 and 1600 minutes, for BSmooth 600 and 260 minutes, respectively. We suggest $\varepsilon = 0.05$ as a good compromise between precision and speed for our method; it is then faster than the other ones. The computation may be parallelized by chromosome to benefit from multiple cores; this has not yet been implemented. The times for DMR calling after estimation are insignificant in comparison (seconds).

Comparative Evaluation

We compare the DMRs discovered by our method to those discovered by BSmooth and RADMeth. Thresholds were chosen to obtain a comparable number of DMCs within DMRs from each method.

For our method, the weights were as described above, precision $\varepsilon = 0.05$, DMC threshold $\Delta = 0.1$ for macaques and $\Delta = 0.4$ for uveal melanoma, minimum DMR size $S = 4$, mean methylation difference $\overline{\Delta} = 0.25$ for macaques and $\overline{\Delta} = 0.5$ for uveal melanoma, and $s = 0$ (no skipping).

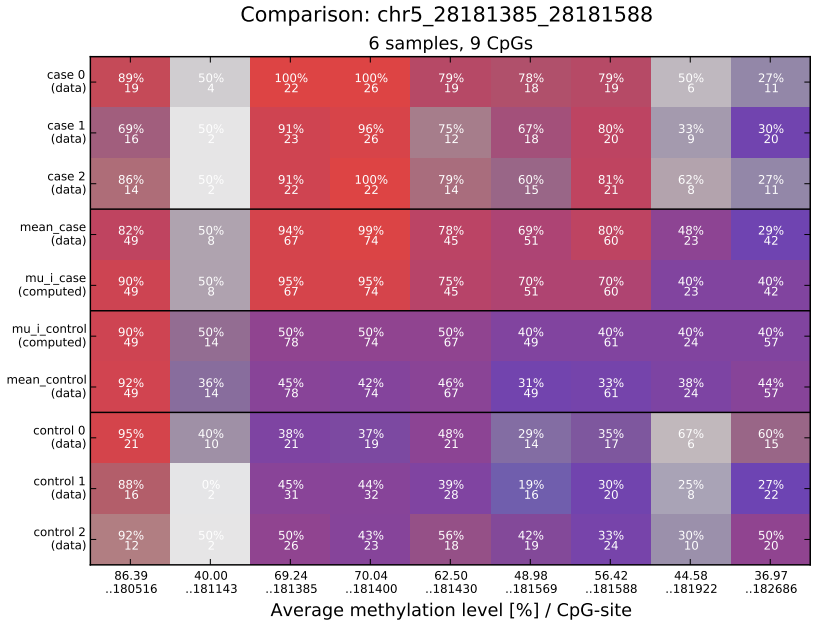

Figure 4. A DMR from chromosome 5 of the macaque data exclusively found by diffmer. Each column represents a CpG-position; each of the three top and bottom rows represents a sample. The two left- and rightmost columns are not included in the DMR and are presented for contrast. Numbers in matrix cells show methylation levels in percent and coverages; color corresponds to methylation level (red: full, blue: none); the saturations corresponds to coverage. The outer and inner middle rows show empirical and computed class methylation levels, respectively. The x-axis shows average methylation levels and (relative) genomic positions for each CpG.

For BSmooth, we followed the guidelines provided by its authors and kept only positions with coverage ≥ 2 in all samples. The t-statistic cutoff was set to 4.5 for uveal melanoma and to 4.0 for the macaque dataset. We reported only DMRs with a minimum length of 4 and a mean methylation difference ≥ 0.19 for macaques and ≥ 0.41 for uveal melanoma.

For RADMeth, we restricted DMRs to be between 1 and 99 bp in length and required an FDR-corrected p-value below 0.01. A minimum DMR size of 4 was required, as for the other methods. For macaques, we did no further filtering because the results were already rare, the uveal melanoma DMRs were additionally required to have a methylation difference of at least 0.4.

Figure 3 shows the number of DMCs in DMRs detected by each method and the size of the intersections of such DMCs (the circle size is proportional to the number of elements in it). For both datasets, we can make similar observations: the results overlap, but do largely not agree (this holds especially for the macaque data). Each method detects a large number of DMRs that are not detected by any of the other methods.

To illustrate that our results not only differ from those of other methods, but that our unique discoveries are meaningful DMRs, we exemplarily show two DMRs (one for each dataset) that are missed by the other methods. Figure 4 shows a DMR from chromosome 5 in the macaque data, and Figure 5 shows a DMR from chromosome 1 in the melanoma data. The plots visually compare measured and computed values by coloring the values (red: full methylation, blue: no methylation). The contrast between the rows and the numbers show clear differential methylation.

DISCUSSION AND CONCLUSION

We presented a method for estimating class methylation levels and differences between them from low-coverage and few-sample WGBS data, based on a Bayesian approach and written as a LASSO-like convex optimization problem. We also gave an efficient dynamic programming solution for a discretized version of the problem. Based on class methylation differences, we gave a versatile method to call DMRs with several threshold parameters. Our initial findings show that existing DMR discovery methods show large differences in reported DMRs and DMCs. We discover a large set of putative DMRs not reported by any other method, and we assure the reader that the presented examples are quite representative.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1287v3 | CC-BY 4.0 Open Access | rec: 2 Sep 2015, publ: 2 Sep 2015

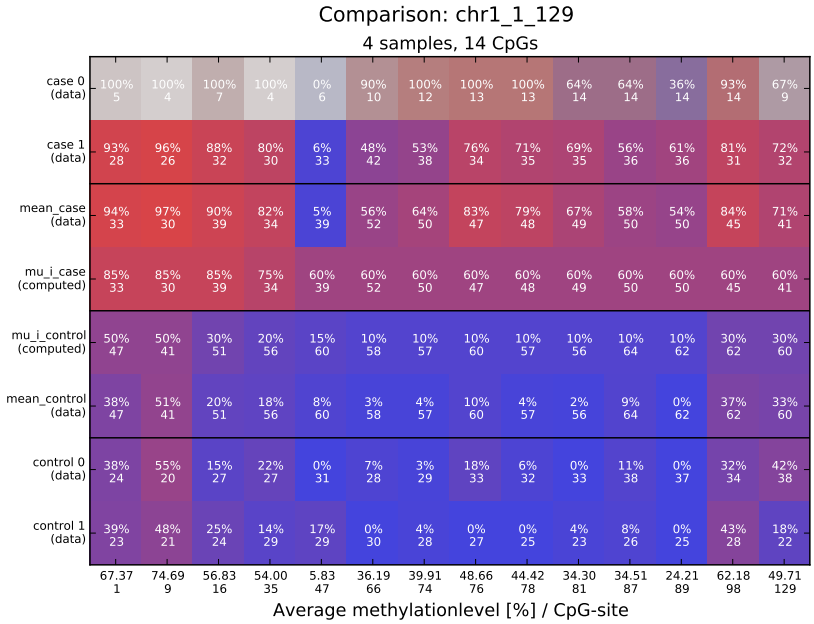

Figure 5. A DMR from chromosome 1 of the melanoma data exclusively found by diffmer (cf. Figure 4)

The statistically inclined reader may have noted that we do not report p-values for DMRs. We do not consider them a necessity, because (a) we intend to use our tool for "discovery" mode (each DMR of interest would have to be verified on a larger patient set by an independent method, such as methylation-sensitive PCR on a larger patient cohort), and (b) we report conservative shrunken estimates of methylation differences, which can be used to sort DMCs or DMRs.

In future work, we will systematically evaluate the discovered DMRs on the uveal melanoma dataset and attempt to interpret them biologically. We will also attempt to find better explanations for the striking differences in reported DMRs between different tools.

ACKNOWLEDGMENTS

We thank Michael Zeschnigk, Bernhard Horsthemke and Ludger Klein-Hitpass for providing access to the uveal melanoma dataset.

REFERENCES

Akman, K., Haaf, T., Gravina, S., Vijg, J., and Tresch, A. (2014). Genome-wide quantitative analysis of DNA methylation from bisulfite sequencing data. *Bioinformatics*, 30(13):1933–1934.

Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, 13(10):705–719.

Dolzhenko, E. and Smith, A. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics*, 15(1):215.

Gartler, S. M. and Goldman, M. A. (2005). X-chromosome inactivation. *eLS. John Wiley & Sons Ltd, Chichester*.

Hansen, K. D., Langmead, B., and Irizarry, R. (2012). Bsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13(10):R83.

Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Zhang, K., Irizarry, R. A., and Feinberg, A. P. (2011). Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 43(8):768–775.

Hirschberg, D. S. (1975). A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, 18(6):341–343.

Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. *Nature*, 366(6453):362–365.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, a. H., Thomson, J. a., Ren, B., and Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322.

Saito, Y., Tsuji, J., and Mituyama, T. (2014). Bisulfighter: Accurate detection of methylated cytosines and differentially methylated regions. *Nucleic Acids Research*, 42(6).

Sun, D., Xi, Y., Rodriguez, B., Park, H. J., Tong, P., Meong, M., Goodell, M. a., and Li, W. (2014). MOABS: model based analysis of bisulfite sequencing data. *Genome biology*, 15(2):R38.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):pp. 267–288.

Tibshirani, R. and Wang, P. (2007). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 9(1):18–29.

Tung, J., Barreiro, L. B., Johnson, Z. P., Hansen, K. D., Michopoulos, V., Toufexis, D., Michelini, K., Wilson, M. E., and Gilad, Y. (2012). Social environment is associated with gene regulatory variation in the rhesus macaque immune system. *Proceedings of the National Academy of Sciences*, 109(17):6490–6495.

Xu, X., Su, S., Barnes, V. A., Miguel, C. D., Pollock, J., Ownby, D., Shi, H., Zhu, H., Snieder, H., and Wang, X. (2013). A genome-wide methylation study on obesity: Differential variability and differential methylation. *Epigenetics*, 8(5):522–533.

SUPPLEMENT

Estimation of Weights

We illustrate exemplarily how the weight α_{diff} in the objective function (8) can be chosen. On the uveal melanoma dataset, we included all CpG-sites with a coverage above 100 in all samples. For these sites, we took the class methylation values $\mu_{i,\oplus}$ and $\mu_{i,\ominus}$ as the mean of the corresponding $\mu_i^{(t)}$. Note that for the available number of two samples per class, this is also the median. We tabulated differences $|\mu_{i,\oplus} - \mu_{i,\ominus}|$ and obtained two candidates for α_{diff}, one via the mean and one via the median of the differences. Figure 6 shows the Q-Q plot for $\alpha_{\text{diff}} = 11.09$ obtained by equating the empirical mean of the differences to the theoretical one $1/\alpha_{\text{diff}}$ of π_{diff}. The agreement is quite satisfactory.

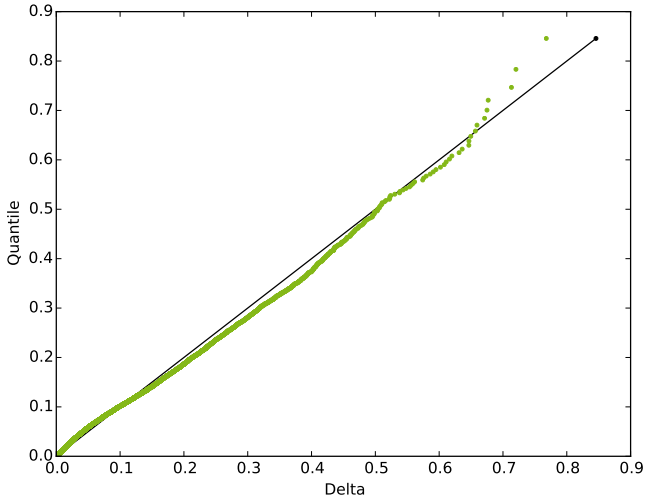

Figure 6. Q-Q plot for the estimate of α_{diff}. X-axis: empirical quantiles (class methylation differences). Y-axis: theoretical quantiles for $\alpha_{\text{diff}} = 11.09$ obtained by equating empirical with theoretical mean.