# MethCP: Differentially Methylated Region Detection with Change Point Models

Boying Gong[1] and Elizabeth Purdom[2,*]

[1] Division of Biostatistics, University of California, Berkeley.
[2] Department of Statistics, University of California, Berkeley.
jorothy_gong@berkeley.edu, epurdom@stat.berkeley.edu

**Abstract.** Whole-genome bisulfite sequencing (WGBS) provides a precise measure of methylation across the genome, yet pose a challenge on identifying regions that are differentially methylated (DMRs) between different conditions. A number of methods have been proposed, mainly by performing tests on individual methylation locus and combining significant positions. While this approach can globally control locus-level error rates, in the detection of a region, these methods often result in poor estimates. We develop a DMR detecting method MethCP for WGBS data based on change point detection, which naturally segments the genome and provide region-level differential analysis. We demonstrate the performance of MethCP on a senescent cell study, an Arabidopsis dataset and a simulated dataset. We compare our method to four developed methods and we show MethCP is more accurate in detecting the complete DM region.

**Availability:** An Bioconductor package MethCP is upcoming (https://github.com/boyinggong/MethCP).
**Key words:** differential methylation, bisulfite sequencing, change point

## 1 Introduction

DNA methylation is an important epigenetic mechanism for regulation of gene expression. DNA methylation is a process by which methyl groups are added to DNA cytosine (C) molecules, with the methylation of promoter sites in particular associated with decreased gene expression and methylation in gene bodies associated with increased gene expression.

Whole-genome bisulfite sequencing (WGBS) allows for precise measurement of DNA methylation across the genome. Briefly, when DNA is treated with bisulfite, the *unmethylated* C nucleotides are converted to uracil (U) leaving methylated C nucleotides unchanged. In the process of PCR amplification, these uracil nucleotides that generated from unmethylated nucleotides will be converted to a thymine (T) and no longer match the original DNA. In this way, they can be distinguished from the methylated nucleotides that remain as a C. Sequencing of bisulfite-treated DNA and mapping of the sequenced reads to a reference genome then provides a quantification of the level of methylation at each C nucleotide. Specifically, for each C nucleotide in the reference genome, the percentage of sequenced reads overlapping that nucleotide that instead have a T is an estimate of the percentage of unmethylated DNA at that position in the original sample.

Analysis of BS-Seq data across samples allows for the comparison of methylation patterns across population groups. Of common interest is to identify regions of the genome where methylation patterns differ across populations of interest. Such regions are called differentially methylated regions (DMRs), analogous to the very common task of differential expression analysis for mRNA sequencing. Indeed, countless methods have been developed for the identification of differential mRNA expression, from mRNA-Seq or microarray studies (see Soneson and Delorenzi (2013) for a review). These methods are ultimately methods based on the comparisons of counts in different populations, such that mRNA DE techniques have been used successfully for other types of sequencing data that also results in counts, such as 16S sequencing of microbiome data and Chip-Seq data (Anders and Huber, 2010). These mRNA methods, however, are not applicable to the detection of DMRs because of the structure of WGBS data. With WGBS, the input data are the counts of C *and* T reads covering a nucleotide (usually summarized by a proportion of C reads), not a single count. Furthermore, the depth of sequencing is generally fairly low so that the denominator of these proportions is fairly small, necessitating careful consideration of the statistical properties of these proportions. Finally, unlike RNA-Seq, the regions are not known *a priori* but must be detected.

Several methods have been developed to identify regions from BS-Seq data that show differential methylation between groups of samples. One common strategy is to perform a per-nucleotide test of differential methylation with a test statistic that appropriately accounts for the proportions and then use these significant results to determine the DMRs. MethylKit (Akalin *et al.*, 2012) performs either a logistic regression test or Fisher's exact test per C nucleotide; RADMeth (Dolzhenko and Smith, 2014) uses a beta-binomial regression, and a log-likelihood ratio test; DSS (Park and Wu, 2016; Feng *et al.*, 2014; Wu *et al.*, 2015) uses a Bayesian hierarchical model with beta-binomial distribution to model the proportions and tests for per-locus significance with a Wald test.

Other methods use the fact that methylated loci are often found in extended regions and include local dependency between neighboring loci to improve their per-locus test. HMM-DM (Yu and Sun, 2016b,a) and HMM-Fisher (Sun and Yu, 2016; Yu and Sun, 2016a) both use Hidden Markov Models along the genome to account for the dependency between loci. BSmooth (Hansen *et al.*, 2012) and BiSeq (Hebestreit *et al.*, 2013) both use local likelihood regression to estimate an underlying methylation curve across the loci, and then test for differences in this mean curve between populations. BSmooth provides only a signal-to-noise ratio as test-statistic, without a known distribution, leaving it to the user to determine an appropriate cutoff. BiSeq, specifically developed for targeted bisulfite sequencing data, fit the beta regression model on the smoothed methylation levels and perform the testing with a Wald test.

For many of these methods, the emphasis is on finding well-performing per-locus tests. The determination of the actual DMR is often just based on whether significant loci are adjacent. Short significant segments are then discarded after combining neighboring significant loci. The methods that account for dependency across the genome have smoothness constraints or transition probability assumption across the genome, implicitly making it more likely that adjacent loci are found to be significant and thus creating more coherent regions. Although methylation levels along the genome are highly correlated, the changes can be either progressive or abrupt. And the cytosine density across the genome is not uniform. Both make the assumption of constant smoothing window or transition probability at times problematic.

We propose a segmentation approach, `MethCP`, for finding DMRs from BS-Seq data. `MethCP` uses as input the results of a per-nucleotide test-statistic, like one of the ones described above, and uses this input to segment the genome into regions and identify which of those regions are DMR. We show via simulations that this method more accurately identifies regions differentially methylated between groups, as compared to competing methods. We also illustrate the performance of `MethCP` on experimental data, and show that its behavior on experimental data mirrors that of the simulations. `MethCP` also has the advantage of being flexible as to the specific test procedure that is fit per-nucleotide. Therefore, `MethCP` can be immediately used with a variety of test statistics that go beyond comparing population groups.

## 2 Methods

`MethCP` assumes as input the results of a per-locus test of significance. Let $T_1, T_2, \cdots, T_K$ be the statistics, indexed by the location of $K$ methylation loci. We assume for now that the test statistics are (asymptotically) normally distributed, such as z-statistics or Wald statistics for testing equality of a proportion between two populations, and in 2.1 we extend this approach for other test-statistics.

The main steps of `MethCP` are to 1) segment the test statistics $T_k$ into regions of similar values, and then 2) assign a p-value per region as to whether the region is a DMR.

We segment the $T_k$ into regions of similar levels of significance based on the Circular Binary Segmentation (CBS) algorithm of Olshen *et al.* (2004), which was originally developed for segmentation of DNA copy number data. Binary segmentation methods involve testing over all of the possible breakpoints (loci) for whether there is a change in the mean of $T$ at location $i$, and the CBS algorithm performs a binary segmentation per chromosome, and adapts the algorithm so as to view the data from a chromosome as if it lies on a circle, and segments the circle into two arcs.

Briefly, for each possible arc defined by, $1 \leq i < j \leq K$, the likelihood ratio test statistic $Z_{ij}$ is calculated by comparing the mean value of $T$ found in the arc from $i+1$ to $j$ with that found in the remaining circle. To find a significant breakpoint, CBS determines whether the statistic $Z = \max_{1 \leq i < j \leq K} |Z_{ij}|$ is significantly larger than 0. If so, this implies a detection that the arc $(i+1, j)$ has a significantly different mean than the remaining arc and the two arcs are declared to be separate segments. The procedure is then applied recursively on each resulting segment until no more significant segments are detected.

Once the segmentation procedure has been completed, the question remains to determine which of these detected regions correspond to significant DMRs. Indeed many of the regions resulting from the segmentation were segmented because a differential region split the undifferentiated region into two. To classify these regions, we use meta-analysis principles to aggregate the single-locus statistics and obtain one single statistic per region, to which we apply significance tests.

Suppose for each locus $i$, the effect size $R_i$ has approximate normal distribution with estimated variance $\hat{\sigma}^2(R_i)$. $R_i$ and $\hat{\sigma}^2(R_i)$ are used to calculate $T_i$ such that $T_i = \frac{R_i}{\hat{\sigma}(R_i)}$, which we have assumed above to follow a Gaussian distribution. For a region with a set of loci $r$, the weighted effect size is given by

$$R_r^* = \frac{\sum_{i \in r} w_i R_i}{\sum_{i \in r} w_i},$$

with estimated variance given by

$$\hat{\sigma}^2(R_r^*) = \frac{\sum_{i \in r} w_i^2 \hat{\sigma}^2(R_i)}{(\sum_{i \in r} w_i)^2},$$

where the weights $w_i$ signify the contribution of locus $i$.

Typically in meta-analysis applications, $w_i$ is set to be $\hat{\sigma}(R_i)^{-1}$(Borenstein *et al.*, 2009). $\hat{\sigma}(R_i)^{-1}$ will be closely related to the overall coverage of the locus, $C_i$, assuming that appropriate methods which account for the variability in the counts are used to calculate $T_i$. Alternatively, for example when $\hat{\sigma}(R_i)$ is not available, we can use $w_i = C_i$, explicitly giving larger weights for high coverage loci.

A region-based test statistic is therefore calculated by $T_r^* = \frac{R_r^*}{\hat{\sigma}(R_r^*)}$. Based on our Gaussian distribution assumptions on the individual $T_i$, we call the region significant if $|T_r^*| > z_{\alpha/2}$, where $\alpha$ is the significance level.

### 2.1 Generalizing beyond $z$-statistics

The above approach relies on input statistics that are Gaussian. This can be limiting, since methods often produce other types of statistics, such as Fisher's exact test implemented by methylKit and log-likelihood ratio test from RADMeth. For this reason, we give a further adaptation in `MethCP` so as to be applicable for any locus-based parametric statistics that result in valid p-values. Let $p_1, p_2, \cdots, p_K$ be the p-values indexed by location of methylation loci. For segmenting the genome into regions, we use the standard transform of the $p$-values to Z-scores,

$$z_i = \left[2\mathbb{1}(R_i \geq 0) - 1\right]\Phi^{-1}(1 - p_i/2),$$

where $\Phi$ is the cumulative distribution function of standard Gaussian. `MethCP` then performs CBS on the $z_i$'s to segment the genome.

Region-level statistics can be obtained by aggregating p-values using Fisher's combined probability test (Fisher, 1934) or Stouffer's weighted Z-method (Stouffer *et al.*, 1949; Whitlock, 2005). Namely, for a region with a set of loci $r$, let

$$T_{\text{Fisher}} = -2\sum_{i \in r} \log p_i,$$

$$T_{\text{Stouffer}} = \frac{\sum_{i \in r} w_i \Phi^{-1}(1 - p_i)}{\sqrt{\sum_{i \in r} w_i^2}},$$

where $w_i$ can be chosen to be constant or given by coverage $C_i$. We test $T_{\text{Fisher}}$ against $\chi^2_{2|r|}$, and $T_{\text{Stouffer}}$ against standard Gaussian for significance.

### 2.2 Quantifying Region Alignment

To quantify the performance of the methods or similarity of DMR sets detected by different methods, we need to define some measures for whether a region was successfully detected. One simple solution is just to calculate measures of specificity and sensitivity based on whether individual loci were correctly called to be in a DMR or not. However, since our goal is to correctly detect regions, this is unsatisfactory. For example, such a strategy would consider equal a method that finds every other locus to be DMRs compared to a method that finds a single, continuous region that is just too small by 50%. The second method, though missing the same number of loci, manages to make a unified region while the first calls a large number of single-locus regions which will require some kind of post-processing to make sense of.

Because no method finds the exact region, in determining the accuracy of an entire region there must still be a trade-off in finding the entire true region versus how many extra loci are included that are not part of the true region. Therefore, we chose to set a parameter $\alpha \in (0, 1]$ that determines the percentage of overlap required in order to be considered as having successfully detected a region. We then calculate true positive rates (TPR) and false positive rates (FPR) that vary with $\alpha$.

Furthermore, we will see that some DMR methods are biased toward longer or shorter regions (Section 3), which can make comparing methods difficult. In order to account for different biases in length of regions found (in the following, we refer to number of CpGs in a region as the length of the region), we calculate the percent overlap between a detected region and a true region using three different denominators: that of the detected region, that of the true region and that of the union of the detected and true ones. The three measures can be interpreted as the local measures of precision, recall and Jaccard index. The first two allowed us to distinguish as to whether methods had a high percentage of the true region detected, versus if a high percentage of the detected region was truly differentially methylated. The local Jaccard index allows us to measure the similarity between detected and true regions symmetrically. Our framework of evaluation is closely related to supervised measures such as directional Hamming distance and segmentation covering in the image segmentation literature (Huang and Dom, 1995; Pont-Tuset and Marques, 2016).

We demonstrate our definitions using the local measure of precision – i.e., the overlap is determined by the proportion of the detected region that intersects the truth. Denote the detected and true region set as $\mathcal{R}^d$ and $\mathcal{S}^t$, respectively. To determine whether a detected region $R_i^d \in \mathcal{R}^d$ was a true positive (TP), we use the local measure of precision and calculate the true positive indicator for $R_i^d$:

$$TP_i^d = I\left\{\max_{S_j^t \in \mathcal{S}^t} \frac{|R_i^d \cap S_j^t|}{|R_i^d|} \geq \alpha\right\},$$

where $|R_i^d|$ is the length of the detected region $R_i^d$, and $\max_{S_j^t \in \mathcal{S}^t} |R_i^d \cap S_j^t|$ is the maximum overlapping length of $R_i^d$ with a true region. Note that we take the maximum over all true regions to account for the fact that a detected region may overlap multiple true regions (and vice versa). From the $TP_i^d$ definitions, we calculate the total true positive (TP), false positive (FP), and false negative (FN) as a function of $\alpha$:

$$\begin{aligned}
TP^d(\alpha) &= \sum_{R_i^d \in \mathcal{R}^d} I\left\{\max_{S_j^t \in \mathcal{S}^t} \frac{|R_i^d \cap S_j^t|}{|R_i^d|} \geq \alpha\right\},\\
FP^d(\alpha) &= \sum_{R_i^d \in \mathcal{R}^d} I\left\{\max_{S_j^t \in \mathcal{S}^t} \frac{|R_i^d \cap S_j^t|}{|R_i^d|} < \alpha\right\},\\
FN^d(\alpha) &= \sum_{S_j^t \in \mathcal{S}^t} I\left\{\max_{R_i^d \in \mathcal{R}^d} \frac{|R_i^d \cap S_j^t|}{|R_i^d|} < \alpha\right\},
\end{aligned}$$

where the false negative is interpreted as the number of true regions that do have overlap greater than $\alpha$ with any detected positives.

The above formulas for TP, FP, and FN can be extended to use local measure of recall or Jaccard index by adjusting the denominator in calculating overlap. For example, measuring the number of TP, we have:

$$TP^t(\alpha) = \sum_{S_j^t \in \mathcal{S}^t} I\left\{ \max_{R_i^d \in \mathcal{R}^d} \frac{|R_i^d \cap S_j^t|}{|S_j^t|} \geq \alpha \right\},$$

$$TP^J(\alpha) = \sum_{R_i^d \in \mathcal{R}^d} I\left\{ \max_{S_j^t \in \mathcal{S}^t} \frac{|R_i^d \cap S_j^t|}{|R_i^d \cup S_j^t|} \geq \alpha \right\}.$$

*Calculating True Negatives* Calculating the total number of true negatives would require a calculation of the number of detected regions that were truly not significant (i.e. equally methylated). However, an equally methylated region is a more nebulous quantity for a region (unlike for loci). Unlike the different DMRs, all equally methylated regions are equivalent from the point of view of all of these methods: arbitrarily defining separate regions within a large block of equally methylated regions could not be detected by any method. Instead, we use the following formula to get an estimation of the total number of true negatives:

$$\frac{\text{\# CpGs in Total - \# CpGs in (TP + FP + FN)}}{\text{Average True DMR Length}}$$

## 3 Results

We implement our method as well as BSmooth, HMM-Fisher, DSS, methylKit on both real and simulated data. Our method, `MethCP`, was run using the statistics of both DSS and methylKit as input (hereinafter referred to as MethCP-DSS and MethCP-methylKit). To be fair between methods, we remove the coverage filter for individual loci, as it varies by methods. Furthermore, the reads in symmetric CpG sites are collapsed. We set the same length filter (3 loci) and absolute mean methylation level difference filter (0.1) for DMRs, where the numbers are the default of the majority of methods. We shorten the smoothing window of BSmooth from default 1000 bps to 500 bps, which gives better results for our simulated dataset. For DSS, we use moving average smoothing, which is recommended in the documentation. For methylKit, the output is DM loci rather than DMRs. We combine neighboring significant loci to DMRS. All other implementation parameters other than the significance level (test-statistics cutoffs) were left at the default values.

### 3.1 Simulation Study

*Generation of Simulated Data* We generate simulated BS-Seq data by the following procedure adapted from Yu and Sun (2016a). We assume there are $K$ loci in the simulated genome and two groups of samples ("treatment" and "control"), each of size $n = 3$ to compare. We designate regions within this genome to be classified as DMR by generating region size (number of CpGs) from a negative binomial distribution $NB(r = 6, p = 0.25)$. We further require that the number of CpGs to be greater than 3 in each region. The starting positions of the DMR were chosen by random sampling. This divided the genome into differentially methylated and equally methylated regions.

To mimic the read coverage and the methylation ratio in real datasets, the actual sequencing counts were generated based on a human senescent cells dataset (Cruickshanks *et al.*, 2013) described in Section 3.2 as follows. To determine the total read coverage, we randomly sampled from the observed coverage distribution of each locus in the human data set. The number of reads determined to be methylated per locus was based on a binomial distribution, with the probability of proportion depending on what treatment group the sample was in and if the locus was in a DMR or not. For samples in the control group or for loci in the equally methylated regions, the binomial probability parameter was chosen from the observed distribution of the per-locus average methylation ratio in the senescent cells dataset. For DM regions, each DM region was randomly assigned one of five beta distributions from which the methylation probability of the treatment samples would follow; in addition, we require that the absolute difference between the mean of binomial distribution in treatment and in the corresponding control group is at least 0.2, which eliminated some of the five beta distributions from consideration. These beta distributions represent five different methylation levels, from poorly methylated to highly methylated (specific parameters of the beta distribution are given in Table 1). Then the locus methylation probability for samples in the treatment group was generated according to the beta distribution chosen for that DMR region. To take into account the high correlation of methylation levels between neighboring CpGs, we simulate smoothing DMR boundaries. For a DMR of length $l$, a region of length $w \sim \text{Unif}(0.1l, 0.3l)$ is added to each side of the DMR where the methylation probability is given by a mixture of treatment and control. The weights of the treatment group decrease as we move to the edge of the DMR. In this paper, We show only results with the smoothing boundary, but simulation without smoothing boundaries give similar results.

Table 1: Parameters of beta distributions for simulating the methylated counts in the treatment group.

| Distribution | (a) | (b) | (c) | (d) | (e) |
|---|---|---|---|---|---|
| $\alpha$ | 2 | 6 | 10 | 14 | 18 |
| $\beta$ | 18 | 14 | 10 | 6 | 2 |
| Mean Probability | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |

*Results of Simulation* Figure 1 shows the summary of the length of the DMRs detected by the six methods (based on the number of methylation loci) using the default significance level (or test-statistic thresholds); we also show the distribution of the lengths of true regions. `MethCP` clearly gives the closest length distribution to that of the true regions. Although we shortened the smoothing window compared to the default, BSmooth and DSS detect much larger regions. In contrast, HMM-Fisher and methylKit both detect small, fragmented regions.

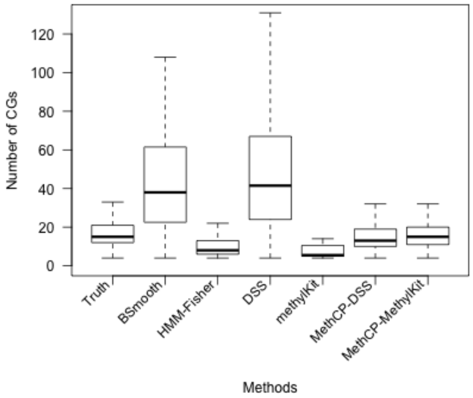
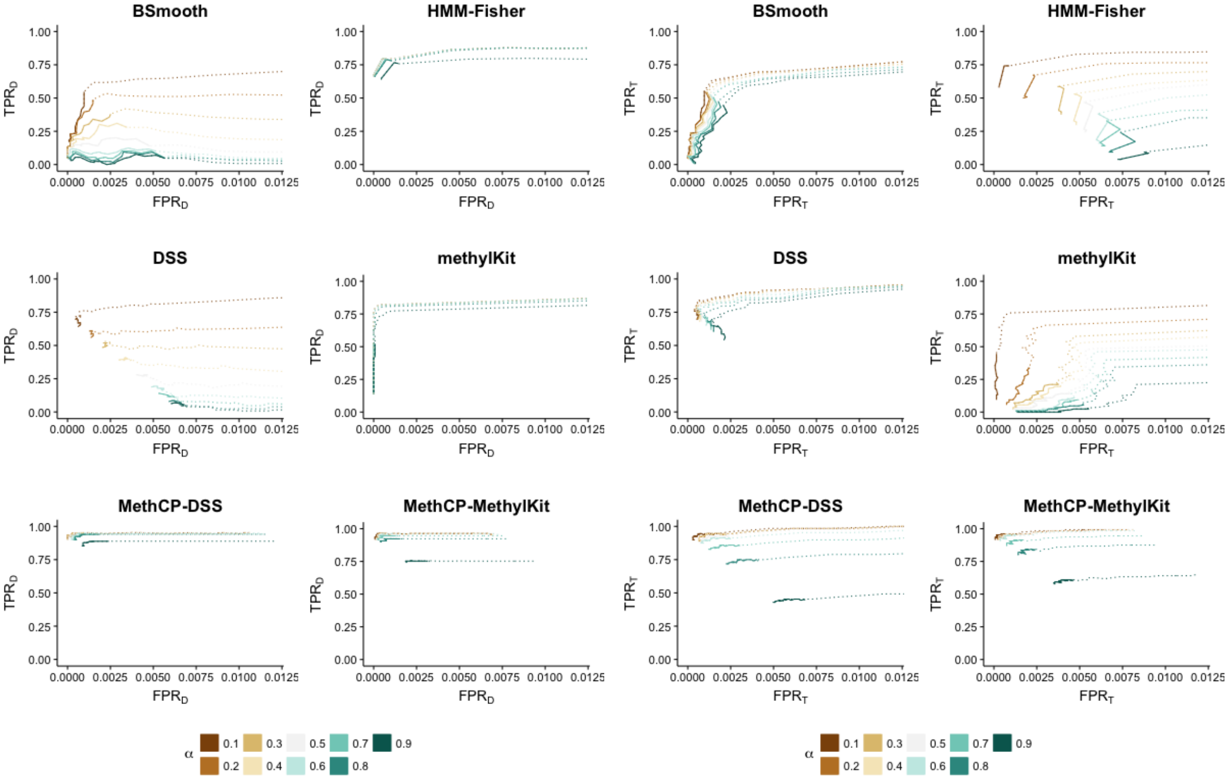


Fig. 1: **Boxplot of Number of CpGs in the DMRs.** Simulated data: Number of CpGs in the true DMRs and in the DMRs detected by the six methods compared.

We evaluated the accuracy of the methods on the simulation data based on false positive rate (FPR) and true positive rate (TPR). We determine a detected region to have successfully found a true region based on whether the percent of overlap with the truth was greater than a threshold $\alpha$. To remove bias due to the different size of the regions, the proportion was calculated in three ways: as a proportion of the size of the detected region (local precision), the size of the true region (local recall), and as a total of the union of the two (local Jaccard similarity), see Section 2.2. We plot the false positive rate versus true positive rate for both the local precision (Figure 2a) and the local recall (Figure 2b). The local precision requires that a large percent of the detected region overlap a true DMR (easier for shorter detected regions and conservative methods), while the local recall requires that a large percent of the true region be overlapped by a detected region (easier for longer detected regions).



(a) Local precision

(b) Local recall

Fig. 2: **ROC curve: Method comparison on the simulated data.** (a) False positive rate (FPR) versus true positive rate (TPR) as we change the parameter $\alpha$ for (a) local precision measure (b) and local recall measure. The solid lines indicate $p$-values thresholds smaller than 0.05 for HMM-Fisher, DSS, MethCP-DSS and MethCP-methylKit, statistics threshold larger than 4 for BSmooth (the author recommendation is 4.6), and $q$-values cutoffs less than 0.05 for methylKit. Thus, in real applications, we only focus on the regions of solid curves. For the completeness of the graph, we extend the curve to larger $p$-values.

MethCP-DSS and MethCP-methylKit detect highly similar regions despite using different test statistics as input. For small $\alpha$, MethCP achieves the highest true positive rate across all of the methods. The fact that both $\mathrm{FPR}^d$ and $\mathrm{FPR}^t$ is close to 1 suggests that most true and detected regions match in pairs, with slight disagreement in the region border. This is not the case for other four methods, which for a given FPR are usually strong in either $\mathrm{TPR}^d$ or $\mathrm{TPR}^t$, but not both, which is evidence that either a proportion of the ground-truth regions are not detected ($\mathrm{TPR}^d$ high but $\mathrm{TPR}^t$ low), or a proportion of detected regions are not overlapping the ground-truth ($\mathrm{TPR}^t$ high but $\mathrm{TPR}^d$ low). DSS and BSmooth behave similarly in that $\mathrm{TPR}^t$ varies little with $\alpha$ while $\mathrm{TPR}^d$ decreases dramatically with the increase of $\alpha$. This is an indication that both methods detect larger regions than the truth, which has been shown in Figure 1. While detecting regions wider than the true regions, BSmooth still misses at least 20% of the true regions, as indicated by the values of $\mathrm{TPR}^t$, while DSS misses a smaller proportion. HMM-Fisher and methylKit are calling fewer and smaller regions significant. MethylKit is the only method among the six that uses an FDR correction procedure. Therefore, the false positive rate is reasonably controlled especially when we use the detected length as the denominator of our measure. The DMRs identified by these two methods are generally a subset of the true regions as indicated by the high values of $\mathrm{TPR}^d$ regardless of the cutoff $\alpha$. However, they also miss a good number of regions as shown in their lower $\mathrm{TPR}^t$ ( Figure 2b). Overall, MethCP achieves the best performance in balancing this tradeoff.

### 3.2 Human Senescent Cells Study

We apply our method to a senescent cells study (Cruickshanks *et al.*, 2013). This whole genome bisulfite sequencing (WGBS) dataset is available from the Gene Expression Omnibus (GEO) with accession number GSE48580. We perform differential analysis between proliferating and senescent cells, three replicates per group. We apply MethCP-DSS to model a total of 52,683,952 CpG loci in chromosome 22. For calculating the region-based statistics, the variances of the statistics are used as weights. To call a region significant, we require the number of CpGs $\geq 3$ and the absolute mean methylation differences between groups greater $\geq 0.1$. With significance level at 0.01, 110,573 hypomethylated DMRs are detected with 149.25 CpGs/10455.11 bps on average, accounting for 31.32% of the total CpGs in the genome. 60,672 hypermethylated DMRs are detected with 20.46 CpGs/550.20 bps on average, accounting for 2.36% of the CpGs. Our result is consistent with the findings of the original paper that replicative senescent human cells exhibit widespread DNA hypomethylation and focal hypermethylation. Among the DMRs MethCP detected, while 54.94% of the hypomethylation overlap with lamin-associated domains (LADs), only 24.54% of the hypermethylation do, where the genomic coordinates of LADs are provided by the ENCODE project(Guelen *et al.*, 2008). This supports the other conclusion of the paper that hypomethylation occurs preferentially at LADs.

### 3.3 Arabidopsis Dataset

To further illustrate the performance of MethCP on real datasets, we ran BSmooth, HMM-Fisher, DSS, methylKit, MethCP-DSS and MethCP-methylKit on a WGBS Arabidopsis Thaliana data (Coleman-Derr and Zilberman, 2012) from GEO with accession number GSE39045. Our analysis aims to detect DMRs between H2A.Z-related wild-type control line and H2A.Z mutant plants, each with 6 biological replicates. Figure 3 shows the boxplot of the size of the DMRs detected (i.e. number of CpGs) under significance level $10^{-2}, 10^{-5}$ and FDR corrected level $10^{-2}$. We see that the sizes of the regions are fairly consistent across the different significance level. The relative sizes between methods resemble that of the simulated data, where DSS and BSmooth detect large regions while HMM-Fisher and methylKit identified small ones. MethCP-DSS and MethCP-methylKit identified highly similar regions despite different test statistics they use. A major function of gene-body DNA methylation is to exclude H2A.Z from the bodies of highly and constitutively expressed genes. Figure 4 shows the location of the DMRs detected. MethCP-DSS and MethCP-methylKit detect DMRs that have higher proportion of overlap with genes and lower percentage of overlap with promoter regions, and therefore better supports the conclusion.

To compare the false positives produced by six methods, we perform two permutations on the control group. The first permutes sequencing counts (methylated and unmethylated count pairs) across the samples for each CpG position. The second permutes loci positions and but keeps the sample labels, thus breaking the spatial correlation between neighboring CpGs. We then split the control group into two and run DMR

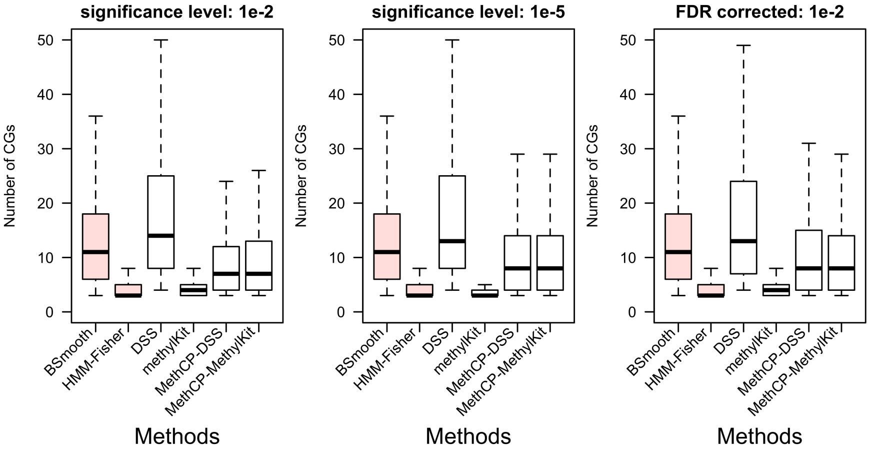


Fig. 3: **Summary of the DMR length of DMR detected for the Arabidopsis data.** Boxplot of the number of CpGs in the DMRs detected by different methods. We summarize the results under significance level $10^{-2}, 10^{-5}$ and FDR corrected level $10^{-2}$. BSmooth and HMM-Fisher are colored because we use the author recommended test-statistic cutoff (4.6) for BSmooth and significance level 0.05 for HMM-Fisher on three plots (Small significance level for HMM-Fisher returns no DMR).

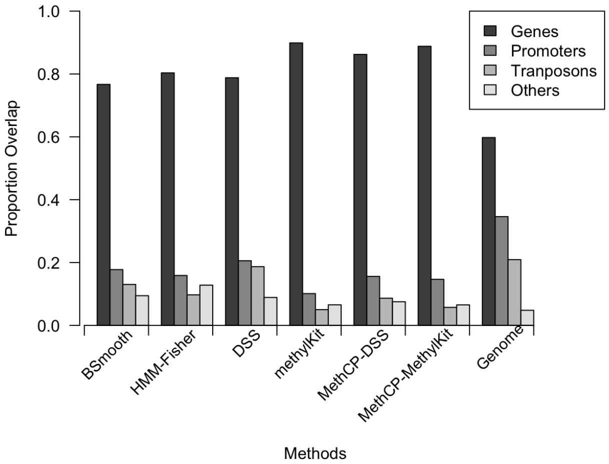
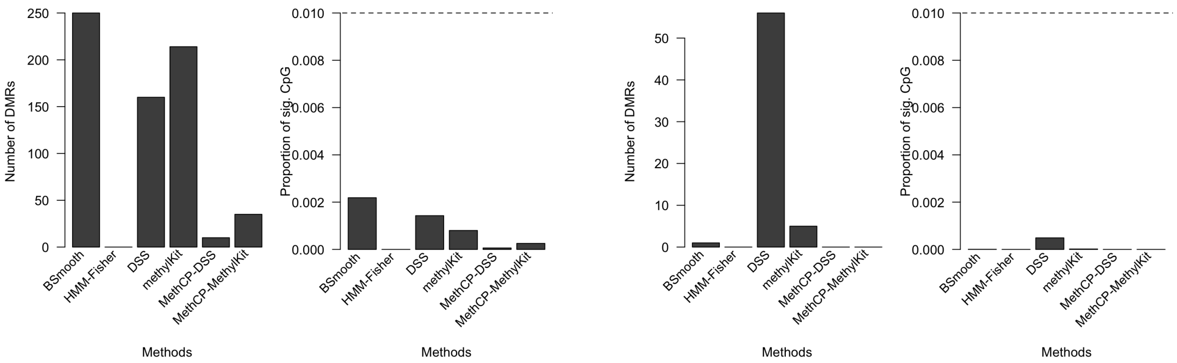


Fig. 4: **Location of DMRs detected for Arabidopsis data.** Proportion of DMRs overlapping with genes, promoters and transposons for the six methods compared.

detecting methods. Figure 5 shows the DMRs and proportion of CpGs detected by each method. In both permutations, MethCP detects fewer false DMRs; indeed for the second permutation, MethCP detects 0 DMRs. If we consider MethCP to detect significant loci based on which loci are DMRs, then MethCP also results in fewer individual loci identified as significant than the other methods. For the first permutation, MethCP detects fewer DMRs and loci than BSmooth, DSS and methylKit.



(a)

(b)

Fig. 5: **Permutation results for Arabidopsis dataset.** (a,b) Number of regions and proportion of CpGs detected when permuting the sequencing counts for each CpG. (c,d) Number of regions and proportion of CpGs detected when permuting the loci position. Methods other than BSmooth uses significance level 0.01, and BSmooth uses the author recommended test-statistic cutoff (4.6). In the proportion of CpGs plot, significance level 0.01 are marked for reference.

We would note that despite detecting fewer false positive DMRs in our permutation analysis, MethCP still remains competitive in terms of the number of DMRs it finds on the real data, as compared to the other methods (Figure 6), indicating that it is not suffering from a lack of power. DSS and MethylKit both find more regions, but we see from our permutation analysis that DSS and MethylKit tend to find many more false positives than MethCP. Furthermore, the regions found by MethylKit are small (Figure 6), suggesting that like in the simulations MethylKit may be missing or fragmenting large parts of the true DMRs.

## 4 Conclusion

We proposed a method MethCP for identifying differentially methylated regions. We presented the results of MethCP-MethylKit and MethCP-DSS on simulated and real datasets. And we showed that MethCP gives better accuracy and a lower number of false positives, as compared to existing methods.

We would also note that while we present only the case of comparing two groups, our framework is in principle flexible for general experimental design assuming an appropriate single-locus test-statistic can be calculated. Thus the method can be expanded immediately to more complicated situations, such comparing multiple groups or measurements of methylation status over time or developmental progression.

## Funding

This work has been supported in part by a DOE BER grant, DE-SC0014081.

## References

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., and Mason, C. E. 2012. methylkit: a comprehensive r package for the analysis of genome-wide dna methylation profiles. *Genome biology* 13, R87.

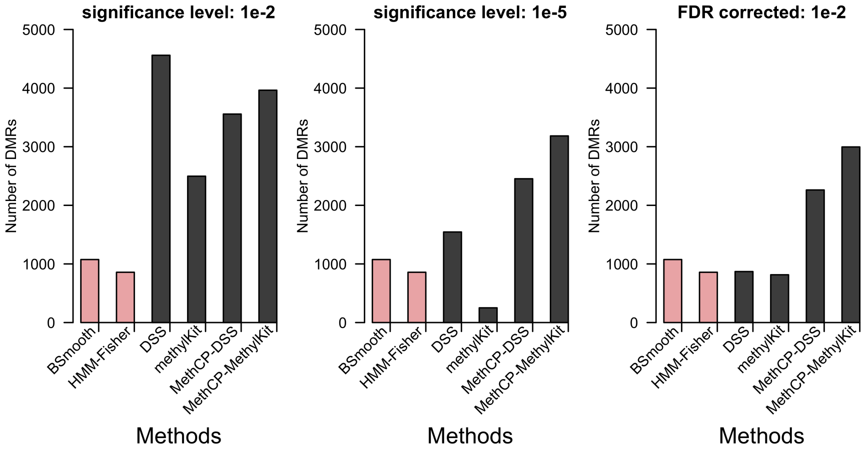


Fig. 6: **Number of DMRs detected for Arabidopsis dataset.** Number of DMRs detected by different methods. We summarize the results under significance level $10^{-2}, 10^{-5}$ and FDR corrected level $10^{-2}$. BSmooth and HMM-Fisher are colored because we use the author recommended test-statistic cutoff (4.6) for BSmooth and significance level 0.05 for HMM-Fisher on three plots (Small significance level for HMM-Fisher returns no DMR).

Anders, S. and Huber, W. 2010. Differential expression analysis for sequence count data. *Genome biology* 11, R106.

Borenstein, M., Hedges, L. V., Higgins, J., and Rothstein, H. R. 2009. *Introduction to meta-analysis*. Wiley Online Library.

Coleman-Derr, D. and Zilberman, D. 2012. Deposition of histone variant h2a. z within gene bodies regulates responsive genes. *PLoS genetics* 8, e1002988.

Cruickshanks, H. A., McBryan, T., Nelson, D. M., VanderKraats, N. D., Shah, P. P., Van Tuyn, J., Rai, T. S., Brock, C., Donahue, G., Dunican, D. S., *et al.* 2013. Senescent cells harbour features of the cancer epigenome. *Nature cell biology* 15, 1495.

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

Feng, H., Conneely, K. N., and Wu, H. 2014. A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research* 42, e69–e69.

Fisher, R. A. 1934. Statistical methods for research workers .

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., Eussen, B. H., de Klein, A., Wessels, L., de Laat, W., *et al.* 2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 453, 948–951.

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

Hebestreit, K., Dugas, M., and Klein, H.-U. 2013. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics* 29, 1647–1653.

Huang, Q. and Dom, B. 1995. Quantitative methods of evaluating image segmentation. In *Image Processing, 1995. Proceedings., International Conference on*, volume 3, pages 53–56. IEEE.

Olshen, A. B., Venkatraman, E., Lucito, R., and Wigler, M. 2004. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics* 5, 557–572.

Park, Y. and Wu, H. 2016. Differential methylation analysis for bs-seq data under general experimental design. *Bioinformatics* 32, 1446–1453.

Pont-Tuset, J. and Marques, F. 2016. Supervised evaluation of image segmentation and object proposal techniques. *IEEE transactions on pattern analysis and machine intelligence* 38, 1465–1478.

Soneson, C. and Delorenzi, M. 2013. A comparison of methods for differential expression analysis of rna-seq data. *BMC bioinformatics* 14, 91.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams Jr, R. M. 1949. The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1 .

Sun, S. and Yu, X. 2016. Hmm-fisher: identifying differential methylation using a hidden markov model and fishers exact test. *Statistical applications in genetics and molecular biology* 15, 55–67.

Whitlock, M. C. 2005. Combining probability from independent tests: the weighted z-method is superior to fisher's approach. *Journal of evolutionary biology* 18, 1368–1373.

Wu, H., Xu, T., Feng, H., Chen, L., Li, B., Yao, B., Qin, Z., Jin, P., and Conneely, K. N. 2015. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic acids research* 43, e141–e141.

Yu, X. and Sun, S. 2016a. Comparing five statistical methods of differential methylation identification using bisulfite sequencing data. *Statistical applications in genetics and molecular biology* 15, 173–191.

Yu, X. and Sun, S. 2016b. Hmm-dm: identifying differentially methylated regions using a hidden markov model. *Statistical applications in genetics and molecular biology* 15, 69–81.