A Comprehensive Evaluation of Methods for Identifying Differentially Methylation Regions in Epigenome-wide Association Studies

Saurav Mallik1, Zhen Gao1, Gabriel J Odam1, Steven Xi Chen1,\*, and Lily Wang1,\*

1Division of Biostatistics, Department of Public Health Sciences, University of Miami Miller School of Medicine, Miami, FL, 33136, USA

\*Contact Author: [lily.wang@med.miami.edu](mailto:lily.wang@med.miami.edu), [steven.chen@med.miami.edu](mailto:steven.chen@med.miami.edu)



## **Abstract**

Recent findings show that a number of tools have been developed for the analysis of Differentially Methylation Regions (DMRs). A very limited number of previous studies has been performed regarding the comparison of the methods for predefined regions. However, there is a lack of systematic evaluation of methods that analyze user defined regions. Hence, in this article, we conducted a comprehensive evaluation of the most popular software tools for user defined DMR analysis, including DMRcate, ProbeLasso, Bumphunting and comb-p. In addition to evaluating power, precision and type I error rate, we also compared several additional characteristics of the analysis results by these different methods, including the size of the DMRs, the amount of correlations between nearby CpGs within the identified DMRs, and overlap between the methods.

Key words: DMR finding tools, A-clustering, DMR co-methylation, DMR overlap, DMR size, Evaluation metrics.



**Introduction**

- DNA methylation

- measured by arrays, beta values

- DMR analysis

- why they are important

- supervised vs. unsupervised methods

- lack of systematic evaluation / comparison of methods

- previous reviews

DNA methylation is one of the most studied epigenetic mechanisms, which are stable heritable traits that can not be explained by DNA sequences. The most widely characterized DNA methylation process is the addition of the methyl group at the 5-carbon of the cytosine ring, which results in 5-methylcytosine (5-mC). When located in gene promoter, DNA methylation typically acts to repress gene transcription.

The gold standard for measuring methylation status is the whole-genome bisulfite sequencing (WGBS). However, the high cost of WGBS has limited its use in large epidemiology studies. Currently, most of the Epigenome-wide Association Studies conducted utilized the array based technologies, which provides an economical, high throughput and comprehensive alternative. For example, <450K array > <location of the probes>

More recently, the EPIC was developed to include more probes on the regulatory regions ([Logue, et al., 2017](#_ENREF_4)) the Illumina Infinium http://it.miami.edu/a-z-listing/endnote/index.htmlprovides coverage for over 850K CpG sites across the genome. <It has been shown concordance between array and bisulfite sequencing result >

It has been shown recently that regions with differentially methylated cytosine were shown to be associated with various diseases (ref). For example,

Therefore, recent research effort has gradually shifted from identifying differentially methylated positions (DMPs) to identification of differentially methylated regions (DMRs). A DMR is defined as a region with consecutive DMPs.

A number of tools have been developed for the analysis of DMRs. Chen et al. (2016) classified DMR identification methods into two classes that considers predefined regions or user defined regions. In methods that analyze predefined regions, the CpG probes are first grouped into genomic regions such as CpG islands, CpG shores, TSS200 using annotation information, and then tested for association with phenotype. In methods that analyze user defined regions, a p-value or corresponding t-statistic is computed at each CpG first, then the regions in the genome with consecutive small p-values or t-statistics are identified based on user specified criteria such as the minimum number of CpGs for the region.

Two previous studies (Klein, Chen) compared analysis methods for predefined regions. However, there is a lack of systematic evaluation of methods that analyze user defined regions.

In this study, we conducted a comprehensive evaluation of the most popular software tools for user defined DMR analysis, including *DMRcate, ProbeLasso, Bumphunting* and *comb-p*. In addition to evaluating power, precision and type I error rate, we also compared several additional characteristics of the analysis results by these different methods, including the size of the DMRs, the amount of correlations between nearby CpGs within the identified DMRs, and overlap between the methods.

**Methods**

Figure 1 shows the workflow of our simulation study. Briefly, we first obtained methylation dataset of 14 healthy controls with similar ages. Next, we performed adjacent site clustering to obtain 3063 clusters of adjacent CpG probes. The 14 samples were randomly divided into two groups and differential methylation of a small subset of the clusters were simulated by adding a small number to beta values in one of the groups.

**Dataset**

To preserve correlation patterns in real datasets, we generated simulation data by using a real dataset as input. The dataset GSE41169) from Horvath et al. (2012) [35], [36] included DNA methylation profiles of whole blood samples of 62 schizophrenia patients and 33 healthy controls from the Dutch population. The Illumina Infinium 450k Human DNA methylation Beadchip v1.2 was applied to measure the methylation status of nearly 485,577 CpGs.

For our study, we selected a total of 14 samples that satisfied two conditions, (1) these are all healthy male samples, and (2) the age-ranges of the patients related to the samples are between 20 and 30. The sample IDs of the 14 selected samples are GSM1009744, GSM1009748, GSM1009666, GSM1009667, GSM1009668, GSM1009688, GSM1009695, GSM1009746, GSM1009742, GSM1009745, GSM1009743, GSM1009681, GSM1009892 and GSM1009893

- beta values

**Adjacent Site Clustering (A-clustering)**

First before clustering analysis, we used DMRcate function rmSNPandCH to remove CpGs that are close to SNPs, cross-hybridizing and located on sex chromosomes (ref- DMRcate, original paper). We also removed those CpGs with little variations across all the samples, i.e. those CpGs with all beta values < 0.05 or CpGs with all beta values > 0.95. On the methylation arrays, beta values are computed based on the ratios of the methylated signal intensity to the sum of both methylated and unmethylated signals after background subtraction, they range from 0 (completely unmethylated ) to 1 (fully methylated).

A-clustering is an algorithm that detect sets of neighboring CpG sites that are correlated with each other. We applied A-clustering to the 14 samples selected above to obtain a total of 3,063 clusters, each consisting of at least 5 adjacent CpGs. The parameters we used are assign.to.clusters(betas = beta.value, dist.thresh = 0.5, bp.merge = 200, dist.type="spearman", method="complete") which corresponded to merging two CpGs with Spearman correlation greater than 0.5 and are within 200 bp into a cluster. The clusters obtained ranged in size of xx CpGs to xx CpGs. Fig 2 shows an example of an A-cluster.

**DMR analysis methods**

All the DMR analysis methods follows three steps (1) fits linear model to each CpGs (2) identify candidate regions in the genome with consecutive CpGs with small p-values (3) compute p-values for the candidate regions. We discuss in details for each of these steps for DMRcate, bumphunter, probeLasso and comb-p in the following.

**DMRcate**

For the the DMRcate method, first a linear model with methylation M value as the outcome variable, group status and any other covariate variables such as age or batch effects is applied to data for each individual CpG by using functions from the limma R package. *limma* is one of the most popular statistical tool for assessing differential gene expression expressions. In the application to methylation study, because *limma* shrinks variances for each CpG toward the global variance estimated from all CpGs using the empirical Bayes method, results from limma are more stable for studies with small sample sizes.

The statistic Y = t2 is calculated for each position, where t is the t-statistics from linear model corresponding to group effect. In the second step, DMRcate applies kernel smoothing using the Gaussian smoother. P-values for each position is then computed by moment matching using the method of Sattererthwaite (ref). The CpG sites with with multiple comparison corrected (via method of BH) are selected as significant CpGs.

Regions for DMRs are identified by collapsing contiguous significant CpGs that are at most lamda nucleotides from each other. The p-value for DMR is computed using Stouffer’s method ([Hoffman, 1965](#_ENREF_2)).

**Bumphunter**

In the bumphunting method, first a linear regression model M value ~ Group is applied to model differential methylation between case and control groups at each CpG site. Here M value is a logit transformation of the beta values, i.e. M value = log2 (beta value / (1- beta value). It has been shown that in the analysis of methylation data, M values have better statistical properties such as homoscedasticity (ref). Next, an optional smoothing step is involved where the loess curve fitting is applied to estimated t-statistic corresponding to the regression coefficients. Candidate regions (bumps) are identified to be clusters of consecutive probes for which all the t-statistics exceed a user defined threshold (argument xx in bumphunter function). Permutation test, which permutes sample labels to create null distribution of candidate regions, are then conducted to estimate statistical significance of the candidate regions.

The bumphunter function is implemented in *minfi* R package. Note that in the identification of regions, spatial correlation structure were used to model correlations of methylation levels between neighboring CpGs. When design of the study involves covariate variables, such as age or batch effect, bootstrap ([Rindskopf, 1997](#_ENREF_7)) option can be used to estimated statistical significance of the bumps.

####Ref Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York, NY: Chapman and Hall, 1993, p. 436.

**Probe Lasso ###Not checked before**

For the case of considering the spacing between the uneven probes, probe Lasso produces flexible (i.e., fault-tolerant) and dynamic lassos which are fit with the content of the local feature. Probe-lassos might be contemplated as consisting of a center and a radius. Just like real object, a probe-lasso can be “thrown” throughout a probe and the radius of it expands to upstream as well as downstream, that are itself centered on the targeted CpG. Notably, important (significant) DMR calling using the non-CG positions exclusively could not be promoted due to their scant distribution on the respective 450K BeadChip. Specially, probe-lasso derivation is completely depending upon user-provided thresholds as well as the input dataset. Probe Lasso computes probe-spacing for every probe belonging to the dataset. Thereafter, these data are grouped together into an epigenetic (or, genetic) types (viz., 4 CGI relations, and 7 gene features) and then converted into the quantile distributions.

A contingency is here prepared based on two user-defined parameters, “lassoStyle” and “lassoRadius”. While “lassoStyle” = max, size of the probe-lasso becomes ≤ (2×lassoRadius) bp. On the other hand, whenever “lassoStyle” = min, the probe-lassos must have the size of ≥ (2×lassoRadius) bp. Since every epigenetic (or, genetic) type has individual probe-spacing, Probe Lasso detects the epigenetic (or, genetic) type which follows the user-defined maximum (or, minimum) lassoRadius as well as derives the quantile in which it locates. Then the derived quantile is utilized to every epigenetic (or, genetic) distribution of the probe-spacings for generating the probelassos which differ depending on the epigenetic (or, genetic) feature.

Probe Lasso ([Butcher and Beck, 2015](#_ENREF_1)) is a popular DMR finding method. The implementation pf the Probe Lasso is found in a Bioconductor package, “ChAMP” ([Morris, et al., 2014](#_ENREF_5)).

**Comb-p ###Please check and update Comp-b method portion carefully as Zen has run this part**

Comb-p is basically a command-line tool and a python library ([Pedersen, et al., 2012](#_ENREF_6)). It can able to handle BED files of the p-values that are irregularly spaced. It computes auto-correlation, conjoins adjacent p-values, carries out false discovery adjustment, identifies the enriched regions (containing the series of adjacent low p-values), and allocates the significance to the enriched regions.

In the comb-p method, first of all, the correlation is measured at altering distance lags that is called as auto-correlation or, ACF). In general, a maximum distance and a single offset which segments the distance into several intervals are here

Accepted, whereas Kechris’ ([Kechris, et al., 2010](#_ENREF_3)) and various ACF implementations depend on the fixed offsets of the adjacent probes. After computing the ACF, it is utilized to apply the Stouffer–Liptak–Kechris correction (symbolized as *slk*) in which each p-value is corrected with respect to the adjacent p-values that is weighted according to the ACF. The resultant BED file contains a new column consisting of the adjusted p-value.

A p-value is pulled lower whenever its neighbors also contain low p-values (as well as little auto-correlation). On the other hand, a p-value is insignificant whenever its neighboring p-values are also relatively high. A q-value is computed depending on the Benjamini–Hochberg false discovery (FDR). The peak-finding algorithm is then applied to determine enriched regions (or, peaks) on the *z\_p* value.

Moreover, comb-p is developed as a single command-line application which identifies multiple independent sub-modules. The comb-p tool allows the uneven data structure towards the genome, fundamental auto-correlation techniques and multiple-testing corrections for enriched regions having many applications to a variety of various techniques.

Of note, the list of parameters utilized here for the four DMR finding methods are provided in Table 1.

Table : Input parameters for the candidate methods along with A-clustering method (i.e., the method for original cluster identification).

Table – different parameters used for these methods

Summary of the methods

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Single CpG | smoothing | Covariate allowed? | P-value for DMR | Software implementation |
| bumphunter | Linear model | Optional loess | Yes | Permutation of sample labels or bootstrap | bumphunter  minfi |
| DMRcate | limma | Gaussian smoothing | Yes | Stouffer’s method | DMRcate |
| ProbeLasso |  |  |  |  |  |

**Evaluation criteria**

- Power: Power is defined as the number of true positive DMRs divided by the total of the number of true positive and false negative DMRs (i.e., power=TP/(TP+FN)).

- Precision: Precision is defined as as the number of true positive DMRs divided by the total of the number of true positive and false positive DMRs (i.e., precision=TP/(TP+FP)).

- DMRsize: DMRsize is stated as the number of CpGs in any resultant DMR obtained by any DMR finding method.

- DMR co-methylation: DMR co-methylation is here defined as the average of Spearman correlation coefficient between the participating CpGs in pair-wise manner.

- Time used: The elapsed time is also measured in Second. The configuration of the used computer is: ###.

- DMR-overlap: DMR-overlap is here stated as the number of common significant resultant DMRs (i.e., the DMRs having p-value less than 0.05 and DMRsize is greater than or equal to 5) obtained by the DMR finding methods pairwise. Of note, it may happen that one DMR determined from any method can overlap with multiple DMRs identified from other method, or vice versa.

- Evaluation metrics: Three evaluation metrics (viz., true positives or TPs, false positives of FPs, and false negatives or FNs) have been utilized for our analysis. The definition of these metrics for our study and the other related information regarding these metrics are demonstrated in Table 2.

Table : The definition of the initial three factors (i.e., true positive (TP), false positive (FP), and false negative (FN)) of the confusion matrix in our simulation study.

Results

**Results of Simulation Study**

In the experiment, we observe the comparative performance of the four DMR finding methods through the alternation of *µ*.

Power

For Smaller µ value (0-0.05)

In the case of smaller µ, only a few TPs had been generated in any resultant DMR finding method for any repetition (run). Bumphunter produced highest number of TPs rather than the other three methods, whereas Probe-lasso generated lowest number of TPs among all. Overall, Bumphunter provided highest average power in this scenario (viz., 0.2023(±0.02) for miu=0.025, and 0.5002(±0.02) for mu=0.05), whereas comp-b yielded second highest power (viz., 0.0852(±0.01) for miu=0.025, and 0.3104(±0.02) for mu=0.05). But, the performance of the remaining two methods (DMRcate and Probe-lasso) were poor.

Overall, no method works well when µ value is low. In addition, it can be said that Bumphunter performs better than the others for smaller value of µ.

For Medium µ value (0.1-0.15)

In the case of medium µ, the number of TPs were increased than these in the smaller µ for the four DMR identification methods. In this case also, Bumphunter was the best performer in producing highest power (viz., 0.7451(±0.03) for miu=0.1, and 0.7805(±0.02) for mu=0.15) among the all methods, whereas DMRcate and Probe-lasso generated lower power (viz., 0.3832(±0.01) and 0.3892(±0.02) respectively for mu=0.1, and 0.4744(±0.02) and 0.4628(±0.02), respectively for mu=0.15) among the all. The remaining method Comb-p second higher average power generator.

Overall, the performance of all the methods improve significantly for medium µ value rather than these for low µ value. In this case also, Bumphunter is the best performer.

For Higher µ value (0.2-0.4)

For higher µ, the number of TPs were increased than these in the medium µ for the four methods. Alike the previous cases, Bumphunter produced highest power (viz., 0.9918(±0) for miu=0.2, 0.9923(±0) for miu=0.3, and 0.9926(±0) for mu=0.4) among the all methods, whereas Probe-lasso generated lower power among the all. Comb-p was the second highest average power generator.

Overall, for higher µ value, all methods perform well. Moreover, alike in the previous cases, Bumphunter is the best performer here.

See Figure 2 and Figure 3 for details. For details about the overall comparative performance of the four DMR finding methods, see Table 3.

Precision

For Smaller µ value (0-0.05)

On the other hand, Comb-p and DMRcate were the best performers in terms of the identified precision (viz., 0.9963(±0.01) and 1(±0) respectively for mu=0.025, and 0.9963(±0) and 0.9831(±0.01), respectively for mu=0.05). Of note, the performance of Comb-p was consistent for smaller µ, wheres Bumphunter produced lowest precision in this case (0.7058(±0.02) for miu=0.025, and 0.8722(±0.01) for mu=0.05) since it generated a lot of FPs rather than these of the other methods.

However, overall, only two methods (Comb-p and DMRcate) work well for µ<0.05.

For Medium µ value (0.1-0.15)

On the other hand, Comb-p was the best performer in terms of the identified precision (viz., 0.9917(±0) for mu=0.1, and 0.9924(±0) for mu=0.15). Specially, Comb-p performed consistently for the medium µ. DMRcate and Bumphunter yielded lowest precision (as well as inconsistent) in this case.

Overall, all the methods more or less perform well for the medium µ value.

For Higher µ value (0.2-0.4)

On the other hand, Comb-p was the best performer in terms of the precision (viz., 0.8199(±0.02) for miu=0.2, 0.8687(±0.03) for miu=0.3, and 0.9099(±0.03) for mu=0.4). Specially, Comb-p and probe-lasso performed consistently for the higher µ. DMRcate generated lowest precision in this case.

Overall, all the methods work well for the higher µ value.

Time

In addition, …….  **###time (need zen’s time calculation on comb-p)**

DMR Overlap

For Smaller µ value (0-0.05)

We also identified the overlap between the four DMR finding methods for smaller µ. Of note, since a DMR of any method can overlap with one or multiple DMRs of the other method in partially or completely, there exists some one-to-many (or, many to one) mapping in the results obtained from the methods. For µ=0.025 and repetition=1, we obtained 8 common DMRs from DMRcate method and 9 common DMRs from Bumphunter method during the overlap operation between DMRcate and Bumphunter method. Similarly, we determined 22 common DMRs from Bumphunter method and 19 common DMRs from Comb-p method during the overlap operation between Bumphunter and Comb-p methods. See Table 4 for details. For other µ values, see supplementary file ST1.

For Medium µ value (0.1-0.15)

In case of the overlap operation, for µ=0.15 and repetition=1, we identified 220 common DMRs from DMRcate method and 354 common DMRs from Bumphunter method during the overlap operation between DMRcate and Bumphunter method. Similarly, we obtained 458 common DMRs from Bumphunter method and 301 common DMRs from Comb-p method during the overlap operation between Bumphunter and Comb-p methods, whereas 198 common DMRs from Probe-lasso method and 200 common DMRs from Comb-p method during the overlap operation between Probe-lasso and Comb-p methods. See Table 4 for details. For other µ values, see supplementary file ST1.

For Higher µ value (0.2-0.4)

During the overlap operation, for µ=0.4 and repetition=1, we identified 374 common DMRs from DMRcate method and 605 common DMRs from Bumphunter method during the overlap operation between DMRcate and Bumphunter method. Similarly, we obtained 575 common DMRs from Bumphunter method and 360 common DMRs from Comb-p method during the overlap operation between Bumphunter and Comb-p methods, whereas 221 common DMRs from Probe-lasso method and 222 common DMRs from Comb-p method during the overlap operation between Probe-lasso and Comb-p methods. See Table 4 for details. For other µ values, see supplementary file ST1.

Co-methylation

For Smaller µ value (0-0.05)

Furthermore, we computed average Spearman’s correlation of the participating CpGs belonging to the top 5 DMRs. Bumphunter generated highest correlation in all these cases, but bumphunter suffered from a lot of outliers. See Figure 4(a), and supplementary file SF1(d) for details.

For Medium µ value (0.1-0.15)

For medium µ, Bumphunter and Probe-lasso perform well in terms of the average Spearman’s correlation measures. But, Bumphunter is still the best performer in maximum repetitions. See Figure 4(b), and supplementary files SF1(a) and SF1(e) for details.

For Higher µ value (0.2-0.4)

For higher µ, Bumphunter and Comb-p work well in terms of the average Spearman’s correlation measures. But, Bumphunter is still the best performer in maximum repetitions. See Figure 4(c), and supplementary files SF1(b), SF1(c) and SF1(f) for details.

Sizes of DMRs

For higher µ value (e.g., µ =0.4), the average number of CpGs in the top 5 DMRs obtained by DMRcate is maximum (~10), whereas the same number obtained by Bumphunter is minimum (~5). See Figure 5 for details.

For other µ values, see supplementary files SF2(a)-SF2(f).

**Results of Real Data**

**Discussion and Conclusion**

Identifying differentially methylation regions is a latest topic of interest. As per recent findings, a very few previous studies has been conducted to compare the DMR finding methods for predefined regions. Moreover, there is a lack of systematic evaluation of methods that analyze user defined regions. Therefore, in this article, we conducted a comprehensive evaluation of the most popular software tools for user defined DMR analysis, including DMRcate, ProbeLasso, Bumphunting and Comb-p. First of all, we applied A-clustering to the 14 healthy control samples with similar ages to obtain a total of 3,063 clusters, each consisting of at least 5 adjacent CpGs. The above samples were then randomly divided into two groups, and differential methylation of a small subset of the clusters were simulated by adding a small number to beta values in one of the groups that contains higher mean beta-value. Next, we ran the aforementioned four methods with default parameters to identify DMRs. Of note, for evaluation, we utilized three metrics (viz., TP, FP and FN), and several criteria such as power, precision, time, DMR overlap, DMR co-methylation, and DMRsize. From the outcome, it has been observed that for the power estimation, Bumphunter is best, where for the precision, Comb-p and DMRcate performed best. For co-methylation, Bumphunter is best. In case of DMRsize, DMRcate generated maximum number of average CpGs determined from the specified number of top resultant DMRs, whereas Bumphunter provided minimum number of average CpGs determined from the specified number of top resultant DMRs. Overall, Bumphunter works more or less best among all the methods in terms of all evaluation criteria, but it took a long elapsed time rather than the others. On the other hand, Probe-lasso was probably faster than the others. Specially, all the methods performed well for medium of higher µ value, but for small µ value, no method works well.

In addition, there are several issues raised over these methods. Firstly, the number of choices for choosing parameters are large for these methods, and these parameters are not properly described how to set them. Although in this study, we used default parameters, it is hoped that the future developers clarify different parameter setting for users. Secondly, for small µ value, the performance of these methods is questionable. Thirdly, the regions might not be precisely specified for Bumphunter. Finally, our comparative study provides the critical review of the latest well-known DMR finding methods through which the future users understand the advantages and shortcomings of the underlying methods and can develop new method to resolve the aforementioned issues.

**Legends for Figures and Tables**

Table 3: True Positives (TP), False Positives (FP), False Negatives (FN), Power, Precision and Elapsed Time (in Second) for the four DMR finding methods in the simulation study.

Table 4: Overlap between the four DMR finding methods for low (0.025), medium (0.15) and large (0.4) *µ* for the four DMR finding methods in the simulation study.

Figure 1: Overall workflow of the analysis.

Figure 2: Power comparison of the four DMR finding methods in the simulation study.

Figure 3: Precision comparison of the four DMR finding methods in the simulation study.

Figure 4: Average pairwise correlation for top five DMRs obtained from the four DMR finding methods in the simulation study.

Figure 5: nCpGs for significant DMRs obtained from the four DMR finding methods in the simulation study.

**References**

Butcher, L.M. and Beck, S. Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data. *Methods* 2015;72:21-28.

Hoffman, J. The American Soldier, Vol 1, Adjustment during Army Life, Vol 2, Combat and Its Aftermath - Stouffer. *Contemp Psychol* 1965;10(11):505-505.

Kechris, K.J., Biehs, B. and Kornberg, T.B. Generalizing Moving Averages for Tiling Arrays Using Combined P-Value Statistics. *Stat Appl Genet Mol* 2010;9(1).

Logue, M.W.*, et al.* The correlation of methylation levels measured using Illumina 450K and EPIC BeadChips in blood samples. *Epigenomics* 2017;9(11):1363-1371.

Morris, T.J.*, et al.* ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics* 2014;30(3):428-430.

Pedersen, B.S.*, et al.* Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics* 2012;28(22):2986-2988.

Rindskopf, D. An introduction to the bootstrap - Efron,B, Tibshirani,RJ. *J Educ Behav Stat* 1997;22(2):245-245.