Bioinformatics, 37(5), 2021, 711–713
doi: 10.1093/bioinformatics/btaa732
Advance Access Publication Date: 17 August 2020
Applications Note

OXFORD

Genome analysis

ipDMR: identification of differentially methylated regions with interval *P*-values

Zongli Xu[1], Changchun Xie[2], Jack A. Taylor[1,3,*] and Liang Niu[2,*]

[1]Epidemiology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC 27709, USA, [2]Division of Biostatistics and Bioinformatics, Department of Environmental and Public Health Sciences, College of Medicine, University of Cincinnati, Cincinnati, OH 45267, USA and [3]Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC 27709, USA

*To whom correspondence should be addressed.

Associate Editor: Peter Robinson

Received on April 19, 2020; revised on August 5, 2020; editorial decision on August 10, 2020; accepted on August 11, 2020

Abstract

Summary: ipDMR is an R software tool for identification of differentially methylated regions (DMRs) using auto-correlated *P*-values for individual CpGs from epigenome-wide association analysis using array or bisulfite sequencing data. It summarizes *P*-values for adjacent CpGs, identifies association peaks and then extends peaks to find boundaries of DMRs. ipDMR uses BED format files as input and is easy to use. Simulations guided by real data found that ipDMR outperformed current available methods and provided slightly higher true positive rates and much lower false discovery rates.

Availability and implementation: ipDMR is available at https://bioconductor.org/packages/release/bioc/html/ENmix.html.

Contact: taylor@niehs.nih.gov or niulg@ucmail.uc.edu

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

1 Introduction

DNA methylation is one of the most studied epigenetic markers, which regulates gene expression by affecting the interactions between DNA and transcription-related proteins. Advanced high-throughput array technologies and bisulfite sequencing have made possible genome-wide profiling of DNA methylation levels in studies with large numbers of samples. Although individual CpGs may have weak association, there is growing evidence that spatially clustered CpGs or 'differentially methylated regions' (DMRs) may have stronger associations with disease (Ziller *et al.*, 2013). In recent years, there is increasing interest in methods that combine information from adjacent CpGs to identify DMRs (Butcher and Beck, 2015; Jaffe *et al.*, 2012; Page *et al.*, 2018; Pedersen *et al.*, 2012; Peters *et al.*, 2015). We here introduce an efficient method to identify DMR using interval *P*-values.

2 Materials and methods

ipDMR identifies DMRs based on user-provided association *P*-values for individual CpGs. It first calculates a *P*-value (see below for detail) for each small interval, i.e. the interval bordered by two adjacent CpG within a user-specified value [default: 1000 base pair (bp)]. Second, it performs the Benjamini–Hochberg (BH) procedure on the interval *P*-values to select those significant intervals at a user-specified false discovery rate (FDR) threshold (seed threshold). It then joins all nearby significant intervals and CpGs if the gap (the number of bps between two intervals/CpGs) is less than the user-specified value (default: 1000 bp). Next, it recalculates *P*-values for each combined region using the original *P*-values for all CpGs in that region. Finally, it performs another BH procedure on these region *P*-values to obtain the FDR-adjusted *P*-values.

The *P*-value for an interval/region that contains n CpGs is calculated as

$$p = \Phi\left(\frac{\sum_{i=1}^{n}\Phi^{-1}(p_i)}{\sqrt{n+2\cdot\sum_{1\le i<j\le n}\hat{\rho}_{ij}}}\right),$$

where p_i is the original *P*-value for CpG i in the interval/region, Φ is the cumulative distribution function for the standard normal distribution, $\hat{\rho}_{ij}$ is the estimated correlation between $\Phi^{-1}(p_i)$ and $\Phi^{-1}(p_j)$. Here, we assume under the null hypothesis that $\left(\Phi^{-1}(p_1),\ \Phi^{-1}(p_2),\ \cdots,\ \Phi^{-1}(p_n)\right)'$ follows a multivariate normal distribution with a zero mean vector and covariance matrix $(\rho_{ij})_{n\times n}$

with $\rho_{ii} = 1$ $(1 \leq i \leq n)$. Therefore, under the null hypothesis, $\frac{\sum_{i=1}^{n} \Phi^{-1}(p_i)}{\sqrt{n+2\cdot\sum_{1\leq i<j\leq n}\rho_{ij}}}$ follows a standard normal distribution.

P-value correlations between CpGs ($\hat{\rho}_{ij}$) are estimated using all possible CpG pairs with distance less than a user-specified cutoff (default: 1000 bp). We first divide these CpG pairs into bins according to a user-specified bin size (default: 50 bp). Then for each bin, we calculate and test the Pearson correlation between the Φ^{-1}-transformed *P*-values of the CpG pairs. For two CpGs with distance d, the estimated correlation $\hat{\rho}_{ij}$ is the Pearson correlation for the corresponding bin that include the distance d. If d is greater than the user-specified distance cutoff or the correlation test *P*-value for a specific bin is greater than 0.05, we set $\hat{\rho}_{ij} = 0$. See Supplementary Materials for an example of the calculation.

We implemented the method into the ENmix R package with function name 'ipDMR'. As with comb-p software (Pedersen *et al.*, 2012), ipDMR uses BED format *P*-value files as input. The function can also generate Manhattan plots that mark DMRs and regional *P*-value plots that provide detailed views of significant DMR regions. Previous studies comparing DMR methods have reported that comb-p is one of the most effective methods (Mallik *et al.*, 2019; Peters *et al.*, 2015). However, comb-p was originally written in python and is not convenient for R users. To facilitate its use in R, we implement a similar method that we make available in the ENmix R package with the function name 'combp'. In order to distinguish results of the python implementation of comb-p in this publication, we term the R implementation as 'ENmix-combp'. The original comb-p software used the Stouffer–Liptak–Kechris method to combine *P*-values. To improve computation efficiency, the Enmix-combp function uses the region *P*-value formula as described earlier for ipDMR.

3 Evaluation

Several studies showed that reproducibility can greatly affect study power for specific CpGs (Sugden *et al.*, 2020). To accommodate these, we performed simulations guided by a blood DNA Methylation 450K dataset for 128 duplicate samples (Xu *et al.*, 2020). We first calculated DNA methylation mean beta values, within-subject variation and between-subject variation for each CpG. Within-subject variation reflects technical variation, and between-subject variation reflects biological variation. Pearson's correlation coefficients were calculated between each pair of adjacent CpGs. Based on Illumina annotation of CpG islands (island, shore, shelf and other) and genomic locations (Exon, 3UTR, 5UTR, Body TSS1500 and other), we parsed all CpGs into 168 671 groups, with an average of 2.7 (range of 1–240) CpGs per group. In each simulation, we randomly selected 20 groups as true DMR regions. For each CpG within these regions, we assigned a true effect size equal to 2/3 of its biological (between subject) standard deviation. We simulated DNA methylation data [458 178 CpGs, low quality CpGs were removed as previously described in Xu *et al.* (2020)] for 100 cases and 100 controls. The simulated data have the same data properties as the real dataset for the following characteristics: mean DNA methylation distribution, CpG to CpG auto-correlation, within- and between-subject variation. Because different CpGs have different variance profiles, particularly different relative magnitudes for within- and between-subject variation, the observed effects at true DMRs are also widely different from CpG to CpG. We tested methylation differences for each CpG between cases and controls using the limma method, and the resulting *P*-values for individual CpGs were then used to detect DMRs with different methods (ipDMR, comb-p and ENmix-combp) using different parameter configurations. We performed 100 round of simulations and summarized the results in Table 1. In all tests, we set 1000 bp as the maximum distance to combine adjacent DMRs and use true positive rate (TPR) and FDR (see Supplementary Results for the definition and examples) to evaluate the effects of various combinations of the seed threshold and bin size (termed as 'steps' in comb-p software).

Table 1. Evaluation results for ipDMR, comb-p and ENmix-combp with different seed threshold and bin sizes

Method	Seed	Bin size	TP (SD)	FD (SD)
ipDMR	0.01	310	0.44 (0.11)	0.21 (0.13)
	0.01	50	0.44 (0.12)	0.19 (0.12)
	0.1	310	0.61 (0.11)	0.59 (0.10)
	0.1	50	0.61 (0.11)	0.57 (0.10)
Comb-p	0.01	310	0.40 (0.11)	0.43 (0.14)
	0.01	50	0.36 (0.11)	0.31 (0.14)
	0.1	310	0.50 (0.11)	0.69 (0.07)
	0.1	50	0.45 (0.10)	0.57 (0.11)
ENmix-combp	0.01	310	0.45 (0.12)	0.47 (0.12)
	0.01	50	0.30 (0.11)	0.17 (0.14)
	0.1	310	0.60 (0.11)	0.74 (0.06)
	0.1	50	0.50 (0.10)	0.44 (0.12)

As shown in Table 1, both comb-p and ENmix-combp are very sensitive to bin size: when bin size is larger, both TP and FD are larger. The overall performance of comb-p is similar to ENmix-combp for the same parameter configuration. The ipDMR is robust to bin size, and outperformed comb-p and ENmix-combp in all four different parameter configurations with higher TPR and lower FDR. Applications in a real dataset are demonstrated in Supplementary Materials.

4 Discussion

The ipDMR is an efficient method to identify DMRs. Compared to comb-p using realistic simulation data, ipDMR has a slightly higher rate for finding true positive DMRs and much lower rate for false DMRs. Although our evaluations here are limited to 450K data, our method should apply equally well to 850K array and bisulfite sequencing data. Compared to comb-p and several other DMR tools (Jaffe *et al.*, 2012; Pedersen *et al.*, 2012; Peters *et al.*, 2015), ipDMR is easier to use and requires the specification of fewer tuning parameters. Our simulations suggest that ipDMR is robust to most parameter specifications except for seed threshold, which provides user-control sensitivity level. Comb-p requires calculation of smoothed *P*-value with a user-specified bin size, while ipDMR uses interval *P*-values to identify DMRs, and thus the computation is more efficient. For example, while it only takes 10 s for ipDMR to infer DMRs in a set of *P*-values from 450 K array data with one CUP core, it takes 69 and 44 s, respectively, for comb-p and ENmix-combp with parallel computing using 22 CUP cores. We should note that all these methods are exploratory tools. Although they can identify robust statistical DMRs, their biological interpretation should be weighted by whether they co-locate with tissue-relevant functional epigenomic loci.

Funding

This work was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences [Z01 ES049033, Z01 ES049032, Z01 ES044005] and National Institute of Environmental Health Sciences Award [P30ES00606].

Conflict of Interest: none declared.

References

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

Jaffe,A.E. *et al.* (2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.*, **41**, 200–209.

Mallik,S. *et al.* (2019) An evaluation of supervised methods for identifying differentially methylated regions in Illumina methylation arrays. *Brief Bioinform.*, **20**, 2224–2235.

Page,C.M. *et al.* (2018) Assessing genome-wide significance for the detection of differentially methylated regions. *Stat. Appl. Genet. Mol. Biol.*, **17**, 20170050.

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

Peters,T.J. *et al.* (2015) De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin*, **8**, 6.

Sugden,K. *et al.* (2020) Patterns of reliability: assessing the reproducibility and integrity of DNA methylation measurement. *Patterns*, **1**, 100014.

Xu,Z. *et al.* (2020) Blood DNA methylation and breast cancer: a prospective case-cohort analysis in the sister study. *J. Natl. Cancer Inst.*, **112**, 87–94.

Ziller,M.J. *et al.* (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature*, **500**, 477–481.