Assessing Genome-Wide Significance for the Detection of Differentially Methylated Regions

Christian M. Page[1,2,3,*,¶], Linda Vos[4,*], Trine B. Rounge[4,5],Hanne F. Harbo[1,2], Bettina K. Andreassen[4,§]

[1]Department of Neurology, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Norway
[2]Department of Neurology, Division of Clinical Neuroscience, Oslo university hospital, Norway
[3]Department of non-communicable diseases, Norwegian Institute of Public Health, Norway
[4]Department of Research, Cancer Registry of Norway, Oslo, Norway
[5]Genetic epidemiology group, Folkhälsan research center, Helsinki, Finland
[§]Corresponding author; email: B.K.Andreassen@kreftregisteret.no
*Equal contributions
[¶]Current adress: Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Oslo, Norway

Abstract

Motivation: DNA methylation plays an important role in human health and disease, and methods for the identification of differently methylated regions are of increasing interest. There is currently a lack of statistical methods which properly address multiple testing, i.e. control genome-wide significance for differentially methylated regions.

Methods: We introduce a scan statistic (DMRScan), which overcomes these limitations. We benchmark DMRScan against two well established methods (bumphunter, DMRcate), using a simulation study based on real methylation data. An implementation of DMRScan is available from Bioconductor.

Results: Our method has higher power than alternative methods across different simulation scenarios, particularly for small effect sizes. DMRScan exhibits greater flexibility in statistical modeling and can be used with more complex designs than current methods.

Conclusion: DMRScan is the first dynamic approach which properly addresses the multiple-testing challenges for the identification of differently methylated regions. DMRScan outperformed alternative methods in terms of power, while keeping the false discovery rate controlled.

Keywords

Differentially methylated regions; Scan statistics; Sliding window; Genomics

Introduction

DNA methylation is an epigenetic marker, which can explain variation in gene expression, as well as cell differentiation and other variability in cell phenotypes[1-3]. It is the most studied

epigenetic modifier on a genome-wide scale[4]. DNA methylation is believed to play an important role in the pathology of complex diseases. In cancer, large changes in the global methylation level have been observed[5]. However, for most other complex diseases, there has been little evidence of such a global change in DNA methylation. This has led to the notion that local methylation differences in smaller regions (called differentially methylated regions; DMRs) may be relevant for these diseases[6]. Although methylation at specific CpG sites may have an effect on its own[7], it is often assumed that multiple methylation sites within a cluster of CpGs are involved in a change of cell characteristics[3]. Several CpGs within a region might contribute to a disease or phenotype, but their individual effects may not be strong enough to pass a genomic-wide significance threshold. In recent genome-wide methylation studies, there has been an increasing focus on identifying significant DMRs by combining methylation information from neighboring CpG sites[8]. The underlying thought is to increase power by reducing the requirements for multiple testing adjustments through accumulation of correlated signals.

There are two types of procedures for determining DMRs. The first procedure is based on underlying biological knowledge with respect to the unit of interest. For instance, the CpG sites can be grouped by their affiliation to genes, regulatory regions, CpG islands or pathways. These *fixed* units can be analyzed separately with respect to the phenotypes of interest, and the units are classified as DMRs if there is enough evidence for association. Multiple testing procedures can be easily applied by taking into account the number of predetermined regions. The second type of aggregation is *dynamic*, where the borders are not pre-determined, but rather data driven, as CpG sites in close proximity are collapsed into regions in order to identify potential DMRs. Adjustment for multiple testing when using this approach is challenging and still developing.

Several methods have been proposed to identify DMRs, such as BSmooth, bumphunter, Comb-p, DMRcate, dmrseq, DMRMark, and ProbeLasso [2, 9-13]. Additionally, there are methods (csaw and PeakSeq [14, 15]) for peak detection involving ChIP-seq data, thus relying on count data. The underlying theory, however, could also be applied to DNA methylation data. Many of these methods are tailored to a specific technology (e.g. dmrseq, DMRMark, BSmooth and ProbeLasso), while some are compatible with almost any measurement technology (bumphunter, Comb-p, and DMRcate). Applying peak detection methods for ChIP-seq on methylation data requires non-trivial adaptations and is outside the scope of this paper. We selected methods based on dynamic aggregation, identifying DMRs which are independent of technology and appropriate to use for both sequencing and chip data. This excludes static methods such as ProbeLasso and methods only applicable to one specific technology, such as dmrseq, BSmooth, and DMRMark. Two widely used methods meeting these criteria were selected for comparison purposes to our method; bumphunter and DMRcate[2, 10]. The bumphunter algorithm is among the most commonly used approach when interrogating DMRs and can be considered as the "gold standard" for DMR calling. The peak calling packages are mostly directed towards ChIP-seq data, and the input data are often structured differently than for methylation data; as such it is difficult to apply directly to methylation data without modifying the source code.

Bumphunter was among the first methods that proposed a multiple-testing adjusted procedure when scanning the epigenome for significant regions[10]. Bumphunter's multiple testing adjustment for the region p-values considers regions where the effect sizes exceed a threshold. There are two ways to adjust the p-values for the selection step, either by permuting the case-control status or with Monte Carlo simulation from a truncated multivariate normal distribution of the same size as the detected region [16]. DMRcate reports a minimum p-value

within a region as well as an aggregated p-value based on Stouffers method[17]. Both these p-values can be hard to interpret, and do not necessarily keep the overall α-level.

There is a wide range of literature on scan statistics, which is based on extreme value theory and uses a well-defined theoretical framework, which allows us to overcome the limitations of current methods and to identify genome-wide significant DMRs. Our introduced method, DMRScan, properly adjusts for multiple testing by keeping the false positives controlled at the α-level significance threshold. Several variants of scan statistics have been successfully applied on different types of genomic data[5, 7]. We propose an adoption of a sliding window approach previously used in peak detection for ChIP-chip tiling arrays[18]. Despite of some similarity to the csaw R-package [14], there are notable differences. The csaw method addresses the issue of FDR control by combining locus-wise p-values to a region-wise p-value using Simes' method. The region-wise p-values are adjusted using a Benjamini-Hochberg FDR correction, while our method relies on Poisson heuristics to assess genome wide significance.

Material and Methods

Bumphunter and DMRcate

Bumphunter[10] identifies all CpG sites over a certain percentile of the test statistic distribution (cut-off parameter). These sites are aggregated together into clusters based on their genomic position. Region-wise p-values are estimated using either permutation or bootstrap approaches. By permuting the outcome variable, a set of null regions are constructed. The candidate regions are compared with the distribution of the null regions in both length and area under the curve. The proportion of null regions with an area under the

curve and a region length being at least as extreme as the candidate region is presented as the family-wise error rate for the given region.

DMRcate[2] applies a Gaussian kernel smoothing on the site-wise test statistic, after using a *limma* model[19] on each CpG. Using the method of Satterthwaite[20], probe-wise p-values are calculated for the smoothed test statistic. After adjustment for multiple testing (by FDR), nearby genome-wide significant probes are aggregated into regions. Using Stouffer's method[17] on the adjusted probe-wise p-values, a region-wise p-value is calculated using all probes within the candidate regions.

DMRScan

DMRScan is a sliding window approach based on extreme value theory, which has earlier been applied to peak detection for transcription factor binding sites[18]. It is based on the observation from Aldous[21], that for a large enough threshold, the number of significant windows in a scan statistic surpassing the threshold will follow a Poisson distribution.

Using extreme value theory, Zhang deduced a relationship between the significance level (α) and the intensity of the Poisson distribution (λ) for the number of peaks above a threshold. Assuming independent tests, we get that: $\alpha = 1 - e^{-\lambda}$. By putting a constraint that no two overlapping windows can both be significant, Zhang constructs independent observations. A natural extension of this is to use different window sizes. To create independent observations, nested or overlapping windows cannot both be significant. In such a case, the smallest window would be regarded as the significant window[18].

The intensity (λ) is dependent on the window threshold (t), the correlation structure of the test statistics, and the window size (k). Using a Monte Carlo simulation with different thresholds,

Zhang was able to derive a relationship between the threshold and the significance level of the test for each window size[18].

For every CpG site, a linear regression analysis was done with methylation level as the dependent variable and case-control status as the explanatory variable. However, there are no restrictions with respect to the statistical model used on each CpG site in order to determine the probe wise statistic. Different link functions can be chosen and additional explanatory or confounding variables can be added with little computational cost. Hence, one is able to select a statistical model which fits the data best. The CpG wise test statistic will be denoted as T_{CpG}. For each window-size k, we used Monte Carlo simulation to determine the minimal threshold t_k based on the significance level α. We chose the window threshold (t_k) of the window statistic (T_{DMR}) such that the expected number of significant tests (E_k) for each window size k was equal (see Appendix 1, eq. 2).

Three variants of DMRScan using different methods to determine the window thresholds t_k were implemented: DMRScan (*MCMC*), DMRScan (*Importance sampling*) and DMRScan (*Siegmund*). In the first two approaches, a Monte Carlo simulation is used to determine the threshold given the dependency structures for the T_{CpG}'s. For DMRScan (*Siegmund*), the thresholds are calculated using an analytic expression.

In DMRScan (*MCMC*), a Monte Carlo simulation was used to determine the number of significant tests over the threshold. In this algorithm, one is free to choose the optimal model for the dependency structure of the test statistic T_{CpG} based on the underlying data.

DMRScan (*Importance sampling*) uses a local average of independent Gaussian variables to describe the dependency structure of the statistic T_{CpG}, assuming a dependency of two probes in both directions. Properties of the standard normal distribution in a fast importance sampling

algorithm were used to simulate the intensity of the number of windows exceeding the threshold. Importance sampling was over 700 times faster than the MCMC algorithm.

A modification of Zhang's method was introduced by Siegmund *et al.*[22] and implemented in DMRScan (as the option "*Siegmund"* in the DMRScan function call). Here, the intensity for the Poisson distribution (λ) is analytically calculated as a function of the desired threshold. This derivation is based on the assumption that the test statistic follows an Ornstein-Uhlenbeck process (OU-process). A closed form solution was first published by Siegmund[23] and later re-formulated in[24] [pp. 112],

$$\lambda = 2\beta L t_k \phi(t_k) v(t_k(2\beta\Delta)^{1/2})$$

Here λ is the intensity of windows over the threshold (t_k), L is the genetic length of the chromosome (in number of CpGs), $\beta = 1/k$ is the autoregressive parameter of the OU-process where k is the window size, Δ is the spacing between observations (assumed to be 1). The function $v(.)$ can be approximated by

$$v(y) \approx \frac{(2/y)(\Phi(y/2) - 0.5)}{(y/2)\Phi(y/2) + \phi(y/2)}$$

The functions $\Phi(.)$ and $\phi(.)$ are the cumulative distribution and the density function of the standard normal distribution, respectively.

Multiple-testing adjusted p-values for the genome-wide significant DMRs can be derived by a combination of empirics and theoretical properties. The variance of the test statistic of the window of interest with window size k is approximated using simulation and theoretical asymptotic p-values are derived using the standard normal distribution (see Appendix, eq 3). Alternatively, empirical p-values can be calculated by comparing the value of the test statistic

T_{DMR} for the window of interest of window size *k* with the distribution of all test statistics T_{DMR} for windows with the same window size *k*.

DMRScan, together with an example dataset is implemented as an R library in Bioconductor. An illustrating example of its usage is given in the supplementary material to this paper.

Results

Simulation study

Procedure

We used real methylation data from chromosome 22 from the Finnish Health in Teens study (Fin-HIT, http://www.finhit.fi/for-researchers/), described in more detail here [25]. The backbone for the CpG regions was known CpG regions at chromosome 22. To evaluate and compare the methods, we tested them on 100 causal regions. This number is a trade-off between few regions (biological plausibility) and having an extensive testing of the methods (many regions). We let the frequency of the causal region be inversely proportional to its length, thus shorter regions were more frequent than longer regions in the simulation. We added an effect by changing the methylation beta-values[26] of the causal CpGs such that the mean difference between cases and controls in that region were equal to the effect size. The beta-values are ranging from 0 to 0.15 and always within the legal limit of 0 to 1. The first simulation was on the original data set with no added effect. The causal regions ranged in size from 5 to 100 sequential CpG sites, reflecting the range which seems biologically relevant and plausible [27]. A CpG island could not have more than one causal region and the maximum distance between the causal CpGs could not exceed the maxGap parameter in all methods.

In each causal region, we added an artificial effect and compared the performance in retrieval of these 100 regions for the five methods (i) bumphunter, (ii) DMRcate, (iii) DMRScan *(MCMC)* with thresholds based on extreme value theory using Monte Carlo simulation, (iv) DMRScan *(Importance Sampling),* where an importance sampling algorithm was used to determine the thresholds, (v) and DMRScan *(Siegmund),* with an analytic expression was used to determine the window thresholds.

For each effect size, we counted the number of true positive and false positive DMRs (Figure 1 A-B). Any DMRs overlapping with a causal region was counted as true positive observation. We also summed the number of significant probes in each DMR, occurring both inside and outside of the causal regions (Figure 1 C-D). Hence, the number of true and false discoveries from both a DMR and CpG perspective were gathered.

DMRScan

When inspecting the test statistics T_{CpG} on a subset of the data, an AR(2) process gave the best description of the dependence structure in our subset. Hence we used an AR(2) process as a null model to determine the thresholds in DMRScan *(MCMC)*.

For window thresholds between 0.8 to 4 with regular increments of 0.2, and different window sizes (k) from 2 to 10, we simulated test statistics from a null model and applied DMRScan with fixed window size and no overlapping significant windows. We determined the number of significant windows for the different window sizes and thresholds. This was done using both the MCMC and the Importance sampling approach. For the different window sizes (k), we chose the window threshold (t_k) such that the expected number of significant tests E[*significant.window*] was equal for all window sizes (see Appendix, eq. 2). Since we placed

the different thresholds on a grid, interpolation was used to determine the minimal threshold keeping the significance level α at a genome wide significance level.

Using the analytic formula of Siegmund, we calculated thresholds t_k for each window size k such that the expected number of significant tests E[*significant.window*] is equal for all window sizes (see Appendix, eq. 2). A detailed explanation for the parameter choices is given in the supplementary materials and methods, and a full list of our parameter choices is listed in Table 1.

Power assessment

We define the power as the proportion of true, genome wide significant causal DMRs. The number of true positive and false positive regions is shown in Figure 1 (A and B), as a function of increasing effect size. All three versions of the DMRScan algorithm had a faster convergence in power compared to bumphunter when calling DMRs. DMRcate outperformed Bumphunter in DMR calling, however, this came at a cost of a higher number of false positive probes (Figure 1 C-D). The false positive probes in DMRcate were in close proximity of the causal regions, but the proportion of false positive probes was considerable as compared to the other methods.

Since the thresholds for the sliding windows are static, the false discovery rate for DMRScan was independent of the added effect size and remained fixed throughout the simulations (Figure 1 B). The number of false positive of DMRscan(siegmund) was approximately equal to that of Bumphunter. For DMRcate, the number of false positive sites increased with increasing effect size, this can be seen in Figure 1(D). On closer investigation, all of the reported false discoveries lay on the edges of a causal region, and no false positive regions independent of any causal DMRs were detected. The false positive discoveries were due to

DMRcate's smoothing effect on the border of the regions, where the smoothing extended the reported regions beyond the causal part. DMRScan with a theoretically derived threshold using Siegmund's model had the lowest false positive rate, which was close to zero. The importance sampling threshold had a marginally higher false positive rate, but a substantially faster convergence in true positives.

We observe the biggest difference between the methods for small effect sizes. Bumphunter had a negligible proportion of true positives for effect sizes under 0.05, while the sliding windows and DMRcate were much more responsive for small effect sizes. DMRcate tended to have a higher false positive rate than the sliding windows approaches, even for very low effect sizes. For the DMRScan with importance sampling and Monte Carlo thresholds, the number of false positive observations was small. Three and 5 of 971 regions (0.5%) were falsely detected, respectively.

Discussion

We have proposed a new method for identifying DMRs, based on Poisson heuristics and a sliding window approach. We compared this to other established methods for identifying DMRs. The approach introduced in this paper is based on an approach presented by Zhang which was originally introduced for ChIPseq analysis. With some modifications, it is now applicable to DNA methylation analysis. However, the method itself may not be restricted to those two areas. Scan statistics can be used for peak detection on any data containing correlated observations.

For most complex diseases, CpG-wise test statistics are not likely to contain distinct peaks like those observed in ChIP-seq. Thus, the thresholds for the region wise test statistics have to be very close to the observed test statistic, T_{DMR}, in order to pick up any signals. When the

threshold lies close to the observed test statistic, the number of false positive windows will be very sensitive to small changes in the threshold.

Having 100 causal regions in one analysis is quite optimistic, but was chosen to provide a good spread on the different length of causal DMRs while maintaining computational efficiency. Longer DMRs were assumed rare and few causal regions spanned more than 40 CpGs.

Since the sliding windows are applied on the test statistic and not on the raw data, they are not as prone to many of the challenges the other methods face, such as probe bias for the methylation microarrays, or varying depth in sequencing studies, which all can be accounted for in the first step of the modeling. Both DMRcate and bumphunter use very specific models to evaluate point-wise methylation, leaving few options for the user to apply more complex designs, like repeated measures, non-linear effects, or logistic regression. This is in contrast to DMRScan, which relies only on the summary statistic, and can be applied on the test statistics from any model as long as the underlying distribution of the test statistic is approximately normal. Additionally, since the marginal summary statistic only has to be calculated once for DMRScan, covariates and confounders can be included without any notable increase in computational time.

When doing whole genome bisulfite sequencing or reduced representation bisulfite sequencing, the methylation data set can be substantially larger than that of chip data. Since DMRcate and DMRScan do not use permutation, they are not affected by this issue as much as bumphunter, where the computational time can be substantial.

The three compared methods use different approaches for constructing p-values for the candidate DMRs. One possible solution, by DMRcate, is to report the minimum p-value, or to

aggregate the p-values using Stouffer's method. Stouffer's method is a way of combining p-values by adding the Z-score normalized by the length of the candidate DMRs. For highly dependent p-values, this may induce inflation in the test statistic, if the sum is not weighted accordingly[28]. Bumphunter uses the minimum p-values in each DMR as its region-wise p-values, which often deflates the p-values. Moreover, an "adjusted p-value" based on a permutation test is given for each region, which is much more conservative. For the bumphunter implementation, Jaffe *et al.* acknowledge that the region-wise adjusted p-values may not always be representative, and that care should be taken when interpreting the findings[10]. By applying a sliding window to call DMRs, we can utilize a well-defined framework to construct p-values for each DMR which are adjusted for multiple testing. Unlike bumphunter and DMRcate, the regions detected by the DMRScan method are always genome-wide significant for the false discovery level set by the user.

DMRcate

An important gain of the applicability of summary statistics in our approach is the possibility to analyze data from already published DNA methylome studies separately or in a meta-analysis setting. In most methylomic or genomic meta-analysis, the individual raw data from each separate study are not accessible, but a summary test statistic for each locus can often be obtained across the different studies. This can open a new opportunity for meta-analysis efforts in identification of DMRs.

Conclusion

DMRScan is a data-driven approach which properly addresses the multiple-testing challenge when claiming genome-wide significance for differentially methylated regions. DMRScan

308 performs better in terms of power compared to previously introduced methods, while keeping
309 the false discovery rate controlled.

310 List of abbreviations

AR(p)	Autoregressive process of order p
ChIP	Chromatin Immunoprecipitation
DMR	Differentially methylated region
E_k	Expected number of significant windows of size k
FDR	False discovery rate
MCMC	Markov Chain Monte Carlo
OU-process	Ornstein-Uhlenbeck process
t_k	Window threshold for sliding windows of size k

311

312 Declarations

313 ***Ethics***

314 The Coordinating Ethics Committees of the Hospital Districts of Helsinki and Uusimaa
315 approved the study.

316 ***Consent for publication:***

317 Informed consent was obtained from all participants and as well as one of their legal
318 guardians.

319 ***Availability of data and materials***

320 The R package is placed at Bioconductor under the name *DMRScan*, along with the example
321 data set used in this paper. The R-code for comparing the methods is available by the author
322 upon request.

323 ***Competing interests***

324 The authors declare that they have no competing interests

Funding

This work was supported by the University of Oslo [grant number 531217/1231]; Folkhälsan Research Foundation; The Academy of Finland [grant number 250704]; The Life and Health Medical Fund [grant number 1-23-28]; The Swedish Cultural Foundation in Finland [grant number 15/0897]; The Signe and Ane Gyllenberg Foundation [grant number 37-1977-43]; and The Yrjö Jahnsson Foundation [grant number 11486].

Author Contributions

CMP; conceived the experiment, made the R scripts, did the analysis, wrote the paper

LV; conceived the experiment, made the R scripts, did the analysis, wrote the paper

TBR; supplied methylation values for the experiment, critically reviewed the manuscript

HFH; contributed to idea and funding, critically reviewed the manuscript

BKA; conceived the experiment, did the analysis, wrote the paper

All authors read and approved the final version of the manuscript.

Acknowledgement

We acknowledge Folkhälsan Research Center and the Fin-HIT study group: Sabina Simola, Stephanie Von Kreamer, Jesper Skand, Catharina Sarkkola, Sajan Raju and Elisabete Weiderpass (Helsinki, Finland) for providing data for benchmarking the different models. Institute for Molecular Medicine Finland (FIMM) provided computational infrastructure and preformed the sequencing to this project. Suzanne Campbell and Marissa LaBlanc for critical evaluation of this manuscript.

References

1. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J *et al*: **Dynamic changes in the human methylome during differentiation**. *Genome research* 2010, **20**(3):320-331.

2. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, R VL, Clark SJ, Molloy PL: **De novo identification of differentially methylated regions in the human genome**. *Epigenetics & chromatin* 2015, **8**:6.
3. Jones PA: **Functions of DNA methylation: islands, start sites, gene bodies and beyond**. *Nature reviews Genetics* 2012, **13**(7):484-492.
4. Rakyan VK, Down TA, Balding DJ, Beck S: **Epigenome-wide association studies for common human diseases**. *Nature reviews Genetics* 2011, **12**(8):529-541.
5. Sun YV, Levin AM, Boerwinkle E, Robertson H, Kardia SL: **A scan statistic for identifying chromosomal patterns of SNP association**. *Genetic Epidemiology* 2006, **30**(7):627-635.
6. Feinberg AP, Irizarry RA, Fradin D, Aryee MJ, Murakami P, Aspelund T, Eiriksdottir G, Harris TB, Launer L, Gudnason V *et al*: **Personalized epigenomic signatures that are stable over time and covary with body mass index**. *Sci Transl Med* 2010, **2**(49):49ra67.
7. Reiner-Benaim A, Davis RW, Juneau K: **Scan statistics analysis for detection of introns in time-course tiling array data**. *Statistical applications in genetics and molecular biology* 2014, **13**(2):173-190.
8. Slieker RC, Bos SD, Goeman JJ, Bovee JV, Talens RP, van der Breggen R, Suchiman HE, Lameijer EW, Putter H, van den Akker EB *et al*: **Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array**. *Epigenetics & chromatin* 2013, **6**(1):26.
9. Hansen KD, Langmead B, Irizarry RA: **BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions**. *Genome biology* 2012, **13**(10):R83.
10. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA: **Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies**. *Int J Epidemiol* 2012, **41**(1):200-209.
11. Butcher LM, Beck S: **Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data**. *Methods* 2015, **72**:21-28.
12. Shen L, Zhu J, Robert Li S-Y, Fan X: **Detect differentially methylated regions using non-homogeneous hidden Markov model for methylation array data**. *Bioinformatics* 2017, **33**(23):3701-3708.
13. Korthauer K, Chakraborty S, Benjamini Y, Irizarry RA: **Detection and accurate False Discovery Rate control of differentially methylated regions from Whole Genome Bisulfite Sequencing**. *bioRxiv* 2017:183210.
14. Lun AT, Smyth GK: **csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows**. *Nucleic acids research* 2015, **44**(5):e45-e45.
15. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB: **PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls**. *Nature biotechnology* 2009, **27**(1):66-75.
16. Benjamini Y, Taylor J, Irizarry RA: **Selection Corrected Statistical Inference for Region Detection with High-throughput Assays**. *bioRxiv* 2016:082321.
17. Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RM, Lumsdaine AA, Lumsdaine MH, Smith MB, Janis IL, Cottrell LS: **Studies in Social Psychology in World War II.: The American Soldier: Adjustment During Army Life**; 1965.
18. Zhang Y: **Poisson approximation for significance in genome-wide ChIP-chip tiling arrays**. *Bioinformatics* 2008, **24**(24):2825-2831.

19. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies**. *Nucleic acids research* 2015:gkv007.
20. Satterthwaite FE: **An approximate distribution of estimates of variance components**. *Biometrics bulletin* 1946, **2**(6):110-114.
21. Aldous D: **Probability approximations via the Poisson clumping heuristic**, vol. 77: Springer Science & Business Media; 1989.
22. Siegmund DO, Zhang NR, Yakir B: **False discovery rate for scanning statistics**. *Biometrika* 2011, **98**(4):979-985.
23. Siegmund D: **Sequential analysis: tests and confidence intervals**: Springer Science & Business Media; 1985.
24. Siegmund D, Yakir B: **The statistics of gene mapping**: Springer Science & Business Media; 2007.
25. Rounge TB, Page CM, Lepisto M, Ellonen P, Andreassen BK, Weiderpass E: **Genome-wide DNA methylation in saliva and body size of adolescent girls**. *Epigenomics* 2016, **8**(11):1495-1505.
26. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM: **Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis**. *BMC bioinformatics* 2010, **11**:587.
27. Bock C: **Analysing and interpreting DNA methylation data**. *Nature Reviews Genetics* 2012, **13**(10):705-719.
28. Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, Zhao J, Zhou X, Reveille JD, Jin L: **Gene and pathway-based second-wave analysis of genome-wide association studies**. *European Journal of Human Genetics* 2010, **18**(1):111-117.

Table 1 Comparison of the parameters between the five models used in the benchmarking.

Figure 1 Comparison of the convergence in power for all five methods, as well as the false positive rate, both as a function of increasing effect size. Top panel (A-B) represents the power to detect causal DMRs for the two different scenarios. The lower panel (C-D) represents the power to detect CpGs within a causal DMR. The dashed lines represent false positives. Bumphunter and DMRScan (Siegmund) had a very similar false positive rate cannot be distinguished as they are directly on top of each other.