OXFORD

Briefings in Bioinformatics, 20(1), 2019, 47–57

doi: 10.1093/bib/bbx097
Advance Access Publication Date: 18 August 2017
Paper

Accounting for differential variability in detecting differentially methylated regions

Ya Wang, Andrew E. Teschendorff, Martin Widschwendter and Shuang Wang

Corresponding author. Shuang Wang, Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, NY 10032, USA. Tel.: 212-342-4165; Fax: 212-305-9408; E-mail: sw2206@columbia.edu

Abstract

DNA methylation plays an essential role in cancer. Differential variability (DV) in cancer was recently observed that contributes to cancer heterogeneity and has been shown to be crucial in detecting epigenetic field defects, DNA methylation alterations happening early in carcinogenesis. As neighboring CpG sites are highly correlated, here, we present a new method to detect differentially methylated regions (DMRs) that uses combined signals from differential methylation and DV between sample groups. We demonstrated in simulation studies the superior performance of the new method than existing methods that use only one type of signals when true DMRs have both. Applications to DNA methylation data of breast invasive carcinoma (BRCA) and kidney renal clear cell carcinoma (KIRC) from The Cancer Genome Atlas (TCGA) and BRCA from Gene Expression Omnibus (GEO) suggest that the new method identified additional cancer-related DMRs that were missed by methods using one type of signals. Replication analyses using two independent BRCA data sets suggest that DMRs detected based on DV are reproducible. Only the new method identified epigenetic field defects when comparing normal tissues adjacent to tumors and normal tissues from age-matched cancer-free women from the GEO BRCA data and confirmed their enrichment in the progression to breast cancer.

Key words: DNA methylation; differential variability; algorithm; differentially methylated regions

Introduction

DNA methylation plays an important role in gene expression [1–4] and cancer [5–8]. Two types of aberrant DNA methylation in cancer have been frequently observed, local hypermethylation in some promoter-related CpG islands that often leads to silencing of downstream tumor suppressor genes [9–13], and global hypo-methylation that usually cause instability of chromosomes [13–16]. Studies have found that abnormal DNA methylation processes are related to many cancer types [17–20] and a range of other human diseases [21–28]. Studies have also found that epigenetic instability of important genomic regions may lead to increased methylation variability in cancer, which also contribute to cancer heterogeneity [29–33]. A study examining DNA methylation profiles of 1505 CpG sites of both normal tissues and tumorigenic tissues observed that there is little variation in the DNA methylation patterns of these normal tissues but greater methylation heterogeneity among tumors [34]. Studies have successfully identified epigenetic field defects in breast cancer based on differential variability (DV) [35]. Epigenetic field defects are notably DNA methylation alterations that usually occur in pre-cancer tissues [36] and are crucial in cancer research because of its potential usage in early cancer detection [35, 37].

Bisulfite microarray and sequencing are two widely used technologies to quantify DNA methylation. Popular array technologies include Illumina Infinium HumanMethylation 27 K, 450 K and 850 K EPIC BeadChips, which produce methylation β-values measuring the proportion of methylated intensities

Ya Wang is a DrPH student at Department of Biostatistics, Mailman School of Public Health, Columbia University, USA.
Andrew E. Teschendorff is a Professor at Department of Women's Cancer, University College London, UK, and CAS Key Lab of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, PR China.
Martin Widschwendter is a Professor at Department of Women's Cancer, University College London, UK.
Shuang Wang is an Associate Professor at Department of Biostatistics, Mailman School of Public Health, Columbia University, USA.
Submitted: 13 April 2017; **Received (in revised form):** 30 June 2017

out of total intensities. Popular sequencing technologies include whole-genome bisulfite sequencing, reduced representation bisulfite sequencing and Agilent SureSelect Human Methyl-Seq (Methyl-seq), which generate either ratio of methylated intensities versus total coverage at each CpG site or number of methylated or unmethylated cytosine.

Methods to identify differentially methylated loci (DML) based on differences in mean methylation levels between two groups are well-studied [38–43]. As DNA methylation levels of neighboring CpG sites are strongly correlated [44, 45] when genomic regions with consecutive CpG sites are associated with cancers [30, 46, 47], methods to detect differentially methylated regions (DMRs) were also developed. However, existing DMR detection methods all focus on mean signals only, and can be generally grouped into three types, to detect site-level signals first and then group adjacent loci into regions using *ad hoc* grouping rules [48–53]; or to define regions first and then test the significance of the defined regions [54–57]; or to use hidden Markov model that assumes three latent methylation states: hypermethylation, hypo-methylation and no differential methylation, and then group adjacent sites with the same state into a region [58, 59]. For array data, for example, bumphunter [49] uses surrogate variable analysis to account for potential batch effects, smoothes site-level signals within a predefined window and defines regions, as adjacent CpG sites with smoothed signals exceed a user-defined threshold. DMRcate [60] uses a tunable Gaussian kernel to smooth site-level differential methylation signals within a given window, then uses the method of Satterthwaite [61] to model the smoothed signals and group neighboring false discovery rate (FDR)-corrected significant CpG sites into regions. Probe Lasso [50] uses a flexible window based on probe density to gather neighboring significant signals to define DMR boundaries. For Bisulfite sequencing data, for example, metilene [51] uses a binary segmentation algorithm to identify candidate DMRs and then use a two-dimensional Kolmogorov–Smirnov (KS) test to assess the significance of candidate DMRs. MethylKit [38] applies logistic regression to predefined regions after normalizing the read coverage across samples. Specific Methylation Analysis and Report Tool (SMART) [62] is an entropy-based framework that first calculates Tukey biweight to quantify methylation specificity at each CpG site, then uses specificity state, Euclidean distance-based methylation similarity, entropy-based methylation similarity and minimum distance requirement to indicate whether methylation patterns of two neighboring CpG sites are similar and then determines DMRs. All of these existing DMR detection methods use mean signals only. In addition to differential methylation, which refers to the difference between mean methylation measures between experimental groups, methods to identify DV, that is, experimental groups differ in terms of methylation variances, were also developed [29, 35, 63–67]. We recently developed NEpiC and pETM methods where NEpiC is a network-based method that combines both mean and variance signals with a much improved power in searching for differentially methylated subnetworks using the protein–protein interaction network [67]; pETM is a penalized Exponential Tilt Model that detects both methylation mean and variance signals at CpG site level with the network-based regularization considering correlations among nearby CpGs [68]. The study that identified epigenetic field defects in breast cancer through comparing DNA methylation levels in normal tissues adjacent to tumors (normal-adjacent) as a surrogate of pre-cancer tissues with those in normal tissues from healthy individuals would have erroneously concluded that there are no significant epigenetic field defects in breast cancer had the authors used a statistical method based on differential methylation only [35]. The authors also observed increased variation in the normal-adjacent tissues driven by a relatively small number of outlier samples exhibiting much-different methylation values from the rest of the normal-adjacent samples [35], when conventional methods focused on mean signals are not able to detect such epigenetic alterations. On region levels, a new method that incorporates DV is needed, especially in detecting epigenetic field defects.

Here, we developed a new DMR detection method that uses combined signal from differential methylation and DV. Simulation studies showed the great performance of the new method. We further demonstrated the performance of the new method through applications to 450 K DNA methylation data of tumor and normal-adjacent tissues of breast invasive carcinoma (BRCA) and kidney renal clear cell carcinoma (KIRC) from The Cancer Genome Atlas (TCGA) project, where some cancer-related genes were missed by the DMR detection methods that use only mean signals or variance signals. By applying the new method to an independent 450 K DNA methylation data of BRCA tumor and normal-adjacent tissues from Gene Expression Omnibus (GEO), we concluded that DMRs detected using variance signals are reproducible. Further applications to the GEO DNA methylation data comparing normal-adjacent tissues from breast cancer patients and normal tissues from age-matched cancer-free women and comparing tumor tissues from breast cancer patients to normal tissues from age-matched cancer-free women not only identified epigenetic field defects in breast cancer but also confirmed that the epigenetic field defects are enriched in the progression to breast cancer [35]. Importantly, the epigenetic field defects were only identified by the developed new DMR detection method that uses mean and variance combined signals.

Methods

As matched case-control study designs with tumor and normal-adjacent tissues are widely used in DNA methylation studies of cancer; here, we focused on studies with a matched case-control design. The proposed new DMR detection method can be easily adapted to other types of designs. There are four steps in the new method (Figure 1): (1) define site-level mean and variance combined signal scores; (2) smooth site-level

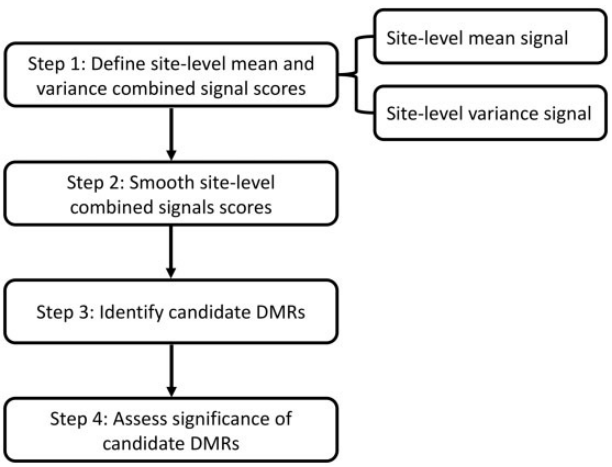

Figure 1. The pipeline of the proposed new DMR detection algorithm.

combined signal scores; (3) identify candidate DMRs; and (4) assess significance of candidate DMRs.

Step 1: Define site-level mean and variance combined signal scores

We define mean and variance combined signal score S_i for CpG site i as follows:

$$S_i = \frac{|T_{mi}|}{T_{mi}} \ (\lambda m_i + (1-\lambda) v_i), \tag{1}$$

where $m_i = \Phi^{-1}(1-p_{mi})$ and $v_i = \Phi^{-1}(1-p_{vi})$. Here, Φ is standard normal quantile function, and p_{mi} and p_{vi} are P-values from the two-sided paired t-test testing if the mean methylation measures are the same between tumor and normal-adjacent tissues and from the one-sided Pitman–Morgan test testing if the variance of the methylation measures in tumor tissues is greater than that in normal-adjacent tissues at CpG site i[69, 70], respectively. We set mean and variance signal scores that are smaller than zero (i.e. sites with $p_{mi} > 0.5$ and $p_{vi} > 0.5$) as zero and remove sites whose mean and variance signal scores are both zero. Here, T_{mi} is the test statistic from a paired t-test, where $\frac{|T_{mi}|}{T_{mi}}$ adds a sign to the combined signal score to indicate whether CpG site i is hyper-methylated (positive sign) or hypo-methylated (negative sign). Similarly, as in our previous work [67], because of potential different scales of the site-level mean and variance signal scores m_i and v_i, we weight the two scores by λ and $1-\lambda$, respectively, to balance the contribution of mean and variance signals to the combined score. We first define the site-level scaling parameter:

$$\lambda_i = \frac{v_i}{m_i + v_i}. \tag{2}$$

At CpG site i, we then average across all sites from the whole genome to obtain the overall scaling parameter λ.

Step 2: Smooth site-level combined signal scores

As methylation levels of CpG sites within 1000 base pairs (bps) are considered highly correlated [44, 55], we assign any two neighboring CpG sites into the same cluster if the genomic distance between them is <1000 bps. We examined how neighboring CpG sites with different distance limits differ in the combined signal scores and summarized results in the Supplementary Data (Investigation of the distance limits to define clusters). We then smooth site-level combined signal scores within a defined cluster using the running median method with a window size of minimum of W sites. The running median method was chosen over the moving average [71] method because of its robustness to outliers. It was chosen over regression-based smoothing methods [48, 49], which have been shown to have similar performance as the moving averaging method [71] because of its computational efficiency. After smoothing, we denote the smoothed combined signal score for CpG site i as $\tilde{S}_i$.

Step 3: Identify candidate DMRs

A candidate DMR is defined to be a region having at least L consecutive CpG sites of the same sign with $|\tilde{S}_i| > k$, where L is a predefined number, and k is a predefined threshold, e.g. the region size to be $L \geq 3$ CpG sites and the threshold to be k =99$^{\text{th}}$ percentile of genome-wide $|\tilde{S}_i|$. Similar criteria of k [48, 49, 71, 72] and L [48, 71] were used in other DMR detection methods.

Step 4: Assess significance of candidate DMRs

We use permutation procedures to assess the significance of candidate DMRs, where we use the following measure to evaluate the strength of evidence for the j^{th} candidate DMR R_j: $A_j = \sum_{i \in R_j} |\tilde{S}_i|$. To assess the significance of the candidate DMR R_j via a permutation procedure under the global null hypothesis adjusting for multiple comparisons, we first shuffle tumor and normal-adjacent status within all pairs and then apply Steps 1–3 to the permuted data set. For the g^{th} permutation that generates n_g regions, we have the evidence of strength for each region as follows: $A_{\text{perm}_g,t}, t = 1, \ldots, n_g$. We repeat the permutation procedure 1000 times, and the empirical p-value of the candidate DMR R_j is calculated as:

$$P_j = \frac{\sum_{g=1}^{1000} \sum_{t=1}^{n_g} I(A_{\text{perm}_g,t} > A_j)}{\sum_{g=1}^{1000} n_g}.$$

To account for multiple comparisons, we calculate the family-wise error rate (FWER) for the candidate DMR R_j as the proportion of permutations with $\max_{t \in [1,n_g]} (A_{\text{perm}_g,t}) > A_j$. The candidate DMR R_j is then considered to be significant if its FWER ≤ 0.05.

The new method outputs a table of candidate DMRs with detailed information of each candidate DMR j: (1) chromosome location, (2) genomic locations of the first and last CpG sites, (3) strength of evidence A_j, (4) number of CpG sites, (5) unadjusted P-value P_j and (6) FWER. Users could also output intermediate results such as mean signal scores and variance signal scores of CpG sites before smoothing as an option.

Comparing methods

We compared the performance of the new method that combines mean and variance signals with those of the DMR detection methods that (1) consider mean signals only including the adapted bump hunting algorithm using two-sided paired t-test, DMRcate, Probe Lasso and the adapted new method with the test statistic from Wilcoxon signed-rank test as the nonparametric version of the mean signals, (2) variance signals only which is the adapted bump hunting algorithm using one-sided Pitman–Morgan test and (3) both mean and variance signals (the adapted new method with test statistic from KS test).

Simulation study

We conducted simulation studies to evaluate type I errors and the performance of the new method. We define type I errors, as the proportions of simulations identified any significant DMRs when data are generated with no DMRs. We use receiver operating characteristic (ROC) curves to evaluate the performance of the new method where we define true positive as significant DMRs with any CpG sites that are in the true DMRs and false positive as significant DMRs with no CpG sites from the true DMRs.

Simulation setup

To simulate methylation measures for tumor and normal-adjacent tissues, we considered 1:1 matched study design with

one tumor sample ($Y = 1$) and one normal-adjacent sample ($Y = 0$) on the matching variable Z. Given Y and Z, we assume logit2 transformed methylation measures [73]; X follows a conditional scaled normal distribution:

$$X|Y = 1, Z = z \sim \sqrt{z}N(\mu, \Delta^T \Sigma \Delta),$$

$$X|Y = 0, Z = z \sim \sqrt{z}N(0, \Sigma),$$

where the matching variable $Z \sim Beta(a, b)$ and Σ is a variance–covariance matrix considering correlations among CpG sites within a predefined cluster. The mean vector $\mu = (\mu_1, \ldots, \mu_h)^T$ and diagonal matrix $\Delta = \text{diag}(\sqrt{\delta_1}, \ldots, \sqrt{\delta_h})$ control the mean and variance signals in a cluster of h consecutive sites. Here, we assume an $AR(1)$ correlation with correlation coefficient ρ, i.e. $\Sigma_{mn} = \sigma \times \rho^{|m-n|}$. We set $\rho = 0.5$ and $Z \sim Beta(1, 1)$ in simulation studies based on our previous experience [43]. In each simulation, we generated X of 10 000 CpG sites from 100 tumor and normal-adjacent pairs, where the genomic locations of these 10 000 sites are the first 10 000 sites of Chromosome 1 on the Illumina 450 K array.

To evaluate type I errors, we set $\mu = 0$, $\sigma = 0.3$, where σ was estimated using methylation measures of the normal-adjacent tissues of the TCGA BRCA data. To evaluate the performance of the new method, we simulated 10 true DMRs with different sizes, varying from 3 to 15 CpG sites, and we considered scenarios when each CpG site in the true DMRs has (1) mean signals only, (2) variance signals only and (3) both mean and variance signals. For all other null CpG sites, we set $\mu = 0$ and $\sigma = 0.3$. For each simulation scenario, we conducted 1000 simulations. In all simulation studies and real data applications, we defined the region size to be $L \geq 3$ CpG sites.

Adaption to case-control designs

We adapted the proposed new DMR detection method for case-control designs, which can adjust for relevant covariates. More specifically, we fit a linear regression model on logit2 transformed methylation β-values, M-values, adjusting for known confounders such as age and gender, and cell composition if necessary, and work on residuals in all subsequent steps. We conducted simulation studies parallel as for matched case-control designs to evaluate the type I errors and the performance. The simulation setup and results are summarized in the Supplementary Data (Simulation studies for case-control designs).

Results

Simulation results

Type I errors are all well controlled at the 0.05 significance level with values 0.055, 0.046, 0.050, 0.041 and 0.041 for the new method, DMR methods based on paired t-test, Pitman–Morgan test, Wilcoxon signed-rank test and KS test, respectively, while that for DMRcate and Probe Lasso are much more conservative with values 0.015 and 0, respectively.

For the ROC curve results (Figure 2), when the significance threshold was set from 0 to 0.05, we notice that when the true DMRs are set to have sites with mean signals only, the new method performs slightly inferior to paired t-test and similarly to KS test, and much better than the Wilcoxon signed-rank test, while Pitman–Morgan test that considers variance signals only could not detect any true DMRs. On the other hand, DMRcate appears to perform better than the new method with higher true-positive rates and zero false-positive rates. This is because DMRcate uses Stouffer transformation [74] of the limma-derived FDRs for individual CpG sites constituting a DMR to assess the overall significance of the DMR, which in general is much smaller than the P-values by the new method assessing significance of candidate DMRs via 1000 permutations. We also noticed that DMRcate may not be able to identify regions with small effect sizes comparing with t-test (with true-positive rates up to around 6 of the 10 regions with signals, while true-positive rates for t-test could be up to around 8); Probe Lasso also has small false-positive rates, but the true-positive rates are smaller than that of the new method, and it also uses Stouffer's method to combine weighted individual P-values, and thus also leads to a much smaller P-values for DMRs compared with the new method. Similarly, when the true DMRs are set to have sites with variance signals only, the new method performs slightly inferior to Pitman–Morgan test that considers variance signals only while all other five comparing methods could not detect any true DMRs. When the true DMRs are set to have sites with both mean and variance signals, the new method performs much better than all of the six comparing methods.

The type I errors and ROC curve results of the adapted algorithm are summarized in the Supplementary Data (Simulation studies for case-control designs).

Real data application

We used two data sets, TCGA BRCA data (tumor and normal-adjacent pairs) and GEO BRCA data (tumor and normal-adjacent pairs, normal controls from age-matched cancer-free women), to demonstrate the performance of the new method from three aspects: (1) identification of DMRs associated with tumor and normal-adjacent status (TCGA BRCA tumor versus normal-adjacent comparison; and GEO BRCA tumor versus normal-adjacent comparison); (2) replication with two independent BRCA data (TCGA BRCA tumor versus normal-adjacent comparison; and GEO BRCA tumor versus normal-adjacent comparison); (3) identification of epigenetic field defects (GEO age-matched cancer-free versus normal-adjacent comparison); (4) enrichment of epigenetic alterations from age-matched cancer-free to normal-adjacent to tumor tissues (GEO age-matched cancer-free versus normal-adjacent comparison, GEO BRCA tumor versus normal-adjacent comparison and GEO age-matched cancer-free versus tumor comparison).

TCGA BRCA data

We applied the new method and the six comparing methods to the TCGA BRCA 450 K DNA methylation data of tumor and normal-adjacent tissues. The original data have DNA methylation measures on 485 577 CpG sites for 96 tumor and normal-adjacent pairs. We conducted standard quality control steps where we removed sites on sex chromosomes and sites overlap with known single-nucleotide polymorphisms (SNPs). We also required at least 95% CpG coverage per sample and 70% sample coverage per CpG sites. We ended up with 326 105 CpG sites for 90 matched tumor and normal-adjacent pairs. We then corrected for the type II probe bias using the 'wateRmelon' package [75].

We found that DMRs identified by the Wilcoxon signed-rank test and KS test are larger than others (both in terms of number of sites and bps) in general, while those by the mean-only method are the smallest (paired t-test, DMRcate and Probe

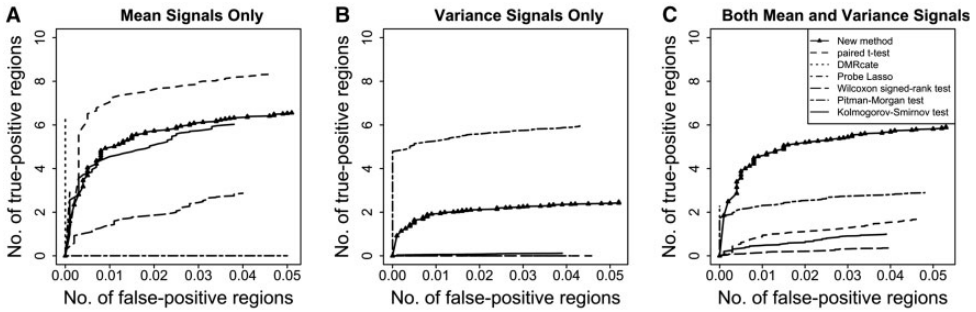

Figure 2. ROC curves from simulation studies where 10 true DMRs have different region sizes ranging from 3 to 15 CpG sites with (**A**) mean signals only; (**B**) variance signals only; and (**C**) both mean and variance signals. DMRs were defined as regions with minimum region size of $L \geq 3$ CpG sites.

Table 1. Significant DMRs identified in the TCGA BRCA data

$L^a \geq 3$)	New method	Wilcoxon signed-rank test	Paired *t*-test	DMRcate	Probe Lasso	Pitman–Morgan test	KS test[b]
Total number of DMRs (total number of DMR-covered CpG sites)	986 (18 654)	135 (4 295)	1473 (21 777)	20 657 (133 410)	7190 (36 936)	610 (13 208)	720 (16 047)
Mean (SD) number of CpG sites per DMR	19 (8)	32 (8)	15 (8)	6 (5)	5 (5)	22 (10)	22 (11)
Mean (SD) number of base pairs per DMR	3373 (1989)	5341 (2387)	2669 (1726)	1185 (1064)	748 (1148)	3804 (2361)	4144 (2492)
Number of overlapping DMRs[c]	–	129	806	986	642	533	588

[a]*L*: minimum region size, i.e. minimum number of CpG sites.
[b]KS test.
[c]Number of overlapping DMRs: a DMR identified by the new method is considered to overlap with DMRs identified by each comparing method if there is any overlap.

Lasso), and those by the new method and Pitman–Morgan test are in between (Table 1). On the CpG site level, 69.2, 13.1, 15.4 and 93.6% of sites in the DMRs that were identified by the mean-only methods: paired *t*-test, DMRcate, Probe Lasso and Wilcoxon signed-rank test were also identified by the new method; 83.6% of sites in the DMRs identified by the Pitman–Morgan test were also identified by the new method, and 77.4% of sites in the DMRs identified by the KS test were also identified by the new method. Further investigation reveals that DMRs identified uniquely by the paired *t*-test and Pitman–Morgan test were all defined by the new method but did not reach significance.

When comparing DMRs identified by the new method to those by the six comparing methods, the new method did not identify any unique DMRs. All DMRs identified by the new method overlap with those identified by DMRcate, where we define overlap if any CpG sites in a DMR identified by the new method are also in a DMR identified by DMRcate. When comparing DMRs identified by the new method to those by the five of the six comparing methods but not DMRcate, the new method uniquely identified 22 DMRs. Among these 22 DMRs, we further examined the top 10 DMRs ranked by the evidence of strength of each region. There are 11 genes in these 10 DMRs, and all were previously reported to be associated with cancer (Table 2). We plotted the top ranked #1 and #2 DMRs out of the 22 uniquely identified DMRs for illustration (Figure 3), where both DMRs were hyper-methylated. In DMR #1, sites in the second half of the region do not have any mean differences between tumor and normal-adjacent tissues but have large variance differences. However, the variance signals are not strong enough to be detected by the variance-only method. For the sites in the first half of DMR #1, there are both mean and variance differences, but are not strong enough to be detected by most of the mean-only or variance-only methods.

Table 2. Eleven genes identified in the top 10 ranked DMRs in TCGA BRCA data

Cancer	Gene
Breast cancer	*LBH* [76]
Chordomas	*NPR3* [77]
Clear cell renal cell carcinoma	*SMPD3* [78]
Gastric cancer	*FGF19* [79]
Head and neck squamous cell carcinomas	*PHF21B* [80]
Hepatocellular carcinoma	*MYADM* [81], *DBX2* [82]
Non-small cell lung cancer	*KCNC3* [83]
Oral squamous cell carcinoma	*ATP8B2* [84]
Prostate cancer	*AQP10* [85], *STEAP2* [86]

We also applied the new method to the TCGA KIRC 450 K DNA methylation data, and observed similar patterns as in the TCGA BRCA data. Results are included in the Supplementary Data (Supplementary Tables S1 and S2 and Supplementary Figure S3).

Replication analysis with GEO BRCA data

We performed a replication analysis using an independent DNA methylation data of BRCA tumor and normal-adjacent tissues from GEO (GSE69914) [35]. The original GEO BRCA DNA methylation data have methylation measures on 385 184 CpG sites from 42 tumor and normal-adjacent pairs, 50 normal/benign controls from age-matched cancer-free women and 263 tumor tissues from independent breast cancer patients. We followed the same quality control steps as for the TCGA BRCA data and kept the same sets of CpG sites as in the TCGA BRCA data for

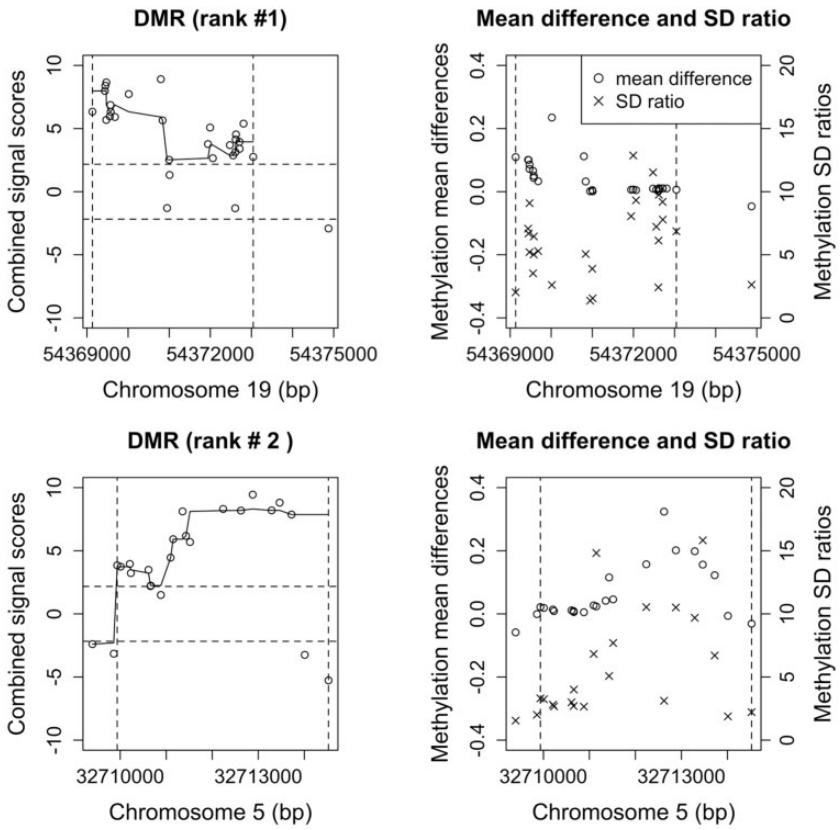

Figure 3. Top two ranked DMRs uniquely identified by the new method in the TCGA BRCA data. DMR #1 (top row) and #2 (bottom row) are located on chromosomes 19 and 5. Vertical dashed lines define boundaries of DMRs. Left column shows the combined signal scores of the sites in the DMRs before (circles) and after (curves) smoothing, in which horizontal dotted lines define the threshold *k* that defines a candidate region. Right column shows the mean differences and SD ratios in methylation measures of sites in the DMRs comparing tumor and normal-adjacent tissues.

comparison purpose. We ended up with 326 105 CpG sites from 42 tumor and normal-adjacent pairs.

We compared results from the TCGA BRCA data and the GEO BRCA data and found that 94.7, 94.4, 87.6, 87.2, 86.3, 80.2 and 95.4% of sites in the DMRs identified in the GEO BRCA tumor versus normal-adjacent comparison were also identified in the TCGA BRCA tumor versus normal-adjacent comparison by the new method, paired *t*-test, DMRcate, Probe Lasso, Wilcoxon signed-rank test, Pitman–Morgan test and KS test, respectively. We plotted two example overlapping DMRs (Figure 4). The first DMR is hypo-methylated and ranks #1 among all DMRs identified by the new method in both BRCA data sets. The second DMR is hyper-methylated and ranks #5 among all DMRs identified by the new method in the TCGA BRCA data and ranks #13 among all DMRs identified by the new method in the GEO BRCA data.

Although sites in DMRs identified using the Pitman–Morgan test have the smallest replication rate (80.2%) as expected, it is still large enough to conclude that DMRs detected using variance signals are reproducible. Sites in DMRs identified in the two BRCA data sets using the new method, paired *t*-test and KS test can be almost perfectly reproduced with replication rates 94.7, 94.4 and 95.4%, respectively. This agrees with the general believe that DMR findings might be more reliable than DML findings, supporting the meaning of detecting DMRs.

Details of the DMRs identified by the new method and the six comparing methods in the GEO BRCA data are summarized in Table 3. We found that 67.5, 7.7, 10.7 and 86.8% of sites in the DMRs that were identified by the mean-only methods: paired *t*-test, DMRcate, Probe Lasso and Wilcoxon signed-rank test were also identified by the new method, 83.0% of sites in the DMRs identified by the Pitman–Morgan test were also identified by the new method and 81.9% of sites in the DMRs identified by KS test were also identified by the new method. Similarly as in the TCGA BRCA and KIRC data, DMRs identified uniquely by the paired *t*-test or Pitman–Morgan test were all defined by the new method but did not reach significance.

We similarly plotted the top #1 and #2 ranked DMRs in the GEO BRCA tumor versus normal-adjacent comparison (Supplementary Figure S4) and further investigated genes in the top 10 DMRs (Supplementary Table S3), where similar patterns were observed as in the TCGA BRCA and KIRC data.

Identification of epigenetic field defects in the GEO BRCA data

Teschendorff et al. [35] showed in their recent paper that the identification of early epigenetic alterations, commonly known as epigenetic field defects, through comparing DNA methylation measures of normal-adjacent tissues from breast cancer patients to normal tissues from age-matched cancer-free women is meaningful in the study of breast cancer development, and the differences are expected to be larger in comparisons between tumor and normal-adjacent tissues, and between tumor and normal tissues from cancer-free women. The original paper investigated the epigenetic field defects on the CpG site level. Here, we further investigated epigenetic field defects on the region level. We kept the same 326 105 CpG sites as in the GEO BRCA tumor versus normal-adjacent comparison.

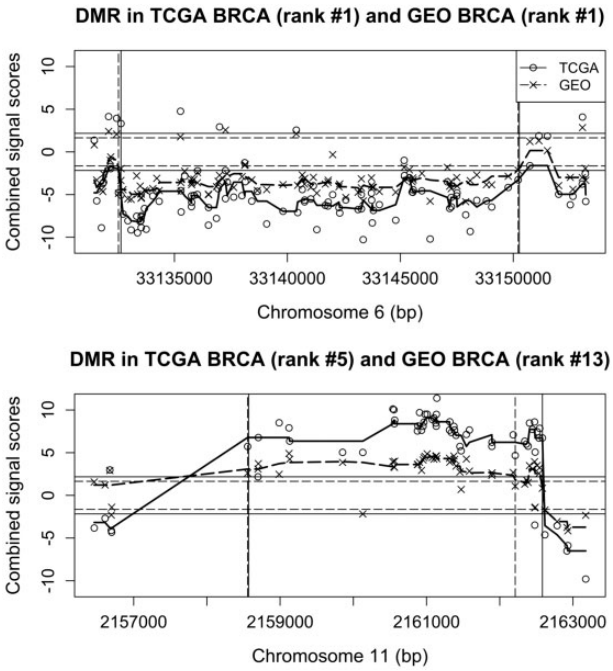

Figure 4. Two examples of overlapping DMRs among all DMRs identified by the new method in both the TCGA and the GEO BRCA data from the tumor versus normal-adjacent comparison. Plotted are the combined site-level signal scores in the DMRs before and after smoothing for the TCGA BRCA data (circles, solid curves) and the GEO BRCA data (crosses, dashed curves). Vertical lines define boundaries of DMRs, and horizontal lines define the threshold k that defines a candidate region.

We first examined the distributions of the estimated genome-wide site-level scaling parameter λ_i from the three comparisons (Supplementary Figure S5) (1) normal-adjacent tissues from breast cancer patients versus normal tissues from age-matched cancer-free women, (2) tumor tissues versus matched normal-adjacent tissues from breast cancer patients and (3) tumor tissues from breast cancer patients versus normal tissues from age-matched cancer-free women in the GEO BRCA data.

As defined in Equation (2), the site-level scaling parameter λ_i reflects the relative strength of the mean and variance signals at CpG site i. CpG sites with $\lambda_i = 0$ do not have any variance signals, CpG sites with $\lambda_i = 1$ do not have any mean signals and CpG sites with $0 < \lambda_i < 1$ have both mean and variance signals, within which sites with $\lambda_i > 0.5$ have stronger variance signals than mean signals. Supplementary Figure S5 suggests that in the normal-adjacent versus normal comparison, there are much fewer sites with both mean and variance signals and a lot more sites with only variance signals comparing with the other two comparisons. The parameter λ that reflects the genome-wide relative signal strength is also the largest in the normal-adjacent versus normal comparison. This suggests that differential variation exists earlier in disease progression, which is consistent with the findings by Teschendorff et al. [35] that there is increased variability in DNA methylation within the normal-adjacent tissues comparing with normal breast tissue from age-matched cancer-free women.

We then examined the identified DMRs in the three comparisons using the GEO BRCA data (1) normal-adjacent versus normal, (2) tumor versus normal-adjacent and (3) tumor versus normal. In the normal-adjacent versus normal comparison that aims for epigenetic field defects, the new method identified two DMRs (Supplementary Figure S6 shows the mean and variance signals of the two DMRs), both hyper-methylated, while all the six comparing methods identified none. Importantly, all 58 CpG sites covered by these two DMRs of epigenetic field defects are also in the DMRs identified in the tumor versus normal-adjacent, and tumor versus normal comparisons (results summarized in Supplementary Table S4 and Supplementary Figure S7). These 58 sites cover two genes, *NKX6-2* and *CCND2*. Both were previously reported to be differentially methylated in breast cancer [87–89]. Moreover, for the two DMRs of epigenetic field defects, the #1 ranked DMR ranks #18 in the tumor versus normal-adjacent comparison and ranks #11 in the tumor versus normal comparison (Figure 5A); the #2 ranked DMR ranks #2 in the other two comparisons (Figure 5A). Sites in these two DMRs have larger combined signal scores in the tumor versus normal-adjacent, and tumor versus normal comparisons than those from the normal-adjacent versus normal comparison. This suggests that there exists epigenetic field defects earlier in disease progression, and the epigenetic field defects are enriched in the progression to breast cancer, confirming what is observed before on the CpG site level [35] using region-based method.

To further investigate whether the epigenetic field defects identified in the normal-adjacent versus normal comparison are because of a few outlier samples as Teschendorff et al. [37] noticed, and whether the notable DNA methylation alterations at the identified CpG sites are real but not technical artifact, we plotted two heat maps of sites in the two DMRs of epigenetic field defects (Figure 5B). It is clear that there is little variation in DNA methylation measures of normal tissues, and increased variation in those of normal-adjacent tissues because of three to five samples and much increased variability in those of tumor tissues. We further selected two sites with large variance signals ($\lambda_i = 0.72$ and 0.76) out of the 58 sites covered by the two DMRs and plotted their methylation measures (Figure 5C). It is clear that at these two sites, there is little variation in DNA methylation measures of normal tissues and increased variation in those of normal-adjacent tissues, mainly because of to three to five outlier samples, and there is no mean difference in DNA methylation measures between the normal and normal-adjacent tissues (Figure 5C). We also notice that the three outlier samples exhibit greater methylation deviations in tumor tissues than in normal-adjacent tissues, indicating an enriched methylation alteration with cancer progression (Figure 5C). We would like to emphasize that the two DMRs of epigenetic field defects in the normal-adjacent versus normal comparison were only identified by the new DMR detection method that uses mean and variance combined signals but were missed by all the other six comparison methods, which suggests the great power achieve by the new DMR detection method.

Discussion

Here, we proposed a new DMR detection method that uses combined signals from differential methylation and DV. Simulation studies showed the correct type I error and the much improved power of the new method when true DMRs have sites with both mean and variance signals. Applications to the TCGA BRCA, TCGA KIRC and GEO BRCA DNA methylation data showed that the majority of genes in the uniquely identified DMRs by the new method were previously reported to be associated with cancers. Replication analysis results using two independent BRCA data sets suggest that DMRs detected with variance signals are reproducible.

Table 3. Significant DMRs identified in the GEO BRCA data

DMRs ($L^a \geq 3$)	New method	Wilcoxon signed-rank test	Paired *t*-test	DMRcate	Probe Lasso	Pitman–Morgan test	KS test[b]
Total number of DMRs (total number of DMR-covered CpG sites)	382 (8769)	126 (3948)	490 (9606)	15 609 (104 713)	4505 (24 053)	22 (653)	213 (5503)
Mean (SD) number of CpG sites per DMR[b]	23 (8)	31 (8)	20 (8)	7 (5)	5 (5)	30 (7)	26 (8)
Mean (SD) number of base pairs per DMR[c]	4047 (2130)	5216 (2279)	3420 (1880)	1127 (999)	713 (1097)	3718 (2457)	4478 (2034)
Number of overlapping DMRs[c]	–	114	284	382	262	19	175

[a]*L*: Minimum region size, i.e. minimum number of CpG sites.
[b]KS test.
[c]Number of overlapping DMRs: a DMR identified by the new method is considered to overlap with DMRs identified by each comparing method if there is any overlap.

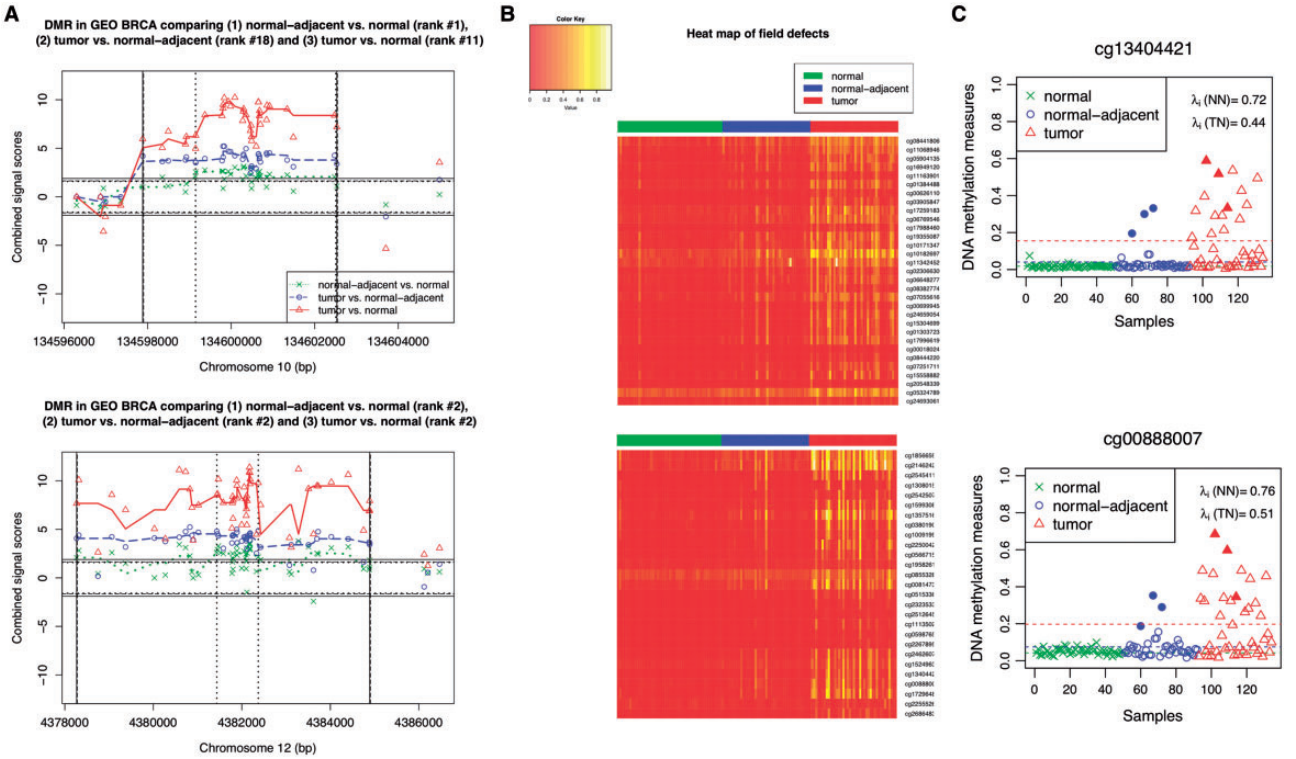

Figure 5. (**A**) The epigenetic field defects, i.e. the two DMRs identified in the GEO BRCA normal-adjacent versus normal comparison (crosses, dotted curves) together with the overlapping DMRs identified in the GEO BRCA tumor versus normal-adjacent comparison (circles, dashed curves), and the GEO BRCA tumor versus normal comparison (triangles, solid curves). Vertical lines define boundaries of DMRs, and horizontal lines define the threshold *k* that defines a candidate region. (**B**) Heat maps of the original DNA methylation measures of the sites from the epigenetic field defects, i.e. the two DMRs identified in the GEO BRCA normal-adjacent versus normal comparison. Green is for 50 normal tissues from age-matched cancer-free women. Blue is for 42 normal-adjacent tissues, and red is for 42 tumor tissues. (**C**) Two CpG sites selected of the 58 sites from the epigenetic field defects, i.e. the two DMRs. Plotted are the original DNA methylation measures of normal tissues from 50 age-matched cancer-free women (crosses), and normal-adjacent tissues (circles) and tumor tissues (triangles) from 42 BRCA patients. The three horizontal lines represent mean methylation levels of the three groups. λ_i (NN) is the site-level scaling parameter from the normal-adjacent versus normal comparison, and λ_i (TN) is that from the tumor versus normal-adjacent comparison. The three outlier samples were marked using solid circles (normal-adjacent tissues) and solid triangles (matching tumor tissues).

Importantly, further application to the DNA methylation data of GEO BRCA normal-adjacent tissues from breast cancer patients and normal tissues from age-matched cancer-free women identified epigenetic field defects in two DMRs only by the new method, while the comparing mean-only and variance-only methods identified none. These two DMRs were also identified, and the methylation alterations were enriched in the comparisons of tumor versus normal-adjacent tissues and tumor versus normal tissues. The identified epigenetic field defects in these two DMRs could potentially be marks for breast cancer early detection with future investigations. Owing to the fact that the identified early DNA methylation alterations in breast cancer are characterized by increased variability because of a few 'outlier' samples when both mean and variance signals are weak and mean-only method and variance-only method could detect no differences, existing methods that focus on mean signals only or adapted methods that focus on variance signals only will be seriously underpowered. This shows the importance of using mean and variance combined signal, especially in identifying epigenetic field defects.

Although we did not consider correcting for differences in variances between batches, in the context of the data presented in

this article, this is not an issue for the following two reasons (1) many previous studies [35, 66, 90] have unequivocally demonstrated that most of the differentially variable loci (DVL) are not batch or technical effects; (2) DVL are indeed generally characterized by fairly large changes in DNA methylation (>30% if not more) affecting a small number of samples, whereas batch effects generally involve smaller (10–15%) changes in DNA methylation, which affect most if not all the samples within a batch.

Furthermore, we did not adjust for cell-type composition in our analysis as Teschendorff et al. [35] have clearly demonstrated that DVL are not driven by changes in cell-type composition: (1) the DVL do not map to markers of adipose cells or immune cells, which are two main types of cell contaminants in breast tissue, (2) changes in cell-type composition between two phenotypes (e.g. normal versus normal-adjacent, or normal versus tumor) only involve relatively smaller changes in DNA methylation (10–15%). In contrast, DVL generally involve much larger changes in DNA methylation (>30%), which only affect a smaller number of samples. Their previous study also demonstrated that (3) the same DVL were found after adjustment for changes in cell-type composition. Put together, it is clear that most of the DVL are unrelated to cell-type composition changes, and that they instead mark pre-cancerous cells on route to becoming cancerous.

One thing we noticed in using methylation variance signals is, when methylation M-values are used, the mean and variance signals may not be completely separated. We have conducted some simulation studies in our previous work and found that if only mean signals are designed in the M-values, there will be both mean and variance signals in β-values after the transformation [68].

In summary, we proposed a new DMR detection method that uses mean and variance combined signals. Although we applied the new method to multiple cancer data sets, the method can be applied to other complex diseases. We focused on methylation array data in this work, but the new method is readily applied to sequencing data with sequencing data being preprocessed to methylation proportions. An R code for the developed DMR detection method together with a tutorial and a sample data set is available for downloading from http://www.columbia.edu/~sw2206/softwares.htm.

Key Points

- The new DMR detection method is able to identify additional DMRs that are related to cancer but were missed by methods that use mean or variance signals only.
- DMRs detected using variance signals are reproducible.
- The new DMR detection method is able to identify epigenetic field defects, and the defects are enriched in the progression to breast cancer.

Supplementary data

Supplementary data are available at *BIB* online.

Reference

1. Baylin SB, Esteller M, Rountree MR, *et al*. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Hum Mol Genet* 2001;**10**:687–92.
2. Fahrner JA, Eguchi S, Herman JG, *et al*. Dependence of histone modifications and gene expression on DNA hypermethylation in cancer. *Cancer Res* 2002;**62**:7213–18.
3. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 2012;**13**:484–92.
4. Phillips T. The role of methylation in gene expression. *Nat Educ* 2008;**1**:116.
5. Das PM, Singal R. DNA methylation and cancer. *J Clin Oncol* 2004;**22**:4632–42.
6. Ehrlich M. DNA methylation in cancer: too much, but also too little. *Oncogene* 2002;**21**:5400–13.
7. Esteller M, Herman JG. Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. *J Pathol* 2002;**196**:1–7.
8. Kulis M, Esteller M. DNA methylation and cancer. *Adv Genet* 2010;**70**:27–56.
9. Koukoura O, Spandidos DA, Daponte A, *et al*. DNA methylation profiles in ovarian cancer: implication in diagnosis and therapy. *Mol Med Rep* 2014;**10**:3–9.
10. Baylin SB. DNA methylation and gene silencing in cancer. *Nat Clin Pract Oncol* 2005;**2**:S4–11.
11. Curradi M, Izzo A, Badaracco G, *et al*. Molecular mechanisms of gene silencing mediated by DNA methylation. *Mol Cell Biol* 2002;**22**:3157–73.
12. Herman JG, Baylin SB. Gene silencing in cancer in association with promoter hypermethylation. *N Eng J Med* 2003;**349**:2042–54.
13. Robertson KD. DNA methylation and human disease. *Nat Rev Genet* 2005;**6**:597–610.
14. Eden A, Gaudet F, Waghmare A, *et al*. Chromosomal instability and tumors promoted by DNA hypomethylation. *Science* 2003;**300**:455.
15. Feinberg AP, Tycko B. The history of cancer epigenetics. *Nat Rev Cancer* 2004;**4**:143–53.
16. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* 2003;**33**:245–54.
17. Ruike Y, Imanaka Y, Sato F, *et al*. Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. *BMC Genomics* 2010;**11**:137.
18. Teschendorff AE, Menon U, Gentry-Maharaj A, *et al*. An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS One* 2009;**4**:e8274.
19. Lasseigne BN, Burwell TC, Patil MA, *et al*. DNA methylation profiling reveals novel diagnostic biomarkers in renal cell carcinoma. *BMC Med* 2014;**12**:235.
20. Hinoue T, Weisenberger DJ, Lange CP, *et al*. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res* 2012;**22**:271–82.
21. Feinberg AP. Phenotypic plasticity and the epigenetics of human disease. *Nature* 2007;**47**:433–40.
22. De Jager PL, Srivastava G, Lunnon K, *et al*. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nat Neurosci* 2014;**17**:1156–63.
23. Lund G, Andersson L, Lauria M, *et al*. DNA methylation polymorphisms precede any histological sign of atherosclerosis in mice lacking apolipoprotein E. *J Biol Chem* 2004;**279**:29147–54.
24. Mill J, Petronis A. Molecular studies of major depressive disorder: the epigenetic perspective. *Mol Psychiatry* 2007;**12**:799–814.
25. Mill J, Petronis A. Pre- and peri-natal environmental risks for attention-deficit hyperactivity disorder (ADHD): the potential role of epigenetic processes in mediating susceptibility. *J Child Psychol Psychiatry* 2008;**49**:1020–30.
26. Mill J, Tang T, Kaminsky Z, *et al*. Epigenomic profiling reveals DNA-methylation changes associated with major psychosis. *Am J Hum Genet* 2008;**82**:696–711.

27. Nestler EJ. Epigenetic mechanisms of drug addiction. *Neuropharmacology* 2014;**76**:259–68.
28. Schanen NC. Epigenetics of autism spectrum disorders. *Hum Mol Genet* 2006;**15**:138–50.
29. Phipson B, Oshlack A. DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biol* 2014;**15**:465.
30. Hansen KD, Timp W, Bravo HC, *et al*. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* 2011;**43**:768–75.
31. Feinberg AP, Irizarry RA. Evolution in health and medicine Sackler colloquium: stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc Natl Acad Sci USA* 2010;**107**:1757–64.
32. Gervin K, Hammerø M, Akselsen HE, *et al*. Extensive variation and low heritability of DNA methylation identified in a twin study. *Genome Res* 2011;**21**:1813–21.
33. Jaffe AE, Feinberg AP, Irizarry RA, *et al*. Significance analysis and statistical dissection of variably methylated regions. *Biostatistics* 2012;**13**:166–78.
34. Fernandez AF, Assenov Y, Martin-Subero JI, *et al*. A DNA methylation fingerprint of 1628 human samples. *Genome Res* 2012;**22**:407–19.
35. Teschendorff AE, Gao Y, Jones A, *et al*. DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nat Commun* 2016;**7**:10478.
36. Slaughter DP, Southwick HW, Smejkal W. "Field cancerization" in oral stratified squamous epithelium. Clinical implications of multicentric origin. *Cancer* 1953;**6**:963–8.
37. Teschendorff AE, Jones A, Widschwendter M. Stochastic epigenetic outliers can define field defects in cancer. *BMC Bioinformatics* 2016;**17**:178.
38. Akalin A, Kormaksson M, Li S, *et al*. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* 2012;**13**:R87.
39. Chen Z, Huang H, Liu J, *et al*. Detecting differentially methylated loci for Illumina Array methylation data based on human ovarian cancer data. *BMC Med Genomics* 2013;**6**:S9.
40. Huang H, Chen Z, Huang X. Age-adjusted nonparametric detection of differential DNA methylation with case-control designs. *BMC Bioinformatics* 2013;**14**:86.
41. Shen J, Wang S, Zhang YJ, *et al*. Genome-wide DNA methylation profiles in hepatocellular carcinoma. *Hepatology* 2012;**55**:1799–808.
42. Sun H, Wang S. Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. *Bioinformatics* 2012;**28**:1368–75.
43. Sun H, Wang S. Network-based regularization for matched case-control analysis of high-dimensional DNA methylation data. *Stat Med* 2013;**32**:2127–39.
44. Eckhardt F, Lewin J, Cortese R, *et al*. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* 2006;**38**:1378–85.
45. Irizarry RA, Ladd-Acosta C, Carvalho B, *et al*. Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res* 2008;**18**:780–90.
46. Irizarry RA, Ladd-Acosta C, Wen B, *et al*. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* 2009;**41**:178–86.
47. Lister R, Pelizzola M, Dowen RH, *et al*. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;**462**:315–22.
48. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* 2012;**13**:R83.
49. Jaffe AE, Murakami P, Lee H, *et al*. Bump hunting to identifying differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol* 2012;**41**:200–9.
50. Butcher LM, Beck S. Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data. *Methods* 2015;**72**:21–8.
51. Jühling F, Kretzmer H, Bernhart SH, *et al*. metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res* 2016;**26**:256–62.
52. Hesse N, Schröder C, Rahmann S. An optimization approach to detect differentially methylated regions from whole genome bisulfite sequencing data. *PeerJ PrePrints* 2015;**3**:e1287v3.
53. Wen Y, Chen F, Zhang Q, *et al*. Detection of differentially methylated regions in whole genome bisulfite sequencing data using local Getis-Ord statistics. *Bioinformatics* 2016;**32**:3396–404.
54. Ayyala DN, Frankhouser DE, Ganbat JO, *et al*. Statistical methods for detecting differentially methylated regions based on MethylCap-seq data. *Brief Bioinform* 2016;**17**:926–37.
55. Sofer T, Schifano ED, Hoppin JA, *et al*. A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure. *Bioinformatics* 2013;**29**:2884–91.
56. Yip WK, Fier H, DeMeo DL, *et al*. A novel method for detecting association between DNA methylation and diseases using spatial information. *Genet Epidemiol* 2014;**38**:714–21.
57. Mayo TR, Schweikert G, Sanguinetti G. M3D: a kernel-based test for spatially correlated changes in methylation profiles. *Bioinformatics* 2015;**31**:809–16.
58. Saito Y, Mituyama T. Detection of differentially methylated regions from bisulfite-seq data by hidden Markov models incorporating genome-wide methylation level distributions. *BMC Genomics* 2015;**16**:S3.
59. Saito Y, Tsuji J, Mituyama T. Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions. *Nucleic Acids Res* 2014;**42**:e45.
60. Peters TJ, Buckley MJ, Statham AL, *et al*. *De novo* identification of differentially methylated regions in the human genome. *Epigenetics Chromatin* 2015;**8**:6.
61. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics* 1946;**2**:110–14.
62. Liu H, Liu X, Zhang S, *et al*. Systematic identification and annotation of human methylation marks based on bisulfite sequencing methylomes reveals distinct roles of cell type-specific hypomethylation in the regulation of cell identity genes. *Nucleic Acids Res* 2016;**44**:75–94.
63. Ahn S, Wang T. A powerful statistical method for identifying differentially methylated markers in complex diseases. *Pac Symp Biocomput* 2013:69–79.
64. Chen Y, Ning Y, Hong C, *et al*. Semiparametric tests for identifying differentially methylated loci with case-control designs using Illumina Arrays. *Genet Epidemiol* 2014;**38**:42–50.
65. Teschendorff AE, Liu X, Caren H, *et al*. The dynamics of DNA methylation covariation patterns in carcinogenesis. *PLoS Comput Biol* 2014;**10**:e1003709.
66. Teschendorff AE, Widschwendter M. Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics* 2012;**28**:1487–94.

67. Ruan P, Shen J, Santella RM, *et al*. NEpiC: a network-assisted algorithm for epigenetic studies using mean and variance combined signals. *Nucleic Acids Res* 2016;**44**:e134.
68. Sun H, Wang Y, Chen Y, *et al*. pETM: a penalized Exponential Tilt Model for analysis of correlated high-dimensional DNA methylation data. *Bioinformatics* 2017;**33**:1765–72.
69. Morgan W. A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika* 1939;**31**:13–19.
70. Pitman EJG. A note on normal correlation. *Biometrika* 1939;**31**: 9–12.
71. Wu H, Xu T, Feng H, *et al*. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Res* 2015;**43**:e141.
72. Hebestreit K, Dugas M, Klein HU. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics* 2013;**29**:1647–53.
73. Du P, Zhang X, Huang CC, *et al*. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 2010;**11**:587.
74. Stouffer SA, Suchman EA, DeVinney LC, *et al*. *The American soldier: adjustment during army life.(Studies in social psychology in World War II)*, Vol. **1**. Princeton, NJ: Princeton University Press, 1949.
75. Pidsley R, Wong CC, Volta M, *et al*. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* 2013;**14**:293.
76. Many AM, Brown AM. Mammary stem cells and cancer: roles of Wnt signaling in plain view. *Breast Cancer Res* 2010;**12**:313.
77. Alholle A, Brini A, Bauer J, *et al*. Genome-wide DNA methylation profiling of recurrent and non-recurrent chordomas. *Epigenetics* 2015;**10**:213–20.
78. Wang J, Li J, Gu J, *et al*. Abnormal methylation status of FBXW10 and SMPD3, and associations with clinical characteristics in clear cell renal cell carcinoma. *Oncol Lett* 2015;**10**:3073–80.
79. Zhao J, Liang Q, Cheung KF, *et al*. Genome-wide identification of Epstein-Barr virus–driven promoter methylation profiles of human genes in gastric cancer cells. *Cancer* 2013;**119**:304–12.
80. Bertonha FB, de Camargo Barros Filho M, Kuasne H, *et al*. PHF21B as a candidate tumor suppressor gene in head and neck squamous cell carcinomas. *Mol Oncol* 2015;**9**:450–62.
81. Song MA, Tiirikainen M, Kwee S, *et al*. Elucidating the landscape of aberrant DNA methylation in hepatocellular carcinoma. *PLoS One* 2013;**8**:e55761.
82. Zhang P, Wen X, Gu F, *et al*. Methylation profiling of serum DNA from hepatocellular carcinoma patients using an Infinium Human Methylation 450 BeadChip. *Hepatol Int* 2013; **7**:893–900.
83. Lokk K, Vooder T, Kolde R, *et al*. Methylation markers of early-stage non-small cell lung cancer. *PLoS One* 2012;**7**: e39813.
84. Yong Deok K, Eun Hyoung J, Yeon Sun K, *et al*. Molecular genetic study of novel biomarkers for early diagnosis of oral squamous cell carcinoma. *Med Oral Patol Oral Cir Bucal* 2015; **20**:e167–79.
85. Raza K, Jaiswal R. Reconstruction and analysis of cancer-specific gene regulatory networks from gene expression profiles. *Int J Bioinforma Biotechnol Biosci* 2013;**3**(2):25–34.
86. Gomes IM, Maia CJ, Santos CR. STEAP proteins: from structure to applications in cancer therapy. *Mol Cancer Res* 2012;**10**: 573–87.
87. Fackler MJ, Umbricht CB, Williams D, *et al*. Genome-wide methylation analysis identifies genes specific to breast cancer hormone receptor status and risk of recurrence. *Cancer Res* 2011;**71**:6195–207.
88. Sharma G, Mirza S, Prasad CP, *et al*. Promoter hypermethylation of p16 INK4A, p14 ARF, CyclinD2 and Slit2 in serum and tumor DNA from breast cancer patients. *Life Sci* 2007;**80**: 1873–81.
89. Virmani A, Rathi A, Heda S, *et al*. Aberrant methylation of the cyclin D2 promoter in primary small cell, non-small cell lung and breast cancers. *Int J Cancer* 2003;**107**:341–5.
90. Teschendorff AE, Jones A, Fiegl H, *et al*. Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome Med* 2012;**4**:24.