HOME: A histogram based machine learning approach for effective identification of differentially methylated regions

Akanksha Srivastava[1], Yuliya V Karpievitch[1,2], Steven R Eichten[3], Justin O Borevitz[3], and Ryan Lister*[1,2]

[1]ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Perth, Australia.

[2]Harry Perkins Institute of Medical Research, Perth, Australia.

[3]ARC Centre of Excellence in Plant Energy Biology, The Australian National University, Canberra, Australia.

Keywords: WGBS, methylation, epigenetics, DMR identification, SVM

*Corresponding author. Email: ryan.lister@uwa.edu.au (R.L.)

Abstract

DNA methylation is a covalent modification of DNA that plays important role in regulating gene expression, cell identity, and organism development. Localized changes in DNA methylation are observed between different cell types, during development and aging, in various disease states, and under different stress conditions, and are often associated with functionally important genomic regions, including promoters and enhancers. The development of whole genome bisulfite sequencing has made it possible to identify methylation differences at single base resolution throughout an entire genome. A persistent challenge in DNA methylome analysis is the accurate identification of differentially methylated regions (DMRs) in the genome between samples. Sensitive and specific identification of DMRs between different conditions requires accurate and efficient algorithms, and while various tools have been developed to tackle this problem, they frequently suffer from limitations in sensitivity and accuracy. Here, we present a novel Histogram Of MEthylation (HOME) based method that exploits the inherent difference in distribution of methylation levels between DMRs and non-DMRs to robustly discriminate between the two via a linear Support Vector Machine. HOME produces accurate DMR boundaries, few spurious DMRs, and provides the ability to determine DMRs in time-series data. HOME can identify DMRs among any number of treatment groups in experiments with or without replicates at high accuracy. We demonstrate that HOME produces more accurate DMRs than the current state-of-the-art methods on both simulated and biological datasets, and provide a user-friendly implementation of the tool.

Introduction

DNA methylation plays an important role in the regulation of various cell functions including genomic imprinting, X-chromosome inactivation and cellular differentiation (Richardson 2002; Khavari et al. 2010; Messerschmidt et al. 2014). However, analysis of DNA methylation presents various challenges as the modification is highly dynamic in space and time (Jones

2012; Lister et al. 2013). DNA methylation levels vary between distinct genomic features such as promoters, enhancers, gene bodies, transposable elements, and repeat elements (Kass et al. 1997; Jones 1999; Meissner et al. 2008; Lister et al. 2009; Law and Jacobsen 2010; Stadler et al. 2011; Bogdanovic et al. 2016). Furthermore, widespread variation in the distribution of DNA methylation has been observed between different cell types, cell lines, tissues, individuals and species (Heyn et al. 2013; Roadmap Epigenomics et al. 2015; Schultz et al. 2015; Eichten et al. 2016; Kawakatsu et al. 2016; Niederhuth et al. 2016). Moreover, the distribution of DNA methylation is not uniform across all cytosines in the genome. In mammals, DNA methylation predominantly occurs in the CG dinucleotide context, however multiple studies have uncovered the presence of non-CG (CH, where H=A, T, or C) methylation in certain cell types including embryonic stem cells and brain cells (Lister et al. 2009; Xie et al. 2012; Lister et al. 2013; Varley et al. 2013). In contrast, DNA methylation in plants occurs in all sequence context, namely CG, CHG, and CHH (Law and Jacobsen 2010). Furthermore, CH methylation is often found at much lower levels compared to CG methylation, as measured by the proportion of reads displaying methylation, making the accurate analysis of CH DNA methylation more challenging given the typical sequencing depth of experiments to date.

High-throughput sequencing methods such as whole genome bisulfite sequencing (WGBS) have been developed to provide detection and quantitative measurement of DNA methylation at single base resolution throughout whole genomes (Cokus et al. 2008; Lister et al. 2009). Sodium bisulfite treatment of genomic DNA converts cytosines, but not methylcytosines, into uracils, and during subsequent PCR amplification of the bisulfite treated DNA the uracils are replaced by thymines. High-throughput sequencing of bisulfite converted DNA and alignment to a reference genome enables the methylation level of any covered cytosine to be computed by counting the number of methylated and unmethylated bases in reads that cover that cytosine position. Sensitive and accurate DMR detection from such data is important in characterization of the differences and dynamics of DNA methylation state, exploration of potential roles in genome regulation, and as disease biomarkers (Guo et al. 2017). However, accurate

DMR detection remains a significant challenge. Most of the existing DMR identification methods such as bsseq (Hansen et al. 2012), RADMeth (Dolzhenko and Smith 2014), MACAU (Lea et al. 2015) and GetisDMR (Wen et al. 2016) are more appropriate to identify DMRs when two or more replicates are available for each of the treatment groups (Shafi et al. 2017). Other methods such as Comet (Saito et al. 2014) and swDMR (Wang et al. 2015) have been developed to identify DMRs for single replicate treatment groups. Two of the recently developed methods, DSS and DSS-single (Feng et al. 2014; Wu et al. 2015) (referred to as DSS hereafter), and Metilene (Juhling et al. 2016) can be used for single or multiple replicate treatment groups and have been shown to outperform the aforementioned methods. However, both of these methods are limited to DMR identification between two treatment groups rather than more complex experimental designs with multiple groups and/or time points. Moreover, multiple characteristics need to be considered for accurate prediction of DMRs, including spatial correlation present between neighboring cytosine sites, sequencing depth that takes into account sampling variability that occurs during sequencing, and biological variation among replicates of treatment groups (Eckhardt et al. 2006; Hansen et al. 2012; Jaffe et al. 2012; Shafi et al. 2017). Most of the DMR identification tools described above do not consider either all or some of the characteristics required for accurate prediction of DMRs.

To overcome these limitations we have developed HOME, a novel DMR finder that takes into account important characteristics such as cytosine spatial correlation, sequencing depth, and biological variation between replicates for predicting accurate DMRs for both single and multiple replicate treatment groups. Moreover, HOME is computationally very efficient for predicting DMRs in the CH context, where the number of potential sites of methylation in the genome are significantly greater than in the CG context. Furthermore, HOME has the functionality to identify DMRs in time-series data to accurately identify temporal changes in DNA methylation state. A detailed comparison of HOME with the most commonly used method, DSS, and a recently developed method, Metilene, demonstrates that HOME achieves high performance on both simulated and biological data. HOME outperforms both DSS and Meti-

lene by predicting more accurate DMR boundaries and having lower false positive and false negative rates.

Results

Algorithm summary

The algorithm developed here approaches the problem of DMR identification from the perspective of binary classification in machine learning, classifying a region as DMR or non-DMR using a Support Vector Machine (SVM) classifier (Cortes 1995), as described below. The classifier is trained separately for the CG and CH contexts, as they show different methylation characteristics. Due to the lack of real-world truth data for DMRs and non-DMRs, we generated a training dataset using publicly available DNA methylome datasets. For the CG context, WGBS datasets of neuronal and non-neuronal cell types isolated from the frontal cortex of 7 week old male mouse prefrontal cortex (Lister et al. 2013) was used. For the CH context, the classifier was trained on methylation data from two neuronal cell types: excitatory pyramidal neurons (EX) and vasoactive intestinal peptide-expressing interneurons (VIP) (Mo et al. 2015).

To generate the training data, different DMR finders such as DSS, bisulfighter and Metilene were used, with default parameters, on the processed methylome data. The DMRs identified by two or more algorithms were further analyzed to confirm the presence of methylation level differences. Among these, the DMRs with an average methylation level difference >0.3 were inspected visually for contiguous methylation level differences and were selected as robust DMRs. The regions not identified as DMRs by all of the DMR finders were randomly selected as potential non-DMRs. Among these, the regions with low average methylation level differences (<0.2) were visually inspected for consistent methylation level differences and were selected as robust non-DMRs. Thereafter, HOME computes the methylation level difference between treatment groups and estimates the significance of methylation level difference (modeled by p-value) for each cytosine site in DMRs and non-DMRs. The p-values were computed

using weighted logistic regression for replicated data, and z-score test for non-replicated data (see Methods). The algorithm accounts for biological variation present between the replicates and uneven read coverage through weighted logistic regression while computing the p-value. The salient information, including the methylation level difference and the measure of significance for the difference in methylation level, is combined to generate a bin value for each cytosine site in DMRs and non-DMRs (see eq.1 in Methods). To account for the spatial correlation between the neighboring cytosines, moving average smoothing is performed on the computed bin values for each cytosine site in DMRs and non-DMRs. The smoothed bin values, ranging from 0 to 1, are then binned into a histogram of 10 bins with the bin size of 0.1, to generate the proposed novel histogram based features for each cytosine, which are then used to train the SVM classifier. Briefly, the histogram features are computed for each cytosine site in DMRs and non-DMRs for a window of fixed size. For a given window, the bin values are binned using a weighted voting approach such that for a given cytosine, its contribution to the bin is computed as a weighted distance from the center cytosine, which is normalized by the maximum allowed distance (eq. 2 in Methods). The use of weighted voting for computation of histogram based features also captures the spatial correlation present among neighboring cytosine sites.

The schematic of the feature generation process described above is illustrated with an example DMR and non-DMR selected from the training dataset in Figure 1 (A-H). The proposed histogram based features (Figure 1D and 1H) show a clear demarcation between DMRs and non-DMRs. In particular, the distributions of non-DMRs show low mean values for the bins representing the higher difference in methylation level (>0.3), indicating low number of votes falling in the bins that correspond to higher methylation differences (Figure 1I). In contrast, DMRs exhibit higher differences in methylation level and have consistently higher mean for bins that correspond to higher methylation differences (Figure 1I). This indicates that the histogram based features are highly discriminative between treatment states, which makes the problem of DMR detection suitable for machine learning analysis.

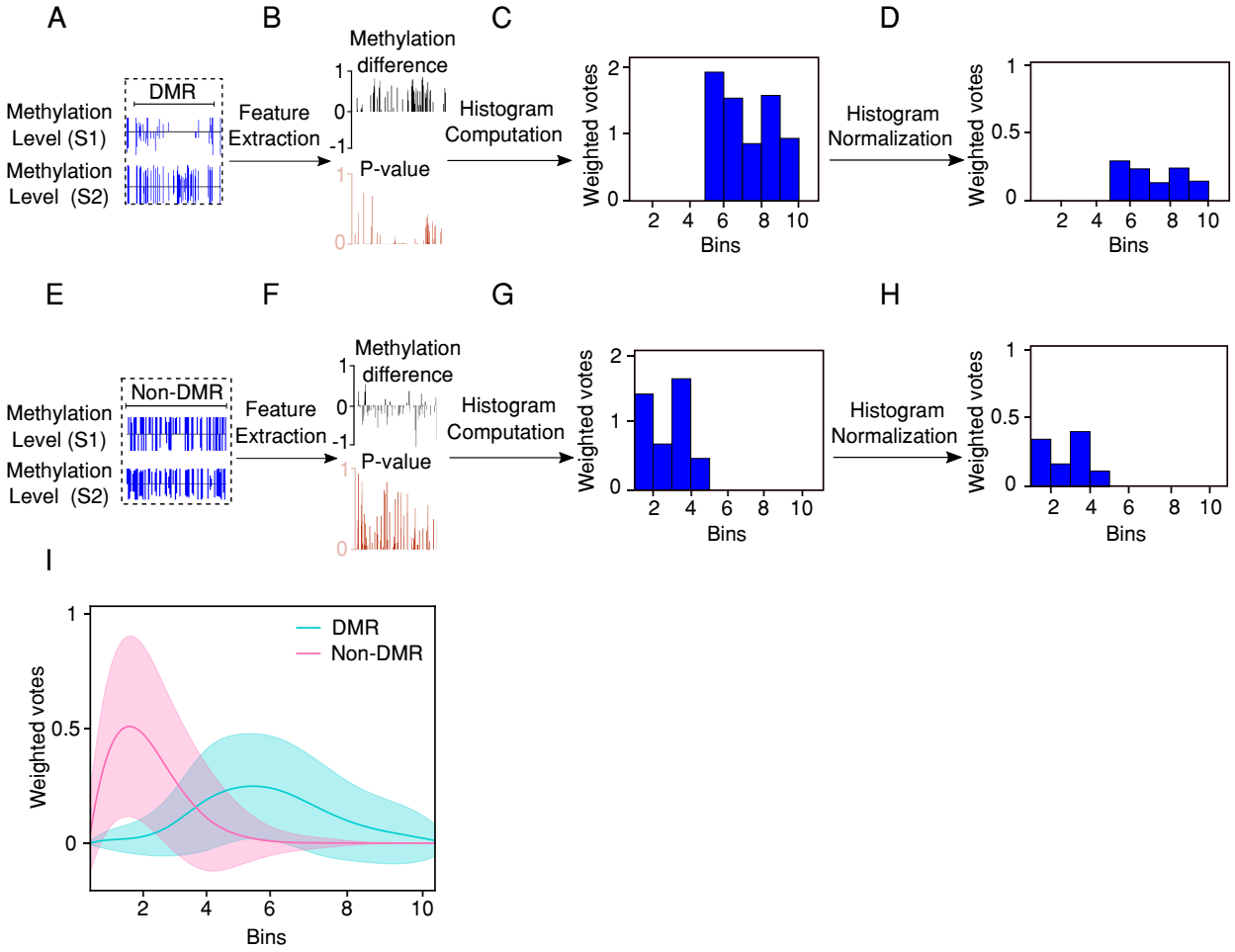

Figure 1. Feature generation overview. **(A)** Methylation level of sample 1 (S1) and sample 2 (S2) for a DMR from the training set. **(B)** Extracted features: p-value and difference in methylation level for each CG site. **(C)** Histogram of scores computed from the extracted features and **(D)** histogram of normalized scores. **(E)** Methylation level of S1 and S2 for a non-DMR from the training dataset. **(F)** Extracted features: p-value and difference in methylation level for each CG. **(G)** Histogram of scores computed from the extracted features and **(H)** histogram of normalized scores. **(I)** Mean and standard deviation of histogram features for complete training data for DMRs (blue) and non-DMRs (pink).

Once the SVM has been trained, the histogram based features for new methylomes can be computed, and HOME scans the entire methylome to provide a prediction score (between 0 and 1) from the SVM classifier for each cytosine site. Then, the individual cytosine sites are grouped together into DMRs based on the user defined prediction score cutoff (default: >0.1) and the distance between neighboring cytosines (default: <500bp). A lower prediction score (<0.1) from the classifier for a cytosine site indicates low confidence in the site being differentially methylated and a higher prediction score (>0.1) indicates high confidence in a site

being methylated. Here, we independently applied HOME to both CG and CH contexts, and compared its performance to two other DMR finders, Metilene and DSS, using both simulated and biological data. Furthermore, we also show that HOME can be used for time-series DMR analysis on biological data.

Analysis of simulated DNA methylation data

The DMRs were simulated with the assumption that the methylation level of cytosine sites follows a beta binomial distribution. DMRs and non-DMRs were simulated with three distinct beta binomial settings, each increasing in their difficulty for identification. For each setting two treatment groups were simulated, each with three replicates. The read coverage for replicates was taken from WGBS datasets of neuronal and non-neuronal cell types (Lister et al. 2013). For all settings, the methylation level of cytosine sites in DMRs were generated from a beta distribution of (6,1.5) and a beta distribution of (1.5,6) for the two treatment groups, respectively. For the first setting (class 1), the non-DMR portion of the genome displayed a fixed methylation level (0.7), with only DMRs showing variation from this value. In the second setting (class 2), non-DMR cytosine methylation level was simulated from beta parameters (2,2), such that the methylation level was not fixed for either DMRs or non-DMRs. The final setting (class 3) was similar to class 2, however use of the non-DMR beta parameters of (1.5,1.5) provided an overall broader distribution under which DMR methylation states can exist by random chance. Therefore, in these simulated datasets it is expected that correct DMR classification will be easiest for class 1 and hardest for class 3. The simulation method described above is similar to approach taken previously (Dolzhenko et al 2014). For class 1, we generated additional data with two treatment groups with single replication to test the efficacy of HOME in the cases where replication is not available.

The performance of HOME, DSS and Metilene were evaluated in terms of the Jaccard index, positive predictive value (PPV), true positive rate (TPR), F1-score and boundary detection accuracy. The Jaccard index is the measure of similarity between simulated and predicted

DMRs, where a Jaccard index of 1 indicates the perfect identification of DMRs and 0 represents complete misidentification of simulated DMRs.

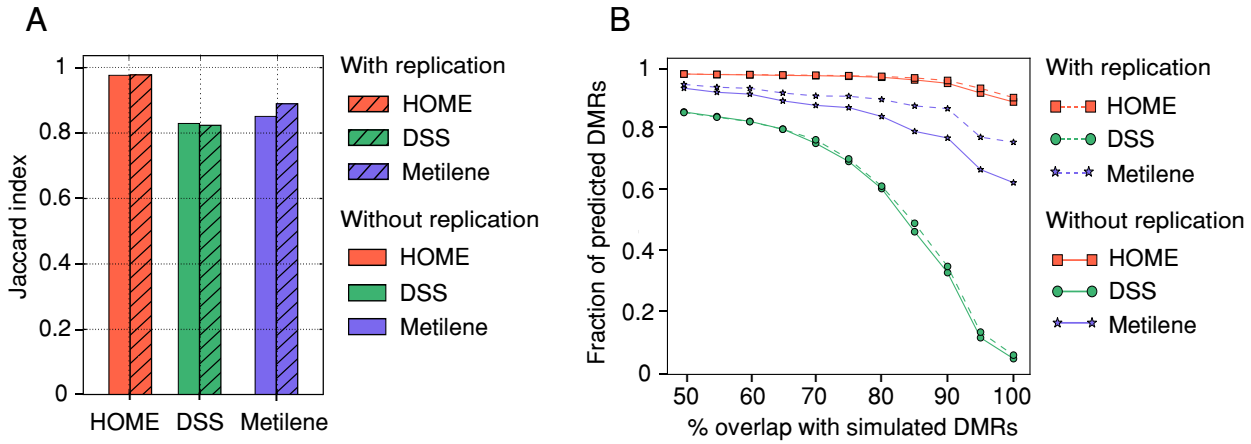

Figure 2. Comparison of DMR identification methods (HOME, DSS and Metilene) on simulated data. (A) Jaccard index for predicted DMRs by HOME (orange), DSS (green) and Metilene (purple) with single and multiple replicates in the treatment groups. **(B)** Percent reciprocal overlap ranging from 50-100% between simulated and predicted DMRs by HOME (orange), DSS (green) and Metilene (purple) with single and multiple replicates in the treatment groups.

Because the default settings of the various DMR finders might not produce the best possible results for each method, an extensive parameter search (Supplementary Materials) for each method was performed to identify the settings for each DMR finder that resulted in the best performance based on the aforementioned evaluation criteria. HOME outperforms DSS and Metilene in terms of the Jaccard index for both replicated and non-replicated data (Figure 2A). The Jaccard index of HOME for replicated and unreplicated data was 0.996 and 0.995, respectively, indicating a high level of identification compared to the known truth. Furthermore, HOME outperformed both DSS and Metilene in all other metrics (PPV, TPR and F1-score), for all classes with distinct biological settings (Supplementary Figure 1A and Supplementary Table 1).

Precise DMR boundary identification is essential to any analysis that follow DMR detection. Imprecise definition of DMR boundaries may lead to inaccurate biological findings, such as when DMRs are linked to promoters or enhancers to understand the connection between methylation changes and phenotype. Therefore, we evaluated the DMR boundary accuracy

for all three DMR finders. To measure the boundary accuracy, we calculated the fraction of CG overlap between simulated and predicted DMRs (see methods). The accuracy was tested for a comprehensive range of reciprocal overlap from 50-100% between simulated and predicted DMRs. HOME produced more accurate DMR boundaries than DSS and Metilene for replicated and non-replicated data (Figure 2B). For example, in the range of 50-80% reciprocal overlap between simulated and predicted DMRs, HOME correctly predicted 99% of simulated DMRs. The fraction of correctly predicted DMRs by HOME decreased slightly to 98% for the range of 80-90% reciprocal overlap, and 90% for the range of 90-100% reciprocal overlap. This means that the 90% of predicted DMRs perfectly matched the boundary of the simulated DMRs. In contrast, for DSS and Metilene the fraction of predicted DMRs decreased with an increasing requirement in the percentage overlap between the predicted and simulated DMRs. This indicates that HOME is highly accurate in predicting the boundaries of the DMRs, outperforming both DSS and Metilene for all three classes with replicates and without replicates (Figure 2B and Supplementary Figure 1B).

Analysis of performance on biological datasets

1. Pairwise differential methylation analysis

We compared the performance of HOME with DSS and Metilene for the CG context on published WGBS data (Mo et al. 2015) for two neuronal cell types: excitatory pyramidal neurons (EX) and parvalbumin-expressing fast-spiking interneurons (PV), each with two replicates. Among the DMR finders that were compared, DMRs identified by HOME appeared to be less fragmented than the DMRs produced by other finders (Figure 3A), despite equivalent DMR merging parameters being used for each finder (Supplementary Materials). The number of DMRs with inter-DMR distance <500bp was lower for DMRs predicted by HOME as compared to DSS and Metilene (Figure 3B), even though a merge distance of 500bp was utilized for both DSS and HOME. Note that there is no parameter to control the merge distance in

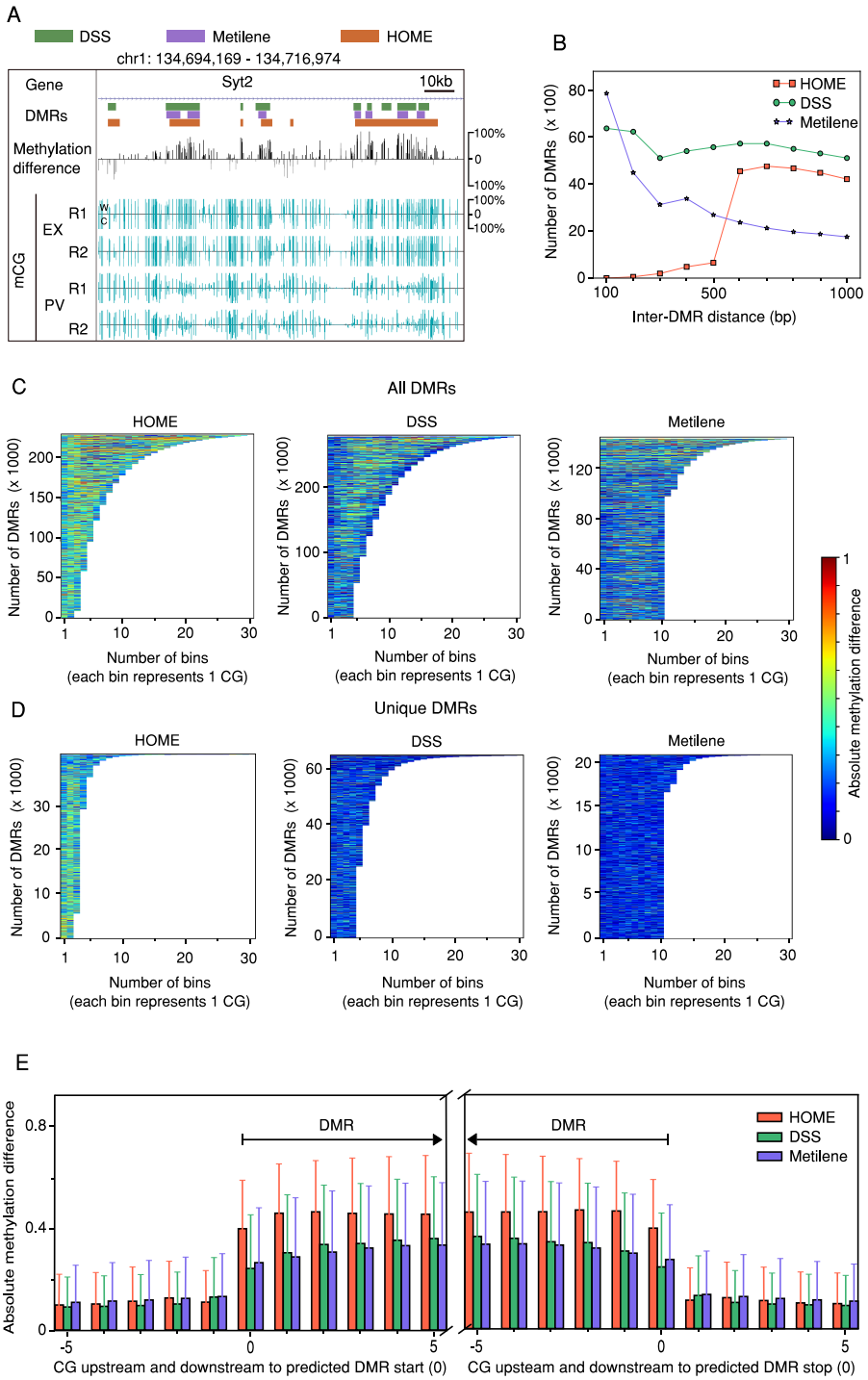

Figure 3. Quality assessment of CG-DMRs predicted in mammalian WGBS data by HOME, DSS and Metilene. (A) Browser representation showing the quality and boundary accuracy of predicted CG context DMRs for the parvalbumin-expressing fast-spiking interneuron (PV) specific gene Syt2. EX, excitatory pyramidal neurons; R1, replicate 1; R2, replicate 2. Methylated CG sites (blue) are marked by upward (Watson strand, W) and downward (Crick strand, C) ticks. **(B)** Distribution of inter-DMR distance for predicted CG-DMRs. The length of DMRs, shown on X-axis, are in the range (bin -100, bin]. **(C)** Heatmap of absolute methylation level difference for all predicted DMRs, and **(D)** uniquely predicted DMRs. DMRs are arranged in descending order of number of CGs. **(E)** Mean and standard deviation of absolute methylation difference for all predicted CG-DMRs for 5 CGs upstream and downstream of the DMR start (left) and stop (right) marked as 0, respectively.

Metilene. HOME predicted DMRs with a consistently higher methylation level difference (>0.3) for CG sites when compared to the other finders (Figure 3C). A more detailed comparison of the DMRs uniquely predicted by each finder showed that DMRs uniquely identified by HOME consistently had a higher methylation level difference (>0.3) for the CG sites located within the DMRs (Figure 3D). In contrast, DMRs uniquely predicted by DSS and Metilene consistently had a lower methylation difference (<0.2) for the CG sites within the DMRs. Furthermore, a one-to-one comparison of the overlapping and uniquely predicted DMRs of DSS and Metilene with HOME showed that for the overlapping DMRs between HOME and DSS, DMRs predicted by HOME had consistently higher methylation level differences as compared to DSS (Supplementary Figure 2A). DMRs uniquely predicted by DSS compared to HOME had low methylation level differences (Supplementary Figure 2B). On the contrary, DMRs uniquely predicted by HOME compared to DSS exhibited consistently higher methylation level differences, indicating that these DMRs, although short, display reasonably high differences in methylation (Supplementary Figure 2B). Similar trends were observed between DMRs identified by Metilene and HOME (Supplementary Figure 3A and 3B).

Given the importance of accurate boundary detection, we next compared boundary precision at high resolution among the three finders. The boundary detection accuracy for all DMRs predicted by HOME, DSS and Metilene was represented by the mean and standard deviation of the absolute methylation level difference between analyzed samples for the 5 CG sites immediately inside and outside the DMR boundary, for each end of the DMR (Figure 3E). The mean methylation level difference for the boundary CG site and CG sites inside the DMRs is higher for HOME than DSS and Metilene (Figure 3E), while the mean methylation level difference for the CG sites immediately outside the DMR boundaries is lower for HOME as compared to DSS and Metilene. This demonstrates the higher DMR boundary detection precision of HOME on real biological datasets, as also observed for the simulated WGBS data (Figure 2B). One-to-one comparisons of the boundary detection accuracy of DSS or Metilene with HOME follow a similar trend for the overlapping predicted DMRs (Supplementary Figure

4A and 4B). However, for the DMRs predicted uniquely by each finder (Supplementary Figure 5A and 5B), the mean methylation level difference between CGs within the DMR and CGs outside the HOME DMR boundaries were higher than for the DMRs identified by other finders. Overall, this indicates that HOME provides more accurate DMR boundary prediction.

To investigate the incidence of false positive DMRs predicted by each finder, we permuted the labels among the EX and PV WGBS samples to generate two artificial datasets: (1) EX replicate 1 and PV replicate 1 comprising treatment group 1 vs EX replicate 2 and PV replicate 2 comprising treatment group 2 and (2) EX replicate 1 and PV replicate 2 vs EX replicate 2 and PV replicate 1, comprising treatment groups 1 and 2 respectively. Due to randomness in the shuffled data, there will be shorter regions with contiguous methylation level differences occurring by chance, and these short regions will be identified by the algorithms as DMRs. However, it is expected that there will be significantly fewer long DMRs in the shuffled data, as long DMRs will be unlikely to occur by random chance. These shuffled datasets were analyzed with HOME, DSS and Metilene, and as expected the number of DMRs predicted by all algorithms in comparison to the unshuffled data was significantly reduced (Supplementary Figure 6A and 6B). In addition, DMRs identified by HOME in the shuffled data were smaller and contained low number of CGs as compared to DSS and Metilene (Supplementary Figure 6A and 6B). Moreover, DMRs identified by HOME exhibited higher methylation level differences (0.2 - 0.3, Supplementary Figure 6C). The DMRs identified by DSS and Metilene tended a higher number of CGs per DMR and had a very low methylation level difference (Supplementary Figure 6C), and therefore are more likely to be false positive DMRs.

To test the performance of HOME for the CH context, we used the same two neuronal cell type (EX and PV) as used for the CG context. A genome browser view of a representative genomic region (Figure 4A) shows that HOME predicts CH-DMRs where there are regions of contiguous methylation level difference at CH sites between EX and PV. High absolute methylation level difference (>0.2) was observed in all of the predicted DMRs (Figure 4B). CH-DMRs predicted by HOME have accurate boundaries with higher mean inter-sample methylation level

difference at and within DMR boundaries, and low mean methylation level difference immediately outside the DMR boundaries (Figure 4C). Furthermore, quantitative analysis for HOME CH-DMR features showed that 36% of the predicted DMRs are longer than 1 kb (Figure 4D). In addition, the distribution of the number of CH cytosines in predicted DMRs showed that 70% of the DMRs contain >50 cytosines (Figure 4E).

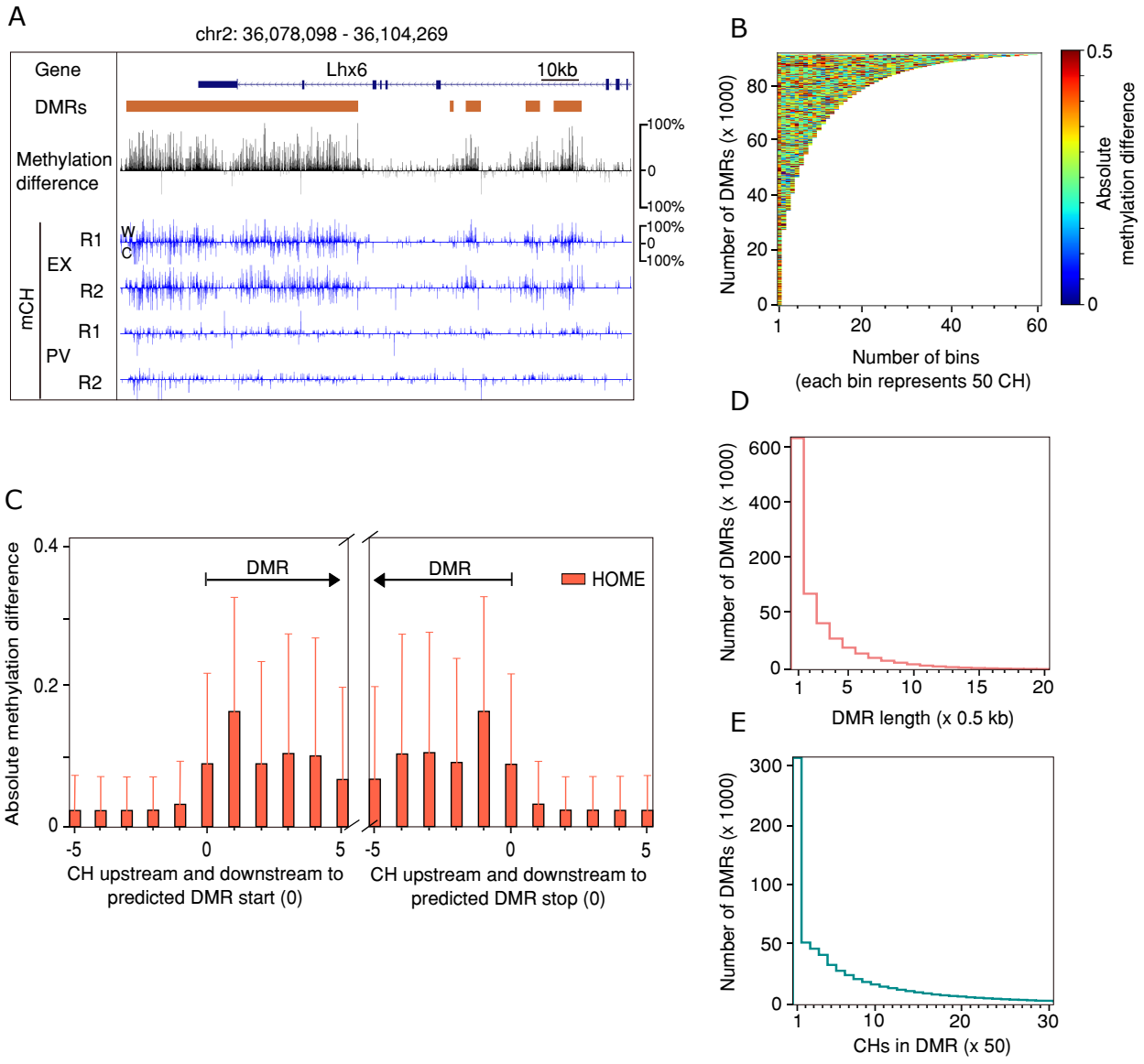

Figure 4. Quality assessment of CH-DMRs predicted in mammalian WGBS data. (A) Genome browser representation of the quality and boundary accuracy of HOME predicted CH context DMRs for the PV interneuron specific gene Lhx6. **(B)** Heatmap showing the absolute methylation difference for HOME CH context DMRs. DMRs are arranged in descending order of number of CHs. **(C)** DMR boundary precision analysis for all HOME predicted CH-DMRs. **(D)** Distribution of the length of HOME predicted CH-DMRs. **(E)** Distribution of the number of CHs in the CH-DMRs predicted by HOME.

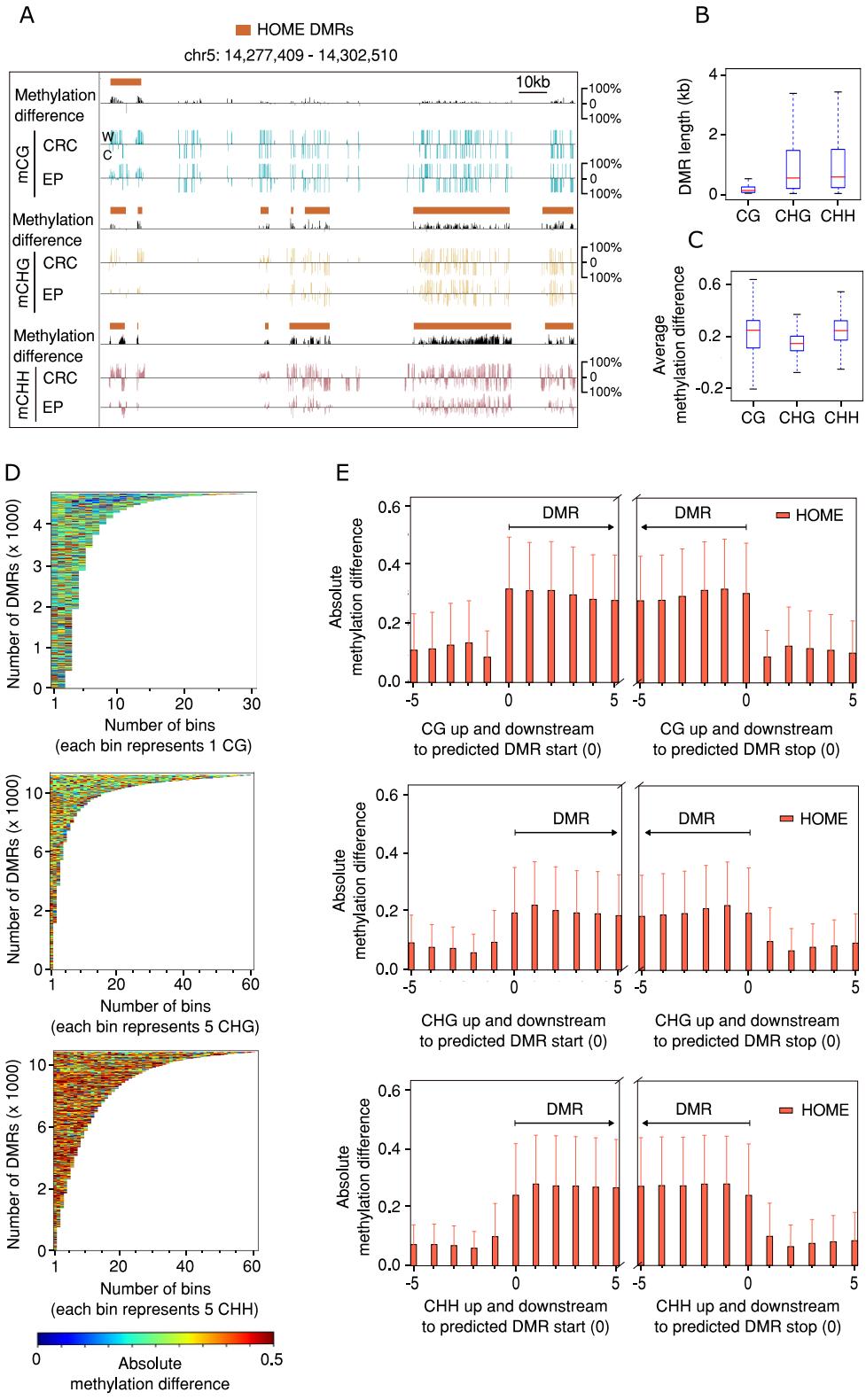

Figure 5. Qualitative and quantitative analysis of all DMRs predicted by HOME in plant WGBS data. (A) Genome browser representation of all HOME predicted DMRs and the underlying methylation difference (black), **(B)** length distribution, and **(C)** average methylation difference distribution for Arabidopsis columella root cap (CRC) and epidermis (EP), for CG, CHG and CHH contexts. **(D)** Heatmap showing the absolute methylation differences, and **(E)** DMR boundary precision analysis for HOME predicted DMRs for CG, CHG and CHH contexts.

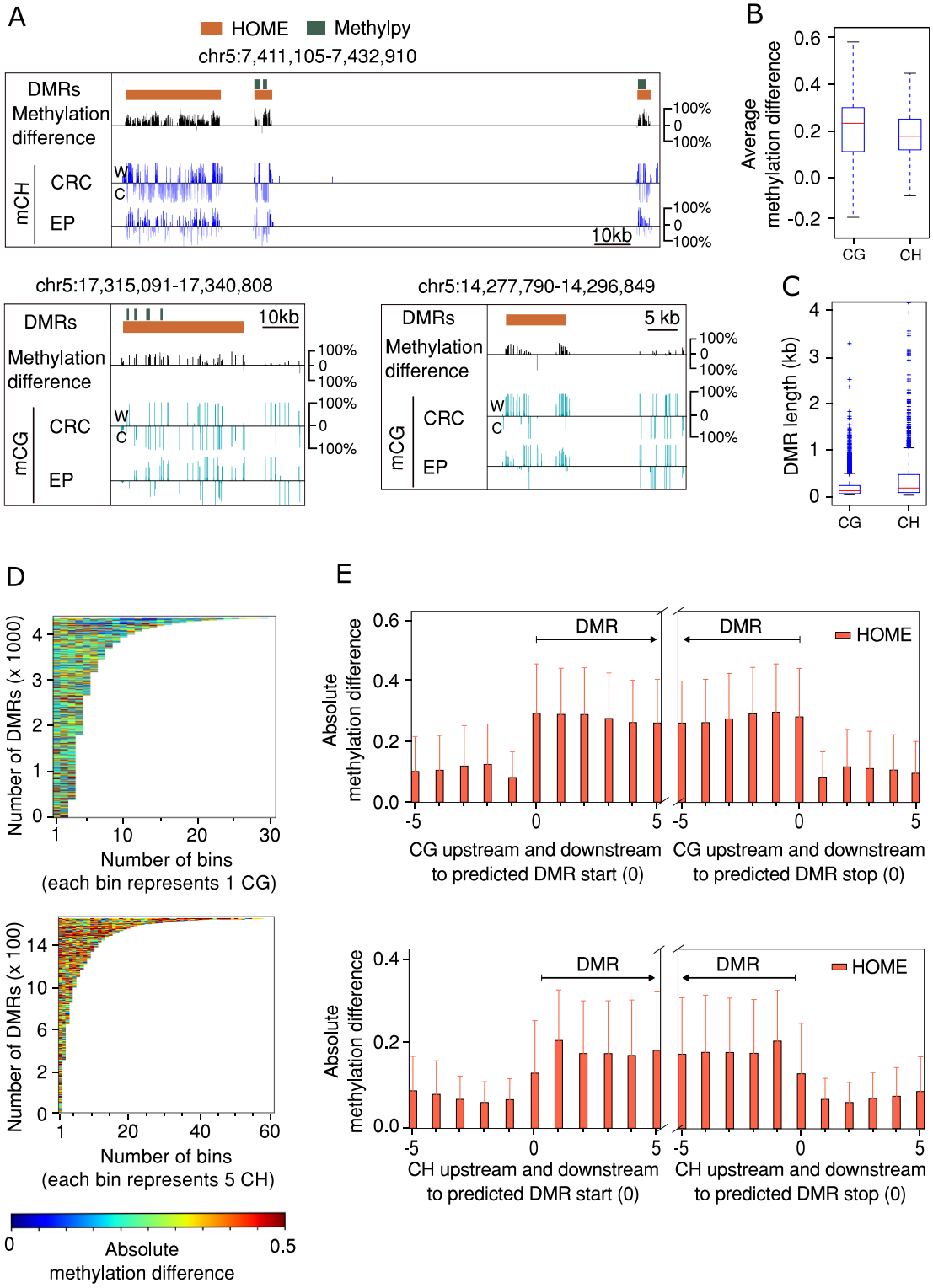

Figure 6. Qualitative and quantitative analysis of uniquely predicted DMRs by HOME in plant WGBS data. **(A)** Genome browser representation of DMRs predicted by HOME and published DMRs predicted by Methylpy for CRC and EP for CH (top panel) and CG context (bottom two panels). **(B)** Average methylation difference distribution, and **(C)** length distribution of uniquely predicted HOME DMRs for CG and CH context. **(D)** Heatmap of the absolute methylation difference, and **(E)** DMR boundary precision analysis for uniquely predicted HOME DMRs for CG and CH contexts.

The performance of HOME was also tested on plant data using published WGBS DNA methylomes of columella root cap (CRC) and epidermis (EP) cell types of the Arabidopsis root meristem (Kawakatsu et al. 2016). Note that we used the same model that is trained on the mammalian dataset for predicting the DMRs in the plant WGBS datasets, highlighting the fact that the model generalizes effectively. HOME separately identified the regions with contiguous methylation level differences for CG, CHG and CHH contexts (Figure 5A). HOME identified a total of 30,930 DMRs, of which 16%, 44% and 40% were identified in the CG, CHG and CHH contexts, respectively. Analysis of the length and average methylation level difference of the predicted DMRs showed that CG-DMRs are shorter and have higher methylation level differences compared to CHG and CHH-DMRs (Figure 5B and 5C). These results are consistent with the results presented in the published study (Kawakatsu et al. 2016). To investigate the quality of predicted DMRs by HOME, we plotted the methylation level difference for all three contexts (Figure 5D). The plots show that the DMRs predicted by HOME for all three contexts exhibit consistently high (>0.2) methylation level differences. The mean and standard deviation of 5 cytosines upstream and downstream of the predicted DMRs start and stop sites, respectively, showed that the methylation level difference is high within the DMR and low just outside the DMRs (Figure 5E), indicating accurate boundary prediction in all contexts in plant data, despite the SVM being trained on mammalian data.

We further compared DMRs predicted by HOME with the published DMRs predicted by Methylpy in Kawakatsu et al. (Kawakatsu et al. 2016). Note that the published DMRs were predicted between five distinct cell types. Whereas, for this experiment, we perform a pairwise comparison between only two cell types (CRC and EP) for DMR prediction by HOME. Therefore, our comparison of HOME and Methylpy is not exact for overlapping DMRs. Nonetheless, 87% of CG-DMRs and 10% of CH-DMRs are uniquely predicted by HOME (Figure 6A). The uniquely predicted DMRs exhibit high methylation level differences for both CG and CH contexts (Figure 6B). Furthermore, the distribution of the length of the DMRs uniquely predicted by HOME for both CG and CH contexts shows a large number of DMRs with the length greater

than 500 bp (Figure 6C), indicating that these uniquely predicted DMRs are long and highly likely to be true DMRs. Moreover, the methylation level difference is consistently high within the DMRs for both CG and CH contexts (Figure 6D), and the DMRs uniquely predicted by HOME exhibit accurate boundary identification (Figure 6E).

2. Time-series differential methylation analysis

A unique feature of HOME is the ability to predict DMRs in time-series data. The HOME time-series analysis algorithm can be successfully used for identification of DMRs in datasets where DNA methylation varies with development stages, for example, seed germination (Narsai et al. 2017), epigenomic reprogramming (Lister et al. 2011; Lee et al. 2014) and mammalian brain development (Lister et al. 2013). To the best of our knowledge, no other published DMR identification method can be directly used to predict DMRs in time-series WGBS data. The performance of HOME on time-series data was tested using human brain frontal cortex development WGBS data (Lister et al. 2013) with four time points: fetal, 2 years, 12 years and 25 years (Figure 7). The processed data with methylated and unmethylated read counts per CG context cytosine was used as an input for HOME. DMRs identified by HOME included expected

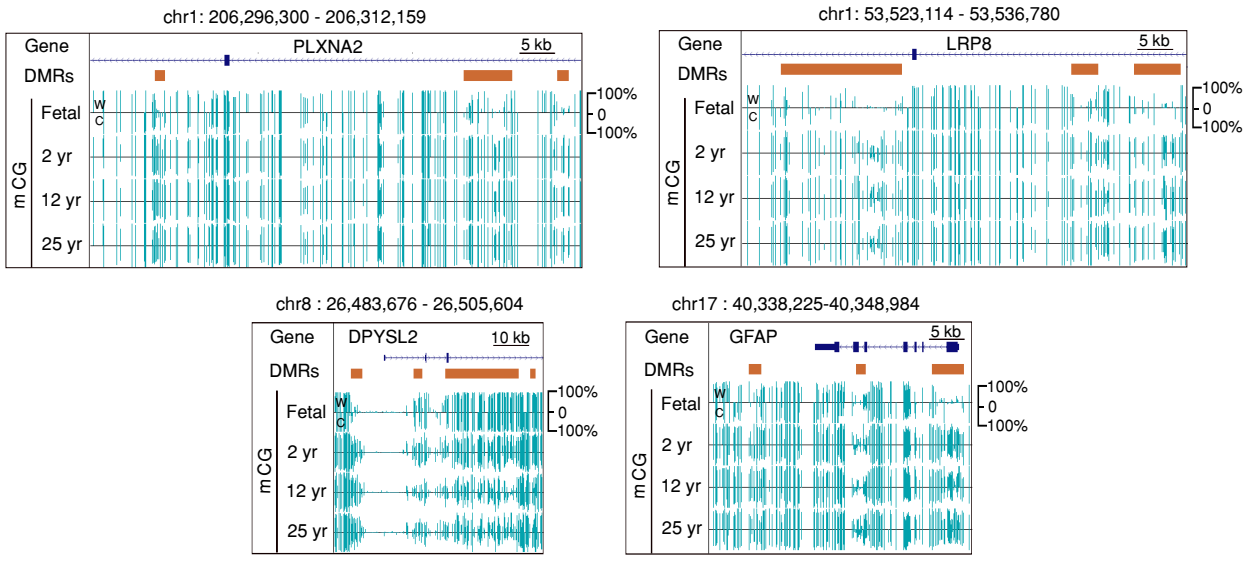

Figure 7. Browser representations of CG-DMRs predicted by HOME in time-series WGBS data with four time points: fetal, 2, 12 and 25 year old mammalian brain.

regions, such as genes related to neuronal development (Figure 7). This added functionality of HOME to predict DMRs in time-series data provides an easy and convenient way to compare many time points. Moreover, the DMRs can be traced through time to determine their stability or stochasticity.

Discussion

Here we present a novel histogram of methylation based method to detect DMRs from single nucleotide resolution DNA methylation data. The method can be used to accurately identify DMRs with precise boundaries for comparison of treatment groups with single or multiple replicates. The key features of HOME are: (i) novel histogram based features which combines important information such as methylation level difference, measure of significance for the difference in methylation, and distance between neighboring cytosines; (ii) robustly trained model that is effective for a wide variety of species; (iii) a flexible method that can be used for prediction of DMRs in both CG and CH contexts with high border accuracy; and (iv) a tool that can identify DMRs in time-series data.

The most important qualities of any DMR finder are accurate prediction of DMR boundaries and low number of spurious DMRs (false positives). HOME outperforms both DSS and Metilene in both of these measures. One of the reasons underlying the low false positive rate of HOME is the use of biological training data for DMRs and non-DMRs to train the classifier. In addition, the histogram based features can robustly discriminate between DMRs and non-DMRs, thereby reducing the probability of detecting spurious DMRs. Histogram based features are also able to capture the information present around each cytosine site with the use of weighted voting, thereby, enabling accurate location of the DMR boundaries.

HOME accounts for biological variation present between the replicates and uneven read coverage through weighted logistic regression while computing the p-value. The spatial correlation present among neighboring cytosine sites is captured by moving average smoothing

and the use of weighted voting for histogram based features. We demonstrate that HOME can be used to predict accurate DMRs in both CG and CH sequence contexts for both mammalian and plant WGBS methylome data by using the same training data. Although the classifier was trained on mammalian WGBS data for CG and CH contexts, HOME can accurately predict DMRs in other species and for specific non-CG contexts (CHG and CHH), demonstrating its versatility.

Finally, another standout feature of HOME is the prediction of DMRs in time-series data. Time-series DNA methylation experiments are commonly used to study a wide range of biological processes such as development (Lister et al. 2013) and stress responses (Dowen et al. 2012). HOME provides an efficient algorithm to directly predict accurate DMRs in studies with multiple timepoints. This added functionality of HOME will greatly facilitate and expand the study of epigenome dynamics in numerous biological systems and disease models. Taken together, HOME is a highly effective and robust DMR finder that accounts for uneven cytosine coverage in WGBS data, accounts for biological variation present between the samples in the same treatment group, predicts DMRs in various genomic contexts, and accurately identifies DMRs among any number of treatment groups in experiments with or without replicates.

Methods

Algorithm availability and use

HOME is written in python. Documentation and source code are freely available in the form of a user friendly package at https://github.com/ListerLab/HOME. All the analysis performed in this paper utilized the default parameters for HOME (version 0.1). The default parameters for HOME are set to be relatively permissive. The user can fine tune the frequently used parameters such as merge distance (–mergedist), minimum length of reported DMRs (–minlength), minimum average methylation difference (–delta) and minimum number of Cs (–minc) in a DMR.

Data acquisition

WGBS data for EX, PV and VIP neuronal cell types were downloaded from NCBI GEO (GSE63137). Arabidopsis root WGBS data were obtained from NCBI GEO repositories GSM2101445 and GSM2101440, respectively. WGBS data from fetal, 2 year, 12 year and 25 year old human brain were obtained from GEO accessions GSM1163695, GSM1167005, GSM1164630 and GSM1164632, respectively.

Training data generation

To generate the training data for the CG context, publicly available methylome data from neuron (GSM1173786) and non-neuron (GSM1173787) cell types isolated from the frontal cortex of 7 weeks old male mouse brain (Lister et al. 2013) was used, considering chromosome 2 only. For the CH context, the published methylation data for chromosome 2 from two neuronal cell types, excitatory pyramidal neurons (EX) with two biological replicates and vasoactive intestinal peptide-expressing interneurons (VIP) with two biological replicates (Mo et al. 2015) were used. The raw data were processed as described previously (Lister et al. 2013) to obtain the number of methylated reads (mc) and total reads (t) for each cytosine. Watson and Crick strand counts were combined for mc and t for CG sites, as the methylation is usually symmetric on both strands (Laurent et al. 2010; Hardcastle 2013; Guo et al. 2014). The strand combination was not performed for CH sites because methylation of the CH sites is strand-specific (Lister et al. 2009; Guo et al. 2014). Thereafter, three existing DMR finders (DSS, bisulfighter and Metilene) were used with default parameters to produce preliminary DMRs. DMRs identified by two or more finders were further analyzed for differences in methylation level, and DMRs with an average methylation level difference >0.3 were visually confirmed and selected as the training set of DMRs. When necessary, boundaries were adjusted by trimming until three consecutive cytosines had the value of the b (Eq.1) > 0.1, to assure accurate boundary cutoff.

The regions not identified as DMRs by all of the DMR finders were randomly selected as potential non-DMRs. Among these, the regions with small average methylation level differences (<0.2) were further analyzed for methylation level differences and visually inspected for consistent methylation level difference and were selected as robust to confirm as non-DMRs.

Histogram computation from training data

HOME uses novel histogram based features for identification of DMRs. The algorithm starts by combining the Watson and Crick strand counts for mc and t for the CG context. For the CH context, no strand combination is performed. Next, the algorithm computes the methylation level difference between the two samples and estimates the p-value for the difference at each cytosine. For CH training data, which has biological replicates within each treatment group, p-values are computed using weighted logistic regression to model methylation levels in relation to the treatment groups and variation between replicates. Since the CG training data has only one sample per treatment group, a z-score test was used for p-value computation. The null hypothesis for modeling p-values for both tests described above is that methylation levels are the same among treatment groups for a given cytosine. The alternative hypothesis is that there is a difference in the methylation levels among the treatment groups. To account for uneven read coverage, a logistic function was used to compute the weight for weighted logistic regression from t, such that the range of weight is between 0 and 1 when calculating the p-value. Thereafter, the absolute difference in methylation level at each cytosine is weighted by its p-value (p) to compute a bin value (b) as shown in Equation 1 below.

$$b = |m_1 - m_2| \cdot e^{(1-p)} \tag{1}$$

where, m_1 and m_2 are the methylation levels of treatment groups under comparison and exponentiation of the $(1-p)$ allows smaller p-values to contribute more to the produced bin value than larger (insignificant) p-values. To account for the spatial correlation between the

neighboring cytosines, moving average smoothing (default: 3 cytosines) is performed on values of b to compute final bin value b_s. The histogram feature is computed for every individual cytosine present in each DMR and non-DMR training data. For a given cytosine, to compute the histogram feature, a fixed window of size w centered around it is used where w is the number of cytosines in a window (w is set to 11 for CG and 51 for CH context). For each window, the bin values b_s are binned using a weighted voting approach such that for a given cytosine, its contribution v to the bin is computed as a weighted distance from the center cytosine which is normalized by the maximum allowed distance as shown in Equation 2 below.

$$v = \begin{cases} 1 - \dfrac{|l - l_c|}{d} & \text{if } |l - l_c| < d, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

where, l is the location of the cytosine being binned, l_c is the location of the center cytosine of w, and d (default: 250bp) is the normalization constant signifying the maximum allowed distance from the center cytosine. Consequently, the cytosines close to the center cytosine will have higher weights and will contribute more to the histogram feature. On the other hand, if the distance between the cytosine being binned and the center cytosine of the window is larger than d, then that cytosine will have zero contribution.

Next, for a given cytosine, a histogram feature is computed by using b_s and v for each cytosine in the window. More specifically, b_s defines the bin of the histogram in which the contribution will be placed and v defines the value of that contribution. Subsequently, the histogram feature vector is normalized such that the feature vector sums to unity.

Training via SVM

The algorithm then uses the normalized histogram feature vectors described above to train a support vector machine (SVM) that predicts the label (DMR or non-DMR) for each individual cytosine. The scikit-learn implementation of SVM (Pedregosa et al. 2011) with linear kernel (with default parameters) was used for training because it is computationally very efficient

and shows comparable performance to the more computationally expensive non-linear RBF kernel (Pedregosa et al. 2011) and random forest classifier (Breiman 2001; Karpievitch et al. 2009; Pedregosa et al. 2011). The F1-scores for linear, RBF and random forest after 5-fold cross-validation on the training dataset were 0.860, 0.864 and 0.860, respectively.

Testing and DMR prediction on new datasets

HOME requires input files containing basic information of methylation, including chromosome numbers, genomic coordinates, type of cytosine (CG, CHG, CHH), and mc and t for cytosines.

1. Pairwise

HOME can be used to predict DMRs from methylomes of two treatment groups with single or multiple replicates. To predict the DMRs, the normalized histogram features are computed for each cytosine on a particular chromosome, which are then provided to the trained SVM model to obtain the prediction scores that are normalized using the logistic function from the generalized linear model (GLM) to lie in the range [0,1]. Individual cytosines are grouped together into preliminary DMRs based on the prediction scores (default: >0.1) and the distance between neighboring cytosines (default: <500bp). To produce the final DMRs, the algorithm performs a boundary refinement of the preliminary DMRs such that boundaries are trimmed until k consecutive cytosines (default: 3) have the value of the b (Eq.1) greater than or equal to the defined threshold (default: 0.1).

2. Time-series and multi-group comparisons

HOME can be used to predict DMRs from time-series and multi-group studies. Given a number of time points or treatments, n, a total of ${}^{n}C_{2}$ pairwise combinations are possible. HOME computes SVM prediction scores for each of these pairwise combinations in the same manner as for pairwise method described above. The prediction scores are then normalized to lie in

the range [0,1] using the logistic function from the generalized linear model (GLM) to allow further analysis among all pairwise comparisons. The scores are summed for each cytosine to get a final score. The cytosines are then grouped into DMRs in the same manner as for pairwise method described above.

Boundary accuracy measurement

To measure the boundary accuracy, the fraction of CG overlap between simulated and predicted DMRs was calculated using BEDtools intersect (Quinlan and Hall 2010) with fraction of overlap (-f) and reciprocal overlap (-r) features. As defined by BEDtools, -r feature require that the fraction of overlap be reciprocal for A and B. That is, if -f is 0.90 and -r is used, this requires that B overlap at least 90% of A and that A also overlaps at least 90% of B. In our case A was simulated DMR and B was predicted DMR.

Acknowledgments

We thank Dr. Egor Dolzhenko for his helpful discussions regarding the generation of simulated data. We thank members of Lister Lab and Borevitz Lab for their suggestions and comments. This work was supported by the Australian Research Council (ARC) Centre of Excellence program in Plant Energy Biology (CE140100008). RL was supported by a Sylvia and Charles Viertel Senior Medical Research Fellowship, ARC Future Fellowship (FT120100862), and Howard Hughes Medical Institute International Research Scholarship (RL).

Author Contributions

AS and RL devised the project. AS conceptualized, designed and implemented HOME with statistical guidance from YVK. RL, SRE and JOB supervised experiments. AS and YVK processed the biological data. AS tested HOME on simulated and biological data. AS and YVK

drafted the manuscript, all authors contributed to writing the manuscript.

Disclosure Declaration

The authors declare no conflicts of interest.

References

Bogdanovic O, Smits AH, de la Calle Mustienes E, Tena JJ, Ford E, Williams R, Senanayake U, Schultz MD, Hontelez S, van Kruijsbergen I et al. 2016. Active DNA demethylation at enhancers during the vertebrate phylotypic period. Nature genetics 48(4): 417-426.

Breiman L. 2001. Random Forests. Mach Learn 45(1): 5-32.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452(7184): 215-219.

Cortes C, Vapnik V, 1995. Support-vector networks. Machine Learning 20(3): 273.

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

Dowen RH, Pelizzola M, Schmitz RJ, Lister R, Dowen JM, Nery JR, Dixon JE, Ecker JR. 2012. Widespread dynamic DNA methylation in response to biotic stress. Proceedings of the National Academy of Sciences of the United States of America 109(32): E2183-2191.

Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA et al. 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. Nature genetics 38(12): 1378-1385.

Eichten SR, Stuart T, Srivastava A, Lister R, Borevitz JO. 2016. DNA methylation profiles of diverse Brachypodium distachyon align with underlying genetic diversity. Genome research 26(11): 1520-1531.

Feng H, Conneely KN, Wu H. 2014. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic acids research 42(8): e69.

Guo S, Diep D, Plongthongkum N, Fung HL, Zhang K, Zhang K. 2017. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nature genetics 49(4): 635-642.

Guo W, Chung WY, Qian M, Pellegrini M, Zhang MQ. 2014. Characterizing the strand-specific distribution of non-CpG methylation in human pluripotent cells. Nucleic acids research 42(5): 3009-3016.

Hansen KD, Langmead B, Irizarry RA. 2012. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome biology 13(10): R83.

Hardcastle TJ. 2013. High-throughput sequencing of cytosine methylation in plant DNA. Plant methods 9(1): 16.

Heyn H, Moran S, Hernando-Herraez I, Sayols S, Gomez A, Sandoval J, Monk D, Hata K, Marques-Bonet T, Wang L et al. 2013. DNA methylation contributes to natural human variation. Genome research 23(9): 1363-1372.

Jaffe AE, Feinberg AP, Irizarry RA, Leek JT. 2012. Significance analysis and statistical dissection of variably methylated regions. Biostatistics 13(1): 166-178.

Jones PA. 1999. The DNA methylation paradox. Trends in genetics : TIG 15(1): 34-37.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature reviews Genetics 13(7): 484-492.

Juhling F, Kretzmer H, Bernhart SH, Otto C, Stadler PF, Hoffmann S. 2016. metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. Genome research 26(2): 256-262.

Karpievitch YV, Hill EG, Leclerc AP, Dabney AR, Almeida JS. 2009. An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++. PloS one 4(9): e7087.

Kass SU, Landsberger N, Wolffe AP. 1997. DNA methylation directs a time-dependent repression of transcription initiation. Current biology : CB 7(3): 157-165.

Kawakatsu T, Stuart T, Valdes M, Breakfield N, Schmitz RJ, Nery JR, Urich MA, Han X, Lister R, Benfey PN et al. 2016. Unique cell-type-specific patterns of DNA methylation in the root meristem. Nature plants 2(5): 16058.

Khavari DA, Sen GL, Rinn JL. 2010. DNA methylation and epigenetic control of cellular differentiation. Cell cycle 9(19): 3880-3883.

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

Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature reviews Genetics 11(3): 204-220.

Lea AJ, Tung J, Zhou X. 2015. A Flexible, Efficient Binomial Mixed Model for Identifying Differential DNA Methylation in Bisulfite Sequencing Data. PLoS genetics 11(11): e1005650.

Lee DS, Shin JY, Tonge PD, Puri MC, Lee S, Park H, Lee WC, Hussein SM, Bleazard T, Yun JY et al. 2014. An epigenomic roadmap to induced pluripotency reveals DNA methylation as a reprogramming modulator. Nature communications 5: 5619.

Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD et al. 2013. Global epigenomic reconfiguration during mammalian brain development. Science 341(6146): 1237905.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462(7271): 315-322.

Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, Antosiewicz-Bourget J, O'Malley R, Castanon R, Klugman S et al. 2011. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature 471(7336): 68-73.

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB et al. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454(7205): 766-770.

Messerschmidt DM, Knowles BB, Solter D. 2014. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. Genes & development 28(8): 812-828.

Mo A, Mukamel EA, Davis FP, Luo C, Henry GL, Picard S, Urich MA, Nery JR, Sejnowski TJ, Lister R et al. 2015. Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. Neuron 86(6): 1369-1384.

Narsai R, Secco D, Schultz MD, Ecker JR, Lister R, Whelan J. 2017. Dynamic and rapid changes in the transcriptome and epigenome during germination and in developing rice (Oryza sativa) coleoptiles under anoxia and re-oxygenation. The Plant journal : for cell and molecular biology 89(4): 805-824.

Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, Li Q, Rohr NA, Rambani A, Burke JM, Udall JA et al. 2016. Widespread natural variation of DNA methylation within angiosperms. Genome biology 17(1): 194.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al. 2011. Scikit-learn: Machine Learning in Python. J Mach Learn Res 12: 2825-2830.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6): 841-842.

Richardson BC. 2002. Role of DNA methylation in the regulation of cell function: autoimmunity, aging and cancer. The Journal of nutrition 132(8 Suppl): 2401S-2405S.

Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J et al. 2015. Integrative analysis of 111 reference human epigenomes. Nature 518(7539): 317-330.

Saito Y, Tsuji J, Mituyama T. 2014. Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions. Nucleic acids research 42(6): e45.

Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H et al. 2015. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature 523(7559): 212-216.

Shafi A, Mitrea C, Nguyen T, Draghici S. 2017. A survey of the approaches for identifying differential methylation using bisulfite sequencing data. Briefings in bioinformatics.

Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D et al. 2011. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480(7378): 490-495.

Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, Cross MK, Williams BA, Stamatoyannopoulos JA, Crawford GE et al. 2013. Dynamic DNA methylation across diverse human cell lines and tissues. Genome research 23(3): 555-567.

Wang Z, Li X, Jiang Y, Shao Q, Liu Q, Chen B, Huang D. 2015. swDMR: A Sliding Window Approach to Identify Differentially Methylated Regions Based on Whole Genome Bisulfite Sequencing. PloS one 10(7): e0132866.

Wen Y, Chen F, Zhang Q, Zhuang Y, Li Z. 2016. Detection of differentially methylated regions in whole genome bisulfite sequencing data using local Getis-Ord statistics. Bioinformatics 32(22): 3396-3404.

Wu H, Xu T, Feng H, Chen L, Li B, Yao B, Qin Z, Jin P, Conneely KN. 2015. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. Nucleic acids research 43(21): e141.

Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, Dempster EL, Ren B. 2012. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. Cell 148(4): 816-831.