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.

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

Abstract

The development of whole genome bisulfite sequencing has made it possible to identify methylation differences at single base resolution throughout an entire genome. However, a persistent challenge in DNA methylome analysis is the accurate identification of differentially methylated regions (DMRs) between samples. Here, we present a novel Histogram Of MEthylation (HOME) based method that takes into account the inherent difference in distribution of methylation levels between DMRs and non-DMRs to robustly discriminate between the two using a Support Vector Machine. We demonstrate that HOME produces more accurate DMRs than the current state-of-the-art methods on both simulated and biological datasets.

Keywords

WGBS, methylation, epigenetics, DMR identification, SVM

Background

DNA methylation plays an important role in the regulation of various cell functions including genomic imprinting, X-chromosome inactivation and cellular differentiation [1-3]. However, analysis of DNA methylation presents various challenges as the modification is highly dynamic in space and time [4, 5]. DNA methylation levels vary between distinct genomic features such as promoters, enhancers, gene bodies, transposable elements, and repeat elements [6-12]. Furthermore, widespread variation in the distribution of DNA methylation has been observed between different cell types, cell lines, tissues, individuals and species [13-18]. 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

34 studies have uncovered the presence of non-CG (CH, where H=A, T, or C) methylation in certain
35 cell types including embryonic stem cells and brain cells [5, 9, 19, 20]. In contrast, DNA
36 methylation in plants occurs in all sequence context, namely CG, CHG, and CHH [10].
37 Furthermore, CH methylation is often found at much lower levels compared to CG methylation,
38 as measured by the proportion of reads displaying methylation, making the accurate analysis of
39 CH DNA methylation more challenging given the typical sequencing depth of experiments to
40 date.

41 High-throughput sequencing methods such as whole genome bisulfite sequencing
42 (WGBS) have been developed to provide detection and quantitative measurement of DNA
43 methylation at single base resolution throughout whole genomes [9, 21]. Sodium bisulfite
44 treatment of genomic DNA converts cytosines, but not methylcytosines, into uracils, and during
45 subsequent PCR amplification of the bisulfite treated DNA the uracils are replaced by thymines.
46 High-throughput sequencing of bisulfite converted DNA and alignment to a reference genome
47 enables the methylation level of any covered cytosine to be computed by counting the number of
48 methylated and unmethylated bases in reads that cover that cytosine position. Sensitive and
49 accurate DMR detection from such data is important in characterization of the differences and
50 dynamics of DNA methylation state, exploration of potential roles in genome regulation, and as
51 disease biomarkers [22]. However, accurate DMR detection remains a significant challenge.
52 Most of the existing DMR identification methods such as bsseq [23], RADMeth [24], MACAU
53 [25] and BiSeq [26] are more appropriate to identify DMRs when two or more replicates are
54 available for each of the treatment groups [27]. Other methods such as Comet [28] and swDMR
55 [29] have been developed to identify DMRs for single replicate treatment groups. Two of the
56 recently developed methods, DSS and DSS-single [30, 31] (referred to as DSS hereafter), and

57 Metilene [32] can be used for single or multiple replicate treatment groups and have been shown
58 to outperform the aforementioned methods. However, both of these methods are limited to DMR
59 identification between two treatment groups and cannot be directly used for more complex
60 experimental designs with multiple groups and/or time points. Moreover, multiple characteristics
61 need to be considered for accurate prediction of DMRs, including spatial correlation present
62 between neighboring cytosine sites, sequencing depth that takes into account sampling variability
63 that occurs during sequencing, and biological variation among replicates of treatment groups [23,
64 27, 33, 34]. Most of the DMR identification tools described above do not consider either all or
65 some of the characteristics required for accurate prediction of DMRs.

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

Results and discussion

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 [35]. Successful employment of supervised or semi-supervised learning algorithm requires access to a high quality training dataset. Due to the lack of a biological dataset with known DMRs and non-DMRs, we generated a training dataset using publicly available DNA methylome datasets. To generate a robust and accurate training data, we followed a coarse-to-fine approach, where we first used consensus regions identified by existing DMR finders, for predicted DMRs, to produce potential DMR and non-DMR training sets. Then, the training sets were further fine-tuned by visual inspection for accurate boundaries and methylation level change. A similar approach was used to prepare training and testing datasets in ChIP-seq peak identification, where the authors identified peaks with one or more existing peak callers and further visually confirmed the peaks [36, 37]. Moreover, similar approaches for training data curation are regularly used in the field of computer vision and machine learning, for example in natural language processing [38], video annotation [38, 39], and face and facial feature annotation [40].

We trained the classifier separately for the CG and CH methylation contexts, as they exhibit different methylation characteristics. 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 [5] was used. For the CH context, the classifier was trained on methylation data from two neuronal cell types: parvalbumin-expressing fast-spiking interneurons (PV) and vasoactive intestinal peptide-expressing interneurons (VIP) [41]. To generate the training data, the DMR

finders DSS, bisulfighter and Metilene were used, with default parameters, on the processed methylome data. The DMRs identified by all algorithms were further analyzed visually on the UCSC genome browser to confirm the presence of contiguous methylation level differences. The regions not identified as DMRs by all of the DMR finders were visually confirmed and selected as true 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-test for non-replicated data (Methods section). The algorithm accounts for biological variation present between the replicates and uneven read coverage through weighted logistic regression while computing the p-value. The important 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 (eq.1 in the Methods section). 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. We tested different bin sizes of 5, 10 and 20, and the F1-scores for 5-fold cross-validation on the training set DMRs and non-DMRs is summarized in Supp. Table 1. As there was marginal improvement in the F1-score between bin size 5 and 10 and no improvement in F1-score from bin size 10 to 20, we decided to use a bin size of 10.

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

124 approach such that for a given cytosine, its contribution to the bin is computed as a weighted
125 distance from the center cytosine, which is normalized by the maximum allowed distance (eq. 2
126 in Methods Section). The use of weighted voting for computation of histogram based features
127 also captures the spatial correlation present among neighboring cytosine sites.

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

138 Once the SVM has been trained, the histogram based features for new methylomes can be
139 computed, and HOME scans the entire methylome to provide a prediction score (between 0 and
140 1) from the SVM classifier for each cytosine site. Then, the individual cytosine sites are grouped
141 together into DMRs based on the user defined prediction score cutoff (default: >0.1) and the
142 distance between neighboring cytosines (default: <500bp). A low prediction score (<0.1) from
143 the classifier for a cytosine site indicates low confidence in the site being differentially
144 methylated and a high prediction score (>0.1) indicates high confidence in a site being
145 methylated. Here, we independently applied HOME to both CG and CH contexts, and compared
146 its performance to two other commonly used DMR finders, Metilene and DSS, using both

147 simulated and biological data. Furthermore, we also show that HOME can be used for time-
148 series DMR analysis on biological data.

149 **Analysis of simulated DNA methylation data**

150 The DMRs were simulated with the assumption that the methylation level of cytosine sites
151 follows a beta binomial distribution. DMRs and non-DMRs were simulated with three distinct
152 beta binomial settings, each increasing in their difficulty for identification. For each setting two
153 treatment groups were simulated, each with three replicates. The read coverage for replicates was
154 taken from WGBS datasets of neuronal and non-neuronal cell types [5]. For all settings, the
155 methylation level of cytosine sites in DMRs were generated from a beta distribution of (6, 1.5)
156 and a beta distribution of (1.5, 6) for the two treatment groups, respectively. For the first setting
157 (class 1), the non-DMR portion of the genome displayed a fixed methylation level (0.7), with
158 only DMRs showing variation from this value. In the second setting (class 2), non-DMR cytosine
159 methylation level was simulated from beta parameters (2, 2), such that the methylation level was
160 not fixed for either DMRs or non-DMRs. The final setting (class 3) was similar to class 2,
161 however, use of the non-DMR beta parameters of (1.5, 1.5) provided an overall broader
162 distribution under which DMR methylation states can exist by random chance. Therefore, in
163 these simulated datasets it is expected that correct DMR classification will be easiest for class 1
164 and hardest for class 3. The simulation method described above is similar to approach taken
165 previously [24]. For clarity, a visual representation of these settings in the UCSC genome
166 browser is shown in Figure 2A.

167 We performed an extensive parameter search to identify settings for each DMR finder
168 that resulted in the best performance (Supplementary Information: Section 1.1). Simulated
169 DMRs are predicted by all the DMR finders for Class 1, however, HOME is more accurate in

170 predicting the boundaries, compared to DSS and Metilene. Predicted DMRs by DSS and
171 Metilene, for class 2 and 3 DMRs, are either missed or have inaccurate boundaries, whereas,
172 HOME identifies all DMRs with highly accurate boundaries (Figure 2A).

173 Precise definition of DMR boundaries is essential for accurate downstream analysis and
174 biological interpretation of differential methylation, for example when associating DMRs to
175 regulatory regions such as promoters or enhancers to explore potential connections between
176 methylation changes and local chromatin state and transcription. Imprecise DMR boundary
177 identification will result in the inappropriate inclusion in the DMR of cytosine sites that are not
178 differentially methylated, or exclusion of bona fide differentially methylated cytosines.
179 Consequently, mean methylation levels calculated for all cytosines within a DMR would be
180 inaccurate, and analysis of genomic or chromatin features at the DMR boundaries would be
181 imprecise. We evaluated the DMR boundary accuracy for all three DMR finders by calculating
182 the fraction of overlap between simulated and predicted DMRs (Methods Section). The accuracy
183 was tested for reciprocal overlap ranging from 50 to 100% between simulated and predicted
184 DMRs. HOME produced more accurate DMR boundaries than DSS and Metilene (Figure 2B).
185 For example, in the range of 50-80% reciprocal overlap between simulated and predicted DMRs,
186 HOME correctly predicted 99% of the simulated DMRs. The fraction of correctly predicted
187 DMRs by HOME decreased to 98% for the range of 80-90% reciprocal overlap, and 90% for the
188 range of 90-100% reciprocal overlap. This means that the 90% of predicted DMRs perfectly
189 matched the boundary of the simulated DMRs. In contrast, for DSS and Metilene the fraction of
190 predicted DMRs decreased with an increasing requirement in the percentage overlap between the
191 predicted and simulated DMRs. This indicates that HOME is highly accurate in predicting the
192 boundaries of the DMRs, outperforming both DSS and Metilene for all three of the simulated

193 DMR classes (Figure 2B).

194 For the comprehensive evaluation of the accuracy of the DMR finders, we measured the
195 performance with respect to CG sites. True positive (TP) was defined when a differentially
196 methylated CG site in the predicted DMR overlapped with a differentially methylated CG site in
197 the simulated DMR, otherwise, it was considered as false positive (FP). False negative (FN) was
198 defined when a differentially methylated CG site in the simulated DMRs was not identified by
199 the DMR identification algorithm. The performance of the DMR finders was measured in terms
200 of precision ($\frac{TP}{TP+FP}$), recall ($\frac{TP}{TP+FN}$), F1-score ($2\ x\ \frac{precision\ x\ recall}{precision+recall}$) and Jaccard Index
201 ($\frac{TP}{TP+FP+FN}$). HOME outperformed both DSS and Metilene in all metrics, for all classes with
202 distinct biological settings (Figure 2C).

203 **Analysis of performance on biological datasets**

204 **1. Pairwise differential methylation analysis**

205 We compared the performance of HOME with DSS and Metilene for the CG context on
206 published WGBS data [41] for two neuronal cell types: excitatory pyramidal neurons (EX) and
207 parvalbumin-expressing fast-spiking interneurons (PV), each with two replicates. The runtime
208 for DSS and HOME was very similar (~2 hours), while Metilene completed in ~4 min. Among
209 the DMR finders that were compared, DMRs identified by HOME were less fragmented and
210 exhibited more accurate boundary detection compared to the DMRs produced by the other
211 finders (Figure 3A), despite equivalent DMR merging parameters being used for each finder
212 (Supplementary Information: Section 1.2). DMRs predicted by HOME showed consistently
213 higher methylation level differences inside the DMR boundaries when compared to the other
214 finders (Figure 3B). For closer inspection of the above observation, the mean and standard

215 deviation of the absolute methylation level difference between analyzed samples for the 5 CG
216 sites immediately inside and outside the DMR boundary is shown (Figure 3C). The mean
217 methylation level difference for the boundary CG site and CG sites inside the DMRs is higher for
218 HOME than DSS and Metilene (Figure 3C), while the mean methylation level difference for the
219 CG sites immediately outside the DMR boundaries is lower for HOME as compared to DSS and
220 Metilene. This demonstrates the higher DMR border detection precision of HOME on real
221 biological datasets, as also observed for the simulated WGBS data (Figure 2B).

222 A more detailed comparison of the DMRs uniquely predicted by each finder showed that
223 DMRs uniquely identified by HOME consistently had a higher methylation level difference for
224 the CG sites located within the DMRs (Figure 3D). In contrast, DMRs uniquely predicted by
225 DSS and Metilene consistently had a lower methylation difference for the CG sites within the
226 DMRs. Furthermore, the boundaries of the DMRs identified by HOME are more precise
227 compared to DSS and Metilene (Figure 3D). To investigate the biological significance of the
228 uniquely predicted DMRs, we tested whether the uniquely predicted DMRs identified by each
229 finder are located in the genomic regions that are relevant to neuronal development and function.
230 For this analysis, phenotype and gene expression annotations provided by the Genomic Regions
231 Enrichment of Annotations Tool (GREAT) were used [42]. Briefly, among the top 25 terms
232 produced by GREAT for gene expression and phenotype, we counted the enrichment terms
233 related to neuronal development and function for uniquely predicted DMRs by each finder. The
234 significance of enrichment terms was ranked according to the binomial distribution-based P-
235 values obtained from GREAT. A similar approach for exploring biological functions of DMRs
236 has been performed previously [43]. The parameter details used for the analysis are provided in
237 Supp. Table 2. Among the top 25 terms for phenotype annotation, 92% of terms for DMRs

238 uniquely predicted by HOME were directly related to neural system functions. In contrast, terms
239 related to neural systems for DMRs uniquely predicted by DSS and Metilene were 56% and 28%
240 respectively. For associated gene expression annotations provided by GREAT, we found that
241 uniquely predicted HOME DMRs were located on or near to genes related to neuronal
242 development and function. Among the top 25 terms for gene expression annotation, 76% of
243 terms for unique HOME DMRs were directly related to neural system function. Terms related to
244 neural systems for unique DMRs by DSS and Metilene DMRs were 20% and 32%, respectively
245 (Figure 3E). The details of phenotype and gene expression annotations are summarized in Supp.
246 Figure 1.

247 Chromatin accessibility has been shown to have strong correlation with DNA
248 methylation, where ATAC-seq peaks are usually associated with DNA hypomethylated regions
249 [41]. Thus, we hypothesised that true DMRs are likely to overlap with differential ATAC-seq
250 peaks detected between EX and PV (n=151,853 differential ATAC-seq peaks, see Methods).
251 Compared with the other analyzed DMR finders, a higher number of differential ATAC-seq
252 peaks overlapping with HOME DMRs were observed compared to DSS and Metilene DMRs
253 (Figure 3E). These results indicate that the DMRs identified by HOME exhibit a higher
254 association with biologically relevant genes, processes, and regulatory events than those
255 identified by the other DMR finders.

256 To investigate the incidence of false positive DMRs predicted by each finder, we
257 permuted the labels among the EX and PV WGBS samples to generate two artificial datasets: (1)
258 EX replicate 1 and PV replicate 1 (comprising treatment group 1) versus EX replicate 2 and PV
259 replicate 2 (comprising treatment group 2), and (2) EX replicate 1 and PV replicate 2 versus EX
260 replicate 2 and PV replicate 1 (comprising treatment groups 1 and 2 respectively). Due to

261 randomness in the shuffled data, it is expected that there will be shorter regions with contiguous
262 methylation level differences occurring by chance, and these short regions will be identified by
263 the algorithms as DMRs. However, it is also expected that there will be significantly fewer long
264 DMRs in the shuffled data, because as the length of the DMRs increases the likelihood of
265 obtaining such DMRs due to random chance decreases rapidly. These shuffled datasets were
266 analyzed with HOME, DSS and Metilene, and as expected the number of DMRs predicted by all
267 algorithms in comparison to the unshuffled data was significantly reduced (Supp. Figure 2). In
268 addition, DMRs identified by HOME in the shuffled data were smaller and contained low
269 number of CGs. In contrast, the DMRs identified by DSS and Metilene were longer and had a
270 higher number of CGs per DMR and therefore are more likely to be false positive DMRs (Supp.
271 Figure 2).

272 To test the performance of HOME for detecting differential methylation in the CH
273 sequence context, we used WGBS datasets of neuronal and non-neuronal cell types isolated from
274 the frontal cortex of 7 week old male mouse prefrontal cortex [5]. Because of a high
275 computational runtime requirement for both DSS and Metilene for CH context analysis, the run
276 did not complete after 12 days of execution and had to be terminated, being deemed an
277 unfeasible analysis to undertake with these algorithms given reasonable timeframes for analysis.
278 Therefore, the results are only shown for HOME (runtime: 4 days). A genome browser view of a
279 representative genomic region showed that the CH DMRs predicted by HOME between neurons
280 (NeuN+) and glia (NeuN-) were regions of contiguous hyper- or hypo-methylation (Figure 4A).
281 The directionality of the DMRs (hyper/hypo) are defined with respect to NeuN+ cells. We
282 observed a high number of hyper-methylated DMRs (516,532) in NeuN+ cells as compared to a
283 very low number of hypo-methylated DMRs (12,732) (Figure 4B). These findings are consistent

284 with previously published results, where CH methylation accumulates to a high level in neurons
285 compared to glia [5, 41]. CH DMRs predicted by HOME have accurate boundaries with higher
286 inter-sample methylation level difference at and within DMR boundaries, and low methylation
287 level difference immediately outside the DMR boundaries (Figure 4B). Although the observed
288 methylation difference in hypo-methylated NeuN+ DMRs was very low (<0.02), gene ontology
289 analysis using GREAT showed that these DMRs were located in genomic regions related to
290 neuronal development and function, suggesting that the HOME CH DMRs are biologically
291 relevant.

292 To assess the generalizability of HOME performance for species that have very distinct
293 methylation patterns and distributions compared to mammals, we next tested the performance of
294 HOME on Arabidopsis WGBS data, comparing DNA methylation in wild-type (WT) and
295 *CHROMOMETHYLASE 2* mutant (*cmt2)*, a well characterised mutant that exhibits differences in
296 CH methylation [44]. CMT2 is a functional non-CG methyltransferase, known to mediate DNA
297 methylation at both CHG and CHH context *in vitro* and *in vivo* [44, 45]. We compared the
298 performance of HOME with DSS, which has previously been used for DMR prediction in plant
299 WGBS datasets [16, 46]. Note that we used the same model that was trained on the mammalian
300 WGBS dataset for predicting the DMRs between *cmt2* and WT, to assess whether the trained
301 model is generalizable. HOME took 12 min to predict CG DMRs compared to 27 min for DSS.
302 For CHG context DMRs, both HOME and DSS showed similar execution time of 27 min, while
303 HOME was >3 times faster compared to DSS for DMR prediction in the CHH sequence context,
304 taking 2 and 7 hours, respectively.

305 The heatmap for all predicted DMRs by HOME and DSS showed that HOME DMRs are
306 more accurate in boundary prediction than DSS, predominantly for the CG and CHG contexts

307 (Figure 5A). Moreover, for the CHG context, HOME detected a large number of DMRs
308 (n=12,558) of a large median length DMRs (680 bp), whereas DSS only predicted small number
309 of DMRs (n=3,038) with a short median length (133 bp), indicating that HOME is more sensitive
310 for DMR detection over a greater range of sizes. For the CHH context, the mean and standard
311 deviation of 5 cytosines upstream and downstream of the predicted DMRs start and stop sites,
312 respectively, showed that the methylation level difference is high within the DMR and low just
313 outside the DMRs for HOME (Figure 5B). In contrast, DMRs predicted by DSS showed similar
314 mean and standard deviation of methylation level difference both inside and outside the DMR
315 boundaries (Figure 5B). These results indicate more accurate boundary prediction by HOME for
316 all contexts. Similarly, genome browser screenshots exemplify how HOME DMRs are more
317 precise than DSS DMRs for all contexts (Supp. Figure 3).

318 Furthermore, the heatmaps of DMRs predicted uniquely by HOME exhibit more accurate
319 boundaries than DSS for all sequence contexts (Figure 6A). We further examined the genomic
320 distribution of the DMRs uniquely predicted by HOME and DSS. Over 80% of HOME CG
321 DMRs were present in gene bodies, while only a small fraction (<20%) of CHG and CHH DMRs
322 overlapped with gene bodies (Figure 6B). These results are consistent with the distribution of
323 DNA methylation in plant genes, where gene bodies mainly exhibit CG methylation [21, 44, 47].
324 Compared to HOME CG DMRs, a smaller fraction (58%) of CG DMRs predicted by DSS
325 overlapped with gene bodies (Figure 6B). Moreover, 44% of CHG DMRs predicted by DSS
326 overlapped with gene bodies as compared to <20% HOME CHG DMRs. Non-CG methylation
327 plays an important role in silencing transposable elements (TEs) [44]. Thus, it is expected that
328 most of the non-CG DMRs detected between WT and *cmt2* will overlap with TEs, given the role
329 of CMT2 in mediating methylating of TEs, particularly long TEs, in the non-CG context. We

330 found that 75% of CHG DMRs and 62% of CHH DMRs predicted by HOME overlapped with
331 TEs, and only 8% of the CG DMRs intersected with TEs (Figure 6B). On the other hand, DMRs
332 predicted by DSS showed a similar percentage overlap with TEs for all contexts (Figure 6B).
333 Previous studies have shown that *cmt2* exhibits loss of non-CG methylation predominantly at
334 long TEs (>1kb) [44]. Hence, we further inspected the location of uniquely predicted DMRs in
335 TEs. The overall genomic distribution of long TEs (>1kb) and short TEs (<1kb) in the genome is
336 20% and 80%, respectively. A higher number of CHG and CHH DMRs predicted by HOME
337 overlapped with long TEs than with short TEs (Figure 6C). The fraction of TEs (long and short)
338 that overlapped with HOME CG DMRs was very small (<2%). For CHH DMRs both HOME
339 and DSS exhibited a similar overlap with TEs. Most CHG DMRs predicted by DSS did not
340 overlap either long or short TEs, while a larger fraction of long TEs overlapped CG DMRs
341 predicted by DSS, compared to short TEs (Figure 6C). Our results suggest that DMRs predicted
342 by HOME fit better with the known biological function of *cmt2*, than DMRs predicted by DSS.

343 **2. Time-series differential methylation analysis**

344 An additional feature of HOME is the ability to predict DMRs in time-series data. The HOME
345 time-series analysis algorithm can be successfully used for identification of DMRs in datasets
346 where DNA methylation varies over time or between development stages, for example, during
347 seed germination [48], cell reprogramming [49, 50], and mammalian brain development [5].
348 Current methods can be used to call pairwise DMRs for each combination of groups in time-
349 series data, and thereafter the DMRs could be merged to obtain final predicted time-series
350 DMRs. However, taking this approach, the complexity of all possible combinations increases
351 and becomes tedious for users, as the number of groups increases in time-series data. With
352 HOME, we provide an easy and convenient way to compare many time points with a single

353 command. Moreover, the output of the HOME time-series module contains many useful metrics
354 that allow users to trace methylation changes through time and determine the stability or
355 stochasticity of the methylation state in the DMRs. For example, the output summarizes the
356 mean methylation level difference and directionality of methylation level change, for each pair
357 combination in the time-series data.

358 We tested the performance of HOME on time-series data using human brain frontal
359 cortex development WGBS data [5] with four time points: fetal, 2 years, 12 years and 25 years
360 (Figure 6, top panel). DMRs identified by HOME included expected regions, such as genes
361 related to neuronal development (Figure 6, top panel). Furthermore, we also tested the
362 performance of HOME on another time-series dataset of mouse embryonic fibroblast (MEF)
363 reprogramming to induced pluripotent stem cells (iPSCs) [51]. The dataset contains 6 time
364 points: MEF, day 3, day 6, day 9, day 12 and iPSCs. The browser representations in Figure 6,
365 bottom panel show that HOME is able to identify DMRs with gradual methylation changes
366 effectively.

367 **Conclusions**

368 Here we present a novel histogram of methylation based machine learning method to detect
369 DMRs from single nucleotide resolution DNA methylation data. The algorithm treats the
370 problem of DMR detection as a binary classification problem and requires a high quality training
371 dataset. Due to the lack of a biological dataset with known DMRs and non-DMRs, we generated
372 our own training dataset using publicly available DNA methylome data. It is important to note
373 that although HOME used DMRs and non-DMRs generated by consensus from existing DMR
374 finders such as DSS and Metilene for generating the training set, it is able to identify a high
375 number of biologically significant unique DMRs. Moreover, HOME showed more accurate

376 DMR prediction and precise DMR boundary identification compared to both DSS and Metilene.
377 The key features of HOME are: (i) novel histogram based features which combines important
378 information such as methylation level difference, measure of significance for the difference in
379 methylation, and distance between neighbouring cytosines; (ii) a robustly trained model that is
380 effective for a wide variety of species; (iii) a flexible method that can be used for prediction of
381 DMRs in both CG and CH contexts with high border accuracy; and (iv) a tool that can identify
382 DMRs in time-series data.

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

391 HOME accounts for biological variation present between the replicates and uneven read
392 coverage through weighted logistic regression while computing the p-values. The spatial
393 correlation present among neighboring cytosine sites is captured by moving average smoothing
394 and the use of weighted voting for histogram based features. We demonstrate that HOME can be
395 used to predict accurate DMRs in both CG and non-CG (CHG and CHH) sequence contexts for
396 both mammalian and plant WGBS methylome data by using the same training data. Although the
397 classifier was trained on mammalian WGBS data for CG and CH contexts, HOME can
398 accurately predict DMRs in plants and for specific non-CG contexts (CHG and CHH),

399 demonstrating its versatility. However, if users wish to retrain the HOME model on their own
400 data, it can easily be done from the approach mentioned below (see Methods section).
401 Finally, another standout feature of HOME is the prediction of DMRs in time-series data.
402 Time-series DNA methylation experiments are commonly used to study a wide range of
403 biological processes such as development [5] and stress responses [52]. HOME provides an
404 efficient algorithm to directly predict accurate DMRs in studies with multiple timepoints. This
405 added functionality of HOME will greatly facilitate and expand the study of epigenome
406 dynamics in numerous biological systems and disease models. Taken together, HOME is a
407 highly effective and robust DMR finder that accounts for uneven cytosine coverage in WGBS
408 data, accounts for biological variation present between the samples in the same treatment group,
409 predicts DMRs in various genomic contexts, and accurately identifies DMRs among any number
410 of treatment groups in experiments with or without replicates.

411 **Methods**

412 **Algorithm availability and use**

413 HOME is computationally efficient algorithm. The algorithm has been parallelized to run on
414 multiple cores. The runtimes for various applications of HOME are summarised in the Supp.
415 Table 3. The execution time of HOME, for all modules, is almost linear to the number of
416 cytosine sites, and increases only slightly when analysing multiple replicates. HOME is written
417 in python and documentation and source code are freely available in the form of a user friendly
418 package at https://github.com/ListerLab/HOME. The default parameters for HOME are set to be
419 useful for most analyses. The user can fine tune the frequently used parameters such as merge
420 distance (*--mergedist*), minimum length of reported DMRs (*--minlength*), minimum average
421 methylation difference (*--delta*) and minimum number of Cs (*--minc*) in a DMR based on the

422 datasets.

423 **Training data generation**

424 To generate the training data for the CG context, publicly available methylome data from neuron
425 and glia cell types isolated from the frontal cortex of 7 weeks old male mouse brain [5] was used,
426 considering chromosome 10 only. For the CH context, the published methylation data for
427 chromosome 2 from two neuronal cell types, excitatory pyramidal neurons (EX) with two
428 biological replicates and vasoactive intestinal peptide-expressing interneurons (VIP) with two
429 biological replicates [41] were used. The details for CG and CH context including number of
430 cytosines and median coverage about the dataset are summarised in Supp. Tables 4 and 5,
431 respectively. The raw data were processed as described previously [5] to obtain the number of
432 methylated reads (*mc*) and total reads (*t*) for each cytosine. Watson and Crick strand counts were
433 combined for *mc* and *t* for CG sites, as the methylation is usually symmetric on both strands [53-
434 55]. The strand combination was not performed for CH sites because methylation of the CH sites
435 is strand-specific [9, 55]. Thereafter, three existing DMR finders (DSS, bisulfighter and
436 Metilene) were used with default parameters to produce preliminary DMRs for the CG context.
437 For the CH context, we ran DSS and Metilene for only half of chromosome 2, due to runtime
438 constraints. DMRs identified by two or more finders were further analyzed visually and DMRs
439 with consistent methylation level difference were selected as the true set of DMRs. The regions
440 not identified as DMRs by all of the DMR finders were visually inspected for inconsistent
441 methylation level difference and selected as true non-DMRs.

442 The details on the number of training DMRs and non-DMRs used for CG and CH context
443 are provided in Supp. Tables 6 and 7, respectively. It is critical to note that for classifier training,
444 each individual cytosine site (in the DMR and non-DMR training set) is an independent

445 observation. Overall, for the CG training dataset, the number of cytosines for DMRs and non-
446 DMRs were 3,994 and 14,534 respectively. For the CH training dataset, the number of cytosines
447 for DMRs and non-DMRs were 25,027 and 1,674,110 respectively.

448 **Histogram computation from training data**

449 HOME uses novel histogram based features for identification of DMRs. The algorithm starts by
450 combining the Watson and Crick strand counts for *mc* and *t* for the CG context. For the CH
451 context, no strand combination is performed. Next, the algorithm computes the methylation level
452 difference between the two samples and estimates the p-value for the difference at each cytosine.
453 For CH training data, which has biological replicates within each treatment group, p-values are
454 computed using weighted logistic regression to model methylation levels in relation to the
455 treatment groups and variation between replicates. More specifically, at a given cytosine site, we
456 model methylation level through weighted logistic regression model. Logistic regression is used
457 to estimate the p-values for the methylation difference with a continuous predictor (methylation
458 level) and a binary outcome representing each treatment group.

459 Weighted logistic regression estimates the p-value for the discriminatory ability of each
460 cytosine to distinguish between treatment groups with the Chi-square test. The test compares
461 how well our model with intercept and methylation level as a predictor fits the data as compared
462 to the null model that includes only the intercept. We used a z-test for p-value estimation for CG
463 training which has only one sample per treatment group. The underlying null hypothesis for
464 modeling p-values for both tests described above is that methylation levels are the same among
465 treatment groups for a given cytosine. The alternative hypothesis is that there is a difference in
466 methylation levels among the treatment groups. To account for uneven read coverage, HOME
467 uses a logistic function to compute the weights for all cytosines for weighted logistic regression.

468 The weights are computed from *t*, such that the range of weight is between 0 and 1 when
469 calculating the p-value. More specifically, if the coverage is low for a particular cytosine, its
470 weight will be lower compared to a cytosine with high coverage. Thereafter, the absolute
471 difference in methylation level at each cytosine is weighted by its p-value (*p*) to compute a bin
472 value (*b*) as shown in Equation 1 below.

473 $$b = |m_1 - m_2|.e^{1-p} \quad (1)$$

474 Where, m_1 and m_2 are the methylation levels of treatment groups under comparison and
475 exponentiation of the *1-p* allows smaller p-values to contribute more to the produced bin value
476 than larger (insignificant) p-values. To account for the spatial correlation between the
477 neighboring cytosines, moving average smoothing (default: 3 cytosines) is performed on values
478 of *b* to compute final bin value b_s.

479 The histogram feature is computed for every individual cytosine present in each DMR
480 and non-DMR training data. For a given cytosine, to compute the histogram feature, a fixed
481 window of size *w* centered around it is used where *w* is the number of cytosines in a window (*w*
482 is set to 11 for CG and 51 for CH context). For each window, the bin values b_s are binned using a
483 weighted voting approach such that for a given cytosine, its contribution *v* to the bin is computed
484 as a weighted distance from the center cytosine which is normalized by the maximum allowed
485 distance as shown in Equation 2 below.

486 $$v = \begin{cases} 1 - \frac{|l - l_c|}{d} & \text{if } |l - l_c| < d \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

487 Where, l is the location of the cytosine being binned, l_c is the location of the center
488 cytosine of *w*, and d (default: 250bp) is the normalization constant signifying the maximum
489 allowed distance from the center cytosine. Consequently, the cytosines close to the center

490 cytosine will have higher weights and will contribute more to the histogram feature. On the other
491 hand, if the distance between the cytosine being binned and the center cytosine of the window is
492 larger than d, then that cytosine will have zero contribution.

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

497 **Training via SVM**

498 The algorithm then uses the normalized histogram feature vectors described above to train a
499 support vector machine (SVM) that predicts the label (DMR or non-DMR) for each individual
500 cytosine. The *scikit-learn* implementation of SVM [56] with linear kernel (with default
501 parameters) was used for training because it is computationally very efficient and shows
502 comparable performance to the more computationally expensive non-linear RBF kernel [56] and
503 random forest classifier [56-58]. The F1-scores for linear, RBF and random forest after 5-fold
504 cross-validation on the training dataset were 0.895, 0.903 & 0.903, respectively for DMRs, and
505 0.972, 0.973 & 0.973, respectively for non-DMRs.

506 **Testing and DMR prediction on new datasets**

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

509 **1. Pairwise**

510 HOME can be used to predict DMRs from methylomes of two treatment groups with single or
511 multiple replicates. To predict the DMRs, the normalized histogram features are computed for

512 each cytosine on a particular chromosome, which are then provided to the trained SVM model to
513 obtain the prediction scores that are normalized using the logistic function from the generalized
514 linear model (GLM) to lie in the range [0,1]. Individual cytosines are grouped together into
515 preliminary DMRs based on the prediction scores (default: >0.1) and the distance between
516 neighboring cytosines (default: <500bp). To produce the final DMRs, the algorithm performs a
517 boundary refinement of the preliminary DMRs such that boundaries are trimmed until *k*
518 consecutive cytosines (default: 3) have the value of the *b* (Eq.1) greater than or equal to the
519 defined threshold (default: 0.1).

520 **2. Time-series and multi-group comparisons**

521 HOME can be used to predict DMRs from time-series and multi-group studies. Given a number
522 of time points or treatments, *n*, a total of ${}^{n}C_{2}$ pairwise combinations are possible. HOME
523 computes SVM prediction scores for each of these pairwise combinations in the same manner as
524 for pairwise method described above. The prediction scores are then normalized to lie in the
525 range [0, 1] using the logistic function from the generalized linear model (GLM) to allow further
526 analysis among all pairwise comparisons. The scores are summed for each cytosine to get a final
527 score. The cytosines are then grouped into DMRs.

528 **Boundary accuracy measurement**

529 To measure the boundary accuracy, the fraction of overlap in bp between simulated and
530 predicted DMRs was calculated using BEDtools *intersect* [59] with fraction of overlap (*-f*) and
531 reciprocal overlap (*-r*) features. As defined by BEDtools, *-r* feature require that the fraction of
532 overlap be reciprocal for A and B. That is, if -f is 0.90 and -r is used, this requires that B overlap
533 at least 90% of A and that A also overlaps at least 90% of B. In our case A was simulated DMR
534 and B was predicted DMR.

535 **Differential ATAC-seq peaks**

536 Fastq files were adapter and quality trimmed and aligned to the mm10 reference genome with
537 bowtie2 [60]. As suggested by Buenrostro et al., fragment ends were offset by +4 base pairs on
538 the positive strand and -5 on the negative strand [61]. Duplicate reads were removed with Picard
539 and peaks were identified with MACS2 [62]. Differential ATAC-Seq peaks between EX and PV
540 samples were detected with MACS2 bdgdiff.

541 **Declarations**

542 **Acknowledgements**

543 We thank Dr. Egor Dolzhenko for his helpful discussions regarding the generation of simulated
544 data. We thank members of Lister Lab and Borevitz Lab for their suggestions and comments.

545 **Funding**

546 This work was supported by the Australian Research Council (ARC) Centre of Excellence
547 program in Plant Energy Biology (CE140100008). RL was supported by a Sylvia and Charles
548 Viertel Senior Medical Research Fellowship, ARC Future Fellowship (FT120100862), and
549 Howard Hughes Medical Institute International Research Scholarship (RL).

550 **Availability of data and materials**

551 WGBS data from 7 weeks old male mouse brain for neuron and glia cell types were obtained
552 from GSM1173786 and GSM1173787, respectively. Datasets for EX, PV and VIP neuronal cell
553 types were downloaded from NCBI GEO (GSE63137). Arabidopsis WT and *cmt2* datasets were
554 obtained from NCBI GEO repositories GSM1242401 and GSM1242405, respectively. Time-
555 series data for fetal, 2 year, 12 year and 25 year old human brain were obtained from GEO
556 accessions GSM1163695, GSM1167005, GSM1164630 and GSM1164632, respectively. Time-

557 series dataset for MEF, day 3, day 6, day 9, day 12 and iPSCs were obtained from GEO
558 accessions GSM2718419, GSM2718420, GSM2718421, GSM2718422, GSM2718423 and
559 GSM2718424, respectively.

560 **Author Contributions**

561 AS and RL devised the project. AS conceptualized, designed and implemented HOME with
562 statistical guidance from YVK. RL, SRE and JOB supervised experiments. AS and YVK
563 processed the biological data. AS tested HOME on simulated and biological data. AS and YVK
564 drafted the manuscript, all authors contributed to writing the manuscript.

565 **Disclosure Declaration**

566 The authors declare no conflicts of interest.

567 **Ethics approval and consent to participate**

568 Not applicable.

569 **Consent for publication**

570 Not applicable.

571 **References**

572 1. Richardson BC: **Role of DNA methylation in the regulation of cell function:**
573 **autoimmunity, aging and cancer.** *J Nutr* 2002, **132:**2401S-2405S.
574 2. Khavari DA, Sen GL, Rinn JL: **DNA methylation and epigenetic control of cellular**
575 **differentiation.** *Cell Cycle* 2010, **9:**3880-3883.

576 3. Messerschmidt DM, Knowles BB, Solter D: **DNA methylation dynamics during**
577 **epigenetic reprogramming in the germline and preimplantation embryos.** *Genes Dev*
578 2014, **28:**812-828.
579 4. Jones PA: **Functions of DNA methylation: islands, start sites, gene bodies and**
580 **beyond.** *Nat Rev Genet* 2012, **13:**484-492.
581 5. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang
582 Y, Dwork AJ, Schultz MD, et al: **Global epigenomic reconfiguration during**
583 **mammalian brain development.** *Science* 2013, **341:**1237905.
584 6. Kass SU, Landsberger N, Wolffe AP: **DNA methylation directs a time-dependent**
585 **repression of transcription initiation.** *Curr Biol* 1997, **7:**157-165.
586 7. Jones PA: **The DNA methylation paradox.** *Trends Genet* 1999, **15:**34-37.
587 8. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X,
588 Bernstein BE, Nusbaum C, Jaffe DB, et al: **Genome-scale DNA methylation maps of**
589 **pluripotent and differentiated cells.** *Nature* 2008, **454:**766-770.
590 9. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee
591 L, Ye Z, Ngo QM, et al: **Human DNA methylomes at base resolution show**
592 **widespread epigenomic differences.** *Nature* 2009, **462:**315-322.
593 10. Law JA, Jacobsen SE: **Establishing, maintaining and modifying DNA methylation**
594 **patterns in plants and animals.** *Nat Rev Genet* 2010, **11:**204-220.
595 11. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E,
596 Wirbelauer C, Oakeley EJ, Gaidatzis D, et al: **DNA-binding factors shape the mouse**
597 **methylome at distal regulatory regions.** *Nature* 2011, **480:**490-495.

598 12. Bogdanovic O, Smits AH, de la Calle Mustienes E, Tena JJ, Ford E, Williams R,
599 Senanayake U, Schultz MD, Hontelez S, van Kruijsbergen I, et al: **Active DNA**
600 **demethylation at enhancers during the vertebrate phylotypic period.** *Nat Genet*
601 2016, **48:**417-426.
602 13. Heyn H, Moran S, Hernando-Herraez I, Sayols S, Gomez A, Sandoval J, Monk D, Hata
603 K, Marques-Bonet T, Wang L, Esteller M: **DNA methylation contributes to natural**
604 **human variation.** *Genome Res* 2013, **23:**1363-1372.
605 14. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour
606 P, Zhang Z, Wang J, Ziller MJ, et al: **Integrative analysis of 111 reference human**
607 **epigenomes.** *Nature* 2015, **518:**317-330.
608 15. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N,
609 Nery JR, Urich MA, Chen H, et al: **Human body epigenome maps reveal noncanonical**
610 **DNA methylation variation.** *Nature* 2015, **523:**212-216.
611 16. Eichten SR, Stuart T, Srivastava A, Lister R, Borevitz JO: **DNA methylation profiles of**
612 **diverse Brachypodium distachyon align with underlying genetic diversity.** *Genome*
613 *Res* 2016, **26:**1520-1531.
614 17. Kawakatsu T, Stuart T, Valdes M, Breakfield N, Schmitz RJ, Nery JR, Urich MA, Han
615 X, Lister R, Benfey PN, Ecker JR: **Unique cell-type-specific patterns of DNA**
616 **methylation in the root meristem.** *Nat Plants* 2016, **2:**16058.
617 18. Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, Li Q, Rohr NA, Rambani A,
618 Burke JM, Udall JA, et al: **Widespread natural variation of DNA methylation within**
619 **angiosperms.** *Genome Biol* 2016, **17:**194.

620 19. Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, Dempster EL, Ren B: **Base-**
621 **resolution analyses of sequence and parent-of-origin dependent DNA methylation in**
622 **the mouse genome.** *Cell* 2012, **148:**816-831.
623 20. Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, Cross MK,
624 Williams BA, Stamatoyannopoulos JA, Crawford GE, et al: **Dynamic DNA methylation**
625 **across diverse human cell lines and tissues.** *Genome Res* 2013, **23:**555-567.
626 21. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson
627 SF, Pellegrini M, Jacobsen SE: **Shotgun bisulphite sequencing of the Arabidopsis**
628 **genome reveals DNA methylation patterning.** *Nature* 2008, **452:**215-219.
629 22. Guo S, Diep D, Plongthongkum N, Fung HL, Zhang K, Zhang K: **Identification of**
630 **methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples**
631 **and tumor tissue-of-origin mapping from plasma DNA.** *Nat Genet* 2017, **49:**635-642.
632 23. Hansen KD, Langmead B, Irizarry RA: **BSmooth: from whole genome bisulfite**
633 **sequencing reads to differentially methylated regions.** *Genome Biol* 2012, **13:**R83.
634 24. Dolzhenko E, Smith AD: **Using beta-binomial regression for high-precision**
635 **differential methylation analysis in multifactor whole-genome bisulfite sequencing**
636 **experiments.** *BMC Bioinformatics* 2014, **15:**215.
637 25. Lea AJ, Tung J, Zhou X: **A Flexible, Efficient Binomial Mixed Model for Identifying**
638 **Differential DNA Methylation in Bisulfite Sequencing Data.** *PLoS Genet* 2015,
639 **11:**e1005650.
640 26. Hebestreit K, Dugas M, Klein HU: **Detection of significantly differentially methylated**
641 **regions in targeted bisulfite sequencing data.** *Bioinformatics* 2013, **29:**1647-1653.

642 27. Shafi A, Mitrea C, Nguyen T, Draghici S: **A survey of the approaches for identifying**
643 **differential methylation using bisulfite sequencing data.** *Brief Bioinform* 2017.
644 28. Saito Y, Tsuji J, Mituyama T: **Bisulfighter: accurate detection of methylated cytosines**
645 **and differentially methylated regions.** *Nucleic Acids Res* 2014, **42:**e45.
646 29. Wang Z, Li X, Jiang Y, Shao Q, Liu Q, Chen B, Huang D: **swDMR: A Sliding Window**
647 **Approach to Identify Differentially Methylated Regions Based on Whole Genome**
648 **Bisulfite Sequencing.** *PLoS One* 2015, **10:**e0132866.
649 30. Feng H, Conneely KN, Wu H: **A Bayesian hierarchical model to detect differentially**
650 **methylated loci from single nucleotide resolution sequencing data.** *Nucleic Acids Res*
651 2014, **42:**e69.
652 31. Wu H, Xu T, Feng H, Chen L, Li B, Yao B, Qin Z, Jin P, Conneely KN: **Detection of**
653 **differentially methylated regions from whole-genome bisulfite sequencing data**
654 **without replicates.** *Nucleic Acids Res* 2015, **43:**e141.
655 32. Juhling F, Kretzmer H, Bernhart SH, Otto C, Stadler PF, Hoffmann S: **metilene: fast and**
656 **sensitive calling of differentially methylated regions from bisulfite sequencing data.**
657 *Genome Res* 2016, **26:**256-262.
658 33. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV,
659 Davies R, Down TA, et al: **DNA methylation profiling of human chromosomes 6, 20**
660 **and 22.** *Nat Genet* 2006, **38:**1378-1385.
661 34. Jaffe AE, Feinberg AP, Irizarry RA, Leek JT: **Significance analysis and statistical**
662 **dissection of variably methylated regions.** *Biostatistics* 2012, **13:**166-178.
663 35. Cortes C, Vapnik, V. : **Support-vector networks.** *Machine Learning* 1995, **20:**273
664

665 36. Rye MB, Saetrom P, Drablos F: **A manually curated ChIP-seq benchmark**
666 **demonstrates room for improvement in current peak-finder programs.** *Nucleic Acids*
667 *Res* 2011, **39:**e25.
668 37. Hocking TD, Goerner-Potvin P, Morin A, Shao X, Pastinen T, Bourque G: **Optimizing**
669 **ChIP-seq peak detectors using visual labels and supervised machine learning.**
670 *Bioinformatics* 2017, **33:**491-499.
671 38. Lingren T, Deleger L, Molnar K, Zhai H, Meinzen-Derr J, Kaiser M, Stoutenborough L,
672 Li Q, Solti I: **Evaluating the impact of pre-annotation on annotation speed and**
673 **potential bias: natural language processing gold standard development for clinical**
674 **named entity recognition in clinical trial announcements.** *J Am Med Inform Assoc*
675 2014, **21:**406-413.
676 39. Fridman L, Reimer B: **Semi-Automated Annotation of Discrete States in Large Video**
677 **Datasets.** In *ArXiv e-prints*; 2016.
678 40. Sagonas C, Antonakos E, Tzimiropoulos G, Zafeiriou S, Pantic M: *300 Faces In-The-*
679 *Wild Challenge: database and results.* 2016.
680 41. Mo A, Mukamel EA, Davis FP, Luo C, Henry GL, Picard S, Urich MA, Nery JR,
681 Sejnowski TJ, Lister R, et al: **Epigenomic Signatures of Neuronal Diversity in the**
682 **Mammalian Brain.** *Neuron* 2015, **86:**1369-1384.
683 42. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM,
684 Bejerano G: **GREAT improves functional interpretation of cis-regulatory regions.**
685 *Nat Biotechnol* 2010, **28:**495-501.

686 43. Wen Y, Chen F, Zhang Q, Zhuang Y, Li Z: **Detection of differentially methylated**
687 **regions in whole genome bisulfite sequencing data using local Getis-Ord statistics.**
688 *Bioinformatics* 2016, **32:**3396-3404.
689 44. Stroud H, Do T, Du J, Zhong X, Feng S, Johnson L, Patel DJ, Jacobsen SE: **Non-CG**
690 **methylation patterns shape the epigenetic landscape in Arabidopsis.** *Nat Struct Mol*
691 *Biol* 2014, **21:**64-72.
692 45. Zemach A, Kim MY, Hsieh PH, Coleman-Derr D, Eshed-Williams L, Thao K, Harmer
693 SL, Zilberman D: **The Arabidopsis nucleosome remodeler DDM1 allows DNA**
694 **methyltransferases to access H1-containing heterochromatin.** *Cell* 2013, **153:**193-
695 205.
696 46. Crisp PA, Ganguly DR, Smith AB, Murray KD, Estavillo GM, Searle I, Ford E,
697 Bogdanovic O, Lister R, Borevitz JO, et al: **Rapid Recovery Gene Downregulation**
698 **during Excess-Light Stress and Recovery in Arabidopsis.** *Plant Cell* 2017, **29:**1836-
699 1863.
700 47. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR:
701 **Highly integrated single-base resolution maps of the epigenome in Arabidopsis.** *Cell*
702 2008, **133:**523-536.
703 48. Narsai R, Secco D, Schultz MD, Ecker JR, Lister R, Whelan J: **Dynamic and rapid**
704 **changes in the transcriptome and epigenome during germination and in developing**
705 **rice (Oryza sativa) coleoptiles under anoxia and re-oxygenation.** *Plant J* 2017,
706 **89:**805-824.

707 49. Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, Antosiewicz-Bourget J,
708 O'Malley R, Castanon R, Klugman S, et al: **Hotspots of aberrant epigenomic**
709 **reprogramming in human induced pluripotent stem cells.** *Nature* 2011, **471:**68-73.
710 50. Lee DS, Shin JY, Tonge PD, Puri MC, Lee S, Park H, Lee WC, Hussein SM, Bleazard T,
711 Yun JY, et al: **An epigenomic roadmap to induced pluripotency reveals DNA**
712 **methylation as a reprogramming modulator.** *Nat Commun* 2014, **5:**5619.
713 51. Knaupp AS, Buckberry S, Pflueger J, Lim SM, Ford E, Larcombe MR, Rossello FJ, de
714 Mendoza A, Alaei S, Firas J, et al: **Transient and Permanent Reconfiguration of**
715 **Chromatin and Transcription Factor Occupancy Drive Reprogramming.** *Cell Stem*
716 *Cell* 2017, **21:**834-845 e836.
717 52. Dowen RH, Pelizzola M, Schmitz RJ, Lister R, Dowen JM, Nery JR, Dixon JE, Ecker
718 JR: **Widespread dynamic DNA methylation in response to biotic stress.** *Proc Natl*
719 *Acad Sci U S A* 2012, **109:**E2183-2191.
720 53. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW,
721 Rigoutsos I, Loring J, Wei CL: **Dynamic changes in the human methylome during**
722 **differentiation.** *Genome Res* 2010, **20:**320-331.
723 54. Hardcastle TJ: **High-throughput sequencing of cytosine methylation in plant DNA.**
724 *Plant Methods* 2013, **9:**16.
725 55. Guo W, Chung WY, Qian M, Pellegrini M, Zhang MQ: **Characterizing the strand-**
726 **specific distribution of non-CpG methylation in human pluripotent cells.** *Nucleic*
727 *Acids Res* 2014, **42:**3009-3016.

728 56. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M,
729 Prettenhofer P, Weiss R, Dubourg V, et al: **Scikit-learn: Machine Learning in Python.**
730 *J Mach Learn Res* 2011, **12:**2825-2830.
731 57. Breiman L: **Random Forests.** *Mach Learn* 2001, **45:**5-32.
732 58. Karpievitch YV, Hill EG, Leclerc AP, Dabney AR, Almeida JS: **An introspective**
733 **comparison of random forest-based classifiers for the analysis of cluster-correlated**
734 **data by way of RF++.** *PLoS One* 2009, **4:**e7087.
735 59. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic**
736 **features.** *Bioinformatics* 2010, **26:**841-842.
737 60. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods*
738 2012, **9:**357-359.
739 61. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ: **Transposition of native**
740 **chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-**
741 **binding proteins and nucleosome position.** *Nat Methods* 2013, **10:**1213-1218.
742 62. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers
743 RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq (MACS).** *Genome*
744 *Biol* 2008, **9:**R137.

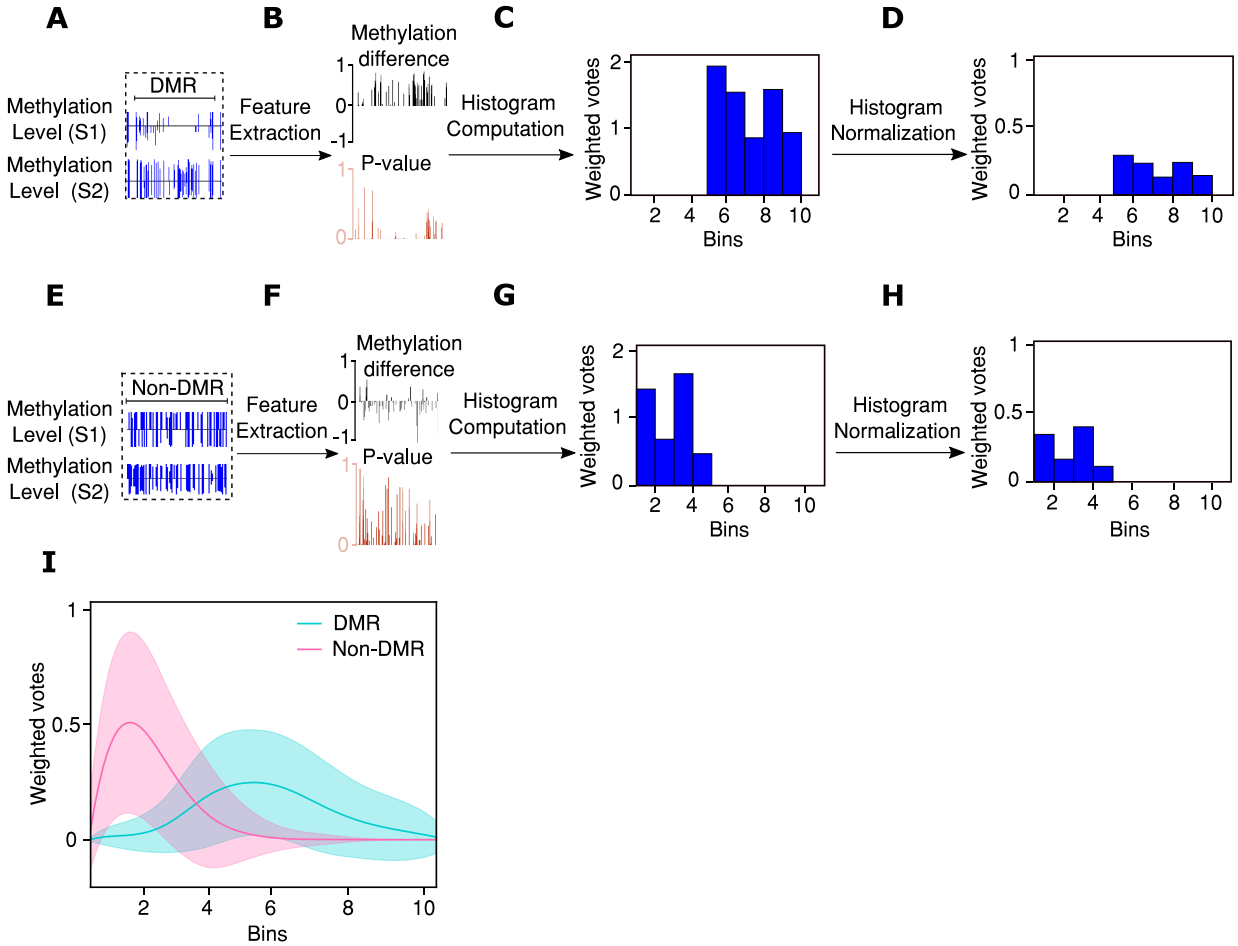

Figure 1. Feature generation overview. (A) Methylation level of sample 1 (S1) and sample 2 (S2) for a DMR from the training set. **(B)** Extracted features: p-value and difference in methylation level for each CG site. **(C)** Histogram of scores computed from the extracted features and **(D)** histogram 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).

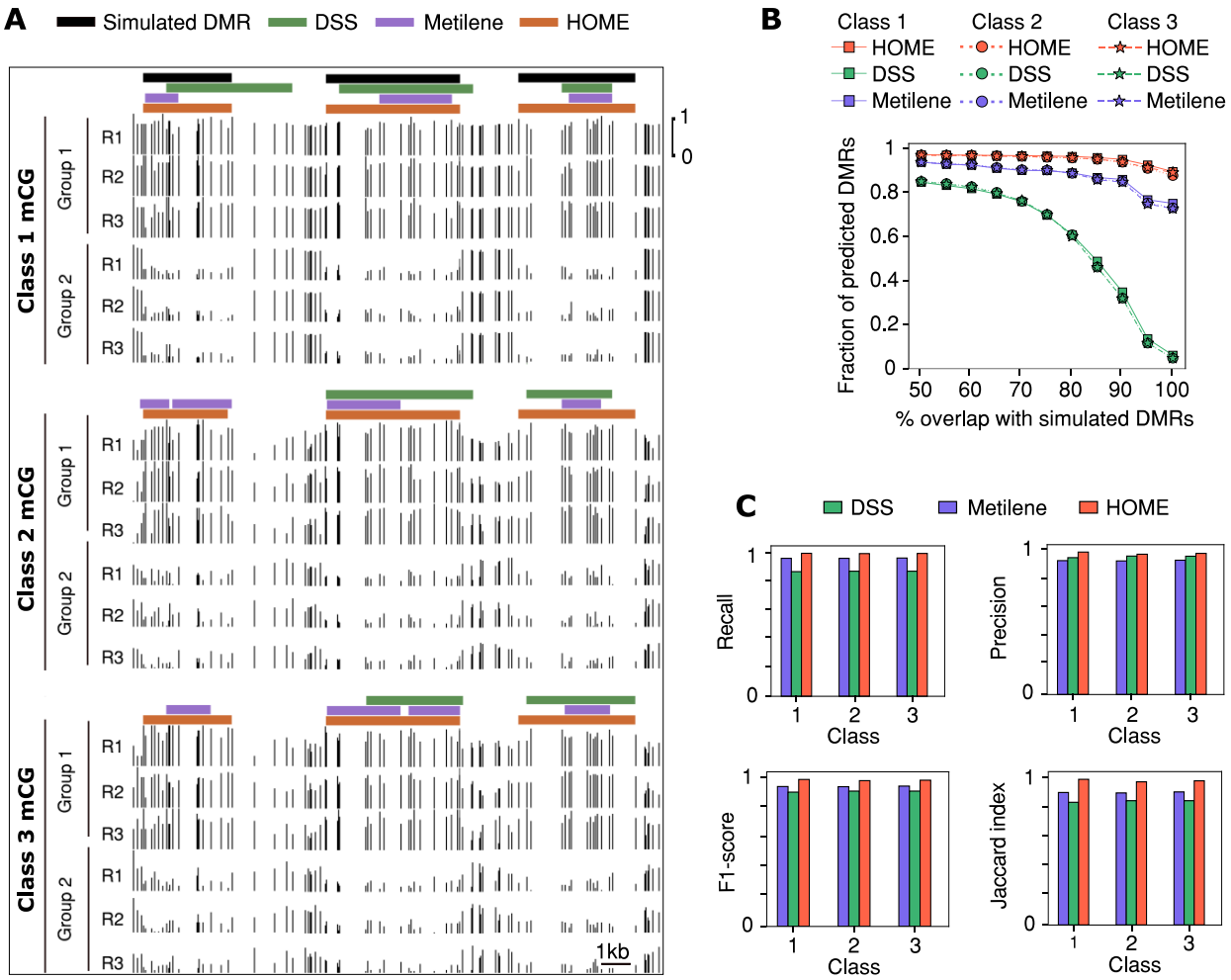

Figure 2. Comparison of DMR detection methods (HOME, DSS and Metilene) on simulated data. (A) Browser representation showing the quality and boundary accuracy of predicted DMRs by HOME, DSS and Metilene on three simulated classes. Simulated DMRs (black) and the scale are common for all three classes. **(B)** Percent reciprocal overlap ranging from 50-100% between simulated and predicted DMRs by HOME, DSS and Metilene for three classes. **(C)** Performance of HOME, Metilene and DSS in terms of recall, precision, F1-score and Jaccard index for three classes.

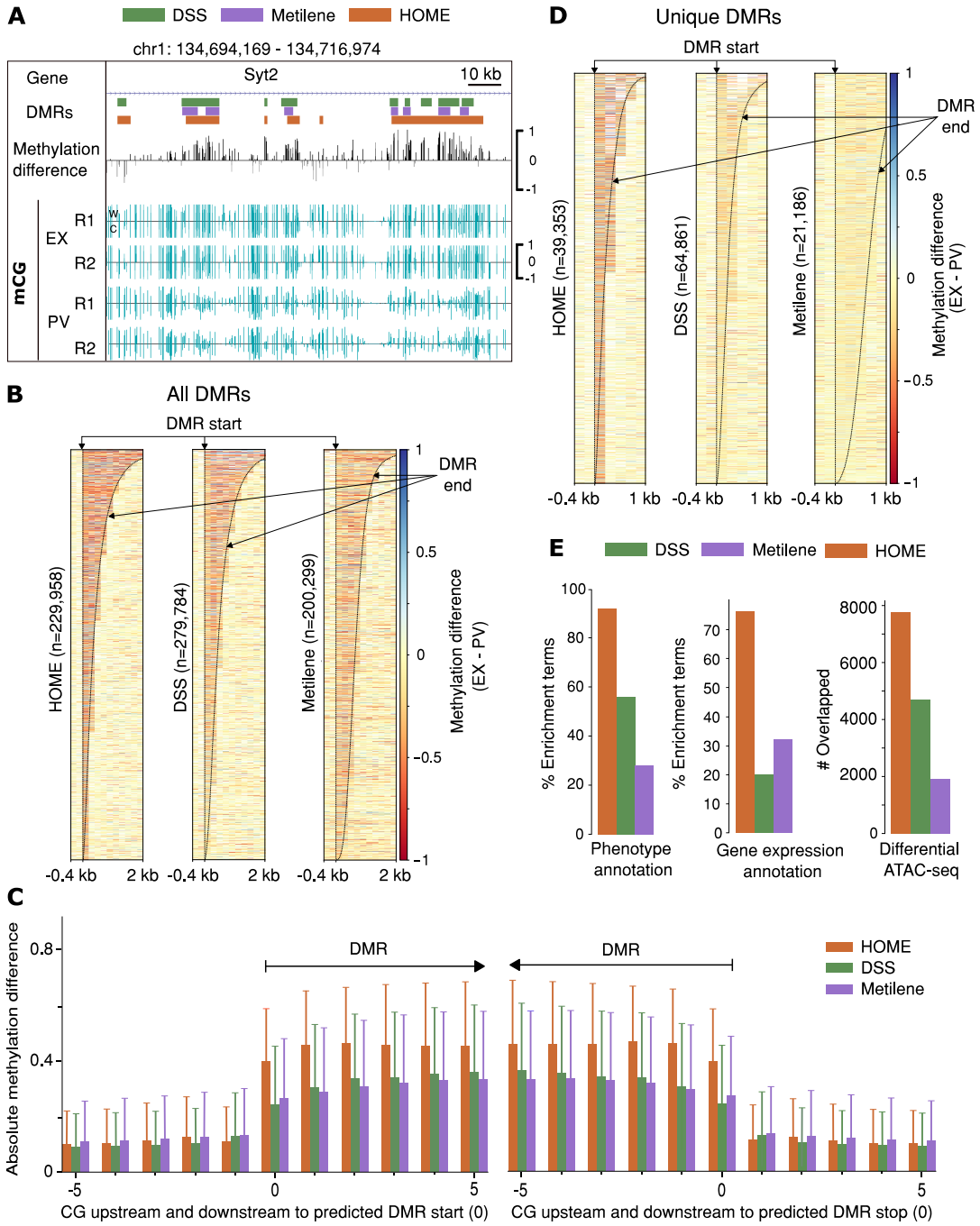

Figure 3. Quality assessment of CG-DMRs predicted in mammalian WGBS data by HOME, DSS and Metilene. (A) Browser representation showing the quality and boundary accuracy of predicted CG context DMRs for the parvalbumin-expressing fast-spiking interneuron (PV) specific gene Syt2. EX, excitatory pyramidal 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)** Heatmap of methylation level difference for all predicted DMRs by HOME, DSS and Metilene. The DMRs are sorted by length. The bin size is 200 bp for all heatmaps. **(C)** 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. **(D)** Heatmap of methylation level difference for uniquely predicted DMRs by HOME, DSS and Metilene. **(E)** Percentage of enrichment terms associated with uniquely predicted DMRs directly related to neural system functions according to mouse phenotype annotations, gene expression annotations as defined in GREAT, and number of differential ATAC-seq peaks between EX and PV.

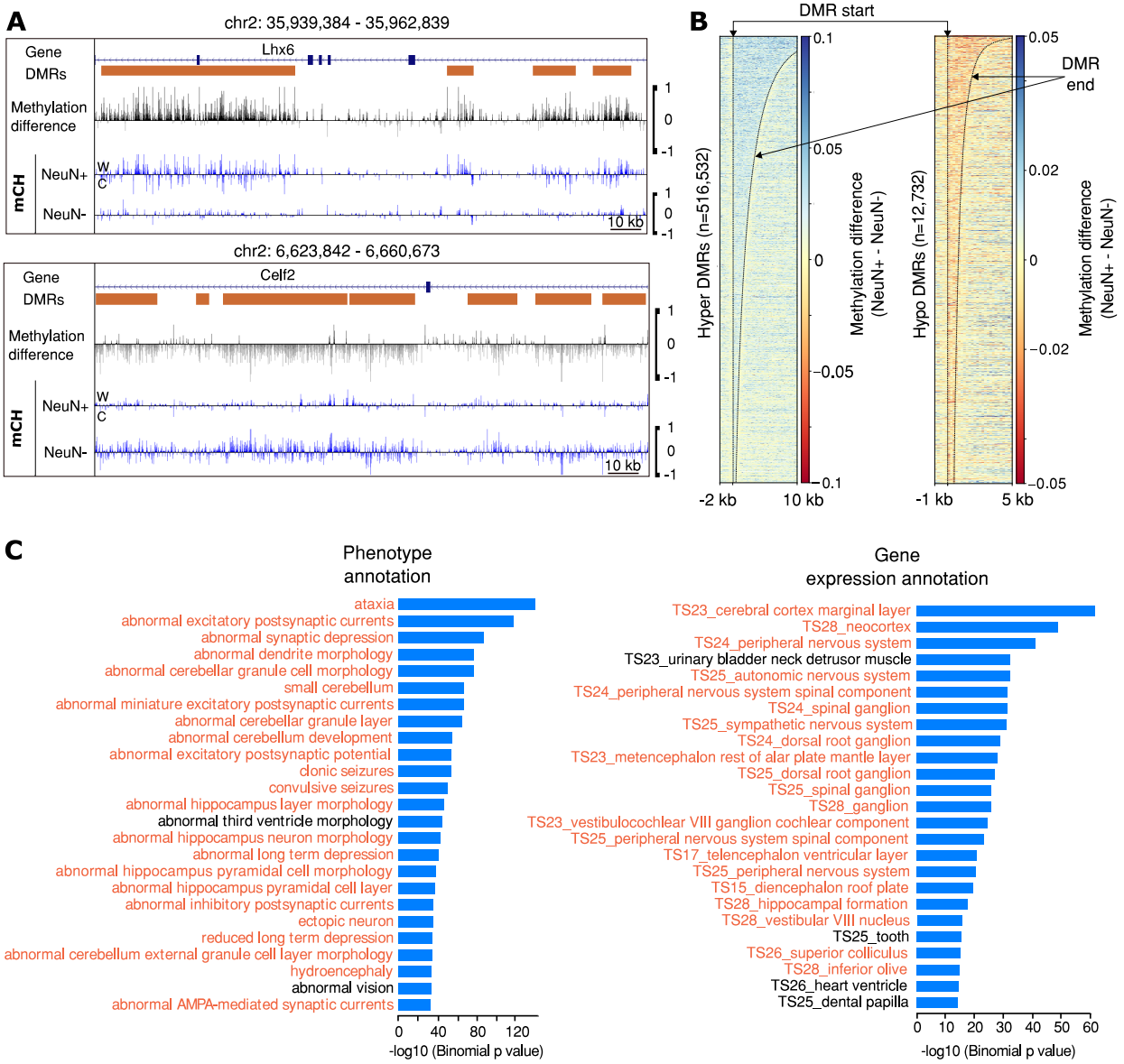

Figure 4. Quality assessment of CH-DMRs predicted in mammalian WGBS data. (A) Genome browser representation of the quality and boundary accuracy of HOME predicted CH context DMRs for neurons (NeuN+) and glia (NeuN-) methylation data. Top panel represents hyper-methylated DMRs in NeuN+ and bottom panel represents hypo-methylated DMRs in NeuN+. **(B)** Heatmap of methylation level difference for hyper-methylated HOME DMRs and hypo-methylated HOME DMRs. **(C)** Biological annotations of hypo-methylated HOME DMRs in NeuN+ cells. displaying the top 25 terms using the mouse phenotype annotation and the MGI gene expression annotation (neuron, or glia, and brain tissue related terms are highlighted in orange).

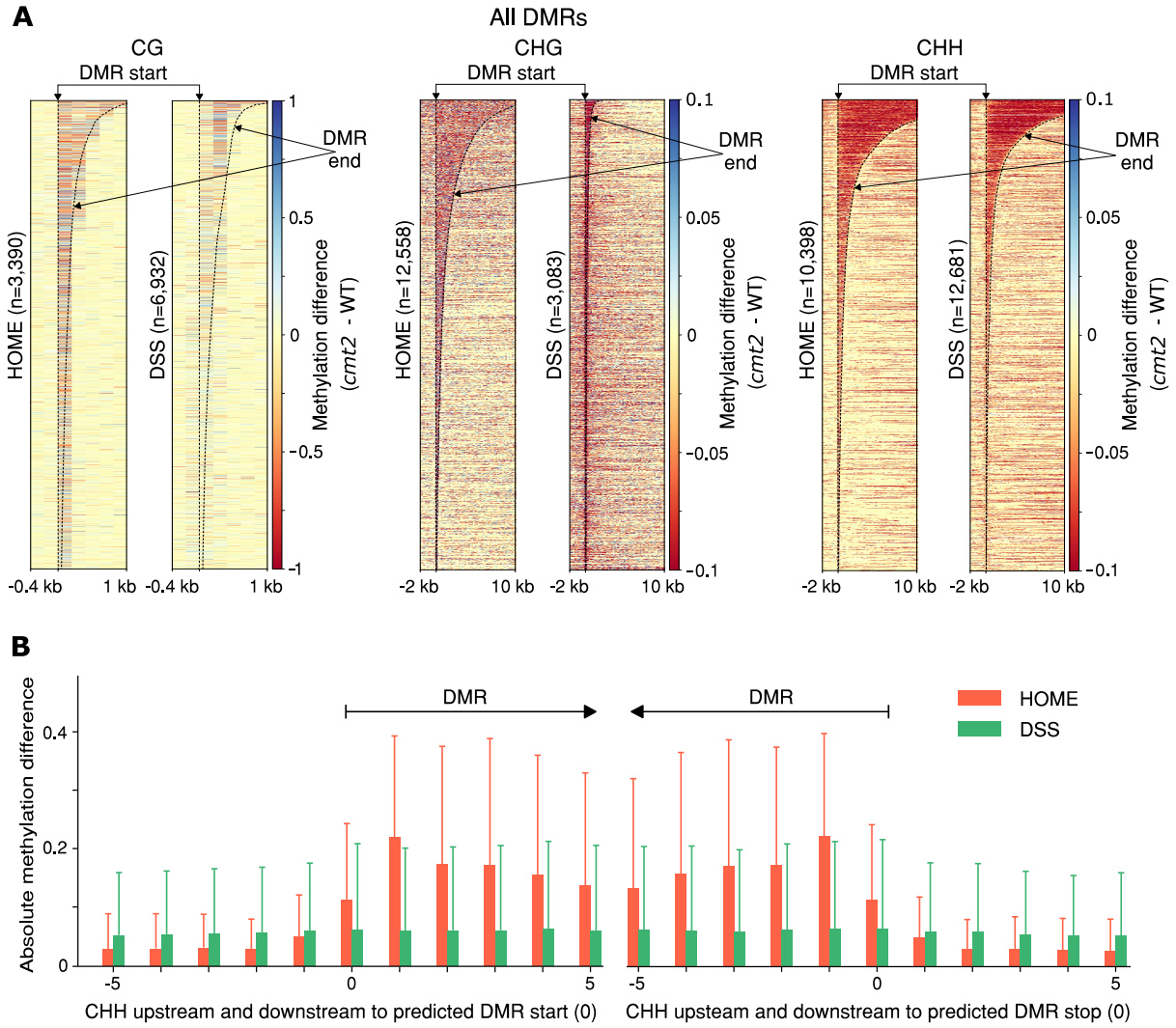

Figure 5. Qualitative analysis of all predicted DMRs by HOME and DSS in plant WGBS data. (A) Heatmap of methylation level difference for all predicted HOME and DSS DMRs, between cmt2 and WT, for CG, CHG and CHH contexts. **(B)** Mean and standard deviation of absolute methylation difference for all predicted CHH DMRs by HOME and DSS for 5 CGs upstream and downstream of the DMR start (left) and stop (right) marked as 0, respectively.

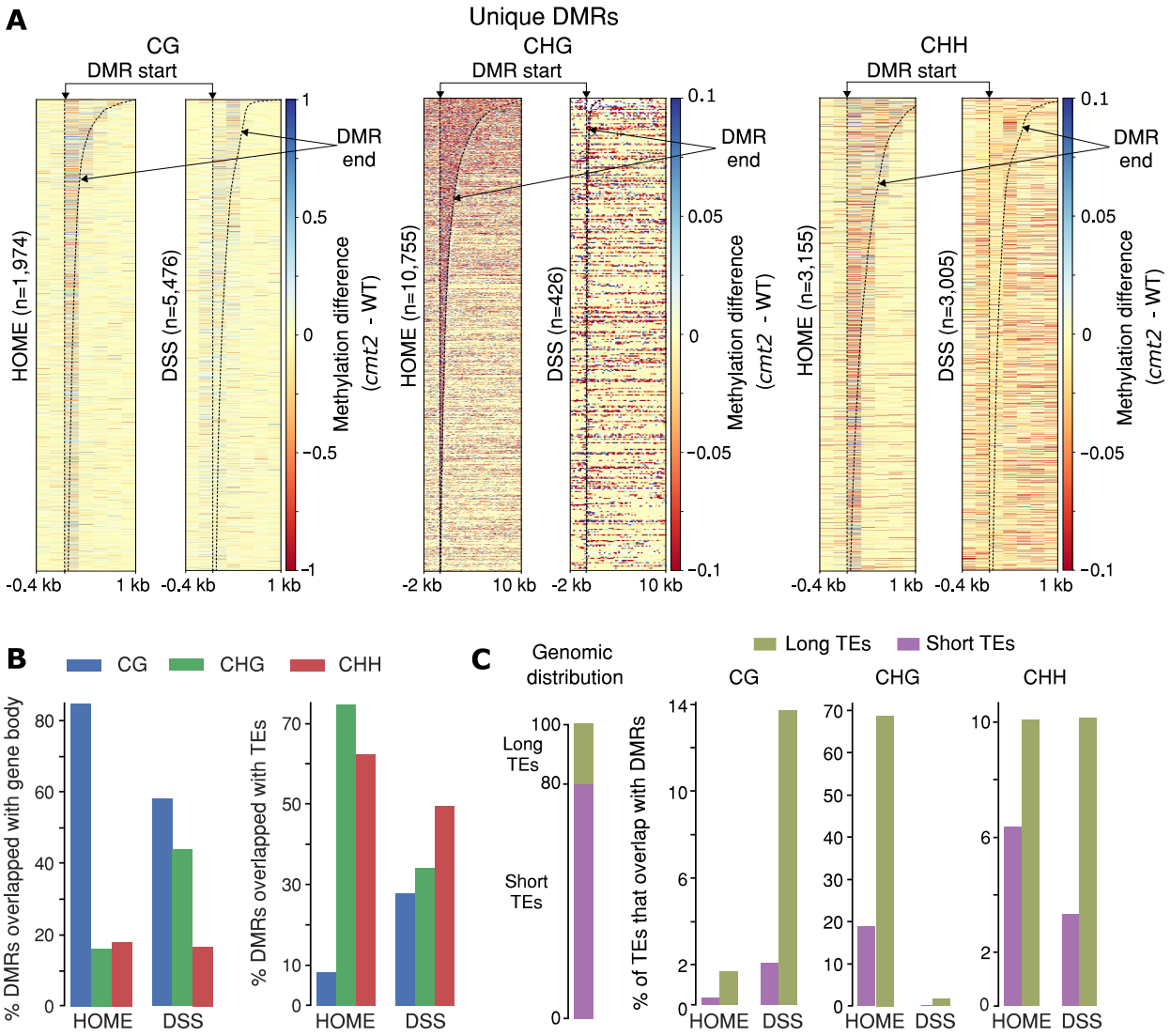

Figure 6. Qualitative analysis of uniquely predicted DMRs by HOME and DSS in plant WGBS data. (A) Heatmap of methylation level difference for all predicted HOME and DSS DMRs, between cmt2 and WT, for CG, CHG and CHH contexts. **(B)** Percentage of uniquely predicted DMRs by HOME and DSS in gene bodies and TEs. **(C)** Percentage of short (<1kb) and long(>1kb) TEs that overlap with uniquely predicted DMRs by HOME and DSS, for CG, CHG and CHH contexts.

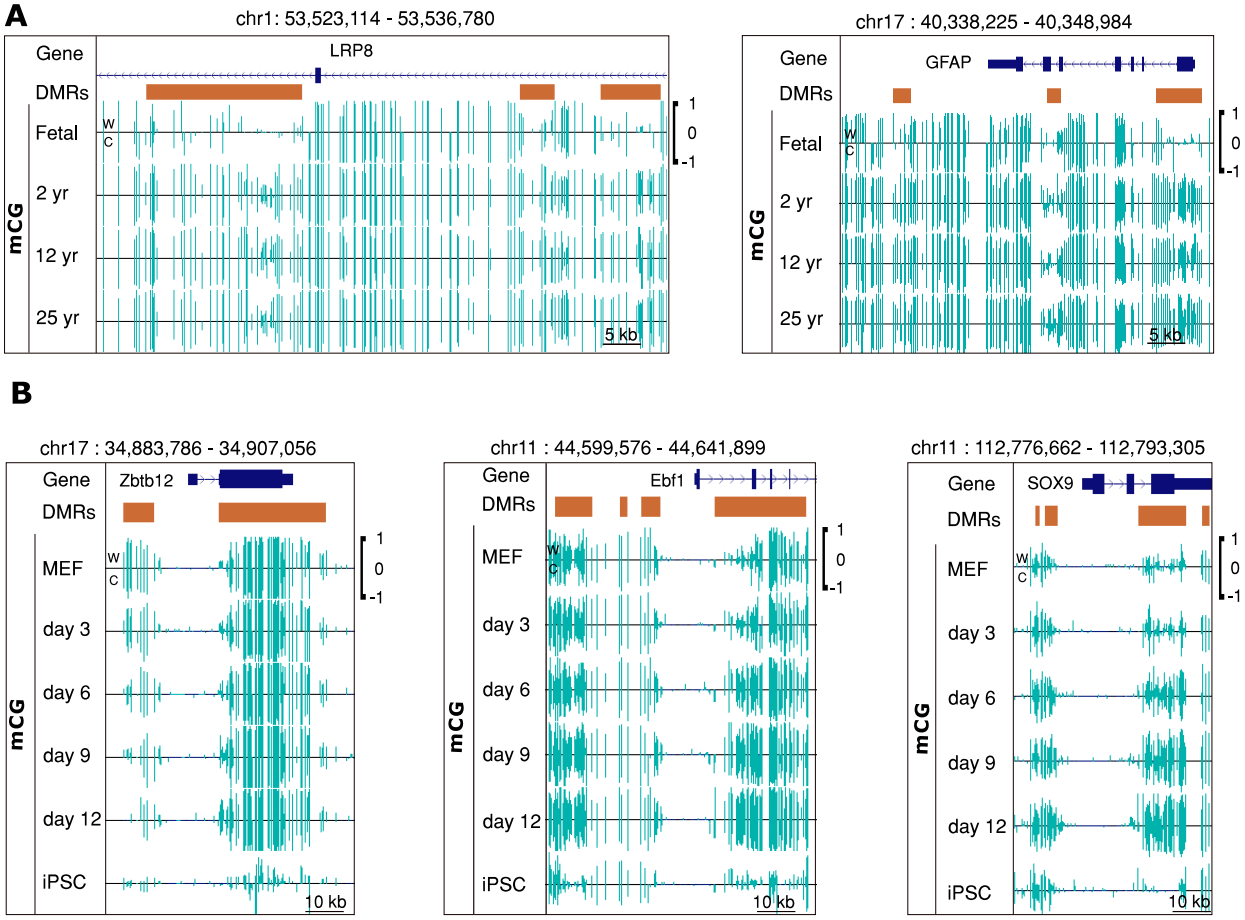

Figure 7. Browser representations of CG DMRs predicted by HOME in time-series WGBS data. (A) Four time points: fetal, 2, 12 and 25 year old mammalian brain, and **(B)** with 6 time points: mouse embryonic fibroblast (MEF), day 3, day 6, day 9, day 12 and induced pluripotent stem cell (iPSC).