Processive and stochastic CHH methylation by plant DRM2 and CMT2 revealed by single-read methylome analysis

Keith D. Harris1 and Assaf Zemach1,*

1 School of Plant Sciences and Food Security, Tel Aviv University, Tel Aviv, Israel
*Corresponding author: Assaf Zemach, assafze@tauex.tau.ac.il

Abstract

Cytosine methylome data is commonly generated through next-generation sequencing (NGS). Analyses of this data average methylation states of individual reads. We propose an alternate method of analysing single-read methylome data. Using this method, we identified patterns that relate to the mechanism of two plant non-CG methylating enzymes, DRM2 and CMT2: DRM2 has higher processivity than CMT2, and DRM2-methylated regions have higher variation among cells. Based on these patterns, we developed a classifier that predicts enzyme activity in different species and tissues. To facilitate further single-read analyses, we developed a genome browser optimised for visualising and analysing NGS data at single-read resolution.

Background

DNA methylation is a conserved epigenetic mechanism that regulates genome stability and expression in diverse eukaryotes [1–4]. This regulation is based on a dynamic addition or removal of a methyl group to/from the fifth carbon of a cytosine residue. DNA methylation appears in distinct genomic features, such as genes and transposable elements (TEs), and in different chromatin states, such as heterochromatin and euchromatin [2, 5–9]. In plants, DNA methylation occurs in three contexts: CG, CHG and CHH (where H is any base except G). These contexts are differentially regulated by four DNA methyltransferase (DNMT) families that share conserved methyl-transferase domain (MTD). METHYLTRANSFERASE1 (MET1) recognises hemi-methylated CG following DNA replication, and methylates the naked cytosine in the daughter strand [10, 11]. CHROMOMETHYLASEs (CMTs), which are plant-specific DNMTs, bind histone H3 lysine 9 (H3K9me2) heterochromatin via a chromodomain (CD) to methylate non-CG contexts [12]. In flowering plants, CMT3 methylates mostly CHG sites, whereas CMT2 methylates mostly CHH sites [13, 14]. The CHH methylation state is additionally regulated by plant DNMT3 orthologues or homologs, i.e. the DOMAINS REARRANGED METHYLASEs (DRMs) [15, 16]. Similar to animal DNMT3, plant DNMT3 and DRMs function as de novo methylases, establishing methylation on unmethylated sites.
The relationship between changes in DNA methylation patterns and gene expression is not trivial, as it involves a non-linear, additive effect of multiple methylation contexts, along with the effect of additional levels of epigenetic regulation, including chromatin structure, histone position and modifications [1] [17] [18]. Additionally, the most common method for studying DNA methylation (bisulfite sequencing, or BS-seq), does not provide information on the methylation states of individual cells. BS-seq involves a chemical reaction that converts unmethylated cytosines into uracil, which are subsequently read as thymine when sequenced [19]. Sequencing data produced by BS-seq consists of short DNA fragments originating from a random subset of cells in the sample tissue; cytosines that have not been converted to thymine are assumed to represent methylated cytosines in the source genome [20] [21]. Hypothetically, each read relates to the methylation state of a single cell in the sample, and so the collection of reads will reflect methylation heterogeneity present in the sample tissue. However, BS-seq methylome data is most commonly averaged among reads overlapping the same region. Thus, the output signal of BS-seq analysis pipelines combine populations that may have fundamentally different methylation levels.

While alternative methods exist for generating methylomes that address this issue, namely single-cell BS-seq, these methods are currently not feasible for all organisms and tissues [22] [25]. As a consequence, most currently available methylome data is not single-cell; analyses that can decode additional dimensions of information from this type of data are of high potential value in producing more insights from new and existing data. One such analysis was recently proposed to produce a heterogeneity signal from CG methylation patterns among cells [26]. This tool calculates the Shannon entropy of reads overlapping a set of CG sites, and identifies unique patterns of methylation within this subsample, as relating to heterogeneity within the sample population of cells.

We were interested in extracting additional information from BS-seq data, specifically relating to CHH methylation, which could identify enzyme-specific patterns of methylation. To this end, we designed a single-read analysis pipeline that extrapolates multiple dimensions of methylation variation, using NGS reads from either a single-region (collection of CHH sites) or from functionally-similar sets of regions. This analysis revealed that DRM2 and CMT2 have distinctive methylation patterns at both single-cell and population levels. DRM2-methylated reads show higher processivity, whereas CMT2-methylated reads and regions are more stochastically methylated. These findings shed new light on the distinct mechanisms of CHH methylating enzymes and can be utilized for prediction purposes. By characterizing these patterns in Arabidopsis thaliana mutants of these enzymes, we developed a classifier that can predict the identity of the enzyme that methylates a particular region. Importantly, the classifier does not rely on a comparison to mutants of the same species or tissue. At a genome level, it can predict the presence or absence of DRM2-like or CMT2-like activity. After validating the classifier, we used it to predict null DRM2 CHH methylation activity and to associate CMT2 methylation pattern to that of DNMT3 methylation signal in the early land plant species and human cells.

Our analyses use BS-seq data at single-read resolution. To facilitate further analyses, we developed a genome browser, Single-Read Browser (SRBrowse), that is optimised for visualizing and analysing NGS data at single-read resolution. The tool, which has a single user interface for browsing and analyses, can directly process local NGS data or NCBI accessions into an optimised format for display in the genome browser.
Results

Designing a single-read analysis pipeline for CHH-methylation

Single-read analyses can be usefully applied to any NGS data where short reads vary in a way that reflects a biological signal. With BS-seq data, the importance of this analysis is that methylation varies between cells meaningfully, and each read hypothetically reflects the state of a single cell. We chose to focus on CHH methylation for a number of reasons: (i) CHH sites are 2-3 times more common than CG/CHG sites. Given that site density dictates the amount of information that can be deduced from a single read, contexts with a higher density allow more data to be retrieved from individual samples of varying coverage; (ii) as opposed to CG sites the methylation of which is usually binary, CHH sites are mostly partially methylated \[20, 21, 27\]. Sites that are either unmethylated or fully methylated have low or zero variation among reads; (iii) CHH methylation is known to vary between tissues \[28\]. In itself, the fact that CHH sites are partially methylated suggests that most CHH sites are differentially methylated between cells of the sample tissue \[20, 21\]; (iv) CHH sites are methylated by two types of DNMTs, DRM and CMT, which their activity is regulated by distinct molecular mechanisms, RdDM and DDM1-dependent respectively \[13, 14\]. Thus, focusing on CHH sites might expose the potential variation between regions of the same sample due to the different mechanisms.

There are a number of factors that limit the maximal region size, mainly: (1) the average read length and coverage of the sequencing library; and (2) the frequency of the specific methylated context. The expected number of reads per region for a given library can be calculated as:

\[
\text{expected reads per region} = \text{coverage} \times (1 - \frac{\text{region size}}{\text{read length}})
\]

This relationship is illustrated in Figure S1a. The frequency of the methylation context is also important to consider, as selecting regions rich in a particular context can limit analyses to a small subset of the data and thus bias the results of the pipeline (Fig. S1b). Different types of analyses can utilize different filter options. For all analyses except where noted otherwise, regions were selected with 5 CHH sites, up to 30bp length. In a wild type A. thaliana sample \[14\], this includes 58% of regions from TEs containing 5 CHH sites with \( >5\% \) methylation (Fig. S1c).

In order to study individual- and population-level variation in methylation, the pipeline segments the genome into short regions of a limited length of similar functional elements or annotations, e.g. TEs, genes, exons, histone marks, chromatin structure, etc. Due to the asymmetry of CHH sites, these regions are defined per strand. For instance, a short region could span 30 base pairs and have 5 CHH sites (Fig. 1h). Reads overlapping with all 5 CHH sites are then scored according to the methylation state of these sites, within the read (Fig. 1i). We defined three features of methylation variation: (1) the standard deviation of CHH site methylation; (2) the standard deviation of read methylation; and (3) stochasticity (Fig. 1b). These features relate to functional differences between methylation patterns: (1) higher variation among sites can reflect fluctuations in the methylation signal and/or CHH subcontext specificity of the enzyme; (2) higher variation among reads can reflect differential regulation of methylation among cells composing the sample; and (3) higher stochasticity can reflect lower processivity in the mechanism of the methylating enzyme, and/or subcontext specificity.
Figure 1: Pipeline for analysing variation of single-read data regarding CHH methylated regions. (a) Schematic representation of pipeline: regions of up to a set span of base pairs with set number of CHH sites are selected; reads overlapping these sites are then quantified by the number of methylated C sites. (b) Schematic representation of different types of variation within methylated regions: (i) regional variation, relating to differences in methylation average between adjacent CHH sites; (ii) read methylation variation, relating to differences in methylation level among reads overlapping the same region, taken to represent the sample population of cells; (iii) within-cell variation, relating to differences between adjacent CHH sites within each read.
Read-level DRM2 CHH methylation activity is more processive than that of CMT2

To compare CMT2 and DRM2 methylation patterns, we used BS-seq data of mutants defective in the respective enzymes [14]. The mutants included 
cmt2 (three mutated alleles), and 
drd1 and 
rm2 (which both affect DRM2 activity [14, 29]). The assumption was that, given 
the complementary activity of these enzymes, the remaining methylation in 
cmt2 should consist of DRM2-methylated reads, whereas methylation in 
drd1 or 
rm2 should consist of 
CMT2-methylated reads [13, 14]. As the methylation pattern in the three 
cmt2 mutants was similar, we combined the methylation data of these samples.

Initially we tested whether the methylation level of individual reads differs between the 
mutants on a genomic scale. To do this, reads were collected from regions conforming to 
the standard parameters of our analysis (30bp max region span, 5 CHH sites). Reads were 
scored based on the methylation of the CHH sites from 0 to 5, removing reads that were 
not methylated at any site.

Figure 2a shows the distribution of read methylation level resulting from the above anal-
ysis in each of the mutants, selected from regions methylated at three different minimal 
levels (5%, 20% and 40%). Methylated reads from 
drd1 and 
rm2 show a different distri-
bution from the combined 
cmt2 data: while 
cmt2 retains a proportion of fully methylated 
reads, 
drd1 and 
rm2 retain mainly partially methylated reads, with a low proportion of 
fully methylated reads. This pattern is present even in regions with higher average methyla-
tion (Fig. 2a, rightmost panel). The difference between lowly- and highly methylated 
drd1 or 
rm2 regions is explained exclusively by the methylation state of partially methylated 
reads (Fig. S2a-b). In comparison, the proportion of partially methylated reads between 
low- and highly-methylated regions in 
cmt2 is similar, while methylation is correlated to 
the proportion of fully methylated reads (Fig. S2a). Due to the relatively low coverage 
of the 
rm2 mutant, we used 
drd1 for the subsequent analyses, but validated the patterns 
identified in 
drd1 using the 
rm2 mutant and other 
rm2 mutants.

Figure 2b demonstrates that reads from 
drd1 and 
cmt2 have different methylation levels. An additional dimension of variation among reads is the processivity of methylation. This is defined as the distribution of methylation within reads, and quantified by counting the 
number of changes in methylation within the read (e.g. a methylated CHH site adjacent 
to an unmethylated CHH site on the same strand) out of the total possible number of 
changes (illustrated in Figure 1b). Figure 2b demonstrates the separation between reads 
from the respective mutants according to their methylation level and processivity: while 
drd1 has reads with lower methylation and lower processivity, 
cmt2 has reads with higher methylation and higher processivity (i.e. low stochasticity). This correlation persist in 
regions with different CHH content, i.e. 5-15 sites per region (Fig. 2b), suggesting that 
this pattern does not depend on the density of CHH sites. This result is also consistent for 
the mutant alleles composing the 
cmt2 sample (i.e. cmt2-4 cmt2-5 and cmt2-6) and 
rm2 (Fig. S2c). Overall, these results suggest that DRM2 CHH methylation activity is more 
processive than that of CMT2.

CMT2 and DRM2 CHH methylation processivity are not dependent 
on demethylase activity

Arabidopsis DNA demethylases regulate DNA methylation levels through direct removing 
of methylated cytosine base from all cytosine sequence contexts [20, 31, 33]. Therefore,
Figure 2: Single-read analysis of CHH methylation in methylases and demethylase mutants. (a) Methylated reads from CHH methylation mutants: all reads with at least one methylated C are plotted, demonstrating differences in methylation intensity between \textit{cmt2} and \textit{drd1}/\textit{drm2}. Each panel presents regions filtered according to averaged methylation level. (b) Separation of reads originating from respective methylation mutants based on read methylation and methylation stochasticity (defined as the randomness of methylation distribution in the read). White areas indicate no data. (c-d) Comparison of demethylase read methylation and respective wild type samples: (c) sites methylated in the \textit{drd1} mutant (i.e. CMT2 methylated sites) that are hypermethylated in the demethylase mutants; (d) sites methylated in the \textit{cmt2} mutant (i.e. DRM2 methylated sites) that are hypermethylated in the demethylase mutants. To select CMT2- and DRM2-methylated sites \textit{ros1}-4 and \textit{rdd} mutants, the \textit{drd1} and \textit{cmt2} mutants from [14] were used, respectively. For \textit{dme}, \textit{drm12} and \textit{cmt2} mutants from [30] were used.
distinct patterns of CHH methylation in the DNMT mutants could result from demethylase activity. To test this hypothetical scenario, we analysed three different demethylase mutants: single mutants repressor of silencing (ros1-4), demeter (dme-2) and the triple mutant ros1-3, demeter-like protein 2 (dml2-1) and dml3-1 (rdd). Regions methylated in either drd1 or cmt2 were analysed as representing regions methylated by the complementary enzyme; the distribution of read methylation at these regions for each of the wild type/demethylase mutant pairs were plotted.

Figure 2c-d summarises this comparison. Figure 2c presents data from CMT2-methylated regions (methylated in drd1), while Figure 2d presents reads from DRM2-methylated regions (methylated in cmt2). For dme, drm2 and cmt2 mutants of vegetative nucleus tissue from [33] were used. For each demethylase, regions were also selected according to hypermethylation (> 10% increase in methylation) relative to the respective wild type sample. The enzyme-specific pattern is present in the demethylase mutants and its respective wild type sample (Fig. 2c-d): in CMT2 regions, both the mutant and wild type have a low proportion of fully methylated reads, with hypermethylation in the mutant correlating to the increase in partially methylated reads (Fig. S3a-c, left panels); in DRM2 regions, both the mutant and wild type have fully methylated reads, with hypermethylation in the mutant correlating to the increase in fully methylated reads (Fig. S3a-c, right panels). This suggests that the enzyme-specific pattern is present prior to demethylase activity; it also suggests that demethylases reduce methylation entirely in specific cells (given the decrease in zero-methylated reads in demethylase mutants).

Variation of CHH methylation among sites and reads distinguishes between CMT2 and DRM2 target regions

Analyses of individual reads can reflect the mechanism of different CHH-methylating enzymes, but the predictive confidence in distinguishing between methylated reads is limited, given that most reads are lowly methylated with a limited range of processivity (Fig. S2c). Hence, to characterise region methylation, we used methylation features that extract information from single-read data (Fig. 1b). The separation of all methylated regions (> = 10% average methylation) from the mutants is shown in Figure 3a-c. Paired with the distributions of features of the actual data (solid lines) are features of read datasets generated using a poisson model of single-C sites, where the chance of methylation per-site, per-read is equal to the mean region methylation (dashed lines). The similarity between the actual and generated data can demonstrate the degree to which the feature distributions of each mutant are explained by stochastic variation at the read level.

CMT2-methylated regions have higher variation among sites, lower variation among reads, and lower average processivity (as suggested by Figure 2b). The read methylation variation and processivity of CMT2-methylated regions overlap with the distributions of generated data (Fig. 3a-c), suggesting that, in the drd1 sample, variation of CMT2-mediated CHH-methylation activity among reads is mainly stochastic. By contrast, DRM2-methylated regions have low variation among sites, higher variation among reads, and higher average processivity (Fig. 3a-c). Variation among reads in DRM2-methylated regions is not stochastic, suggesting that DRM2-mediated CHH methylation in these regions is differentially regulated. Of these factors, variation among reads best predicts the methylating enzyme of the region (Table 1).

To understand the relationship between these methylation features in the mutants, features were analysed at the level of whole functional elements (in this case, TEs), by averaging
Figure 3: Separation of CHH methylated sites according to methylating enzyme. (a-c) K-density estimate plots of each type of variation in CHH methylation mutants. Actual distribution of data are shown in solid lines, dashed lines show the methylation features of a random poisson model based on the data of each sample. (d) Separation of whole transposable elements (TEs) according to among-site and read methylation variation in CHH methylation mutants (no minimal methylation level for regions or TEs). Linear regressions for each mutant drawn as solid lines along with squared correlation coefficients. P-value < $1 \times 10^{-10}$ for both regressions. (e) Differences between CWA and non-CWA subcontexts (CCH/CHY, where Y is C or T) in CHH methylation mutants. Lines represent the average read methylation sd. of a region for a given methylation level. (f) Resulting separation of regions from CHH methylation mutants based on the “enzyme score” feature. (g) Distribution of whole TEs according to enzyme score in CHH methylation mutants ($cmt2$ separated into its composing replicates). As opposed to (d), in (f) and (g) a minimal region methylation of 10% was used to filter regions. (h) Classifier receiver operating characteristic curve demonstrating separation between regions of CHH methylation mutants as shown in (f).
the features of individual regions contained within each element. This reduces noise caused by low coverage of individual regions. Only TEs with at least two regions with the required minimal coverage and methylation (>10%) are plotted. Read methylation variation and among-site variation were plotted in all TEs (Fig. 3d) and in specific TE superfamilies (Fig. S4a). In the \textit{drd1} mutant these features are correlated with a slope of 1, with among-site variation increasing linearly with read methylation variation (Fig. 3d). On the other hand, in the \textit{cmt2} mutant the two features are correlated with a smaller slope (0.21), with most TEs having a low average among-site variation (Fig. 3d). This was consistent in each of the TE superfamilies (Fig S4a).

\textit{CMT2} shows specificity for the CHH subcontext CWA (i.e. CTA or CAA) \cite{20, 21, 34}. As this could contribute to higher variation when analysing all CHH subcontexts, CWA and non-CWA subcontexts were analysed separately. For this comparison a larger region size of 50 bp was used, given the lower density of CWA sites (4.5 times lower than that of CHH). CWA contexts had higher variation among reads in \textit{drd1}, whereas in \textit{cmt2} these levels are comparable to all CHH subcontexts (Fig. 3d, S4a). Among-site variation remains similar. The increase in read methylation sd. can be explained by the higher methylation of CWA-methylated regions. In addition, CWA-methylated regions in \textit{drd1} have higher read methylation sd. relative to regions methylated at the same level in \textit{drd1}, but still lower than in \textit{cmt2} (Fig. S4b). In non-CWA and all CHH-sites \textit{drd1} read methylation variation is similar to that of stochastic methylation. Methylated reads from CWA-methylated regions show a similar pattern in terms of methylation level and methylation processivity, as opposed to non-CWA methylated regions (Fig. S4c-d). This suggests that the \textit{CMT2} methylation pattern observed in CHH-methylated sites in \textit{drd1} is composed of two distinct patterns; however, even at sites for which \textit{CMT2} shows specificity, it has relatively lower processivity compared to DRM2 (Fig S4b, left panel).

Based on ANOVA of the methylation features in the CHH methylation mutants, we designed a classifier to score regions and whole functional elements in terms of CHH methyla-
ting enzyme. The results of the model are presented in Table 1. The separation of the mutants used to define the classifier based on enzyme score is presented in Figure 3 (regions) and Figure 3c (whole elements), along with the ROC curve of region prediction (Fig. 3a). Enzyme score ranges from 0 to 1, with lower values indicating a similarity to DRM2 methylation, and higher values indicating a similarity to CMT2 methylation. Each mutant has a single enzyme score peak, and these peaks are aligned for mutants affecting the same enzyme (Fig. 3). The separation of the mutant samples in Figure 3b, and the ROC curve of the classifier demonstrate the potential of using the classifier to predict enzyme identity. \textit{A. thaliana cmt2} mutants from previous studies and \textit{drm2}-related mutants from multiple species show similar distributions of enzyme score, suggesting that this distribution is not specific to the mutant samples used to construct the classifier (Fig. S4b-c).

Table 1: CHH methylation mutant methylation features OLS model result ($r^2 = 0.700$).

|               | coefficient | std err | $t$   | $P > |t|$ |
|---------------|-------------|---------|-------|------|
| Intercept     | 0.3160      | 0.006   | 53.058| 0.000|
| Read methylation sd. | 2.412      | 0.015   | 164.994| 0.000|
| Stochasticity | -0.6728     | 0.007   | -95.785| 0.000|
| Among-site sd.| -1.7761     | 0.016   | -110.476| 0.000|
DRM2 and CMT2 methylation patterns are distinct from chromatin structure-dependent patterns

DRM2 and CMT2 function at distinct chromatin environments; DRM2 via RdDM is targeted mainly to euchromatic TEs, whereas CMT2 via H3K9me2 is targeted preferentially to heterochromatic TEs [13, 14, 29]. Accordingly, it is possible that DRM2’s and CMT2’s distinct CHH methylation activities are influenced by the genomic chromatin environment rather than by their intrinsic enzymatic activity. In order to test the role of genomic environment on DRM2 and CMT2 methylation patterns, we correlated enzyme score with GC content (a prominent indicator for chromatin structure [14, 18, 35]), for individual CHH-methylated regions in \textit{drd1} and \textit{cmt2} mutants (Fig. 4a). While regions methylated in \textit{drd1} showed on average higher GC content than regions methylated in \textit{cmt2}, no correlation was found between GC content and enzyme score in either mutant (Fig. 4a). We also correlated enzyme score in \textit{cmt2}, \textit{drd1} and \textit{drm2} with the following hetero- and eu-chromatic histone marks, H3K9me2 and H3K4me3, respectively [36]. As with GC content, the mutants are separated by both histone marks and enzyme score, but there is no correlation between histone marks and enzyme score within each mutant (Fig. 4b-c). These results suggest that processive and stochastic CHH methylation are intrinsic enzymatic activities of DRM2 and CMT2, respectively.

Plant and human DNMT3s show similar CHH methylation patterns to that of angiosperm CMT2

By identifying methylation patterns associated with either DRM2 or CMT2, the classifier can predict the presence or absence of the activity of either enzyme in samples from different tissues and species. Currently, in the absence of mutants showing partially reduced CHH methylation, it is unclear whether the total methylation pattern present in a given sample is derived from one or more CHH-methylating enzymes, and this distinction can not be made based on average methylation alone.

Mutants that possess only one active CHH methylating enzyme present a simplified case for the classifier, given that there is only one pattern. To assess the ability of the classifier to identify distinct patterns in more diverse samples, the classifier was applied to wild type samples of multiple species (Figure 5, with the distribution of TEs in individual features presented in Figure S5). Of the species analysed, \textit{A. thaliana}, \textit{Oryza sativa} and \textit{Solanum lycopersicum} had two peaks, while \textit{Physcomitrella patens} and \textit{Marchantia polymorpha} had only one peak. In \textit{A. thaliana}, the two peaks associate with each of the CHH methylation mutants in Figure S5. \textit{S. lycopersicum} and \textit{O. sativa} have two peaks, similarly to \textit{A. thaliana}, which are also aligned to the \textit{A. thaliana} CHH methylation mutants; however, the ratio between the peaks is different. This is also apparent in Figure S5x, both features are defined in a similar range but with different mean values. This ratio relates to the frequency of TEs regulated by either CMT2 or DRM2. For example, rice is known for its exceptional number of MITEs (130k) targeted by RNA directed DNA methylation (RdDM) and DRMs [38].

Both \textit{P. patens} and \textit{M. polymorpha} have a single peak that overlaps the \textit{A. thaliana} \textit{drd1} mutant (Fig. S5x). In addition, reads from these species have high stochasticity, similar to those from the \textit{drd1} mutant (Fig S6), suggesting similar processivity. \textit{P. patens} has one dominant CHH methylation enzyme, DNMT3, with trivial methylation activity by DRMs [16]. Finding a single enzyme distribution in \textit{P. patens} that is similar to that of CMT2, substantiates the trivial CHH methylation by PpDRMs and suggests that the CHH
methylation activity of DNMT3 is comparable to that of CMT2. Similar to *P. patens*, *M. polymorpha* contains DRM and DNMT3 and is missing CMT2 [39]. The classifier identified a single enzyme peak in the *M. polymorpha* methylome that overlaps that of PpDNMT3 and CMT2 (Fig. 5a), predicting that, similarly to *P. patens*, DNMT3 rather than DRMs are its main CHH methylases. Figure 5 suggests that *P. patens* and *M. polymorpha* read methylation variation is higher than that of *A. thaliana* cmt2, but lower than that of *A. thaliana* drd1. In addition, both *P. patens* and *M. polymorpha* have higher among-site variation compared to *A. thaliana* cmt2 (Fig. 5b). However, this difference relates partly to differences in methylation level of these samples: when comparing region features for a given region methylation level, *P. patens*, *M. polymorpha* and *A. thaliana* drd1 show more minor differences in among-site and read methylation variation (Fig. 5c-d).

While CG is the predominant methylation context in animals, non-CG methylation can be enhanced in particular tissues, such as the brain [3, 40]. Non-CG methylation in mammals (also called CH methylation) is mediated by DNMT3s. In human, two DNMT3s, DNMT3a and DNMT3b, were found to mediate CH methylation [3]. Thus, to test how many CH methylation patterns exist in human data and their relationship to plant ones, we next ran our single-read method on human methylome derived brain tissues. Applied to human CH data using the same parameters as for CHH analyses, our classifier detected a single peak of enzyme score that overlaps that of plant DNMT3 and CMT2 enzymes (Fig. 5e). Distributions of variations of among site and read methylation also show only a single peak of activity (Fig. 5f). These results suggest that CH methylation in neurons is mediated by a single dominant enzymatic pattern that is similar to that of plant DNMT3 and CMT2.

Conclusively, these results demonstrate the use of the enzyme classifier in predicting the presence or absence of CMT2- or DRM2-like methylating activity at a genomic scale.
Figure 5: Identification of CMT2 and DRM2 signatures in multiple species. (a) Distribution of TEs in multiple species according to enzyme score. Single peaks indicate the identification of only one enzyme signal; two peaks indicate the presence of two enzyme signals. (b) TE methylation variation features in *P. patens* and *M. polymorpha* that lack a peak aligned with DRM2 activity. (c-d) Differences of region among-site sd. and read methylation sd. among *M. polymorpha*, *P. patens* and *A. thaliana* mutant samples, with respect to region methylation: lines represent the average among-site sd. (c) or read methylation sd. (d) of a region for a given methylation level. All regions are plotted (i.e. without a minimal methylation level). (e) Analysis of enzyme score of exon CH methylation in human samples from frontal cortex [37]. (f) Methylation variation features of individual exons in human samples from (e).
Tissue-specific samples have different proportions of CMT2/DRM2-methylated regions

The enzyme classifier relies on methylation features the range of which may be biased by sample composition. For example, read methylation may vary less within homogeneous samples, if methylation patterns are similar between cells. Given that read methylation variation is the strongest predictor of enzyme identity ($r^2 = 0.589$), the effectiveness of the classifier may be limited in such samples.

In order to assess the ability of the classifier to function in tissue-specific samples, we used datasets from two studies that produced methylomes of sperm and vegetative nucleus cells [30], and root tissue subsamples [28]. Given that in each of these studies, altered regulation of CMT2/DRM2 activity was observed in one or more of the samples, this analysis also provided a means of validating the predictions of the classifier.

Figure 6a-b demonstrates the differences in CHH methylation regulation among tissue-specific samples of A. thaliana: all samples contain two peaks; however, in some tissues enzyme activity shifts, with both sperm and root tip having more DRM2-methylated TEs. drm12 and cmt2 mutants from the same study [30] were also analysed, and both vegetative nucleus and sperm mutants are distinguishable based on enzyme score analysis (Fig. 6c-d). As noted above, the wild type sperm sample has less CMT2 activity compared to the vegetative nucleus and other wild type A. thaliana samples analysed, confirming previous findings showing a reduced CHH methylation in heterochromatic TEs targeted by CMT2 [30].

Individual TEs and regions that are methylated in both sperm and vegetative nucleus are on average more similar to DRM2-methylated regions (Fig. S7a-b). In addition, relative to other DRM2-related mutants, the enzyme score peak of the sperm drm12 mutant is more DRM2-like (Fig. 6d). Interestingly, the sperm drm12 has higher read methylation variation, but lower among-site variation (Fig. 6b). This reflects the signal in the wild type sperm sample, in which overall among-site variation is low compared to other A. thaliana wild type tissues (Fig. 6b).

Compared to wild type sperm, drm12 sperm TEs are more regulated by CMT2 (Fig. 6f). This change is partly due to the loss of DRM2-methylated regions from TEs, rather than CMT2 methylating previously DRM2-methylated regions. However, the same change is observed also when comparing individual regions: regions retaining methylation in drm12 sperm, in particular regions that are defined by the classifier as regulated by DRM2 in the wild type sample, are shifted towards a CMT2-like signal (Fig. S7c, left panel). In the vegetative nucleus, the same shift is observed (Fig. S7d, left panel). In the cmt2 mutant, in both sperm and vegetative nucleus, no change is observed at the level of individual regions (Fig. S7c-d, right panels).

We also analysed three root samples that differed in their enzyme score distributions for whole TEs: root tip (RT), columella root cap (CRC), and lower columella (LC) [28]. The root tip sample contains both CRC and LC, but is in itself different from other wild type tissues analysed (Fig. 3h). Similarly to sperm, it has a lower average among-site variation. In terms of enzyme score, all three samples have two peaks (Fig. S7e), but the distributions are not overlapping. A comparison of enzyme score of individual TEs between samples shows that in the LC sample, all TEs are shifted towards a DRM2-like signal (Fig. S7f): TEs with intermediate signals (e.g. regulated by both enzymes) are more shifted than those with more defined DRM2/CMT2 signals (e.g. regulated by one enzyme). The same shift is present also in the CRC sample relative to RT (Fig. S7g). Overall, these results
Figure 6: Analysis of tissue-specific methylation patterns. (a) Distribution of TEs according to enzyme score in *A. thaliana* samples from seedling and specific tissues. (b) Separation of TEs in *A. thaliana* wild type tissues according to read and among-site CHH methylation variation. (c-d) Distribution of TEs according to enzyme score in *A. thaliana* vegetative nucleus (c) and sperm (d) wild type samples, alongside *drm12* and *cmt2* mutants of the respective tissues. (e-f) Correlation between enzyme score of individual TEs in wild type compared to mutant tissues. Only elements that were methylated in both wild type and respective mutant were selected. Each dot represents a single TE, with the distance from dashed line (*f(x) = x*) indicating a shift in enzyme score between samples (i.e. change in the average CHH enzyme methylating the element). (g) Schematic representation of the methylation patterns of CMT2 and DRM2 identified in this study: CMT2-mediated methylation has lower processivity and has a unimodal distribution among cells; DRM2-mediated methylation has higher processivity and has a bimodal distribution among cells.
demonstrate the ability of the classifier to predict changes in the activity level of DRM2 and CMT2 in different tissues.

**SRBrowse, a tool for visualising and analysing BS-seq data at single-read resolution**

Most popular genome browsers, including UCSC genome browser and Integrative Genomics Viewer (IGV) [41], are limited in their ability to load high resolution data on-the-fly without creating a large memory footprint, significantly increasing the load times of browser displays or requiring pre-processing of data. In order to visualise single read data, it is necessary to convert aligned read data (e.g. SAM/BAM files) into track files (such as GFF) or compressed indexed files (BED, TDF, etc.) suitable for fast retrieval.

In order to make browsing and analysing BS-seq data at single-read resolution more accessible, we developed a genome browser specifically designed for visualising BS-seq data at a single-read level, which we called SRBrowse. SRBrowse is web browser-based and can run on a local computer or server with minimal software requirements (see the repository README file on installation). Importantly, SRBrowse allows users to load data into the browser view and monitor its alignment progress from the same interface. Typical steps for loading and displaying data appear in Figure S8. All data loaded through SRBrowse is aligned using bowtie2 and stored in indexed files allowing fast access to the read data.

**Discussion**

We presented a novel analysis pipeline for extracting additional layers of information from NGS BS-seq data. The pipeline uses data of individual read methylation states and the distribution of DNA methylation within reads to identify patterns that augment information regarding the averaged methylation signal. Using this pipeline, we were able to define characteristic features of reads and regions methylated by two non-CG methylating enzymes CMT2 and DRM2 in *A. thaliana*. Specifically, we found that *A. thaliana* mutants of CMT2 and DRM2 present stereotypical CHH methylation patterns that are robust to background methylation and consistent among different mutated alleles and species (Fig. 5, S6). These patterns are also independent of demethylase-activity: in the absence of demethylase activity, the same distinct patterns are observed in regions regulated by each enzyme (Fig. 2c-d, S3). On the one hand, cmt2 mutants have mainly highly methylated reads, and methylation is concentrated within specific regions; on the other hand, drm2 and drd1 mutants have mainly lowly methylated reads, and methylation is distributed stochastically within and among reads. In other words, our analysis suggests that, compared to DRM2, CMT2 has lower processivity and presents more stochastic variation of methylation level among cells. In contrast, DRM2-methylated regions present distinct subpopulations of methylation states, with less stochastic variation (Fig. 6g).

Variation in DRM2 and CMT2 methylation characteristics could relate to the distinct genomic targets of these enzymes. DRM2 methylates mostly short euchromatic-TE sequences located next to genes and CMT2 methylates mainly long heterochromatic-TE sequences [13, 14]. Thus, DRM2 subpopulations of highly processive methylation and unmethylation states could relate to the ability of DRM2 methylation to regulate genes within particular cell types or under certain conditions (Fig. 6k), for example in the formation of lateral root development [42]. In contrast, CMT2 low but uniform methylation correlates with constant
need to silence heterochromatic TEs (Fig. 6g). Furthermore, given that we found no correlation between chromatin structure markers and CMT2 and DRM2 single read methylation patterns (Fig. 4), this suggests that the biological significance of CMT2 and DRM2 distinct methylation patterns are determined by their intrinsic enzymatic characteristics rather than by the chromatin environment.

Based on the variation of these patterns between CHH-methylating mutants, we designed a classifier that scores short regions of 30 base pairs and collections of regions within functional elements (such as genes, exons or transposable elements). This score provides an arbitrary scale to differentiate between DRM2-like and CMT2-like CHH methylation patterns. The comparison among species highlights the ability of the classifier to predict the presence or absence of CMT2/DRM2 in species for which mutants have not yet been developed, such as *M. polymorpha*. The classifier is robust to differences in sample heterogeneity, and is able to differentiate between methylation patterns even within highly specific tissue samples (Fig. 6).

While DRM2 and plants as well as human DNMT3 are monophyletic and distantly related to CMT2, DRM2 is the only enzyme that has a rearranged catalytic domain [16]. Therefore, our findings that DNMT3 and CMT2 have similar CHH methylation characteristics suggest that different DNMTs can have similar methylation mechanisms, and substantiates the hypothesis that DNMT3 CHH methylation activity in early land plants has been replaced by CMT2 in angiosperm [16]. Moreover, the DRM2 unique high-processivity methylation activity can be associated to its exclusive rearranged catalytic domain rather than to its general homology to DNMT3 enzymes.

**Conclusion**

Overall, the analyses of methylation profiles we present here demonstrate the potential of studying patterns of variation in BS-seq data, through single-read analysis, which provide new biological insights on the writing, erasing, and readout mechanisms of CHH methylation. The tool we developed can facilitate further studies of methylomes at single-read resolution.

**Methods**

**BS-seq alignment**

All code used for the read analysis pipeline will be deposited in a publicly available code repository (https://github.com/zemachlab/srbrowse). For aligning reads from BS-seq data we used bowtie2 with a Node.js-based wrapper. The method we used for aligning BS-seq data is based on a previously described pipeline [14]. The wrapper converts the reference assembly to C-to-T and G-to-A sequences before bowtie2 indexing; each strand is converted manually so that each index consists both of forward and reverse strand versions of each scaffold. BS-seq reads are converted either C-to-T or G-to-A depending on whether the read is a left or right mate (in the case of paired ends reads), with the original read data stored for collecting methylation information after alignment. Bowtie2 was run with end-to-end search algorithm. For all datasets a minimum score of 0 was used (i.e. no mismatches or gaps). Reads that mapped to more than one position were discarded. Aligned reads were then sorted and exact duplicates removed.
Analyses of BS-seq data

Analyses consist of three stages: (1) identifying short regions (10-100bp) according to the region selection parameters (see 2.1); (2) extracting reads overlapping with each region from the selected samples; and (3) averaging read data. Region selection parameters were optimised to ensure sufficient data for low coverage samples (Fig. S1a-c). Functional elements are first selected according to an annotation provided in gff/gff3 format. Next, sites of specific methylation contexts are identified based on the reference sequence of the element (e.g. CHH sites), per strand. Separation to strands is important for asymmetrical contexts such as CHH. Regions are defined by iterating through sites until N sites within the defined region size are found. Reads from aligned BS-seq samples are retrieved from indexed lists of reads, and stored as binary arrays where each site is represented as either unmethylated (0) or methylated (1).

The read methylation data of a specific region were analysed to produce methylation features (Fig. 1b) of individual reads and their associated regions. For individual reads these features are (1) read methylation, the mean of the vector components, and (2) read stochasticity, the number of changes between methylation states between adjacent sites (for illustration, see Figure 1b). Importantly, features of individual reads do not refer to the overall methylation of the read, but only to methylation at the sites included in the specific region. For regions, the methylation features are (1) mean read methylation; (2) standard deviation of read methylation; (3) mean read stochasticity; and (4) standard deviation of site methylation. The last feature is not based on single-read data but rather the averaged methylation signal at each site.

The output data of this analysis can be either at the level of individual reads, individual regions, or whole functional elements. For reads data, the output is an array of reads for each sample from all regions matching the selection parameters. For regions, the output is an array of regions with features derived from averaged read vectors as explained above. The regions also have positional data relating to their parent element, length, and any other genomic features of interest (GC content, etc.). For whole functional elements, the output is an array of elements such as exons or transposable elements, where methylation features relate to the average of all regions identified with the functional element. The exclusion of regions based on methylation or coverage prior to averaging whole elements is important to reduce background of unmethylated regions. Unless stated otherwise, regions were selected with a minimum of 10% methylation average and 4 overlapping reads. Functional elements that contained at least two such regions were selected. While increasing the minimum regions per element improves coverage per element, it can bias the analysis to longer elements.

Statistical analyses

All statistical analyses we performed on either read, region or element data resulting from the above pipeline, using Python 3.6 along with the following libraries: matplotlib, numpy, scipy, statsmodels, pandas and seaborn. K-density plots were produced using seaborn.distplot (which implements statsmodels.nonparametric.kde.KDEUnivariate) with gaussian kernel shape and Scott’s Rule of Thumb bandwidth. For ANOVA of methylation features we used methylated regions from the CHH methylation mutants drd1 and cmt2 (composed of data from cmt2-4, cmt2-5 and cmt2-6 mutant alleles). Methylation features of the regions were provided as independent variables, and sample source (0 for drd1, 1 for cmt2) as the dependent variable. The results of the ordinary least squares model are presented in Table S1. For the classifier, the resulting coefficients of the independent variables were scaled
so that enzyme score is defined between 0 and 1. Linear regression for scatter plots were conducted using the scipy.stats.linregress function.

Data sources

The following assemblies were used for aligning reads: GCF_000001735.4 (A. thaliana), GCF_000002425.4 (P. patens), O. sativa v7.0, GCF_000188115.4 (S. lycopersicum), M. polymorpha v3.0, GCF_000001405.39 (Homo sapiens). The following annotations were used for genes and TEs: Araport 11 TE annotation from from TAIR [43] for A. thaliana; P. patens TE annotation was downloaded from CoGe, and information from a P. patens Repeatmasked assembly (v3.3) was downloaded from Phytozome to increase the resolution of LTR-TEs families; TEs were annotated de-novo for M. polymorpha using REPET v3.0 [44, 45]; Repeatmasked assemblies were downloaded from Phytozome for O. sativa (323) [46] and S. lycopersicum (ITAG 3.2) [47]; GCA_000001405.28 gene annotation [48] was used for H. sapiens.

Whole genome BS-seq data from the following studies was used (for a full list of accesses see supplementary data): GSE41302 for A. thaliana cmt2, drm2, drd1 mutants [14], GSE64569 for A. thaliana ros1 mutants [19], GSE33071 for A. thaliana rdd triple mutant [50], GSE38935 for A. thaliana dme mutants [51], GSE87170 for A. thaliana sperm and vegetative nucleus wild type and cmt2 and drm12 mutants [30], GSE79746 for A. thaliana drm2 and cmt2 mutants [52], GSE39901 for A. thaliana cmt2 mutant [29], GSE43857 for A. thaliana ecotypes Gro-3, Kz-9 and Neo-6 [53], PRJNA350766 and GSE118153 for P. patens wild type samples [16, 54], SRP101412 for M. polymorpha wildtype (thallus) samples [39], GSE81436 for O. sativa wild type sample [38], GSE108527 for O. sativa drm2 ddm1 mutant [55], SRP008329 for S. lycopersicum wild type sample [56], SRP081115 for S. lycopersicum slrpd1 mutant [34], and GSE47966 for H. sapiens frontal cortex samples [37].

Acknowledgements

We would like to thank Ohad Roth for contributing to an early draft of the manuscript. We would also like to thank all members of the Zemach lab and Nir Ohad’s lab for their critical feedback on the study.

References

[1] Huiming Zhang, Zhaobo Lang, and Jian-Kang Zhu. “Dynamics and function of DNA methylation in plants”. In: Nature Reviews Molecular Cell Biology 19.8 (2018), pp. 489–506.
[2] Jason T Huff and Daniel Zilberman. “Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes”. In: Cell 156.6 (2014), pp. 1286–1297.
[3] Yupeng He and Joseph R Ecker. “Non-CG methylation in the human genome”. In: Annual review of genomics and human genetics 16 (2015), pp. 55–77.
[4] Robert J Schmitz, Zachary A Lewis, and Mary G Goll. “DNA methylation: shared and divergent features across eukaryotes”. In: Trends in Genetics (2019).
[5] Assaf Zemach et al. “Genome-wide evolutionary analysis of eukaryotic DNA methylation”. In: Science 328.5980 (2010), pp. 916–919.
[6] Suhua Feng et al. “Conservation and divergence of methylation patterning in plants and animals”. In: Proceedings of the National Academy of Sciences 107.19 (2010), pp. 8689–8694.

[7] Adam J Bewick et al. “Diversity of cytosine methylation across the fungal tree of life”. In: Nature ecology & evolution 3.3 (2019), pp. 479–490.

[8] Adam J Bewick et al. “Evolution of DNA methylation across insects”. In: Molecular biology and evolution 34.3 (2017), pp. 654–665.

[9] Ryan Lister et al. “Human DNA methylomes at base resolution show widespread epigenomic differences”. In: nature 462.7271 (2009), pp. 315–322.

[10] Mary Grace Goll and Timothy H Bestor. “Eukaryotic cytosine methyltransferases”. In: Annu. Rev. Biochem. 74 (2005), pp. 481–514.

[11] Julie A Law and Steven E Jacobsen. “Establishing, maintaining and modifying DNA methylation patterns in plants and animals”. In: Nature Reviews Genetics 11.3 (2010), pp. 204–220.

[12] Jiamu Du et al. “Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants”. In: Cell 151.1 (2012), pp. 167–180.

[13] Hume Stroud et al. “Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis”. In: Nature structural & molecular biology 21.1 (2014), p. 64.

[14] Assaf Zemach et al. “The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin”. In: Cell 153.1 (2013), pp. 193–205.

[15] Julie A Law et al. “SHH1, a homeodomain protein required for DNA methylation, as well as RDR2, RDM4, and chromatin remodeling factors, associate with RNA polymerase IV”. In: PLoS genetics 7.7 (2011).

[16] Rafael Yaari et al. “RdDM-independent de novo and heterochromatin DNA methylation by plant CMT and DNMT3 orthologs”. In: Nature communications 10.1 (2019), pp. 1–10.

[17] Dirk Schubeler. “Function and information content of DNA methylation”. In: Nature 517.7534 (2015), pp. 321–326.

[18] Jaemyung Choi et al. “DNA methylation and histone H1 cooperatively repress transposable elements and aberrant intragenic transcripts”. In: bioRxiv (2019), p. 527523.

[19] Marianne Frommer et al. “A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.” In: Proceedings of the National Academy of Sciences 89.5 (1992), pp. 1827–1831.

[20] Ryan Lister et al. “Highly integrated single-base resolution maps of the epigenome in Arabidopsis”. In: Cell 133.3 (2008), pp. 523–536.

[21] Shawn J Cokus et al. “Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning”. In: Nature 452.7184 (2008), pp. 215–219.

[22] Sébastien A Smallwood et al. “Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity”. In: Nature methods 11.8 (2014), pp. 817–820.

[23] Hongshan Guo et al. “Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing”. In: Genome research 23.12 (2013), pp. 2126–2135.
[24] Yanchang Wei et al. “DNA methylation analysis and editing in single mammalian oocytes”. In: *Proceedings of the National Academy of Sciences* 116.20 (2019), pp. 9883–9892.

[25] Chongyuan Luo et al. “Robust single-cell DNA methylome profiling with snmC-seq2”. In: *Nature communications* 9.1 (2018), pp. 1–6.

[26] Qing Huan et al. “HeteroMeth: a database of cell-to-cell heterogeneity in DNA methylation”. In: *Genomics, proteomics & bioinformatics* 16.4 (2018), pp. 234–243.

[27] Chad E Niederhuth et al. “Widespread natural variation of DNA methylation within angiosperms”. In: *Genome biology* 17.1 (2016), p. 194.

[28] Taiji Kawakatsu et al. “Unique cell-type-specific patterns of DNA methylation in the root meristem”. In: *Nature plants* 2.5 (2016), pp. 1–8.

[29] Hume Stroud et al. “Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome”. In: *Cell* 152.1-2 (2013), pp. 352–364.

[30] Ping-Hung Hsieh et al. “Arabidopsis male sexual lineage exhibits more robust maintenance of CG methylation than somatic tissues”. In: *Proceedings of the National Academy of Sciences* 113.52 (2016), pp. 15132–15137.

[31] Yeonhee Choi et al. “DEMETER, a DNA glycosylase domain protein, is required for endosperm gene imprinting and seed viability in Arabidopsis”. In: *Cell* 110.1 (2002), pp. 33–42.

[32] Zhizhong Gong et al. “ROS1, a repressor of transcriptional gene silencing in Arabidopsis, encodes a DNA glycosylase/lyase”. In: *Cell* 111.6 (2002), pp. 803–814.

[33] Tzung-Fu Hsieh et al. “Genome-wide demethylation of Arabidopsis endosperm”. In: *Science* 324.5933 (2009), pp. 1451–1454.

[34] Quentin Gouil and David C Baulcombe. “DNA methylation signatures of the plant chromomethyltransferases”. In: *PLoS genetics* 12.12 (2016), e1006526.

[35] Eran Segal et al. “A genomic code for nucleosome positioning”. In: *Nature* 442.7104 (2006), pp. 772–778.

[36] François Roudier et al. “Integrative epigenomic mapping defines four main chromatin states in Arabidopsis”. In: *The EMBO journal* 30.10 (2011), pp. 1928–1938.

[37] Ryan Lister et al. “Global epigenomic reconfiguration during mammalian brain development”. In: *Science* 341.6146 (2013), p. 1237905.

[38] Feng Tan et al. “Analysis of chromatin regulators reveals specific features of rice DNA methylation pathways”. In: *Plant physiology* 171.3 (2016), pp. 2041–2054.

[39] Marc W Schmid et al. “Extensive epigenetic reprogramming during the life cycle of Marchantia polymorpha”. In: *Genome biology* 19.1 (2018), p. 9.

[40] Keith D Harris et al. “DNA methylation is maintained with high fidelity in the honey bee germline and exhibits global non-functional fluctuations during somatic development”. In: *Epigenetics & chromatin* 12.1 (2019), p. 62.

[41] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. “Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration”. In: *Briefings in bioinformatics* 14.2 (2013), pp. 178–192.
[42] Zaigham Shahzad et al. “Cryptic variation in RNA-directed DNA-methylation controls lateral root development when auxin signalling is perturbed”. In: *Nature Communications* 11.1 (2020), pp. 1–11.

[43] Tanya Z Berardini et al. “The Arabidopsis information resource: making and mining the “gold standard” annotated reference plant genome”. In: *genesis* 53.8 (2015), pp. 474–485.

[44] Hadi Quesneville et al. “Combined evidence annotation of transposable elements in genome sequences”. In: *PLoS computational biology* 1.2 (2005).

[45] Timothée Flutre et al. “Considering transposable element diversification in de novo annotation approaches”. In: *PloS one* 6.1 (2011).

[46] Shu Ouyang et al. “The TIGR rice genome annotation resource: improvements and new features”. In: *Nucleic acids research* 35.suppl1 (2007), pp. D883–D887.

[47] Tomato Genome Consortium et al. “The tomato genome sequence provides insights into fleshy fruit evolution”. In: *Nature* 485.7400 (2012), p. 635.

[48] Fiona Cunningham et al. “Ensembl 2019”. In: *Nucleic acids research* 47.D1 (2019), pp. D745–D751.

[49] June-Sik Kim et al. “ROS1-dependent DNA demethylation is required for ABA-inducible NIC3 expression”. In: *Plant physiology* 179.4 (2019), pp. 1810–1821.

[50] Weiqiang Qian et al. “A histone acetyltransferase regulates active DNA demethylation in Arabidopsis”. In: *Science* 336.6087 (2012), pp. 1445–1448.

[51] Christian A Ibarra et al. “Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes”. In: *Science* 337.6100 (2012), pp. 1360–1364.

[52] Kaushik Panda et al. “Full-length autonomous transposable elements are preferentially targeted by expression-dependent forms of RNA-directed DNA methylation”. In: *Genome biology* 17.1 (2016), p. 170.

[53] Taiji Kawakatsu et al. “Epigenomic diversity in a global collection of Arabidopsis thaliana accessions”. In: *Cell* 166.2 (2016), pp. 492–505.

[54] Daniel Lang et al. “The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution”. In: *The Plant Journal* 93.3 (2018), pp. 515–533.

[55] Feng Tan et al. “DDM1 represses noncoding RNA expression and RNA-directed DNA methylation in heterochromatin”. In: *Plant physiology* 177.3 (2018), pp. 1187–1197.

[56] Silin Zhong et al. “Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening”. In: *Nature biotechnology* 31.2 (2013), p. 154.