TUM School of Life Sciences
Technische Universität München

A computational framework for studying epimutational processes in plants

Yadollah Shahryary Dizaji

Complete reprint of the dissertation approved by the TUM School of Life Sciences of the Technical University of Munich for the award of the

Doktor der Naturwissenschaften (Dr. rer. nat.).

Chair:
Prof. Dr. Aurélien Tellier

Examiners:
1. Prof. Dr. Frank Johannes
2. Prof. Dr. Markus List

The dissertation was submitted to the Technical University of Munich on 29.04.2024 and accepted by the TUM School of Life Sciences on 03.08.2024.

In memory of my father

Acknowledgments

I would like to express my sincere gratitude to my adviser, Prof. Dr. Frank Johannes for his continuous support during my PhD research, for his patience with me, for constantly motivating me during the research, and for his enthusiasm for my projects. I have benefited tremendously from his knowledge and guidance, without which it would not have been possible for me to complete PhD program. I could not have hoped for a better adviser and mentor than Prof. Dr. Frank Johannes.

I would also like to thank my mentor Prof. Dr. Maria Colomé-Tatché for her guidance during my research and Prof. Bob Schmitz for his helpful scientific inputs. My gratitude also goes to the "Technical University of Munich- Institute for Advanced Study" and Deutsche Forschungsgemeinschaft (DFG) for funding this work within the research no. 291763 and Sonderforschungsbereich924.

I would like to thank my colleagues who gave me the opportunity to discuss my works, and for the great working atmosphere, scientific discussions, and engaging coffee breaks, especially Dr. Rashmi Hazarika, for her advice and support during my PhD projects. I also wish to acknowledge the help provided by Susanna Fink for her administrative assistance.

I am particularly grateful to my family for their support encouragement. I do not have the words to tell you how truly fortunate I feel to have you in my life.

Finally, I would like to express my profound gratitude to my wife, Fahimeh, for her unwavering patience and encouragement throughout the process of completing my thesis. Her support was invaluable, especially during times when the tasks felt overwhelmingly challenging. Fahimeh, your strength and encouragement have meant the world to me. I love you.

Abstract

DNA methylation, an important chromatin modification, plays crucial roles in silencing transposable elements and regulating certain genes. However, the methylation status of individual cytosines or clusters of cytosines can sometimes change stochastically, leading to what is termed "spontaneous epimutations". These epimutations can accumulate during plant development and aging, and some even pass through the gametes to subsequent generations. A goal in the field of plant epigenetics is to obtain accurate estimates of the rate of spontaneous epimutations, identify genetic and environmental factors that can modulate the rate, and to delineate the molecular mechanisms underlying epimutational processes.

The overarching goal of this thesis is to develop a computational framework to facilitate quantitative insights into epimutational processes in plants. My approach to this problem is to implement a stochastic model within a computational workflow that starts with pedigree-based whole genome bisulfite sequencing (WGBS) data and ends with statistical estimates of the spontaneous epimutation rate.

To achieve this, my first aim is to develop MethylStar, a bioinformatic pipeline for the high-throughput analysis of WGBS datasets. MethylStar is a fast, stable and flexible preprocessing pipeline for WGBS data. MethylStar integrates well-established tools for read trimming, alignment and methylation state calling in a highly parallelized environment, manages computational resources and performs automatic error detection. It offers easy installation through a dockerized container with all preloaded dependencies and also features a user-friendly interface designed for experts/non-experts. Application of MethylStar to various WGBS datasets demonstrates favorable performance in terms of speed and memory requirements compared with existing pipelines.

My second aim is to develop AlphaBeta, a stochastic model that takes pedigree-based WGBS data as input to estimate epimutation rates. AlphaBeta starts with base-level or region-level DNA methylation state calls for each of the samples in the pedigree. AlphaBeta fits an explicit epimutation model to the DNA methylation divergence data, and relates this information to the temporal divergence of the samples, as calculated from the pedigree topology. I show that the software can be applied to data from multi-generational mutation accumulation lines, derived either through sexual or clonal propagation. Furthermore, I demonstrate that AlphaBeta can also be used to estimate somatic epimutation rates in long-lived perennials, such as trees. In this case, AlphaBeta interprets the tree branching topology as a phylogeny of somatic cell lineages with the leaves representing the end-points of these lineages. The software calculates DNA

methylation divergence between leaves and relates this information to their temporal divergence, as determined from coring data on branch/stem ages.

Application of MethylStar and AlphaBeta to published and new data reveals that spontaneous epimutations accumulate neutrally at the genome-wide scale, originate mainly during somatic development and that they can be used as a molecular clock for age-dating trees.

Zusammenfassung

DNA-Methylierung, eine wichtige Chromatinmodifikation, spielt eine entscheidende Rolle bei der Stilllegung von Transposons und der Regulation bestimmter Gene. Allerdings kann sich der Methylierungsstatus einzelner Cytosine oder Cytosincluster manchmal stochastisch ändern, was als "spontane Epimutationen" bezeichnet wird. Diese Epimutationen können sich während der Entwicklung und Alterung von Pflanzen ansammeln, und einige gelangen sogar über die Gameten in nachfolgende Generationen. Ein Ziel im Bereich der Pflanzenepigenetik ist es, genaue Schätzungen der Rate spontaner Epimutationen zu erhalten, genetische und umweltbedingte Faktoren zu identifizieren, die die Rate modulieren können, und die molekularen Mechanismen zu erläutern, die epimutationalen Prozessen zugrunde liegen.

Das übergeordnete Ziel dieser Dissertation ist die Entwicklung eines rechnergestützten Rahmens, um quantitative Einblicke in epimutationale Prozesse bei Pflanzen zu ermöglichen. Mein Ansatz für dieses Problem besteht darin, ein stochastisches Modell innerhalb eines rechnergestützten Workflows zu implementieren, der mit auf Stammbaumdaten basierender whole genome bisulfite sequencing (WGBS) beginnt und mit statistischen Schätzungen der spontanen Epimutationsrate endet.

Um dies zu erreichen, ist mein erstes Ziel die Entwicklung von MethylStar, einer bioinformatischen Pipeline für die Hochdurchsatzanalyse von WGBS-Datensätzen. MethylStar ist eine schnelle, stabile und flexible pre-processing für WGBS-Daten. MethylStar integriert etablierte Werkzeuge für das Trimmen von Reads, die Ausrichtung und die Klassifizierung des Methylierungszustands in in einem hochparallelisierten Rahmen. Verwaltet Rechenressourcen und führt automatische Fehlererkennung durch. Es bietet eine einfache Installation durch einen dockerisierten Container mit allen vorgeladenen Abhängigkeiten und verfügt auch über eine benutzerfreundliche Schnittstelle, die für Experten/Nicht-Experten konzipiert ist. Die Anwendung von MethylStar auf verschiedene WGBS-Datensätze zeigt eine sehr gute Leistung in Bezug auf Geschwindigkeit und Speicheranforderungen im Vergleich zu bestehenden Pipelines.

Mein zweites Ziel ist die Entwicklung von AlphaBeta, einem stochastischen Modell, das auf Stammbaumdaten basierende WGBS-Daten verwendet, um Epimutationsraten zu schätzen. AlphaBeta beginnt der Klassifizierung des DNA-Methylierungszustands auf Nukleotid oder Regionsebene für jede der Proben im Stammbaum. AlphaBeta passt ein explizites Epimutationsmodell an die DNA-Methylierungsdivergenzdaten an und stellt diese Informationen in Bezug zur zeitlichen Divergenz der Proben, wie sie aus der Stammbaumtopologie berechnet wird. Ich zeige, dass die Software auf Daten von über mehrere Generationen gehenden Mutationsakkumulationslinien angewendet werden

kann, die entweder durch sexuelle oder klonale Vermehrung abgeleitet wurden. Darüber hinaus demonstriere ich, dass AlphaBeta auch zur Schätzung somatischer Epimutationsraten bei langlebigen mehrjährigen Pflanzen wie Bäumen verwendet werden kann. In diesem Fall interpretiert AlphaBeta die Baumverzweigungstopologie als eine Phylogenie somatischer Zelllinien, wobei die Blätter die Endpunkte dieser Linien darstellen. Die Software berechnet die DNA-Methylierungsdivergenz zwischen den Blättern und stellt diese Informationen in Bezug zu ihrer zeitlichen Divergenz, wie sie aus Kernbohrungsdaten über das Alter von Ästen/Stämmen bestimmt wird.

Die Anwendung von MethylStar und AlphaBeta auf veröffentlichte und neue Daten zeigt, dass spontane Epimutationen auf genomweiter Skala neutral akkumulieren, hauptsächlich während der somatischen Entwicklung entstehen und als molekulare Uhr zur Altersbestimmung von Bäumen verwendet werden können.

Contents

1 Introduction

1.1 DNA methylation in Plants

DNA cytosine methylation (mC) represents a crucial nucleotide modification, prevalent in the majority of plant genomes. Its primary functions include the repression of transposable elements (TEs) and repeated sequences, along with playing a role in gene regulation [1]. Plants undergo cytosine methylation at both symmetrical CG and CHG sites, as well as at asymmetrical CHH sites, with H representing A, T, or C. The mechanisms responsible for initiating and preserving methylation across these distinct sequence contexts are thoroughly understood [2] (Fig.1.1), and demonstrate a broad level of conservation among various plant species [3–5].

The de novo establishment of methylation across all three sequence types is predominantly driven by the RNA-directed DNA methylation (RdDM) pathway. This process involves the use of 24-nucleotide (nt) small RNAs (sRNAs) serving as guiding entities for the action of DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2). For sRNA production and their precise targeting to specific genomic regions, RdDM is dependent on two RNA polymerases, Pol IV and Pol V (reviewed in Law Jacobsen, [1]; Matzke et al., [6]). Beyond their involvement in heterochromatin formation, sRNAs play a critical role in the modulation of gene activity through both transcriptional and post-transcriptional silencing strategies. They are also integral to various aspects of plant development, reproduction, and the capacity for phenotypic adaptation [7].

A variety of specific pathways play a role in maintaining DNA methylation once it has been established. Within a CG sequence, proteins from the VARIANT IN METHYLATION (VIM1) family detect hemimethylated CG sites, subsequently attracting METHYLTRANSFERASE (MET1) to perform CG methylation on the new strand through template duplication. Disrupting MET1 leads to a total loss of CG methylation across the genome [8]. Predominantly, CHG methylation is preserved by the unique plant methyltransferase CHROMOMETHYLASE 3 (CMT3), which establishes a reinforcing feedback loop with both histone H3 lysine 9 dimethylation (H3K9me2) and the histone methyltransferase SUVH4 [9]. Additionally, at certain CHG and CHH locations, CMT2, which also necessitates H3K9me2, sustains methylation through de novo activity [1]. Examination of diverse DNA methylation mutants has revealed extensive interactions among these distinct pathways [10, 11].

In addition to DNA methyltransferases, the regulation of DNA methylation patterns involves the crucial role of the chromatin remodeler DECREASE IN DNA METHYLATION 1 (DDM1). DDM1 indirectly influences methylation by facilitating access for

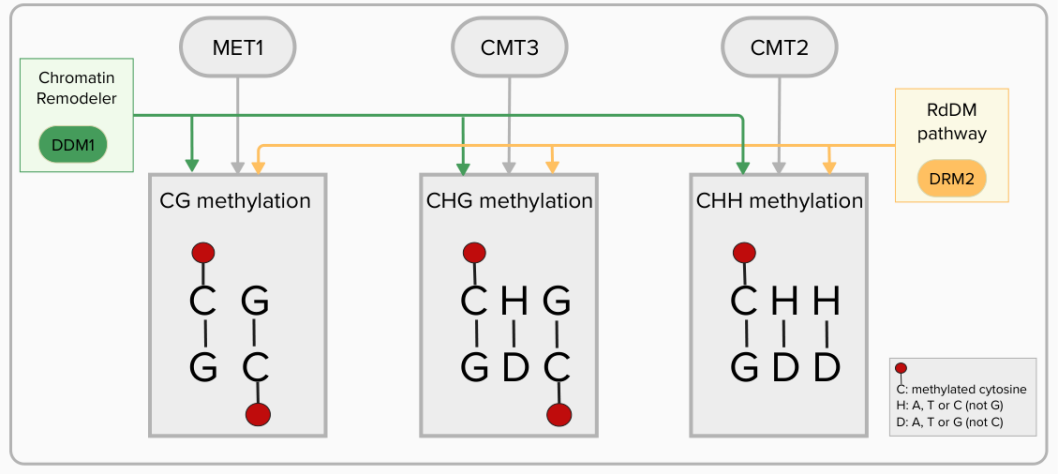

Figure 1.1: Overview of DNA methylation pathways and their targets contexts. Figure source: modified from Robert M. et al. [12]

DNA methyltransferases to H1-containing histones, primarily within long transposable elements (TEs) found in heterochromatic regions [13]. The loss of DDM1 leads to a significant decrease in DNA methylation across all three sequence contexts, accompanied by a widespread overaccumulation of TE-related transcripts [14–16]. DDM1 also plays a vital role in maintaining histone H3 methylation patterns. The loss of DNA methylation correlates with the replacement of methylation at lysine 9 with methylation at lysine 4 [17, 18], which aligns with the transcriptional activation of regions that are otherwise repressed. These findings emphasize the intricate interplay between DDM1, DNA methylation, and histone modifications in shaping the epigenetic landscape and gene expression dynamics.

1.2 Measuring DNA methylation using whole genome bisulphite sequencing (WGBS)

The elucidation of DNA methylation pathways in plants has advanced significantly due to the capability to decipher DNA methylation at single-base resolution with high-throughput techniques. Whole genome bisulfite sequencing (WGBS) stands out as the benchmark procedure for such analyses. This technique encompasses the exposure of genomic DNA (gDNA) to sodium bisulfite, followed by subsequent next-generation sequencing. During this process, unmethylated cytosines undergo conversion to uracils due to the action of bisulfite, which are ultimately read as thymidines in the sequencing phase. Conversely, methylated cytosines remain unaffected by the treatment, allowing them to be identified as cytosines in the resulting sequence data. Utilizing a range of

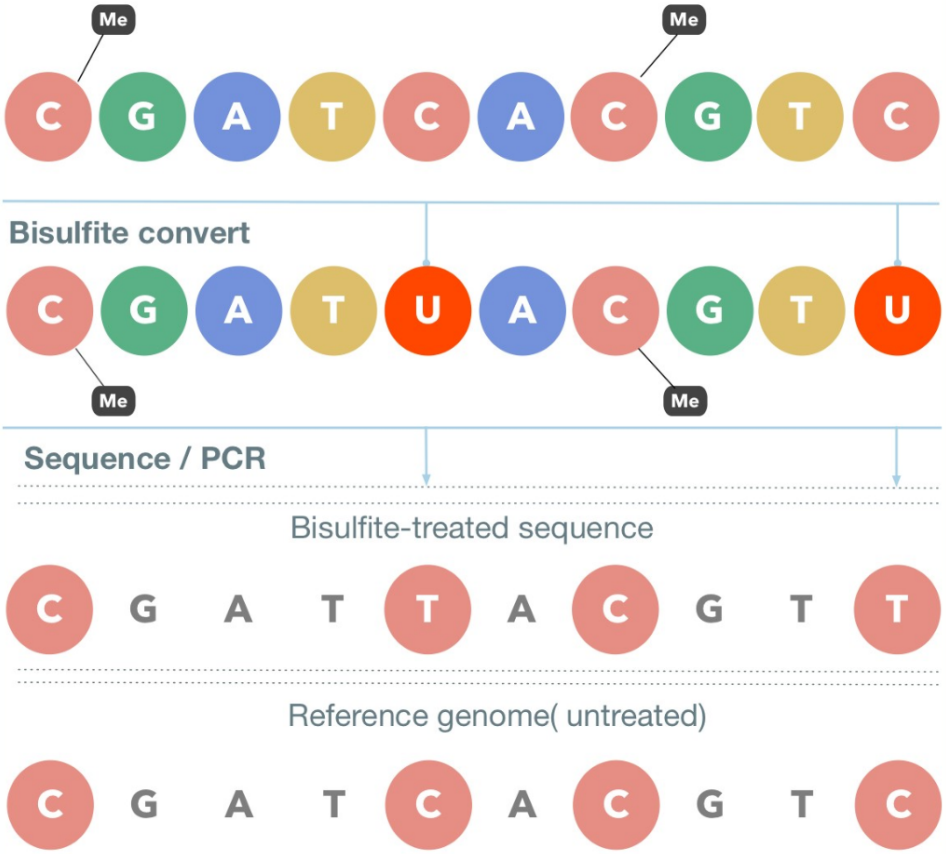

Figure 1.2: Overview of WGBS, Methylated cytosines ('Me') remain unchanged after treatment, while unmethylated cytosines convert to uracil (U), and are read as thymine (T) during PCR. This allows for comparison with the reference genome to identify DNA methylation patterns.

computational tools to conduct comparative analyses between treated and untreated sequences facilitates the accurate determination of the methylation status of individual cytosine residues (Fig. 1.2).

1.3 The molecular origin of spontaneous epimutations

Advanced base-resolution methylation techniques have enabled detailed monitoring of cytosine methylation changes over developmental phases and generational spans. It has become apparent that, despite stringent control mechanisms, the methylation status of cytosines, whether individual or in clusters, does not always maintain integrity across cell division cycles. Spontaneous alterations in cytosine methylation, a phenomenon

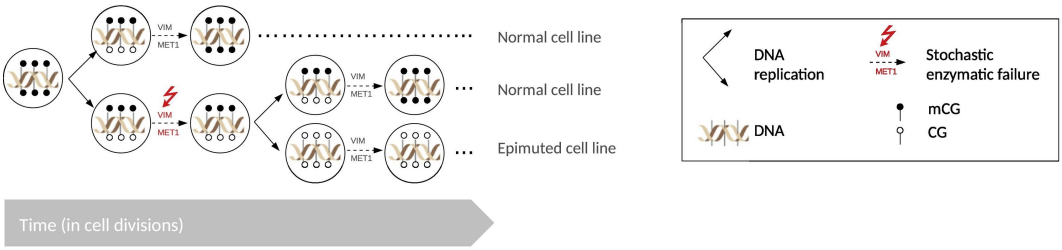

Figure 1.3: Model of molecular basis of spontaneous CG epimutations. CG epimutations can spontaneously occur due to occasional MET1 methyltransferase malfunction. The illustration depicts two cell division cycles. During DNA replication, VARIANT IN METHYLATION proteins identify hemimethylated CG sites, guiding METHYLTRANSFERASE 1 to induce CG methylation on the fresh strand via a "template-copying" method. Any lapses in enzymatic function or unintended methyltransferase actions might result in enduring methylation reductions or additions in offspring cells and their subsequent cell lineages (referred to as the epimutated cell line). To keep things clear, only the CG methylation decrease is shown. Figure source: Yao, N., et al. [23]

described as "spontaneous epimutation" have been identified in both animals and plants. These epimutations, known to accumulate over the course of development and aging [19], are predominantly observed at CG dinucleotides in plant species.

A basic molecular model suggests that the random loss of CG methylation arises from the flawed enzymatic function of MET1 during the process of DNA replication. In instances where MET1 fails to methylate the new DNA strand, permanent methylation losses could be passed down to daughter cells and their subsequent generations, including germline cells (Fig. 1.3). Occasional de novo methylation activities of MET1 [20] may account for unexpected gains in CG methylation, a notion supported by the similar roles observed in the mammalian equivalent, DNA (cytosine-5)-methyltransferase 1 (DNMT1) [21]. Notably, in mammals, such activities are largely limited to retrotransposons [22].

Discrepancies in the maintenance of CG methylation are notably evident in somatic contexts, particularly when they stem from the activity within shoot apical meristems (SAMs). These meristems, a specialized and small collection of stem cells, are integral to the development of the plant's aerial components. The fact that only a select few meristematic cells act as the progenitors for new lateral branches [24], as well as for leaves and flowers, means that this limited cellular representation leads to fixation of CG epimutations within newly developing vegetative lineages. This dynamic, known as "somatic epigenetic drift" (Fig. 1.4A), often culminates in pronounced chimerism [25]. This phenomenon is particularly observable in perennial species, where sequential methylation changes in consecutively layered sections of the plant are evident [26, 27].

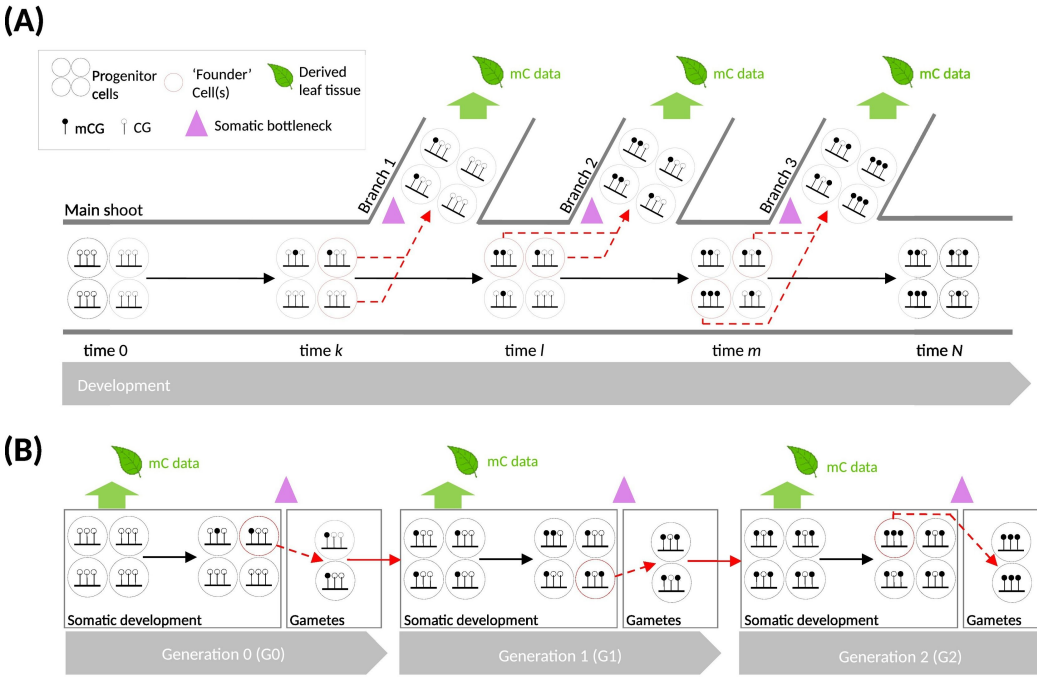

Figure 1.4: Somatic origin and transgenerational inheritance of CG epimutations in plants. **A)** During mitotic divisions of progenitor cells originating from SAM, a lack of CG methylation preservation results in unexpected somatic epimutations passed onto the side branches. For clarity, we've depicted only the spontaneous addition of methylation on a singular genome. Due to a limited number of these progenitor cells serving as starting points for new side branches, there's a notable constraint on CG epimutations at every branching juncture. Consequently, these become permanently set in fresh vegetative lineages due to "somatic drift." This phenomenon amplifies the differences between leaves from distinct terminal branches—for instance, leaf methylomes from Branches 1 and 2 resemble each other more than those from Branches 1 and 3. **B)** Cells that evolve into gametes have their origins in SAM precursors, typically at advanced stages of growth. This implies that CG epimutations acquired during the somatic phase can be passed on to future generations. Plants allow for a consistent transfer across generations since the CG methylome is not radically altered in reproductive cells and initial zygotic stages. For clarity, only a pair of gametes from a singular lineage (either male or female) are presented per generation. Figure source: Yao, N. et el. [23]

This methylation change pattern reflects the occurrence of rare somatic DNA nucleotide mutations [26] and [28–32], implying a shared meristematic origin for these stochastic genetic and epigenetic variations.

Unlike animals, plants do not possess a designated germline. Cells destined to form gametes are instead derived from SAM precursors relatively late in development [33], implicating that somatically acquired CG epimutations are often transmitted to subse-

quent generations (Fig. 1.4B). Ensuring stable intergenerational transfer necessitates a lack of reprogramming of the CG methylome in sex cells and the early zygote, a condition that is typically met in plant species [34].

This heritability underscores the contribution of epimutations to the DNA methylation diversity within plant populations, providing a substrate for selection over evolutionary timescales. The realization that these epimutations, which occur at rates significantly higher than genetic mutations and exhibit apparent neutrality and are inherited with stability, underscoring their potential as a reservoir for genetic variability, which might play a significant role in long-term evolutionary processes and species diversification.

Therefore, it is essential to study and quantify these epimutational processes within controlled experimental systems to understand their broader implications on plant biology, evolution, and potentially, agricultural breeding strategies. This understanding can provide insights into the evolutionary mechanisms at play and could be leveraged to harness plant diversity and resilience in changing environments.

1.4 Experimental systems to study epimutational processes

In multi-generational studies of epimutational phenomena, a key challenge lies in differentiating "germline" epimutations from other types of methylation alterations, such as those linked to genetic variation that segregates or temporary environmental disturbances [35]. To address this, mutation accumulation (MA) lines cultivated under stringent laboratory conditions serve as an important experimental framework. Originating from a single isogenic progenitor, MA lines are propagated independently across numerous generations. The advancement of these lines can be executed through clonal methods or sexual reproduction, including self-fertilization or mating between siblings (Fig. 1.5A). In the context of MA lines produced clonally, maintaining isogenicity in the progenitor is not a prerequisite, given that the genome remains "fixed" , thus eliminating the factor of genetic segregation.

The relational structure among diverse MA lineages can be graphically depicted through a pedigree or phylogenetic tree (Fig. 1.5A), with its topology usually predefined due to deliberate choices in experimental design, including the timing of branch points and the lengths of branches. When paired with multi-generational methylome assessments, MA lines offer the unique capability for "real-time" monitoring of "germline" epimutations, all while maintaining a nearly constant genomic background. This setup aids in deriving estimates for the rates of epimutation per generation [36]. Currently, sequenced methylomes from an extensive array of sexually propagated MA lines are accessible for A. thaliana [37–42] and rice [43]. Additionally, various other MA lines are in the process of being developed for epimutation analyses across different genotypes, environmental conditions, and plant species.

In addition to experimentally derived MA lines, natural mutation accumulation systems are also observable in the context of plant development and aging. Long-lived

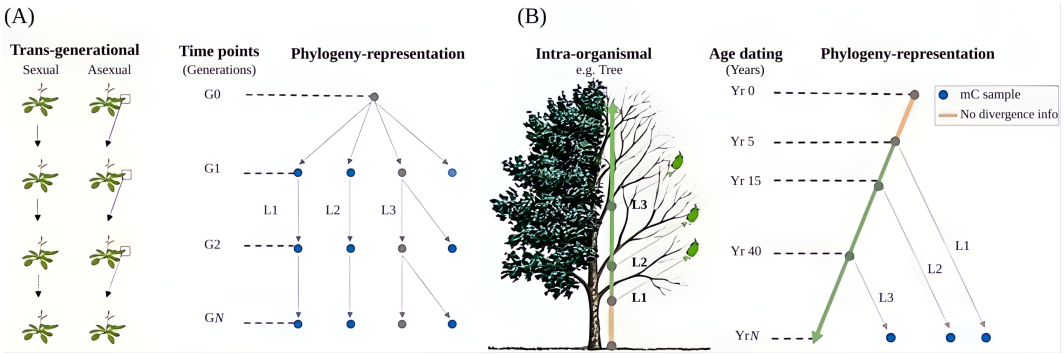

Figure 1.5: Mutation Accumulation systems.**(A)** Development of mutation accumulation (MA) lines spanning multiple generations (from G0 to GN) can be achieved either through sexual methods like self-breeding or mating among siblings or through asexual means such as clonal reproduction. These various lineages, from L1 to L3, can be depicted as a phylogenetic tree. Both the points of divergence and the lengths of the branches are usually predetermined from the experimental setup. mC sampling can be carried out at chosen generational intervals, either directly from the progenitor plant material or from their sibling counterparts. **(B)** A tree can be seen as a representation of intra-species phylogeny. The branching structure or topology is typically known, and the points and lengths of branches can be determined using coring methods. For clarity, only three branches (L1 to L3) are emphasized. Using measurements from leaf mC, one can compute CG methylation differences. Likewise, by retracing the history of the branches to their most recent common point of divergence, divergence durations (expressed in years) between pairs of leaves can be assessed. This can be done only up to the tree's earliest point of branching (in this example, Yr = 5) and not for periods before that (indicated by the orange segment). Figure source:Yao, N. et al. [23]

perennials, like trees, provide a particularly illustrative example. The branching structure of trees can be treated as a phylogeny of somatic lineages, carrying valuable information about the epimutational history of each branch [44]. To determine the branch-point times and branch lengths, one can utilize coring data or other dating methods (Fig. 1.5B).

By integrating this information with contemporary leaf methylome measurements, it is possible to infer the rate of somatic epimutations as it varies with age [26]. This approach allows for the examination of the dynamic changes in epigenetic patterns over the lifespan of the plant and offers insights into the accumulation and persistence of somatic epimutations over time. The exploration of natural mutation accumulation systems in perennial plants provides an invaluable opportunity to understand the complexities of epigenetic processes in the context of plant growth and aging.

1.5 Challenges with quantifying epimutation rates

Efforts to determine the rate of spontaneous epimutations across various plant systems are significantly impeded due to a shortage of analytical methods. Simplistic approaches that simply count the number of methylation alterations over a given time period are impractical in this context, as measurements of DNA methylation are exceedingly prone to noise.

From a technological perspective, this noise originates from heightened errors in sequencing and alignment of bisulfite reads, as well as inefficiencies in bisulfite conversion. On the biological front, augmented measurement errors may arise from heterogeneity within tissues in terms of 5mC patterns [45], and the dependency of DNA methylomes on variations in environmental and laboratory conditions, potentially resulting in transcription-dependent methylation changes [46].

To navigate these obstacles, a model-based estimation approach that accurately accounts for measurement errors in the data is imperative. This approach should describe the temporal accumulation of epimutations through a clearly defined statistical model.

2 Aims and outline

The overarching goal of this thesis is to implement such a model within a general computational workflow that starts with the raw WGBS data and ends with statistical estimates of the spontaneous epimutation rate. To achieve this, my first aim is to construct a bioinformatic pipeline for the high-throughput analysis of WGBS data. My approach to this problem is summarized in section 4.1 ("Publication 1"). The second aim of this thesis is to use the WGBS data from mutation accumlation systems (see Fig. 1.5) to estimate the rate of spontaneous epimutations. The computational approach to this problem is presented in section 4.2 ("Publication 2"). Key methodological aspects contained in the two publicaitons are highlighted in section 3 ("Materials and methods"), and a discussion and outlook based on these publications is given in section 5 ("Discussion and outlook").

3 Materials and methods

3.1 Bioinformatic processing and analysis of WGBS data

Whole Genome Bisulfite Sequencing (WGBS) has emerged as a gold standard for examining genome-wide DNA methylation, a critical epigenetic modification with profound implications in various biological processes, including gene regulation, genomic imprinting, and cellular differentiation. The generation of high-throughput WGBS data provides an unparalleled depth of insight into the methylation landscape across the genome. However, the voluminous and intricate nature of WGBS data necessitates robust bioinformatics processing and analysis to effectively decipher and interpret the underlying methylation patterns.

To navigate the complexities of WGBS data, an array of bioinformatic tools and pipelines have been developed, each designed to efficiently handle distinct aspects of the data processing and analysis workflow. This array of tools encompasses a variety of functions, such as data normalization (e.g. RnBeads [47], SWAN [48], ChAMP [49]), the identification of differentially methylated regions (DMRs) with tools like Methylkit [50], DMRcaller [51], Methylpy [52], and metilene [53]. Additionally, there are tools designed for the imputation of methylomes from bulk whole-genome bisulfite sequencing (WGBS) data, such as METHimpute [54], and others for the imputation of single-cell methylomes, like Melissa [55] and deepCpG [56]. Also included are tools for addressing dropouts in single-cell data, such as SCRABBLE [57]. Among these, the MethylStar pipeline [58] stands out as a comprehensive solution, offering an integrated suite of analytical processes for the precise and reliable examination of bisulfite sequencing data.

MethylStar includes an interactive command-line user interface, designed with Python, which simplifies the process of configuring software settings and running the pipeline. This easy navigation allows both experts and non-experts to handle large batches of samples efficiently, without the necessity to type commands at the terminal.

MethylStar automatically manages computational resources and performs error detection. It integrates well-established tools and operates in a highly parallelized environment. This setup ensures that errors are minimized and any that do occur are promptly identified and addressed, contributing to the stability and reliability of the software in processing large datasets.

The ensuing sections delineate the core components of the MethylStar pipeline, shedding light on their functions and contributions to the overall data analysis process. The following parts provide a more detailed description of each component of the MethylStar

pipeline, focusing on the input and output at each stage and how these processes work. See Figure 3.1 for the workflow of the MethylStar pipeline.

3.1.1 Quality Control (QC) of Raw Data

Quality control in sequencing data constitutes a rigorous assessment of various quality metrics, including read length and base quality. This evaluative step guarantees the utilization of only high-quality data for subsequent analyses. MethylStar executes QC assessment to ascertain and uphold the integrity of the sequencing data.

Tool used: FastQC [59].

Starting point: The pipeline begins with raw sequencing data, typically in the form of FASTQ files. FASTQ files are a standard format in genomics and bioinformatics, providing a way to store sequence data along with quality information.

Process: The pipeline initiates by assessing raw sequencing data (FASTQ format) through FastQC, which evaluates key quality metrics such as per-base sequence quality, sequence duplication levels, and overrepresented sequences. This step is pivotal in identifying potential quality issues in the sequencing data, thereby informing the necessity for subsequent trimming and alignment adjustments.

Output: HTML reports and plots detailing various quality metrics.

3.1.2 Adapter Trimming and Quality Filtering

Next, the pipeline employs Trim Galore! to process Fastq files. Adapter trimming is a pivotal preprocessing step that involves the exclusion of adapter sequences inadvertently incorporated during the sequencing process. The presence of these sequences can lead to analytical complications, underscoring the necessity of their removal.

Tool used: Trim Galore! [60].

Input: The FASTQ files post-initial QC

Process: Trim Galore! It focuses on two main tasks: trimming adapter sequences from sequencing reads and filtering out low-quality bases. The removal of adapter sequences is vital for preventing inaccurate results in downstream analyses, such as sequence alignment. Simultaneously, the software filters out bases with low Phred quality scores, ensuring that only high-quality data is used for further analysis. This dual functionality of Trim Galore! ensures that the resulting data is not only of high quality but also free from sequences that could lead to erroneous interpretations, thereby bolstering the reliability of genomic and transcriptomic studies.

Output: The output from this step is a set of trimmed and cleaned FASTQ files, ready for the next stage.

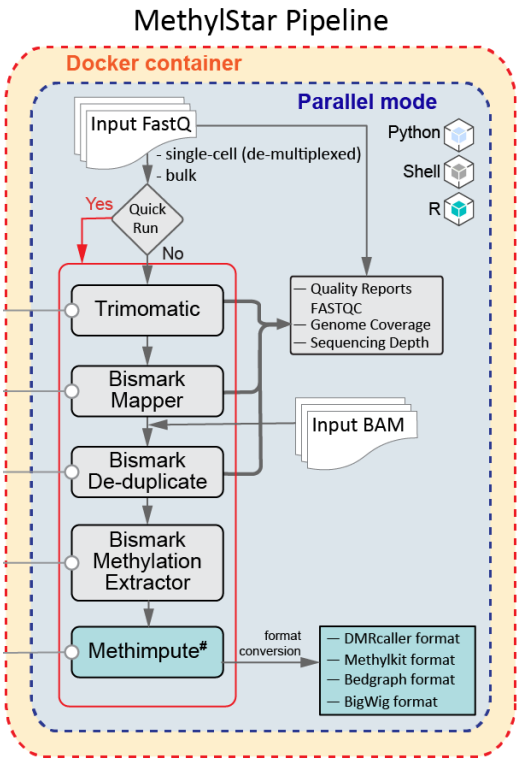

Figure 3.1: Basic workflow of MethylStar showing the pipeline architecture. Figure source: modified from Shahryary et al. [58]

3.1.3 Post-trimming Quality Control

While Trim Galore! effectively trims adapters and removes low-quality sequences, it is imperative to validate the quality of the resulting data. Post-trimming quality control assesses parameters such as the distribution of sequence quality scores, GC content, sequence length distribution, and the presence of overrepresented sequences. By conducting this analysis, researchers can ensure that the trimming process has not inadvertently introduced biases or errors, and that the data is of sufficient quality for accurate and reliable downstream analysis.

Tool used: FastQC [59].

Input: Trimmed and cleaned FASTQ files from the previous step.

Process: This step repeats the initial QC process using FastQC but with the trimmed data. It reassures that the trimming process has been effective and that the data quality is suitable for the next steps.

Output: Updated HTML reports and plots assessing the quality of the trimmed reads.

3.1.4 Alignment to Reference Genome

Read alignment involves the precise mapping of sequenced reads onto a reference genome. MethylStar employs Bismark for read alignment, a tool specially designed for bisulfite-treated sequencing reads. Bismark negotiates the challenges posed by bisulfite conversions (C to T or G to A) by transforming the reference genome to its bisulfite-converted form. This approach allows for the effective alignment of reads, concurrently accounting for bisulfite conversions, and enabling the accurate determination of methylation states by monitoring converted and non-converted cytosines.

Tool used: Bismark Mapper [61].

Input: Quality-filtered and trimmed FASTQ files.

Process: Bismark performs alignment by first converting the reference genome and the reads to their bisulfite-treated equivalents. It then aligns the reads to this converted reference genome. Bismark handles the complexities of bisulfite sequencing data, like the C-to-T conversions, ensuring accurate alignment and preparation for methylation state analysis.

Output: BAM files representing the aligned reads.

3.1.5 Post-alignment Processing

When Bisulfite sequencing is performed, the DNA is often fragmented and then amplified through a process like PCR (Polymerase Chain Reaction). This amplification can result in multiple copies of the same DNA fragment. These duplicates can skew the data analysis, leading to inaccurate representation of methylation levels. Bismark-Deduplication is the process of eradicating duplicate reads originating from PCR amplification during library preparation, a step crucial for obviating analytical bias. In the MethylStar pipeline, deduplication is conducted post-alignment, ensuring that each read is uniquely accounted for in the ensuing analysis, thereby enhancing the robustness and precision of the analysis.

Tool used: Bismark-deduplication [61] and Samtools [62].

Input: BAM files from the alignment step.

Process: In Bismark, deduplication works by comparing the start coordinates of uniquely aligned sequencing reads in the reference genome. If multiple reads have identical start coordinates, indicating PCR duplicates, all but one of these reads are discarded, ensuring that only unique, non-duplicated reads are retained for accurate methylation analysis. Similarly, SAMtools contributes by marking or removing these duplicates based on their alignment position and orientation. This integrated approach ensures that only unique, non-duplicated reads are used in the analysis, preventing the overrepresentation of any DNA fragment and providing a more accurate reflection of the genome's methylation status.

Output: Sorted and indexed BAM files.

3.1.6 Methylation Extractor

The Bismark Methylation Extractor processes the results of bisulfite-treated DNA sequencing from the post-alignment processing phase.

Tool used: Bismark Methylation Extractor [61].

Input: BAM files from the alignment step.

Process: The supplementary bismark methylation extractor script within Bismark is designed to analyze Bismark result files, extracting methylation data for each cytosine. It outputs the positions of these cytosines, categorized by their genomic context (CpG, CHG, or CHH). The script distinguishes between methylated and non-methylated cytosines, denoting methylated cytosines as forward reads (+) and non-methylated cytosines as reverse reads (-).

Output: CX report format.

Additionally, MethylStar create files compatible with genome browsers, bedGraph, and coverage file formats, facilitating diverse downstream analyses.

After this part, the remaining data is used for further analysis, such as determining the methylation status of cytosines in the genome (see Methylation state calling).

3.2 Methylation state calling

The end point of the WGBS data processing pipeline described in 3.1 is cytosine-level statistics of the number of reads aligning to a given cytosine that report that the cytosine is methylated over the total number of reads aligning to that cytosine. This information is typically used to define so-called "methylation levels":

$$methylationlevels = R/S$$

where R is number of reads reporting methylation and S total number of reads.

However, it is often of interest to obtain discrete methylation state calls; that is, information if a cytosine is methylated or unmethylated.

3.2.1 Classical Binomial model

A common way to call cytosine-level methylation states from WGBS data is to use a binomial model. At a given cytosine position the binomial model calculates the probability of observing k number of methylated reads out of n total reads where($n \geq k$), given the null hypothesis that the cytosine is actually unmethylated. Let the random variable X follows the binomial distribution with parameters $n \in N$ and $p \in [0, 1]$. The probability of obtaining exactly k successes (i.e. methylated reads) in n independent

Bernoulli trials is given by the probability mass function:

$$P(K = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.1}$$

for $k = 0, 1, 2, ..., n$ where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient. Here p is the success probability which is typically take to fixed at $1 - C$, where C is the bisulfite conversion rate (e.g. Becker et al.. 2011 [63], Schmitz et al. 2011 [38], van der Graaf et al. 2015 [37]). The bisulphite conversation rate is typically determined from counting the fraction of methylated reads aligning to the chloroplast genome, which is known to be unmethylated.

Using the binomial model, the sample size of the test is the total number of reads (n) aligning to the cytosine. In samples with low sequencing depth this approach has low statistical power and will lead to substantial undercalling of methylated cytosines (i.e. false negatives). Methimpute [54] is an alternative approach that largely overcomes this problem.

3.2.2 Hidden Markov Model

Methimpute is a hidden markov model with binomial emission densities. It is capable of making accurate methylation calls even for cytosines with missing or low read counts. Another advantage of Methimpute is its ability to identify not only epihomozygotous unmethylated (UU) or epihomozygous methylated sites (MM), but also epiheterozygous sites (MU). Knowledge of epiheterozygous sites is absolutely necessary when studying epimutation accumulation in clonal species / lineages, as most epimutations are of the form $UU \leftrightarrow MU$, or $MM \leftrightarrow MU$. Methimpute actually reduces to a standard binomial model in samples with large read coverage, which is the case in the present study. In such situations, Methimpute does not impute anything, but just performs binomial calling similar to the standard binomial model. The only difference is that Methimpute makes use of its underlying hidden markov structure for statistical inference.

In its two-state implementation, the algorithm operates through a HMM, where the hidden states correspond to the unmethylated (UU) and methylated (MM) statuses of cytosines. The model parameters, including the binomial parameters p_U and p_M, and the transition matrix, are estimated during the training phase. The algorithm calculates posterior probabilities γ_U and γ_M, which represent the probability that a given cytosine belongs to one of the hidden states. These probabilities serve as a measure of confidence in the methylation status call. Based on the posterior probabilities, METHimpute provides discrete methylation status calls for every cytosine in the genome. A cytosine's maximum posterior probability represents its most likely methylation status, and the magnitude of this probability can be used as a measure of confidence in the underlying status call. In addition to methylation status calls, METHimpute outputs recalibrated

methylation levels per cytosine, calculated as:

$$m^* = p_U \cdot \gamma_U + p_M \cdot \gamma_M \tag{3.2}$$

This ensures that the methylation levels are adjusted based on the model's estimations, providing a more accurate representation of the methylation status.

3.3 Epimutation rate estimation

Having obtained methylation state calls for all samples in any one of the mutation accumulation systems introduced above, the next object is to estimate the rate at which spontaneous epimutations arise per unit time. To achieve this, the AlphaBeta software was developed, facilitating the precise analysis of these epigenetic alterations. This tool proved essential in quantifying epimutation rates across various plant species, such as Arabidopsis thaliana, poplar, and dandelion, thereby significantly advancing our comprehension of plant epigenetics. In what follows I will outline my approach to this problem.

3.3.1 Calculating mC divergence function

In the context of the ith sequenced sample within the pedigree, let's denote s_{ik} as the observed methylation state at the kth locus (where k ranges from $1...N$). These N loci could refer to individual cytosines or predefined regions containing clusters of cytosines. We adopt a coding system where s_{ik} can take on values of $1, 0.5$, or 0. These values correspond to the diploid epigenotype at that specific locus: m/m for 1, m/u for 0.5, and u/u for 0. Here, m represents a methylated epiallele, while u signifies an unmethylated epiallele. Utilizing this coding scheme, we proceed to calculate the mean absolute 5mC divergence denoted as D. This calculation allows us to quantify the divergence between any two samples, say i and j, within the pedigree. The formulation of this divergence metric involves assessing the differences in methylation patterns between these samples.

$$D_{ij} = \sum_{k=1}^{N} I(s_{ik}, s_{jk}) N^{-1}, \tag{3.3}$$

where $I(\cdot)$ is an indicator function, such that

$$I(s_{ik}, s_{jk}) = \begin{cases} 0 \text{ if } s_{ik} = s_{jk} \\ \frac{1}{2} \text{ if } s_{ik} = 0.5 \text{ and } s_{jk} \in \{0, 1\} \\ \frac{1}{2} \text{ if } s_{jk} = 0.5 \text{ and } s_{ik} \in \{0, 1\} \\ 1 \text{ if } s_{ik} = 0 \text{ and } s_{jk} = 1 \\ 1 \text{ if } s_{jk} = 1 \text{ and } s_{ik} = 0. \end{cases}$$

AlphaBeta performs automated computations of both D_{ij} and Δt across all distinct pairs of samples. This calculation is carried out using the methylation state calls and the corresponding pedigree coordinates of each individual sample as input.

3.3.2 Modeling 5mC divergence

The model for 5mC divergence is formulated as:

$$D_{ij} = c + D_{ij}^{\bullet}(M_{\Theta}) + \epsilon_{ij}, \tag{3.4}$$

Here, the term $\epsilon_{ij} \sim N(0, \sigma^2)$ represents the normally distributed residual error, c denotes the intercept, and $D_{ij}^{\bullet}(M_{\Theta})$ signifies the anticipated divergence between samples i and j. This divergence is a function of an underlying epimutation model $M(\cdot)$ with parameter vector Θ, as detailed below.

The expected divergence, $D_{ij}^{\bullet}(M_{\Theta})$, is computed using the following equation:

$$\begin{aligned} D_{ij}^{\bullet}(M_{\Theta}) = & \sum_{n \in v} \sum_{l \in v} \sum_{m \in v} I(l, m) \\ & \cdot Pr(s_{ik} = l, s_{jk} = m | s_{ijk} = n, M_{\Theta}) \\ & \cdot Pr(s_{ijk} = n | M_{\Theta}), \end{aligned} \tag{3.5}$$

where s_{ijk} represents the methylation state at locus k of the most recent common ancestor of samples i and j, and $v = \{0, 0.5, 1\}$.

Given the conditional independence of samples s_i and s_j, we can further express:

$$\begin{aligned} Pr(s_{ik}, s_{jk} | s_{ijk}, M_{\Theta}) = & Pr(s_{ik} | s_{ijk}, M_{\Theta}) \\ & \cdot Pr(s_{jk} | s_{ijk}, M_{\Theta}). \end{aligned} \tag{3.6}$$

In order to assess these conditional probabilities, it becomes necessary to establish an explicit form for the epimutational model, M_{Θ}. To motivate this, we introduce $\mathbf{G}$ as a 3×3 transition matrix, summarizing the probability of transitioning from epigenotype l to m within the time interval $[t, t+1]$:

$$\mathbf{G} = \begin{array}{c} \begin{array}{ccc} \text{u/u (t+1)} & \text{m/u (t+1)} & \text{m/m (t+1)} \end{array} \\ \begin{bmatrix} f_{11}(\alpha, \beta, w) & f_{12}(\alpha, \beta, w) & \cdot \\ f_{21}(\alpha, \beta, w) & \cdot & \cdot \\ \cdot & \cdot & f_{33}(\alpha, \beta, w) \end{bmatrix} \end{array} \begin{array}{l} \text{u/u (t)} \\ \text{m/u (t)} \\ \text{m/m (t)} \end{array} \tag{3.7}$$

The elements of this matrix depend on parameters such as the gain rate α, loss rate β, and selection coefficient $w \in [0, 1]$. Depending on the propagation method, $\mathbf{G}$ has distinct forms.

For diploid systems propagated by selfing:

$$\mathbf{G} = \begin{bmatrix} (1-\alpha)^2 & 2(1-\alpha)\alpha & \alpha^2 \\ \frac{1}{4}(\beta+1-\alpha)^2 & \frac{1}{2}(\beta+1-\alpha)(\alpha+1-\beta) & \frac{1}{4}(\alpha+1-\beta)^2 \\ \beta^2 & 2(1-\beta)\beta & (1-\beta)^2 \end{bmatrix} \circ \mathbf{W}, \tag{3.8}$$

For systems propagated clonally or somatically:

$$\mathbf{G} = \begin{bmatrix} (1-\alpha)^2 & 2(1-\alpha)\alpha & \alpha^2 \\ \beta(1-\alpha) & (1-\alpha)(1-\beta)+\alpha\beta & \alpha(1-\beta) \\ \beta^2 & 2(1-\beta)\beta & (1-\beta)^2 \end{bmatrix} \circ \mathbf{W}, \tag{3.9}$$

Here, $\circ$ denotes the Hadamard product, and $\mathbf{W}$ is a matrix of selection coefficients depending on the nature of selection against epialleles u or m.

$$\begin{bmatrix} w & \frac{(w+1)}{2} & 1 \\ w & \frac{(w+1)}{2} & 1 \\ w & \frac{(w+1)}{2} & 1 \end{bmatrix} \text{ or } \begin{bmatrix} 1 & \frac{(w+1)}{2} & w \\ 1 & \frac{(w+1)}{2} & w \\ 1 & \frac{(w+1)}{2} & w \end{bmatrix}$$

Utilizing this framework, we are able to categorize four distinct models, denoted as *ABneutral*, *ABmm*, *ABuu*, and *ABnull*. The *ABneutral* model postulates that the accumulation of spontaneous 5mC gains and losses is selectively neutral ($w = 1$, α and/or $\beta > 0$). In this scenario, all epigenotype transitions from time t to $t + 1$ are governed solely by the rates α and β, and in the case of selfing, also by the Mendelian segregation of epialleles u and m.

In contrast, the selection models *ABmm* and *ABuu* propose that the accumulation of epimutations is influenced by selection against spontaneous gains or losses of 5mC, respectively ($w < 1$, α and/or $\beta > 0$). For instance, in the case of selection favoring epiallele u (*ABuu* model), the fitness of epihomozygote m/m and epiheterozygote m/u is reduced by factors of w and $(w + 1)/2$, respectively. This fitness reduction is directly incorporated into the transition matrix by adjusting the transition probabilities to these epigenotypes accordingly [64]. Similar principles apply when selection is in favor of epiallele m. As a reference, the *ABnull* model is defined as the null model of no accumulation, characterized by $\alpha = 0$, $\beta = 0$, and $w = 1$.

To ensure the row sum of $\mathbf{G}$ (i.e., transition probabilities) remains unity in the presence of selection, we redefine $\mathbf{G}$ using the normalization:

$$\mathbf{G}' = \begin{bmatrix} (\sum_i \mathbf{G}_{1i})^{-1} & 0 & 0 \\ 0 & (\sum_i \mathbf{G}_{2i})^{-1} & 0 \\ 0 & 0 & (\sum_i \mathbf{G}_{3i})^{-1} \end{bmatrix} \cdot \mathbf{G} \tag{3.10}$$

Drawing on Markov chain theory, the conditional probability $Pr(s_{ik}|s_{ijk}, M_\Theta)$ can be expressed in terms of $\mathbf{G}'$ as outlined below:

$$\sum_n Pr(s_{ik} = 0|s_{ijk} = n, M_\Theta) = \sum_{r=1}^{3} (\mathbf{G}'^{t_i - t_{ij}})_{r1} \tag{3.11}$$

$$\sum_n Pr(s_{ik} = 0.5|s_{ijk} = n, M_\Theta) = \sum_{r=1}^{3} (\mathbf{G}'^{t_i - t_{ij}})_{r2} \tag{3.12}$$

$$\sum_n Pr(s_{ik} = 1|s_{ijk} = n, M_\Theta) = \sum_{r=1}^{3} (\mathbf{G}'^{t_i - t_{ij}})_{r3} \tag{3.13}$$

where t_i corresponds to the time-point of sample i, and t_{ij} signifies the time-point of the most recent common ancestor shared between samples i and j ($t_{ij} \leq t_i, t_j$). The expressions for $Pr(s_{jk}|s_{ijk}, M_\Theta, t_j)$ can be derived correspondingly by substituting t_i with t_j in the aforementioned equation. Notably, the calculation of these conditional probabilities necessitates iterative matrix multiplication. Nonetheless, a direct assessment of these equations is feasible by leveraging the fact that

$$\mathbf{G}'^{t_i - t_{ij}} = \mathbf{p}\mathbf{V}^{t_i - t_{ij}}\mathbf{p}^{-1} \text{ and } \mathbf{G}'^{t_j - t_{ij}} = \mathbf{p}\mathbf{V}^{t_j - t_{ij}}\mathbf{p}^{-1}, \tag{3.14}$$

where $\mathbf{p}$ signifies the eigenvector of matrix $\mathbf{G}'$, and $\mathbf{V}$ denotes a diagonal matrix of eigenvalues. For selfing and clonal/somatic systems, these eigenvalues and eigenvectors can be analytically obtained.

Finally, to deduce $D^{\bullet}_{ij}(M_\Theta)$, we also require $Pr(s_{ijk} = n|M_\Theta)$; in other words, the probability that locus k in the most recent common ancestor of samples i and j is in state n ($n \in \{0, 0.5, 1\}$). To achieve this, consider the methylome of the pedigree founder at time $t = 1$, and let $\pi = [p_1\ p_2\ p_3]$ denote a row vector of probabilities corresponding to states u/u, u/m, and m/m, respectively. Leveraging Markov Chain theory, we obtain

$$Pr(s_{ijk} = 0|M_\Theta) = \left[\pi\, \mathbf{G}'^{(t_{ij}-1)}\right]_1 \tag{3.15}$$

$$Pr(s_{ijk} = 0.5|M_\Theta) = \left[\pi\, \mathbf{G}'^{(t_{ij}-1)}\right]_2 \tag{3.16}$$

$$Pr(s_{ijk} = 1|M_\Theta) = \left[\pi\, \mathbf{G}'^{(t_{ij}-1)}\right]_3 \tag{3.17}$$

In many instances, the most recent common ancestor corresponds to the pedigree founder itself, yielding $t_{ij} = 1$. In scenarios where the methylome of the pedigree founder has been measured, the probabilities p_1, p_2, and p_3 can be directly estimated from the data using $x_1 N^{-1}$, $x_2 N^{-1}$, and $x_3 N^{-1}$, respectively. Here, x_1, x_2, and x_3 represent the number of loci observed in states u/u, u/m, m/m, and N denotes the total number of loci. Typically, x_2 remains unknown since most DMP and DMR callers do not output intermediate methylation states (epiheterozygous calls). Hence, we adopt the following:

$$p_1 = \frac{x_1}{N}, \quad p_2 = \gamma\frac{x_3}{N}, \quad p_3 = (1 - \gamma)\frac{x_3}{N} \tag{3.18}$$

where $\gamma \in [0, 1]$ stands as an unknown parameter.

3.3.3 Parameter estimation

To estimate the values of Θ, our objective is to minimize the expression:

$$\nabla \sum_{q=1}^{M} \big(D_q - D_q^{\bullet}(M_\Theta) - c\big)^2 = \mathbf{0}, \tag{3.19}$$

In this equation, the summation spans across all M distinct pairs of sequenced samples within the pedigree. The minimization process employs the "Nelder-Mead" algorithm, utilizing the optimx package in R. However, our experience reveals that achieving convergence is not consistently stable. This issue likely arises due to the intricate and highly non-linear nature of the function $D_q^{\bullet}(M_\Theta)$. To address this, we introduce an additional constraint for minimization:

$$\nabla \sum_{q=1}^{M} \big(D_q - D_q^{\bullet}(M_\Theta) - c\big)^2 + M\big(\tilde{p_1} - p_1(t_\infty, M_\Theta)\big)^2 = \mathbf{0}. \tag{3.20}$$

Here, $p_1(t_\infty, M_\Theta)$ signifies the equilibrium proportion of u/u loci in the genome as t approaches infinity. In the context of a selfing system with $w = 1$, we find:

$$p_1(t_\infty, M_\Theta) = \frac{\beta((1-\beta)^2 - (1-\alpha)^2 - 1)}{(\alpha+\beta)((\alpha+\beta-1)^2 - 2)}, \tag{3.21}$$

For a clonal/somatic system, it takes the form:

$$p_1(t_\infty, M_\Theta) = \frac{\beta^2}{(\alpha+\beta)^2} \tag{3.22}$$

For situations where $0 \leq w < 1$, the equations become more intricate and are excluded from presentation here. Importantly, $\tilde{p_1}$ represents an empirical estimation of these equilibrium proportions. In cases where the methylomes of samples can be assumed to be in equilibrium, we find that $p_1(t = 1) = p_1(t = 2) = \cdots = p1(t_\infty)$, signifying that the proportion of loci in the genome exhibiting the u/u state remains dynamically stable for any time t. Under this presumption, $\tilde{p_1}$ can be replaced with $\overline{p_1}$, the average proportion of u/u loci calculated from all samples within the pedigree.

4 Publications: summaries and contributions

4.1 Publication 1. MethylStar

MethylStar: A fast and robust pre-processing pipeline for bulk or single-cell whole-genome bisulfite sequencing data.

Shahryary, Y., Hazarika, R.R. & Johannes, F. MethylStar: A fast and robust pre-processing pipeline for bulk or single-cell whole-genome bisulfite sequencing data.

Published in BMC Genomics 21, 479 (2020).
DOI: 10.1186/s12864-020-06886-3

Summary

As Whole-Genome Bisulfite Sequencing (WGBS) becomes increasingly central to large-scale epigenetic studies, the necessity for specialized software to effectively manage growing data volumes is paramount. This ensures accurate analysis and interpretation of methylation patterns important for understanding their roles in biological processes.

While existing WGBS tools like RnBeads [47], SWAN [48], ChAMP [49], Methylkit [50], and DMRcaller [51] are adept at data normalization and identifying differentially methylated regions, they depend on robust pre-processing steps such as quality control, sequence sorting, adapter trimming, and genome alignment. However, comprehensive solutions like gemBS [65] and Methylpy [52], which aim to integrate these stages, often face challenges with user-friendly setups and efficient processing of large datasets.

MethylStar offers an advanced pre-processing pipeline adept at handling both bulk and single-cell WGBS data. Utilizing sophisticated algorithms and parallel computing, it accelerates the analysis of raw sequencing reads, effectively correcting bisulfite conversion efficiencies and reducing batch effects. By incorporating essential functions such as read trimming, alignment, and methylation state calling within this parallel processing framework, MethylStar significantly enhances computational throughput. Its adaptable framework also supports customized configurations to accommodate diverse species and experimental designs, making it exceptionally versatile for a range of epigenetic research initiatives. Additionally, MethylStar is engineered for ease of use, featuring a dockerized container that

simplifies installation by preloading all necessary dependencies. Its user-friendly interface is accessible to both experts and novices alike, promoting broader adoption and making it an invaluable tool for epigenetic research.

Authors' contributions

FJ, RRH and YS conceptualized the method. YS and RRH developed, implemented and tested the pipeline. RRH, FJ and YS wrote the paper. FJ supervised the project.

My detailed contributions:

Software implementation and optimization: I was responsible for the full life-cycle development of the MethylStar pipeline, focusing on software implementation, performance optimization, and parallelization. This involved ensuring that the pipeline was not only functionally robust but also optimized for high-throughput data processing, leveraging parallel computing techniques to enhance its speed and efficiency.

Data analysis and interpretation: I conducted thorough analyses of the WGBS data, interpreting results to inform pipeline refinements.

Bug fixing and optimization: I was responsible for identifying and rectifying bugs in MethylStar, improving its performance and reliability.

Manuscript preparation: I wrote the first draft of the paper, with subsequent input from Frank Johannes and the other co-authors.

4.2 Publication 2. AlphaBeta

AlphaBeta: computational inference of epimutation rates and spectra from high-throughput DNA methylation data in plants.

Shahryary, Y., Symeonidi, A., Hazarika, R.R., Johanna, D., Talha, M., Brigitte, H., Thomas, v.G., Maria, C.T, Koen J.F.V., Gerald, T., Robert J.S., Johannes, F. et al. AlphaBeta: computational inference of epimutation rates and spectra from high-throughput DNA methylation data in plants.

Published in Genome Biology 21, 260 (2020).
DOI: 10.1186/s13059-020-02161-6

Summary

DNA methylation, an important chromatin modification, plays crucial roles in silencing transposable elements and regulating certain genes. However, the methylation status of individual cytosines or clusters of cytosines can sometimes change stochastically, leading to what is termed as "spontaneous epimutations". These epimutations can accumulate during plant development and aging, and some even pass through the gametes to subsequent generations. A goal in the field of plant epigenetics is to obtain accurate estimates of the rate of spontaneous epimutations, identify genetic and environmental factors that can modulate the rate, and to delineate the molecular mechanisms underlying epimutational processes. To begin to address this challenge, I have developed AlphaBeta, a computational method for estimating the rate and spectrum of spontaneous epimutations using pedigree-based DNA methylation data as input.

The software starts with the assumption that base-level or region-level DNA methylation state calls are available for each of the samples in the pedigree. These calls can be produced with MethylStar (previous chapter), or with alternative methods. AlphaBeta fits an explicit epimutation model to the DNA methylation divergence data, and relates this information to the temporal divergence of the samples, as calculated from the pedigree topology. I show that the software can be applied to

data from multi-generational mutation accumulation lines, derived either through sexual or clonal propagation. Furthermore, I demonstrate that AlphaBeta can also be used to estimate somatic epimutation rates in long-lived perennials, such as trees. In this case, AlphaBeta interprets the tree branching topology as a phylogeny of somatic cell lineages with the leaves representing the end-points of these lineages. In this case, the software calculates DNA methylation divergence between leaves and relates this information to their temporal divergence, as determined from coring data on branch/stem ages.

Application of AlphaBeta to published and new data revealed that spontaneous epimutations accumulate neutrally at the genome-wide scale, originate mainly during somatic development and that they can be used as a molecular clock for age-dating trees.

Authors' contributions

FJ and MCT conceptualized the method. YS, FJ, and RRH implemented and documented the method. FJ, YS, AS, RRH, JD, TM, BTH, and TvG analyzed the data. KV, GT, and RJS contributed materials. FJ, YS and RRH drafted the paper.

My detailed contributions:

Model Implementation: I was responsible for the development and coding of the AlphaBeta model.

Bioconductor R Package: I worked on integrating the AlphaBeta model into the Bioconductor R package, ensuring its functionality, and submitted the package to the Bioconductor repository center.

Testing: I conducted extensive testing of the AlphaBeta model across various datasets to ensure its accuracy and reliability.

Tutorial: I developed the tutorial and user guide for the AlphaBeta tool, facilitating its use by other researchers.

Drafting the Paper: I wrote the first draft of the paper, with subsequent input from Frank Johannes and the other co-authors.

5 Discussion and outlook

5.1 Improving high-throughput analysis with MethylStar

The development of MethylStar marks a notable contribution to the processing of whole-genome bisulfite sequencing (WGBS) data. Recognizing the existence of other pipelines with similar capabilities, the focus shifts to the ongoing refinement and adaptation of MethylStar, particularly in the dynamic field of genomics.

In advancing MethylStar towards its next iteration, a detailed and technical approach is required to address the complexities of genomic data processing. The integration of workflow management systems such as Nextflow [66] or Snakemake [67] into MethylStar represents a significant step forward in enhancing its usability and efficiency. While Nextflow and Snakemake are powerful tools for managing complex genomic workflows, they primarily serve as workflow orchestrators without specialized functionalities for bisulfite sequencing data. Nextflow for instance, offers a DSL (Domain-Specific Language) for parallel and distributed computational pipelines, ideal for managing complex workflows like those required for genomic data processing. Its containerization support, through technologies like Docker [68](MethylStar version 1.0 supports this technique) and Singularity [69], ensures that MethylStar can be run consistently across different computing environments. This is crucial for genomic research, where reproducibility and consistency across different labs and studies are paramount. Additionally, Nextflow's compatibility with cloud platforms and high-performance computing clusters aligns well with the proposed enhancements for MethylStar, especially in terms of scalability and adaptability to diverse computational resources. Similarly, Snakemake, another workflow management system, could also be an excellent fit for MethylStar. Known for its simplicity and flexibility, Snakemake allows for the

creation of complex workflows using a human-readable, Python-based language. It integrates seamlessly with conda and Bioconda [70], which could facilitate the management of dependencies and the installation of MethylStar. Furthermore, Snakemake's ability to automatically determine the interdependencies of tasks and efficiently schedule their execution would optimize the pipeline's performance, particularly in dealing with large and complex genomic datasets.

In the development of a user-friendly and accessible interface for MethylStar, the incorporation of a web-based platform for running analyses or generating reports would be a significant advancement. This web-based component could transform how users interact with MethylStar, making it more accessible and convenient for a broader range of researchers. A web-based platform for running MethylStar analyses would allow users to access the tool from any device with an internet connection, without the need for installing specific software or managing complex computational resources. This approach could democratize access to advanced genomic analysis, enabling researchers from institutions with limited computational infrastructure to engage in high-level genomic research. Users could simply upload their data to the platform, configure their analysis parameters through an intuitive web interface, and initiate the processing. The platform would handle the computational aspects on the backend, leveraging cloud computing resources as necessary.

A web-based reporting system could also support collaborative research efforts. Researchers could easily share their results with collaborators or advisors through secure web links, fostering collaborative analysis and discussion. Additionally, for educational purposes, such a platform could be an invaluable tool, allowing students and trainees to gain hands-on experience with genomic data analysis in a more accessible and controlled environment.

Transitioning to cloud-based computing platforms like AWS (Amazon Web Services) [71] or Google Cloud Platform [72] could provide MethylStar with enhanced scalability and flexibility. These platforms offer robust infrastructure for storing and processing large genomic datasets, allowing users to access scalable computational resources on-demand. This transition would be particularly beneficial for researchers and institutions that lack the necessary in-house computational infras-

tructure. By leveraging cloud services, MethylStar could offer a more accessible and scalable solution for genomic data processing, catering to a wider range of research needs.

Finally, ensuring that MethylStar is compatible with common cluster submission systems, such as SLURM [73] or PBS [74], is essential for its integration into high-performance computing environments. This compatibility would allow researchers to easily submit and manage MethylStar jobs on these systems, which are widely used in genomic research institutions for large-scale data processing. Enhancing MethylStar's functionality in this regard would streamline its integration into existing research workflows, making it a more versatile and user-friendly tool for genomic researchers.

Another challenge arises in integrating methylation calling for Oxford Nanopore Technologies (ONT) [75] data into the MethylStar pipeline. Integrating methylation calling for ONT data into the MethylStar pipeline requires a nuanced approach, focusing on the unique characteristics of ONT sequencing. This integration involves selecting and adapting software tools that can effectively handle the longer reads and specific error profiles associated with ONT data, which pose distinct challenges in methylation analysis.

A key aspect of this integration is the use of long-read alignment tools. These tools, for instance, Minimap2 [76], are essential for mapping the long reads of ONT data to a reference genome. However, for methylation analysis, these tools need to be specifically adapted to account for ONT's unique error profiles and methylation patterns. This means enhancing the alignment algorithms to accurately map long reads while distinguishing true methylation signals from sequencing errors inherent in ONT data.

Furthermore, error correction plays a crucial role in this process. ONT sequencing is known for higher error rates, making sophisticated error correction algorithms a necessity. Software like Nanopolish [77], which has capabilities for working with ONT data, could be a starting point. However, its algorithms would need significant enhancements to interpret the methylation context accurately from the long-read data. The goal is to develop tools that can precisely identify methylation states in the context of ONT's specific sequencing characteristics.

Additionally, there is potential in leveraging existing Whole Genome Bisulfite Sequencing (WGBS) tools as a transitional approach. Tools originally designed for bisulfite sequencing, such as BSMAP [78] or BRAT-nova [79], could be adapted or used as a foundation for developing new algorithms tailored to ONT data. The challenge here lies in modifying these tools to accommodate the length and complexity of ONT reads and their distinct methylation calling requirements.

In summary, the integration of MethylStar with ONT data for methylation analysis is a multifaceted task. It involves combining the capabilities of long-read alignment, sophisticated error correction, and methylation calling algorithms, along with adapting existing WGBS tools.

An additional advancement could be the use of deep-learning-based methods for methylation calling and imputation. MethylStar currently employs Methimpute, a first-order HMM, to achieve this. While Methimpute learns DNA methylation transitions in different cytosine contexts, it ignores other potentially informative sequence features in the genomic neighborhood during methylation calling and/or imputation. Deep learning tools such as DeepCpG [56] and DeepSignal-plant [80] employ a combination of convolutional and recurrent neural networks for enhanced accuracy in predicting methylation states across a broad range of DNA methylation datasets. DeepCpG was initially designed for imputation of single-cell DNA methylation data, which is inherently sparse due to technical limitations in NGS library preparation. Although DeepCpG has been trained on human data, it should be possible to retrain it on plant genomic data.

However, one challenge will be the analysis of non-CG methylation, which is essentially absent in mammals but is the primary type of methylation in plant genomes. Non-CG methylation patterns are not only more complex, often occurring in highly repetitive regions of the genome, but are also measured with substantially more uncertainty due to cell-to-cell heterogeneity. Indeed, the initial motivation for the development of DeepSignal-plant for ONT methylation calling was precisely to account for non-CG methylation, as previous ONT calling tools were exclusively designed around CpG data from mammalian systems. A second challenge in extending tools like DeepCpG to plants will be the generality of the trained model. Plant methylome landscapes differ substantially from others,

partly due to vast differences in genome architectures and the differential evolution of DNA methylation pathways [81]. One strategy to address this latter challenge could be to train a cross-species model and apply deep learning approaches such as transfer learning.

In summary, MethylStar Version 2 is an updated and improved version of the original tool. This update builds on the strong base of MethylStar, adding new features and enhancements that respond to the latest needs in genomic research. The enhancements in Version 2 include more powerful machine learning algorithms, better handling of multiple tasks at once, and improved compatibility with new sequencing technologies and computer systems. Version 2 also improves the user interface with a web-based platform that makes it easier to access from anywhere. By integrating well-known workflow management tools like Nextflow or Snakemake and using cloud computing, MethylStar Version 2 strengthens its ability to support genomic researchers with a tool that is more efficient, flexible, and easy to use.

5.2 Improving somatic epimutation analysis with AlphaBeta

The AlphaBeta statistical approach was originally designed to model epimutational processes in plants propagated by selfing. Fundamental to its assumptions is that epimutations pass through a single-cell bottleneck at the beginning of each generation, i.e., via the male and female gametes. We have subsequently used the same assumptions in a clonal version of this model to estimate somatic epimutations in trees. In this case, we treated the tree branching topology as a pedigree (or phylogeny) of cell lineages. Model fit diagnostics suggest that the model performed reasonably well. However, closer inspection of the biology of how the branching topology in trees is actually initiated from cell lineages suggests that several of our initial model assumptions are violated.

In higher plants, lateral branches are initiated from axillary buds, which are formed at the axils of developing leaves along the stem. The buds themselves derive from a cluster of cells known as axillary meristems (AM). Recent studies using live imaging support a "detached-meristem" model, where AMs derive from a few precursor cells sequestered from the periphery of the shoot apical meristem

(SAM) [24,82]. These precursors maintain their meristematic characteristics in the leaf axil, eventually increase in number, and develop into active AMs generating new organs. As such, they become the SAM of the newly emerging shoot and share all the functional and morphological properties with the original SAM (e.g., organization, capability of self-maintenance, and organ formation [83]. Thus, shoot branching represents a developmental transition from a SAM at the original shoot to the establishment of a new SAM on the emerging lateral branch. The sampling of AM precursors presents a major cellular bottleneck during shoot branching. However, unlike during sexual reproduction, the bottleneck is not defined by the sampling of a single cell, but rather by the selection of multiple cells (usually around 2-3). It can be shown formally [84] that this more relaxed epigenetic drift gives rise to a mixture of cell phylogenies that lead to somatic epigenetic heterogeneity in the SAM over time, as well as in the tissues derived from the SAM (e.g., leaves).

This has at least two implications: First, statistical models that account for moderate epigenetic drift could provide an improved fit to the data, and possibly lead to slightly different somatic epimutation rate estimates in trees. Second, DNA methylation data based on bulk sequencing of leaf tissues may mask the underlying cell-to-cell DNA methylation heterogeneity. Alternative sequencing approaches based on single-cell or low-input WGBS may be necessary in this setting, although such approaches remain technically challenging and often lead to highly sparse data. Nonetheless, it is worth exploring these latter approaches in future implementations of AlphaBeta.

Bibliography

[1] Law, J.A., Jacobsen, S.E.: Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics **11**(3), 204–220 (2010). doi:10.1038/nrg2719. Accessed 2019-08-01

[2] Stroud, H., Greenberg, M.V.C., Feng, S., Bernatavichute, Y.V., Jacobsen, S.E.: Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell **152**(1-2), 352–364 (2013). doi:10.1016/j.cell.2012.10.054

[3] Bewick, A.J., Hofmeister, B.T., Powers, R.A., Mondo, S.J., Grigoriev, I.V., James, T.Y., Stajich, J.E., Schmitz, R.J.: Diversity of cytosine methylation across the fungal tree of life. Nature Ecology & Evolution **3**(3), 479 (2019). doi:10.1038/s41559-019-0810-9. Accessed 2019-04-13

[4] Feng, S., Cokus, S.J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M.G., Hetzel, J., Jain, J., Strauss, S.H., Halpern, M.E., Ukomadu, C., Sadler, K.C., Pradhan, S., Pellegrini, M., Jacobsen, S.E.: Conservation and divergence of methylation patterning in plants and animals. Proceedings of the National Academy of Sciences of the United States of America **107**(19), 8689–8694 (2010). doi:10.1073/pnas.1002720107

[5] Niederhuth, C.E., Bewick, A.J., Ji, L., Alabady, M.S., Kim, K.D., Li, Q., Rohr, N.A., Rambani, A., Burke, J.M., Udall, J.A., Egesi, C., Schmutz, J., Grimwood, J., Jackson, S.A., Springer, N.M., Schmitz, R.J.: Widespread natural variation of DNA methylation within angiosperms. Genome Biology **17** (2016). doi:10.1186/s13059-016-1059-0. Accessed 2019-06-18

[6] Matzke, M., Kanno, T., Daxinger, L., Huettel, B., Matzke, A.J.M.: Rna-mediated chromatin-based silencing in plants **21**, 367–376 (2009). doi:10.1016/j.ceb.2009.01.025

[7] Borges, F., Martienssen, R.A.: The expanding world of small rnas in plants **16**, 727–741 (2015). doi:10.1038/nrm4085

[8] Saze, H., Scheid, O.M., Paszkowski, J.: Maintenance of cpg methylation is essential for epigenetic inheritance during plant gametogenesis **34**, 65–69 (2003). doi:10.1038/ng1138

[9] Johnson, L.M., Bostick, M., Zhang, X., Kraft, E., Henderson, I., Callis, J., Jacobsen, S.E.: The sra methyl-cytosine-binding domain links dna and histone methylation **17**, 379–384 (2007). doi:10.1016/j.cub.2007.01.009

[10] Stroud, H., Do, T., Du, J., Zhong, X., Feng, S., Johnson, L., Patel, D.J., Jacobsen, S.E.: Non-cg methylation patterns shape the epigenetic landscape in arabidopsis **21**, 64–72 (2014). doi:10.1038/nsmb.2735

[11] To, T.K., Kakutani, T.: Crosstalk among pathways to generate dna methylome **68**, 102248 (2022). doi:10.1016/j.pbi.2022.102248

[12] Erdmann, R.M., Picard, C.L.: RNA-directed DNA methylation. PLOS Genetics **16**(10), 1009034 (2020). doi:10.1371/journal.pgen.1009034

[13] Zemach, A., Kim, M.Y., Hsieh, P.-H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S.L., Zilberman, D.: The arabidopsis nucleosome remodeler ddm1 allows dna methyltransferases to access h1-containing heterochromatin **153**, 193–205 (2013). doi:10.1016/j.cell.2013.02.033

[14] Kakutani, T., Jeddeloh, J.A., Richards, E.J.: Characterization of anarabidopsis thalianadna hypomethylation mutant **23**, 130–137 (1995). doi:10.1093/nar/23.1.130

[15] Lippman, Z., Gendrel, A.-V., Black, M., Vaughn, M.W., Dedhia, N., Richard McCombie, W., Lavine, K., Mittal, V., May, B., Kasschau, K.D., Carrington, J.C., Doerge, R.W., Colot, V., Martienssen, R.: Role of transposable elements in heterochromatin and epigenetic control. **430**(6998), 471–476 (2004). doi:10.1038/nature02651

[16] Vongs, A., Kakutani, T., Martienssen, R.A., Richards, E.J.: Arabidopsis thaliana dna methylation mutants. Science **260**(5116), 1926–1928 (1993). doi:10.1126/science.8316832

[17] Gendrel, A.-V., Lippman, Z., Yordan, C., Colot, V., Martienssen, R.A.: Dependence of heterochromatic histone h3 methylation patterns on the arabidopsis gene ddm1 **297**, 1871–1873 (2002). doi:10.1126/science.1074950

[18] Soppe, W.J.J.: Dna methylation controls histone h3 lysine 9 methylation and heterochromatin assembly in arabidopsis **21**, 6549–6559 (2002). doi:10.1093/emboj/cdf657

[19] Field, A.E., Robertson, N.A., Wang, T., Havas, A., Ideker, T., Adams, P.D.: DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. Molecular Cell **71**(6), 882–895 (2018). doi:10.1016/j.molcel.2018.08.008. Accessed 2019-04-13

[20] Zubko, E., Gentry, M., Kunova, A., Meyer, P.: De novo dna methylation activity of methyltransferase 1 (met1) partially restores body methylation in arabidopsis thaliana **71**, 1029–1037 (2012). doi:10.1111/j.1365-313x.2012.05051.x

[21] Jeltsch, A., Adam, S., Dukatz, M., Emperle, M., Bashtrykov, P.: Deep enzymology studies on dna methyltransferases reveal novel connections between flanking sequences and enzyme activity **433**, 167186 (2021). doi:10.1016/j.jmb.2021.167186

[22] Haggerty, C., Kretzmer, H., Riemenschneider, C., Kumar, A.S., Mattei, A.L., Bailly, N., Gottfreund, J., Giesselmann, P., Weigert, R., Brändl, B., Giehr, P., Buschow, R., Galonska, C., von Meyenn, F., Pappalardi, M.B., McCabe, M.T., Wittler, L., Giesecke-Thiel, C., Mielke, T., Meierhofer, D., Timmermann, B., Müller, F.-J., Walter, J., Meissner, A.: Dnmt1 has de novo activity targeted to transposable elements **28**, 594–603 (2021). doi:10.1038/s41594-021-00603-8

[23] Yao, N., Schmitz, R.J., Johannes, F.: Epimutations define a fast-ticking molecular clock in plants **37**, 699–710 (2021). doi:10.1016/j.tig.2021.04.010

[24] Burian, A., Barbier de Reuille, P., Kuhlemeier, C.: Patterns of stem cell divisions contribute to plant longevity **26**, 1385–1394 (2016). doi:10.1016/j.cub.2016.03.067

[25] Szymkowiak, E.J., Sussex, I.M.: What chimeras can tell us about plant development **47**, 351–376 (1996). doi:10.1146/annurev.arplant.47.1.351

[26] Hofmeister, B.T., Denkena, J., Colomé-Tatché, M., Shahryary, Y., Hazarika, R., Grimwood, J., Mamidi, S., Jenkins, J., Grabowski, P.P., Sreedasyam, A., Shu, S., Barry, K., Lail, K., Adam, C., Lipzen, A., Sorek, R., Kudrna, D., Talag, J., Wing, R., Hall, D.W., Jacobsen, D., Tuskan, G.A., Schmutz, J., Johannes, F., Schmitz, R.J.: A genome assembly and the somatic genetic and epigenetic mutation rate in a wild long-lived perennial populus trichocarpa. Genome Biology **21**(1), 259 (2020). doi:10.1186/s13059-020-02162-5

[27] Herrera, C.M., Bazaga, P., Pérez, R., Alonso, C.: Lifetime genealogical divergence within plants leads to epigenetic mosaicism in the shrub lavandula latifolia (lamiaceae) **231**, 2065–2076 (2021). doi:10.1111/nph.17257

[28] Wang, L., Ji, Y., Hu, Y., Hu, H., Jia, X., Jiang, M., Zhang, X., Zhao, L., Zhang, Y., Jia, Y., Qin, C., Yu, L., Huang, J., Yang, S., Hurst, L.D., Tian, D.: The architecture of intra-organism mutation rate variation in plants **17**, 3000191. doi:10.1371/journal.pbio.3000191

[29] Hanlon, V.C.T., Otto, S.P., Aitken, S.N.: Somatic mutations substantially increase the per-generation mutation rate in the conifer picea sitchensis **3**, 348–358 (2019). doi:10.1002/evl3.121

[30] Schmid-Siegert, E., Sarkar, N., Iseli, C., Calderon, S., Gouhier-Darimont, C., Chrast, J., Cattaneo, P., Schütz, F., Farinelli, L., Pagni, M., Schneider, M., Voumard, J., Jaboyedoff, M., Fankhauser, C., Hardtke, C.S., Keller, L., Pannell, J.R., Reymond, A., Robinson-Rechavi, M., Xenarios, I., Reymond, P.: Low number of fixed somatic mutations in a long-lived oak tree. Nature Plants **3**(12), 926 (2017). doi:10.1038/s41477-017-0066-9. Accessed 2019-06-26

[31] Orr, A.J., Padovan, A., Kainer, D., Külheim, C., Bromham, L., Bustos-Segura, C., Foley, W., Haff, T., Hsieh, J.-F., Morales-Suarez, A., Cartwright, R.A., Lanfear, R.: A phylogenomic approach reveals a low somatic mutation rate in a long-lived plant. bioRxiv, 727982 (2019). doi:10.1101/727982. Accessed 2019-09-04

[32] Yu, L., Boström, C., Franzenburg, S., Bayer, T., Dagan, T., Reusch, T.B.H.: Somatic genetic drift and multi-level selection in modular species. bioRxiv, 833335 (2019). doi:10.1101/833335. Accessed 2019-11-09

[33] McDaniel, C.N., Poethig, R.S.: Cell-lineage patterns in the shoot apical meristem of the germinating maize embryo **175**, 13–22 (1988). doi:10.1007/bf00402877

[34] Kawashima, T., Berger, F.: Epigenetic reprogramming in plant sexual reproduction **15**, 613–624 (2014). doi:10.1038/nrg3685

[35] Taudt, A., Colomé-Tatché, M., Johannes, F.: Genetic sources of population epigenomic variation. Nature Reviews. Genetics **17**(6), 319–332 (2016). doi:10.1038/nrg.2016.45

[36] Johannes, F., Schmitz, R.J.: Spontaneous epimutations in plants. New Phytologist **221**(3), 1253–1259 (2019). doi:10.1111/nph.15434. Accessed 2019-02-04

[37] Graaf, A.v.d., Wardenaar, R., Neumann, D.A., Taudt, A., Shaw, R.G., Jansen, R.C., Schmitz, R.J., Colomé-Tatché, M., Johannes, F.: Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. Proceedings of the National Academy of Sciences **112**(21), 6676–6681 (2015). doi:10.1073/pnas.1424254112. Accessed 2019-08-01

[38] Schmitz, R.J., Schultz, M.D., Lewsey, M.G., O'Malley, R.C., Urich, M.A., Libiger, O., Schork, N.J., Ecker, J.R.: Transgenerational epigenetic instability is a source of novel methylation variants. Science (New York, N.Y.) **334**(6054), 369–373 (2011). doi:10.1126/science.1212959

[39] Hofmeister, B.T., Lee, K., Rohr, N.A., Hall, D.W., Schmitz, R.J.: Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation. Genome Biology **18**(1), 155 (2017). doi:10.1186/s13059-017-1288-x. Accessed 2019-08-01

[40] Jiang, C., Mithani, A., Belfield, E.J., Mott, R., Hurst, L.D., Harberd, N.P.: Environmentally responsive genome-wide accumulation of de novo Arabidopsis thaliana mutations and epimutations. Genome Research **24**(11), 1821–1829 (2014). doi:10.1101/gr.177659.114. Accessed 2019-08-01

[41] Ganguly, D.R., Crisp, P.A., Eichten, S.R., Pogson, B.J.: The Arabidopsis DNA Methylome Is Stable under Transgenerational Drought Stress.

Plant Physiology **175**(4), 1893–1912 (2017). doi:10.1104/pp.17.00744. Accessed 2019-06-12

[42] Yao, N., Zhang, Z., Yu, L., Hazarika, R., Yu, C., Jang, H., Smith, L.M., Ton, J., Liu, L., Stachowicz, J.J., Reusch, T.B.H., Schmitz, R.J., Johannes, F.: An evolutionary epigenetic clock in plants. Science **381**(6665), 1440–1445 (2023). doi:10.1126/science.adh9443

[43] Zheng, X., Chen, L., Xia, H., Wei, H., Lou, Q., Li, M., Li, T., Luo, L.: Transgenerational epimutations induced by multi-generation drought imposition mediate rice plant's adaptation to drought condition. Scientific Reports **7**, 39843 (2017). doi:10.1038/srep39843. Accessed 2019-08-01

[44] Lanfear, R.: Do plants have a segregated germline? PLOS Biology **16**(5), 2005439 (2018). doi:10.1371/journal.pbio.2005439. Accessed 2019-06-26

[45] Horvath, R., Laenen, B., Takuno, S., Slotte, T.: Single-cell expression noise and gene-body methylation in Arabidopsis thaliana. Heredity, 1 (2019). doi:10.1038/s41437-018-0181-z. Accessed 2019-06-18

[46] Secco, D., Wang, C., Shou, H., Schultz, M.D., Chiarenza, S., Nussaume, L., Ecker, J.R., Whelan, J., Lister, R.: Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements. eLife **4** (2015). doi:10.7554/eLife.09343

[47] Müller, F., Scherer, M., Assenov, Y., Lutsik, P., Walter, J., Lengauer, T., Bock, C.: Rnbeads 2.0: comprehensive analysis of dna methylation data. Genome biology **20**(1), 55 (2019)

[48] Maksimovic, J., Gordon, L., Oshlack, A.: Swan: Subset-quantile within array normalization for illumina infinium humanmethylation450 beadchips. Genome biology **13**(6), 44 (2012)

[49] Tian, Y., Morris, T.J., Webster, A.P., Yang, Z., Beck, S., Feber, A., Teschendorff, A.E.: Champ: updated methylation analysis pipeline for illumina beadchips. Bioinformatics **33**(24), 3982–3984 (2017)

[50] Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A., Mason, C.E.: methylkit: a comprehensive r package for the

analysis of genome-wide dna methylation profiles. Genome biology **13**(10), 87 (2012)

[51] Catoni, M., Tsang, J.M., Greco, A.P., Zabet, N.R.: Dmrcaller: a versatile r/bioconductor package for detection and visualization of differentially methylated regions in cpg and non-cpg contexts. Nucleic acids research **46**(19), 114–114 (2018)

[52] Schultz, M.D., He, Y., Whitaker, J.W., Hariharan, M., Mukamel, E.A., Leung, D., Rajagopal, N., Nery, J.R., Urich, M.A., Chen, H., Lin, S., Lin, Y., Jung, I., Schmitt, A.D., Selvaraj, S., Ren, B., Sejnowski, T.J., Wang, W., Ecker, J.R.: Human body epigenome maps reveal noncanonical dna methylation variation. Nature **523**(7559), 212–216 (2015)

[53] Jühling, F., Kretzmer, H., Bernhart, S.H., Otto, C., Stadler, P.F., Hoffmann, S.: metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. Genome research **26**(2), 256–262 (2016)

[54] Taudt, A., Roquis, D., Vidalis, A., Wardenaar, R., Johannes, F., Colomé-Tatché, M.: METHimpute: imputation-guided construction of complete methylomes from WGBS data. BMC Genomics **19**(1), 444 (2018). doi:10.1186/s12864-018-4641-x. Accessed 2019-08-02

[55] Kapourani, C.-A., Sanguinetti, G.: Melissa: Bayesian clustering and imputation of single-cell methylomes. Genome biology **20**(1), 61 (2019)

[56] Angermueller, C., Lee, H.J., Reik, W., Stegle, O.: Deepcpg: accurate prediction of single-cell dna methylation states using deep learning. Genome biology **18**(1), 67 (2017)

[57] Peng, T., Zhu, Q., Yin, P., Tan, K.: Scrabble: single-cell rna-seq imputation constrained by bulk rna-seq data. Genome biology **20**(1), 88 (2019)

[58] Shahryary, Y., Hazarika, R.R., Johannes, F.: Methylstar: A fast and robust pre-processing pipeline for bulk or single-cell whole-genome bisulfite sequencing data. BMC Genomics **21**(1), 479 (2020)

[59] FastQC. https://www.bioinformatics.babraham.ac.uk/projects/fastqc

[60] Martin, M.: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**(1), 10. doi:10.14806/ej.17.1.200

[61] Krueger, F., Andrews, S.R.: Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. Bioinformatics (Oxford, England) **27**(11), 1571–1572 (2011)

[62] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and, R.D.: The sequence alignment/map format and samtools. Bioinformatics **25**(16), 2078–2079 (2009). doi:10.1093/bioinformatics/btp352

[63] Becker, C., Hagmann, J., Müller, J., Koenig, D., Stegle, O., Borgwardt, K., Weigel, D.: Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. Nature **480**(7376), 245–249 (2011). doi:10.1038/nature10555. Accessed 2019-08-01

[64] Colomé-Tatché, M., Johannes, F.: Signatures of Dobzhansky–Muller Incompatibilities in the Genomes of Recombinant Inbred Lines. Genetics **202**(2), 825–841 (2016). doi:10.1534/genetics.115.179473. Accessed 2020-06-14

[65] Merkel, A., Fernández-Callejo, M., Casals, E., Marco-Sola, S., Schuyler, R., Gut, I.G., Heath, S.C.: gemBS: high throughput processing for DNA methylation data from bisulfite sequencing. Bioinformatics **35**(5), 737–742 (2018). doi:10.1093/bioinformatics/bty690. http://oup.prod.sis.lan/bioinformatics/article-pdf/35/5/737/27994742/bty690.pdf

[66] Nextflow. https://www.nextflow.io/docs/latest/index.html

[67] Snakemake. https://snakemake.readthedocs.io/en/stable/

[68] Docker. https://docs.docker.com/

[69] Singularity. https://docs.sylabs.io/guides/2.6/user-guide/index.html#

[70] Bioconda. https://bioconda.github.io/

[71] AWS. http://aws.amazon.com/

[72] GCC. https://cloud.google.com/

[73] SLURM. https://slurm.schedmd.com/documentation.html

[74] PBS. https://openpbs.org/

[75] Oxford Nanopore Technologies. https://nanoporetech.com/platform/technology

[76] Li, H.: New strategies to improve minimap2 alignment accuracy. arXiv e-prints, 2108–03515 (2021). doi:10.48550/arXiv.2108.03515. 2108.03515

[77] NANOPOL. https://github.com/jts/nanopolish

[78] Xi, Y., Li, W.: Bsmap: whole genome bisulfite sequence mapping program. BMC bioinformatics **10**(1), 232 (2009)

[79] Harris, E.Y., Ounit, R., Lonardi, S.: Brat-nova: fast and accurate mapping of bisulfite-treated reads. Bioinformatics **32**(17), 2696–2698 (2016)

[80] Ni, P., Huang, N., Nie, F., Zhang, J., Zhang, Z., Wu, B., Bai, L., Liu, W., Xiao, C.-L., Luo, F., Wang, J.: Genome-wide detection of cytosine methylations in plant from nanopore data using deep learning. Nature Communications **12**(1) (2021). doi:10.1038/s41467-021-26278-9

[81] Bewick, A.J., Schmitz, R.J.: Gene body dna methylation in plants. Current Opinion in Plant Biology **36**, 103–110 (2017). doi:10.1016/j.pbi.2016.12.007

[82] Shi, B., Zhang, C., Tian, C., Wang, J., Wang, Q., Xu, T., Xu, Y., Ohno, C., Sablowski, R., Heisler, M.G., Theres, K., Wang, Y., Jiao, Y.: Two-step regulation of a meristematic cell population acting in shoot branching in arabidopsis. PLOS Genetics **12**(7), 1006168 (2016). doi:10.1371/journal.pgen.1006168

[83] Nicolas, A., Laufs, P.: Meristem initiation and de novo stem cell formation. Frontiers in Plant Science **13** (2022). doi:10.3389/fpls.2022.891228

[84] Chen, Y., Burian, A., Johannes, F.: Somatic epigenetic drift during shoot branching: a cell lineage-based model. bioRxiv (2024). doi:10.1101/2024.01.24.577071. https://www.biorxiv.org/content/early/2024/01/29/2024.01.24.577071.full.pdf

List of Figures

Acronyms

5mC	5-methyl cytosine. 1
CMT2	CHROMOMETHYLASE 2. 1
CX-reports	Cytosine context (CG, CHG, CHH) report for all cytosines. 1
DDM1	Deficient in DNA Methylation 1. 1
DMRs	Differentially Methylated Regions. 1
EpiDiverse	Epigenetic Diversity in Ecology. 1
epiGBS	epigenotyping by sequencing. 1
FDR	False Discovery Rate. 1
HMM	Hidden Markov Model. 1
IHEC	International Human Epigenome Consortium. 1
MA lines	Mutation Accumulation lines. 1
NGS	Next Generation Sequencing. 1
NIH ROADMAP	National Institutes of Health Roadmap for Medical Research. 1
PBAT	Post-bisulfite Adaptor Tagging. 1
PCR	Polymerase Chain Reaction. 1

QC	Quality Control. 1
RdDM	RNA-directed DNA methylation pathway. 1
RRBS	Reduced Representation Bisulphite Sequencing. 1
SYSCID	A European Network for Systematic Gastrointestinal Disease Data Collection. 1
TEs	Transposable elements. 1
WGBS	Whole-Genome Bisulfite Sequencing. 1

A Appendix I: MethylStar paper reprint

Methlystar: A fast and robust pre-processing pipeline for bulk or single-cell whole-genome bisulfite sequencing data.

Shahryary *et al. BMC Genomics* (2020) 21:479
https://doi.org/10.1186/s12864-020-06886-3

BMC Genomics

SOFTWARE Open Access

MethylStar: A fast and robust pre-processing pipeline for bulk or single-cell whole-genome bisulfite sequencing data

Yadollah Shahryary[1,2], Rashmi R. Hazarika[1,2] and Frank Johannes[1,2*]

Abstract

Background: Whole-Genome Bisulfite Sequencing (WGBS) is a Next Generation Sequencing (NGS) technique for measuring DNA methylation at base resolution. Continuing drops in sequencing costs are beginning to enable high-throughput surveys of DNA methylation in large samples of individuals and/or single cells. These surveys can easily generate hundreds or even thousands of WGBS datasets in a single study. The efficient pre-processing of these large amounts of data poses major computational challenges and creates unnecessary bottlenecks for downstream analysis and biological interpretation.

Results: To offer an efficient analysis solution, we present MethylStar, a fast, stable and flexible pre-processing pipeline for WGBS data. MethylStar integrates well-established tools for read trimming, alignment and methylation state calling in a highly parallelized environment, manages computational resources and performs automatic error detection. MethylStar offers easy installation through a dockerized container with all preloaded dependencies and also features a user-friendly interface designed for experts/non-experts. Application of MethylStar to WGBS from Human, Maize and *A. thaliana* shows favorable performance in terms of speed and memory requirements compared with existing pipelines.

Conclusions: MethylStar is a fast, stable and flexible pipeline for high-throughput pre-processing of bulk or single-cell WGBS data. Its easy installation and user-friendly interface should make it a useful resource for the wider epigenomics community. MethylStar is distributed under GPL-3.0 license and source code is publicly available for download from github https://github.com/jlab-code/MethylStar. Installation through a docker image is available from http://jlabdata.org/methylstar.tar.gz

Keywords: DNA methylation, Whole genome bisulfite sequencing, NGS, Pipeline, Single cell

Background

Whole-Genome Bisulfite Sequencing (WGBS) is a Next Generation Sequencing (NGS) technique for measuring DNA methylation at base resolution. As a result of continuing drops in sequencing costs, an increasing number of laboratories and international consortia (e.g. IHEC, SYSCID, BLUEPRINT, EpiDiverse, NIH ROADMAP, Arabidopsis 1001 Epigenomes, Genomes and physical Maps) are adopting WGBS as the method of choice to survey DNA methylation in large population samples or in collections of cell lines and tissue types, either in bulk or at the single-cell level [1, 2]. Such surveys can easily generate hundreds or even thousands of WGBS datasets in a single study. A broad array of software solutions for the downstream analysis of bulk and single-cell WGBS data have been developed in recent years. These include tools for data normalization (e.g. RnBeads [3], SWAN [4], ChAMP [5]), detection of differentially methylated regions (DMRs) (e.g. Methylkit [6], DMRcaller [7], Methylpy [8], metilene [9]), imputation

*Correspondence: frank@johanneslab.org
[1]Technical University of Munich, Institute for Advanced Study (IAS), Lichtenbergstr. 2a, 85748 Garching, Germany
[2]Technical University of Munich, Department of Plant Sciences, Liesel-Beckmann-Str. 2, 85354 Freising, Germany

of methylomes from bulk WGBS data (e.g. METHimpute [10]), imputation of single-cell methylomes (e.g. Melissa [11], deepCpG [12]) and dropouts in single-cell data (e.g. SCRABBLE [13]).

However, these downstream analysis tools are dependent on the output of a number of data pre-processing steps, such as quality control (e.g. FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc), QualiMap [14], NGS QC toolkit [15]), de-multiplexing of sequence reads, adapter trimming (e.g. Trimmomatic [16], TrimGalore (https://github.com/FelixKrueger/TrimGalore) , Cutadapt [17]), alignment of reads to a reference genome and generation of methylation calls (e.g. BSseeker2 [18], BSseeker3 [19], Bismark [20], BSMap [21], bwa-meth (https://github.com/brentp/bwa-meth/) , BRAT-nova [22], BiSpark [23], WALT [24], segemehl [25]). From a computational standpoint, data pre-processing is by far the most time-consuming step in the entire bulk or single-cell WGBS analysis workflow (Fig.1). In an effort to help streamline the pre-processing of WGBS data several pipelines have been published in recent years. These include nf-core/methylseq [26], gemBS [27], Bicycle [28] and Methylpy, some of which are currently employed by several epigenetic consortia. gemBS, Bicycle and Methylpy integrate data pre-processing and analysis steps using their own custom trimming and/or alignment tools (see Table 1). By contrast, nf-core/methylseq implements well-established NGS tools, such as TrimGalore for read trimming and Bismark and bwa-meth/MethylDackel for alignment. The nf-core/methylseq framework is built using Nextflow [29], and aims to provide reproducible pipeline templates that can be easily adapted by both developers as well as experimentalists. Despite these efforts, the installation and execution of these pipelines is not trivial and often require substantial bioinformatic support. Moreover, managing the run times of these pipelines for large numbers of WGBS datasets (i.e. in the order of hundreds or thousands) relies on substantial manual input, such as launching of parallel jobs on a compute cluster and collecting output files from temporary folders.

In an attempt to address these issues, we have developed MethylStar, a fast, stable and flexible pre-processing pipeline for WGBS data. MethylStar integrates well-established NGS tools for read trimming, alignment and methylation state calling in a highly parallelized environment, manages computational resources and performs automatic error detection. MethylStar offers easy installation through a dockerized container with all preloaded dependencies and also features a user-friendly interface designed for experts/non-experts. Application of MethylStar to WGBS from Human, Maize and *A. thaliana* shows favorable performance in terms of speed and memory requirements compared with existing pipelines.

Implementation

Core pipeline NGS components

In its current implementation, MethylStar integrates processing of raw fastq reads for both single- and paired-end data with options for adapter trimming, quality control (fastQC) and removal of PCR duplicates (Bismark software suite). Read alignment and cytosine context extraction is performed with the Bismark software suite. Alignments can be performed for WGBS and Post-bisulfite adaptor tagging (PBAT) approaches for single-cell libraries. Bismark was chosen because it features one of the most sensitive aligners, resulting in comparatively high mapping efficiency, low mapping bias and good genomic coverage [30, 31]. Finally, cytosine-level methylation calls are (optionally) obtained with METHimpute, a Hidden Markov Model for inferring the methylation status/level of individual cytosines, even in the presence of low sequencing depth and/or missing data. All the different data processing steps have been optimized for speed and performance (see below), and can run on local machines as well as on larger compute nodes.

User interface

MethylStar features a lightweight python-based user interface, which is particularly useful for bench-scientists who are not familiar with command-line scripting. The aim of the interface is to improve useability and to reduce human error arising from typing mistakes or from the misspecification of parameter settings during pipeline configuration. The interface offers configuration templates that can be easily re-used for subsequent samples/projects, thus ensuring consistency and repeatability of data analysis projects. Unlike many web-based or graphical-based interfaces, the MethylStar interface does not require additional resources and/or dependencies. Users navigate through an index menu and run selected pipeline components by typing the menu index of choice. We designed the interface for both experts and non-experts. Non-experts are able to execute all pipeline commands without having to edit a single bash script, while advanced users can easily configure additional parameters and install software/tools (e.g. most recent/legacy version of a software) to integrate with MethylStar by simply specifying path variables. Finally, users can configure email addresses to receive automatic notifications when a job completed or failed. A video demonstrating the use of the interface can be found at https://github.com/jlab-code/MethylStar#MethylStar_tutorial_on_YouTube.

Pipeline architecture, optimization of parallel processes and memory usage

The pipeline architecture comprises three main layers (Fig. 1). The first layer is the interactive command-line user interface implemented in Python to simplify the

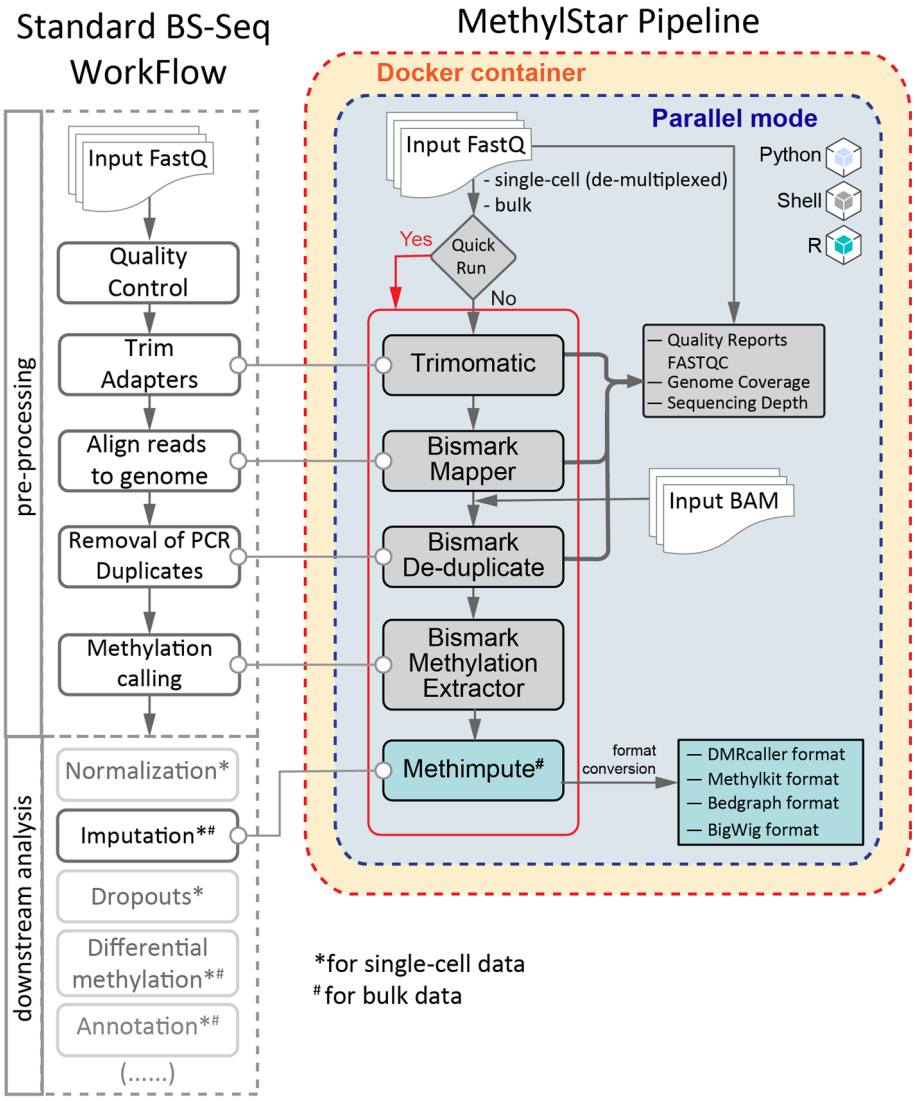

Fig. 1 Basic workflow of MethylStar showing the pipeline architecture. The left panel shows a standard BS-Seq workflow and on the right are the different components of the MethylStar pipeline integrated as 3 different layers viz. Python, Shell and R. All steps of the pipeline have been parallelized using GNU parallel. MethylStar offers the option for "Quick run" (indicated in red) which runs all steps sequentially in one go or each component can be executed separately. MethylStar incorporates all pre-processing steps of a standard BS-Seq workflow and generates standard outputs that can be used for input into several downstream analysis tools

process of configuring software settings and running MethylStar. The second layer consists of shell scripts, and handles low-level processes, efficiently coordinates the major software components and manages computational resources. The final layer is implemented in R, and is used to call METHimpute and to generate output files that are compatible with a number of publicly available DMR-callers such as Methylkit, DMRcaller and bigWig files for visualization in Genome Browsers such as JBrowse [32]. All outputs are provided in standard data formats for downstream analysis.

All components/steps of the pipeline have been parallelized using GNU Parallel (https://www.gnu.org/software/parallel/) (Fig. 1). The user can either set the number of parallel jobs manually for each pipeline component, or can opt to use the inbuilt parallel option from the "configuration" option of the menu. The inbuilt parallel implementation is also available under the "Quick Run" option. This latter option detects the number of parallel processes/jobs automatically for each pipeline component based on available system cores/threads and memory, thus allowing the user to run the entire steps of the pipeline in one go.

In the parallel implementation of all pipeline steps, we use genome size (in base pairs) as an additional factor in the optimization of computational resources. For

Table 1 Table showing different features of MethylStar as compared to other BS-seq pipelines

	Methylpy	MethylStar	methylseq	gemBS	Bicycle
Pipeline Features					
Multi-threading		√	√	√	√
language	Python	Python, shell, R	Java	C, Python	Java
distribution	github, PyPI (Apache license)	GitHub (GNU GPL3)	Github (MIT license)	GitHub (GNU GPL3)	Github (GNU GPL3)
Installation & configuration	pip install, install dependencies	Docker, install dependencies	Docker, Singularity, Conda	Docker, Singularity	Docker
User-interface	-	√	-	-	-
Single/paired-end	√	√	√	√	√
Input data	Single-cell, WGBS, singlecell NOMe-seq, PBAT	WGBS, Single-cell (PBAT)	WGBS	RRBS, WGBS, PBAT	WGBS
Pipe steps					
adapter trimming	Cutadapt	Trimmomatic	TrimGalore	-	bicycle analyzemethylation
alignment	bowtie/bowtie2	Bismark	Bismark, bwa-meth	gem3	bicycle align/ bowtie/bowtie2
remove PCR duplicates	Picard	Bismark	Bismark, Picard	Bscall	bicycle analyzemethylation
methylation calling	√	ProcessBismarAln, Bismark	Bismark, MethylDackel	Bscall	bicycle analyzemethylation, GATK
imputation of missing cytosines	-	METHimpute	-	-	-
DMR calling	√	-	-	-	bicycle analyze differential methylation
SNP calling	-	-	-	Bscall	-
Alignment QC	-	Bismark	Qualimap	√	√
summary reports	√	FastQC	Bismark, MultiQC, Preseq	√	√
Methylation visualization	BigWig	BigWig, bedGraph	-	BigWig, bedGraph	BigWig

example, in the analysis of *A. thaliana* samples (genome size ∼135 mega base pairs), our parallel implementation of Trimmomatic (a java tool) sets the optimal number of jobs to 12 on a system with 88 cores and 386 GB RAM. This setting allocates (12 jobs ×8 threads) =96 threads for trimming (java threads) and (12 jobs ×1 threads) =12 threads to the gzip tools (default no. of threads fixed to 8 in the pipeline). By contrast, for read trimming in Maize (genome size ∼2500 mega base pairs), the optimal number of jobs is set to 5. In the parallel implementation of Bismark alignment step under a similar system configuration, while running paired-end reads from *A. thaliana*, we optimally set the number of jobs to 4. This setting allocates (4 jobs ×8 files/threads) =32 threads to Bowtie2 and (4 jobs ×8 files/threads ×2) =64 threads to the bismark alignment tool (default no. of threads fixed to 8 in the internal bismark parallel argument). In a similar way, for deduplicate_bismark, the optimal number of jobs is set to (1/4th of total 88 cores) =22. For bismark_methylation_extractor it is set as 4, which allocates (4 jobs ×8 threads) =32 threads each to itself and to Bowtie tools as well as a few additional cores to gzip and samtools streams. In this way, the maximum number of threads never exceeds the total number of available cores, which in turn allows other jobs such as file compression, I/O operations to be performed simultaneously. Under the

"Quick Run" option we have parallelized R processes such as the extraction of methylation calls from BAM files (post PCR duplicates removal) by bypassing the Bismark methylation extractor step and by passing these calls directly onto METHimpute for imputation of missing cytosines (Fig. 1).

Automatic error handling and detection

MethylStar issues user-friendly messages related to configuration errors such as non-existing paths to input/output folders, low disk space, incorrect file extensions, non-empty folders. In addition, we have introduced checkpoints for each individual component of the pipeline so that a job can be resumed easily from the nearest checkpoint in the unlikely event of system failure (e.g. disk issues, file corruption, user interruption). MethylStar accepts intermediate files such as BAM files, CX-reports etc., and is able to process these new files together with pre-existing files in the folder. MethylStar issues user-friendly warnings before resuming each run. For instance, if a given folder is non-empty it will ask for user permission to continue, and issues a message that files with pre-existing names will be overwritten.

Running MethylStar

The user can choose to run each pipeline component individually, and customize software settings at each step by editing the configuration file, which is available as an option through the interactive command-line user interface. The user interface displays the available options as an index menu, and users can execute specific pipeline steps. Some of the key configuration parameters include setting file paths to input and output data, options for handling large batches of samples, file format conversions, as well as options for deleting auxiliary files that are generated during intermediate analysis steps. Our interactive user interface aids in the fast execution of complex commands and will be particularly effective for users who are less familiar with command line scripting. As an alternative, MethylStar also features a "Quick Run option", which allows the user to run all pipeline steps in one go using default configuration settings (Fig. 1).

Installation and documentation

MethylStar can be easily installed via a Docker image. This includes all the softwares, libraries and packages within the container, and thus solves any dependency issues. Advanced users can edit the existing docker container and build their own image.

Detailed description about installation and running the pipeline is available at https://github.com/jlab-code/MethylStar.

Results and discussion

Benchmarking of speed

To demonstrate MethylStar's performance we analyzed bulk WGBS data from a selection of 200 *A. thaliana* ecotypes (paired-end, 295 GB, ~ 8.63× depth, 85.66% genome coverage, GSE54292), 75 Maize strains (paired-end, 209 GB, ~ 0.36× depth, ~22.12% genome coverage, GSE39232) and 88 Human H1 cell lines (single-end, 82 GB, ~ 0.12× depth, ~10.62% genome coverage, GSM429321). MethylStar was compared with Methylpy, nf-core/methylseq and gemBS. All pipelines were run with default parameters on a computing cluster with a total of 88 cores (CPU 2.2 GHz with 378 GB RAM). Speed performance was assessed for a series of batch sizes (*A. thaliana*: 50, 100, 150, 200 samples; Human H1 cell line: 22, 44, 66, 88 samples; Maize: 15, 30, 45, 60, 75 samples) and was restricted to a fixed number of jobs (=32), (Fig. 2a-c and Additional file 1: Table S2). Although gemBS achieved the fastest processing times for the *A. thaliana* samples, MethylStar clearly outperformed the other pipelines when applied to the more complex genomes of Maize and Human, which are computationally more expansive and resource-demanding (Fig. 2b-c). For instance, for 88 Human WGBS samples (82 GB of data), MethylStar showed a 75.61% reduction in processing time relative to gemBS, the second fastest pipeline (~909 mins vs. ~3727 mins). Extrapolating from these numbers, we expect that for 1000 Human WGBS samples, MethylStar could save about ~22.24 days of run time (4× faster). To show that MethylStar can also be applied to single-cell WGBS data, we analyzed DNA methylation of 200 single cells from Human early embryo tissue (paired-end, 845 GB, ~ 0.38× depth, ~9.97% genome coverage, GSE81233) split into batches of 100 and 200 (Fig. 2d and Additional file 1: Table S2). MethylStar's processing times were compared to Methylpy which also supports single-cell data. For 100 cells, MethylStar required only ~2225 mins as compared to ~5518 mins required by Methylpy. Hence, MethylStar presents an efficient analysis solution for deep single-cell WGBS experiments.

To demonstrate that MethylStar's processing speed does not come at the expense of poor read alignments, we analysed the read mapping statistics of 50 samples each of *A. thaliana*, Maize, Human H1 cell line and single-cell Human data using MethylStar, Methylpy, nf-core/methylseq and gemBS. Our results show that MethylStar and nf-core/methylseq, both of which employ the Bismark alignment tool, provide the most accurate and sensitive alignments. This observation that is consistent with recent benchmarking results [30, 31]. By contrast, Methylpy and gemBS use their own inbuilt aligners and generally display poorer alignment statistics. Interestingly, although gemBS was the fastest pipeline for the *A.*

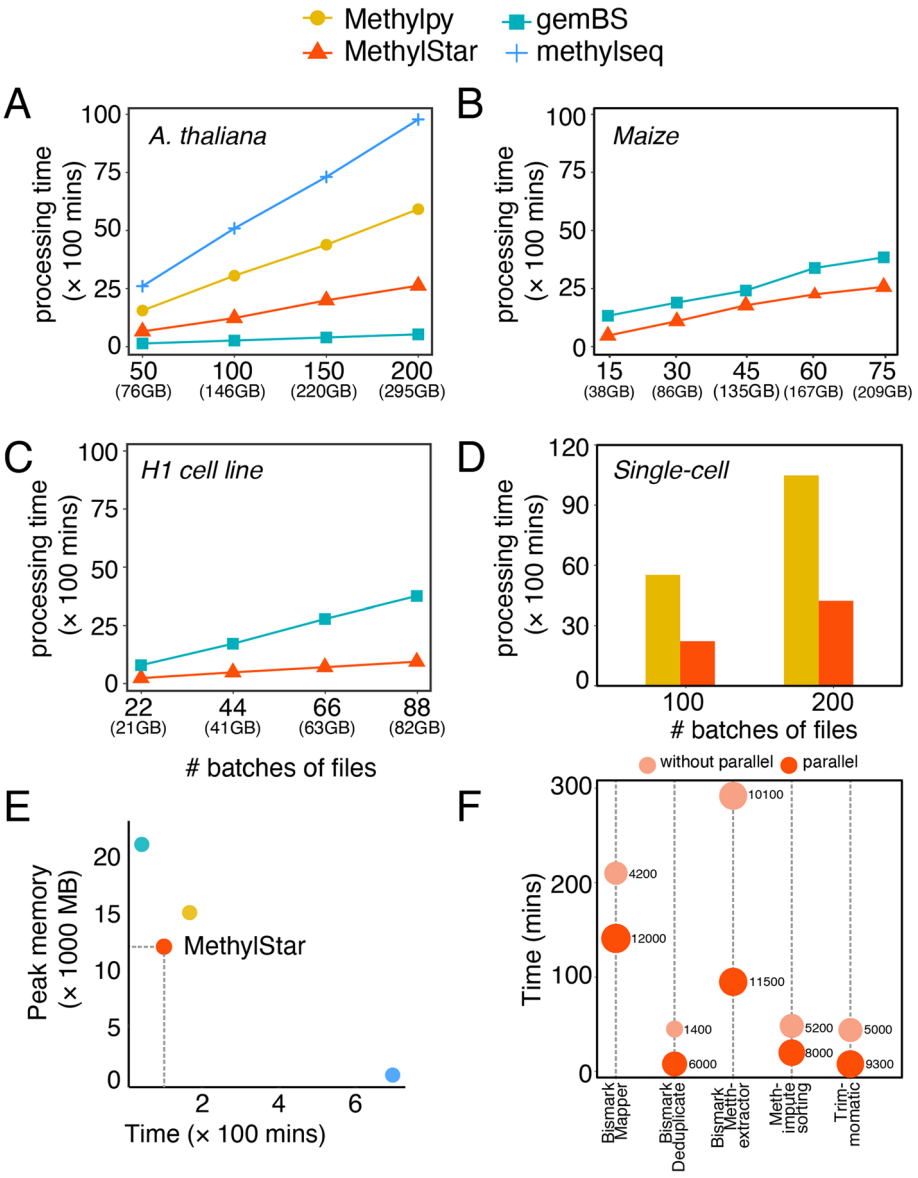

Fig. 2 Performance of MethylStar as compared with other BS-Seq analysis pipelines viz. Methylpy, nf-core/methylseq and gemBS in (**a**) *A. thaliana* (**b**) Maize (**c**) H1 cell line and (**d**) scBS-Seq samples. CPU processing time taken by METHimpute was not included in the current benchmarking process as there is no equivalent method in the other pipelines to compare with. Because of the very long run times observed for the *A. thaliana* data, Methylpy and Methylseq were no longer considered for benchmarking of speed in Maize and H1 cell line samples. All pipelines were run using 32 jobs. (**e**) Peak memory usage as a function of time for 10 random *A. thaliana* samples. (**f**) Time taken by each component of MethylStar. X-axis shows the individual components of MethylStar where the dot with lighter shade of orange indicates -without parallel and darker shade of orange indicates - with parallel implementation of MethylStar. On the y-axis is the time in mins. The size of the dot indicates the peak memory usage in MB by each component

thaliana samples, the percentage of ambiguously mapped reads was considerably higher than that of MethylStar, thus demonstrating a trade-off between speed and mapping performance. We also noticed that the percentage of ambiguously mapped reads by gemBS was even further increased in the case of the Maize samples (Additional file 1: Fig. S1 and Table S1). This could indicate that gemBS's alignment performance is particularly challenged in complex plant genomes, although this hypothesis should be explored in more detail.

Memory usage statistics

Along with benchmarking of speed, we also evaluated the performance of the MethylStar, gemBS, nf-core/methylseq and Methylpy pipelines in terms of system memory utilization using the MemoryProfiler (https://

github.com/pythonprofilers/memory_profiler) python module (Fig. 2e). We assessed the CPU time versus peak/max memory of all the 4 pipelines (default settings) on a computing cluster (specifications above). For 10 random samples from the above *A. thaliana* benchmarking dataset (paired-end, 16 GB, GSE54292) MethylStar and Methylpy showed the best balance between peak memory usage (∼12000 MB and ∼15000 MB, respectively) and total run time (∼177 mins and ∼333 mins, respectively). In contrast, nf-core/methylseq and gemBS exhibited strong trade-offs between memory usage and speed, with nf-core/methylseq showing the lowest peak memory usage (∼700 MB) but the longest CPU time (∼697 mins), and gemBS the highest peak memory usage (∼21000 MB) but the shortest run time (∼42 mins) (Fig. 2e and Additional file 1: Table S5).

Furthermore, we inspected the run times of MethylStar's individual pipeline components, both with and without parallel implementation (Fig. 2f and Additional file 1: Table S3). Our results clearly show that the parallel implementation is considerably faster for all components; however, it is accompanied by a higher peak memory usage. For instance, the implementation of the Bismark alignment step required ∼141 mins (with parallel) as compared to ∼210 mins (without parallel), a ∼33% reduction in processing time. However, in exchange, peak memory usage was increased by ∼65%. Thus, with sufficient computational resources, MethylStar's parallel implementation of Bismark alignment can be very effective in handling large numbers of read alignments in considerably less amount of time (Fig. 2f).

We further benchmarked memory usage using 10 random samples from the above Maize dataset (paired-end, 23 GB, GSE39232). For this analysis, we focused on gemBS and MethylStar due to their shorter processing times for these datasets as compared to nf-core/methylseq and Methylpy. For these Maize dataset, gemBS's peak memory usage was ∼110000 MB as compared to ∼81000 MB for MethylStar (∼1.3 times less memory), (Additional file 1: Table S4) with a total run time of ∼667 mins and ∼508 mins, respectively. We observed a 76% reduction in processing times of Maize samples using the parallel implementation of MethylStar pipeline (Additional file 1: Table S4) as compared to the without parallel implementation. Taken together, these benchmarking results clearly show that MethylStar exhibits favorable performance in terms of processing time and memory, and that it is therefore an efficient solution for the pre-processing of large numbers of samples even on a computing cluster with limited resources.

Conclusion

MethylStar is a fast, stable and flexible pipeline for the high-throughput analysis of bulk or single-cell WGBS data. Its easy installation and user-friendly interface should make it a useful resource for the wider epigenomics community.

Availability and requirements

Project name: MethylStar
Project home page: https://github.com/jlab-code/MethylStar
Operating system(s): Cross-platform
Programming language: Python, Shell, R
License: GPL-3.0

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-06886-3.

Additional file 1: Supplementary figures and data tables (pdf format) showing mapping statistics, processing times and memory usage of different pipelines benchmarked.

Abbreviations

WGBS: Whole-genome bisulfite sequencing; NGS: Next generation sequencing; DMRs: Differentially methylated regions; QC: Quality control; PCR: Polymerase chain reaction; PBAT: Post-bisulfite adaptor tagging; CX-reports: Cytosine context (CG, CHG, CHH) report for all cytosines; BAM: Binary alignment map; RAM: Random-access memory; CPU: Central processing unit; MB: Mega bytes; GB: Giga bytes; I/O: Input/output

Acknowledgements

We thank Markus List for his suggestion to use a docker container for version control.

Authors' contributions

FJ, RRH and YS conceptualized the method. YS and RRH developed, implemented and tested the pipeline. RRH, FJ and YS wrote the paper. FJ supervised the project. All authors have read and approved the manuscript.

Funding

FJ, YS, RRH acknowledge support from the Technical University of Munich-Institute for Advanced Study funded by the German Excellent Initiative and the European Seventh Framework Programme under grant agreement no. 291763. FJ and YS were also supported by the SFB Sonderforschungsbereich924 of the Deutsche Forschungsgemeinschaft(DFG).

Availability of data and materials

Not applicable

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Received: 13 March 2020 Accepted: 6 July 2020
Published online: 13 July 2020

References

1. Luo C, Keown CL, Kurihara L, Zhou J, He Y, Li J, Castanon R, Lucero J, Nery JR, Sandoval JP, Bui B, Sejnowski TJ, Harkins TT, Mukamel EA, Behrens MM, Ecker JR. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science. 2017;357(6351):600–4.

2. Zhu P, Guo H, Ren Y, Hou Y, Dong J, Li R, Lian Y, Fan X, Hu B, Gao Y, Wang X, Wei Y, Liu P, Yan J, Ren X, Yuan P, Yuan Y, Yan Z, Wen L, Yan L, Qiao J, Tang F. Single-cell DNA methylome sequencing of human preimplantation embryos. Nat Genet. 2018;50(1):12–9.
3. Müller F, Scherer M, Assenov Y, Lutsik P, Walter J, Lengauer T, Bock C. Rnbeads 2.0: comprehensive analysis of DNA methylation data. Genome Biol. 2019;20(1):1–2.
4. Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. Genome Biol. 2012;13(6):44.
5. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, Teschendorff AE. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. Bioinformatics. 2017;33(24):3982–4.
6. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. 2012;13(10):1–9.
7. Catoni M, Tsang JM, Greco AP, Zabet NR. DMRcaller: a versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. Nucleic Acids Res. 2018;46(19):114.
8. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H, Lin S, Lin Y, Jung I, Schmitt AD, Selvaraj S, Ren B, Sejnowski TJ, Wang W, Ecker JR. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015;523(7559):212–6.
9. Jühling F, Kretzmer H, Bernhart SH, Otto C, Stadler PF, Hoffmann S. metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. Genome Res. 2016;26(2):256–62.
10. Taudt A, Roquis D, Vidalis A, Wardenaar R, Johannes F, Colomé-Tatché M. METHimpute: imputation-guided construction of complete methylomes from WGBS data. BMC Genomics. 2018;19(1):1–4.
11. Kapourani C-A, Sanguinetti G. Melissa: Bayesian clustering and imputation of single-cell methylomes. Genome Biol. 2019;20(1):1–15.
12. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 2017;18(1):1–13.
13. Peng T, Zhu Q, Yin P, Tan K. SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data. Genome Biol. 2019;20(1):88.
14. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32(2):292–4.
15. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLOS ONE. 2012;7(2):30619. https://doi.org/10.1371/journal.pone.0030619.
16. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
17. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal. 2011;17(1):10–2.
18. Guo W, Fiziev P, Yan W, Cokus S, Sun X, Zhang MQ, Chen P-Y, Pellegrini M. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics. 2013;14(1):774.
19. Huang KYY, Huang Y-J, Chen P-Y. Bs-Seeker3: ultrafast pipeline for bisulfite sequencing. BMC Bioinformatics. 2018;19(1):111.
20. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27(11):1571–2.
21. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009;10(1):1–9.
22. Harris EY, Ounit R, Lonardi S. BRAT-nova: fast and accurate mapping of bisulfite-treated reads. Bioinformatics. 2016;32(17):2696–8.
23. Soe S, Park Y, Chae H. BiSpark: a Spark-based highly scalable aligner for bisulfite sequencing data. BMC Bioinformatics. 2018;19(1):1–9.
24. Chen H, Smith AD, Chen T. WALT: fast and accurate read mapping for bisulfite sequencing. Bioinformatics. 2016;32(22):3507–9.
25. Otto C, Stadler PF, Hoffmann S. Lacking alignments? the next-generation sequencing mapper segemehl revisited. Bioinformatics. 2014;30(13):1837–43.
26. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38(3):276–8.
27. Merkel A, Fernández-Callejo M, Casals E, Marco-Sola S, Schuyler R, Gut IG, Heath SC. gemBS: high throughput processing for DNA methylation data from bisulfite sequencing. Bioinformatics. 2018;35(5):737–42. https://doi.org/10.1093/bioinformatics/bty690, https://doi.org/oup.prod.sis.lan/bioinformatics/article-pdf/35/5/737/27994742/bty690.pdf.
28. Graña O, López-Fernández H, Fdez-Riverola F, González Pisano D, Glez-Peña D. Bicycle: a bioinformatics pipeline to analyze bisulfite sequencing data. Bioinformatics. 2017;34(8):1414–5. https://doi.org/10.1093/bioinformatics/btx778, https://doi.org/oup.prod.sis.lan/bioinformatics/article-pdf/34/8/1414/25119980/btx778.pdf.
29. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nature Biotechnol. 2017;35(4):316–9.
30. Chatterjee A, Stockwell PA, Rodger EJ, Morison IM. Comparison of alignment software for genome-wide bisulphite sequence data. Nucleic Acids Res. 2012;40(10):79. https://doi.org/10.1093/nar/gks150.
31. Omony J, Nussbaumer T, Gutzat R. DNA methylation analysis in plants: review of computational tools and future perspectives. Brief Bioinform. 2020;21(3):906–18.
32. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. Jbrowse: a next-generation genome browser. Genome Res. 2009;19(9):1630–8.

Publisher's Note

B Appendix II: Alphabeta paper reprint + SI

AlphaBeta: computational inference of epimutation rates and spectra from high-throughput DNA methylation data in plants.

Shahryary *et al. Genome Biology* (2020) 21:260
https://doi.org/10.1186/s13059-020-02161-6

Genome Biology

SOFTWARE Open Access

AlphaBeta: computational inference of epimutation rates and spectra from high-throughput DNA methylation data in plants

Yadollah Shahryary[1,2], Aikaterini Symeonidi[1], Rashmi R. Hazarika[1,2], Johanna Denkena[3], Talha Mubeen[1,2], Brigitte Hofmeister[4], Thomas van Gurp[7], Maria Colomé-Tatché[3,5,6], Koen J.F. Verhoeven[7], Gerald Tuskan[8], Robert J. Schmitz[2,9*] and Frank Johannes[1,2*]

*Correspondence:
schmitz@uga.edu;
frank@johanneslab.org
[1]Technical University of Munich, Department of Plant Sciences, Liesel-Beckmann-Str. 2, 85354 Freising, Germany
[2]Technical University of Munich, Institute for Advanced Study, Lichtenbergstr. 2a, 85748 Garching, Germany
Full list of author information is available at the end of the article

Abstract

Stochastic changes in DNA methylation (i.e., spontaneous epimutations) contribute to methylome diversity in plants. Here, we describe *AlphaBeta*, a computational method for estimating the precise rate of such stochastic events using pedigree-based DNA methylation data as input. We demonstrate how *AlphaBeta* can be employed to study transgenerationally heritable epimutations in clonal or sexually derived mutation accumulation lines, as well as somatic epimutations in long-lived perennials. Application of our method to published and new data reveals that spontaneous epimutations accumulate neutrally at the genome-wide scale, originate mainly during somatic development and that they can be used as a molecular clock for age-dating trees.

Keywords: Epimutation, DNA methylation, Plants, Trees, Epigenetics, Epimutation rate, Evolution, Molecular clock, Epigenetic clock, Bioinformatics software tool, R/Bioconductor package

Introduction

Cytosine methylation is an important chromatin modification and a pervasive feature of most plant genomes. It has major roles in the silencing of transposable elements (TEs) and repeat sequences and is also involved in the regulation of some genes [1]. Plants methylate cytosines at symmetrical CG and CHG sites, but also extensively at asymmetrical CHH sites, where H= A, T, C. The molecular pathways that establish and maintain methylation in these three sequence contexts are well-characterized [2] and are broadly conserved across plant species [3–7]. Despite its tight regulation, the methylation status of individual cytosines or of clusters of cytosines is not always faithfully maintained across

cell divisions. As a result, cytosine methylation is sometimes gained or lost in a stochastic fashion, a phenomenon that has been termed "spontaneous epimutation." In both animals and plants, spontaneous epimutations have been shown to accumulate throughout development and aging [8], probably as a byproduct of the mitotic replication of small stem cell pools that generate and maintain somatic tissues.

However, in plants, spontaneous epimutations are not only confined to somatic cells, but occasionally pass through the gametes to subsequent generations [9, 10]. In the model plant *Arabidopsis thaliana* (*A. thaliana*), these transgenerationally heritable (i.e., "germline") epimutations are mainly restricted to CG sites and appear to be absent or not detectable at CHG and CHH sites [11–14]. Initial estimates in *A. thaliana* indicate CG "germline" epimutations are about five orders of magnitude more frequent than genetic mutations ($\sim 10^{-4}$ vs. $\sim 10^{-9}$ per site per haploid genome per generation) [12, 14–16]. Because of these relatively high rates, CG methylation differences accumulate rapidly in the *A. thaliana* genome and generate substantial methylation diversity among individuals in the course of only a few generations [12, 17–19] [20].

A key experimental challenge in studying epimutational processes in a multi-generational setting is to be able to distinguish "germline" epimutations from other types of methylation changes, such as those associated with segregating genetic variation or transient environmental perturbations [21]. Mutation accumulation (MA) lines grown in controlled laboratory conditions are a powerful experimental system to achieve this. MA lines are derived from a single isogenic founder and are independently propagated for a large number of generations. The lines can be advanced either clonally or sexually, i.e., self-fertilization or sibling mating (Fig. 1a). In clonally produced MA lines, the isogenicity of the founder is not required because the genome is "fixed" due to the lack of genetic segregation.

The kinship among the different MA lineages can be presented as a pedigree (Fig. 1a). The structure (or topology) of these pedigrees is typically known, a priori, as the branch-point times and the branch lengths are deliberately chosen as part of the experimental design. In conjunction with multi-generational methylome measurements, MA lines therefore permit "real-time" observations of "germline" epimutations against a nearly invariant genomic background and can facilitate estimates of the per-generation epimutation rates [11]. Sequenced methylomes from a large number of sexually derived MA lines are currently available in *A. thaliana* [12–14, 18, 22, 23] and rice [24], and various other MA lines are currently under construction for epimutation analysis in different genotypes, environmental conditions, and plant species.

Beyond experimentally derived MA lines, natural mutation accumulation systems can also be found in the context of plant development and aging. An instructive example is long-lived perennials, such as trees, whose branching structure can be interpreted as a pedigree (or phylogeny) of somatic lineages that carry information about the epimutational history of each branch [25]. In this case, the branch-point times and the branch lengths can be determined ad hoc using coring data or other types of dating methods (Fig. 1a). By combining this information with contemporary leaf methylome measurements, it is possible to infer the rate of somatic epimutations as a function of age (see also co-submission, [26]).

Attempts to infer the rate of spontaneous epimutations in these diverse plant systems are severely hampered by the lack of available analytical tools. Naive approaches that

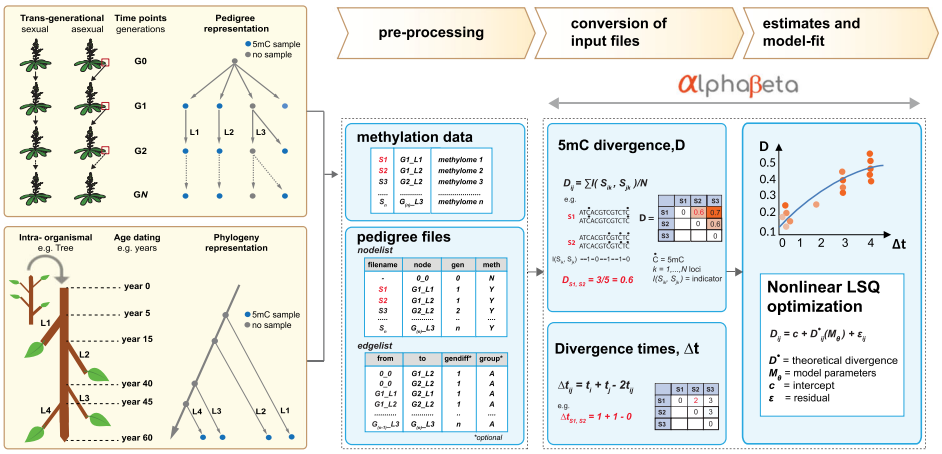

Fig. 1 Overview of the AlphaBeta computational pipeline. **a** Top panel: Construction of multi-generational (G0 to GN) mutation accumulation (MA) lines through sexual (selfing or sibling mating) or asexual (clonal) propagation. The different lineages (L1 to L3) can be represented as a pedigree. The pedigree branch point times and the branch lengths are typically known, a priori, from the experimental design. 5mC sampling can be performed at selected generations, either from plant material of direct progenitors or from siblings of those progenitors. The data can be used to estimate the rate and spectrum of "germline" epimutations. Bottom panel: Long-lived perennials, such as trees, can be viewed as a natural mutation accumulation system. The tree branching structure can be treated as an intra-organismal phylogeny of somatic lineages. 5mC samples can be performed on leaf tissues from selected branches. Along with coring data, the leaf methylomes can be used to estimate the rate and spectrum of somatic epimutations. **b** Data pre-processing: *AlphaBeta* requires methylation data and pedigree data as input. File conversion: Using the input files, *AlphaBeta* calculates the 5mC divergence (D) as well as divergence time (Δt) between all sample pairs. Model estimation: *AlphaBeta* fit competing epimutation models to the data. The model parameters are estimated using numerical non-linear least squares optimization. Model comparisons allow for tests of selection and neutrality

try to count the number of epimutations per some unit of time cannot be used in this setting, because DNA methylation measurements are far too noisy. On the technological side, this noise stems from increased sequencing and alignment errors of bisulphite reads and bisulphite conversion inefficiencies. On the biological side, increased measurement error may result from within-tissue heterogeneity in 5mC patterns [27] and the fact that DNA methylomes are in part transcriptionally responsive to variation in environmental/laboratory conditions [28]. To overcome these challenges, we previously implemented a model-based estimation method, which was originally designed for the analysis of selfing-derived mutation accumulation lines [12]. This approach appropriately accounts for measurement error in the data by describing the time-dependent accumulation of epimutations through an explicit statistical model (Fig. 1b). Fitting this model to pedigree-based 5mC measurements yields estimates of the rate of spontaneous methylation gains and losses and provides a quantitative basis for predicting DNA methylation dynamics over time.

Here, we generalize this method and present *AlphaBeta*, the first software package for inferring the rate and spectrum of "germline" and somatic epimutations in plants. *AlphaBeta* can be widely applied to multi-generational data from sexually or asexually derived MA lines, as well as to intra-generational data from long-lived perennials such as trees. Drawing on novel and published data, we demonstrate the power and versatility of our approach and make recommendations regarding its implementation.

The AlphaBeta method

We start from the assumption that 5mC measurements have been obtained from multiple sampling time-points throughout the pedigree. These measurements can come from whole genome bisulphite sequencing (WGBS) [29] [30], reduced representation bisulphite sequencing (RRBS) [31], or epigenotyping-by-sequencing (epiGBS) [32] technologies, and possibly also from array-based methods. We only require that a "sufficiently large" number of loci has been measured. Moreover, with multigenerational data, we allow measurements to come from plant material of direct progenitors, or else from individual or pooled siblings of those progenitors (Fig. 1a).

Calculating 5mC divergence

For the ith sequenced sample in the pedigree, let s_{ik} be the observed methylation state at the kth locus ($k = 1 \cdots N$). Here, the N loci can be individual cytosines or pre-defined regions (i.e., cluster of cytosines). We assume that s_{ik} takes values 1, 0.5, or 0, according to whether the diploid epigenotype at that locus is m/m, m/u, u/u, respectively, where m is a methylated and u is an unmethylated epiallele. Using this coding, we calculate the mean absolute 5mC divergence, D, between any two samples i and j in the pedigree as follows:

$$D_{ij} = \sum_{k=1}^{N} I(s_{ik}, s_{jk}) N^{-1}, \tag{1}$$

where $I(\cdot)$ is an indicator function, such that

$$I(s_{ik}, s_{jk}) = \begin{cases} 0 \text{ if} s_{ik} = s_{jk} \\ \frac{1}{2} \text{ if} s_{ik} = 0.5 \text{ and} s_{jk} \in \{0, 1\} \\ \frac{1}{2} \text{ if} s_{jk} = 0.5 \text{ and} s_{ik} \in \{0, 1\} \\ 1 \text{ if} s_{ik} = 0 \text{ and} s_{jk} = 1 \\ 1 \text{ if} s_{jk} = 1 \text{ and} s_{ik} = 0. \end{cases}$$

The software automatically calculates D_{ij} and Δt for all unique sample pairs using as input the methylation state calls and the pedigree coordinates of each sample (Fig. 1b).

Modelling 5mC divergence

We model the 5mC divergence as

$$D_{ij} = c + D^{\bullet}_{ij}(M_{\Theta}) + \epsilon_{ij}. \tag{2}$$

Here, $\epsilon_{ij} \sim N(0, \sigma^2)$ is the normally distributed residual error, c is the intercept, and $D^{\bullet}_{ij}(M_{\Theta})$ is the expected divergence between samples i and j as a function of an underlying epimutation model $M(\cdot)$ with parameter vector Θ (see below). We have that

$$\begin{aligned} D^{\bullet}_{ij}(M_{\Theta}) = & \sum_{n \in \nu} \sum_{l \in \nu} \sum_{m \in \nu} I(l, m) \\ & \cdot Pr(s_{ik} = l, s_{jk} = m | s_{ijk} = n, M_{\Theta}) \\ & \cdot Pr(s_{ijk} = n | M_{\Theta}), \end{aligned}$$

where s_{ijk} is the methylation state at the kth locus of the most recent common ancestor of samples i and j, and $\nu = \{0, 0.5, 1\}$. Since samples s_i and s_j are conditionally independent, we can further write:

$$Pr(s_{ik}, s_{jk}|s_{ijk}, M_\Theta) = Pr(s_{ik}|s_{ijk}, M_\Theta) \cdot Pr(s_{jk}|s_{ijk}, M_\Theta).$$

To be able to evaluate these conditional probabilities, it is necessary to posit an explicit form for the epimutational model, M_Θ. To motivate this, we define **G** to be a 3×3 transition matrix, which summarizes the probability of transitioning from epigenotype l to m in the time interval $[t, t+1]$:

$$\mathbf{G} = \begin{array}{c} \begin{matrix} u/u(t+1) & m/u(t+1) & m/m(t+1) \end{matrix} \\ \begin{bmatrix} f_{11}(\alpha,\beta,w) & f_{12}(\alpha,\beta,w) & \cdot \\ f_{21}(\alpha,\beta,w) & \cdot & \cdot \\ \cdot & \cdot & f_{33}(\alpha,\beta,w) \end{bmatrix} \end{array} \begin{matrix} u/u\ (t) \\ m/u\ (t) \\ m/m(t) \end{matrix}$$

The elements of this matrix are a function of gain rate α (i.e., the probability of a stochastic epiallelic switch from an unmethylated to a methylated state within interval $[t, t+1]$), the loss rate β (i.e., the probability of a stochastic epiallelic switch from a methylated to an unmethylated state), and the selection coefficient w ($w \in [0, 1]$). It can be shown that for a diploid system propagated by selfing, **G** has the form

$$\begin{bmatrix} (1-\alpha)^2 & 2(1-\alpha)\alpha & \alpha^2 \\ \frac{1}{4}(\beta+1-\alpha)^2 & \frac{1}{2}(\beta+1-\alpha)(\alpha+1-\beta) & \frac{1}{4}(\alpha+1-\beta)^2 \\ \beta^2 & 2(1-\beta)\beta & (1-\beta)^2 \end{bmatrix} \circ \mathbf{W},$$

and for systems that are propagated clonally or somatically **G** is:

$$\begin{bmatrix} (1-\alpha)^2 & 2(1-\alpha)\alpha & \alpha^2 \\ \beta(1-\alpha) & (1-\alpha)(1-\beta)+\alpha\beta & \alpha(1-\beta) \\ \beta^2 & 2(1-\beta)\beta & (1-\beta)^2 \end{bmatrix} \circ \mathbf{W},$$

where $\circ$ is the Hadamard product and **W** is a matrix of selection coefficients of the form

$$\begin{bmatrix} w & \frac{(w+1)}{2} & 1 \\ w & \frac{(w+1)}{2} & 1 \\ w & \frac{(w+1)}{2} & 1 \end{bmatrix} \text{ or } \begin{bmatrix} 1 & \frac{(w+1)}{2} & w \\ 1 & \frac{(w+1)}{2} & w \\ 1 & \frac{(w+1)}{2} & w \end{bmatrix}$$

depending on whether selection is against epiallele u or m, respectively.

Using this formalism, we can distinguish four different models, which we denote by *ABneutral*, *ABmm*, *ABuu*, and *ABnull*. Model *ABneutral* assumes that the accumulation of spontaneous 5mC gains and losses is selectively neutral ($w = 1$, α and/or $\beta > 0$). In this special case, all epigenotype transitions from time t to $t+1$ are only governed by the rates α and β, and—in the case of selfing—also by the Mendelian segregation of epialleles u and m. The selection models *ABmm* and *ABuu*, by contrast, assume that epimutation accumulation is in part shaped by selection against spontaneous losses or gains of 5mC, respectively ($0 \leq w < 1$, α and/or $\beta > 0$). For example, with selection in favor of epiallele u (model *ABuu*), the fitness of epihomozygote m/m and epiheterozygote m/u are reduced by a factor of w and $(w+1)/2$, respectively. We incorporate this fitness loss directly into the transition matrix by weighing the transition probabilities to these epigenotypes accordingly [33]. Similar arguments hold for the case where selection is for epiallele m. As a reference, we define model *ABnull* as the null model of no accumulation, with $\alpha = 0$, $\beta = 0$, and $w = 1$.

To ensure that the rows of **G** (i.e., the transition probabilities) still sum to unity in the presence of selection, we redefine **G** using the normalization:

$$\mathbf{G}' = \begin{bmatrix} (\sum_i \mathbf{G}_{1i})^{-1} & 0 & 0 \\ 0 & (\sum_i \mathbf{G}_{2i})^{-1} & 0 \\ 0 & 0 & (\sum_i \mathbf{G}_{3i})^{-1} \end{bmatrix} \cdot \mathbf{G}$$

Based on Markov chain theory, the conditional probability $Pr(s_{ik}|s_{ijk}, M_\Theta)$ can then be expressed in terms of $\mathbf{G}'$ as follows:

$$\sum_n Pr(s_{ik} = 0|s_{ijk} = n, M_\Theta) = \sum_{r=1}^{3} (\mathbf{G}'^{t_i - t_{ij}})_{r1}$$

$$\sum_n Pr(s_{ik} = 0.5|s_{ijk} = n, M_\Theta) = \sum_{r=1}^{3} (\mathbf{G}'^{t_i - t_{ij}})_{r2}$$

$$\sum_n Pr(s_{ik} = 1|s_{ijk} = n, M_\Theta) = \sum_{r=1}^{3} (\mathbf{G}'^{t_i - t_{ij}})_{r3}$$

where t_i is the time-point corresponding to sample i and t_{ij} is the time-point of the most recent common ancestor shared between samples i and j, ($t_{ij} \leq t_i, t_j$), and r is a row index. Expressions for $Pr(s_{jk}|s_{ijk}, M_\Theta, t_j)$ can be derived accordingly, by simply replacing t_i by t_j in the above equation. Note that the calculation of these conditional probabilities requires repeated matrix multiplication. However, a direct evaluation of these equations is also possible using the fact that

$$\mathbf{G}'^{t_i - t_{ij}} = \mathbf{p}\mathbf{V}^{t_i - t_{ij}}\mathbf{p}^{-1} \text{ and } \mathbf{G}'^{t_j - t_{ij}} = \mathbf{p}\mathbf{V}^{t_j - t_{ij}}\mathbf{p}^{-1},$$

where **p** is the eigenvector of matrix $\mathbf{G}'$ and **V** is a diagonal matrix of eigenvalues. For selfing and clonal/somatic systems, these eigenvalues and eigenvectors can be obtained analytically.

Finally, to derive $D^{\bullet}_{ij}(M_\Theta)$, we also need to supply $Pr(s_{ijk} = n|M_\Theta)$; that is, the probability that locus k in the most recent common ancestor of samples i and j is in state n ($n \in \{0, 0.5, 1\}$). To do this, consider the methylome of the pedigree founder at time $t = 1$, and let $\pi = [p_1\ p_2\ p_3]$ be a row vector of probabilities corresponding to states u/u, u/m and m/m, respectively. Using Markov Chain theory, we have

$$Pr(s_{ijk} = 0|M_\Theta) = \left[\pi\ \mathbf{G}'^{(t_{ij}-1)}\right]_1$$

$$Pr(s_{ijk} = 0.5|M_\Theta) = \left[\pi\ \mathbf{G}'^{(t_{ij}-1)}\right]_2$$

$$Pr(s_{ijk} = 1|M_\Theta) = \left[\pi\ \mathbf{G}'^{(t_{ij}-1)}\right]_3$$

In many situations, the most recent common ancestor happens to be the pedigree founder itself, so that $t_{ij} = 1$. In the case where the methylome of the pedigree founder has been measured, the probabilities p_1, p_2 and p_3 can be estimated directly from the data using $x_1 N^{-1}$, $x_2 N^{-1}$ and $x_3 N^{-1}$, respectively. Here, x_1, x_2, and x_3 are number of loci that are observed to be in states u/u, u/m, m/m, and N is the total number of loci. Typically, however, x_2 is unknown as most DMP and DMR callers do not output epiheterozygous states (i.e., intermediate methylation calls). Instead, we therefore use

$$p_1 = \frac{x_1}{N}, \qquad p_2 = \gamma\frac{x_3}{N}, \qquad p_3 = (1-\gamma)\frac{x_3}{N}$$

where $\gamma \in [0, 1]$ is an unknown parameter.

Model inference

To obtain estimates for Θ, we seek to minimize the least-squares using

$$\nabla \sum_{q=1}^{M} \left(D_q - D_q^{\bullet}(M_\Theta) - c \right)^2 = \mathbf{0}, \tag{3}$$

where the summation is over all M unique pairs of sequenced samples in the pedigree. Minimization is performed using the "Nelder-Mead" algorithm as part of the optimx package in R. However, from our experience, convergence is not always stable, probably because the function $D_q^{\bullet}(M_\Theta)$ is complex and highly non-linear. We therefore include the following minimization constraint:

$$\nabla \sum_{q=1}^{M} \left(D_q - D_q^{\bullet}(M_\Theta) - c \right)^2 \tag{4}$$

$$+M \left(\tilde{p_1} - p_1(t_\infty, M_\Theta) \right)^2 = \mathbf{0}. \tag{5}$$

Here, $p_1(t_\infty, M_\Theta)$ is the equilibrium proportion of u/u loci in the genome as $t \to \infty$. For a selfing system with $w = 1$, we have that

$$p_1(t_\infty, M_\Theta) = \frac{\beta((1-\beta)^2 - (1-\alpha)^2 - 1)}{(\alpha+\beta)((\alpha+\beta-1)^2 - 2)},$$

and for a clonal/somatic system, it is:

$$p_1(t_\infty, M_\Theta) = \frac{\beta^2}{(\alpha+\beta)^2}.$$

For the case where $0 \leq w < 1$, the equations are more complex and are omitted here. Note that the value $\tilde{p_1}$ is an empirical guess at these equilibrium proportions. For samples whose methylomes can be assumed to be at equilibrium, we have that $p_1(t=1) = p_1(t=2) = \cdots = p1(t_\infty)$, meaning that the proportion of loci in the genome that are in state u/u are (dynamically) stable for any time t. Under this assumption, $\tilde{p_1}$ can be replaced by $\overline{p_1}$, which is the average proportion of u/u loci calculated from all pedigree samples.

Confidence intervals

We obtain confidence intervals for the estimated model parameters by boostrapping the model residuals. The procedure has the following steps: (1) For the qth sample pair q ($q = 1, \cdots, M$), we define a new response variable $B_q = \hat{D_q} + \hat{\epsilon}_k$, where $\hat{D_q}$ is the fitted divergence for the qth pair and $\hat{\epsilon}_k$ is drawn at random and with replacement from the $1 \times M$ vector of fitted model residuals. (2) Refit the model using the new response variable and obtain estimates for the model parameters. (3) Repeat steps 1 to 2 a large number of times to obtain a bootstrap distribution. (4) Use the bootstrap distribution from 3 to obtain empirical confidence intervals.

Testing for selection

To assess whether a selection model provides a significantly better fit to the data compared to a neutral model, we define

$$RSS_F = \sum_{q=1}^{M} \epsilon_q(\hat{\Theta})^2$$

and

$$RSS_R = \sum_{q=1}^{M} \epsilon_q(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{c}|w = 1)^2$$

to be the estimated residual sums of squares of the full model and reduced (i.e., neutral) model, respectively, with corresponding degrees of freedom df_F and df_R. To test for selection, we evaluate the following F-statistic:

$$F = \frac{(RSS_R - RSS_F)}{RSS_F} \cdot \frac{df_F}{df_N},$$

where $df_N = df_F - df_R$. Under the Null $F \sim F(df_N, df_F)$.

Application

To illustrate the utility of our method, we used *AlphaBeta* to study "germline" epimutations in selfing- and asexually derived MA lines of *Arabidopsis* (*A. thaliana*) and dandelion (*Taraxacum officinale*), as well as somatic epimutations in a single poplar tree (*Populus trichocarpa*). Our goal was to demonstrate the wide range of application of our method and to highlight several novel insights into the nature of spontaneous epimutations in plants.

Analysis of spontaneous epimutations in selfing-derived *A. thaliana* MA lines

We first analyzed three *A. thaliana* MA pedigrees (MA1_1, MA1_3, MA3, see Fig. 2a). We chose these MA pedigrees because they differ markedly in their topologies, 5mC sampling strategies, sequencing method, and depth (Fig. 2a, b, Additional file 1: Table S1). All MA pedigrees were derived from a single Col-0 founder accession. The first MA pedigree (MA1_1) was originally published by Becker et al. [13]. The pedigree data consists of 11 independent lineages with sparsely collected WGBS samples ($\sim$ 19.2X coverage) from generations 3, 31, and 32, and a maximum divergence time (Δt) of 64 generations. MA1_3 was previously published by van der Graaf et al. [12]. This data consists of single lineage with dense MethylC-seq measurements ($\sim$ 13.8X coverage) from generations 18 to 30, and a maximum Δt of 13 generations. Finally, we present a new pedigree (MA3), which consists of 2 lineages with dense MethylC-seq measurements ($\sim$ 20.8X coverage) from generations 0 to 11, and a maximum Δt of 22 generations. Unlike MA1_1 and MA1_3, MA3 has 5mC measurements from progenitor plants of each sampled generation, rather than from siblings of those progenitors (Fig. 2a). Further information regarding the samples, sequencing depths, and platforms is provided in Additional file 1: Table S1. A detailed description of data pre-processing and methylation state calling can be found in the "Materials and data pre-processing" section.

Spontaneous epimutations accumulate neutrally over generations

We started by plotting genome-wide (global) 5mC divergence (D) against divergence time (Δt). D increases as a function of Δt in all pedigrees (Fig. 2d). A characteristic pattern is the rapid, non-linear increase in D for the first $\sim$ 8 generations followed by a nearly linear increase. As pointed out before [12], the initial non-linearity is driven by the stable segregation and fixation of epiheterozygote loci that originate from the pedigree founder, a phenomenon that has been well-described in the classical genetic theory of experimental line crosses [34–37]. By contrast, the subsequent linear increase in D is mainly due

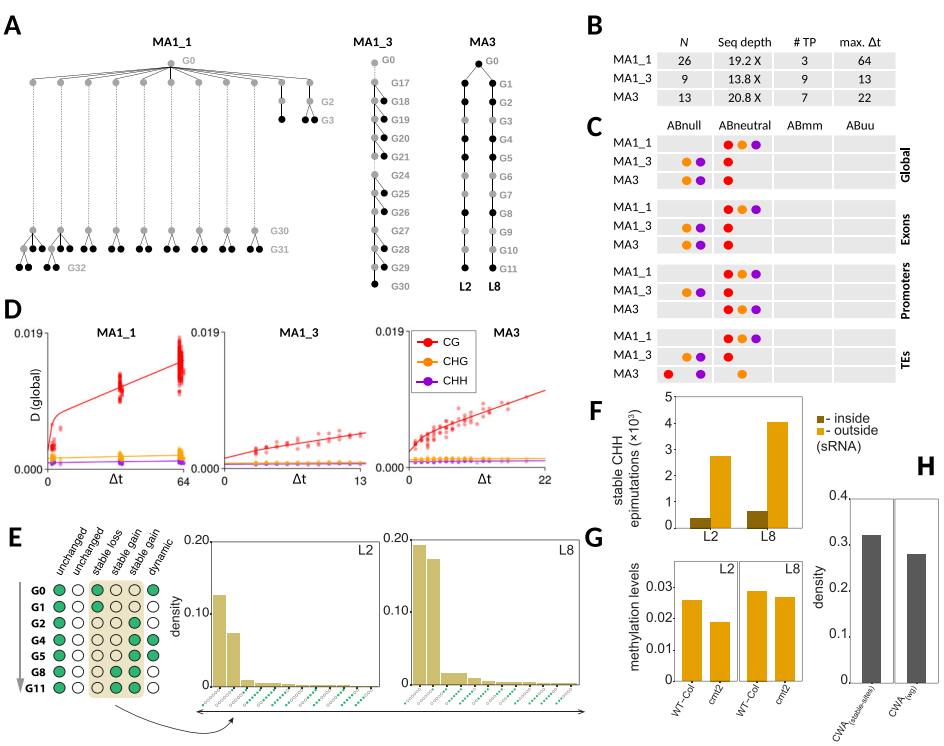

Fig. 2 Analysis of "germline" epimutations in *A. thaliana* mutation accumulation (MA) lines. **a** Three different MA pedigrees were analyzed. All three pedigrees were derived from a single Columbia (Col-0) inbred genotype. Two of the pedigrees were previously published (MA1_1, Becker et al. 2011; MA1_3, van der Graaf et al. 2015), and one pedigree (MA3) is new. These three MA pedigrees were chosen because they differ in their topologies, 5mC measurement strategies, and the temporal resolution of the 5mC samples. **b** Overview of the data: *N* is the total number of sequenced samples; Seq depth is the average sequence depths of the samples; # TP is the number of unique time-points (or generations) that are sampled; max. (Δt) is the maximum divergence time (in generations) in the pedigree. **c** Application of models Abnull, Abneutral, ABmm, and Abuu. The best fitting model is indicated for each MA pedigree, sequence context (CG, CHG, and CHH), and genomic feature (global, exons, promoters, TEs). **d** Shown are the fits of the best fitting models for each pedigree and context. **e** Schematic representation of transgenerationally stable CHH epimutations. The barplots indicate the density of stable CHH epimutations in lineages L2 and L8 of the MA3 pedigree. **f** CHH sites featuring stable epimutations tend to fall outside of sRNA clusters in lineages L2 and L8. **g** Analysis of *cmt2* mutant and Col-0 wt from Stroud et al. [2] show loss of methylation in the mutant at the stable CHH epimutation sites, indicating that these loci are targeted by CMT2. **e** Compared to the whole genome (wg), stable CHH loci with stable epimutations are enriched for CWA trinucleotides, which is a preferred substrate for CMT2 binding

to the accumulation of new epimutations that arise de novo during inbreeding. The co-occurrence of these two processes is restricted to mutation accumulation systems that are propagated sexually. In clonally or asexually derived MA lines, the non-linear increase in *D* should be absent, as can indeed be seen in our later analysis of poplar and dandelion (see below).

Another striking insight from the 5mC divergence patterns is that the increase in *D* is particularly pronounced for context CG but appears to be low, or even absent, at CHG and CHH loci. Similar observations have previously led to the hypothesis that the inheritance of spontaneous epimutations may be restricted to CG dinucleotides [11, 12], perhaps as a consequence of the preferential reinforcement of CHG and CHH methylation during sexual reproduction [38, 39]. Using heuristic arguments, it had been further suggested that CG epimutations accumulate neutrally, at least at the level of individual cytosines,

meaning that 5mC gains and loss in this context are under no selective constraints [12]. However, these hypotheses have never been tested explicitly due to a lack of analytical tools.

To address this, we fitted models *ABneutral*, *ABmm*, *ABuu*, and *ABnull* to the divergence data of each pedigree (Fig. 2c). As mentioned above (see the "The AlphaBeta method" section), model *ABneutral* assumes that spontaneous 5mC gains and losses accumulate neutrally across generations, *ABmm* assumes that the accumulation is partly shaped by selection against spontaneous losses of 5mC, *ABuu* assumes that the accumulation is partly shaped selection against spontaneous gains, and *ABnull* is the null model of no accumulation. Formal model comparisons revealed that *ABneutral* provides the best fit to the 5mC divergence data in context CG in all pedigrees (Fig. 2c, Additional files 2, 3, and 4: Tables S2-S4). This was true at the genome-wide scale (global) as well as at the sub-genomic scale (exons, promoters, TEs). Globally, *ABneutral* explained between 77 and 90% of the total variance in D, indicating that a neutral epimutation model provides a good and sufficient description of the molecular process that generates heritable 5mC changes at level of individual cytosines over time. Interestingly, we also detected, for the first time, highly significant accumulation of neutral epimutations in contexts CHG and CHH (Fig. 2c, Additional files 2, 3, and 4: Tables S2-S4). However, the detection of these accumulation patterns was mainly restricted to MA1_1, the largest of the three pedigrees in terms of both sample size (N=26) and divergence times (max. Δt=64), and to some extent also to MA3, the second largest of the three pedigrees ($N = 13$, max. Δt=22).

The detected accumulation of CHH epimutations was somewhat surprising, given that cytosine methylation in this context is typically targeted by the RNA-directed DNA methylation pathway (RdDM). The de novo action of this pathway should prevent the formation of trans-generationally stable epimutations, particularly those originating from DNA methylation loss [40]. To explore this observation in more detail, we inspected specific CHH sites that showed stable methylation status changes over generation time (Fig. 2e). Our analysis revealed that these CHH sites actually fall outside of known sRNA clusters and are therefore unlikely involved in RdDM (Fig. 2f). Instead, they appear to be targeted by CHROMOMETHYLASE 2 (CMT2), an enzyme that maintains methylation at a subset of CHG and CHH sites, independently of RdDM. Support for this hypothesis comes from the fact that these CHH sites are enriched for trinucleotide context CWA (W = A, T) (Fig. 2g), which is a preferred substrate for CMT2 binding [41]. Moreover, a re-analysis of a *cmt2* methylation mutant from Stroud et al. [2] revealed a marked reduction in cytosine methylation at these CHH sites relative to wt (Fig. 2h), providing additional evidence for a maintenance role of CMT2 at these loci.

Taken together, these results provide a possible molecular explanation for the accumulation of CHH epimutations over generation time, at least for specific CHH subcontexts. However, the ability to consistently detect these accumulation patterns from multi-generational pedigree data should be explored more systematically in future studies, particularly as a function of sample size, divergence time, and measurement uncertainly in 5mC divergence.

The rate and spectrum of spontaneous CG, CHG, and CHH epimutations

We examined the estimated epimutation rates corresponding to the best fitting models from above (Fig. 3a, Additional files 2, 3, and 4: Tables S2-S4). Globally, we found that

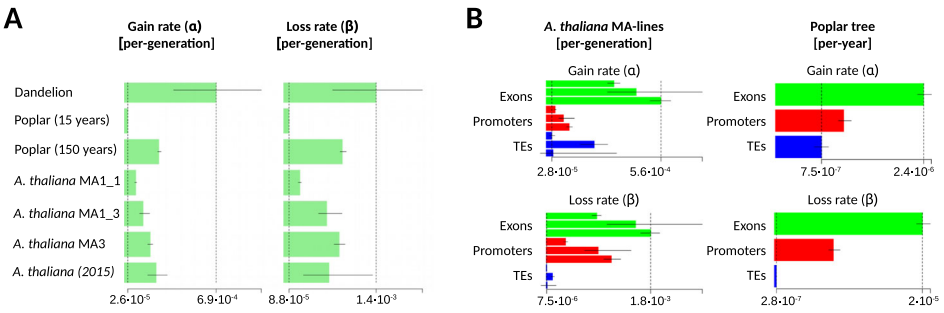

Fig. 3 Comparisons of the CG epimutation rates and spectra. **a** Shown are the estimates (± 95% confidence intervals) of the genome-wide (global) CG epimutation rates for the different pedigrees. For comparison, we also show the range of previous estimates from *A. thaliana* MA lines (*A. thaliana* (2015)); see van der Graaf et al. (2015). In poplar, the estimated per-year epimutation rates were converted to per-generation rates by assuming a generation time of 15 years and 150 years. The gain and loss rates are all well within one order of magnitude of each other, and differences are mostly within estimation error. The dashed vertical lines mark off the lower and upper range of the point estimates. **b** Side-by-side comparison of "germlines" and somatic epimutation rate estimates (± 95% confidence intervals) in *A. thaliana* MA lines and poplar, respectively, for selected genomic features. The rank ordering of the magnitude of these rates is similar. For *A. thaliana*, the order of presentation of the pedigrees is MA3, MA1_3, and MA1_1 (from bottom to top within each feature). Feature-specific rates could not be obtained in dandelion since no annotated assembly is currently available

the CG methylation gain rate (α) is $1.4 \cdot 10^{-4}$ per CG per haploid genome per generation on average (range $8.6 \cdot 10^{-5}$ to $1.94 \cdot 10^{-4}$) and the loss rate (β) is $5.7 \cdot 10^{-4}$ on average (range $2.5 \cdot 10^{-4}$ to $8.3 \cdot 10^{-4}$). Using data from pedigree MA1_1, we also obtained the first epimutation rate estimates for contexts CHG and CHH. The gain and loss rates for CHG were $3.5 \cdot 10^{-6}$ and $5.8 \cdot 10^{-5}$ per CHG per haploid genome per generation, respectively; and for CHH, they were $1.9 \cdot 10^{-6}$ and $1.6 \cdot 10^{-4}$ per CHH per haploid genome per generation. Hence, transgenerationally heritable CHG and CHH epimutations arise at rates that are about 1 to 2 orders of magnitude lower than CG epimutations in *A. thaliana*, which is reflected in the relatively slow increase of 5mC divergence in non-CG contexts over generation time (Fig. 2d).

In addition to global estimates, we also assessed the gain and loss rates for selected genomic features (exons, promoters, TEs). In line with previous analyses [12], we found striking and consistent rate differences, with exon-specific epimutation rates being 2 to 3 orders of magnitude higher than TE-specific rates (Fig. 3b, Additional files 2, 3, and 4: Tables S2-S4). Interestingly, this trend was not only restricted to CG sites, but was also present in contexts CHG and CHH. This later finding points to yet unknown sequence or chromatin determinants that affect the 5mC fidelity of specific regions across cell divisions, independently of CG, CHG, and CHH methylation pathways.

We note that the CG epimutation rates reported here differ slightly from our previous estimates [12] (Fig. 3a, Additional files 3 and 4: Tables S3-S4). This small discrepancy is mainly the result of differences in the data pre-processing. Application of *AlphaBeta* to published pre-processed samples yielded similar results to those reported previously (data not shown), indicating that the statistical inference itself is consistent. Unlike past approaches, we here utilized the recent *MethylStar* pipeline [42] for data pre-processing and methylation state calling. The use of this pipeline leads to a substantial increase in the number of high-confidence cytosine methylation calls for downstream epimutation analysis (Additional file 5: Table S5). This boost in sample size is reflected in the lower

variation in α and β estimates across MA pedigree compared with previous reports [12] (Fig. 3a, Additional files 2 and 3: Tables S2-S3).

Analysis of spontaneous somatic epimutations in poplar

Despite the above quantitative insights into the rate and spectrum of spontaneous epimutation in *A. thaliana*, it remains unclear how and where these epimutations actually originate in the plant life cycle. One hypothesis is that they are the result of imperfect 5mC maintenance during the mitotic replication of meristematic cells which give rise to all above and below ground tissues, including the "germline" (Additional file 6: Figure S1). As the germline is believed to be derived quite late in development from somatic precursors, somatic epimutations that accumulate during aging can subsequently be passed to offspring. An alternative hypothesis is that heritable epimutations originate as a byproduct of sRNA-mediated reinforcement errors in the sexual cell linages. One way to distinguish these two possibilities is to study epimutational processes in systems that bypass or exclude sexual reproduction.

Long-lived perennials, such as trees, represent a powerful system to explore this. A tree's branching structure can be interpreted as an intra-organismal phylogeny of different somatic cell lineages. It is therefore possible to track mutations and epimutations and their patterns of inheritances across different tree sectors. Recently, there has been a surge of interest in characterizing somatic nucleotide mutations in trees using whole genome sequencing data [43–46]. These studies have shown that fixed mutations arise sequentially in different tree sectors, thus pointing at a shared meristematic origin.

To facilitate the first insights into epimutational processes in long-lived perennials, we applied *AlphaBeta* to MethylC-seq leaf samples ($\sim$ 41.1X coverage) from 8 separate branches of a single poplar (*Populus trichocarpa*) tree (see also co-submission, [26]). The tree features two main stems (here referred to as tree 13 and tree 14), which were originally thought to be two separate trees (Fig. 4a, b). However, both stems are stump sprouts off an older tree that was knocked down about 350 years ago. In other words, tree 13 and tree 14 are clones that have independently diverged for a long time. Four branches from each tree were chosen and aged by coring at the points where each branch meets the main stem as well as at the terminal branch (Fig. 4a, b, see the "Materials and data pre-processing" section). Age dating of the bottom sector of the tree proved particularly challenging because of heart rot, rendering estimates of the total tree age imprecise. However, an estimate based on diameter measurements places the minimum age of the tree at about 250 years.

Inferring total tree age from leaf methylome data

We used the coring-based age measurements from each of the branches along with the branch points to calculate divergence times (Δt) between all pairs of leaf samples (Fig. 4c). We did this by tracing back their ages (in years) along the branches to their most recent common branch point (i.e., "founder cells") (Additional file 6: Figure S1). The calculation of the divergence times for pairs of leaf samples originating from tree 13 and tree 14 was not possible since the total age of the tree was unknown. To solve this problem, we included the total age of the tree as an additional unknown parameter into our epimutation models. Our model estimates revealed that the total age of the tree is approximately 330 years (Fig. 4e), an estimate that fits remarkably well with the hypothesized age window

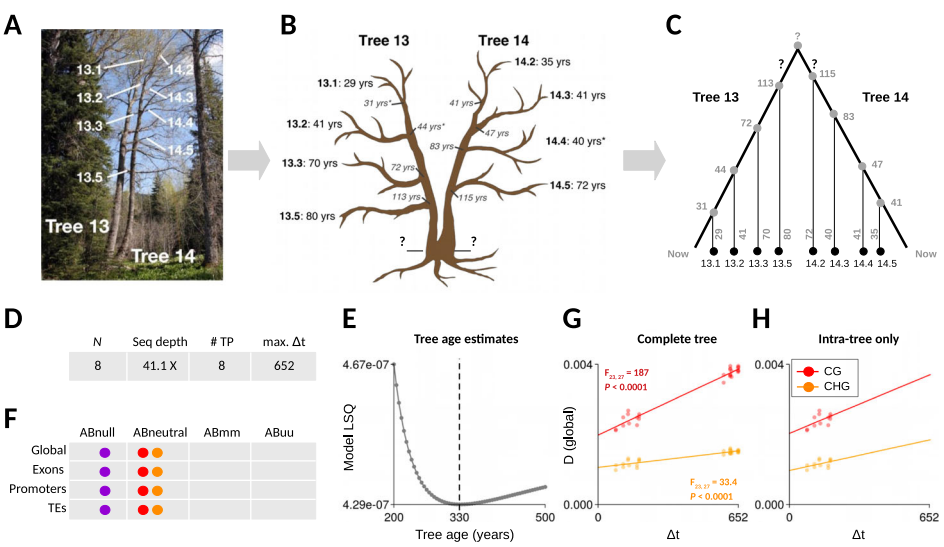

Fig. 4 Analysis of somatic epimutations in poplar. **a** A single poplar (*P. trichocarpa*) tree was analyzed. Tree 13 and 14 are two main stems that have diverged early in development. Four branches from each tree were chosen and aged by coring. **b** Shown are the coring sites along with the coring-based branch ages. Age coring proved technically challenging at the bottom of the tree and led to unintelligible ring counts. An educated guess places the age of the tree between 250 and 350 years. **c** The tree can be presented as an intra-organismal phylogeny. Leaf methylomes were collected from each of the selected branches and served as input for *AlphaBeta*. **d** Overview of the data: *N* is the total number of sequenced samples; Seq depth is the average sequence depths of the samples; # TP is the number of unique time-points that are sampled; max. (Δt) is the maximum divergence time (in years) between leaf samples. **e** *AlphaBeta* was fitted to the global CG methylation divergence data of the complete tree data treating tree age as an unknown parameter. Model residual (LSQ) was minimized at an age of 330 years, which is our estimate of the age of the tree. **f** Model comparisons indicate that somatic epimutations accumulate neutrally in context CG (red) and CHG (orange) during aging, both at the global scale as well as within specific genomic features (exons, promoters, TEs). **g**, **h** Shown are the fits of model *ABneutral* to the global CG (red) and CHG (orange) methylation divergence data of the complete tree (intra-tree + inter-tree, **g**), as well as for tree 13 and tree 14 separately (intra-tree, **h**)

(between 250 and 350 years). Furthermore, the model fits provided overwhelming evidence that somatic epimutations, in poplar, accumulate in a selectively neutral fashion during aging, both at the genome-wide scale (globally) as well as at the sub-genomic scale (exons, promoters, TEs) (Fig. 4f, see also co-submission [26]). This was true for CG and CHG contexts (Fig. 4g). The fact that the accumulation of CHG epimutations is so clearly detectable in poplar, but only inconsistently in *A. thaliana* MA lines, could indicate that somatically acquired CHG methylation changes experience some level of reprogramming during sexual reproduction. But this hypothesis should be tested more directly using cell-type-specific sequencing approaches. To rule out that the somatic accumulation patterns in poplar are not dominated by our estimate of tree age, we also examined the accumulation patterns within tree 13 and tree 14 separately. We found similar accumulation slopes as well as epimutation rates (Fig. 4h, see also co-submission [26]).

Epimutation spectra have a somatic origin

We examined the somatic epimutation rate estimates from the complete tree analysis. At the genome-wide scale, we found that the 5mC gain and loss rates in context CG are $1.7 \cdot 10^{-6}$ and $5.8 \cdot 10^{-6}$ per site per haploid genome per year, respectively, and $3.3 \cdot 10^{-7}$ and $4.1 \cdot 10^{-6}$ in context CHG. Interestingly, these *per-year* CG epimutation rates are only about two orders of magnitude lower than the *per-generation* rates in *A. thaliana*

MA lines. Assuming an average generation time of about 15 to 150 years in poplar [47], its expected per-generation CG epimutation rate would be between $\sim 10^{-5}$ and $\sim 10^{-4}$, which is within the same order of magnitude to that of *A. thaliana* ($\sim 10^{-4}$) (Fig. 3a). This close similarity is remarkable given that poplar is about ~ 100 times larger and its life cycle ~ 1000 times longer than that of *A. thaliana*. Similar insights were reached in a recent comparison of the per-generation nucleotide mutation rates between Oak (*Quercus rubur*) and *A. thaliana* [45], which were also found to be remarkably close to each other. Taken together, these findings support the emerging hypothesis that meristematic cells of long-lived perennials undergo fewer cell divisions per unit time than annuals, so that the cumulative life-time number of cell divisions is similar [46]. This hypothesis should be tested more directly using cell count assays.

To assess whether the accumulation dynamics of somatic epimutations in poplar differs between genomic features, we examined in more detail the estimated rates and spectra for exons, promoters, and TEs (Fig. 3b). Focusing on context CG, we found considerable rate differences. The gain rates for exons, promoters, and TEs were $2.4 \cdot 10^{-6}$, $1.1 \cdot 10^{-6}$, and $7.5 \cdot 10^{-7}$ per site per haploid genome per year, respectively, and the loss rates were $2 \cdot 10^{-5}$, $8 \cdot 10^{-6}$, and $2.8 \cdot 10^{-7}$. Intriguingly, the rank order of these rates was similar to what we had observed for germline epimutations in *A. thaliana*, with exons showing the highest combined rates, followed by promoters and then TEs (Fig. 3b). These findings indicate that the epimutation spectrum is deeply conserved across angiosperms and that it is mainly shaped during somatic development, rather than being a byproduct of selective reinforcement of DNA methylation in the germline or early zygote. Identifying *cis*- and *trans*-determinants that affect local epimutation rates seems to be an important next challenge [11].

Analysis spontaneous epimutations in asexually derived dandelion MA lines

Our analysis of *A. thaliana* and poplar revealed strong similarities in epimutation rates and spectra. To facilitate further inter-specific comparisons, particulary across different mating systems, we generated novel MA lines in an asexual dandelion (*Taraxacum officinale*) genotype (AS34) [48] (Fig. 5a). Apomictic dandelions are triploid and produce asexually via clonal seeds in a process that involves unreduced egg cell formation (diplospory), parthenogenic embryo development, and autonomous endosperm formation, resulting in genetically identical offspring [49]. Using single-seed descent from a single apomictic triploid founder genotype, 8 replicated lineages were propagated for 6 generations, and 5mC measurements were obtained from each generation (Fig. 5a).

The total dataset was relatively large, with 48 sequenced samples and a maximum divergence time of 14 generations (Fig. 5b). 5mC measurements were obtained using epigenotyping-by-sequencing (epiGBS) [32] (see the "Materials and data pre-processing" section). Since there is currently no published dandelion reference assembly, local assemblies were generated de novo from the epiGBS short reads and served as basis for cytosine methylation calling [32]. With this approach, ~ 24000 measured cytosines were shared between any two sample pairs on average and were used to calculate pair-wise CG methylation divergence D.

Plotting D against divergence time (Δt) revealed considerable measurement variation across samples (Fig. 5c). This large variation could have several possible sources: First, methylation state calling was based on local assemblies rather than on reference-based

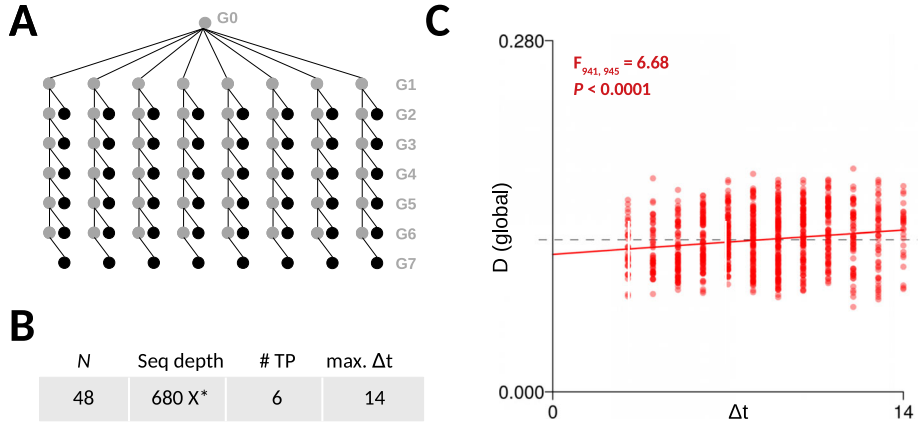

Fig. 5 Analysis of CG epimutations in apomictic dandelion. **a** Using single-seed descent from a single apomictic triploid founder genotype, 8 replicated lineages were propagated for 6 generations. DNA methylation measurements were obtained using epigenotyping by sequencing (epiGBS). **b** Overview of the data: *N* is the total number of sequenced samples; Seq depth is the average sequence depths of the samples; # TP is the number of unique time-points that are sampled; max. Δt is the maximum divergence time (in generations) between samples. *Note: the calculation of average read coverage was based only on interrogated cytosines as epiGBS does not yield any genome-wide data. **c** Model fits to the CG divergence data. Highly significant increases in 5mC divergence (*D*) over generation time (Δ*t*) were detected in all sequence contexts, despite the relatively large variation in 5mC divergence patterns (see text)

alignments. Second, epiheterozygotes in this triploid genotype could not be effectively distinguished on the basis of the observed methylation levels, which introduce uncertainties in the calculation *D*. Third, early implementations of the epiGBS protocol could not distinguish PCR duplicates, a problem that has since been solved [50].

Despite these limitations, application of *AlphaBeta* to the CG divergence data revealed strong statistical evidence for epimutation accumulation over time ($F_{941,945}$=6.68, $p < 0.0001$). Consistent with *A. thaliana* and poplar, a neutral epimutation model (*ABneutral*) provided the best fit to the data. Based on these model fits, we estimate the global CG gain rate and loss rate at $6.9 \cdot 10^{-4}$ and $1.4 \cdot 10^{-3}$ per CG site per haploid genome per generation, respectively (Fig. 3). We note that these "per-haploid" rate estimates are slightly biased upward, since we applied *AlphaBeta*'s diploid models to data from a triploid species, but this model mis-specification should have little impact in the analysis of asexually reproducing systems in which genetic segregation is absent. Keeping this caveat in mind, our results show that the dandelion per-generation CG epimutation rates are close to those obtained in *A. thaliana* and poplar (Fig. 3a), and at least within the same order of magnitude. This finding reinforces the notion that epimutational process are largely conserved across angiosperms, which is probably a direct consequence of the fact that the DNA methylation maintenance machinery is itself highly conserved [5, 51]. Moreover, our findings in dandelion lend further support to the hypothesis that sexual reproduction has no major impact on the formation and inheritance of spontaneous epimutations. Future studies should test this hypothesis more directly by studying the epimutation landscape of a fixed genotype that has been propagated in parallel both sexually and asexually.

Discussion

Accurate estimates of the rate and spectrum of spontaneous epimutations are essential for understanding how DNA methylation diversity arises in the context of plant evolution, development, and aging. Here, we presented *AlphaBeta*, a computational method for

obtaining such estimates from pedigree-based high-throughput DNA methylation data. Our method requires that the topology of the pedigree is known. This requirement is typically met in the experimental construction of mutation accumulation lines (MA lines) that are derived through sexual or clonal reproduction. However, we demonstrated that *AlphaBeta* can also be used to study somatic epimutations in long-lived perennials, such as trees, using leaf methylomes and coring data as input. In this case, our method treats the tree branching structure as an intra-organismal phylogeny of somatic lineages and uses information about the epimutational history of each branch.

To demonstrate the versatility of our method, we applied *AlphaBeta* to very diverse plant systems, including multi-generational DNA methylation data from selfing- and asexually derived MA lines of *A. thaliana* and dandelion, as well as intra-generational DNA methylation data of a poplar tree. Our analysis led to several novel insights about epimutational processes in plants. One of the most striking findings was the close similarity in the epimutation landscapes between these very different systems. Close similarities were observed in the per-generation CG epimutation rates between *A. thaliana*, dandelion, and poplar both at the genome-wide as well as at the subgenomic scale. Any detected rate differences between these different systems were all within one order of a magnitude of each other, and as such practically indistinguishable from experimental sources of variation. As a reference, variation in epimutation rate estimates across different *A. thaliana* mutation accumulation experiments vary up to 75% of an order of a magnitude. Clearly, larger sample sizes are needed along with controlled experimentally comparisons to be able to identify potential biological causes underlying subtle epimutation rate differences between species, mating systems, genotypes, or environmental treatments. Furthermore, the close similarity between sexual and asexual (or somatic) systems reported here provide indirect evidence that transgenerationally heritable epimutations originate mainly during mitotic rather than during meiotic cell divisions in plants.

Our application of *AlphaBeta* to poplar also provided the first proof-of-principle demonstration that leaf methylome data, in combination with our statistical models, can be employed as a molecular clock to age-date trees or sectors of trees. Analytically, this is similar to inferring the branch lengths of the underlying pedigree (or phylogeny). With sufficiently large sample sizes, it should be possible to achieve this with relatively high accuracy and extend this inference to the entire tree structure. The comparatively high rates of somatic and germline epimutations are instrumental in this as they provide increased temporal resolution over classical DNA sequence approaches, which rely on rare de novo nucleotide mutations. Our methodological approach should be applicable, more generally, to any perennial or long-lived species. We are currently extending the *AlphaBeta* tool set to facilitate such analyses.

Analytically, *AlphaBeta* is not restricted to the analysis of plant data. The method could also be used to study epimutational processes in tumor clones based on animal single-cell WGBS data. Such datasets are rapidly emerging [52]. In this context, *AlphaBeta* could be instrumental in the inference of clonal phylogenies and help calibrate them temporally. Such efforts may complement current pseudotemporal ordering (or trajectory inference) methods and lineage tracing strategies in single-cell methylation data [53, 54].

The implementation of *AlphaBeta* is relatively straight-forward. The starting point of the method are methylation state calls for each cytosine. These can be obtained from any methylation calling pipeline. In the data applications presented here, we used *AlphaBeta*

in conjunction with *MethylStar* [42], which is an efficient pre-processing pipeline for the analysis of WGBS data and features a HMM-based methylation state caller [55]. Application of this pipeline leads to up a substantial increase in the number of high-confidence cytosine methylation calls for epimutation rate inference compared with more conventional methods. We therefore recommend using *AlphaBeta* in conjunction with *MethylStar*. Software implementing *AlphaBeta* is available as a Bioconductor R package at https://bioconductor.org/packages/release/bioc/html/AlphaBeta.html.

Materials and data pre-processing

A. thaliana MA lines data

Plant material

For MA3, seeds were planted and grown in 16-h day lengths and samples were harvested from young above ground tissue. Tissue was flash frozen in liquid nitrogen and DNA was isolated using a Qiagen Plant DNeasy kit (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. For MA1_1 and MA1_3, a detailed description of growth conditions and plant material can be found in the original publications [12, 13].

Sequencing and data processing

For MA3, MethylC-seq libraries were prepared according to the protocol described in Urich et al. [56]. Libraries were sequenced to 150 bp per read at the Georgia Genomics & Bioinformatics Core (GGBC) on a NextSeq500 platform (Illumina). Average sequencing depth was 20.8X among samples (Additional file 1: Table S1). For MA1_1 and MA1_3, FASTQ files (*.fastq) were downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64463. All data processing and methylation state calling was performed using the *MethylStar* pipeline [42]. Summary statistic for each sample can be found in Additional file 1: Table S1. All sequences have been submitted to the GEO repository with the following GEO accession number GSE153055.

Poplar data

Tree coring

The tree used in this study was located at Hood River Ranger District [Horse Thief Meadows area], Mt. Hood National Forest, 0.6 mi south of Nottingham Campground off OR-35 at unmarked parking area, 500′ west of East Fork Trail nbr. 650 across river, ca. 45.355313, -121.574284. Tree cores were originally collected from the main stem and five branches in April 2015 at breast height (∼ 1.5 m) for standing tree age using a stainless-steel increment borer (5 mm in diameter and up to 28 cm in length). Cores were mounted on grooved wood trim, dried at room temperature, sanded, and stained with 1% phloroglucinol following the manufacturer's instructions (https://www.forestry-suppliers.com/Documents/1568_msds.pdf).

Annual growth rings were counted to estimate age. For cores for which accurate estimates could not be made from the 2015 collection, additional collections were made in spring 2016. However, due to difficulty in collecting by climbing, many of the cores did not reach the center of the stem or branches (pith) and/or the samples displayed heart rot. Combined with the difficulty in demarcating rings in porous woods such as poplar *Populus*, accurate measures of tree age or branch age were challenging.

Sequencing and data processing

A single MethylC-seq library was created for each branch from leaf tissue. Libraries were prepared according to the protocol described in Urich et al. [56]. Libraries were sequenced to 150 bp per read at the Georgia Genomics & Bioinformatics Core (GGBC) on a NextSeq500 platform (Illumina). Average sequencing depth was 41.1x among samples. MethylC-seq reads were aligned using Methylpy v1.3.2 [57]. Alignment was to the new Stettler14 assembly of *P. trichocarpa*, as described in [26]. Starting from the BAM files (*.bam), the *MethylStar* pipeline [42] was used for further data processing and methylation state calling. All sequences have been deposited in SRA (see [26]).

Dandelion MA lines data

Plant material

Starting from a single founder individual, eight replicate lineages of the apomictic common dandelion (*Taraxacum officinale*) genotype AS34 [48] were grown for six generations via single-seed descent under common greenhouse conditions. Apomictic dandelions are triploid and produce asexually via clonal seeds in a process that involves unreduced egg cell formation (diplospory), parthenogenic embryo development, and autonomous endosperm formation, resulting in genetically identical offspring [49]. Seeds were collected from each of the 48 plants in the six-generation experiment and stored under controlled conditions (15 °C and 30% RH). After the 6th generation, from each plant in the pedigree, a single offspring individual was grown in a fully randomized experiment under common greenhouse conditions. Leaf tissue from a standardized leaf was collected after 5 weeks, flash frozen in liquid nitrogen, and stored at − 80 °C until processing.

Sequencing and data processing

DNA was isolated using the Macherey-Nagel Nucleospin Plant II kit (cell lysis buffer PL1). DNA was digested with the PstI restriction enzyme and epiGBS sequencing libraries were prepared as described elsewhere [32]. Based on genotyping-by-sequencing [58], epiGBS is a multiplex reduced representation bisulphite sequencing (RRBS) approach with an analysis pipeline that allows for local reference construction from bisulphite reads, which makes the method applicable to species for which a reference genome is lacking [32]. PstI is a commonly used restriction enzyme for genotyping-by-sequencing; however, its activity is sensitive to CHG methylation in CTGCAG recognition sequence. This makes the enzyme better at unbiased quantification of CG methylation than of CHG methylation [32]. After quantification of the sequencing libraries using a multiplexed Illumina MiSeq Nano run, samples were re-pooled to achieve equal representation in subsequent epiGBS library sequencing. The experimental samples were sequenced on two Illumina HiSeq 2500 lanes (125 cycles paired-end) as part of a larger epiGBS experiment which consisted of a total of 178 samples that were randomized over the two lanes. Because of inadequate germination or due to low sequencing output (library failure), four of the 48 samples were not included in the downstream analysis. All sequences have been deposited in SRA under Bioproject: PRJNA608438. The biosamples include SAMN14266774 to 778, SAMN14266797 to 802, SAMN14266821 to 826, SAMN14266845 to 850, SAMN14266869 to 872, SAMN14266874, SAMN14266893 to

894, SAMN14266896 to 897, SAMN14266916 to 921, and SAMN14266940 to 945. These 44 samples have been submitted as part of a bigger experiment of 178 samples total.

DNA methylation analysis

Sequencing reads were demultiplexed (based on custom barcodes) and mapped against a dandelion pseudo-reference sequence that was generated de novo from PstI-based epiGBS [32]. This pseudo-reference contains the local reference of PstI-based epiGBS fragments as inferred from the bisulphite reads. Methylation variant calling was based on SAMtools mpileup and custom python scripts, following a similar approach as described in van Gurp et al. [32]. For downstream analysis, we included only those cytosines that were called in at least 80% of the samples. In addition, cytosine positions that did not pass the filtering criteria for all generations were removed.

To obtain methylation status calls, we implemented a one-tail binomial test as previously described [12]. Multiple testing correction was performed using the Benjamini-Yekutiely method [59], and the false discovery rate (FDR) was controlled at 0.05. All statistical tests for obtaining methylation status calls of the samples were conducted within the SciPy ecosystem.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-02161-6.

Additional file 1: Table S1. WGBS information for MA pedigrees MA1_1, MA1_3 and MA3.

Additional file 2: Table S2. Epimutation rate estimates and model selection results for pedigree MA1_1.

Additional file 3: Table S3. Epimutation rate estimates and model selection results for pedigree MA1_3.

Additional file 4: Table S4. Epimutation rate estimates and model selection results for pedigree MA3.

Additional file 5: Table S5. Pre-processing of WGBS data using MethylStar increases the number of high-confident cytosines that can be used for epimutation analysis compared with previous pre-processing approaches.

Additional file 6: Figure S1. Developmental origin of somatic epimutations in plants.

Additional file 7: Review history.

Abbreviations

WGBS: Whole-genome bisulfite sequencing; TEs: Transposable elements; MA lines: Mutation accumulation lines; RRBS: Reduced representation bisulphite sequencing; epiGBS: Epigenotyping by sequencing; 5mC: 5-Methyl cytosine; RdDM: RNA-directed DNA methylation pathway; CMT2: CHROMOMETHYLASE 2; HMM: Hidden Markov model; FDR: False discovery rate

Acknowledgements

We thank Kay Schneitz for discussing plant development with us, Cristina Cipriani for early tests of the optimX package, and Keith Slotkin for the sRNA data.

Review history

The review history is available as Additional file 7.

Peer review information

Anahita Bishop was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

FJ and MCT conceptualized the method. FJ, YS, and RRH implemented and documented the method. FJ, YS, AS, RRH, JD, TM, BTH, and TvG analyzed the data. KV, GT, and RJS contributed materials. FJ wrote the paper with input from all coauthors. The authors read and approved the final manuscript.

Funding

FJ, RJS, YS, RRH, and TM acknowledge support from the Technical University of Munich-Institute for Advanced Study funded by the German Excellent Initiative and the European Seventh Framework Programme under grant agreement no. 291763. RJS acknowledges the support from the National Science Foundation (IOS-1546867). RJS is a Pew Scholar in the Biomedical Sciences, supported by the Pew Charitable Trusts. FJ and YS were also supported by the SFB Sonderforschungsbereich924 of the Deutsche Forschungsgemeinschaft (DFG). Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials
AlphaBeta [60] is an open source R package licensed under GPL-3. It is freely and openly available from the Github website (https://github.com/jlab-code/AlphaBeta) under GNU General Public License v3.0, and it is part of Bioconductor [61]. Schmitz RJ. AlphaBeta: Computational inference of epimutation rates and spectra from high-throughput DNA methylation data in plants. GSE153055. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE153055 (2020) [62]. Van Gurp TP, Wagemaker NCAM, Verhoeven KJF. Epimutation accumulation experiment in two Taraxacum officinale apomicts. BioProject PRJNA608438. https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA608438.

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Technical University of Munich, Department of Plant Sciences, Liesel-Beckmann-Str. 2, 85354 Freising, Germany. [2]Technical University of Munich, Institute for Advanced Study, Lichtenbergstr. 2a, 85748 Garching, Germany. [3]Institute of Computational Biology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany. [4]Institute of Bioinformatics, 120 East Green Street, Athens, 30602 USA. [5]European Research Institute for the Biology of Ageing, University of Groningen, University Medical Centre Groningen, A. Deusinglaan 1, 9713 AV Groningen, Netherlands. [6]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Emil-Erlenmeyer-Forum 2, 85354 Freising, Germany. [7]Netherlands Institute of Ecology (NIOO-KNAW), Department of Terrestrial Ecology, Wageningen, Wageningen, The Netherlands. [8]The Center for Bioenergy Innovation, Oak Ridge National Laboratory, Oak Ridge, USA. [9]Department of Genetics, The University of Georgia, 120 East Green Street, 30602 Athens, USA.

Received: 3 December 2019 Accepted: 2 September 2020
Published online: 06 October 2020

References
1. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010;11(3):204–20. https://doi.org/10.1038/nrg2719.
2. Stroud H, Greenberg MVC, Feng S, Bernatavichute YV, Jacobsen SE. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell. 2013;152(1-2):352–64. https://doi.org/10.1016/j.cell.2012.10.054.
3. Bewick AJ, Hofmeister BT, Powers RA, Mondo SJ, Grigoriev IV, James TY, Stajich JE, Schmitz RJ. Diversity of cytosine methylation across the fungal tree of life. Nat Ecol Evol. 2019;3(3):479. https://doi.org/10.1038/s41559-019-0810-9.
4. Feng S, Cokus SJ, Zhang X, Chen P-Y, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, Ukomadu C, Sadler KC, Pradhan S, Pellegrini M, Jacobsen SE. Proc Natl Acad Sci USA. 2010;107(19):8689–94. https://doi.org/10.1073/pnas.1002720107.
5. Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, Li Q, Rohr NA, Rambani A, Burke JM, Udall JA, Egesi C, Schmutz J, Grimwood J, Jackson SA, Springer NM, Schmitz RJ. Widespread natural variation of DNA methylation within angiosperms. Genome Biol. 2016;17:. https://doi.org/10.1186/s13059-016-1059-0.
6. Takuno S, Ran J-H, Gaut BS. Evolutionary patterns of genic DNA methylation vary across land plants. Nat Plants. 2016;2(2):15222. https://doi.org/10.1038/nplants.2015.222.
7. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science (New York, NY). 2010;328(5980):916–9. https://doi.org/10.1126/science.1186366.
8. Field AE, Robertson NA, Wang T, Havas A, Ideker T, Adams PD. DNA methylation clocks in aging: categories, causes, and consequences. Mol Cell. 2018;71(6):882–95. https://doi.org/10.1016/j.molcel.2018.08.008.
9. Calarco JP, Borges F, Donoghue MTA, Van Ex F, Jullien PE, Lopes T, Gardner R, Berger F, Feijó JA, Becker JD, Martienssen RA. Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. Cell. 2012;151(1):194–205. https://doi.org/10.1016/j.cell.2012.09.001.
10. Walker J, Gao H, Zhang J, Aldridge B, Vickers M, Higgins JD, Feng X. Sexual-lineage-specific DNA methylation regulates meiosis in Arabidopsis. Nat Genetics. 2018;50(1):130. https://doi.org/10.1038/s41588-017-0008-5.
11. Johannes F, Schmitz RJ. Spontaneous epimutations in plants. New Phytologist. 2019;221(3):1253–9. https://doi.org/10.1111/nph.15434.
12. Graaf AVD, Wardenaar R, Neumann DA, Taudt A, Shaw RG, Jansen RC, Schmitz RJ, Colomé-Tatché M, Johannes F. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. Proc Natl Acad Sci. 2015;112(21):6676–81. https://doi.org/10.1073/pnas.1424254112.
13. Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, Weigel D. Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. Nature. 2011;480(7376):245–9. https://doi.org/10.1038/nature10555.
14. Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, Schork NJ, Ecker JR. Transgenerational epigenetic instability is a source of novel methylation variants. Science (New York, NY). 2011;334(6054):369–73. https://doi.org/10.1126/science.1212959.
15. Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science (New York, NY). 2010;327(5961):92–4. https://doi.org/10.1126/science.1180677.
16. Weng M-L, Becker C, Hildebrandt J, Neumann M, Rutter MT, Shaw RG, Weigel D, Fenster CB. Fine-grained analysis of spontaneous mutation spectrum and frequency in Arabidopsis thaliana. Genetics. 2019;211(2):703–14. https://doi.org/10.1534/genetics.118.301721.
17. Vidalis A, Živković D, Wardenaar R, Roquis D, Tellier A, Johannes F. Methylome evolution in plants. Genome Biol. 2016;17(1):264. https://doi.org/10.1186/s13059-016-1127-5.

18. Hofmeister BT, Lee K, Rohr NA, Hall DW, Schmitz RJ. Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation. Genome Biol. 2017;18(1):155. https://doi.org/10.1186/s13059-017-1288-x.
19. Hagmann J, Becker C, Müller J, Stegle O, Meyer RC, Wang G, Schneeberger K, Fitz J, Altmann T, Bergelson J, Borgwardt K, Weigel D. Century-scale methylome stability in a recently diverged Arabidopsis thaliana lineage. PLoS Genet. 2015;11(1):1004920. https://doi.org/10.1371/journal.pgen.1004920.
20. Schmid MW, Heichinger C, Schmid DC, Guthörl D, Gagliardini V, Bruggmann R, Aluri S, Aquino C, Schmid B, Turnbull LA, Grossniklaus U. Contribution of epigenetic variation to adaptation in Arabidopsis. Nat Commun. 2018;9(1):1–12. https://doi.org/10.1038/s41467-018-06932-5.
21. Taudt A, Colomé-Tatché M, Johannes F. Genetic sources of population epigenomic variation. Nat Rev Genet. 2016;17(6):319–32. https://doi.org/10.1038/nrg.2016.45.
22. Jiang C, Mithani A, Belfield EJ, Mott R, Hurst LD, Harberd NP. Environmentally responsive genome-wide accumulation of de novo Arabidopsis thaliana mutations and epimutations. Genome Res. 2014;24(11):1821–9. https://doi.org/10.1101/gr.177659.114.
23. Ganguly DR, Crisp PA, Eichten SR, Pogson BJ. The Arabidopsis DNA methylome is stable under transgenerational drought stress. Plant Phys. 2017;175(4):1893–912. https://doi.org/10.1104/pp.17.00744.
24. Zheng X, Chen L, Xia H, Wei H, Lou Q, Li M, Li T, Luo L. Transgenerational epimutations induced by multi-generation drought imposition mediate rice plant's adaptation to drought condition. Sci Rep. 2017;7:39843. https://doi.org/10.1038/srep39843.
25. Lanfear R. Do plants have a segregated germline? PLOS Biol. 2018;16(5):2005439. https://doi.org/10.1371/journal.pbio.2005439.
26. Hofmeister BT, et al. A genome assembly and the somatic genetic and epigenetic mutation rate in a wild long-lived perennial Populus trichocarpa. Genome Biol. 2020. https://doi.org/10.1186/s13059-020-02162-5.
27. Horvath R, Laenen B, Takuno S, Slotte T. Single-cell expression noise and gene-body methylation in Arabidopsis thaliana. Heredity. 2019;1:. https://doi.org/10.1038/s41437-018-0181-z.
28. Secco D, Wang C, Shou H, Schultz MD, Chiarenza S, Nussaume L, Ecker JR, Whelan J, Lister R. Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements. eLife. 2015;4:. https://doi.org/10.7554/eLife.09343.
29. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. Nature. 2008;452(7184):215–9. https://doi.org/10.1038/nature06745.
30. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008;133(3):523–36. https://doi.org/10.1016/j.cell.2008.03.029.
31. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 2005;33(18):5868–77. https://doi.org/10.1093/nar/gki901.
32. van Gurp TP, Wagemaker NCAM, Wouters B, Vergeer P, Ouborg JNJ, Verhoeven KJF. epiGBS: reference-free reduced representation bisulfite sequencing. Nat Methods. 2016;13(4):322–4. https://doi.org/10.1038/nmeth.3763.
33. Colomé-Tatché M, Johannes F. Signatures of Dobzhansky–Muller incompatibilities in the genomes of eecombinant inbred lines. Genetics. 2016;202(2):825–41. https://doi.org/10.1534/genetics.115.179473.
34. Broman KW. Genotype probabilities at intermediate generations in the construction of recombinant Inbred Lines. Genetics. 2012;190(2):403–12. https://doi.org/10.1534/genetics.111.132647.
35. Johannes F, Colomé-Tatché M. Quantitative epigenetics through epigenomic perturbation of isogenic lines. Genetics. 2011;188(1):215–27. https://doi.org/10.1534/genetics.111.127118.
36. Bartlett MS, Haldane JBS. The theory of inbreeding with forced heterozygosis. J Genet. 1935;31(3):327. https://doi.org/10.1007/BF02982404.
37. Ronald Aylmer Fisher. The theory of inbreeding. Edinburgh: Oliver and Boyd; 1949.
38. Kawashima T, Berger F. Epigenetic reprogramming in plant sexual reproduction. Nat Rev Genet. 2014;15(9):613–24. https://doi.org/10.1038/nrg3685.
39. Gehring M. Epigenetic dynamics during flowering plant reproduction: evidence for reprogramming? New Phytol. https://doi.org/10.1111/nph.15856.
40. Teixeira FK, Heredia F, Sarazin A, Roudier F, Boccara M, Ciaudo C, Cruaud C, Poulain J, Berdasco M, Fraga MF, Voinnet O, Wincker P, Esteller M, Colot V. A role for RNAi in the selective correction of DNA methylation defects. Science. 2009;323(5921):1600–4. https://doi.org/10.1126/science.1165313.
41. Gouil Q, Baulcombe DC. DNA methylation signatures of the plant chromomethyltransferases. PLOS Genet. 2016;12(12):1006526. https://doi.org/10.1371/journal.pgen.1006526.
42. Shahryary Y, Hazarika RR, Johannes F. Methylstar: a fast and robust pre-processing pipeline for bulk or single-cell whole-genome bisulfite sequencing data. BMC Genomics. 2020;21(1):479.
43. Wang L, Ji Y, Hu Y, Hu H, Jia X, Jiang M, Zhang X, Zhao L, Zhang Y, Jia Y, Qin C, Yu L, Huang J, Yang S, Hurst LD, Tian D. The architecture of intra-organism mutation rate variation in plants. PLOS Biol. 2019;17(4):3000191. https://doi.org/10.1371/journal.pbio.3000191.
44. Hanlon VCT, Otto SP, Aitken SN. Somatic mutations substantially increase the per-generation mutation rate in the conifer Picea sitchensis. Evol Lett. https://doi.org/10.1002/evl3.121.
45. Schmid-Siegert E, Sarkar N, Iseli C, Calderon S, Gouhier-Darimont C, Chrast J, Cattaneo P, Schütz F, Farinelli L, Pagni M, Schneider M, Voumard J, Jaboyedoff M, Fankhauser C, Hardtke CS, Keller L, Pannell JR, Reymond A, Robinson-Rechavi M, Xenarios I, Reymond P. Low number of fixed somatic mutations in a long-lived oak tree. Nat Plants. 2017;3(12):926. https://doi.org/10.1038/s41477-017-0066-9.
46. Orr AJ, Padovan A, Kainer D, Külheim C, Bromham L, Bustos-Segura C, Foley W, Haff T, Hsieh J-F, Morales-Suarez A, Cartwright RA, Lanfear R. A phylogenomic approach reveals a low somatic mutation rate in a long-lived plant. bioRxiv. 2019727982. https://doi.org/10.1101/727982.

47. Ingvarsson PK. Multilocus patterns of nucleotide polymorphism and the demographic history of Populus tremula. Genetics. 2008;180(1):329–40. https://doi.org/10.1534/genetics.108.090431.
48. Verhoeven KJF, Van Dijk PJ, Biere A. Changes in genomic methylation patterns during the formation of triploid asexual dandelion lineages. Mol Ecol. 2010;19(2):315–24. https://doi.org/10.1111/j.1365-294X.2009.04460.x.
49. Koltunow A. Apomixis: embryo sacs and embryos formed without meiosis or fertilization in ovules,. Plant Cell. 1993;5(10):1425–37.
50. Moorsel S. J. v., Schmid MW, Wagemaker NCAM, Gurp T. v., Schmid B, Vergeer P. Evidence for rapid evolution in a grassland biodiversity experiment. bioRxiv. 2018262303. https://doi.org/10.1101/262303.
51. Bewick AJ, Niederhuth CE, Ji L, Rohr NA, Griffin PT, Leebens-Mack J, Schmitz RJ. The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants. Genome Biol. 2017;18(1):65. https://doi.org/10.1186/s13059-017-1195-1.
52. Gaiti F, Chaligne R, Gu H, Brand RM, Kothen-Hill S, Schulman RC, Grigorev K, Risso D, Kim K-T, Pastore A, Huang KY, Alonso A, Sheridan C, Omans ND, Biederstedt E, Clement K, Wang L, Felsenfeld JA, Bhavsar EB, Aryee MJ, Allan JN, Furman R, Gnirke A, Wu CJ, Meissner A, Landau DA. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. Nature. 2019;1:. https://doi.org/10.1038/s41586-019-1198-z.
53. Danese A, Richter ML, Fischer DS, Theis FJ, Colomé-Tatché M. EpiScanpy: integrated single-cell epigenomic analysis. bioRxiv. 2019648097. https://doi.org/10.1101/648097.
54. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019;37(5):547–54. https://doi.org/10.1038/s41587-019-0071-9.
55. Taudt A, Roquis D, Vidalis A, Wardenaar R, Johannes F, Colomé-Tatché M. METHimpute: imputation-guided construction of complete methylomes from WGBS data. BMC Genomics. 2018;19(1):444. https://doi.org/10.1186/s12864-018-4641-x.
56. Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nat Protocol. 2015;10(3):475–83. https://doi.org/10.1038/nprot.2014.114.
57. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H, Lin S, Lin Y, Jung I, Schmitt AD, Selvaraj S, Ren B, Sejnowski TJ, Wang W, Ecker JR. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015;523(7559):212–6. https://doi.org/10.1038/nature14465.
58. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLOS ONE. 2011;6(5):19379. https://doi.org/10.1371/journal.pone.0019379.
59. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001;29(4):1165–88. https://doi.org/10.1214/aos/1013699998.
60. Shahryary Y, Johannes F, Hazarika R. jlab-code/AlphaBeta. 2020. https://doi.org/10.5281/zenodo.3992612.
61. Shahryary Y, Johannes F, Hazarika R. Bioconductor AlphaBeta Software Package. 2020. https://doi.org/10.18129/B9.bioc.AlphaBeta.
62. Schmitz RJ. AlphaBeta: Computational inference of epimutation rates and spectra from high-throughput DNA methylation data in plants. GSE153055. 2020. https://urldefense.proofpoint.com/v2/url?u=https-3A__www.ncbi.nlm.nih.gov_geo_query_acc.cgi-3Facc-3DGSE153055&d=DwIGaQ&c=vh6FgFnduejNhPPD0fl_yRaSfZy8CWbWnlf4XJhSqx8&r=Z3BY_DFGt24T_Oe13xHJ2wIDudwzO_8VrOFSUQlQ_zsz-DGcYuoJS3jWWxMQECLm&m=nMao27rggwqbBJbvu1-d0yavK1ZEszYRhgNn0-mmx8g&s=HsUT2FBGvJLvyqtcALnMlH07FzdJt3Uw2EtIold06B0&e=.

Publisher's Note

B.1 Additional file 1 — Table S1

Table S1

samples	platform	library layout	Mean Coverage(X)
MA1_1_G31_109_r1	Illumina GAIIx	Paired-end	10.75
MA1_1_G31_109_r2	Illumina GAIIx	Paired-end	7.53
MA1_1_G31_119_r1	Illumina GAIIx	Paired-end	9.91
MA1_1_G31_119_r2	Illumina GAIIx	Paired-end	10.5
MA1_1_G31_29_r1	Illumina GAIIx	Paired-end	7.91
MA1_1_G31_29_r2	Illumina GAIIx	Paired-end	13.53
MA1_1_G31_29_r3	Illumina GAIIx	Paired-end	7.86
MA1_1_G31_39_r1	Illumina GAIIx	Paired-end	12.29
MA1_1_G31_39_r2	Illumina GAIIx	Paired-end	6.98
MA1_1_G31_49_r1	Illumina GAIIx	Paired-end	8.71
MA1_1_G31_49_r2	Illumina GAIIx	Paired-end	11.26
MA1_1_G31_59_r1	Illumina GAIIx	Paired-end	11.05
MA1_1_G31_59_r2	Illumina GAIIx	Paired-end	7.07
MA1_1_G31_79_r1	Illumina GAIIx	Paired-end	11.28
MA1_1_G31_79_r2	Illumina GAIIx	Paired-end	13.77
MA1_1_G31_89_r1	Illumina GAIIx	Paired-end	9.28
MA1_1_G31_89_r2	Illumina GAIIx	Paired-end	7.94
MA1_1_G31_99_r1	Illumina GAIIx	Paired-end	10.83
MA1_1_G31_99_r2	Illumina GAIIx	Paired-end	7.18
MA1_1_G31_99_r3	Illumina GAIIx	Paired-end	12.96
MA1_1_G32_39_r1	Illumina GAIIx	Paired-end	9.81
MA1_1_G32_39_r2	Illumina GAIIx	Paired-end	9.43
MA1_1_G32_49_r1	Illumina GAIIx	Paired-end	8.65
MA1_1_G32_49_r2	Illumina GAIIx	Paired-end	6.6
MA1_1_G3_26_r1	Illumina GAIIx	Paired-end	7.99
MA1_1_G3_87_r1	Illumina GAIIx	Paired-end	10.97
MA1_1_G3_87_r2	Illumina GAIIx	Paired-end	7.66
MA1_3_G18_12_r1	Illumina NextSeq 500	Single-end	5.86
MA1_3_G19_12_r1	Illumina NextSeq 500	Single-end	6.25
MA1_3_G20_12_r1	Illumina NextSeq 500	Single-end	6.56
MA1_3_G21_12_r1	Illumina NextSeq 500	Single-end	8.55
MA1_3_G25_12_r1	Illumina NextSeq 500	Single-end	6.46
MA1_3_G26_12_r1	Illumina NextSeq 500	Single-end	6.75
MA1_3_G28_12_r1	Illumina NextSeq 500	Single-end	8.18
MA1_3_G29_12_r1	Illumina NextSeq 500	Single-end	6.67
MA1_3_G30_12_r1	Illumina NextSeq 500	Single-end	7.2
MA3_G0	Illumina NextSeq 500	Single-end	9.44
MA3_G11_L2	Illumina NextSeq 500	Single-end	10.31
MA3_G11_L8	Illumina NextSeq 500	Single-end	23.52
MA3_G1_L2	Illumina NextSeq 500	Single-end	8.85
MA3_G1_L8	Illumina NextSeq 500	Single-end	8.47
MA3_G2_L2	Illumina NextSeq 500	Single-end	9.91
MA3_G2_L8	Illumina NextSeq 500	Single-end	9.36
MA3_G4_L2	Illumina NextSeq 500	Single-end	9.8
MA3_G4_L8	Illumina NextSeq 500	Single-end	9.38
MA3_G5_L2	Illumina NextSeq 500	Single-end	8.82
MA3_G5_L8	Illumina NextSeq 500	Single-end	8.98
MA3_G8_L2	Illumina NextSeq 500	Single-end	8.97
MA3_G8_L8	Illumina NextSeq 500	Single-end	9.43

Table S1: WGBS information for MA pedigrees MA1_1, MA1_3 and MA3**.**

B.2 Additional file 2 — Table S2

Table S2

A. thaliana (MA1_1)

context	annotation	alpha	beta	beta/alpha	FM	RM	F-value	df RM	df FM	P--value
CG	global	8.605897E-05	0.0002497981	2.903	ABneutral	Abnull	461.7119	350	346	2.70488E-137
CG	exon	0.0003329146	0.0008854249	2.660	ABneutral	Abnull	506.9881	350	346	2.99272E-143
CG	promoter	4.390624E-05	0.0003396206	7.735	ABneutral	Abnull	574.9184	350	346	2.19508E-151
CG	TE	2.777663E-05	7.500164E-06	0.270	ABneutral	Abnull	122.3618	350	346	6.001881E-65
CG	global				ABselectUU	Abneutral	0.4806	347	346	0.4885955
CG	exon				ABselectUU	Abneutral	0.0453	347	346	0.8316087
CG	promoter				ABselectUU	Abneutral	1.8765	347	346	0.1716251
CG	TE				ABselectUU	Abneutral	0.7210	347	346	0.3963924
CG	global				ABselectMM	Abneutral	0.4731	347	346	0.4920452
CG	exon				ABselectMM	Abneutral	0.1176	347	346	0.7318225
CG	promoter				ABselectMM	Abneutral	0.4049	347	346	0.5249724
CG	TE				ABselectMM	Abneutral	0.1476	347	346	0.7011182
CHG	global	3.533486E-06	5.84628E-05	16.545	ABneutral	Abnull	16.4449	350	346	2.399368E-12
CHG	exon	1.537786E-06	0.0002547714	165.674	ABneutral	Abnull	20.6207	350	346	2.958294E-15
CHG	promoter	2.408792E-06	0.0001227248	50.949	ABneutral	Abnull	19.3769	350	346	2.124209E-14
CHG	TE	2.124037E-05	3.025496E-05	1.424	ABneutral	Abnull	9.4822	350	346	2.768493E-07
CHG	global				ABselectUU	Abneutral	0.0000	347	346	1
CHG	exon				ABselectUU	Abneutral	0.0164	347	346	0.8980838
CHG	promoter				ABselectUU	Abneutral	0.0000	347	346	1
CHG	TE				ABselectUU	Abneutral	0.0000	347	346	1
CHG	global				ABselectMM	Abneutral	0.0000	347	346	1
CHG	exon				ABselectMM	Abneutral	0.0228	347	346	8.80E-01
CHG	promoter				ABselectMM	Abneutral	0.0000	347	346	1
CHG	TE				ABselectMM	Abneutral	0.0000	347	346	1
CHH	global	1.905105E-06	0.0001614412	84.741	ABneutral	Abnull	11.6186	350	346	7.294017E-09
CHH	exon	1.09903E-06	0.0006017335	547.513	ABneutral	Abnull	18.1246	350	346	1.577799E-13
CHH	promoter	1.349004E-06	0.0002622809	194.426	ABneutral	Abnull	19.3769	350	346	2.124209E-14
CHH	TE	5.54414E-06	6.181445E-05	11.150	ABneutral	Abnull	9.4822	350	346	2.768493E-07
CHH	global				ABselectUU	Abneutral	0.9755	347	346	0.3240002
CHH	exon				ABselectUU	Abneutral	0.0000	347	346	1
CHH	promoter				ABselectUU	Abneutral	0.0000	347	346	1
CHH	TE				ABselectUU	Abneutral	0.0000	347	346	1
CHH	global				ABselectMM	Abneutral	0.0000	347	346	1
CHH	exon				ABselectMM	Abneutral	0.0000	347	346	1
CHH	promoter				ABselectMM	Abneutral	0.0000	347	346	1
CHH	TE				ABselectMM	Abneutral	0.0000	347	346	1

FM = Full model
RM = Reduced model
df = degrees of freedom
Best performing model

Table S2: Epimutation rate estimates and model selection results for pedigree MA1_1

B.3 Additional file 3 — Table S3

Table S3

A. thaliana (MA1_3)

context	annotation	alpha	beta	beta/alpha	FM	RM	F-value	df RM	df FM	P--value
CG	global	0.0001411325	0.0006480785	4.592	ABneutral	Abnull	26.0964	35	31	1.545386E-09
CG	exon	0.0004417924	0.001566664	3.546	ABneutral	Abnull	36.7694	35	31	2.352825E-11
CG	promoter	8.473369E-05	0.0009098872	10.738	ABneutral	Abnull	22.7134	35	31	7.66475E-09
CG	TE	0.0002361935	0.0001108124	0.469	ABneutral	Abnull	7.9028	35	31	0.0001635841
CG	global				ABselectUU	Abneutral	0.6312	32	31	0.4329535
CG	exon				ABselectUU	Abneutral	0.1957	32	31	0.6612462
CG	promoter				ABselectUU	Abneutral	0.4402	32	31	0.5119528
CG	TE				ABselectUU	Abneutral	0.0135	32	31	0.9084139
CG	global				ABselectMM	Abneutral	0.6343	32	31	0.4318425
CG	exon				ABselectMM	Abneutral	0.2816	32	31	0.5994397
CG	promoter				ABselectMM	Abneutral	0.2889	32	31	0.5947381
CG	TE				ABselectMM	Abneutral	0.0163	32	31	0.8993372
CHG	global	NA	NA	NA	ABneutral	Abnull	0.6354	35	31	0.6410907
CHG	exon	NA	NA	NA	ABneutral	Abnull	0.4926	35	31	0.7411286
CHG	promoter	NA	NA	NA	ABneutral	Abnull	0.7356	35	31	0.5747893
CHG	TE	NA	NA	NA	ABneutral	Abnull	0.3041	35	31	0.8729979
CHG	global				ABselectUU	Abneutral	0.0271	32	31	0.8703212
CHG	exon				ABselectUU	Abneutral	0.0000	32	31	1
CHG	promoter				ABselectUU	Abneutral	0.1223	32	31	0.7289703
CHG	TE				ABselectUU	Abneutral	0.0759	32	31	0.7847896
CHG	global				ABselectMM	Abneutral	0.0312	32	31	0.8609342
CHG	exon				ABselectMM	Abneutral	0.0000	32	31	1.00E+00
CHG	promoter				ABselectMM	Abneutral	0.1063	32	31	0.7466025
CHG	TE				ABselectMM	Abneutral	0.0779	32	31	0.7820306
CHH	global	NA	NA	NA	ABneutral	Abnull	0.6695	35	31	0.6180636
CHH	exon	NA	NA	NA	ABneutral	Abnull	0.5049	35	31	0.732369
CHH	promoter	NA	NA	NA	ABneutral	Abnull	1.2939	35	31	0.2938961
CHH	TE	NA	NA	NA	ABneutral	Abnull	0.3691	35	31	0.8287454
CHH	global				ABselectUU	Abneutral	0.0018	32	31	0.9661093
CHH	exon				ABselectUU	Abneutral	0.0000	32	31	1
CHH	promoter				ABselectUU	Abneutral	0.0016	32	31	0.9687817
CHH	TE				ABselectUU	Abneutral	0.7337	32	31	0.3982717
CHH	global				ABselectMM	Abneutral	0.0079	32	31	0.9296158
CHH	exon				ABselectMM	Abneutral	0.0000	32	31	1
CHH	promoter				ABselectMM	Abneutral	0.0003	32	31	0.9871071
CHH	TE				ABselectMM	Abneutral	0.0325	32	31	0.8581938

FM = Full model
RM = Reduced model
df = degrees of freedom
Best performing model

Table S3: Epimutation rate estimates and model selection results for pedigree MA1_3

B.4 Additional file 4 — Table S4

Table S4

A. thaliana (MA3)

context	annotation	alpha	beta	beta/alpha	FM	RM	F-value	df RM	df FM	P--value
CG	**global**	0.0001942304	0.0008340002	4.294	ABneutral	Abnull	136.2307	65	61	1.103063E-29
CG	**exon**	0.0005635082	0.001829657	3.247	ABneutral	Abnull	210.1967	65	61	6.177088E-35
CG	**promoter**	0.0001128601	0.001141507	10.114	ABneutral	Abnull	108.4535	65	61	5.190251E-27
CG	**TE**	NA	NA	NA	ABneutral	Abnull	1.0608	65	61	0.383711
CG	**global**				ABselectUU	Abneutral	0.0000	62	61	1
CG	**exon**				ABselectUU	Abneutral	0.0000	62	61	1
CG	**promoter**				ABselectUU	Abneutral	0.1053	62	61	0.7467267
CG	**TE**				ABselectUU	Abneutral	0.0000	62	61	1
CG	**global**				ABselectMM	Abneutral	0.0001	62	61	0.9934013
CG	**exon**				ABselectMM	Abneutral	0.0000	62	61	1
CG	**promoter**				ABselectMM	Abneutral	0.0000	62	61	1
CG	**TE**				ABselectMM	Abneutral	0.1116	62	61	0.739439
CHG	**global**	NA	NA	NA	ABneutral	Abnull	2.0182	65	61	0.1030624
CHG	**exon**	NA	NA	NA	ABneutral	Abnull	0.0141	65	61	0.9995992
CHG	**promoter**	1.783862E-06	0.0001055335	59.160	ABneutral	Abnull	5.0076	65	61	0.001479819
CHG	**TE**	5.517707E-05	0.0001307905	2.370	ABneutral	Abnull	5.7078	65	61	0.0005720866
CHG	**global**				ABselectUU	Abneutral	0.9363	62	61	0.3370451
CHG	**exon**				ABselectUU	Abneutral	0.0000	62	61	1
CHG	**promoter**				ABselectUU	Abneutral	0.0707	62	61	0.7912417
CHG	**TE**				ABselectUU	Abneutral	0.5221	62	61	0.4727021
CHG	**global**				ABselectMM	Abneutral	0.8975	62	61	0.3471843
CHG	**exon**				ABselectMM	Abneutral	0.0000	62	61	1.00E+00
CHG	**promoter**				ABselectMM	Abneutral	0.0000	62	61	1
CHG	**TE**				ABselectMM	Abneutral	0.8804	62	61	0.3517958
CHH	**global**	NA	NA	NA	ABneutral	Abnull	1.0187	65	61	0.404864
CHH	**exon**	NA	NA	NA	ABneutral	Abnull	0.1767	65	61	0.9495845
CHH	**promoter**	1.71043E-08	4.1274E-06	241.308	ABneutral	Abnull	2.9878	65	61	0.02558529
CHH	**TE**	NA	NA	NA	ABneutral	Abnull	0.2803	65	61	0.8896153
CHH	**global**				ABselectUU	Abneutral	0.7073	62	61	0.4036315
CHH	**exon**				ABselectUU	Abneutral	0.4033	62	61	0.5277759
CHH	**promoter**				ABselectUU	Abneutral	0.0000	62	61	1
CHH	**TE**				ABselectUU	Abneutral	0.0000	62	61	1
CHH	**global**				ABselectMM	Abneutral	1.2129	62	61	0.275082
CHH	**exon**				ABselectMM	Abneutral	0.2588	62	61	0.6127902
CHH	**promoter**				ABselectMM	Abneutral	0.0000	62	61	1
CHH	**TE**				ABselectMM	Abneutral	0.0684	62	61	0.7945567

FM = Full model
RM = Reduced model
df = degrees of freedom
Best performing model

Table S4: Epimutation rate estimates and model selection results for pedigree MA3

B.5 Additional file 5 — Table S5

Table S5

	coverage > 3				posteriorMax >=0.99			
MA	**AllC**	**CG**	**CHG**	**CHH**	**AllC**	**CG**	**CHG**	**CHH**
MA1_1	13417233	2344610	2513665	8558958	23428047	3030753	3797292	16600002
MA1_3	25009841	3450911	3937253	20731677	29141823	3753716	4183152	21204955
MA3	25679815	3086050	3542387	19051378	24357974	3655356	4131622	16570996

Previous methods
MethylStar

Table S5: Pre-processing of WGBS data using MethylStar increases the number of high-confident cytosines that can be used for epimutation analysis compared with previous pre-processing approaches.

B.6 Additional file 6 — Figure S1

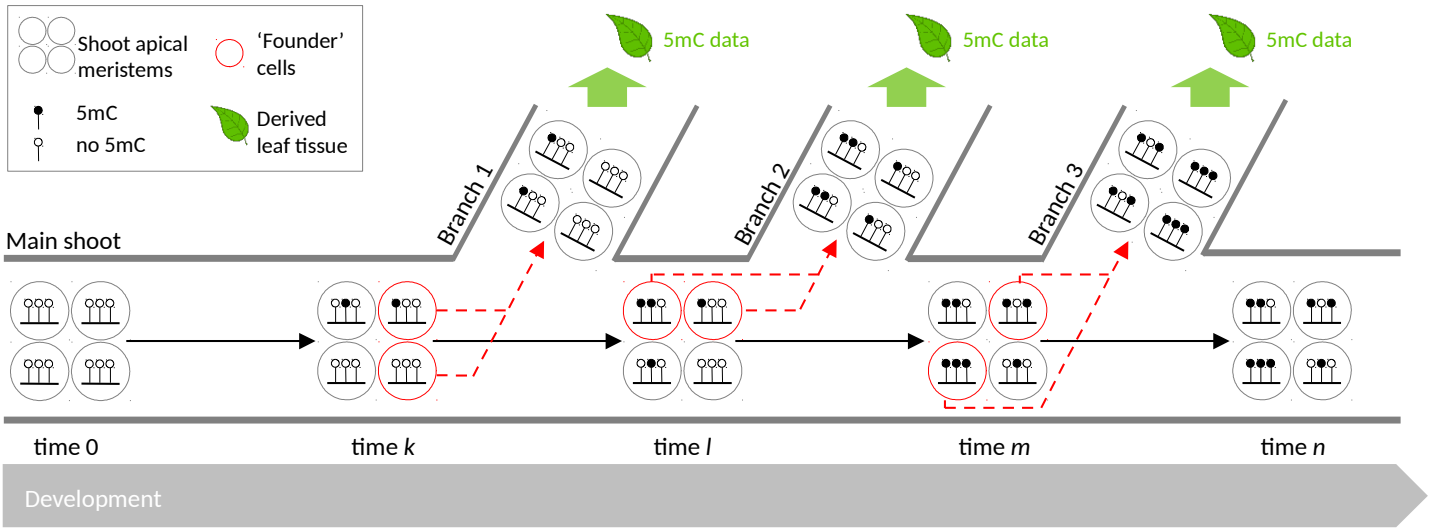

Fig. S1: **Developmental origin of somatic epimutations in plants.** The failure to maintain the methylation status of cytosines during the mitotic maintenance of shoot apical meristematic cell pools leads to spontaneous somatic epimutations. Shown here are only spontaneous gains of methylation, for simplicity. A small set of 'founder' cells gives rise to lateral branches at developmental times *k*, *l,* and *m*. The random sampling of founder cells creates a bottleneck which increases the frequency of somatic epimutations in the cell populations of lateral branches. Somatic epimutation accumulation in shoot apical meristems thus leads to increased 5mC divergence between leafs originating from different lateral branches (e.g. leaf methylomes from Branch 1 and 2 are more similar than those from Branch 1 and 3).