Methylomes reveal recent evolutionary changes in populations of two plant species

Kevin Korfmann[ab1], Andreas Zauchner[a1], Bing Huo[a], Corinna Grünke[a], Yitong Wang[a], Aurélien Tellier[a†2], Ramesh Arunkumar [a†2]

[a]Section of Population Genetics, School of Life Sciences, Technical University of Munich, Liesel-Beckmann Strasse 2, 85354 Freising, Germany

[b]Present address: Department of Biology 77 Klamath Hall 1210 University of Oregon Eugene, OR 97403

[1] These authors contributed equally

[2] These authors contributed equally

[†] Authors for correspondence (Aurélien Tellier - aurelien.tellier@tum.de, Ramesh Arunkumar – arunkumar.ramesh@tum.de)

Abstract

Plant DNA methylation changes occur hundreds to thousands of times faster than DNA mutations and can be transmitted transgenerationally, making them useful for studying population-scale patterns in clonal or selfing species. However, a state-of-the-art approach to use them for inferring population genetic processes and demographic histories is lacking. To address this, we compare evolutionary signatures extracted from CG methylomes and genomes in *Arabidopsis thaliana* and *Brachypodium distachyon*. While methylation variants (SMPs) are less effective than genetic variants (SNPs) for identifying population differentiation in *A. thaliana*, they can classify phenotypically divergent *B. distachyon* subgroups that are otherwise genetically undistinguishable. The site frequency spectra generated using methylation sites from varied genomic locations and evolutionary conservation exhibit similar shapes, indicating minimal noise when all CG sites are analyzed jointly. Nucleotide diversity is three orders of magnitude higher for methylation variants than for genetic variants in both species, driven by the higher epimutation rate. Correlations between SNPs and SMPs in nucleotide diversity and allele frequencies at gene exons are weak or absent in *A. thaliana*, possibly because the two sources of variation reflect evolutionary forces acting at different timescales. Linkage disequilibrium quickly decays within 250 bp for methylation variants in both plant species. Finally, we developed a deep learning-based demographic inference approach. We identified recent population expansions in *A. thaliana* and *B. distachyon* using methylation variants that were not identified when using genetic variants. Our study demonstrates the unique evolutionary insights methylomes provide that genetic variation alone cannot reveal.

Keywords: Methylation site frequency spectra, demographic inference, deep learning, Tajima's *D*, linkage disequilibrium, nucleotide diversity, clocklike regions, gene body methylation, *Arabidopsis thaliana*, *Brachypodium distachyon*

Background

Understanding how variation is partitioned within and between populations is a central goal in evolutionary biology. Increasing evidence suggests that epigenetic modifications can vary within and between populations [1]. Nevertheless, our knowledge of how epigenomes evolve through neutral and selective forces and, consequently, their capacity to reveal population demographic histories is limited. DNA cytosine methylation, a conserved epigenetic modification in eukaryotes [2], occurs in plants within three sequence contexts: CG, CHG, and CHH (H represents A, T, or C) [3]. DNA methylation is known to be involved in regulating gene expression, suppressing transposable element activity, and stabilizing the genome [3]. Long-term changes in the methylome are mainly shaped by genomic alterations such as the proliferation of transposable elements and mutations affecting DNA methylation pathways, whereas heritable stochastic epimutations that occur independent of genomic background drive short-term changes [4].

The unique characteristics of epimutations make them appealing markers for evolutionary investigations. In plants, CG methylation is maintained during cell division by METHYLTRANSFERASE 1 (MET1) [3]. Spontaneous stochastic loss or gain of methylation at CG sites can be transmitted stably across multiple generations [5]. Van der Graaf *et al*. have

estimated the forward (gain of DNA methylation) and backward (loss of methylation) epimutation rates per CG site in the model plant *Arabidopsis thaliana* to be approximately 2.56×10^{-4} and 6.3×10^{-4}, respectively, per generation per haploid methylome [6]. These rates are several orders of magnitude higher than the mutation rate of 7×10^{-9} base substitutions per site per generation in the same population [7]. The increased rates of change at the epigenetic level may facilitate evolutionary inference over recent timescales, where the utility of slower-evolving genetic markers is limited [8, 9].

As transiently induced epimutations can be mistaken for stably inherited markers, a recent study exposed genetically identical *A. thaliana* to various environmental conditions to identify genomic regions—termed clocklike regions—where CG divergence rates are independent of environmental stimuli [10]. While these regions are valuable for evolutionary studies, they are challenging to identify in most species. Therefore, the same study observed that gene body methylated regions (gbM) exhibit a pattern similar to that of CG divergence as the clocklike regions [10]. gbM refers to a subclass of genes that display elevated CG methylation. They are mostly found in moderately and constitutively expressed housekeeping genes [11]. As they can be identified based on the distribution of methylation in the gene bodies, their use may facilitate the investigation of evolutionary processes in nonmodel plant species.

Previous attempts to explore how methylation variants evolve have either neglected demographic (neutral) models, mixed environmentally sensitive and insensitive methylation sites, or did not adequately assess the interdependence of genetic and methylation variants. Charlesworth and Jain's analytic framework for hypermutable sites [12] has been applied to create a methylation site frequency spectrum (mSFS) [4]. Previous studies generated mSFS for genic CG sites and the methylation states of gbM genes, finding either no selection or only weak selection acting on the methylome in *A. thaliana* [4, 13] and maize [14]. However, these studies do not incorporate an explicit demographic model while assessing selection. Another study found that positive selection acted on methylation variants linked to metabolite regulation in *A. thaliana*, even after accounting for demographic histories [15]. Other studies in *Brachypodium distachyon* [16], *Fragaria vesca* [17, 18], lettuce [19], and grapevines [20] have compared evolutionary inferences using genetic and methylation variants by analyzing all CG sites as the environmental sensitivity of individual sites is unknown in most plant species. Recently, Sellinger *et al*. utilized methylation as a secondary marker alongside genetic variants to enhance demographic inference over short and long timescales [8]. The conceptual novelty was to consider that the underlying genealogy at a locus (*i.e.* the coalescent tree tracing back the ancestry of the sample back in time) is common for SNPs and SMPs (under a neutral model of evolution). This represents the first population genomics inference model combining DNA mutations, epimutations, and recombination along the genome, and the basis for developing a Sequential Markovian Coalescent method (SMCm). However, that study exhibits a fair amount of noise in the inference of past evolutionary history when it was applied to methylome data from *A. thaliana* [8]. These inaccuracies are due to a lack of knowledge about the optimal SMP markers to use, the relevant genomic context, the relationship between SNPs and SMPs, and the SMCm method being very sensitive and not accounting well for methylated region effects (so-called differentially methylated regions or DMRs). A related question is whether DMRs can be used as well for evolutionary inferences.

Given these limitations, our goal was to assess how well methylation variants reveal known evolutionary patterns inferred using genetic variants. We chose *A. thaliana* and *B. distachyon*

for our analyses because their evolutionary histories are well-described [16, 21]. We found that the capacity of methylation variants to differentiate populations is contingent on when populations diverge. We also show that CG sites across the methylome capture within- and between-population patterns comparable to environmentally insensitive sites (clocklike regions and gbM) [10] in samples grown in a common garden. We found weak or no local correlations between genetic and methylation variants, indicating they reflect different aspects of population evolutionary histories. Additionally, the linkage between neighboring methylation variants was minimal. Conventional genome-based demographic inference methods, such as the Sequential Markovian Coalescent [22], show limitations in using methylation variants due to high sensitivity to violations of the Markovian properties of the underlying coalescent model along the genome [8]. Moreover, since such methods require knowing the variant accumulation rates and distributions—often unknown for methylation in many species—we developed a deep learning method that uses methylation as a neutral marker to estimate past demography. Unlike conventional convolutional networks, we use (to our knowledge) the first Encoder-Only Transformer architecture [23-25] method, focusing on the placement of mutations on a hidden ancestral recombination graph. We track estimates of coalescent events in a time-discretized transition matrix by using heterogeneous sites per genomic window as input, thereby applying a coarse-grained approach, similar to tsABC [22]. By aligning training and real data distributions, we show that rapidly changing methylation variants can reveal recent demographic shifts in selfing and clonal species.

Results

CG methylation variants have variable performance in identifying population structure

We analyzed genetic and methylation diversity using 63 samples each from the Central European (CEU) and the Iberian non-relicts (IBnr) *A. thaliana* populations (Table S1). These populations were used by Strütt *et al*. [22]. Across the two populations, 31-33% of CG methylation sites in clocklike regions were polymorphic (Figure S1A), which is similar to a previous estimate for CG sites in *A. thaliana* mutation accumulation lines [26]. In contrast, 1.6-1.7% of coding sites and 2.7-2.8% of non-coding sites were polymorphic (Figure S1A). We compared our results from *A. thaliana* with those from 29 samples from the clonally propagating grass *B. distachyon*. Wilson *et al*. used SNPs to identify clonal groups among Turkish *B. distachyon*, with samples within each group showing high levels of genetic similarity [27]. Eichten *et al*. identified high levels of consistency between the known genetic relationships for the clonal groups and the extent of CG divergence [16]. Here, we restrict our focus to SMPs and DMRs occurring within gbM regions and two clonal groups of *B. distachyon* with average bisulfite sequencing depths greater than 10 [16] (Group 1: 22 samples, Group 6: 7 samples) (Table S1). Between 4-7% of CG methylation sites within gbM regions were polymorphic in *B. distachyon* (Figure S1B). The median length of DMRs in *B. distachyon* was 107 bp (1st percentile: 44 bp, 99% percentile: 717 bp) which is smaller than the median length in *A. thaliana* DMRs, which was 181 bp (1st percentile: 78 bp, 99% percentile: 1125 bp) (Figure S2).

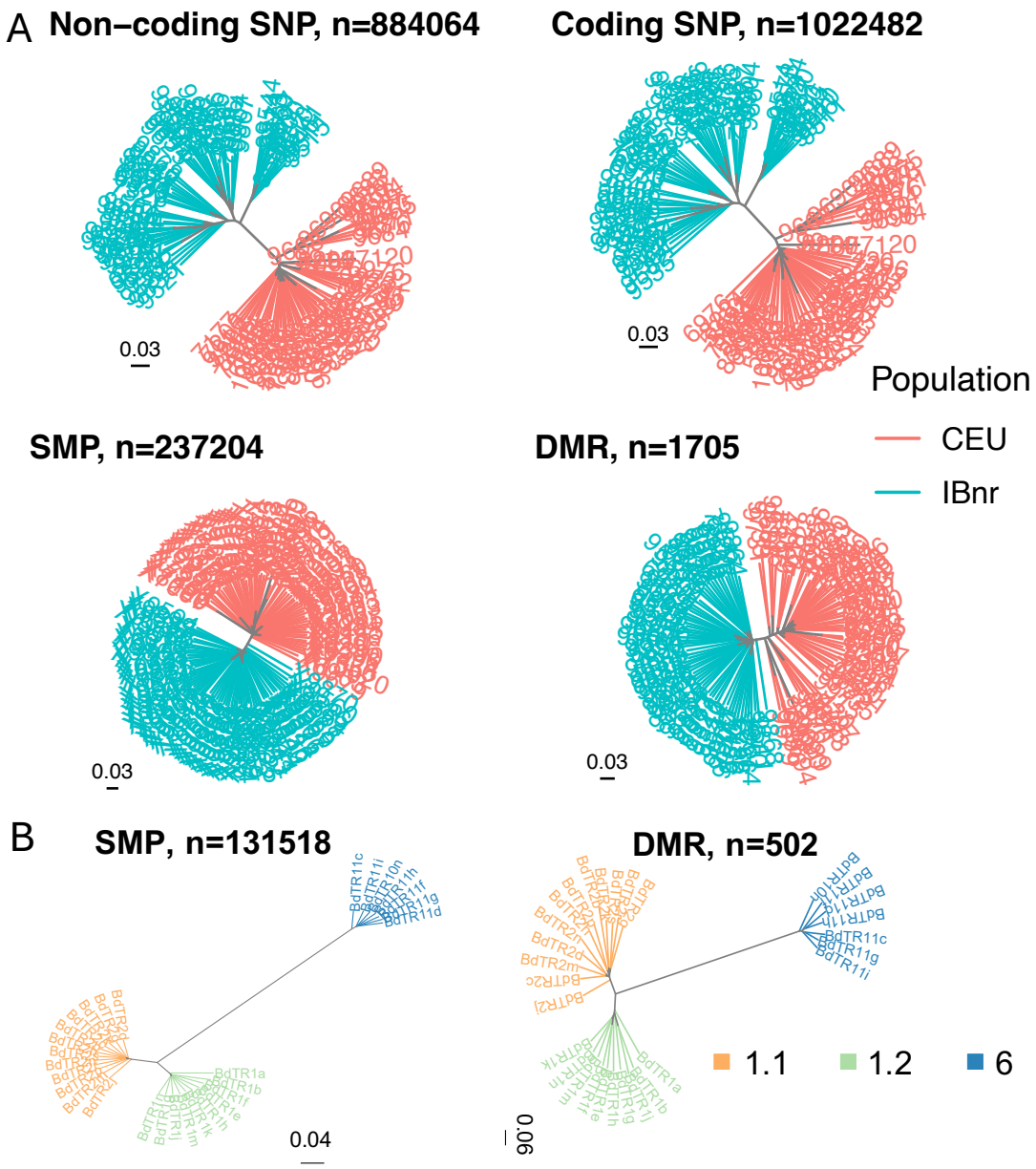

Figure 1. Methylomes outperform genetic variants in detecting population subdivision in *Brachypodium distachyon* but not *Arabidopsis thaliana*. (**A**) Neighbor-joining trees were generated using p-distance matrices for non-coding single nucleotide polymorphisms (SNP), coding SNPs, clocklike region single methylation polymorphisms (SMP), and clocklike region differentially methylated regions (DMR) for Central European (CEU) and the Iberian non-relicts (IBnr) *A. thaliana*. (**B**) Neighbor-joining trees generated using p-distance matrices estimated for SMP and DMR in gbM regions of Turkish *Brachypodium distachyon* clonal groups.

We estimated the proportion of sites at which two samples were different (p-distance) and clustering to uncover how the net total of genetic and methylation variation differed between samples from the CEU and IBnr populations. In this approach, a clear separation between the CEU and IBnr populations was apparent with both coding and non-coding single nucleotide polymorphisms (SNP) (Figure 1A). Although single methylation polymorphisms (SMP) and differentially methylated regions (DMR) largely separated the two populations, some samples were separated by only minimal distances (Figure 1A). Clustering indicated nine or more

groups within the SNP data (Figure S3A-B), whereas only two groups were apparent in the SMP and DMR datasets (Figure S3C-D).

Furthermore, we performed discriminant analysis of principal components (DAPC) analyses to identify how correlated changes in subsets of genetic and methylation variants can be used to discriminate the two populations. DAPC indicated greater discriminatory power with coding and non-coding SNPs compared to SMPs and DMRs (Figure S3E). Although p-distance matrix-based comparisons could have been influenced by the greater number of SNPs (Figure 1A), marker numbers were comparable for the DAPC, and yet SMPs and DMRs showed poor discriminatory capability (Figure S3E). SMPs and DMRs that occur within gbM genes have comparable performance to those occurring within clocklike regions in discriminating the two populations (Figure S4). High methylation rates in *A. thaliana* may have caused multiple changes at the same site (so-called homoplasy accounted for as a finite site model [4, 12, 28]), resulting in the underestimation of the true levels of divergence. Indeed, the divergence between IBnr and CEU populations is estimated to have occurred after the last glacial maximum ca. 10,000-20,000 years ago [29-31].

We then performed a similar analysis using the *B. distachyon* accessions. In contrast to the clonal grouping based on the genetic profiles [27], SMPs and DMRs were capable of additionally distinguishing the two subgroups of Group 1 (Figures 1B, S5). The trees estimated using the p-distance matrices were shorter for SMPs compared to DMRs (Figure 1B), possibly due to increased error in this dataset because of the shallow bisulfite sequencing depths. However, both trees were longer compared to those in *A. thaliana* (Figure S4A) indicating greater methylome differentiation in *B. distachyon*. Both our results using gbM regions and the results from Eichten *et al*. that use the entire collection of CG sites [16] reveal a difference in methylation profiles between the two subgroups of Group 1. We obtained phenotype data for the two subgroups [32] and found significant differences in the longitude (Figure S6), elevation of collection site, seed lengths, widths, yield production, and germination percentages (Figure S7). Our findings suggest that methylation variants can reveal subpopulation structure in recent timescales, where there has been little time for genetic differences and multiple methylation changes to accumulate.

Genomic contexts have little influence on the shape of SMP distributions

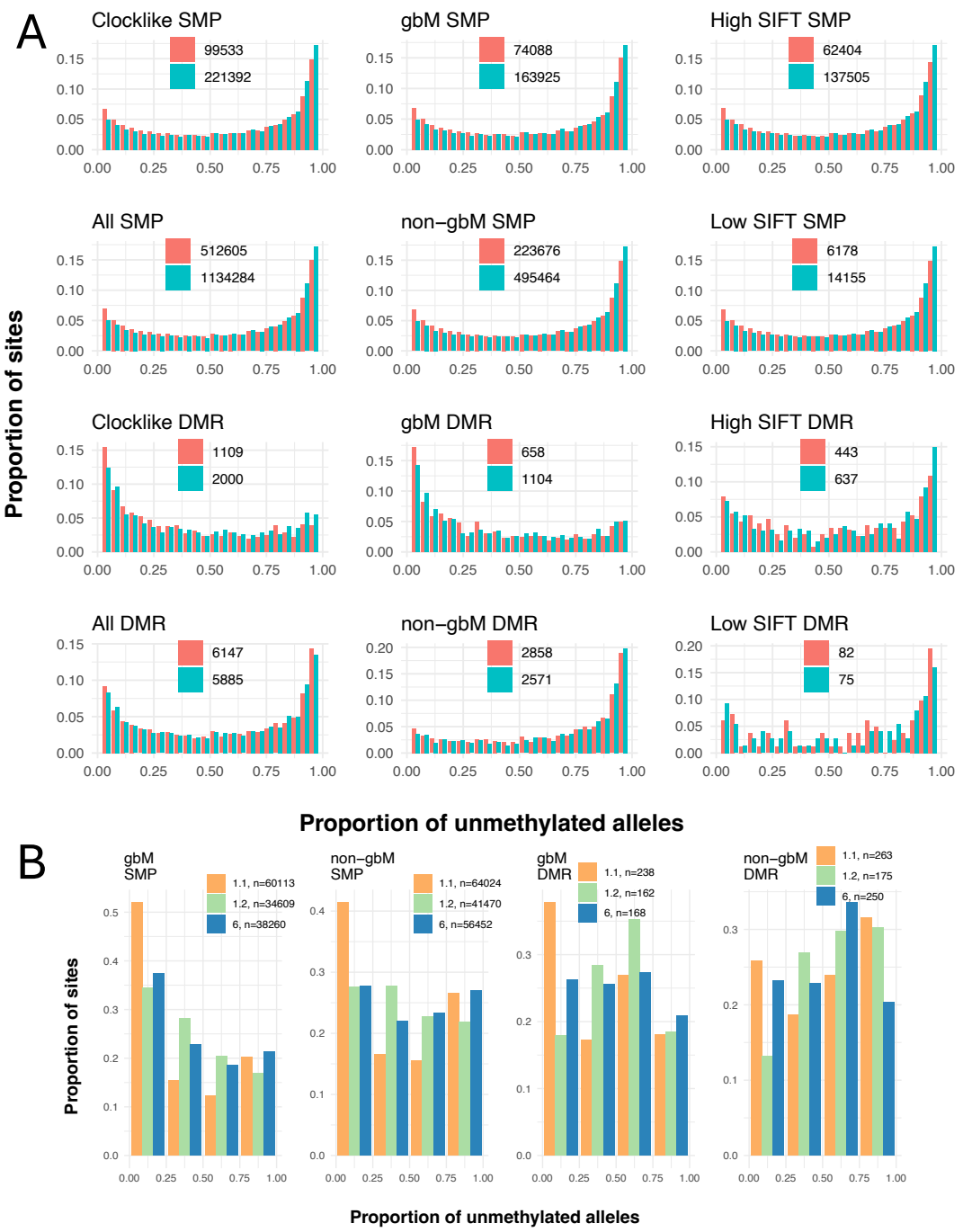

Figure 2. Methylation site frequency spectra generated using single methylation polymorphisms were similar regardless of genomic contexts. Methylation site frequency spectra (mSFS) generated using differentially methylated regions (DMR) exhibited greater variability. (**A**) mSFS for the Central European (CEU, Salmon) and the Iberian non-relicts (IBnr, Cyan) *Arabidopsis thaliana* populations using CG sites that occur in clocklike, gene body methylated (gbM), non-gbM, and low and high Sorting Intolerant from Tolerant (SIFT) score regions. Only sites with minor allele frequencies (MAF) >0.02 were used to generate the mSFS. (**B**) mSFS for *Brachypodium distachyon* clonal groups 1.1, 1.2, and 6 for single methylation polymorphisms (SMP) and DMRs occurring within gbM and non-gbM regions. Only sites with minor allele frequencies (MAF) >0.04 were used to generate the mSFS. For (A) and (B), the numbers in each panel indicate the number of sites used for the analyses.

We sought to understand how genomic contexts influenced the methylation site frequency spectra (mSFS) in *A. thaliana*. The mSFS reveals how methylation rates, demographic changes, and natural selection influence changes in the methylome over time [4, 12]. We found a greater frequency of unmethylated alleles at most polymorphic methylation sites for the CEU and IBnr populations, but CEU consistently had higher frequencies of unmethylated alleles (Figure 2A). The shapes of the mSFS for SMPs are consistent with previous findings that the backward methylation rate promoting the change from methylated to unmethylated is higher than the forward rate [6] and are consistent with previous mSFS generated by combining *A. thaliana* samples from extended geographic regions [4, 13]. The shape of the mSFS for SMPs in clocklike and gbM regions were similar, supporting their interchangeability. We also studied the mSFS for intervals containing genes with low and high Sorting Intolerant From Tolerant (SIFT) scores, which are positively correlated with the level of evolutionary constraint that genes are under [33]. Our results indicate that the level of gene constraint has little impact on the mSFS shape (Figure 2A).

When compared to SMPs, the DMR mSFS for *A. thaliana* were more dependent on the underlying set of intervals used to generate them. For DMRs that occur in intervals containing non-gbM genes, and high and low SIFT-score genes, more samples have the unmethylated state reflecting the underlying SMP distributions (Figure 2A). This is also true when DMRs across the entire genome are analyzed (Figure 2A). In contrast, for DMRs in clocklike regions and gbM genes, more samples exhibit the methylated state (Figure 2A). The mSFS shapes for DMRs show greater variability compared to those for SMPs, which may reflect underlying differences in their molecular basis.

In contrast to the mSFS for SMPs in *A. thaliana* (Figure 2A), the mSFS for all three clonal groups of *B. distachyon* indicated an excess of methylated alleles within gbM and non-gbM SMPs, although it was less pronounced for the latter (Figure 2B). There was less consistency in the shapes of the mSFS for DMRs in gbM and non-gbM regions (Figure 2B), similar to our observation when *A. thaliana* DMRs were analyzed (Figure 2A).

Methylation and genetic variants show unique evolutionary signatures in *A. thaliana*

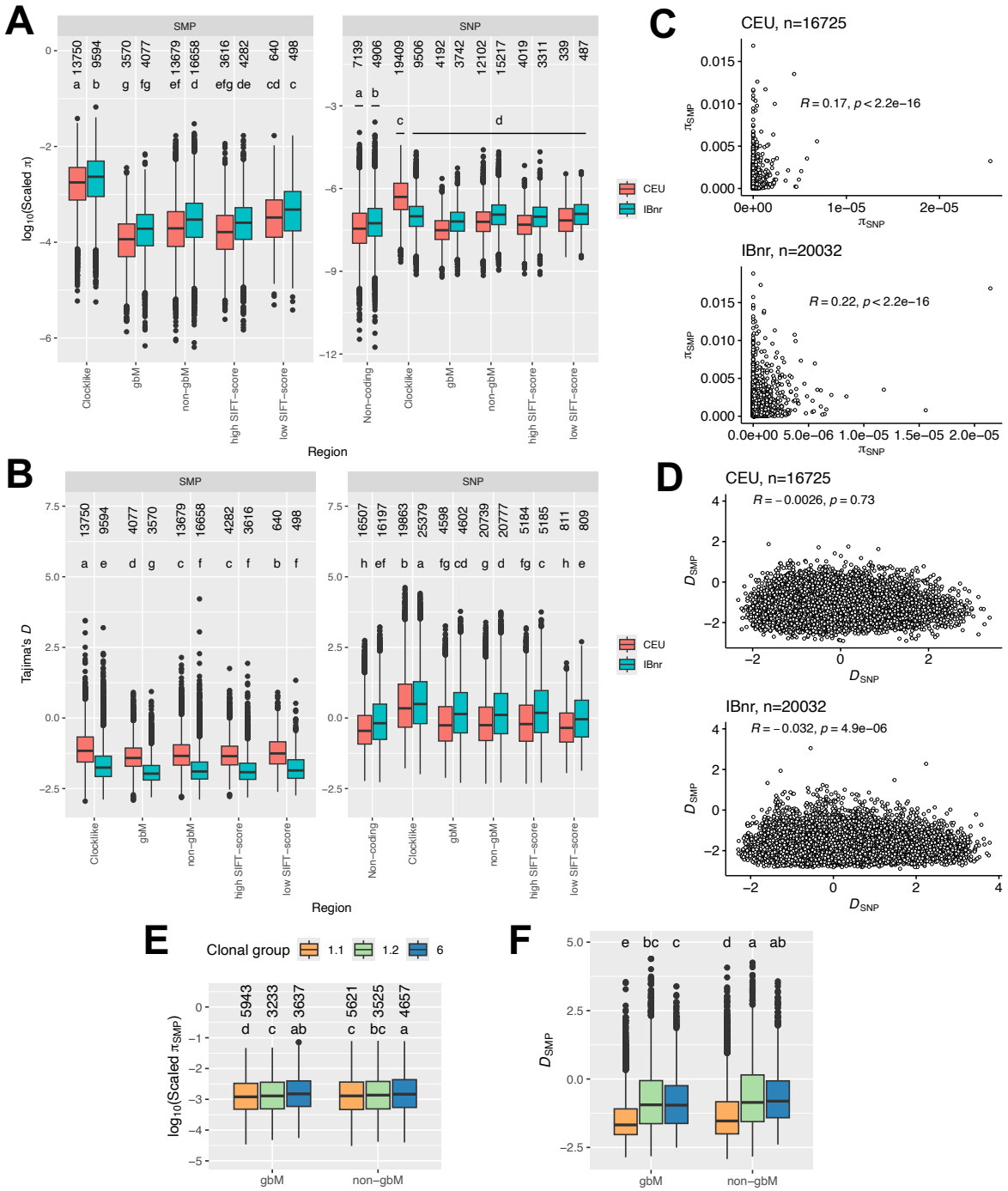

Figure 3. Unique evolutionary signatures in genomes and methylomes. (**A**) Estimates of pairwise nucleotide diversity (π) scaled by the number of cytosines per region for single methylation polymorphisms (π_{SMP}) and the number of bases for single nucleotide polymorphisms (π_{SNP}) for the Central European (CEU) and the Iberian non-relicts (IBnr) *Arabidopsis thaliana* populations. (**B**) Estimates of Tajima's *D* per region for CEU and IBnr. Comparisons of (**C**) π_{SNP} versus π_{SMP} and (**D**) Tajima's *D* for SNPs (D_{SNP}) and SMPs (D_{SMP}) per gene for CEU and IBnr. The numbers in the titles of each panel indicate the number of sites used for the analyses. Correlation tests were conducted to assess the associations between π and Tajima's *D*. (**E**) π_{SMP} scaled by the number of cytosines per region for *Brachypodium distachyon* clonal groups 1.1, 1.2, and 6. (**F**) D_{SMP} per region for *B. distachyon*. For (A), (B), (E), and (F), the numbers above the boxplots indicate the number of intervals.

Letters indicate group-wise differences in means following Tukey's honest significant difference tests. For (F), the color scheme and number of sites are the same as in (E).

We next sought to understand how genetic and methylation polymorphisms are distributed within *A. thaliana* populations. We first calculated the pairwise nucleotide diversity (π) for SMPs (π_{SMP}) and SNPs (π_{SNP}) from noncoding, clocklike, gbM, non-gbM, and high and low SIFT genic regions. We scaled estimates by the total number of sites (polymorphic and monomorphic) to ensure that the results were comparable (that is per site estimates). π_{SMP} is three orders of magnitude higher than π_{SNP} (Figure 3A), likely driven by the high rates of change in methylomes [6]. π_{SNP} in the CEU population and π_{SMP} in CEU and IBnr populations for clocklike regions were significantly higher than in other regions (Figure 3A), possibly indicating better detection of environmental insensitivity with more polymorphic sites. π_{SMP} was higher for gbM and high-SIFT genes than at non-gbM and low-SIFT, respectively, but no such difference was seen for π_{SNP} (Figure 3A), indicating that constraint at the genic level has a minor effect on the levels of local methylation diversity.

We then compared how selective and demographic processes influence genomes and methylomes by estimating Tajima's *D* for CEU and IBnr. For methylomes, we used a modified Tajima's *D* using the D^m [28] that accounts for the hypermutability of the methylation sites. D_{SNP} was slightly negative for noncoding sites in the CEU and IBnr populations (Figure 3B), reflecting the known history of population expansions [22]. The higher D_{SNP} for IBnr (Figure 3B) is consistent with larger effective population sizes over recent timescales [22]. In contrast, IBnr had significantly lower D_{SMP} compared to CEU (Figure 3B). This could be due to variation in methylation rates, purifying selection, or demographic and selective (purifying selection) histories. Therefore, we used SMCm [8], which employs fixed mutation rates and compares the distributions of mutation and methylation markers, to estimate the forward and backward methylation rates for the IBnr and CEU populations. Methylation rates were only slightly higher in CEU but the ratio of backward to forward rates was similar in both populations (Figure S8). Hypermutability should result in greater frequencies of intermediate variants [12], but we did not observe such a pattern (Figure 2A). Tajima's *D* is based on a comparison of the number of segregating sites versus π. The disparity in D_{SMP} appears to be driven primarily by the variation in the number of segregating sites per region (Figure S9). This difference suggests that the demographic histories of the two populations may be slightly different.

We then estimated Tajima's *D* in different genomic contexts for CEU and IBnr. Clocklike regions had significantly higher D_{SNP} and D_{SMP} compared to other regions (Figure 3B). The D^m [28] method assumes that forward and backward methylation rates are equivalent. Although the ratio of the backward to forward methylation rates in clocklike regions is 1.36, the ratios are 2.11 and 2.84 for gbM and non-clocklike regions, respectively [10], indicating the assumption of D^m is less strongly violated for clocklike regions. D_{SMP} for non-gbM regions was slightly but significantly higher compared to the estimates for gbM regions (Figure 3B), suggesting that ignoring the differences between regions can introduce minimal noise when inferring evolutionary signatures from methylomes. The differences in D_{SMP} between high-SIFT vs low-SIFT were not consistent in both populations (Figure 3B). There was only a weak positive correlation for π (Figure 3C) generated for SNPs vs SMPs that occur in coding regions and no significant correlation for Tajima's *D* (Figure 3D).

We compared π and Tajima's *D* for *B. distachyon* SMPs generated using SNPs for the Anatolian and Turkish lineage (B_East) [34, 35] to which all three clonal groups belong. Scaled π for SMPs was larger for the *B. distachyon* clonal groups compared to the *A. thaliana* populations, but there were no consistent differences between gbM and non-gbM regions or among clonal groups (Figure 3A, 3E). Similar to *A. thaliana* (Figure 3A), scaled π for SMPs in all three clonal groups was three orders of magnitude larger than the median SNP-based scaled π (~1.8×10^{-6}) for the B_East lineage [34, 35] (Figure 3E). Similar to *A. thaliana* (Figure 3B), D_{SMP} in the three clonal groups was negative. D_{SMP} was slightly but significantly larger in non-gbM regions compared to gbM regions (Figure 3F). These negative values contrast with the positive D_{SNP} for genetic variants in the B_East lineage [34]. Unlike in *A. thaliana* (Figure S9), this contrast cannot be attributed primarily to the differences in the number of segregating sites (Figure S10), meaning that it cannot be explained solely by the extent of genotyping. D_{SMP} was lower for Group 1.1 compared to the other two groups (Figure 3F), which is consistent with the higher frequency of rare alleles in this group (Figure 2B).

Linkage disequilibrium quickly erodes for methylation variants

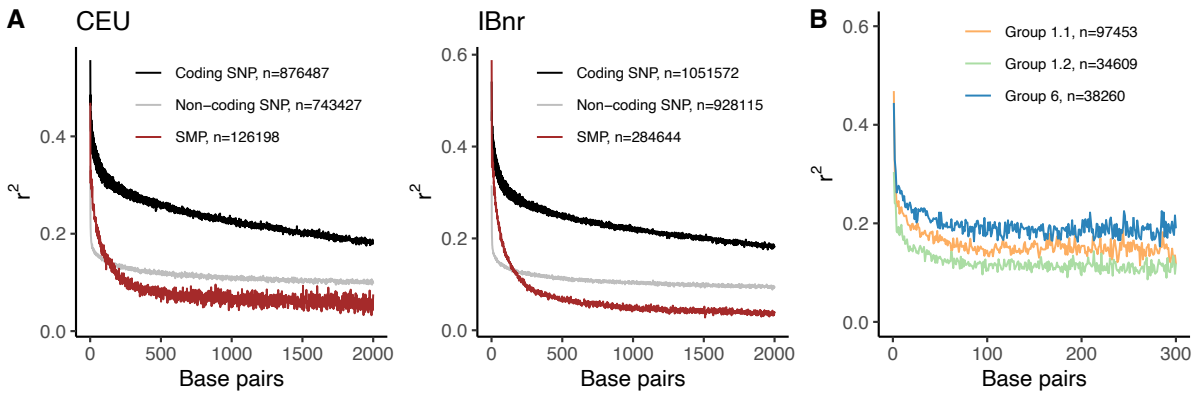

Figure 4. Weak linkage disequilibrium (LD) between neighboring single methylation polymorphisms (SMP). (**A**) Pairwise r^2 estimated using coding single nucleotide polymorphisms (SNP), non-coding SNPs, and SMPs for Central European (CEU) and Iberian non-relicts (IBnr) *Arabidopsis thaliana* populations. The number of sites used for comparisons is indicated within the plots. (**B**) Pairwise r^2 was estimated using SMPs for the clonal groups. For (A) and (B), the number of sites used for comparisons is indicated within the plots.

An important consideration for population genetic analysis is the presence of sites with independent evolutionary histories [8]. Therefore, we assessed the levels of pairwise linkage disequilibrium (LD) to identify the number of independent polymorphic SNPs and SMPs. Pairwise LD decay was stronger for noncoding SNPs compared to coding SNPs in *A. thaliana* (Figure 4A). However, LD decay was much stronger for SMPs than for SNPs with r^2 dropping below 0.1 within the first 250 bp (Figure 4A). LD decays to 0.1-0.25 for SMPs sites within gbM regions for *B. distachyon* within the first 100 bp but remains constant thereafter (Figure 4B) and is stronger compared to LD in *A. thaliana* over large distances (Figure 4A).

Demographic inference using genetic and methylation variants

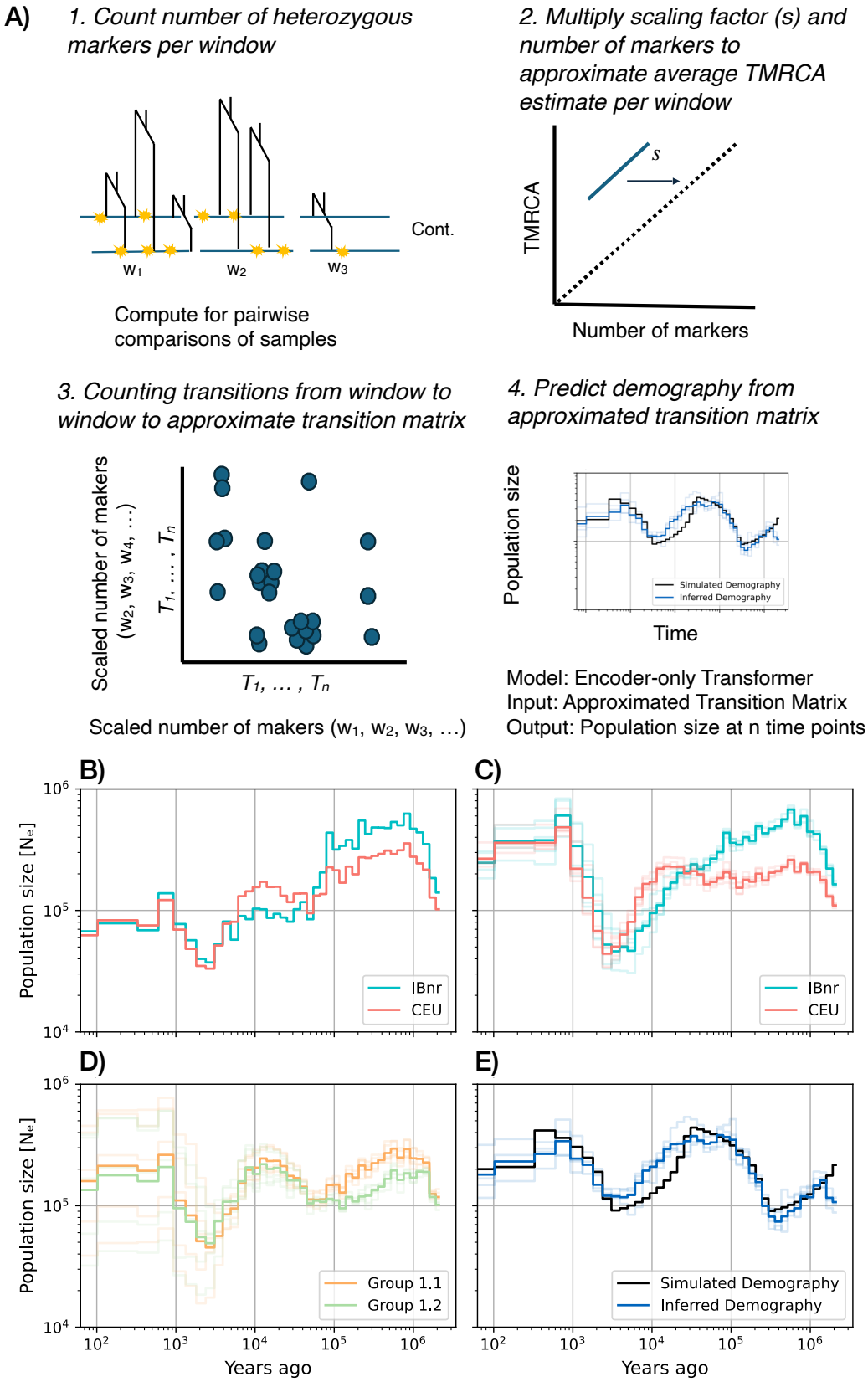

Figure 5. Demographic inference using mutation or methylation markers. **(A)** Schematic of the approximated transition matrix inference framework. The genome is first divided into 20 Kbp bins, and heterozygous markers are counted for all pairwise comparisons in a sample of

10 individuals. These counts are then scaled by an empirically chosen factor to ensure sufficient values in the approximated transition matrix by aligning the number of markers per window to the TMRCA in that window. The framework constructs this matrix by concatenating all pairwise comparisons into a long sequence and applying a sliding window of size 2 and step size 1. Each pair of consecutive values in this window represents coordinates in the transition matrix, which is incremented at those positions. This process results in an unnormalized transition matrix that captures the genetic relationships between samples. Finally, an encoder-only transformer neural network takes this noisy matrix as input, implicitly learning the fixed scaling factor, and predicts population sizes at various time points. Inference using over 42 time steps using (**B**) single nucleotide polymorphisms (SNP) and (**C**) single methylation polymorphisms (SMP) for the central European (CEU) and the Iberian non-relicts (IBnr) *Arabidopsis thaliana* populations. (**D**) Demographic inference using SMP for Groups 1.1 and 1.2 of *Brachypodium distachyon*. SMP inference was done using sequences from all five chromosomes (light lines, average in bold), while SNP inference is based on chromosome one. (**E**) Inference of simulated sawtooth demography using five replicates, 4Mbp segments, and 10 samples.

We developed a novel inference method that focuses on large genome blocks (20 Kbp windows) likely spanning multiple coalescent genealogies (Figure 5A), to estimate the times to the most recent common ancestor (TMRCA). Our method assumes that the number of heterozygous markers per window is correlated with the average TMRCA. Essentially, longer coalescent times allow more mutations and epimutations to accumulate. To mitigate the possibility that older genealogies (long TMRCA) can be unexpectedly short due to immediate recombination, our approach uses 20 Kbp windows to average out genealogies. We apply a scaling to the number of markers to estimate the TMRCA (Figure S11). Although the chosen scaling factor will not perfectly offset all variable demographic samples—given that we trained using a fixed recombination rate and a fixed window size—offsets are minimal and addressed during training. An encoder-only transformer neural network processes this data, learns the scaling factor, and predicts population sizes over time. This is done by summarizing the TMRCA of two consecutive windows in the empirical transition matrix. The differences between each pair of samples are used to infer their TMRCA. The TMRCA values for all pairs across multiple intervals (windows) in the genome are then represented as a transition matrix. This method combines raw genotype data restructuring with deep learning to infer demographic history.

For deep learning inference, the distributions of heterozygous markers in training and real data must overlap to ensure in-distribution inference. For the training data, we collapsed each training transition matrix into a single scalar by summing the flattened matrix and plotted the distribution of unnormalized transition matrices (Figure S12A). To ensure the real data *from A. thaliana* and *B. distachyon* matched the training data, we analyzed 20 randomly sampled genomic regions. For the SNP data, we selected start positions uniformly between 0-28 Mbp, whereas for the methylation data, the start positions were chosen uniformly between 0-0.5 Mbp. For both data types, we determined each region's length by sampling from a uniform distribution between 1-3 Mbp. Summing the resulting transition matrices for the IBnr, CEU, Group 1.1, and Group 1.2 regions (Figure S12B-D) produced values within the training data range (Figure S12A). Despite the differential properties of SMPs and SNPs, the distribution of

heterozygous markers per 20 Kbp window was similar across populations (Figure S12B-C), supporting the use of a common training dataset for deep learning.

We applied our coarse−grained inference approach to *A. thaliana* and *B. distachyon*. We generated approximated transition matrices from 10 samples and 4 Mbp sequence lengths. The 45 pairwise comparisons totaling 180 Mbp with 20 Kbp windows yielded 9,000 pairs of TMRCA values per matrix. We detected a population bottleneck ~100,000 years ago (100 kya) in CEU and IBnr using SNPs (Figure 5B), which was consistent with the findings of a previous study [22]. This signal was not clearly detected with SMPs (Figure 5C), indicating their limited use for ancient events. Both SNPs and SMPs indicated another bottleneck between 1-10 kya (Figure 5B-C), but SMPs also revealed a subsequent population expansion (Figure 5C) undetected by SNPs (Figure 5B). The population contraction began earlier in IBnr, followed by a slightly stronger expansion compared to CEU (Figure 5C), which may explain its more negative Tajima's *D* estimate for methylation variants (Figure 3B). Using SMPs, we recovered two bottlenecks (~100 kya and between 1-10 kya) in *B. distachyon* (Figure 5D) that were previously reported using SNPs for samples combined from multiple clonal groups in this species [34]. The steeper expansion of IBnr compared to CEU (*A. thaliana*) and Group 1.1 compared to Group 1.2 (*B. distachyon*) in the last 10,000 years likely led to the observed excess of rare alleles (Figures 2B, 3F).

We validated our demographic inference with two approaches. First, we simulated the sawtooth demography observed in both species, confirming that our deep learning method accurately detected multiple bottleneck and expansion events (Figure 5E). Second, we simulated methylation markers with high rates of change (10^{-4}, 10^{-5}, or 10^{-6} per site per generation) and backward rates 2.5 times larger than forward rates, based on the estimates provided in [6], to test whether deviations from the infinite sites model affect inference. We found that our method accurately recovered demographic histories beyond 0.6 kya and before 5, 22, and 442 kya for methylation rates of 10^{-4}, 10^{-5}, and 10^{-6} per site per generation, respectively (Figure S13). Conservatively, this suggests that inferences 20 kya and beyond are unlikely to be accurate for CEU and IBnr due to their estimated rates of methylation change being around 10^{-5} per site per generation (Figure S8). Note that the demographic estimates with SNPs are likely accurate for older events (>20kya). Nonetheless, we are confidently estimating recent population expansion with SMPs (Figure 5C), highlighting the potential of methylation markers to reveal insights beyond what genetic variants alone can provide.

Discussion

We observed contrasting patterns in the capacity of methylation variants to distinguish population subdivisions in *A. thaliana* and *B. distachyon*. In *A. thaliana*, the infinite sites model is likely violated for methylation, resulting in the erosion of methylome differentiation likely due to the older divergence between populations. In *B. distachyon*, samples from different altitudes with varying seed morphologies exhibited distinct methylation profiles in a common garden, despite their genetic similarity. While SMP distributions are similar across different genomic contexts, DMRs are more context dependent, making the former preferable for population-scale analyses. We found weak negative correlations between π and no local correlations for Tajima's *D* between genetic variants and methylation variants in *A. thaliana*, underscoring their distinct molecular and evolutionary properties, similar to the conclusions reached by Schmitz

et al. (2013) [36]. We found slightly stronger linkage between neighboring SMPs compared to a previous estimate in wild *A. thaliana* sampled across the Northern Hemisphere [36], likely due to our focus on individual populations with less genetic structure [22]. SMPs and SNPs recover recent and older demographic changes, respectively, providing a comprehensive reconstruction of population evolutionary histories.

The differing ability to detect population subdivisions in *A. thaliana* and *B. distachyon* using methylation variants could lead to differences in divergence times, methylation rates, or local adaptation. As the post-glacial re-colonization of the Iberian Peninsula and the spread northward started ~16 kya [29], multiple changes at the same methylation site (homoplasy) during this period may have caused an underestimation of the true levels of divergence in *A. thaliana*. The greater capacity to detect group subdivision is *B. distachyon* could reflect a recent separation by altitude contributing to methylome divergence. We indeed rule out the role of adaptation such as found in *F. vesca* [17], because we generated inferences using gbM regions that are highly conserved in angiosperms [11]. The observed divergence could be due to low rates of methylation change, but we could not estimate methylation rates for *B. distachyon* as our rate estimate relies on comparing the distributions of genetic and methylation variants. The most parsimonious explanation is that the two subgroups diverged recently. As the B_East lineage split from the western Mediterranean lineage 23 kya [34, 35], the split between the closely related subgroups is likely to have been far more recent, ensuring that the infinite sites model was not violated for methylation variants. Our findings underscore the necessity for methods distinguishing neutral demographic changes from selective pressures on the methylome, elucidating why SMPs behave differently across species.

Our results support the use of individual CG sites from all genomic contexts for population-scale analyses. Our between-population comparisons using CG sites in gbM regions successfully identified the known population structure within *B. distachyon*, similar to a previous study using genic CG sites [16]. Further supporting this conclusion is the consistent similarity in the mSFS shape when using SMPs, regardless of their genomic context. The *A. thaliana* [21] and *B. distachyon* [16] methylomes, generated under similar environmental conditions, ensure that sites affected solely by transient environmental effects would not be recognized as SMPs. Another study using all CG sites from lettuce methylomes, cultivated under uniform environmental conditions, also demonstrated the capacity to distinguish multiple species [19]. Some noise can be introduced by analyzing the whole methylome as non-gbM regions show a slight skew towards variants with intermediate frequencies, but this is likely to be offset by the increased number of sites for inference. The distribution of DMR states is tied to the genomic context. gbM results from the combined activity of MET1, which is enhanced by nearby existing methylation, as well as the histone variant H2A.Z and the DNA demethylase ROS1 [37]. The molecular basis underlying the clock-like, non-gbM regions, as well as regions with high and low SIFT scores, remains unknown. However, given that our mSFS for these regions differ, this suggests potential for differences in their molecular basis. Studying DMRs can nevertheless be valuable for imputing methylation states when sequencing depths are low [38], while the impact of such imputation on population genetics inference needs to be evaluated. Overall, using a restricted set of CG sites may be more appropriate when analyzing accessions grown under variable environments together or for tracking long-term methylome divergence [10].

The weak local correlations between genetic and methylation diversity and the association between gene constraint and methylation diversity in *A. thaliana* can partly be attributed to methylation affecting the mutability of neighboring sites [39-41] and transposable element polymorphisms influencing methylation variation [42, 43]. However, only about 12% of methylation quantitative trait loci (QTLs) in *A. thaliana* are within 10Kbp of the epiallelic target [36]. In maize, approximately 50% of methylation QTLs occur in *cis* [44], indicating a potentially stronger trend in this species. The higher rate of methylation changes compared to mutation accumulation [6, 7] may also explain why methylation differences persist despite genetic homogeneity. Our simulations also show that coalescent histories of markers with vastly different rates of change diverge, which can explain the general lack of correlation between π and Tajima's *D* for local genetic and methylation diversity. Our results are similar to those of previous studies in lettuce [19] and cotton [45], which also found higher levels of polymorphism in the methylome compared to the genome. The lack of correlation or similarity between the Tajima's *D* estimates from the genome and methylome in *A. thaliana* and *B. distachyon* supports the notion that the SNPs and SMPs likely reflect different parts of the common underlying coalescent genealogy at a locus: SNPs reflect the older part of the tree, whereas SMPs would reflect the topology and branch lengths in the more recent past. It is then not surprising that for recent positive selection events (selective sweeps), the signatures are similar for both types of markers, and SMPs may even be more accurate in detecting them [15]. Our study mostly focuses on local correlations in diversity, rather than genome-wide correlations, where life history traits and genetic differentiation levels also play a significant role [46-52].

Our study demonstrates that methylation variants can independently detect and allow inference of historical demographic changes. This represents a significant improvement over a previous approach [8], where methylation was used only as a secondary marker to supplement genetic variant-based inferences. Furthermore, our improved deep learning method does not require precise methylation rate parameters and is less sensitive to the presence of invariant methylation sites (compared to the SMCm in [8]). Instead, it trains on diverse scenarios to match observed variant distributions. For *B. distachyon*, methylation variants revealed expansions also detected by genetic variants [34]. In *A. thaliana*, genetic variant-based demographic inferences fail to detect recent population expansions (this study and [53]), which can be uncovered through methylation variants instead. Our simulations confirmed the utility of rapidly mutating markers for inferring demographic changes. Overall, our study highlights the value of methylation markers for understanding recent shifts in population size.

Conclusions

Our findings demonstrate that CG methylation can provide new insights into population structure in clonal and selfing species, similar to its previously demonstrated utility in *A. thaliana* from America and the clonal seagrass *Zostera marina* [10]. We demonstrate the utility of methylation variants for uncovering very recent demographic events, where SNP markers are limited due to their slower rate of change. We also illustrate its usefulness in clonal groups, where the paucity of SNPs would render such inferences either impossible or necessitate very deep and costly levels of DNA sequencing. Conversely, at older timescales, methylation markers are unlikely to be informative because the high rates of change would have resulted

in the violation of the infinite sites model. We show that SNPs and SMPs are complementary for generating demographic inferences over short and long timescales, highlighting the value of a recently developed method that leverages both sources of variation for inferring demography [8]. Our work demonstrates that methylomes contain unique evolutionary insights unattainable through studying genomic variation in isolation.

Methods

Genetic and methylation variant calling in Arabidopsis thaliana

We obtained the *A. thaliana* 1001 genomes [30] SNP calls from https://1001genomes.org/data/GMI-MPI/releases/v3.1/. We only retained biallelic SNPs for our analyses. As C/T variant calls can become confounded with methylation calls for data generated using bisulfite sequencing, we generated sample-specific genome references for mapping bisulfite sequencing reads. We substituted the genetic variant calls for each sample into the TAIR10 genome reference [54] using SAMtools [55], BCFtools consensus [55], and fastq split [56]. We obtained bisulfite sequencing data for the 1001 *A. thaliana* from [21]. We first removed poor-quality reads using Trimmomatic [57] with the following options: ILLUMINACLIP:TruSeq3-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 for single-end reads and ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:True LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:36 for paired-end reads. We subsequently mapped reads to the sample-specific reference genomes using Bismark with the aid of the HISAT2 aligner [58]. We ran Bismark deduplicate_bismark to remove optical duplicates and ran bismark_methylation_extractor to extract methylation calls. We studied cytosine sites with a coverage of at least five reads similar to what was done in [21]. To identify the methylation state of a site, we performed a binomial test with the non-conversion rate being used as the hypothesized probability of success following the method in [59]. The non-conversion rate was calculated as the proportion of methylation in the chloroplast as done in [59]. Sites were assigned as methylated if the false discovery rate corrected *P*-values following the binomial test was less than 0.01. Otherwise, they were designated as unmethylated. We used jDMR [60] to identify DMRs in the CG contexts as this method accounts for coverage when calculating likelihood scores for methylation states. For this, we called the methylation status per site using METHimpute (function: callMethylationSeparate) [38], extracted cytosine clusters in the CG context ensuring that there was a minimum of eight cytosines in each cluster (function: CfromFASTAv4 and makeReg with default options makeRegnull = c(FALSE), N.boot = 10^5, N.sim.C = "all", fp.rate = 0.01, set.tol = 0.01), and identified the methylation states of the regions in each sample (function: runMethimputeRegions), ensuring that we used a minimum of eight cytosines and a coverage of three.

63 IBnr and 67 CEU samples overlapped between the SNP and methylation variant call files. We randomly subsampled 63 samples from the CEU population to ensure the numbers were comparable to those of the IBnr population. For our analyses, we only used sites where at least 80% of the samples within each population were genotyped. The proportions of variant SNP and SMP sites were calculated by dividing the number of sites with minor allele frequencies greater than 0.02 by the total number of variant and invariant sites.

Population (epi)genetic metric estimation for Arabidopsis thaliana

To estimate π and Tajima's *D*, we created fasta files for each genomic region containing the sequences for each sample. We substituted the variants from the SNP vcf into the TAIR10 reference genome using the approach described previously for each sample, extracted the region of interest, and converted it to a fasta format. The regions could be the clocklike regions described in [10], genic regions comprising all exons, or non-coding regions comprising of all regions other than exons. Genic regions were further split into gbM genes using the list for *A. thaliana* provided by [61] and high SIFT-score (>0.75) and low SIFT-score (<0.3) genes using the list provided by [62].

We estimated π_{SNP} and D_{SMP} using pegas [63] and estimated π_{SMP} and D_{SMP} using the D^m [28]. We only analyzed intervals that had at least three polymorphic sites and sequences with lengths greater than 10bp. We calculated the number of cytosines in each interval using METHimpute [38]. To compare π and Tajima's *D* among the different genomic contexts, we built a linear model where these metrics were the dependent variable, and the concatenated population and genomic context groupings were the predictor variables. We then applied a Tukey honest significant differences test using the agricolae R package (https://myaseen208.com/agricolae//index.html) to identify significant differences between individual predictor levels. We performed a correlation test to compare the π and Tajima's *D* for SMPs and SNPs per exon.

We used VCFtools [64] to estimate the allele frequencies for SMPs and DMRs. We used PopLDdecay [65] to estimate pairwise LD using the r^2 metric analyzing all sites with a minor allele frequency greater than 0.005.

Between population comparisons for Arabidopsis thaliana

We combined the CEU and IBnr vcfs and used VCF2Dis (https://github.com/BGI-shenzhen/VCF2Dis) to estimate the p-distance matrix. We used this matrix to generate neighbor-joining trees using FastME [66] and plotted it using ggtree [67]. We also performed DAPC to discriminate samples from the CEU and IBnr populations. We thinned the dataset so that only sites that were separated by 1Kbp were sampled to reduce the computational time. We converted vcf files into a genlight format using the *genomic_converter* function in the R package radiator (https://thierrygosselin.github.io/radiator/). We used the find.clusters function in adegenet [68] to identify the number of clusters in each dataset allowing for a maximum of 20 clusters for SMPs/DMRs and 30 clusters for SNPs. As DAPC regions require fully typed data, we imputed missing values using the mean value. Then, we performed cross-validation using adegenet [68] to identify the number of informative principal components. We retained a maximum of 15 principal components, used 90% of the data for training, and calculated group-wise assignment success. We centered but did not scale variables. We used 30 replicates for each principal component. Finally, we performed DAPC using adegenet [68].

Methylation rate estimation *for Arabidopsis thaliana*

We generated the multihetsep files for SNPs for the same 5Mbp regions that were used for SMPs (described in the section immediately above). We combined the SNP and SMP multihetsep files. We randomly selected 20 samples from each population to ensure reasonable runtimes. We used the *Methylation_rate_estimation* function in SMCm to estimate

forward and backward rates [8]. We set the mutation rate as 7×10^{-9} per site per generation and the recombination rate as 3.5×10^{-8} per site per generation. We ignored regional influences on methylation rates and estimated rates under a site-only model. We used the known estimates of forward (2.56×10^{-4} per generation per haploid methylome) and backward methylation rates (6.3×10^{-4} per generation per haploid methylome) for *A. thaliana* [6] as priors.

Genetic and methylation variant calling in Brachypodium distachyon

We obtained whole genome sequencing data (29 samples) for *B. distachyon* from [27] and whole genome bisulfite sequencing data from [16]. We used Bd21-3 v1.2 genome reference and the GenBank annotation from the Joint Genome Institute (http://phytozome.jgi.doe.gov/) for variant calling. We used Trimmomatic [57] to trim reads using the paired-end options as stated above for the whole genome sequencing data and the single-end option for the whole genome bisulfite sequencing data. We mapped whole genome sequencing data reads to the reference using bwa-mem [69]. We used Picard Toolkit (https://broadinstitute.github.io/picard/) to add read groups and mark duplicate reads. We used SAMtools [55] to remove unpaired reads, remove reads with mapping quality of less than 20, and sort reads. We called genetic variants using GATK HaplotypeCaller [70]. We only retained SNPs with depths greater than five and quality scores greater than 30 using BCFtools [55]. Using this SNP variant call file, we perform methylation variant calling in the same way as was done for *A. thaliana*. We obtained the sample localities from [32] and created a map showing their occurrence with Map Maker (https://maps.co/).

Within and between population analyses in Brachypodium distachyon

We used a minor allele frequency threshold of 10% to identify polymorphic sites in each clonal group because fewer samples than *A. thaliana* were analyzed. For the analysis of SMPs and DMRs, we only used variants that occurred within CG-gbM regions identified in [71]. We estimated mSFS, LD, p-distance matrices, π, and Tajima's *D* in a similar way as we did for *A. thaliana*. We thinned the SMP dataset so that only sites that were separated by 3Kbp were sampled to reduce the computational time for the DAPC analyses.

Generating the approximated transition matrix for demographic inference

We created a training and testing dataset consisting of a combined 60×10^3 simulations (80%/20%, respectively). Of these 10×10^3 were constant demographies (*i.e.*, no demographic change) with a random population size drawn between 10×10^3 and 200×10^3 , and the other 50×10^3 simulations were random walk demographies, with time steps being defined for $i \in \{0,1,...,K-1\}$ using:

$$(\text{population_time})_i = \frac{\exp\left(\frac{i}{K-1}\cdot\ln(1+\delta t_{\max})\right)-1}{\delta}$$

where K=42 is the number of time points, $\delta = 0.001$ is the growth rate, and $t_{\max} = 2.1\times10^6$ is the maximum time.

The population size was drawn for each of the K=42 time points in the following way: an initial population size was randomly selected between a minimum (N_min = 10×10^3) and maximum (N_max = 1.2×10^6) value. For each subsequent time point, a new population size was proposed by multiplying the previous size by a random factor between 0.1 and 10. If this proposed size fell outside the allowed range (N_*min* to *N_max*), new proposals were generated until an acceptable value was found. This process was repeated for all 42 time points, resulting in a sequence of population sizes that exhibit both gradual and abrupt changes while remaining within the specified bounds. Nevertheless, a sliding window average with a size of 6 and a step size of 1 was applied to smooth extreme changes. This approach for sampling demographies was adapted from Sanchez *et al*. [72].

The obtained population sizes and times were passed onto *msprime* [73] to simulate the corresponding ancestral recombination graphs using a recombination rate of 3.4×10^{-8} per base pair and generation. For SNPs, branches were mutated at a rate of 7×10^{-9} per base pair and generation, leading to a ρ over θ ratio of 4.86.

Simulated and real data are summarized as a genotype matrix (*GM*) with dimensions $n \times m$, where $n,m \in \mathbb{N}$ represent the number of samples and mutations, respectively. Each element g_{ij} of *GM*, where $i \in \{1,...,n\}$ and $j \in \{1,...,m\}$, denotes the presence ($g_{ij} = 1$) or absence ($g_{ij} = 0$) of mutation j in sample i. We generated a corresponding vector of positions of size m for each segregating site containing floating point numbers of the location in the genome.

For each random population size parameter, we used a genome of length 4Mbp,10 samples (ploidy=1), and the *GM* and positions vector to construct a density estimate of mutations along the genome (in theory, corresponding to the respective TMRCA at that region). To ensure the presence of enough markers, we segmented the genome into 20Kbp pieces and counted the number of heterozygous markers between all pairwise $\binom{10}{2}$ combinations. The 4Mbp sequences of all the comparisons were concatenated and log-transformed. Likewise, the population time steps were also log-scaled. To make sure that the number of markers per window fell approximately into the time windows, we scaled the markers by an arbitrarily

chosen scaling factor (=$e^{8.2}$) (Figure S13). The importance of the given scaling factor is negligible, due to all the transformation steps occurring in the pre-training, and therefore any noisy representation presented to the learning algorithm will be part of the training.

Each *msprime* simulation (under a random demographic history) yielded one *approximated* transition matrix, which was obtained by counting the transitions (scaled number of heterozygous sites in window t → window t+1) from one 20 Kbp window to the next using the following algorithm: a sliding window of size 2 and step size 1 was applied along the sequence and each pair of consecutive values in this window represents coordinates in the transition matrix, which was incremented at those positions leading to a filled matrix which remained unnormalized. Lastly, the entire transition matrix was incremented by one at each entry and scaled by the natural log to squeeze all values approximately into the same range to avoid training instability.

Model description, training, and inference

The inference approach can be described as a supervised learning approach using the *approximated* transition matrices *x* with defined dimensions $R^{L=42\times K=42}$ as the input and the population size at each time step as the predicted variable. The input is a square matrix where both dimensions (K and L) are 42. The model architecture is an encoder-only transformer [74] (https://github.com/lucidrains/x-transformers). It consists of three main components: an input projection layer, a transformer encoder, and an output projection layer. It can be formalized as:

1. Input Projection: $x_{proj} = W_{in}x + b_{in}$, where $W_{in} \in R^{D_{enc}\times K_{in}}, b_{in} \in R^{D_{enc}}, x \in R^{L\times K_{in}}$

2. Transformer Encoder: $z = \text{Encoder}(x_{proj})$, where $z \in R^{L\times D_{enc}}$

   The Encoder employs a configuration with dimension D_{enc} = 512, depth N_{enc} = 12, number of attention heads H_{enc} = 10 heads, Gated Linear Unit activation in feed-forward layers [75], residual attention connections [76], and rotary positional embeddings [77].

3. Dimension Reduction: $z_{reduced} = \sum_{i=1}^{L} z_i$, where $z_{reduced} \in R^{D_{enc}}$

4. Output Projection: $y = W_{out}z_{reduced} + b_{out}$, where $W_{out} \in R^{D_{out}\times D_{enc}}, b_{out} \in R^{D_{out}}$

   In this formalization, sequence length *L* = the input dimension of the transition matrix K_{in} = 42, which represents the square input matrix of 42 x 42 dimensions, and the output dimension D_{out} = 42.

The neural network model underwent training for 50 epochs, employing mean squared error as the loss function. The training process leverages the Accelerate framework from Hugging Face to enhance performance [78]. An Accelerator object was configured with mixed precision using bfloat16 and gradient accumulation over 12 steps (with batch size 32). Throughout the training regiment, both the training and test losses were continuously monitored to evaluate model performance and assess convergence.

Generating data for demographic inference

We first generated unique variant call files for each sample from the population variant call files using BCFtools [55]. We then used the generate_multihetsep.py from MSMC-tools (https://github.com/stschiff/msmc-tools/tree/master) to combine all the samples per chromosome. For *A. thaliana* SNPs, we used the entire chromosome 1 as we needed a long contiguous fragment to generate the approximated transition matrix. For SMPs, we used 5Mbp segments in each of the five chromosomes (Table S2) as the SMP density was far higher. We used the list of genotyped cytosine positions as a genomic location mask for the SMPs. For *B. distachyon*, we included singletons to support the demographic inference given the small sample sizes in this species. The data were polarized by identifying the minor (less frequent) alleles and encoding them as derived, whereas the major alleles were encoded as ancestral. The resulting genotype matrix represents this polarized, unphased data, where each row corresponds to an individual and each column corresponds to a genomic position.

Methylation simulation model

We explored the boundaries of our course-grained demographic inference by adding a new mutation model to msprime [73]. We simulated a sawtooth demographic model with cyclical population size changes over time and a baseline effective population size of 2×10^5 individuals. Simulations were conducted on a 5Mbp sequence with a recombination rate of 3.4×10^{-8} per base pair per generation (ploidy=1), to isolate the influence of methylation dynamics.

A custom matrix mutation simulation was implemented to model methylation and demethylation processes, employing 2 allelic states: 0 (unmethylated) and 1 (methylated). The model incorporates 3 key rates: the site introduction rate (μ), methylation rate (β), and demethylation rate (δ). The methylation and demethylation rates are measured as factors of the site entry rate, where β = the methylation factor$\cdot\mu$ and δ = the demethylation factor$\cdot\mu$. The transition probabilities between the unmethylated and methylated states are determined by the ratio $\beta/(\beta+\delta)$ for methylation and $\delta/(\beta+\delta)$ for demethylation. The resulting 2×2 transition matrix encapsulates these probabilities:

$$\begin{bmatrix} 1 - [\beta/(\beta+\delta)] & \beta/(\beta+\delta) \\ \delta/(\beta+\delta) & 1 - [\delta/(\beta+\delta)] \end{bmatrix}$$

This framework allowed us to explore various methylation scenarios by adjusting β and δ relative to μ. We investigated cases where $\beta = \mu$ and $\delta = 2.5\mu$, $\beta = 10\mu$ and $\delta = 25\mu$, and $\beta = 100\mu$ and $\delta = 250\mu$, representing low, medium, and high methylation dynamics respectively, with the demethylation rate 2.5 times higher than the methylation rate. By varying the epimutation rate compared to the mutation rate, we test the robustness of our approach in correlating the number of markers in one Kbp windows with the average coalescent time. This proportionality, and the robustness of our method, is expected to decrease for higher epimutation rates, as these would violate the infinite site model for older coalescent times more strongly (due to homoplasy, [4]).

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The scripts for analyzing and plotting all the results, except for the demographic inference, have been published on GitHub (arunkumarramesh/methevolve) and archived using Zenodo (https://doi.org/10.5281/zenodo.13991812). The demographic inference method is available on GitHub (kevinkorfmann/aTMi) and has been archived using Zenodo (https://doi.org/10.5281/zenodo.13332833).

Competing interests

The authors declare that they have no competing interests.

Funding

RA was funded by the Peter and Traudl Engelhorn Foundation.

Authors' contributions

RA and AT designed the research. RA, KK, AW, BH, CG, and YW performed the research. RA, AT, KK, and AW wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Thibaut Paul Patrick Sellinger for helping with the estimation of methylation rates. This work was supported by the de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) and ELIXIR-DE (Forschungszentrum Jülich and W-de.NBI-001, W-de.NBI-004, W-de.NBI-008, W-de.NBI-010, W-de.NBI-013, W-de.NBI-014, W-de.NBI-016, W-de.NBI-022).

References

1. Richards CL, Alonso C, Becker C, Bossdorf O, Bucher E, Colome-Tatche M, Durka W, Engelhardt J, Gaspar B, Gogol-Doring A, et al: **Ecological plant epigenetics: Evidence from model and non-model species, and the way forward.** *Ecol Lett* 2017, **20:**1576-1590.
2. Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, et al: **Conservation and divergence of methylation patterning in plants and animals.** *Proc Natl Acad Sci U S A* 2010, **107:**8689-8694.
3. Zhang H, Lang Z, Zhu JK: **Dynamics and function of DNA methylation in plants.** *Nat Rev Mol Cell Biol* 2018, **19:**489-506.
4. Vidalis A, Zivkovic D, Wardenaar R, Roquis D, Tellier A, Johannes F: **Methylome evolution in plants.** *Genome Biol* 2016, **17:**264.
5. Cortijo S, Wardenaar R, Colome-Tatche M, Gilly A, Etcheverry M, Labadie K, Caillieux E, Hospital F, Aury JM, Wincker P, et al: **Mapping the epigenetic basis of complex traits.** *Science* 2014, **343:**1145-1148.
6. van der Graaf A, Wardenaar R, Neumann DA, Taudt A, Shaw RG, Jansen RC, Schmitz RJ, Colome-Tatche M, Johannes F: **Rate, spectrum, and evolutionary dynamics of spontaneous epimutations.** *Proc Natl Acad Sci U S A* 2015, **112:**6676-6681.
7. Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M: **The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana.** *Science* 2010, **327:**92-94.
8. Sellinger T, Johannes F, Tellier A: **Improved inference of population histories by integrating genomic and epigenomic data.** *eLife* 2023, **12:**RP89470.
9. Yao N, Schmitz RJ, Johannes F: **Epimutations Define a Fast-Ticking Molecular Clock in Plants.** *Trends Genet* 2021, **37:**699-710.
10. Yao N, Zhang Z, Yu L, Hazarika R, Yu C, Jang H, Smith LM, Ton J, Liu L, Stachowicz JJ, et al: **An evolutionary epigenetic clock in plants.** *Science* 2023, **381:**1440-1445.
11. Muyle AM, Seymour DK, Lv Y, Huettel B, Gaut BS: **Gene Body Methylation in Plants: Mechanisms, Functions, and Important Implications for Understanding Evolutionary Processes.** *Genome Biol Evol* 2022, **14:**evac038.
12. Charlesworth B, Jain K: **Purifying selection, drift, and reversible mutation with arbitrarily high mutation rates.** *Genetics* 2014, **198:**1587-1602.
13. Muyle A, Ross-Ibarra J, Seymour DK, Gaut BS: **Gene body methylation is under selection in Arabidopsis thaliana.** *Genetics* 2021, **218:**iyab061.
14. Xu G, Lyu J, Li Q, Liu H, Wang D, Zhang M, Springer NM, Ross-Ibarra J, Yang J: **Evolutionary and functional genomics of DNA methylation in maize domestication and improvement.** *Nat Commun* 2020, **11:**5539.
15. Shirai K, Sato MP, Nishi R, Seki M, Suzuki Y, Hanada K: **Positive selective sweeps of epigenetic mutations regulating specialized metabolites in plants.** *Genome Res* 2021, **31:**1060-1068.
16. Eichten SR, Srivastava A, Reddiex AJ, Ganguly DR, Heussler A, Streich JC, Wilson PB, Borevitz JO: **Extending the Genotype in Brachypodium by Including DNA Methylation Reveals a Joint Contribution with Genetics on Adaptive Traits.** *G3 (Bethesda)* 2020, **10:**1629-1637.

17. Le Vève A, Sammarco I, Latzel V, Lafon Placette C: **DNA methylation shows footprints of altitude selection in the clonal plant species Fragaria vesca.** *bioRxiv* 2024, **2024.03.19.585697**.
18. Sammarco I, Diez Rodriguez B, Galanti D, Nunn A, Becker C, Bossdorf O, Munzbergova Z, Latzel V: **DNA methylation in the wild: epigenetic transgenerational inheritance can mediate adaptation in clones of wild strawberry (Fragaria vesca).** *New Phytol* 2024, **241:**1621-1635.
19. Cao S, Sawettalake N, Li P, Fan S, Shen L: **DNA methylation variations underlie lettuce domestication and divergence.** *Genome Biol* 2024, **25:**158.
20. Rodriguez-Izquierdo A, Carrasco D, Anand L, Magnani R, Catarecha P, Arroyo-Garcia R, Rodriguez Lopez CM: **Epigenetic differences between wild and cultivated grapevines highlight the contribution of DNA methylation during crop domestication.** *BMC Plant Biol* 2024, **24:**504.
21. Kawakatsu T, Huang SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, Castanon R, Nery JR, Barragan C, He Y, et al: **Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions.** *Cell* 2016, **166:**492-505.
22. Strütt S, Sellinger T, Glemin S, Tellier A, Laurent S: **Joint inference of evolutionary transitions to self-fertilization and demographic history using whole-genome sequences.** *Elife* 2023, **12:**e82384.
23. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I: **Attention Is All You Need.** *arXiv* 2017, **1706.03762**.
24. Devlin J, Chang M-W, Lee K, Toutanova K: **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.** *arXiv* 2018, **1810.04805**.
25. Korfmann K, Gaggiotti OE, Fumagalli M: **Deep Learning in Population Genetics.** *Genome Biol Evol* 2023, **15**.
26. Becker C, Hagmann J, Muller J, Koenig D, Stegle O, Borgwardt K, Weigel D: **Spontaneous epigenetic variation in the Arabidopsis thaliana methylome.** *Nature* 2011, **480:**245-249.
27. Wilson PB, Streich JC, Murray KD, Eichten SR, Cheng R, Aitken NC, Spokas K, Warthmann N, Gordon SP, Accession C, et al: **Global Diversity of the Brachypodium Species Complex as a Resource for Genome-Wide Association Studies Demonstrated for Agronomic Traits in Response to Climate.** *Genetics* 2019, **211:**317-331.
28. Wang J, Fan C: **A neutrality test for detecting selection on DNA methylation using single methylation polymorphism frequency spectrum.** *Genome Biol Evol* 2014, **7:**154-171.
29. Hewitt G: **Post-glacial re-colonization of European biota.** *Biological Journal of the Linnean Society* 1999, **68:**87-112.
30. The1001GenomesConsortium: **1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana.** *Cell* 2016, **166:**481-491.
31. Sharbel TF, Haubold B, Mitchell-Olds T: **Genetic isolation by distance in Arabidopsis thaliana: biogeography and postglacial colonization of Europe.** *Mol Ecol* 2000, **9:**2109-2118.
32. Filiz E, Ozdemir BS, Budak F, Vogel JP, Tuna M, Budak H: **Molecular, morphological, and cytological analysis of diverse Brachypodium distachyon inbred lines.** *Genome* 2009, **52:**876-890.

33. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Research* 2003, **31:**3812-3814.
34. Minadakis N, Williams H, Horvath R, Caković D, Stritt C, Thieme M, Bourgeois Y, Roulin AC: **The demographic history of the wild crop relative Brachypodium distachyon is shaped by distinct past and present ecological niches.** *Peer Community Journal* 2023, **3:**e84.
35. Stritt C, Gimmi EL, Wyler M, Bakali AH, Skalska A, Hasterok R, Mur LAJ, Pecchioni N, Roulin AC: **Migration without interbreeding: Evolutionary history of a highly selfing Mediterranean grass inferred from whole genomes.** *Mol Ecol* 2022, **31:**70-85.
36. Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ, Ecker JR: **Patterns of population epigenomic diversity.** *Nature* 2013, **495:**193-198.
37. Briffa A, Hollwey E, Shahzad Z, Moore JD, Lyons DB, Howard M, Zilberman D: **Millennia-long epigenetic fluctuations generate intragenic DNA methylation variance in Arabidopsis populations.** *Cell Syst* 2023, **14:**953-967 e917.
38. Taudt A, Roquis D, Vidalis A, Wardenaar R, Johannes F, Colome-Tatche M: **METHimpute: imputation-guided construction of complete methylomes from WGBS data.** *BMC Genomics* 2018, **19:**444.
39. Xia J, Han L, Zhao Z: **Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome.** *BMC Genomics* 2012, **13:**S7.
40. Mugal CF, Ellegren H: **Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content.** *Genome Biol* 2011, **12:**R58.
41. Kusmartsev V, Drozdz M, Schuster-Bockler B, Warnecke T: **Cytosine Methylation Affects the Mutability of Neighboring Nucleotides in Germline and Soma.** *Genetics* 2020, **214:**809-823.
42. Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, Lister R: **Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation.** *Elife* 2016, **5:**e20777.
43. Shen Y, Zhang J, Liu Y, Liu S, Liu Z, Duan Z, Wang Z, Zhu B, Guo YL, Tian Z: **DNA methylation footprints during soybean domestication and improvement.** *Genome Biol* 2018, **19:**128.
44. Eichten SR, Briskine R, Song J, Li Q, Swanson-Wagner R, Hermanson PJ, Waters AJ, Starr E, West PT, Tiffin P, et al: **Epigenetic and genetic influences on DNA methylation variation in maize populations.** *Plant Cell* 2013, **25:**2783-2797.
45. Zhao T, Guan X, Hu Y, Zhang Z, Yang H, Shi X, Han J, Mei H, Wang L, Shao L, et al: **Population-wide DNA methylation polymorphisms at single-nucleotide resolution in 207 cotton accessions reveal epigenomic contributions to complex traits.** *Cell Res* 2024.
46. Avramidou EV, Korakaki E, Malliarou E, Boutsios S: **Studying the Genetic and the Epigenetic Diversity of the Endangered Species Juniperus drupacea Labill. towards Safeguarding Its Conservation in Greece.** *Forests* 2023, **14:**1271.
47. Avramidou EV, Moysiadis T, Ganopoulos I, Michailidis M, Kissoudis C, Valasiadis D, Kazantzis K, Tsaroucha E, Tsaftaris A, Molassiotis A, et al: **Phenotypic,**

Genetic, and Epigenetic Variation among Diverse Sweet Cherry Gene Pools. *Agronomy* 2021, **11:**680.

48. Wang MZ, Li HL, Li JM, Yu FH: **Correlations between genetic, epigenetic and phenotypic variation of an introduced clonal herb.** *Heredity (Edinb)* 2020, **124:**146-155.
49. Langford N, Fargeot L, Blanchet S: **Spatial covariation between genetic and epigenetic diversity in wild plant and animal populations: a meta-analysis.** *J Exp Biol* 2024, **227**.
50. Medrano M, Alonso C, Bazaga P, Lopez E, Herrera CM: **Comparative genetic and epigenetic diversity in pairs of sympatric, closely related plants with contrasting distribution ranges in south-eastern Iberian mountains.** *AoB Plants* 2020, **12:**plaa013.
51. Fargeot L, Loot G, Prunier JG, Rey O, Veyssiere C, Blanchet S: **Patterns of Epigenetic Diversity in Two Sympatric Fish Species: Genetic vs. Environmental Determinants.** *Genes (Basel)* 2021, **12:**107.
52. Lele L, Ning D, Cuiping P, Xiao G, Weihua G: **Genetic and epigenetic variations associated with adaptation to heterogeneous habitat conditions in a deciduous shrub.** *Ecol Evol* 2018, **8:**2594-2606.
53. Durvasula A, Fulgione A, Gutaker RM, Alacakaptan SI, Flood PJ, Neto C, Tsuchimatsu T, Burbano HA, Pico FX, Alonso-Blanco C, Hancock AM: **African genomes illuminate the early history and transition to selfing in Arabidopsis thaliana.** *Proc Natl Acad Sci U S A* 2017, **114:**5213-5218.
54. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al: **The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools.** *Nucleic Acids Res* 2012, **40:**D1202-1210.
55. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H: **Twelve years of SAMtools and BCFtools.** *Gigascience* 2021, **10:**giab008.
56. Nassar LR, Barber GP, Benet-Pages A, Casper J, Clawson H, Diekhans M, Fischer C, Gonzalez JN, Hinrichs AS, Lee BT, et al: **The UCSC Genome Browser database: 2023 update.** *Nucleic Acids Res* 2023, **51:**D1188-D1195.
57. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics* 2014, **30:**2114–2120.
58. Krueger F, Andrews SR: **Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.** *Bioinformatics* 2011, **27:**1571-1572.
59. 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:**523-536.
60. Hazarika RR, Serra M, Zhang Z, Zhang Y, Schmitz RJ, Johannes F: **Molecular properties of epimutation hotspots.** *Nat Plants* 2022, **8:**146-156.
61. Williams CJ, Dai D, Tran KA, Monroe JG, Williams BP: **Dynamic DNA methylation turnover in gene bodies is associated with enhanced gene expression plasticity in plants.** *Genome Biol* 2023, **24:**227.
62. Hamala T, Tiffin P: **Biased Gene Conversion Constrains Adaptation in Arabidopsis thaliana.** *Genetics* 2020, **215:**831-846.

63. Paradis E: **pegas: an R package for population genetics with an integrated-modular approach.** *Bioinformatics* 2010, **26:**419-420.
64. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al: **The variant call format and VCFtools.** *Bioinformatics* 2011, **27**.
65. Zhang C, Dong SS, Xu JY, He WM, Yang TL: **PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files.** *Bioinformatics* 2019, **35:**1786-1788.
66. Lefort V, Desper R, Gascuel O: **FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program.** *Mol Biol Evol* 2015, **32:**2798-2800.
67. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY: **ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data.** *Methods in Ecology and Evolution* 2017, **8:**28-36.
68. Jombart T: **adegenet: a R package for the multivariate analysis of genetic markers.** *Bioinformatics* 2008, **24:**1403–1405.
69. Li H: **Aligning sequence reads , clone sequences and assembly contigs with BWA-MEM.** 2013, **00:**1-3.
70. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Research* 2010, **20:**1297–1303.
71. Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, Li Q, Rohr NA, Rambani A, Burke JM, Udall JA, et al: **Widespread natural variation of DNA methylation within angiosperms.** *Genome Biol* 2016, **17:**194.
72. Sanchez T, Cury J, Charpiat G, Jay F: **Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation.** *Mol Ecol Resour* 2021, **21:**2645-2660.
73. Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, Zhu S, Eldon B, Ellerman EC, Galloway JG, et al: **Efficient ancestry and mutation simulation with msprime 1.0.** *Genetics* 2022, **220**.
74. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al: *pyTorch: An Imperative Style, High-Performance Deep Learning Library.* Curran Associates, Inc.; 2019.
75. Shazeer N: **GLU Variants Improve Transformer.** *arXiv* 2020**:**2002.05202.
76. He R, Ravula A, Kanagal B, Ainslie J: **RealFormer: Transformer Likes Residual Attention.** *arXiv* 2021**:**2012.11747.
77. Su J, Ahmed M, Lu Y, Pan S, Bo W, Liu Y: **RoFormer: Enhanced transformer with Rotary Position Embedding.** *Neurocomputing* 2024, **568:**127063.
78. Gugger S, Debut L, Wolf T, Schmid P, Mueller Z, Mangrulkar S, Sun M, Bossan B: **Accelerate: Training and inference at scale made simple, efficient and adaptable.** *https://githubcom/huggingface/accelerate* 2022.