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†], Ramesh Arunkumar [a†]

[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

[†] 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 single nucleotide polymorphisms (SNPs) for identifying population differentiation in *A. thaliana*, they can classify phenotypically divergent *B. distachyon* subgroups that are otherwise genetically indistinguishable. The site frequency spectra generated using methylation sites from varied genomic locations and evolutionary conservation exhibit an excess of rare alleles. Nucleotide diversity estimates were three orders of magnitude higher for methylation variants than for SNPs 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 100 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 SNPs. Our study demonstrates the unique evolutionary insights methylomes provide that SNPs 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 (Richards, et al. 2017). 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 (Feng, et al. 2010), occurs in plants within three sequence contexts: CG, CHG, and CHH (H represents A, T, or C) (Zhang, et al. 2018). DNA methylation is known to be involved in regulating gene expression, suppressing transposable element activity, and stabilizing the genome (Zhang, et al. 2018). 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 (Vidalis, et al. 2016).

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) (Zhang, et al. 2018). Spontaneous stochastic loss or gain of methylation at CG sites can be transmitted stably across multiple generations (Cortijo, et al. 2014). 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 (van der Graaf, et al. 2015). 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 (Ossowski, et al. 2010). 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 (Yao, et al. 2021; Sellinger, et al. 2023).

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 (Yao, et al. 2023). 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 in clocklike regions (Yao, et al. 2023). gbM refers to a subclass of genes that display elevated CG methylation. They are mostly found in moderately and constitutively expressed housekeeping genes (Muyle, et al. 2022). 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 (Charlesworth and Jain 2014) has been applied to create a methylation site frequency spectrum (mSFS) (Vidalis, et al. 2016). 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* (Vidalis, et al. 2016; Muyle, et al. 2021) and maize (Xu, et al. 2020). 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 (Shirai, et al. 2021). Other studies in *Brachypodium distachyon* (Eichten, et al. 2020), *Fragaria vesca* (Le Vève, et al. 2024; Sammarco, et al. 2024), lettuce (Cao, et al. 2024), and grapevines (Rodriguez-Izquierdo, et al. 2024) 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

113 marker alongside SNPs to enhance demographic inference over short and long timescales
114 (Sellinger, et al. 2023). The conceptual novelty was to consider that the underlying genealogy
115 at a locus (*i.e.* the coalescent tree tracing back the ancestry of the sample back in time) is
116 common for SNPs and SMPs (under a neutral model of evolution). This represents the first
117 population genomics inference model combining DNA mutations, epimutations, and
118 recombination along the genome, and the basis for developing a Sequential Markovian
119 Coalescent method (SMCm). However, that study exhibits a fair amount of noise in the
120 inference of past evolutionary history when it was applied to methylome data from *A. thaliana*
121 [8]. These inaccuracies are due to a lack of knowledge about the optimal SMP markers to use,
122 the relevant genomic context, the relationship between SNPs and SMPs, and the SMCm
123 method being very sensitive and not accounting well for methylated region effects (so-called
124 differentially methylated regions or DMRs). A related question is whether DMRs can be used
125 for evolutionary inferences as well.
126
127 Given these limitations, our goal was to assess how well methylation variants reveal known
128 evolutionary patterns inferred using SNPs. We chose *A. thaliana* and *B. distachyon* for our
129 analyses because their evolutionary histories are well-described (Kawakatsu, et al. 2016;
130 Eichten, et al. 2020). We found that the capacity of methylation variants to differentiate
131 populations is contingent on when populations diverge. We also show that CG sites across
132 the methylome capture within- and between-population patterns comparable to
133 environmentally insensitive sites (clocklike regions and gbM) (Yao, et al. 2023) in samples
134 grown in a common garden. We found weak or no local correlations between genetic and
135 methylation variants, indicating they reflect different aspects of population evolutionary
136 histories. Additionally, the linkage between neighboring methylation variants was minimal.
137 Conventional genome-based demographic inference methods, such as the Sequential
138 Markovian Coalescent (Strütt, et al. 2023), show limitations in using methylation variants due
139 to high sensitivity to violations of the Markovian properties of the underlying coalescent model
140 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 (Vaswani, et al. 2017; Devlin, et al. 2018; Korfmann, et al. 2023) 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 (Strütt, et al. 2023). 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 identify population structure in recently diverged groups

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. 2023). Across the two populations, 31-33% of CG methylation sites were polymorphic (Figure S1A), which is similar to a previous estimate for CG sites in *A. thaliana* mutation accumulation lines (Becker, et al. 2011). In contrast, 1.6-1.7% of coding sites and 2.7-2.8% of non-coding sites were polymorphic (Figure S1A). Interestingly, 80–81% of sites in clocklike regions and 60% of sites in gbM regions were polymorphic, which is significantly higher compared to CG sites found elsewhere in the genome (Figure S1A).

We compared our results from *A. thaliana* with those from 18 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 clonal group showing high levels of genetic similarity (Wilson, et al. 2019). Eichten *et al*. identified high levels of consistency between the

169 known genetic relationships for the clonal groups and the extent of CG divergence (Eichten,
170 et al. 2020). Through the analysis of a larger number of Turkish accessions, Skalska *et al*.
171 identified two distinct genetic clusters: one in coastal regions and the other in central regions,
172 the latter roughly corresponding to the high-elevation Anatolian plateau (Skalska, et al. 2020).
173 In this study, we focus exclusively on a single clonal group, as defined by Eichten *et al*. (clonal
174 group 1), and analyze only the 18 samples from the central Turkish region that have average
175 bisulfite sequencing depths greater than 10 (Eichten, et al. 2020) (Table S1). Between 0.8-
176 2.8% of CG methylation sites were polymorphic in *B. distachyon* (Figure S1B). Similar to *A.*
177 *thaliana*, the frequency of segregating CG methylation sites was higher within gbM regions,
178 with 3–4% of sites segregating (Figure S1B). The median length of DMRs in *B. distachyon*
179 was 107 bp (1st percentile: 44 bp, 99% percentile: 717 bp) which is smaller than the median
180 length in *A. thaliana* DMRs, which was 181 bp (1st percentile: 78 bp, 99% percentile: 1125 bp)
181 (Figure S2).
182

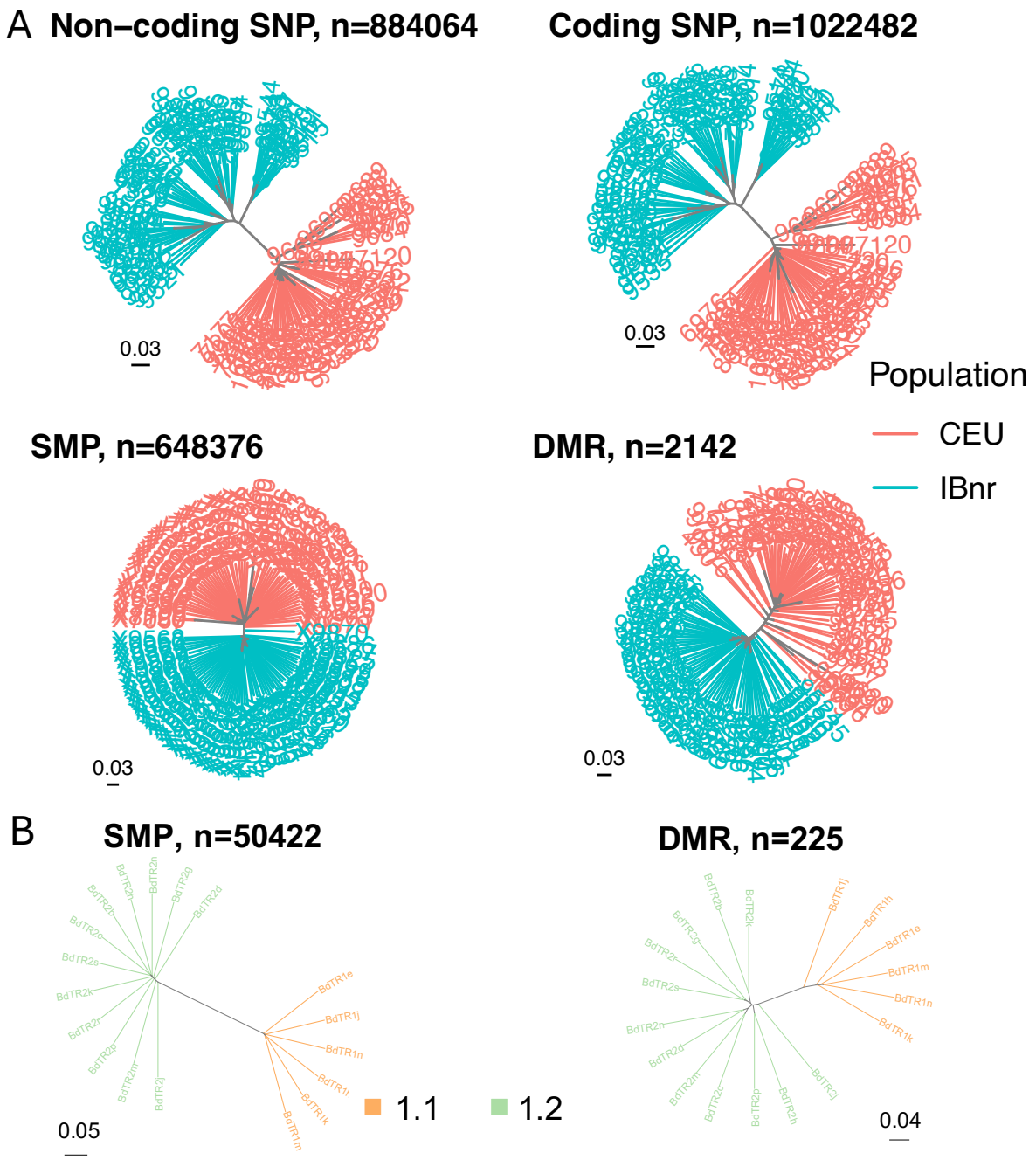


Figure 1. Methylomes outperform single nucleotide polymorphisms (SNP) in detecting population subdivision in *Brachypodium distachyon* but not *Arabidopsis thaliana*. (**A**) Neighbor-joining trees were generated using p-distance matrices for non-coding SNPs, coding SNPs, clocklike region single methylation 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 a Turkish *Brachypodium distachyon* clonal group.

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 (Charlesworth and Jain 2014; Wang and Fan 2014; Vidalis, et al. 2016)), 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 (Hewitt 1999; Sharbel, et al. 2000; The1001GenomesConsortium 2016).

218 We then performed a similar analysis using the *B. distachyon* accessions. Unlike the *A.*
219 *thaliana* populations, there were just 63 segregating SNP sites within the *B. distachyon* clonal
220 group 1, suggesting a very recent origin for this genetically homogeneous clone. We expected
221 that methylomes would be more useful in revealing population structure during the very early
222 stages of clonal expansion and divergence. Even though both samples were genetically
223 identical (Wilson, et al. 2019) and originated from a similar elevational gradient in the central
224 Turkish region, SMPs and DMRs revealed two distinct subgroups (Figures 1B, S5). Both our
225 results using gbM regions and the results from Eichten *et al*. that use the entire collection of
226 CG sites (Eichten, et al. 2020) reveal a difference in methylation profiles between the two
227 subgroups of Group 1. The SMP and DMR based trees were longer compared to those in *A.*
228 *thaliana* (Figure S4A) indicating greater methylome differentiation in *B. distachyon*. We
229 obtained phenotype data for the two subgroups (Filiz, et al. 2009) and found significant
230 differences in the longitude, seed lengths, seed yield, seed production, and germination
231 percentages (Figure S6-S7). As expected, we did not find any significant difference in
232 elevation between the subgroups from the Central Turkish region (Figure S7). Our findings
233 suggest that methylation variants can reveal subpopulation structure in recent timescales,
234 where there has been little time for genetic differences and multiple methylation changes to
235 accumulate. However, they become saturated with multiple changes occurring at the same
236 site as time passes obscuring population differentiation.
237
238 **Excess of rare alleles for SMP and DMR**

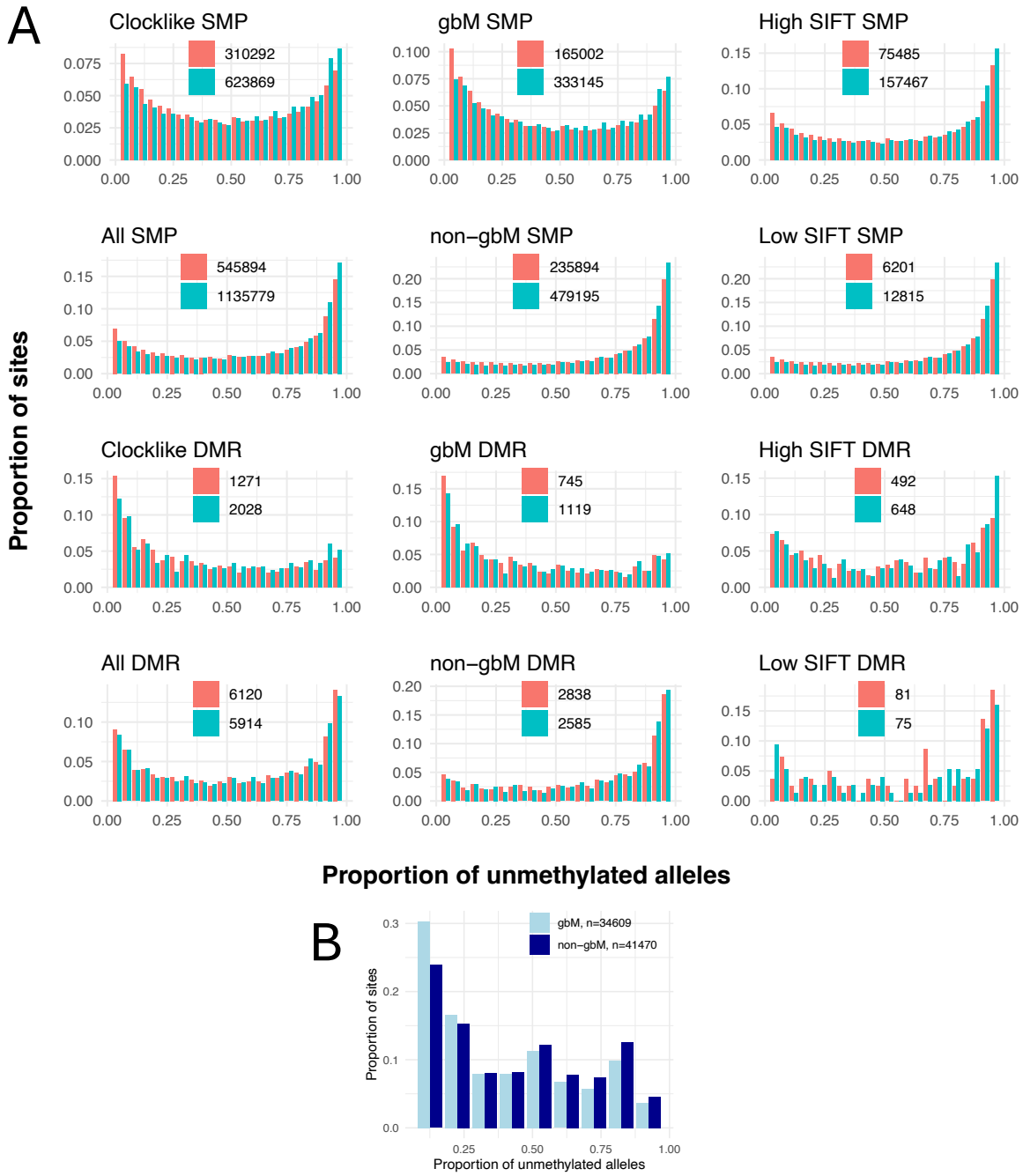


Figure 2. Methylation site frequency spectra generated using single methylation polymorphisms were similar regardless of genomic contexts. Methylation site frequency spectra (mSFS) generated using differentially methylated regions (DMR) exhibited greater variability. (**A**) mSFS for the Central European (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 group 1.2 for single methylation polymorphisms (SMP) and DMRs occurring within gbM and non-gbM regions. Only polymorphic sites 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 (Charlesworth and Jain 2014; Vidalis, et al. 2016). When considering all CG SMP or DMRs jointly, we found a greater frequency of unmethylated alleles for the CEU and IBnr populations, but IBnr consistently had higher frequencies of unmethylated alleles for SMP sites at all genomic regions (Figure 2A). The overall 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 (van der Graaf, et al. 2015) and are consistent with previous mSFS generated by combining *A. thaliana* samples from extended geographic regions (Vidalis, et al. 2016; Muyle, et al. 2021).

The relative frequencies of methylated and unmethylated alleles varied across individual genomic regions. SMPs in gbM regions for CEU and DMRs in gbM regions in both populations showed an excess of methylated alleles (Figure 2A). gbM are defined based on an excess of CG methylation and this is reflected in the mSFS. The number of methylated and unmethylated alleles was similar for SMP sites in clocklike regions, which could partly result from their comparable forward and backward methylation rates (Yao, et al. 2023). However, the differing direction of the skew in CEU and IBnr indicates some differences in the methylation rates for the populations for these sites. 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 (Ng and Henikoff 2003). Our results indicate that the level of gene constraint has little impact on the

275 mSFS shape (Figure 2A). While individual genomic regions differ in the proportion of
276 methylated sites they contain, it is evident that all cases exhibit an excess of rare variants,
277 suggesting a shared influence of demographic changes on all classes of variants.
278
279 In contrast to the mSFS for SMPs in *A. thaliana* (Figure 2A), the mSFS of *B. distachyon* group
280 1.2 indicated a slight excess of methylated alleles within gbM and non-gbM SMPs, although it
281 was less pronounced for the latter (Figure 2B). Similar to *A. thaliana*, an excess of rare alleles
282 was observed, suggesting a common demographic influence.
283
284 **SMP and SNP show unique evolutionary signatures**

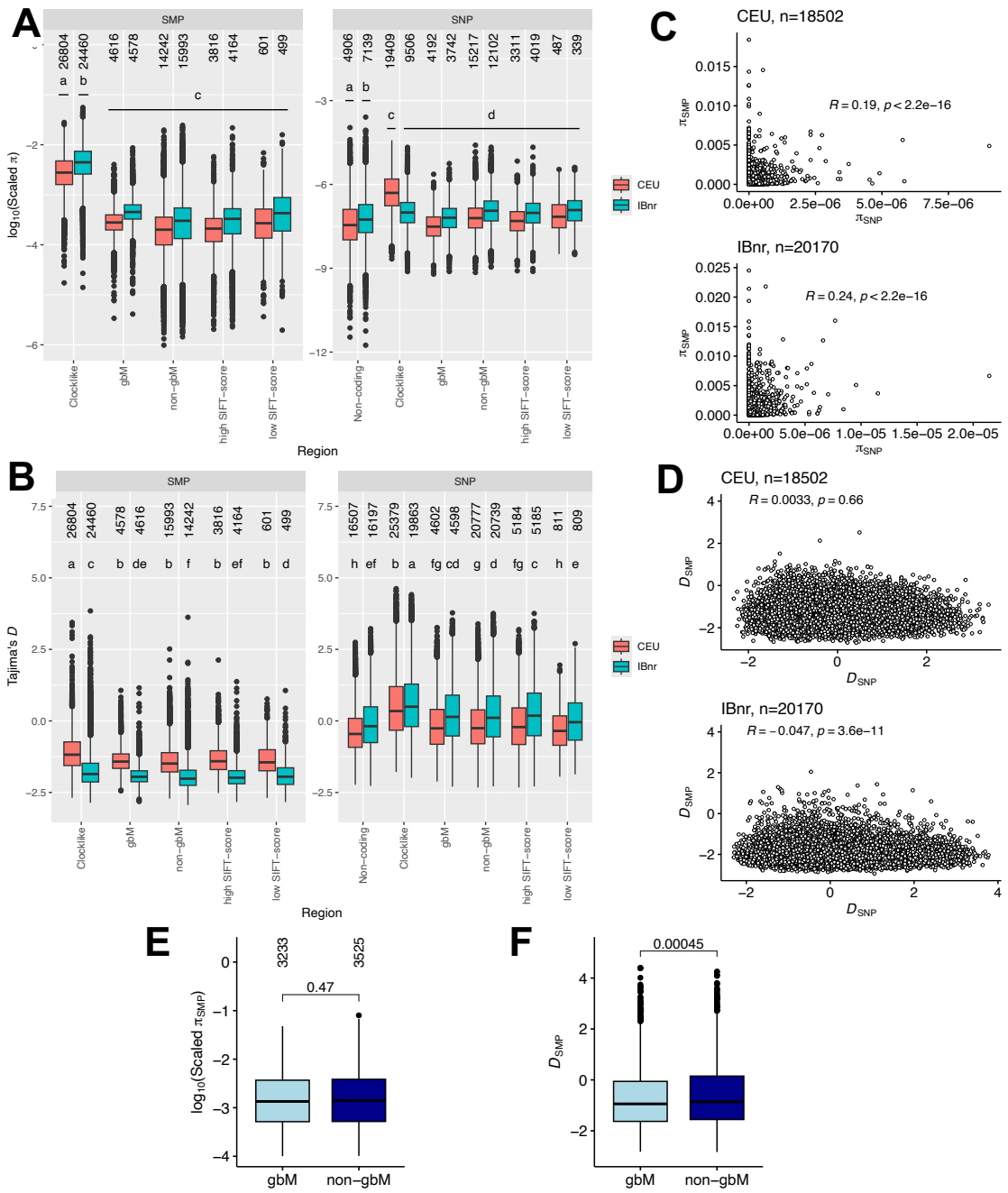
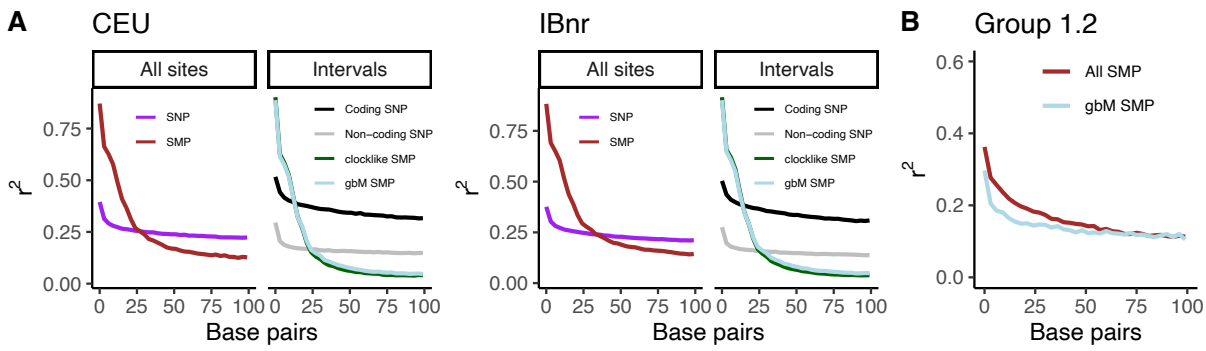


Figure 3. Unique evolutionary signatures in genomes and methylomes. (**A**) Estimates of pairwise nucleotide diversity (π) scaled by the number of cytosines per region for single methylation polymorphisms (π_{SMP}) and the number of bases for single nucleotide polymorphisms (π_{SNP}) for the Central 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

293 sites used for the analyses. Correlation tests were conducted to assess the associations
294 between π and Tajima's *D*. (**E**) π_{SMP} scaled by the number of cytosines per region for
295 *Brachypodium distachyon* clonal group 1.2. (**F**) D_{SMP} per region for *B. distachyon* group 1.2.
296 For (A), (B), (E), and (F), the numbers above the boxplots indicate the number of intervals.
297 For (A) and (B), letters indicate group-wise differences in means following Tukey's honest
298 significant difference tests. For (E) and (F), *P*-values from Wilcoxon tests are shown. For (F),
299 the color scheme and number of sites are the same as in (E).
300
301 We next sought to understand how SNPs and methylation polymorphisms are distributed
302 within *A. thaliana* populations. We first calculated the pairwise nucleotide diversity (π) for
303 SMPs (π_{SMP}) and SNPs (π_{SNP}) from noncoding, clocklike, gbM, non-gbM, and high and low
304 SIFT genic regions. We scaled estimates by the total number of sites (polymorphic and
305 monomorphic) to ensure that the results were comparable (that is per-site estimates). π_{SMP} is
306 three orders of magnitude higher than π_{SNP} (Figure 3A), likely driven by the high rates of
307 change in methylomes (van der Graaf, et al. 2015). π_{SNP} in the CEU population and π_{SMP} in
308 CEU and IBnr populations for clocklike regions were significantly higher than in other regions
309 (Figure 3A). This may indicate improved detection of environmental insensitivity—the criterion
310 for defining clocklike regions—when more polymorphic sites are present. π_{SMP} was similar for
311 gbM and non-gbM regions indicating that the levels of polymorphisms are not affected by gbM
312 (Figure 3A). π_{SMP} was also similar for regions containing genes with high-SIFT scores or low-
313 SIFT scores (Figure 3A) indicating that constraint at the genic level has negligible effects on
314 the levels of local methylation diversity.
315
316 We then compared how selective and demographic processes influence genomes and
317 methylomes by estimating Tajima's *D* for CEU and IBnr. For methylomes, we used a modified
318 Tajima's *D* using the D^m (Wang and Fan 2014) that accounts for the hypermutability of the
319 methylation sites. D_{SNP} was slightly negative for noncoding sites in the CEU and IBnr
320 populations (Figure 3B), reflecting the known history of population expansions (Strütt, et al.

321 2023). The higher D_{SNP} for IBnr (Figure 3B) is consistent with larger effective population sizes
322 over recent timescales (Strütt, et al. 2023). In contrast, IBnr had significantly lower D_{SMP}
323 compared to CEU (Figure 3B). This could be due to variation in methylation rates, purifying
324 selection, or demographic and selective (purifying selection) histories. Therefore, we used
325 SMCm (Sellinger, et al. 2023), which employs fixed mutation rates and compares the
326 distributions of mutation and methylation markers, to estimate the forward and backward
327 methylation rates for the IBnr and CEU populations. Methylation rates were only slightly higher
328 in CEU but the ratio of backward to forward rates was similar in both populations (Figure S8).
329 Hypermutability should result in greater frequencies of intermediate variants (Charlesworth
330 and Jain 2014), but we did not observe such a pattern (Figure 2A). Tajima's *D* is based on a
331 comparison of the number of segregating sites versus π. The disparity in D_{SMP} appears to be
332 driven primarily by the variation in the number of segregating sites per region (Figure S9). This
333 difference suggests that the demographic histories of the two populations may be slightly
334 different.
335
336 We then estimated Tajima's *D* in different genomic contexts for CEU and IBnr. Clocklike
337 regions had significantly higher D_{SNP} and D_{SMP} compared to other regions (Figure 3B). The D^m
338 method assumes that forward and backward methylation rates are equivalent (Wang and Fan
339 2014). Although the ratio of the backward to forward methylation rates in clocklike regions is
340 1.36, the ratios are 2.11 and 2.84 for gbM and non-clocklike regions, respectively (Yao, et al.
341 2023), indicating the assumption of D^m is less strongly violated for clocklike regions. The
342 differences in D_{SMP} between high-SIFT vs low-SIFT or between gbM and non-gbM were not
343 consistent in both populations (Figure 3B). This contrasts with consistently lower D_{SNP} for
344 regions containing genes with low-SIFT score that are under lower selective constraints
345 compared to regions with genes with higher-SIFT regions (Figure 3B). While ignoring the
346 differences between regions can introduce some noise when inferring evolutionary signatures
347 from methylomes, our results suggest any such effect is minimal. There was only a weak

348 positive correlation for π (Figure 3C) generated for SNPs vs SMPs that occur in coding regions
349 and negligible correlation for Tajima's *D* (Figure 3D).
350
351 We compared π and Tajima's *D* for *B. distachyon* SMPs generated using SNPs for the
352 Anatolian and Turkish lineage (B_East) (Stritt, et al. 2022; Minadakis, et al. 2023), to which
353 clonal group 1.2 belongs. Scaled π for SMPs was larger for the clonal group 1.2 compared to
354 the *A. thaliana* populations, but there were no consistent differences between gbM and non-
355 gbM regions (Figure 3A, 3E). Similar to *A. thaliana* (Figure 3A), scaled π for SMPs for clonal
356 group 1.2 was three orders of magnitude larger than the median SNP-based scaled π
357 (~1.8×10^{-6}) for the B_East lineage (Stritt, et al. 2022; Minadakis, et al. 2023) (Figure 3E).
358 Similar to *A. thaliana* (Figure 3B), D_{SMP} for clonal group 1.2 was negative. D_{SMP} was slightly
359 but significantly larger in non-gbM regions compared to gbM regions (Figure 3F). These
360 negative values contrast with the positive D_{SNP} for SNPs in the B_East lineage (Minadakis, et
361 al. 2023), further suggesting that methylomes exhibit unique signatures.
362
363 **Linkage disequilibrium quickly erodes for methylation variants**



364
365 **Figure 4**. Weak linkage disequilibrium (LD) between neighboring single methylation
366 polymorphisms (SMP). (**A**) Pairwise r^2 estimated using all single nucleotide polymorphisms
367 (SNP), coding SNPs, non-coding SNPs, all SMPs, and SMPs occurring within clocklike
368 regions, SMPs occurring within gene body methylated regions (gbM) for Central European

369 (CEU) and Iberian non-relicts (IBnr) *Arabidopsis thaliana* populations. (**B**) Pairwise r^2 was
370 estimated using SMPs for *Brachypodium distachyon* clonal group 1.2.
371
372 An important consideration for population genetic analysis is the presence of sites with
373 independent evolutionary histories (Sellinger, et al. 2023). Therefore, we assessed the levels
374 of pairwise linkage disequilibrium (LD) to identify the number of independent polymorphic
375 SNPs and SMPs. Pairwise LD decay was quicker for SMPs compared to SNPs in *A. thaliana*
376 (Figure 4A). LD decay was much stronger for SMPs occurring within clocklike or gbM regions
377 than for coding and noncoding SNPs with r^2 dropping below 0.1 within the first 100 bp (Figure
378 4A). LD decays to 0.1 for SMPs sites within gbM regions for *B. distachyon* within the first 100
379 bp (Figure 4B).
380
381 **Demographic inference using genetic and methylation variants**

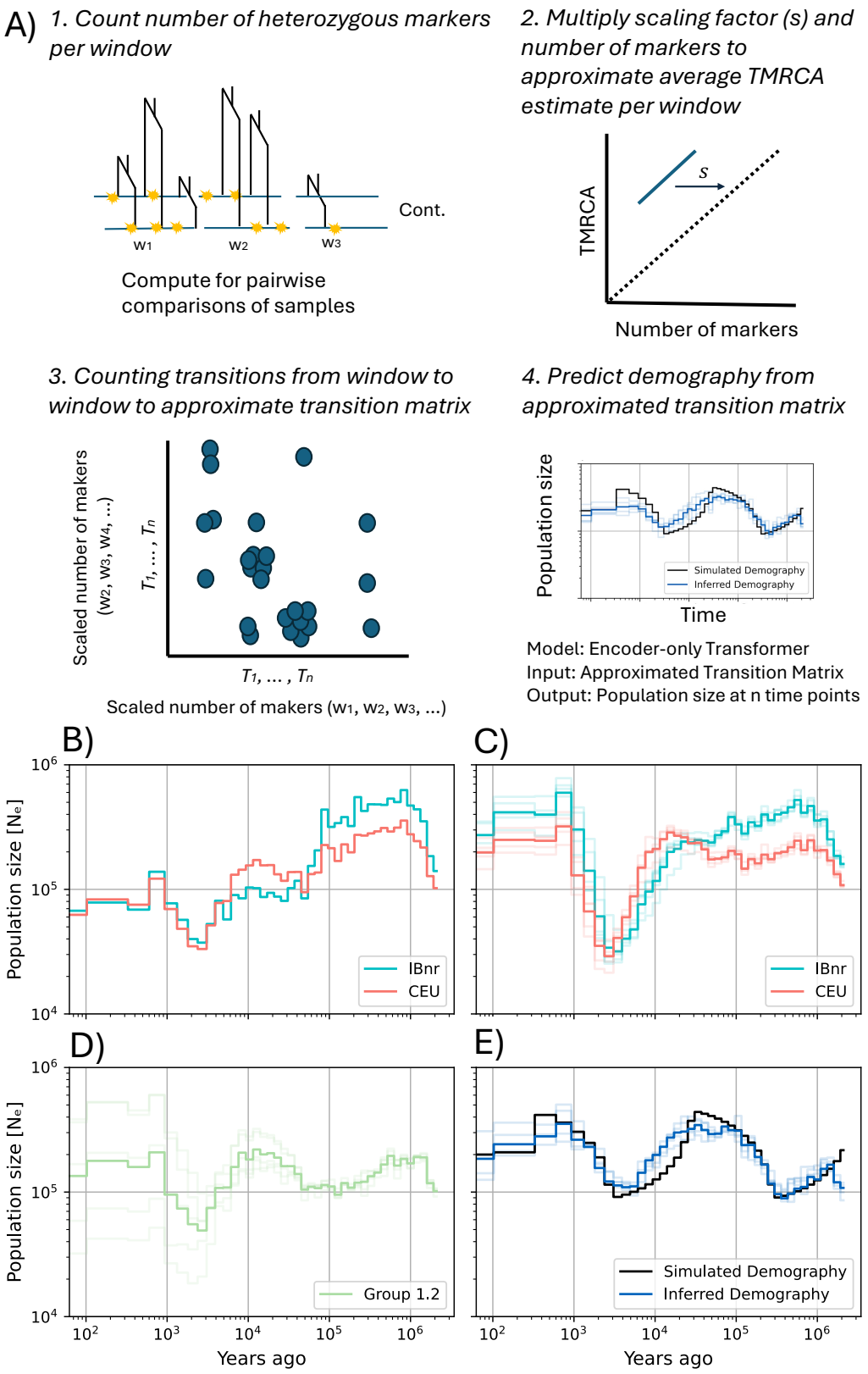


Figure 5. Demographic inference using mutation or methylation markers. (**A**) Schematic of the approximated transition matrix inference framework. The genome is first divided into 20 Kbp bins, and heterozygous markers are counted for all pairwise comparisons in a sample of

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 *Brachypodium distachyon* groups 1.2. 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 S10). 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 S11A). 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, and Group 1.2 regions (Figure S11B-D) produced values within the training data range (Figure S11A). Despite the differential properties of SMPs and SNPs, the distribution of heterozygous markers per 20 Kbp window was similar across populations (Figure S11B-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 (Strütt, et al. 2023). This signal was not clearly detected with SMPs (Figure 5C),

442 indicating their limited use for ancient events. Both SNPs and SMPs indicated another
443 bottleneck between 1-10 kya (Figure 5B-C), but SMPs also revealed a subsequent population
444 expansion (Figure 5C) undetected by SNPs (Figure 5B). The steeper expansion of IBnr
445 compared to CEU (*A. thaliana*) in the last 10,000 years likely led to the observed excess of
446 rare methylation alleles (Figures 2A, 3B). Using SMPs, we recovered two bottlenecks (~100
447 kya and between 1-10 kya) in *B. distachyon* (Figure 5D).
448
449 We validated our demographic inference method with two approaches. First, we simulated the
450 sawtooth demography observed in both species, confirming that our deep learning method
451 accurately detected multiple bottleneck and expansion events (Figure 5E). Second, we
452 simulated methylation markers with high rates of change (10^{-4}, 10^{-5}, or 10^{-6} per site per
453 generation) and backward rates 2.5 times larger than forward rates, based on the estimates
454 provided in (van der Graaf, et al. 2015), to test whether deviations from the infinite sites model
455 affect inference. We found that our method accurately recovered demographic histories
456 beyond 0.6 kya and before 5, 22, and 442 kya for methylation rates of 10^{-4}, 10^{-5}, and 10^{-6} per
457 site per generation, respectively (Figure S12). Conservatively, this suggests that inferences
458 20 kya and beyond are unlikely to be accurate for CEU and IBnr due to their estimated rates
459 of methylation change being around 10^{-5} per site per generation (Figure S8). Note that the
460 demographic estimates with SNPs are likely accurate for older events (>20kya).
461
462 Finally, we obtained short tandem repeat (STR) calls from (Reinar, et al. 2021) to test whether
463 this independent, fast-mutating marker also provides evidence for population expansions in
464 the *A. thaliana* populations. STRs occurring within noncoding regions were obtained from 36
465 and 54 samples from the CEU and IBnr, respectively, that overlapped with our methylation
466 dataset. We used the Bottleneck software (Cornuet and Luikart 1996) to test for a
467 heterozygosity deficit, which can serve as evidence of a population expansion (Luikart and
468 Cornuet 1998; Girod, et al. 2011). Since we were working with highly inbred effectively haploid
469 samples, we combined the calls from two randomly chosen pairs of samples to create diploid

470 data. We consistently detected a significant heterozygosity deficit using the sign test, the
471 standardized difference test, and the Wilcoxon signed-rank test in CEU and IBnr (Table S2).
472 These results indicate that we are confidently estimating recent population expansion using
473 SMPs (Figure 5C), highlighting the potential of methylation markers to provide insights beyond
474 those offered by SNPs alone.
475
476
477 **Discussion**
478 We observed contrasting patterns in the capacity of methylation variants to distinguish
479 population subdivisions in *A. thaliana* and *B. distachyon*. In *A. thaliana*, the infinite sites model
480 is likely violated for methylation, resulting in the erosion of methylome differentiation likely due
481 to the older divergence between populations. In contrast, genetically identical *B. distachyon*
482 samples exhibited distinct methylation profiles, most likely driven by their very recent
483 divergence. While individual genomic locations varied in whether they had more methylated
484 or unmethylated alleles—likely driven by the underlying molecular architecture—CG
485 methylation sites and DMRs consistently showed an excess of rare alleles in *A. thaliana* and
486 *B. distachyon*, indicating a common demographic influence. We found weak negative
487 correlations between π and no local correlations for Tajima's *D* between SNPs and methylation
488 variants in *A. thaliana*, underscoring their distinct molecular and evolutionary properties,
489 similar to the conclusions reached by (Schmitz, et al. 2013). Linkage quickly decays between
490 neighboring SMPs, which was also seen in wild *A. thaliana* sampled across the Northern
491 Hemisphere (Schmitz, et al. 2013). SMPs and SNPs recover recent and older demographic
492 changes, respectively, providing a comprehensive reconstruction of population evolutionary
493 histories.
494
495 The differing ability to detect population subdivisions in *A. thaliana* and *B. distachyon* using
496 methylation variants could lead to differences in divergence times, methylation rates, or local
497 adaptation. As the post-glacial re-colonization of the Iberian Peninsula and the spread

498 northward started ~16 kya (Hewitt 1999), multiple changes at the same methylation site
499 (homoplasy) during this period may have caused an underestimation of the true levels of
500 divergence in *A. thaliana*. The lack of congruence between SMP- and SNP-based
501 comparisons in A. thaliana also suggests that there has been limited post-expansion gene
502 flow, as otherwise, the two profiles would be more similar. The greater capacity to detect group
503 subdivision is *B. distachyon* could reflect a recent separation by longitude contributing to
504 methylome divergence. We indeed rule out the role of adaptation such as found in *F. vesca*
505 (Le Vève, et al. 2024), because we generated inferences using gbM regions that are highly
506 conserved in angiosperms (Muyle, et al. 2022). The observed divergence could be due to low
507 rates of methylation change, but we could not estimate methylation rates for *B. distachyon* as
508 our rate estimate relies on comparing the distributions of SNP and methylation variants. The
509 most parsimonious explanation is that the two subgroups diverged recently. As the B_East
510 lineage split from the western Mediterranean lineage 23 kya (Stritt, et al. 2022; Minadakis, et
511 al. 2023), the split between the closely related subgroups is likely to have been far more recent,
512 ensuring that the infinite sites model was not violated for methylation variants. Our findings
513 underscore the necessity for methods distinguishing neutral demographic changes from
514 selective pressures on the methylome, elucidating why SMPs behave differently across
515 species.
516
517 The differences between the shapes of the mSFS for the different genomic contexts indicate
518 that potential for differences in their molecular basis. gbM results from the combined activity
519 of MET1, which is enhanced by nearby existing methylation, as well as the histone variant
520 H2A.Z and the DNA demethylase ROS1 (Briffa, et al. 2023). The molecular basis underlying
521 the clocklike, non-gbM regions, as well as regions with high and low SIFT scores, remains
522 unknown. Interestingly, the shapes of the mSFS from SMPs in clocklike regions vary between
523 IBnr and CEU, and both also differ from the mSFS of DMRs in the same regionsThis pattern,
524 combined with their much higher rates of polymorphism compared to the remaining genomic
525 regions, suggests that these intervals likely have a unique molecular basis that warrants

526 further research. As clocklike regions accumulate methylation differences in a linear fashion
527 over time, they can be useful for studying neutral signatures (Yao, et al. 2023). While gbM
528 sites may not be entirely neutral—partly because they are known to affect gene expression
529 (Muyle, et al. 2022)—the similarity of their mSFS to those of clocklike regions, along with the
530 fact that ~60% of clocklike regions occur within gbM regions (Yao, et al. 2023), suggests
531 shared molecular features and that gbM sites may serve as a suitable substitute when
532 clocklike regions have not been defined.
533
534 Our results support the use of individual CG sites from all genomic contexts for population-
535 genetic analyses. Our between-population comparisons using CG sites in gbM regions
536 successfully identified the known population structure within *B. distachyon*, similar to a
537 previous study using genic CG sites (Eichten, et al. 2020). Sites from various genomic
538 contexts showed variation in the prevalence of methylated versus unmethylated alleles.
539 However, a consistent feature was the excess of rare alleles across these sites. Given the
540 rapid changes in markers such as methylation, determining the ancestral state is challenging.
541 The skew toward rare alleles in the mSFS suggests that demographic parameters exert a
542 consistent impact, even when the molecular mechanisms controlling different genomic regions
543 vary. Moreover, the *A. thaliana* (Kawakatsu, et al. 2016) and *B. distachyon* (Eichten, et al.
544 2020) methylomes, generated under similar environmental conditions, ensure that sites
545 affected solely by transient environmental effects would not be recognized as SMPs. Another
546 study using all CG sites from lettuce methylomes, cultivated under uniform environmental
547 conditions, also demonstrated the capacity to distinguish multiple species (Cao, et al. 2024).
548 DMRs may be preferable when low sequencing depth limits SMP identification (Taudt, et al.
549 2018), but their unclear evolutionary basis and susceptibility to multiple forces (Sellinger, et al.
550 2023) make them unreliable when used on their own.
551
552

553 The weak local correlations between genetic and methylation diversity and the association
554 between gene constraint and methylation diversity in *A. thaliana* can partly be attributed to
555 methylation affecting the mutability of neighboring sites (Mugal and Ellegren 2011; Xia, et al.
556 2012; Kusmartsev, et al. 2020) and transposable element polymorphisms influencing
557 methylation variation (Stuart, et al. 2016; Shen, et al. 2018). However, only about 12% of
558 methylation quantitative trait loci (QTLs) in *A. thaliana* are within 10Kbp of the epiallelic target
559 (Schmitz, et al. 2013). In maize, approximately 50% of methylation QTLs occur in *cis* (Eichten,
560 et al. 2013), indicating a potentially stronger trend in this species. The higher rate of
561 methylation changes compared to mutation accumulation (Ossowski, et al. 2010; van der
562 Graaf, et al. 2015) may also explain why methylation differences persist despite genetic
563 homogeneity. Our simulations also show that coalescent histories of markers with vastly
564 different rates of change diverge, which can explain the general lack of correlation between π
565 and Tajima's *D* for local genetic and methylation diversity. Our results are similar to those of
566 previous studies in lettuce (Cao, et al. 2024) and cotton (Zhao, et al. 2024), which also found
567 higher levels of polymorphism in the methylome compared to the genome. The lack of
568 correlation or similarity between the Tajima's *D* estimates from the genome and methylome in
569 *A. thaliana* and *B. distachyon* supports the notion that the SNPs and SMPs likely reflect
570 different parts of the common underlying coalescent genealogy at a locus: SNPs reflect the
571 older part of the tree, whereas SMPs would reflect the topology and branch lengths in the
572 more recent past. It is then not surprising that for recent positive selection events (selective
573 sweeps), the signatures are similar for both types of markers, and SMPs may even be more
574 accurate in detecting them (Shirai, et al. 2021). Our study mostly focuses on local correlations
575 in diversity, rather than genome-wide correlations, where life history traits and genetic
576 differentiation levels also play a significant role (Lele, et al. 2018; Medrano, et al. 2020; Wang,
577 et al. 2020; Avramidou, et al. 2021; Fargeot, et al. 2021; Avramidou, et al. 2023; Langford, et
578 al. 2024).
579

580 Our study demonstrates that methylation variants can independently detect and allow
581 inference of historical demographic changes. This represents a significant improvement over
582 a previous approach (Sellinger, et al. 2023), where methylation was used only as a secondary
583 marker to supplement SNP-based inferences. Furthermore, our improved deep learning
584 method does not require precise methylation rate parameters and is less sensitive to the
585 presence of invariant methylation sites (compared to the SMCm in (Sellinger, et al. 2023)).
586 Instead, it trains on diverse scenarios to match observed variant distributions. Our simulations
587 confirmed the utility of rapidly mutating markers for inferring demographic changes. Our STR
588 analysis confirm that there have been recent population expansions in the two *A. thaliana*
589 populations. While population expansions have previously been detected using SNPs in global
590 and European *A. thaliana* (The1001GenomesConsortium 2016; Lee, et al. 2017) and samples
591 combined from multiple *B. distachyon* clonal groups (Minadakis, et al. 2023), our approach
592 offers the advantage of identifying such signatures within a single population without requiring
593 excessively deep sequencing or a large sample size. Additionally, it enables the inference of
594 demographic history in studies that only generate methylation data and details the limits of
595 inference when using methylation markers. Overall, our study highlights the value of
596 methylation markers for understanding recent shifts in population size.
597
598
599 **Conclusions**
600 Our findings demonstrate that CG methylation can provide new insights into population
601 structure in clonal and selfing species, similar to its previously demonstrated utility in *A.*
602 *thaliana* from America and the clonal seagrass *Zostera marina* (Yao, et al. 2023). We
603 demonstrate the utility of methylation variants for uncovering very recent demographic events,
604 where SNP markers are limited due to their slower rate of change. We also illustrate its
605 usefulness in clonal groups, where the paucity of SNPs would render such inferences either
606 impossible or necessitate very deep and costly levels of DNA sequencing. Conversely, at older
607 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 (Sellinger, et al. 2023). 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 (The1001GenomesConsortium 2016) 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 SNP calls for each sample into the TAIR10 genome reference (Lamesch, et al. 2012) using SAMtools (Danecek, et al. 2021), BCFtools consensus (Danecek, et al. 2021), and fastq split (Nassar, et al. 2023). We obtained bisulfite sequencing data for the 1001 *A. thaliana* from (Kawakatsu, et al. 2016). We first removed poor-quality reads using Trimmomatic (Bolger, et al. 2014) 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 (Krueger and Andrews 2011). We ran Bismark deduplicate_bismark to remove optical duplicates and ran bismark_methylation_extractor to extract methylation calls. Some samples were sequenced using both single-end and paired-end reads. For these samples, the cytosine counts were merged after methylation calling. We studied cytosine sites with a coverage of at least five reads similar to what was done in

636 (Kawakatsu, et al. 2016). To identify the methylation state of a site, we performed a binomial
637 test with the non-conversion rate being used as the hypothesized probability of success
638 following the method in (Lister, et al. 2008). The non-conversion rate was calculated as the
639 proportion of methylation in the chloroplast as done in (Lister, et al. 2008). Sites were assigned
640 as methylated if the false discovery rate corrected *P*-values following the binomial test was
641 less than 0.01. Otherwise, they were designated as unmethylated. We used jDMR (Hazarika,
642 et al. 2022) to identify DMRs in the CG contexts as this method accounts for coverage when
643 calculating likelihood scores for methylation states. For this, we called the methylation status
644 per site using METHimpute (function: callMethylationSeparate) (Taudt, et al. 2018), extracted
645 cytosine clusters in the CG context ensuring that there was a minimum of eight cytosines in
646 each cluster (function: CfromFASTAv4 and makeReg with default options makeRegnull =
647 c(FALSE), N.boot = 10^5, N.sim.C = "all", fp.rate = 0.01, set.tol = 0.01), and identified the
648 methylation states of the regions in each sample (function: runMethimputeRegions), ensuring
649 that we used a minimum of eight cytosines and a coverage of three.
650
651 63 IBnr and 67 CEU samples overlapped between the SNP and methylation variant call files.
652 We randomly subsampled 63 samples from the CEU population to ensure the numbers were
653 comparable to those of the IBnr population. For our analyses, we only used sites where at
654 least 80% of the samples within each population were genotyped. The proportions of variant
655 SNP and SMP sites were calculated by dividing the number of sites with minor allele
656 frequencies greater than 0.02 by the total number of variant and invariant sites.
657
658 *Population (epi)genetic metric estimation for Arabidopsis thaliana*
659 To estimate π and Tajima's *D*, we created fasta files for each genomic region containing the
660 sequences for each sample. We substituted the variants from the SNP vcf into the TAIR10
661 reference genome using the approach described previously for each sample, extracted the
662 region of interest, and converted it to a fasta format. The regions could be the clocklike regions
663 described in (Yao, et al. 2023), genic regions comprising all exons, or non-coding regions

comprising all regions other than exons. Genic regions were further split into gbM genes using the list for *A. thaliana* provided by (Williams, et al. 2023) and high SIFT-score (>0.75) and low SIFT-score (<0.3) genes using the list provided by (Hamala and Tiffin 2020).

We estimated π_{SNP} and D_{SMP} using pegas (Paradis 2010) and estimated π_{SMP} and D_{SMP} using the D^m (Wang and Fan 2014). 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 (Taudt, et al. 2018). 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 (Danecek, et al. 2011) to estimate the allele frequencies for SMPs and DMRs. We used PopLDdecay (Zhang, et al. 2019) to estimate pairwise LD using the r^2 metric analyzing all sites with a minor allele frequency greater than 0.05. r^2 values across every three base pair interval were averaged prior to plotting the decay curve.

We detected a total of 6,012 and 6,188 polymorphic noncoding STRs in the CEU and IBnr populations, respectively. To reduce computational time, we randomly sampled 100 STRs without replacement for each run and repeated this process 10 times for each population. We performed the sign test, the standardized difference test, and the Wilcoxon signed-rank test. The tests were conducted using the two-phase model (TPM) and the stepwise mutation model (SMM), as these models best describe mutational processes in STRs. The variance for TPM was set to 30, with 70% SMM in the TPM. We ran 1,000 iterations.

692 *Between population comparisons for Arabidopsis thaliana*

693 We combined the CEU and IBnr vcfs and used VCF2Dis (https://github.com/BGI-
694 shenzhen/VCF2Dis) to estimate the p-distance matrix. We used this matrix to generate
695 neighbor-joining trees using FastME (Lefort, et al. 2015) and plotted it using ggtree (Yu, et al.
696 2017). We also performed DAPC to discriminate samples from the CEU and IBnr populations.
697 We thinned the dataset so that only sites that were separated by 1Kbp were sampled to reduce
698 the computational time. We converted vcf files into a genlight format using the
699 *genomic_converter* function in the R package radiator
700 (https://thierrygosselin.github.io/radiator/). We used the find.clusters function in adegenet
701 (Jombart 2008) to identify the number of clusters in each dataset allowing for a maximum of
702 20 clusters for SMPs/DMRs and 30 clusters for SNPs. As DAPC regions require fully typed
703 data, we imputed missing values using the mean value. Then, we performed cross-validation
704 using adegenet (Jombart 2008) to identify the number of informative principal components.
705 We retained a maximum of 15 principal components, used 90% of the data for training, and
706 calculated group-wise assignment success. We centered but did not scale variables. We used
707 30 replicates for each principal component. Finally, we performed DAPC using adegenet
708 (Jombart 2008).

709

710 Methylation rate estimation *for Arabidopsis thaliana*

711 We generated the multihetsep files for SNPs for the same 5Mbp regions that were used for
712 SMPs (described in the section immediately above). We combined the SNP and SMP
713 multihetsep files. We randomly selected 20 samples from each population to ensure
714 reasonable runtimes. We used the *Methylation_rate_estimation* function in SMCm to estimate
715 forward and backward rates (Sellinger, et al. 2023). We set the mutation rate as 7×10^{-9} per
716 site per generation and the recombination rate as 3.5×10^{-8} per site per generation. We ignored
717 regional influences on methylation rates and estimated rates under a site-only model. We used
718 the known estimates of forward (2.56×10^{-4} per generation per haploid methylome) and

719 backward methylation rates (6.3×10^{-4} per generation per haploid methylome) for *A. thaliana*
720 (van der Graaf, et al. 2015) as priors.
721
722 *Genetic and methylation variant calling in Brachypodium distachyon*
723 We obtained whole genome sequencing data (29 samples) for *B. distachyon* from (Wilson, et
724 al. 2019) and whole genome bisulfite sequencing data from (Eichten, et al. 2020). We used
725 Bd21-3 v1.2 genome reference and the GenBank annotation from the Joint Genome Institute
726 (http://phytozome.jgi.doe.gov/) for variant calling. We used Trimmomatic (Bolger, et al. 2014)
727 to trim reads using the paired-end options as stated above for the whole genome sequencing
728 data and the single-end option for the whole genome bisulfite sequencing data. We mapped
729 whole genome sequencing data reads to the reference using bwa-mem (Li 2013). We used
730 Picard Toolkit (https://broadinstitute.github.io/picard/) to add read groups and mark duplicate
731 reads. We used SAMtools (Danecek, et al. 2021) to remove unpaired reads, remove reads
732 with a mapping quality of less than 20, and sort reads. We called SNPs using GATK
733 HaplotypeCaller (McKenna, et al. 2010). We only retained SNPs with depths greater than five
734 and quality scores greater than 30 using BCFtools (Danecek, et al. 2021). Using this SNP
735 variant call file, we perform methylation variant calling in the same way as was done for *A.*
736 *thaliana*. We obtained the sample localities from (Filiz, et al. 2009) and created a map showing
737 their occurrence with Map Maker (https://maps.co/).
738
739 We used a minor allele frequency threshold of 10% to identity polymorphic sites in *B.*
740 *distachyon* due to the smaller number of samples being analyzed when compared to *A.*
741 *thaliana*. For the analysis of SMPs and DMRs, we only used variants that occurred within CG-
742 gbM regions identified in (Niederhuth, et al. 2016). We estimated mSFS, LD, p-distance
743 matrices, π, and Tajima's *D* in a similar way as we did for *A. thaliana*. We used the first principal
744 component for the DAPC analyses.
745

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:

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

where K=42 is the number of time points, $\delta = 0.001$ is the growth rate, and $t_{\text{max}} = 2.1\times10^6$ 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. 2021).

The obtained population sizes and times were passed onto *msprime* (Baumdicker, et al. 2022) 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 S12). 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 (Paszke, et al. 2019) (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 (Shazeer 2020), residual attention connections (He, et al. 2021), and rotary positional embeddings (Su, et al. 2024).

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

826 Face to enhance performance (Gugger, et al. 2022). An Accelerator object was configured with
827 mixed precision using bfloat16 and gradient accumulation over 12 steps (with batch size 32).
828 Throughout the training regiment, both the training and test losses were continuously
829 monitored to evaluate model performance and assess convergence.
830
831 *Generating data for demographic inference*
832 We first generated unique variant call files for each sample from the population variant call
833 files using BCFtools (Danecek, et al. 2021). We then used the generate_multihetsep.py from
834 MSMC-tools (https://github.com/stschiff/msmc-tools/tree/master) to combine all the samples
835 per chromosome. For *A. thaliana* SNPs, we used the entire chromosome 1 as we needed a
836 long contiguous fragment to generate the approximated transition matrix. For SMPs, we used
837 5Mbp segments in each of the five chromosomes (Table S3) as the SMP density was far
838 higher. We used the list of genotyped cytosine positions as a genomic location mask for the
839 SMPs. For *B. distachyon*, we included singletons to support the demographic inference given
840 the small sample sizes in this species. The data were polarized by identifying the minor (less
841 frequent) alleles and encoding them as derived, whereas the major alleles were encoded as
842 ancestral. The resulting genotype matrix represents this polarized, unphased data, where
843 each row corresponds to an individual and each column corresponds to a genomic position.
844
845 *Methylation simulation model*
846 We explored the boundaries of our course-grained demographic inference by adding a new
847 mutation model to msprime (Baumdicker, et al. 2022). We simulated a sawtooth demographic
848 model with cyclical population size changes over time and a baseline effective population size
849 of 2×10^{5} individuals. Simulations were conducted on a 5Mbp sequence with a recombination
850 rate of 3.4×10^{-8} per base pair per generation (ploidy=1), to isolate the influence of methylation
851 dynamics.
852

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, (Vidalis, et al. 2016)).

Declarations

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.14851293). 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

Avramidou EV, Korakaki E, Malliarou E, Boutsios S. 2023. Studying the Genetic and the Epigenetic Diversity of the Endangered Species Juniperus drupacea Labill. towards Safeguarding Its Conservation in Greece. Forests 14:1271.

Avramidou EV, Moysiadis T, Ganopoulos I, Michailidis M, Kissoudis C, Valasiadis D, Kazantzis K, Tsaroucha E, Tsaftaris A, Molassiotis A, et al. 2021. Phenotypic, Genetic, and Epigenetic Variation among Diverse Sweet Cherry Gene Pools. Agronomy 11:680.

906 Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, Zhu S, Eldon
907 B, Ellerman EC, Galloway JG, et al. 2022. Efficient ancestry and mutation simulation
908 with msprime 1.0. Genetics 220.
909 Becker C, Hagmann J, Muller J, Koenig D, Stegle O, Borgwardt K, Weigel D. 2011.
910 Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. Nature
911 480:245-249.
912 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina
913 sequence data. Bioinformatics 30:2114–2120.
914 Briffa A, Hollwey E, Shahzad Z, Moore JD, Lyons DB, Howard M, Zilberman D. 2023.
915 Millennia-long epigenetic fluctuations generate intragenic DNA methylation variance in
916 Arabidopsis populations. Cell Syst 14:953-967 e917.
917 Cao S, Sawettalake N, Li P, Fan S, Shen L. 2024. DNA methylation variations underlie
918 lettuce domestication and divergence. Genome Biol 25:158.
919 Charlesworth B, Jain K. 2014. Purifying selection, drift, and reversible mutation with
920 arbitrarily high mutation rates. Genetics 198:1587-1602.
921 Cornuet JM, Luikart G. 1996. Description and power analysis of two tests for detecting
922 recent population bottlenecks from allele frequency data. Genetics 144:2001-2014.
923 Cortijo S, Wardenaar R, Colome-Tatche M, Gilly A, Etcheverry M, Labadie K, Caillieux E,
924 Hospital F, Aury JM, Wincker P, et al. 2014. Mapping the epigenetic basis of complex
925 traits. Science 343:1145-1148.
926 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter
927 G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics
928 27.
929 Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T,
930 McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools.
931 Gigascience 10:giab008.
932 Devlin J, Chang M-W, Lee K, Toutanova K. 2018. BERT: Pre-training of Deep Bidirectional
933 Transformers for Language Understanding. arXiv 1810.04805.
934 Eichten SR, Briskine R, Song J, Li Q, Swanson-Wagner R, Hermanson PJ, Waters AJ, Starr
935 E, West PT, Tiffin P, et al. 2013. Epigenetic and genetic influences on DNA methylation
936 variation in maize populations. Plant Cell 25:2783-2797.

Eichten SR, Srivastava A, Reddiex AJ, Ganguly DR, Heussler A, Streich JC, Wilson PB, Borevitz JO. 2020. Extending the Genotype in Brachypodium by Including DNA Methylation Reveals a Joint Contribution with Genetics on Adaptive Traits. G3 (Bethesda) 10:1629-1637.

Fargeot L, Loot G, Prunier JG, Rey O, Veyssiere C, Blanchet S. 2021. Patterns of Epigenetic Diversity in Two Sympatric Fish Species: Genetic vs. Environmental Determinants. Genes (Basel) 12:107.

Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, et al. 2010. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A 107:8689-8694.

Filiz E, Ozdemir BS, Budak F, Vogel JP, Tuna M, Budak H. 2009. Molecular, morphological, and cytological analysis of diverse Brachypodium distachyon inbred lines. Genome 52:876-890.

Girod C, Vitalis R, Leblois R, Freville H. 2011. Inferring population decline and expansion from microsatellite data: a simulation-based evaluation of the Msvar method. Genetics 188:165-179.

Gugger S, Debut L, Wolf T, Schmid P, Mueller Z, Mangrulkar S, Sun M, Bossan B. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate.

Hamala T, Tiffin P. 2020. Biased Gene Conversion Constrains Adaptation in Arabidopsis thaliana. Genetics 215:831-846.

Hazarika RR, Serra M, Zhang Z, Zhang Y, Schmitz RJ, Johannes F. 2022. Molecular properties of epimutation hotspots. Nat Plants 8:146-156.

He R, Ravula A, Kanagal B, Ainslie J. 2021. RealFormer: Transformer Likes Residual Attention. arXiv:2012.11747.

Hewitt G. 1999. Post-glacial re-colonization of European biota. Biological Journal of the Linnean Society 68:87-112.

Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405.

Kawakatsu T, Huang SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, Castanon R, Nery JR, Barragan C, He Y, et al. 2016. Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell 166:492-505.

969 Korfmann K, Gaggiotti OE, Fumagalli M. 2023. Deep Learning in Population Genetics.
970 Genome Biol Evol 15.
971 Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for
972 Bisulfite-Seq applications. Bioinformatics 27:1571-1572.
973 Kusmartsev V, Drozdz M, Schuster-Bockler B, Warnecke T. 2020. Cytosine Methylation
974 Affects the Mutability of Neighboring Nucleotides in Germline and Soma. Genetics
975 214:809-823.
976 Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K,
977 Alexander DL, Garcia-Hernandez M, et al. 2012. The Arabidopsis Information Resource
978 (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40:D1202-1210.
979 Langford N, Fargeot L, Blanchet S. 2024. Spatial covariation between genetic and
980 epigenetic diversity in wild plant and animal populations: a meta-analysis. J Exp Biol
981 227.
982 Le Vève A, Sammarco I, Latzel V, Lafon Placette C. 2024. DNA methylation shows
983 footprints of altitude selection in the clonal plant species Fragaria vesca. bioRxiv
984 2024.03.19.585697.
985 Lee CR, Svardal H, Farlow A, Exposito-Alonso M, Ding W, Novikova P, Alonso-Blanco C,
986 Weigel D, Nordborg M. 2017. On the post-glacial spread of human commensal
987 Arabidopsis thaliana. Nat Commun 8:14458.
988 Lefort V, Desper R, Gascuel O. 2015. FastME 2.0: A Comprehensive, Accurate, and Fast
989 Distance-Based Phylogeny Inference Program. Mol Biol Evol 32:2798-2800.
990 Lele L, Ning D, Cuiping P, Xiao G, Weihua G. 2018. Genetic and epigenetic variations
991 associated with adaptation to heterogeneous habitat conditions in a deciduous shrub.
992 Ecol Evol 8:2594-2606.
993 Li H. 2013. Aligning sequence reads , clone sequences and assembly contigs with BWA-
994 MEM. 00:1-3.
995 Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008.
996 Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell
997 133:523-536.
998 Luikart G, Cornuet JM. 1998. Empirical Evaluation of a Test for Identifying Recently
999 Bottlenecked Populations from Allele Frequency Data. Conservation Biology 12:228-
1000 237.

1001 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K,
1002 Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce
1003 framework for analyzing next-generation DNA sequencing data. Genome Research
1004 20:1297–1303.
1005 Medrano M, Alonso C, Bazaga P, Lopez E, Herrera CM. 2020. Comparative genetic and
1006 epigenetic diversity in pairs of sympatric, closely related plants with contrasting
1007 distribution ranges in south-eastern Iberian mountains. AoB Plants 12:plaa013.
1008 Minadakis N, Williams H, Horvath R, Caković D, Stritt C, Thieme M, Bourgeois Y, Roulin
1009 AC. 2023. The demographic history of the wild crop relative Brachypodium distachyon is
1010 shaped by distinct past and present ecological niches. Peer Community Journal 3:e84.
1011 Mugal CF, Ellegren H. 2011. Substitution rate variation at human CpG sites correlates
1012 with non-CpG divergence, methylation level and GC content. Genome Biol 12:R58.
1013 Muyle A, Ross-Ibarra J, Seymour DK, Gaut BS. 2021. Gene body methylation is under
1014 selection in Arabidopsis thaliana. Genetics 218:iyab061.
1015 Muyle AM, Seymour DK, Lv Y, Huettel B, Gaut BS. 2022. Gene Body Methylation in Plants:
1016 Mechanisms, Functions, and Important Implications for Understanding Evolutionary
1017 Processes. Genome Biol Evol 14:evac038.
1018 Nassar LR, Barber GP, Benet-Pages A, Casper J, Clawson H, Diekhans M, Fischer C,
1019 Gonzalez JN, Hinrichs AS, Lee BT, et al. 2023. The UCSC Genome Browser database:
1020 2023 update. Nucleic Acids Res 51:D1188-D1195.
1021 Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function.
1022 Nucleic Acids Research 31:3812-3814.
1023 Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, Li Q, Rohr NA, Rambani A, Burke JM,
1024 Udall JA, et al. 2016. Widespread natural variation of DNA methylation within
1025 angiosperms. Genome Biol 17:194.
1026 Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel
1027 D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in
1028 Arabidopsis thaliana. Science 327:92-94.
1029 Paradis E. 2010. pegas: an R package for population genetics with an integrated-modular
1030 approach. Bioinformatics 26:419-420.

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. 2019. pyTorch: An Imperative Style, High-Performance Deep Learning Library: Curran Associates, Inc.

Reinar WB, Lalun VO, Reitan T, Jakobsen KS, Butenko MA. 2021. Length variation in short tandem repeats affects gene expression in natural populations of Arabidopsis thaliana. Plant Cell 33:2221-2234.

Richards CL, Alonso C, Becker C, Bossdorf O, Bucher E, Colome-Tatche M, Durka W, Engelhardt J, Gaspar B, Gogol-Doring A, et al. 2017. Ecological plant epigenetics: Evidence from model and non-model species, and the way forward. Ecol Lett 20:1576-1590.

Rodriguez-Izquierdo A, Carrasco D, Anand L, Magnani R, Catarecha P, Arroyo-Garcia R, Rodriguez Lopez CM. 2024. Epigenetic differences between wild and cultivated grapevines highlight the contribution of DNA methylation during crop domestication. BMC Plant Biol 24:504.

Sammarco I, Diez Rodriguez B, Galanti D, Nunn A, Becker C, Bossdorf O, Munzbergova Z, Latzel V. 2024. DNA methylation in the wild: epigenetic transgenerational inheritance can mediate adaptation in clones of wild strawberry (Fragaria vesca). New Phytol 241:1621-1635.

Sanchez T, Cury J, Charpiat G, Jay F. 2021. Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. Mol Ecol Resour 21:2645-2660.

Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ, et al. 2013. Patterns of population epigenomic diversity. Nature 495:193-198.

Sellinger T, Johannes F, Tellier A. 2023. Improved inference of population histories by integrating genomic and epigenomic data. eLife 12:RP89470.

Sharbel TF, Haubold B, Mitchell-Olds T. 2000. Genetic isolation by distance in Arabidopsis thaliana: biogeography and postglacial colonization of Europe. Mol Ecol 9:2109-2118.

Shazeer N. 2020. GLU Variants Improve Transformer. arXiv:2002.05202.

Shen Y, Zhang J, Liu Y, Liu S, Liu Z, Duan Z, Wang Z, Zhu B, Guo YL, Tian Z. 2018. DNA methylation footprints during soybean domestication and improvement. Genome Biol 19:128.

1063 Shirai K, Sato MP, Nishi R, Seki M, Suzuki Y, Hanada K. 2021. Positive selective sweeps of
1064 epigenetic mutations regulating specialized metabolites in plants. Genome Res
1065 31:1060-1068.
1066 Skalska A, Stritt C, Wyler M, Williams HW, Vickers M, Han J, Tuna M, Savas Tuna G, Susek
1067 K, Swain M, et al. 2020. Genetic and Methylome Variation in Turkish Brachypodium
1068 Distachyon Accessions Differentiate Two Geographically Distinct Subpopulations. Int J
1069 Mol Sci 21.
1070 Stritt C, Gimmi EL, Wyler M, Bakali AH, Skalska A, Hasterok R, Mur LAJ, Pecchioni N,
1071 Roulin AC. 2022. Migration without interbreeding: Evolutionary history of a highly selfing
1072 Mediterranean grass inferred from whole genomes. Mol Ecol 31:70-85.
1073 Strütt S, Sellinger T, Glemin S, Tellier A, Laurent S. 2023. Joint inference of evolutionary
1074 transitions to self-fertilization and demographic history using whole-genome
1075 sequences. eLife 12:e82384.
1076 Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, Lister R. 2016. Population scale
1077 mapping of transposable element diversity reveals links to gene regulation and
1078 epigenomic variation. eLife 5:e20777.
1079 Su J, Ahmed M, Lu Y, Pan S, Bo W, Liu Y. 2024. RoFormer: Enhanced transformer with
1080 Rotary Position Embedding. Neurocomputing 568:127063.
1081 Taudt A, Roquis D, Vidalis A, Wardenaar R, Johannes F, Colome-Tatche M. 2018.
1082 METHimpute: imputation-guided construction of complete methylomes from WGBS
1083 data. BMC Genomics 19:444.
1084 The1001GenomesConsortium. 2016. 1,135 Genomes Reveal the Global Pattern of
1085 Polymorphism in Arabidopsis thaliana. Cell 166:481-491.
1086 van der Graaf A, Wardenaar R, Neumann DA, Taudt A, Shaw RG, Jansen RC, Schmitz RJ,
1087 Colome-Tatche M, Johannes F. 2015. Rate, spectrum, and evolutionary dynamics of
1088 spontaneous epimutations. Proc Natl Acad Sci U S A 112:6676-6681.
1089 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I.
1090 2017. Attention Is All You Need. arXiv 1706.03762.
1091 Vidalis A, Zivkovic D, Wardenaar R, Roquis D, Tellier A, Johannes F. 2016. Methylome
1092 evolution in plants. Genome Biol 17:264.
1093 Wang J, Fan C. 2014. A neutrality test for detecting selection on DNA methylation using
1094 single methylation polymorphism frequency spectrum. Genome Biol Evol 7:154-171.

1095 Wang MZ, Li HL, Li JM, Yu FH. 2020. Correlations between genetic, epigenetic and
1096 phenotypic variation of an introduced clonal herb. Heredity (Edinb) 124:146-155.
1097 Williams CJ, Dai D, Tran KA, Monroe JG, Williams BP. 2023. Dynamic DNA methylation
1098 turnover in gene bodies is associated with enhanced gene expression plasticity in
1099 plants. Genome Biol 24:227.
1100 Wilson PB, Streich JC, Murray KD, Eichten SR, Cheng R, Aitken NC, Spokas K, Warthmann
1101 N, Gordon SP, Accession C, et al. 2019. Global Diversity of the Brachypodium Species
1102 Complex as a Resource for Genome-Wide Association Studies Demonstrated for
1103 Agronomic Traits in Response to Climate. Genetics 211:317-331.
1104 Xia J, Han L, Zhao Z. 2012. Investigating the relationship of DNA methylation with
1105 mutation rate and allele frequency in the human genome. BMC Genomics 13:S7.
1106 Xu G, Lyu J, Li Q, Liu H, Wang D, Zhang M, Springer NM, Ross-Ibarra J, Yang J. 2020.
1107 Evolutionary and functional genomics of DNA methylation in maize domestication and
1108 improvement. Nat Commun 11:5539.
1109 Yao N, Schmitz RJ, Johannes F. 2021. Epimutations Define a Fast-Ticking Molecular Clock
1110 in Plants. Trends Genet 37:699-710.
1111 Yao N, Zhang Z, Yu L, Hazarika R, Yu C, Jang H, Smith LM, Ton J, Liu L, Stachowicz JJ, et al.
1112 2023. An evolutionary epigenetic clock in plants. Science 381:1440-1445.
1113 Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. ggtree: an r package for visualization and
1114 annotation of phylogenetic trees with their covariates and other associated data.
1115 Methods in Ecology and Evolution 8:28-36.
1116 Zhang C, Dong SS, Xu JY, He WM, Yang TL. 2019. PopLDdecay: a fast and effective tool for
1117 linkage disequilibrium decay analysis based on variant call format files. Bioinformatics
1118 35:1786-1788.
1119 Zhang H, Lang Z, Zhu JK. 2018. Dynamics and function of DNA methylation in plants. Nat
1120 Rev Mol Cell Biol 19:489-506.
1121 Zhao T, Guan X, Hu Y, Zhang Z, Yang H, Shi X, Han J, Mei H, Wang L, Shao L, et al. 2024.
1122 Population-wide DNA methylation polymorphisms at single-nucleotide resolution in
1123 207 cotton accessions reveal epigenomic contributions to complex traits. Cell Res.
1124