This is a repository copy of *An evolutionary epigenetic clock in plants*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/203985/

Version: Accepted Version

Article:
Yao, N. orcid.org/0000-0003-1129-2769, Zhang, Z. orcid.org/0000-0003-2325-2342, Yu, L. orcid.org/0000-0003-3748-5470 et al. (10 more authors) (2023) An evolutionary epigenetic clock in plants. Science, 381 (6665). pp. 1440-1445. ISSN 0036-8075

https://doi.org/10.1126/science.adh9443

© 2023 The Authors. Except as otherwise noted, this author-accepted version of a journal article published in Science is made available via the University of Sheffield Research Publications and Copyright Policy under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

An evolutionary epigenetic clock in plants

Yao N[1#], Zhang Z[2#], Yu L[3], Hazarika R[2], Yu C[2], Jang H[1], Smith LM[4], Ton J[4], Liu L[5], Stachowicz J[6], Reusch TBH[3*], Schmitz RJ[1*], Johannes F[2*]

1 Department of Genetics, University of Georgia, Athens, USA
2 Plant Epigenomics, Technical University of Munich, Freising, Germany
3 Marine Evolutionary Ecology, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany
4 School of Biosciences, University of Sheffield, UK
5 Department of Statistics, University of Georgia, Athens, USA
6 Department of Evolution and Ecology, University of California, Davis, USA

# Contributed equally
* Corresponding authors

One-sentence summary: A fast-ticking evolutionary epigenetic clock in plants facilitates phylogenetic insights into the recent past.

Abstract: Molecular clocks are the basis for dating the divergence between lineages over macro-evolutionary timescales (~10^5-10^8 years). However, classical DNA-based clocks tick too slowly to inform us about the recent past. Here, we demonstrate that stochastic DNA methylation changes at a subset of cytosines in plant genomes possess a clock-like behavior. This 'epimutation-clock' is orders of magnitude faster than DNA-based clocks and enables phylogenetic explorations on a scale of years to centuries. We show experimentally that epimutation-clocks recapitulate known topologies and branching times of intra-species phylogenetic trees in the selfing plant *A. thaliana* and the clonal seagrass *Z. marina,* which represent two major modes of plant reproduction. This discovery will open new possibilities for high-resolution temporal studies of plant biodiversity.

Main Text:

Reconstructing the tree of life is a central goal in evolutionary biology. A key challenge is to infer the approximate date when two lineages diverged from each other in the past (*1*, *2*). In addition to fossil and archaeological evidence, molecular clocks have emerged as an important tool to perform such dating (*2*, *3*). Constant-rate clock calibrations such as those originally introduced by Zuckerkandl and Pauling (*4*) are based on the premise that neutral mutations in DNA (or proteins) accumulate at a fixed rate, so that nucleotide differences increase with time. If the mutation rate is known, it becomes possible to deduce when two lineages shared their most recent common ancestor (MRCA). Although the modern use of molecular clocks relies on a number of strong modeling assumptions (*5*), in practice, they are often the only means to obtain temporal information for parts of a phylogeny where fossil or archaeological records are lacking (*3*).

The low mutation rate found in most species limits the use of DNA-based clocks. With a rate of ~10^{-9}-10^{-8} (per site per year), they may offer sufficient temporal resolution over larger timescales (~10^4 to 10^8 years) but are less accurate in recent time (< 10^3 years from the present), as too few mutations accumulate to permit reliable tree inference and dating (*6*). However, it may be of interest to infer shallow divergence times of just a few decades to hundreds of years, for example, when assessing associations with species range shifts, colonization events, or environmental changes (*7*). In self-fertilizing or clonal species with short life cycles, new lineages

can diverge rapidly due to extensive genetic drift, restricted gene flow, and divergent natural selection (*8*). In the recent past, many such events have co-occurred with the emergence of modern civilizations and may in part even be driven by human activities (e.g., migration or trade). To improve the resolution in studying shallow divergences and their timing, a new class of molecular clocks is needed, whose tick-rate is orders of magnitude faster than that of DNA. We recently proposed that DNA cytosine methylation could provide a biomolecular basis for such a clock in plants (*9*), but this possibility has not been explored rigorously.

DNA cytosine methylation is a conserved base modification in eukaryotes (*10*). Stochastic enzymatic failure or off-target DNA methyltransferase activity at CG dinucleotides leads to lasting methylation losses or gains (i.e., epimutations) in daughter cells and their decedent cell lineages (*11*). Such CG methylation changes have been observed within the lifetime of mammals and have been extensively exploited as a DNA methylation clock of aging (*12*). However, unlike in mammals, many somatically-acquired CG epimutations are stably inherited across clonal and sexual generations in plants (*13–15*), and thus hold high-resolution information about the evolutionary histories of cell lines or clonal and sibling lineages (*11*, *16*). Estimates in several plant species indicate that CG epimutations are effectively neutral at the genome-wide scale and occur at a rate that is ~10,000 - 100,000 times higher than the genetic mutation rate per unit time (*13*, *14*, *17–20*). Here we show that the rapid accumulation of CG epimutations in plant genomes defines a fast-ticking evolutionary clock (henceforth 'epimutation-clock'), which can be used for the reconstruction and dating of phylogenies.

Discovery of clock-like regions in A. thaliana

We set out to construct a robust epimutation-clock by first searching for genomic regions whose CG epimutation rates are invariant to genetic and environmental perturbations. To do this, we used the selfing plant *Arabidopsis thaliana* as a model to generate mutation accumulation lines (MA-lines) from seven diverse natural accessions (i.e., genotypes) as founders (**Fig. 1A, Fig. S1A and Materials and Methods** (*21*)). These MA pedigrees had a depth of 17 generations with nine whole genome bisulfite sequencing (WGBS) measurements per pedigree, on average (see **Fig. S1A** (*21*) for details). To evaluate the impact of environmental factors, we also generated *A. thaliana* MA lines grown under biotic stress (repeated exposure to *Pseudomonas syringae* and salicylic acid) and combined these data with published MA-lines grown under abiotic stress (exposure to high salinity and drought conditions, (*22*, *23*), **Fig. 1A, Fig. S1B-D** (*21*)). The depth of these latter MA pedigrees varied from 7-13 generations and had ~9 WGBS measurements per pedigree.

Epimutation analysis of these different MA-lines revealed specific genomic regions, whose CG epimutation rates were largely invariable across the 14 different genetic or environmental perturbation experiments (**Fig. 1B-E, Fig. S2, Table S1-S4 and Materials and Methods** (*21*)). These clock-like regions were in stark contrast with other regions where rates were 5-fold more variable on average (**Fig. 1B, Table S1, S5** (*21*)). We found that clock-like regions displayed specific epigenomic features and comprised about 16.1% of all CGs in the genome (~896,323 of CG total sites, **Fig. 1B, Table S2-S3, Fig. S3** (*21*)). The majority of clock-like regions (~60%) were located within gene body methylated (gbM) genes (**Fig. S3B** (*21*)). In *A. thaliana*, gbM genes comprise a subclass of ~5,000 genes that display elevated CG methylation (*24*). Although the precise function of gbM is unclear, the nucleotide sequences and steady-state methylation levels of these genes are generally conserved across diverse plant species (*25*, *26*). Interestingly, we recently identified these same regions as epimutation hotspots in *A. thaliana (27)*. Their average epimutation rates exceed the genome-wide average by one order of magnitude (~10^{-3} versus 10^{-4} per CG site, per haploid genome, per generation). Hence, the biomolecular properties of these regions form the basis for a robust epimutation-clock, whose fast mCG substitution rate can facilitate high-resolution

inference about divergence events in the recent past. We here use the term "mCG substitution rate" to refer to the number of fixed CG methylation changes that occur over a specific time interval.

Clock calibration and phylogenetic inference in A. thaliana

To confirm the existence of such clock-like regions, we analyzed the largest *A. thaliana* MA pedigree available (here named MA1_1 and MA1_2 (*13*, *14*), see **Fig. 2A**). This MA pedigree features 15 independent lineages with a maximum depth of 32 generations. WGBS samples were available from generations 3, 31, and 32. In total, we detected 46,597 segregating CG epimutations within the clock-like regions after ~31 generations. By contrast, only 99 segregating SNPs were detected genome-wide over this timescale (*28*). This shows that epimutations in the relatively small clock-like regions are overall much more abundant than genome-wide SNPs. Using pairwise distances based on the GTR2 substitution model (**Materials and Methods** (*21*)), we performed neighbor-joining clustering of the samples on the basis of their mCG status within the clock-like regions and were able to recapitulate the known topology (**Fig. 2B**). Hence, the rapid accumulation of epimutations is highly informative about divergence events as recent as a few generations.

Estimates of the mCG substitution rates were highly consistent across lineages (**Fig. 2C-D**), yielding rates of $4.43 \pm 0.229 \times 10^{-4}$ and $4.34 \pm 0.214 \times 10^{-4}$ (± SE, **Supplementary text**, **Table S6** (*21*)) for MA1_1-G31 and MA1_2-G31, respectively. We then applied the mean mCG substitution rate of MA1_1-G31 to the methylation data of MA1_2-G31 to infer the time until its MRCA. This estimate indicated that the MRCA lived approximately 30.4 ± 2.94 generations ago (95% CI, **Fig. 2E, Table S7-S8, Fig. S4** (*21*)), which is remarkably close to the actual depth of the pedigree (31 generations). By contrast, attempts to date the MRCA using available SNP data from the MA lines were biased and more variable, yielding an estimate of 28.2 ± 8.22 generations (95% CI), an uncertainty of nearly ~29.19% of the total age of the phylogeny (**Fig. 2E, Table S7, Materials and Methods** (*21*)). Together these results indicate that CG epimutations are much more robust and informative over these short timescales than DNA mutations, a conclusion that is strongly supported by theoretical arguments and extensive simulation studies (*6*, *29*), **Table S9-S10, Fig. S5-7** (*21*)).

We sought to extrapolate these insights to natural settings. Hagmann et al. (*30*) sequenced the genomes and DNA methylomes of 13 *A. thaliana* accessions collected around the Great Lakes and the East Coast of the United States (**Fig. 3A**), which all belong to a large haplogroup HPG1 (*30*, *31*). As *A. thaliana* is not native to North America, it has been hypothesized that these lineages were introduced with the arrival of the early European settlers (*30*, *32*). The average genetic distance between the 13 lineages and the HPG1 pseudo reference genome is about 245 nucleotide positions, providing evidence for their close kinship. In line with previous work (*30*, *33–36*), clustering accessions based on CG methylation produced nearly identical phylogenetic trees compared to SNP-based clustering, which indicates that they can capture the same evolutionary relationships between lineages (**Fig. 3B-C**). Attempts to date the MRCA for 12 non-recombinant lineages (*30*, *32*) based on the available SNP data and three published SNP substitution rates from mutation accumulation experiments (*28*, *32*, *37*) yielded estimates ranging from the year 1792 ± 59.2 years (95% CI) to 1766 ± 66.2 years (95% CI, **Fig. 3F**, **Table S11** (*21*)). As an alternative method, we also combined SNP data from these 12 taxa (ingroups) with that of different herbarium samples (outgroups) as input for Bayesian tip calibration with BEAST (*38*). The mean values of the estimated divergence time of the 12 taxa ranged from the year ~1949 to the year ~1792, depending on the set of herbarium outgroups used (**Fig. 3G, Table S12, Supplementary text** (*21*)). This extensive uncertainty is partly corroborated by our simulation studies, which indicate that a DNA mutation rate of 10^{-9} would yield large estimation errors for founding events occurring as recently as a few centuries ago (**Table S9**, **Supplementary text** (*21*)).

To help resolve the temporal questions about the origin of these lineages, we reanalyzed the WGBS data and applied the experimentally calibrated epimutation-clock. Our analysis placed the MRCA of these 12 lineages in the year 1863 ± 19 years (95% CI, **Fig. 3D-F, Table S11** (*21*)). This estimate provides solid evidence for a very recent MRCA. As North American herbarium samples of *A. thaliana* can be found that are older than this date (*32*), our results support the notion that these lineages belong to a clade of a larger phylogeny whose breadth has not been sufficiently sampled and/or are the result of serial founding events on this continent. Broader sampling of North American accessions would be necessary to further study the introduction time of *A. thaliana*.

Inference and timing of recent clonal phylogenies in Z. marina

The rapid formation and radiation of new plant lineages are particularly prevalent in species with facultative clonal reproduction through the production of runners, stolons or tillers. These comprise an estimated 40% of plant species on earth (*39*), many of which are of significant ecological relevance. The marine flowering plant *Zostera marina* is an important example of these. As most seagrasses, *Z. marina* is the foundation of entire ecosystems (*40*). It has important ecosystem functions in carbon storage, biodiversity enhancement, coastal protection, and is currently being developed as a seagrass genomic model (*41*). Along with sexual reproduction carried out fully under water, including subaqueous pollination, clones can grow several football fields large (*41*). Previously, the accumulation of somatic genetic variation (as SNPs) has been observed in large mega-clones (*42*). However, a dating of clones less than a decade old is elusive, but highly useful to elucidate the demographic structure, longevity, and ultimately vulnerability of seagrasses to a changing ocean environment.

To demonstrate that the epimutation-clock is a powerful tool for the reconstruction and dating of clonal phylogenies over short time scales, we took advantage of two *Z. marina* clones that had been cultured since 2004 for ecological experiments in Bodega Bay, California (**Materials and Methods, Supplementary text and Table S13-14** (*21*)). These two clones (clone R and clone G) were initiated and propagated independently from each other in large tanks under ambient light, temperature and flow-through of ocean water (**Fig. 4A**). Sixteen ramets were sampled from each clone in 2021, corresponding to a clonal age of 17 years (**Materials and Methods** (*21*)). WGBS was performed for 15 ramets of clone R, as one ramet did not belong to this clone (*43*). As for clone G, all 16 ramets were sequenced using WGBS, and technical replicates were generated by sequencing five aliquots of one sample independently (**Materials and Methods** (*21*)). As extensive epigenomic information is lacking for the de novo identification of clock-like regions in *Z. marina*, we used gbM genes as a proxy of the epimutation-clock (**Materials and Methods, Fig. S4, S8, Table S8** (*21*)). For comparison, we also generated deep (100x) re-sequencing data for SNP identification in a subset of these ramets (**Materials and Methods** (*21*)).

Our analysis revealed that the clones have generated 20,713 (clone R) and 21,008 (clone G) fixed CG methylation changes in gbM genes over the course of 17 years, on average, compared with only 31 and 47 fixed SNP changes (**Materials and Methods** (*21*)). Again, this constitutes a large excess of epimutations relative to SNPs per unit of time, even though the effective genome size of the former is orders of magnitude smaller. Although the true clonal phylogeny is unknown in this experiment, an epimutation-clock based analysis revealed high-confidence phylogenetic trees in both clones (**Fig. 4B-C, Supplementary text** (*21*)). We were unable to recapitulate the phylogenetic trees using fixed SNPs (**Fig. 4D-E**). Instead, we observed that the bootstrap support values in the SNP-based trees were low (~0.6, **Supplementary text** (*21*)), which indicates that there is little phylogenetic information in the few SNPs captured among clonal ramets over these time scales.

To assess our ability to date the clones based on DNA methylation data, we calibrated our epimutation-clock in clone R, using its known age of exactly 17 years (**Fig. 4 F-H, Materials and Methods** (*21*)). Application of the calibrated clock to the phylogenetic tree of clone G estimates the MRCA at 18.14 ± 3.01 years ago (95% CI), which aligns well with the actual founding date of this clone. (**Fig. 4H, Table S15** (*21*)). Applying the same approach using fixed SNPs led to variable estimates 21.39 ± 11.59 years ago (95% CI), an uncertainty of one decade over this short timescale (**Fig. 4H**).

Discussion and outlook

The existence of a fast-ticking evolutionary epigenetic clock in plants opens novel research avenues at the interface between evolutionary biology and ecology (*9*). Our proof-of-principle study focused on the application of this clock to the reconstruction and dating of shallow, intra-species phylogenies that arise within clonal or selfing species over short timescales. The use of epimutation-clocks may be extended to plant species with longer life cycles as the necessary clock-calibration could rely on DNA methylation data from F2 intercrosses, rather than from mutation accumulation lines. A promising extension of our work is to combine the clock-like regions with modified coalescent methods to infer recent demographic changes (i.e. bottlenecks) and selection at the population level. The rapid rate at which epigenetic diversity arises within the clock-like regions will facilitate unprecedented temporal resolution. It can shed new light on microevolutionary questions that have been challenging to resolve, such as the timing of introduction of invasive species, the rate of poleward and upslope spread of species after the retreat of the Pleistocene glaciers, and the consequences of anthropogenic activities for population divergence. In an era of climate change, plant biodiversity is transforming at a fast pace. The ability to monitor rapid population dynamics at the molecular level will help us gauge the fate of this diversity going into the future.

References

1. S. Kumar, S. B. Hedges, Advances in Time Estimation Methods for Molecular Data. *Mol. Biol. Evol.* **33**, 863–869 (2016).

2. E. Zuckerkandl, Molecular disease, evolution, and genetic heterogeneity. *Horiz. Biochem. Biophys.*, 189–225 (1962).

3. S. Kumar, Molecular clocks: four decades of evolution. *Nat. Rev. Genet.* **6**, 654–662 (2005).

4. E. Zuckerkandl, L. Pauling, Evolutionary Divergence and Convergence in Proteins. *Evolving Genes and Proteins* (1965), pp. 97–166.

5. S. Y. W. Ho, S. Duchêne, Molecular-clock methods for estimating evolutionary rates and timescales. *Mol. Ecol.* **23**, 5947–5965 (2014).

6. Z. Yang, On the best evolutionary rate for phylogenetic analysis. *Syst. Biol.* **47**, 125–133 (1998).

7. P. Pyšek, D. M. Richardson, Invasive Species, Environmental Change and Management, and Health. *Annu. Rev. Environ. Resour.* **35**, 25–55 (2010).

8. S. C. H. Barrett, R. I. Colautti, C. G. Eckert, Plant reproductive systems and evolution during biological invasion. *Mol. Ecol.* **17**, 373–383 (2008).

9. N. Yao, R. J. Schmitz, F. Johannes, Epimutations define a fast-ticking molecular clock in plants. *Trends Genet.* (2021), doi:10.1016/j.tig.2021.04.010.

10. S. Feng, S. J. Cokus, X. Zhang, P.-Y. Chen, M. Bostick, M. G. Goll, J. Hetzel, J. Jain, S. H. Strauss, M. E. Halpern, C. Ukomadu, K. C. Sadler, S. Pradhan, M. Pellegrini, S. E. Jacobsen, Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 8689–8694 (2010).

11. F. Johannes, R. J. Schmitz, Spontaneous epimutations in plants. *New Phytol.* **221**, 1253–1259 (2019).

12. S. Horvath, K. Raj, DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).

13. R. J. Schmitz, M. D. Schultz, M. G. Lewsey, R. C. O'Malley, M. A. Urich, O. Libiger, N. J. Schork, J. R. Ecker, Transgenerational epigenetic instability is a source of novel methylation variants. *Science*. **334**, 369–373 (2011).

14. C. Becker, J. Hagmann, J. Müller, D. Koenig, O. Stegle, K. Borgwardt, D. Weigel, Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. *Nature*. **480**, 245–249 (2011).

15. A. Wibowo, C. Becker, J. Durr, J. Price, S. Spaepen, S. Hilton, H. Putra, R. Papareddy, Q. Saintain, S. Harvey, G. D. Bending, P. Schulze-Lefert, D. Weigel, J. Gutierrez-Marcos, Partial maintenance of organ-specific epigenetic marks during plant asexual reproduction leads to heritable phenotypic variation. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E9145–E9152 (2018).

16. A. Vidalis, D. Živković, R. Wardenaar, D. Roquis, A. Tellier, F. Johannes, Methylome evolution in plants. *Genome Biol.* **17**, 264 (2016).

17. A. van der Graaf, R. Wardenaar, D. A. Neumann, A. Taudt, R. G. Shaw, R. C. Jansen, R. J. Schmitz, M. Colomé-Tatché, F. Johannes, Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proceedings of the National Academy of Sciences*. **112**, 6676–6681 (2015).

18. Y. Shahryary, A. Symeonidi, R. R. Hazarika, J. Denkena, T. Mubeen, B. Hofmeister, T. van Gurp, M. Colomé-Tatché, K. J. F. Verhoeven, G. Tuskan, R. J. Schmitz, F. Johannes, AlphaBeta: computational inference of epimutation rates and spectra from high-throughput DNA methylation data in plants. *Genome Biol.* **21**, 1–22 (2020).

19. J. Denkena, F. Johannes, M. Colomé-Tatché, Region-level epimutation rates in Arabidopsis thaliana. *Heredity* , 1–13 (2021).

20. B. T. Hofmeister, K. Lee, N. A. Rohr, D. W. Hall, R. J. Schmitz, Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation. *Genome Biol.* **18**, 155 (2017).

21. Yao N, Zhang Z, Yu L, Hazarika R, Yu C, Jang H, Smith LM, Ton J, Liu L, Stachowicz J, Reusch TBH, Schmitz RJ, Johannes F, Supplementary Materials. *Science* (2023).

22. C. Jiang, A. Mithani, E. J. Belfield, R. Mott, L. D. Hurst, N. P. Harberd, Environmentally responsive genome-wide accumulation of de novo Arabidopsis thaliana mutations and epimutations. *Genome Res.* **24**, 1821–1829 (2014).

23. D. R. Ganguly, P. A. Crisp, S. R. Eichten, B. J. Pogson, The Arabidopsis DNA Methylome Is Stable under Transgenerational Drought Stress. *Plant Physiol.* **175**, 1893–1912 (2017).

24. R. K. Tran, J. G. Henikoff, D. Zilberman, R. F. Ditt, S. E. Jacobsen, S. Henikoff, DNA methylation profiling identifies CG methylation clusters in Arabidopsis genes. *Curr. Biol.* **15** (2005), doi:10.1016/j.cub.2005.01.008.

25. S. Takuno, B. S. Gaut, Body-Methylated Genes in Arabidopsis thaliana Are Functionally Important and Evolve Slowly. *Molecular Biology and Evolution*. **29** (2012), pp. 219–227.

26. S. Takuno, B. S. Gaut, Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proceedings of the National Academy of Sciences*. **110** (2013), pp. 1797–1802.

27. R. R. Hazarika, M. Serra, Z. Zhang, Y. Zhang, R. J. Schmitz, F. Johannes, Molecular properties of epimutation hotspots. *Nat Plants*. **8**, 146–156 (2022).

28. S. Ossowski, K. Schneeberger, J. I. Lucas-Lledó, N. Warthmann, R. M. Clark, R. G. Shaw, D. Weigel, M. Lynch, The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science*. **327**, 92–94 (2010).

29. S. Klopfstein, T. Massingham, N. Goldman, More on the Best Evolutionary Rate for Phylogenetic Analysis. *Syst. Biol.* **66**, 769–785 (2017).

30. J. Hagmann, C. Becker, J. Müller, O. Stegle, R. C. Meyer, G. Wang, K. Schneeberger, J. Fitz, T. Altmann, J. Bergelson, K. Borgwardt, D. Weigel, Century-scale Methylome Stability in a Recently Diverged Arabidopsis thaliana Lineage. *PLoS Genet.* **11**, e1004920 (2015).

31. A. Platt, M. Horton, Y. S. Huang, Y. Li, A. E. Anastasio, N. W. Mulyati, J. Agren, O. Bossdorf, D. Byers, K. Donohue, M. Dunning, E. B. Holub, A. Hudson, V. Le Corre, O. Loudet, F. Roux, N. Warthmann, D. Weigel, L. Rivero, R. Scholl, M. Nordborg, J. Bergelson, J. O. Borevitz, The scale of population structure in Arabidopsis thaliana. *PLoS Genet.* **6**, e1000843 (2010).

32. M. Exposito-Alonso, C. Becker, V. J. Schuenemann, E. Reiter, C. Setzer, R. Slovak, B. Brachi, J. Hagmann, D. G. Grimm, J. Chen, W. Busch, J. Bergelson, R. W. Ness, J. Krause, H. A. Burbano, D. Weigel, The rate and potential relevance of new mutations in a colonizing plant lineage. *PLoS Genet.* **14**, e1007155 (2018).

33. R. J. Schmitz, M. D. Schultz, M. A. Urich, J. R. Nery, M. Pelizzola, O. Libiger, A. Alix, R. B. McCosh, H. Chen, N. J. Schork, J. R. Ecker, Patterns of population epigenomic diversity. *Nature*. **495**, 193–198 (2013).

34. V. N. Ibañez, M. van Antro, C. Peña-Ponton, S. Milanovic-Ivanovic, C. A. M. Wagemaker, F. Gawehns, K. J. F. Verhoeven, Environmental and genealogical effects on DNA methylation in a widespread apomictic dandelion lineage. *J. Evol. Biol.* **36**, 663–674 (2023).

35. I. Sammarco, B. D. Rodríguez, D. Galanti, A. Nunn, C. Becker, O. Bossdorf, Z. Münzbergová, V. Latzel, DNA methylation in the wild: epigenetic transgenerational inheritance can mediate adaptation in clones of wild strawberry (Fragaria vesca) (2023), doi:10.21203/rs.3.rs-2642365/v1.

36. J. Xu, G. Chen, P. J. Hermanson, Q. Xu, C. Sun, W. Chen, Q. Kan, M. Li, P. A. Crisp, J. Yan, L. Li, N. M. Springer, Q. Li, Population-level analysis reveals the widespread occurrence and phenotypic consequence of DNA methylation variation not tagged by genetic variation in maize. *Genome Biol.* **20**, 243 (2019).

37. M.-L. Weng, C. Becker, J. Hildebrandt, M. Neumann, M. T. Rutter, R. G. Shaw, D. Weigel, C. B. Fenster, Fine-Grained Analysis of Spontaneous Mutation Spectrum and Frequency in Arabidopsis thaliana. *Genetics*. **211**, 703–714 (2019).

38. M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, A. Rambaut, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol*. **4**, vey016 (2018).

39. O. Honnay, B. Bossuyt, Prolonged clonal growth: escape route or route to extinction? *Oikos*. **108**, 427–432 (2005).

40. R. K. F. Unsworth, L. C. Cullen-Unsworth, B. L. H. Jones, R. J. Lilley, The planetary role of seagrass conservation. *Science*. **377**, 609–613 (2022).

41. J. L. Olsen, P. Rouzé, B. Verhelst, Y.-C. Lin, T. Bayer, J. Collen, E. Dattolo, E. De Paoli, S. Dittami, F. Maumus, G. Michel, A. Kersting, C. Lauritano, R. Lohaus, M. Töpel, T. Tonon, K. Vanneste, M. Amirebrahimi, J. Brakel, C. Boström, M. Chovatia, J. Grimwood, J. W. Jenkins, A. Jueterbock, A. Mraz, W. T. Stam, H. Tice, E. Bornberg-Bauer, P. J. Green, G. A. Pearson, G. Procaccini, C. M. Duarte, J. Schmutz, T. B. H. Reusch, Y. Van de Peer, The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea. *Nature*. **530**, 331–335 (2016).

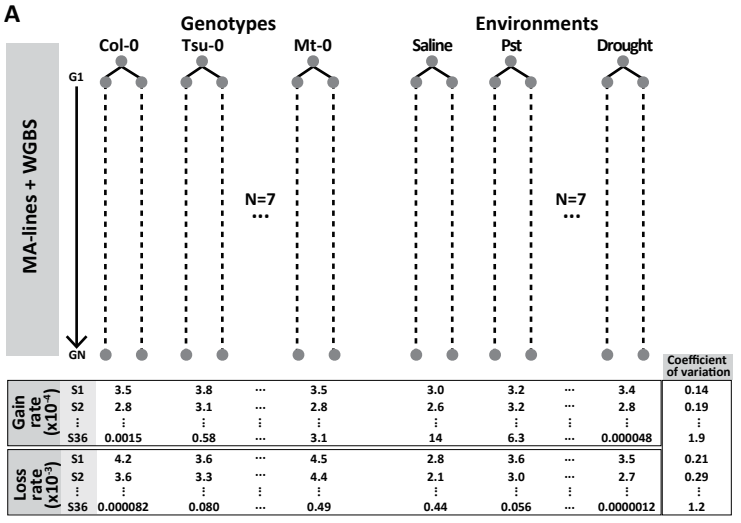
42. L. Yu, C. Boström, S. Franzenburg, T. Bayer, T. Dagan, T. B. H. Reusch, Somatic genetic drift and multilevel selection in a clonal seagrass. *Nat Ecol Evol*. **4**, 952–962 (2020).

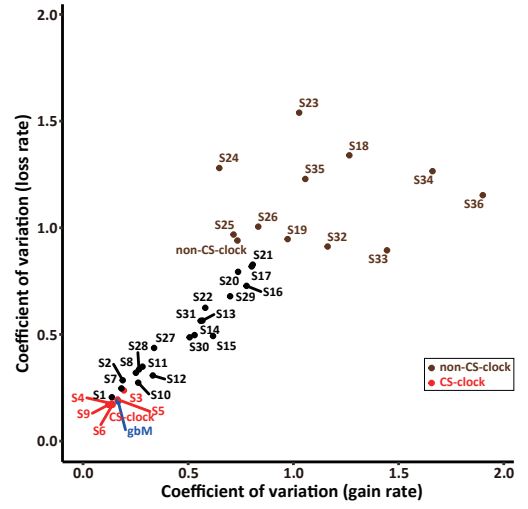
43. L. Yu, J. J. Stachowicz, K. DuBois, T. B. H. Reusch, Detecting clonemate pairs in multicellular diploid clonal species based on a shared heterozygosity index. *Mol. Ecol. Resour.* (2022), doi:10.1111/1755-0998.13736.

44. N. Yao, *Evolutionary epigenetic clock: software code* (2023; https://zenodo.org/record/8109580).

Acknowledgments
We thank Jill Anderson, Andrea Sweigart, Kelly Dyer, James Leebens-Mack and Detlef Weigel for their useful comments, Moi Expósito-Alonso for the SNP data, and Susanne Landis (www.scienstration.com) for her help preparing the figures. **Funding:** TBHR acknowledges support from the HFSP project (PI TBHR; ADAPTASEX; RGP42/2020). RJS acknowledges support from the National Science Foundation (MCB-2242696) and the National Institutes of Health (R01GM134682). FJ acknowledges support from the Deutsche Forschungsgemeinschaft (DFG). FJ RRH and RJS. acknowledge support from the Technical University of Munich Institute for Advanced Study, funded by the German Excellence Initiative and the European Seventh Framework Programme under grant agreement no. 291763. Z.Z. and L.Y. hold fellowships from the China Scholarship Council (no. CSC202006380020) and (no. CSC201704910807), respectively. **Author Contributions:** FJ and RJS conceived the project. NY developed phylogenetic models that incorporate epimutations with support from LL. ZZ analyzed WGBS data and identified clock-like regions. RH, LY, CY and HJ analyzed data. LMS, JT, JS, and RJS produced MA lines. LY and TBHR collected and produced *Z. marina* clones and accompanying WGBS and SNP data. FJ, RJS, ZZ, NY wrote the manuscript with contributions from all authors. **Competing Interests:** None to declare. **Data and materials availability:** The *A. thaliana* sequence data is located at NCBI GEO (MA- accessions - GSE223810 and MA-Pst, MA-SA, and MA-control-Pst&SA - GSE223861). *Z. marina* WGBS sequence data is located at NCBI SRA under BioProject numbers PRJNA933356, PRJNA943354, and PRJNA943356. Code is located via GitHub (https://github.com/schmitzlab/Evolutionary-epigenetic-clock) and Zenodo (https://zenodo.org/record/8109580) (*44*).

Supplementary Materials
Materials and Methods
Supplementary Text
Figs. S1 to S8
Tables S1 to S15
References (*1–55*)





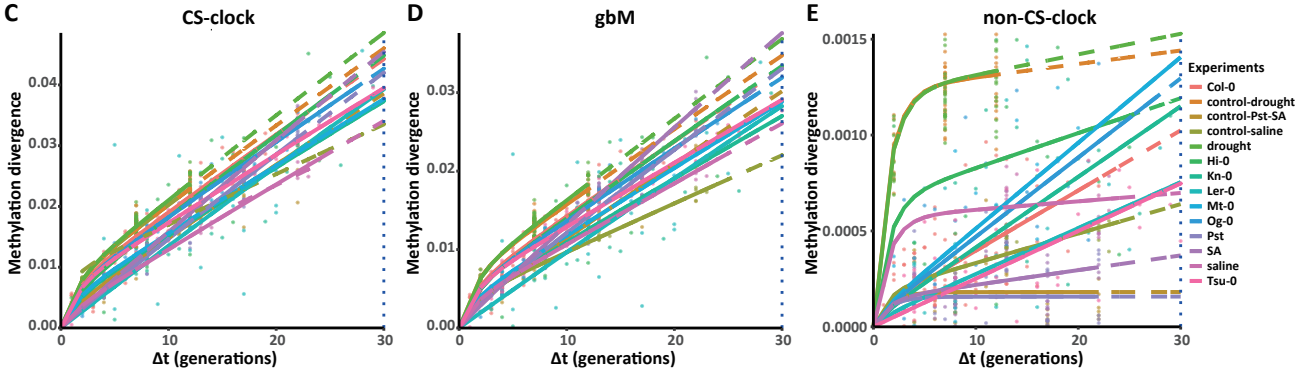


Fig. 1. Discovery of epigenetic clock regions. (A) We sought to identify genomic regions whose CG epimutation rates are invariable across genetic and environmental perturbations. To that end, we created seven MA pedigrees in various genetic backgrounds by using different natural accessions as founders (Col-0, Hi-0, Kn-0, Ler-0, Mt-0, Og-0, and Tsu-0). In addition, we generated MA lines grown under biotic stress (repeated exposure to *Pseudomonas syringae* and salicylic acid) and combined these data with published MA-lines grown under abiotic stress (exposure to high salinity and drought conditions). Using multigenerational WGBS data from these MA lines, we estimated CG epimutation rates in each of 36 annotated Arabidopsis chromatin states (CS). Detailed information about pedigree topologies and WGBS sampling can be found in Fig. S1. For each CS, the coefficient of variation (CV) in the gain and the loss rates were calculated across pedigrees. **(B)** Low CV for the gain and the loss rates were identified in a cluster of CS including S3, S4, S5, S6, and S9. Genomic regions indexed by these CS were defined as epigenetic clock regions (CS-clock). For comparison, a cluster of CS including S18, S19, S23, S24, S25, S26, S32, S33, S34, S35, and S36 displayed high CVs, which were defined as non-clock regions (non-CS-clock). **(C-E)** CG epimutation accumulation is plotted as a function of divergence time (Δt) for each MA pedigrees. For CS-clock regions (C), epimutation accumulation patterns were much more invariable across pedigrees than for non-CS-clock regions (E). Epimutation accumulation in gbM genes (D) closely resembled CS-clock regions, indicating that gbM genes can serve as a proxy for an epigenetic clock.

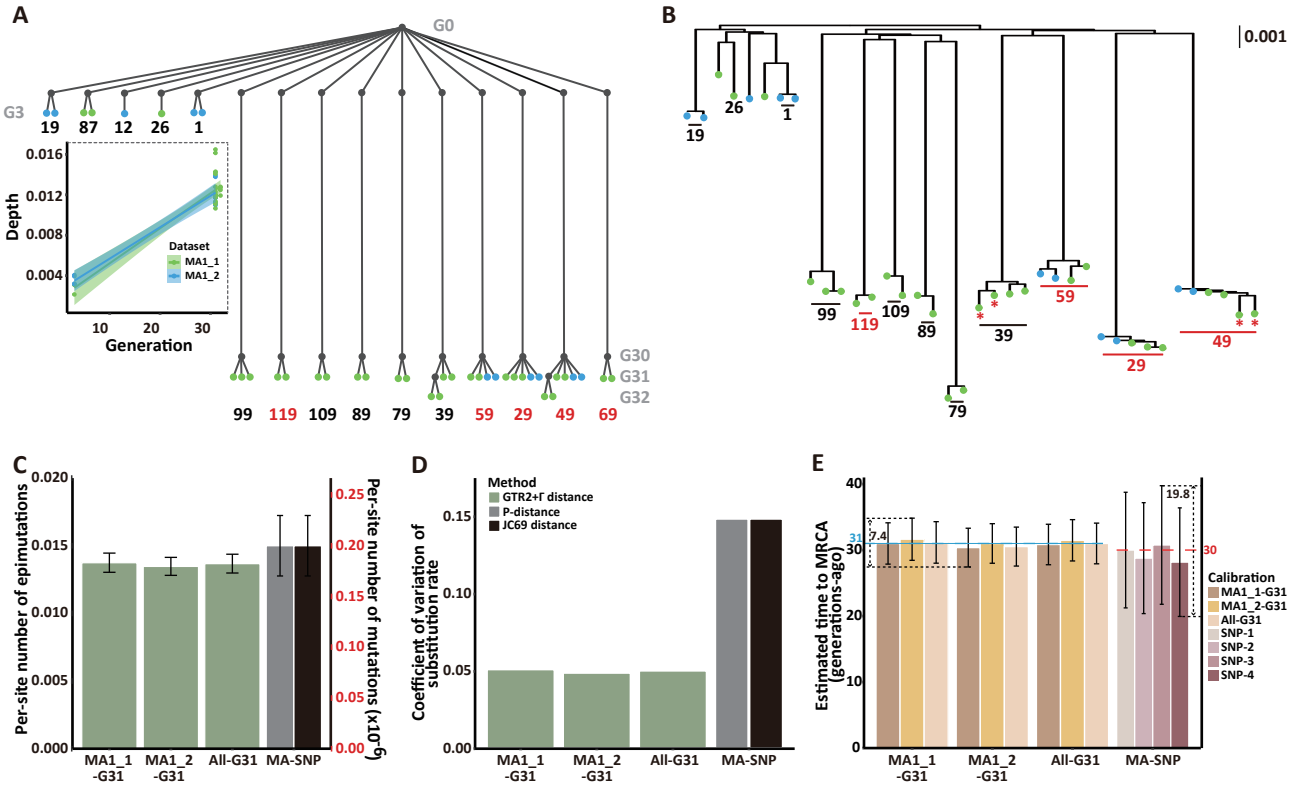


Fig. 2. Evolutionary histories of *A. thaliana* MA lines. (A) The MA lines shared a common ancestor and were maintained by single-seed descent for 3, 31, and 32 generations (G3, G31, and G32). Their DNA methylomes were sequenced by Becker et al. (MA1_1, green dots) and Schmitz et al. (MA1_2, blue dots) separately. In addition, Ossowski et al. sequenced the genomes of five G30 individuals from 29, 49, 59, 69, and 119 (orange, 27). **(B)** We inferred the phylogeny of these MA lines using our epimutation clock. The inferred topology recapitulates the known evolutionary relationships among the sequenced samples. The G32 individuals are marked with a red *. **(C)** We estimated the number of accumulated substitutions on each lineage from the phylogeny (i.e., depth of tip node, Materials and Methods). The error bars represent the standard error of accumulated substitutions per lineage. The average depth of G31 individuals is consistent with the clock-like accumulation of epimutations in the different MA lineages. For comparison, the number of SNPs per G30 lineage from the previous study (27) is also shown. **(D)** The coefficient of variation (CV) of the estimated substitution rate. The CV of the substitution rate from segregating SNPs is significantly higher than the CV of the substitution rate from the epimutation clock. **(E)** With the estimated substitution rates, we inferred the time to the most recent common ancestor (MRCA). The error bars show 95% confidence intervals. The blue line indicates 31 generations. The red dash line indicates 30 generations. The epimutation clock not only estimates the correct time to the MRCA but also shows higher consistency than the times estimated with segregating SNPs. The SNP-1 is the mean of SNP substitution rates that was calibrated from the SNP phylogeny of the MA lines. SNP-2, SNP-3, and SNP-4 are four published estimates of SNP substitution rates (27, 31).

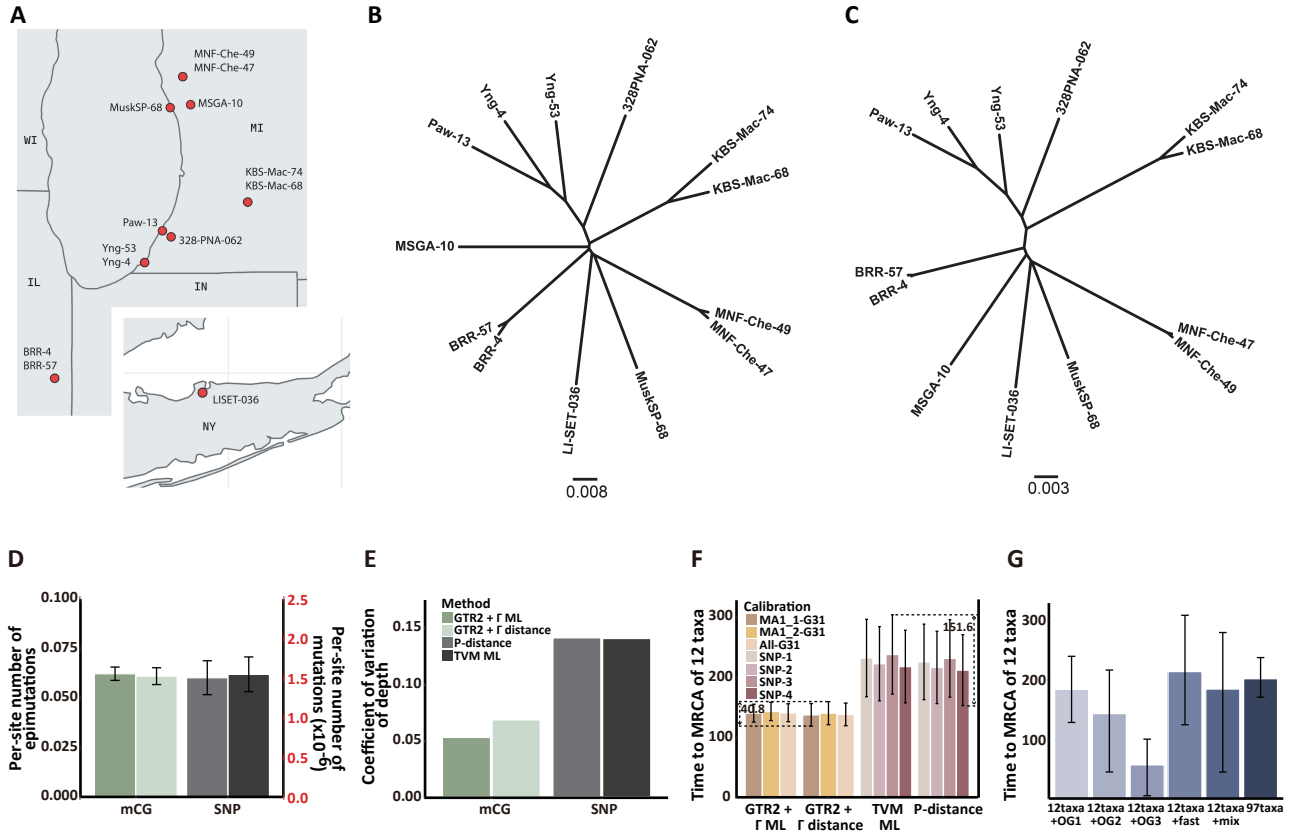


Fig. 3. Evolutionary histories of *A. thaliana* North American accessions. (A) We re-analyzed the DNA methylomes of 13 samples from (30) and used them for phylogenetic inference. The selected samples all belong to a single haplogroup (HPG1, 29, 30). **(B)** The phylogeny was inferred with the same epimutation clock regions that were used in MA1_1 and MA1_2. **(C)** We reanalyzed the published set of segregating SNPs to generate the phylogeny of North American accessions (30). **(D)** The number of substitutions per lineage (± standard error). We removed Yng-53, a hybrid taxon (30, 31), and estimated the number of substitutions on the rest of the 12 lineages. **(E)** The coefficient of variation of depth suggests the SNP-based estimation has higher uncertainty than the estimation based on epimutation clock regions. **(F)** The estimated times between a modern taxon and the most recent common ancestor (MRCA) of 12 taxa. The error bars indicate 95% confidence intervals. All substitution rates of the epigenetic clock, which were calibrated from different data sources, indicate that the MRCA lived in 1860 ± 20 years. However, using four different published substitution rates of SNPs from (30) shows the MRCA lived between 1700 and 1840 (Materials and Methods). **(G)** Dating divergence event of 12 taxa with Bayesian tip calibration. We re-analyzed segregating SNPs from a total of 70 modern and 27 herbarium specimen samples from a previous study (31). In the phylogeny with all 97 samples, the MRCA of 12 taxa lived between 1769 and 1835 (95% high posterior density interval (HPD), mean=202 years ago). However, the phylogenies with only the 12 modern taxa and a part of herbarium specimen samples produced larger HPD intervals. The mean values of tMRCA varied with the sample collection dates and substitution rates on the lineages. Outgroups “OG1”, “OG2”, and “OG3” are specimens that were collected in 19 century, 1900-1950, and 1950-1985. The outgroup

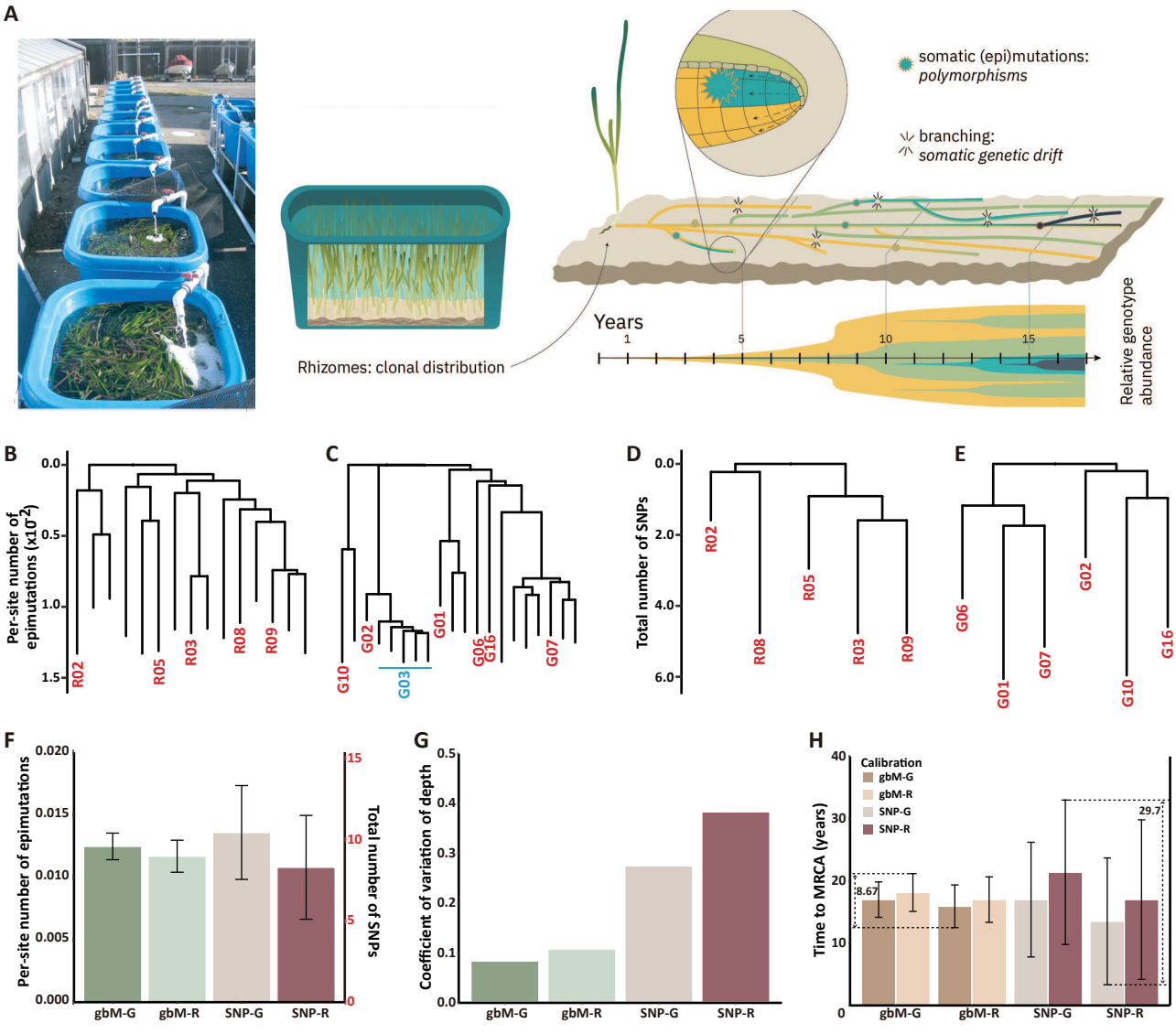


Fig. 4. Evolutionary histories of *Z. marina* experimental clones. (A) Two independent *Z. marina* clonal lines were established from Bodega Harbor in large tanks under ambient light, temperature, and flow-through of ocean water. They were maintained via clonal reproduction for 17 years. **(B-C)** DNA methylomes were sequenced for 15 and 16 ramets of the R and G clones (five technique replicates were generated with the same ramet G03, blue). Using CG sites within gbM genes as a proxy of the epimutation clock, we inferred the phylogeny of the R and G clones separately. We re-sequenced genomes from five samples from the R clone and six samples from the G clone. Only 31 and 47 segregating SNPs were identified. We inferred the phylogeny using the segregating SNPs, and the trees of the R clone and the G clone are shown in **(D)** and **(E)**. **(F-G)**The average depth of two gbM-based trees is 1.246 x 10^{-2} and 1.168 x 10^{-2} with CVs less than 11%. However, the average depth in the SNP trees is much more variable, with CVs are over 27%. **(H)** The estimated times to MRCA were further calibrated with these depths and substitution rates.

Science
AAAS

Supplementary Materials for

An evolutionary epigenetic clock in plants

Yao N[1#], Zhang Z[2#], Yu L[3], Hazarika R[2], Yu C[2], Jang H[1], Smith LM[4], Ton J[4], Liu L[5], Stachowicz J[6], Reusch TBH[3*], Schmitz RJ[1*], Johannes F[2*]

Correspondence to: treusch@geomar.de; schmitz@uga.edu; f.johannes@tum.de

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S8
Tables S1 to S15
References

Materials and Methods

Maximum likelihood estimation of pairwise divergence

Let x and y be a pair of aligned sequences with a common ancestor o. Let the time between o and x be t_1 and the time between time between o and y be t_2. It is assumed that the evolution of each site (i.e., a locus) in the alignment follows an m-state continuous time Markov chain (CTMC), whose state space is $w = \{w_1, w_2, w_3 \ldots w_m\}$. By the forward Kolmogorov equations for the finite state space CTMC, the transition probability matrix can be expressed as $P(t) = e^{Qt}$ for any $t \geq 0$, where Q is the substitution rate matrix. Let $t = t_1 + t_2$, and let i and j be the states of a site in sequence x and y, respectively. When this Markov chain is time reversible, the probability of the site k is

$$P(i, j|t_1, t_2, Q) = P(i, j|t, Q) = \pi_i P_{ij}(t) \quad \textbf{(eq. 1)}$$

where P_{ij} is the ij-th element in the transition probability matrix $P(t)$, and π_i is the equilibrium frequency of state i (*1*). We assume that sites evolve independently and have the same substitution rate. The joint probability of data D is the product of the probabilities for individual sites, i.e.,

$$P(D|t, Q) = \Pi_{i \in w} \Pi_{,j=w} P(i, j|t, Q)^{n_{ij}}$$

where $n_{i,j}$ is the count of the sites whose states are i in the sequence x and j in the sequence y. It follows that the log-likelihood function for these two sequences (*1*) is

$$l(Q, t|D) = \sum_{i \in w} \sum_{j \in w} n_{ij} log\left(P(i, j|t, Q)\right) \quad \textbf{(eq. 2)}$$

The pairwise divergence between two sequences is defined by

$$d = \mu t = -\sum_{i \in w} \pi_i Q_{ii}\, t \quad \textbf{(eq. 3)}$$

where Q_{ii} is a diagonal element in Q. The substitution rate μ in (eq. 3) is the number of substitutions per site per unit time. The maximum likelihood estimates of the pairwise divergence d can be obtained by maximizing the log-likelihood function in (eq. 2).

The substitution rate, though assumed constant, often varies across sites, due to the possible differences in chromatin structure and selection pressures (*2–4*). Let r be the relative substitution rate of a site, and $\phi(r)$ is the probability density function (PDF) of r. Because the rate matrix Q is positive definite, we have $Q = UDU^{-1}$ where U is the matrix of eigenvectors and D is a diagonal matrix in which the diagonal elements are the eigenvalues $\{\lambda_1, \ldots, \lambda_m\}$. Thus, the transition probability matrix is $P(t) = e^{Qt} = UD^*U^{-1}$ where D^* is a diagonal matrix in which the ith diagonal value is equal to $e^{\lambda_i t}$. For each type of observed substitution from i to j, the probability of observing i and j in two sequences at a site is given by

$$P(i,j|t,Q) = \int_0^\infty \pi_i P_{ij}(t \cdot r) \cdot \phi(r)\mathrm{d}r = \pi_i \sum_{k \in w}^{w} u_{ik} u_{kj}^{-1} M_r(\lambda_k t)$$

(eq. 4)

The λ_k is the k-th eigenvalue of Q (i.e., the k-th element in diagonal matrix D). The u_{ik} is the ik-th element in U. The u_{kj}^{-1} is the kj-th element in U^{-1}. The $M_r(\lambda_k t)$ is the moment-generating function (MGF) of $\lambda_k t$, which is defined by

$$M_r(\lambda_k t) = \int_0^\infty e^{\lambda_k t \cdot r}\, \phi(r)\mathrm{d}r$$

(eq. 5)

The continuous gamma model (+Γ, (*2*) assumes the relative substitution rates r among sites can be modeled by a gamma distribution with a probability density function:

$$\phi(r) = g(r;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta r} r^{\alpha-1}$$

(eq. 6)

where $\Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1}\mathrm{d}t$. Thus, the moment generating function is equal to

$$M_r(\lambda_k t) = (1 - \lambda_k t/\beta)^{-\alpha}$$

(eq. 7)

If we set the mean of this rate $E(r) = \frac{\alpha}{\beta} = 1$, we have $\alpha = \beta$ and $Var(r) = \frac{\alpha}{\beta^2} = \frac{1}{\alpha}$. The invariable site + gamma model (I + Γ, (*5*) assumes a fraction of sites (with proportion p_0) are invariant during evolution, the rest of the sites follows a gamma distribution $g(r;\alpha,\beta)$. Thus, a moment-generating function of $\lambda_k t$ is

$$M_r(\lambda_k t) = p_0 + (1 - p_0)(1 - \lambda_k t/\beta)^{-\alpha}$$

(eq. 8)

When we set $E(r) = \frac{(1-p_0)\alpha}{\beta} = 1$, we have $\beta = (1 - p_0)\alpha$.

The three models above are a set of nested models. The continuous I + Γ model can be reduced to a continuous Γ model by fixing the invariable sites' proportion $p_0 = 0$. When $\alpha \to \infty$, the continuous Γ model will converge to the constant rate model. With the new likelihood function for the continuous Γ or I + Γ model, we can find the maximum likelihood estimates of the parameters in the rate matrix Q as well as the pairwise divergence. The theory in this section will be used to derive the substitution models for the evolution of CpG methylation.

A substitution model for methylome evolution in diploid self-fertilizing plants

We model methylome evolution in diploid self-fertilizing plants using a two-state (state space $s = \{UU, MM\}$) continuous-time Markov chain (CTMC), where the state UU is unmethylated homozygous, and MM is methylated homozygous. We do not consider epiheterozygous states UM or MU, because epiheterozygous sites hardly provide any evolutionary information when mutation-drift equilibrium is reached. Now, let $\{X_t\}$ be the two-state continuous-time Markov chain of epigenotypes, where $X_t = 0\ (i.e., UU)$ or 1 (i.e., MM) at time $t \geq 0$. We use the term "substitution" in a broad sense here to denote the transition between

epigenotype states UU and MM. Let A be the substitution rate from state 0 to state 1, and B be the substitution rate from 1 to 0. The rate matrix is given by

$$Q = \begin{bmatrix} -A & A \\ B & -B \end{bmatrix}$$

(eq. 9)

The forward Kolmogorov differential equations indicate that the transition probability matrix is given by

$$P(t) = \begin{bmatrix} P_{00}(t) & P_{01}(t) \\ P_{10}(t) & P_{11}(t) \end{bmatrix} = e^{Qt}$$

(eq. 10)

where $P_{ij}(t) = P(X_t = j | X_0 = i)$ is the probability of state $X_t = j$ at time t given the initial state $X_0 = i$. Moreover, the equilibrium probabilities (π_0, π_1) can be derived by solving the equations

$$\begin{cases} -A\pi_0 + B\pi_1 = 0 \\ A\pi_0 - B\pi_1 = 0, \text{ where } \pi_0 = \frac{B}{A+B} \text{ and } \pi_1 = \frac{A}{A+B}. \\ \pi_0 + \pi_1 = 1 \end{cases}$$

Let μ be the number of substitutions per locus per generation. When the distribution of substitution rates is uniform, with (eq. 3), we have

$$d = \mu t = -\sum_{i \in S} \pi_i Q_{ii} t = \frac{2k}{(1+k)} At.$$

(eq. 11)

where $k = \frac{B}{A}$.

Thus, the transition probability matrix can be expressed as $P = e^{Q^* \tau}$,where $Q^* = \frac{1}{\mu} Q$ is the standardized rate matrix and $\tau = \mu t$ is the number of substitutions per site. When the substitution rates follow the continuous gamma distribution (GTR2 + Γ), we can estimate the pairwise divergence $\hat{d}$ by the observed difference S and a gamma parameter α (estimated separately) between two sequences:

$$\hat{d}(S, \alpha) = \frac{\alpha}{c} \cdot [(1 - cS)^{-\frac{1}{\alpha}} - 1]$$

(eq. 12)

where $c = \frac{(1+k)^2}{2k}$. Let n be the length of the sequence. The large number variance (*6*) and coefficient of variation of d are given by

$$Var(\hat{d}) \approx Var(S) \cdot \left(\frac{\mathrm{d}\hat{d}}{\mathrm{d}S}\right)^2 = \frac{S(1-S)}{n} \cdot (1 - cS)^{-2-\frac{2}{\alpha}}$$

(eq. 13)

$$CV(\hat{d}) = \frac{\sqrt{Var(\hat{d})}}{E(\hat{d})} \approx \frac{\sqrt{\frac{S(1-S)}{n}} \cdot (1 - cS)^{-1-\frac{1}{\alpha}}}{\hat{d}(S, \alpha)}$$

(eq. 14)

When invariable sites are added to this continuous gamma model (eq. 8), we have the GTR2 + I + Γ model. When we set the expectation of the relative substitution rate $E(r) = 1 = (1 - p_0)\alpha/\beta$, the $\beta = (1 - p_0)\alpha$. The estimate of the pairwise divergence is

$$\hat{d}(S, \alpha, p_0) = \frac{\alpha(1 - p_0)}{c} \cdot [\left(1 - \frac{cS}{1 - p_0}\right)^{-\frac{1}{\alpha}} - 1] \quad \textbf{(eq. 15)}$$

where $c = \frac{(1+k)^2}{2k}$. Its variance and coefficient of variation are

$$Var(\hat{d}) \approx Var(S) \cdot \left(\frac{\mathrm{d}\hat{d}}{\mathrm{d}S}\right)^2 = \frac{S(1 - S)}{n} \cdot (1 - \frac{cS}{(1 - p_0)})^{-2-\frac{2}{\alpha}} \quad \textbf{(eq. 16)}$$

$$CV(\hat{d}) = \frac{\sqrt{Var(\hat{d})}}{E(\hat{d})} \approx \frac{\sqrt{\frac{S(1 - S)}{n}} \cdot (1 - \frac{cS}{(1 - p_0)})^{-1-\frac{1}{\alpha}}}{\hat{d}(S, \alpha, p_0)} \quad \textbf{(eq. 17)}$$

As mentioned in the last section, when $p_0 = 0$, the GTR2 + I + Γ model can be reduced to the GTR2 + Γ model. We used a Python script to calculate the these pairwise divergence above, which is available at https://github.com/schmitzlab/Evolutionary-epigenetic-clock.

A substitution model for methylome evolution in diploid clonal plants

To understand the epigenotype dynamics in clonal plants, we proposed a three-state continuous time Markov chain (CTMC) model, referred to as the "baseline model" (Supplementary Text). Let $s = \{UU, \{UM, MU\}, MM\}$ be the state space of this model." UU is unmethylated homozygous and MM is methylated homozygous. In unphased diploid methylomes, the two epiheterozygotes UM and MU are considered as a single state, because they are indistinguishable in short-read WGBS data. We further assume that 1) methylation gain epimutations ($U \to M$) and methylation loss epimutations ($M \to U$) occur spontaneously and independently from each other, and 2) gain and loss epimutations can have different rates, as has been demonstrated in several plant species (*7*). As a result of assumption 1), direct transitions between UU and MM are negligible, because such transitions require two independent methylation or demethylation events on the same locus at the same time. In this section, 0, 1, and 2 denote UU, $\{UM, MU\}$, and MM, respectively.

Let $\{X_t\}$ be the continuous time Markov chain of epigenotypes and $X_t = 0$, 1, or 2 for any time $t \geq 0$. Let a be the number of methylation gain epimutations per CG site per year per haploid methylome. Let b be the number of methylation loss epimutations per CG site per year per haploid methylome. Then, the transition rate matrix of X_t is

$$Q = \begin{bmatrix} -2a & 2a & 0 \\ b & -a - b & a \\ 0 & 2b & -2b \end{bmatrix} \quad \textbf{(eq. 18)}$$

As in the two-state model in the previous section, the transition probability function $P(t) = e^{Qt}$ and the equilibrium frequencies of epigenotypes are given by

$$[\pi_0, \pi_1, \pi_2] = [\frac{b^2}{(a+b)^2}, \frac{2ab}{(a+b)^2}, \frac{a^2}{(a+b)^2}]$$

(eq. 19)

Let μ be the total number of substitutions per year per site per *diploid methylome*. Similarly, with (eq. 3), the value of μ can be calculated from the transition rate matrix Q and the equilibrium frequencies as follows:

$$\mu = -\sum_{i \in s} \pi_i Q_{ii} = \frac{4ab}{a+b}$$

(eq. 20)

For this model, we used a Python script to obtain the maximum likelihood estimation of the pairwise divergence (see eq. 2-8, script available at https://github.com/schmitzlab/Evolutionary-epigenetic-clock). By inputting gamma parameters and proportions of invariable sites, the script can also estimate the pairwise divergence under the gamma model and the invariable site + gamma model.

Inferring Neighbor-joining trees with methylation data

To construct reliable distance-based phylogenetic trees, we used the neighbor-joining method (*8*) and the bootstrap method. Firstly, for each dataset (aligned methylation data or SNP data), we generated 500 bootstrap datasets (*9*). Next, we selected appropriate formulas according to the data type and among-site rate heterogeneity and calculated the divergence between sequences within each bootstrap sample. The neighbor-joining method was implemented with Biopython (*10*) and scikit-bio (*11*). The script is available at https://github.com/schmitzlab/Evolutionary-epigenetic-clock. Then, a total of 500 bootstrap trees were obtained and summarized with Dendropy (*12*) and sumTrees (*13*) for the final consensus tree.

Inferring maximum likelihood tree and rate heterogeneity parameters with IQtree2

We reconstructed maximum likelihood trees with IQtree2 (*14*). For the reliability of results, we used the "-B 5000" option to construct each phylogeny with 5,000 rounds of ultrafast bootstrapping (UFBoot, (*15*). We used the "-m" option to specify the rate-variation-among-site models (Table S6, S11). When inferring the trees, the parameters in the rate-variation-among-site models (proportion of invariable sites and gamma distribution parameter) were estimated simultaneously.

Estimating substitution rate and divergence time from inferred phylogenies

Ohta and Kimura's seminal work laid the foundation for this field (*6*) by demonstrating how to use known divergence times to estimate the mean substitution rates on each lineage from a known phylogenetic tree. Their study also highlighted that the constancy of the substitution rate can be measured by its variance or standard error. Inspired by their study, we estimated the divergence time in two steps. Firstly, we used an inferred tree with known divergence time to estimate the substitution rate. Secondly, we fixed the substitution rate as a constant, and used it to calibrate divergence events in another tree.

For the first step, genetic divergence between the root node (the most recent common ancestor, i.e., MRCA) and a tip node (taxon) can be represented by the path length between them in a phylogenetic tree, which was estimated from CG methylation data. The path length is also denoted by the *depth* of the tip node. Under the assumption of a molecular clock, the n tip nodes should have the same depth, which is denoted by D_1. In real data analysis, the depths $\{d_{1,1}, d_{1,2}, d_{1,3} \ldots d_{1,n}\}$ of the n tip nodes are a random sample from a probability distribution. The expectation, variance, standard error and coefficient of variation of D_1 are $E(D_1)$, $Var(D_1)$, and

$$SE(D_1) = \sqrt{Var(D_1)} \quad \textbf{(eq. 21)}$$

$$CV(D_1) = \frac{\sqrt{Var(D_1)}}{E(D_1)} \quad \textbf{(eq. 22)}$$

For these n tips, the time to their MCRA is equal to a known constant T. We can define the substitution rate on *a lineage* with a random variable $\mu = \frac{D_1}{T}$. The expectation, variance, and coefficient of variation of μ are

$$E(\mu) = \frac{E(D_1)}{T} \quad \textbf{(eq. 23)}$$

$$Var(\mu) = Var(D_1)\left(\frac{\mathrm{d}(\mu)}{\mathrm{d}D_1}\right)^2 = \frac{Var(D_1)}{T^2} \quad \textbf{(eq. 24)}$$

$$CV(\mu) = \frac{\sqrt{Var(\mu)}}{E(\mu)} = \frac{\sqrt{Var(D_1)/T^2}}{E(D_1)/T} = \frac{\sqrt{Var(D_1)}}{E(D_1)} = CV(D_1) \quad \textbf{(eq. 25)}$$

In the second step, we estimated the divergence time in another inferred phylogeny using the substitution rate estimated from the first step. For another rooted phylogeny (or a subtree with common ancestor) with m tips, let $\{d_{2,1}, d_{2,2}, d_{2,3} \ldots d_{2,m}\}$ be the depth of these tips. Similarly, they are a set of samples of random variable D_2. If the substitution rates from two sets of lineages follow the same distribution, we can use the estimated substitution rate to calibrate the time between a pair of nodes, such as between a tip and the root, or between a pair of tips. This idea has been widely utilized in many methods (*6*, *16*, *17*), since the emergence of molecular dating as a field of research.

The estimated time between a tip and the root (i.e., time to MRCA of all tips) can be defined by the random variable

$$t = \frac{D_2}{\bar{\mu}} \quad \textbf{(eq. 26)}$$

where the $\bar{\mu}$ is the sample mean value of substitution rates estimated from the first step (which is a constant instead of a random variable). The expectation, variance, coefficient of variation, and 95% confidence interval are

$$E(t) = \frac{E(D_2)}{\bar{\mu}} \quad \textbf{(eq. 27)}$$

$$Var(t) = Var(D_2)\left(\frac{\mathrm{d}(t)}{\mathrm{d}D_2}\right)^2 = \frac{Var(D_2)}{\bar{\mu}^2} \quad \textbf{(eq. 28)}$$

$$CV(t) = \frac{\sqrt{Var(t)}}{E(t)} = \frac{\sqrt{Var(D_2)/\bar{\mu}^2}}{E(D_2)/\bar{\mu}} = \frac{\sqrt{Var(D_2)}}{E(D_2)} = CV(D_2) \quad \textbf{(eq. 29)}$$

$$CI(t) = E(t) \pm 1.96 \times \sqrt{Var(t)} \quad \textbf{(eq. 30)}$$

We can calculate the values of statistics in (eq. 21-30), by replacing the expectation and variance of D_1 and D_2 with the sample mean and sample variance. In Fig 2C, Fig 3D, Fig 4F, we show the mean values and standard error (the error bars) of "a single taxon's depth" with the sample mean and (eq. 21). In Fig. 2D, Fig 2E, and Fig 4G, the CVs were calculated with (eq. 22) and (eq. 25). In Fig 2E, Fig 3F, and Fig 4H, we showed the mean values and 95% CIs of "time between a taxon and the root" with (eq. 27) and (eq. 30). We implemented the methods above with R scripts which is available at https://github.com/schmitzlab/Evolutionary-epigenetic-clock.

Simulations

To investigate the accumulation of genetic mutations and epimutations in diploid selfing and clonal plants, we generated forward simulation data with a Python script (available at https://github.com/schmitzlab/Evolutionary-epigenetic-clock) and package, simuPop (*18*). We simulated eight sets of mutation accumulation (MA) lines for each propagation type (Table S9-S10). To emulate real-world MA experiments, each set of MA lines shared a single common ancestor, whose genome or methylome were assumed to be at equilibrium. Then, as the progeny of each common ancestor, 50 independent MA lines were simulated with an identical propagation type (selfing or clonal), substitution rates, genome sizes (or the number of CpG sites), and generation time for 500 generations (from G0 to G500, Fig. S5). During the simulation, the generation time was assumed to be constant and discrete, and there is no overlapping between generations. The substitution rates in the simulation were defined by the probability of having a kind of substitution between two generations (see Supplementary Text for details). The GTR2 model (*14*) was employed between every two generations to introduce epimutations into the simulated epimutation clock regions. For the simulated genomes, the distribution of DNA point mutations followed the K80 model (*19*). We established the haploid size of the simulated genome at 100 MB (totaling 2×10^6 nucleotides per diploid genome), a magnitude comparable to the mappable genome size of *A. thaliana* (*20*). Each of the simulated epigenetic clock regions includes 2×10^5 CpG sites for each diploid methylome). This number is as large as 10~20% of the epigenetic clock regions we identified from *A. thaliana* and *Z. marina*. Generally, longer sequences can provide higher statistical robustness (eq. 15-17). This simulation provided insight into the best performance of SNP-based methods and the performance of methylation-based methods in the most disadvantaged situations.

From G0 to G500, we measured the pairwise divergence between each progeny and the common ancestor (i.e., the depth in the tree of life) every ten generations. As phased methylomes are usually not available, we used unphased methylation data from simulated epimutation clock

regions to reflect the performance of epigenetic-clock-based methods. The P-distance, MK distance (*21*), and GTR2 distance were used for the epimutation clock regions from simulated selfing plants. The baseline distance was applied to the epimutation clock regions from simulated clonal plants. For simulated genomes, the SNP data was phased. In each set of the 50 MA lines, the sample average, sample standard deviation, and coefficient of variation (CV) of depth can be calculated as described previously. All calculations were implemented with Python scripts (*10*, *22*), https://github.com/schmitzlab/Evolutionary-epigenetic-clock)

Discovery of epigenetic clock-like regions in *A. thaliana*

To identify genomic regions where epimutations accumulate in a clock-like fashion, we analyzed WGBS data from a total of 14 different *A. thaliana* MA pedigrees (Fig. 1A, Fig. S1). We created seven MA pedigrees in different genetic backgrounds. To that end, we used seven different natural accessions as founders and propagated 2-3 lineages per founder by single seed descent for 17 generations separately (Fig. S1A). The founder accessions were Col-0, Hi-0, Kn-0, Ler-0, Mt-0, Og-0, and Tsu-0. We refer to these MA pedigrees as MA-accessions. For the MA-accessions, seeds were planted and grown in 16-h-day lengths, and samples were collected from young above-ground tissue. Leaf tissue was flash frozen in liquid nitrogen and DNA was extracted using a Qiagen Plant DNeasy kit (Qiagen, Valencia, CA, USA) based on the manufacturer's instructions. To evaluate the impact of environmental factors, we also generated *A. thaliana* MA lines that were grown under multi-generational biotic stress (repeated exposure to *Pseudomonas syringae* or salicylic acid), as well as matched controls. We refer to these MA pedigrees as MA-Pst, MA-SA, MA-control-Pst&SA, respectively. These MA lines were propagated as four lineages per treatment from the same ancestor for 12 generations. The majority of the sequenced samples were derived from G2, G6, and G11 (Fig. S1C). DNA methylation sampling strategy is a sibling design, which means the samples are obtained from siblings of progenitors. The treatments for the samples from G1 to G11 are on leaves at 3, 4, and 5 weeks of age, except for the sequenced samples. DNA was extracted from leaf tissue. Two further MA pedigrees were grown under multi-generational abiotic stress (saline and drought stress), as well as their controls. We refer to these MA pedigrees as MA-saline, MA-drought, MA-control-saline, and MA-control-drought respectively (Fig. S1B, Fig. S1D). Details concerning the construction of these pedigrees and DNA extraction protocols can be found in their original publications (*23*, *24*). We used the above 14 MA pedigrees for the discovery of clock-like regions.

For the MA-accessions, MethylC-seq libraries were prepared based on the protocol described in a previous study (*25*). Libraries were sequenced at Genewiz on a NovaSeq 6000 platform (Illumina) in a paired-end or single-end 150bp format. The sequences of the MA-accessions have been submitted to the GEO repository under number GSE223810. And for the sequences of MA-Pst, MA-SA, and MA-control-Pst&SA have been submitted to the GEO repository under number GSE223861. For MA-saline and MA-control-saline, and MA-drought and MA-control-drought, FASTQ files were downloaded from the GEO (BioProject numbers PRJNA259090 and PRJNA368978 respectively). All WGBS data were processed using the MethylStar pipeline (Shahryary et al., 2020), using TAIR10 (*26*) as a reference. When applicable, different files corresponding to the same sample were merged at the BAM stage using "samtools merge -n" (Samtools version 1.11). For all samples, cytosine-level methylation states were called as either methylated ("M"), unmethylated ("U") or intermediate methylated ("I") using Methimpute (*27*). Methimpute is a hidden markov model with binomial emission

densities. It is capable of making accurate methylation calls even for cytosines with missing or low read counts. Another advantage of Methimpute is its ability to identify not only epihomozygotous unmethylated (UU) or epihomozygous methylated sites (MM), but also epiheterozygous sites (MU). Knowledge of epiheterozygous sites is absolutely necessary when studying epimutation accumulation in clonal species / lineages, as most epimutations are of the form UU <-> MU, or MM <-> MU.

It is important to note that Methimpute actually reduces to a standard binomial model in samples with large read coverage, which is the case in the present study. In such situations, Methimpute does not impute anything, but just performs binomial calling similar to the standard binomial model that has been employed in many other studies (e.g. Becker et al.. 2011, Schmitz et al. 2011, van der Graaf et al. 2015). The only difference is that Methimpute makes use of its underlying hidden markov structure for statistical inference. To demonstrate this latter point we here conducted a side-by-side comparison between Methimpute and the standard binomial model using the MA-pedigree "MA-accession Col-0" as an example. Focusing on the CS clock-regions, we found that Methimpute's methylation calls are highly correlated with those from the standard binomial model (corr = 0.98). Further, we used the CG methylation state calls obtained from the standard Binomial model to estimate epimutation rates in "MA-accession Col-0". We found that the estimated gain and loss rates are remarkably close to each other (Table S4). In this table, "Original_two-state" and "Using_binom_test" are based on two-state methylation calls (UU and MM) only, as epiheterozygotes (MU) cannot be called with the standard binomial model. The omission of the MU state in these approaches could account for the slight downward bias in the estimated epimutation rates when compared with the estimates provided in "Original".

The genome coordinates of the 36 chromatin states (CS) were retrieved from the Plant Chromatin State Database (*28*). Using these coordinates, we partitioned the methylomes of the MA-pedigrees into different CS. Following Hazarika et al. (*29*), we estimated global as well as CS-specific CG methylation gain rates (α) and loss rates (β) in all 14 MA-pedigrees separately. Estimation was performed using the R package AlphaBeta (version 1.10.0, (*7*). In all cases, we fit a neutral model (ABneutral) to the data. Having obtained estimates of the gain rate and the loss rate for the 36 CS in each MA pedigree, we calculated the coefficient of variation (CV) of these rates across pedigrees. The CV for a given CS is defined as the ratio of the population standard deviation to the population mean. During calculation, the population standard deviation and the population mean were replaced with their unbiased estimates, sample standard deviation and sample mean from each pedigree. A high CV value for a given CS, would indicate that epimutation rates in that particular CS are susceptible to either environmental and/or genetic perturbations. Conversely, a low CV value for a given CS would indicate that epimutation rates in that particular CS are robust. The CV values obtained for the gain and loss rates are displayed in Fig. 1B, Fig. S2B, and Table S1. We observed a cluster of five CS that showed low CV values for both gain and loss rates. These CS were CS3, CS4, CS5, CS6 and CS9. Genomic regions indexed by these CS were selected and combined to define epigenetic clock-like regions. The rest of the genomic regions were defined as "non-CS-clock".

We explored the annotation enrichment in CS corresponding to clock-like regions. To this end, annotation files for genes and TEs in gff3 format were downloaded from Ensembl Plants (http://plants.ensembl.org/info/data/ftp). The list of gbM genes was obtained from the Supplementary Data 3 in the previous work (*30*).

Experimental calibration of an epimutation clock in *A. thaliana*

For clock calibration, we used WGBS data from MA pedigrees MA1_1 and MA1_2 (*31*, *32*), Fig. 2A). These MA lines were propagated by single-seed descent for 30 generations and are the largest experimental MA system currently available in *A. thaliana* (*33*). Detailed descriptions of the growth conditions and plant material of MA1-1 and MA1-2 can be found in the original publications (*31–33*). FASTQ files of MA1_1 and MA1_2 was downloaded from GEO (BioProject number PRJNA271082).

We processed all WGBS data with the MethylStar pipeline (*34*), using TAIR10 as a reference (*26*). For all samples, cytosine-level methylation states were called as either methylated ("M"), unmethylated ("U") or intermediate methylated ("I") using a three-state Hidden Markov Model (*27*). Given that most epimutations are neutral (*7*), and the MA lines had been maintained by single-seed selfing propagation for multiple generations, we assumed that the *A. thaliana* methylome is at (epi)mutation-drift equilibrium. The number of fixed epimutations per locus per generation should be approximately equal to the number of epimutations per CG sites per generation per haploid methylome. Thus, homozygous sites are sufficient to infer evolutionary histories. We selected all CG sites within the clock-like regions from the MA 1_1 and MA 1_2 WGBS datasets and removed epiheterogeneous sites. In total, 452,341 epihomozygous sites were used for phylogenetic inference. Epigenotype calls with posterior probability lower than 0.8 or coverage below 1 were marked as missing data.

We merged the MA1_1 and MA1_2 pedigrees, as they are derived from the same founder individual. We assumed all CG sites on the epimutation clock regions evolved at the same rate. Then we constructed a neighbor-joining tree (*8*) for all individuals based on the GTR2 distance (Fig 2B). We used individuals from G3 as the outgroups in the inferred phylogenies. The inferred phylogeny provided important information for data cleaning. 1) For the G3 lineages, the number of fixed epimutations is significantly higher than that in other lineages. This is likely due to the unfixed epiheterozygous sites in the methylome of their common ancestor. Moreover, three generations are not sufficiently long enough to reach mutation-drift equilibrium. 2) Line 79 had a much higher depth than other G31 individuals. A previous study observed the line 69 has more epimutations as well, which might be related to a SNP in a protein-coding gene (*31*). However, the mechanism behind the higher depth observed in line 79 is unknown and needs further study.

We excluded G3 and line 79 individuals and used the rest of G31 and G32 in subsequent analysis. Based on a series of rate heterogeneity models, we used IQtree2 (*14*) to estimate the maximum likelihood trees (ML trees) and the parameters. Then, we substituted these parameters into distance calculation formulas, and obtained a series of distance-based phylogenetic trees. Thus, each ML tree has a corresponding distance tree that is biologically equivalent to it. We set the midpoint as the root of each phylogeny. Using these phylogenies, we estimated the substitution rates on each G31 lineage from MA1_1 or MA1_2 with the methods mentioned in the previous sections (eq. 21-25, Materials and Methods). Based on the mean values of per-lineage substitution rates from each data source, we further estimated the times to the MRCA on each lineage (eq. 26-30, Materials and Methods).

Inferring divergence times in *A. thaliana* natural populations with an epimutation clock

We analyzed published WGBS data from 13 different North American accessions, which were collected in the field from locations around the Great Lakes and the East Coast of the USA. Details concerning these samples, including DNA extraction protocols can be found in the original paper (*35*). For the North American accessions, FASTQ files were downloaded from the

European Nucleotide Archive under accession number PRJEB5331. All WGBS data were processed using the MethylStar pipeline (*34*), using TAIR10 as a reference (*26*). For all samples, cytosine-level methylation states were called as either methylated ("M"), unmethylated ("U") or intermediate methylated ("I") using a three-state Hidden Markov Model (*27*).

To conduct a preliminary analysis of the evolutionary history among the 13 accessions (taxa), we used the method previously used to explore the relationship between MA1_1 and MA1_2. Specifically, assuming a constant evolutionary rate across all CG sites, we constructed the phylogenetic trees using the same epimutation clock regions, GTR2 distance formula and neighbor-joining method. We set their midpoints as the root, as the geological outgroup from Long Island (NY) is not significantly diverged from the other samples.

In previous studies, Yng-53 was identified as a hybrid accession (*35*, *36*). Therefore, we removed it from the dataset and only focused on the remaining 12 non-recombinant accessions. For the phylogenetic trees, we applied the same set of methods that we used on G31 individuals in the MA lines. The midpoint of each inferred phylogeny was set as the root. Using the methods we described in previous sections, we estimated the number of substitutions on each lineage (eq. 21-25) and further estimated time between the root and each taxon from each inferred phylogeny (eq. 26-30, Materials and Methods).

Inferring evolutionary histories with *A. thaliana* SNP data

We analyzed SNP data from the same *A. thaliana* MA lines (*20*) and North American accessions (*35*, *36*). For MA lines, the genomes of five (29, 49, 59, 69, and 119) G30 individuals were collected and sequenced (*20*), Fig. 2A). We denoted this dataset by "MA-SNP". We used the observed number of SNPs per site (P-distance) and the JC69 distance to construct the neighbor-joining trees (see previous sections for details of neighbor-joining tree and bootstrap methods).

For 13 samples from Hagmann et al. (*35*), we explored their evolutionary relationships with a neighbor-joining tree based on P-distance between their segregating SNP sites (Fig. 2C). The P-distance NJ tree were constructed with MEGA (*37*). For the 12 non-recombinant taxa, we further reconstructed their NJ tree with P-distance. Also, we inferred maximum likelihood trees from segregating SNPs with IQtree2 (*14*). We allowed IQtree2 to optimize the best-fit model by using the option "-m MFP". Each of these phylogenies was rooted with the mid-point. For each tip in the phylogenies, we measured its depth and estimated its time to the root with the same methods we introduced in previous sections (Fig. 2D-F).

We also applied Bayesian tip-calibration to date the divergence event of these 12 taxa using BEAST (*38–40*). We used the SNP data collected from modern and herbarium specimen samples together with their collection dates from a previous publication (*36*). Using SNP data and sample collection dates as input, BEAST can output a Bayesian phylogeny and times of divergence events (i.e., internal nodes) inside the phylogeny. We applied tip-calibration to 97 taxa from the original dataset, three taxa were removed as the collection dates were missing in the original paper's supplementary information. We set the same priors and the models that were used in the original paper (Table. S12). We used these 12 taxa and different outgroups as input of Bayesian tip-calibration to determine the time of the MRCA of the 12 non-recombinant modern taxa (Table. S12). To avoid having the Bayesian model converge to an incorrect tree topology, we set these 12 taxa as a mono-clade in the prior. To ensure the tree priors (which are related to the population demographic histories) does not influence the results, we used the "skygrid" model in the prior, which was applied in the previous study (*36*). Then, we confirmed the

analysis by changing the tree prior model into “constant population size”. For each combination of input datasets (12 taxa and a set of outgroups) and priors, we launched two independent BEAST runs to guarantee the MCMC samplings sufficiently converged. With BUEAti (*39*), we generated the XML files that included the SNP data and setting as the direct input for BEAST software. The chain lengths of the MCMC sampling were fixed at one billion. During MCMC sampling, BEAST randomly attempted possible phylogenies and model parameters, and computed the corresponding posterior probabilities. Every 10,000 iterations, the phylogeny and parameters were stored in the output files. We trimmed the first 10% of samples as burn-in and used the rest of samples for the following analysis. The model parameters were summarized with Tracer (*40*). The phylogenies were summarized with treeAnnotator (*39*) for the maximum clade credibility trees, in which the node heights were defined with time of the MRCA. We extracted the mean and 95% highest posterior density intervals of the 12 taxa’s divergence times from the summarized results (Table S12, Fig. 2G).

Identifying gbM genes in *Z. marina*

Chromatin states have not been defined for *Z. marina.* Therefore, we used CpG sites from gene–body methylated (gbM) genes, where epimutations that behave in a clock-like manner are highly enriched. We extracted 20,244 genes that are less than 10kb in the *Z. marina* genome (Zmarina-668-v2.0) using the latest version of gene annotation (Zmarina-668-v3.1, available at https://phytozome-next.jgi.doe.gov). The gbM gene identification followed a previously established pipeline (*30*). The thresholds of P-value and Q-value were set as 0.05. Minor modifications included processing of WGBS reads with ‘paired-end-pipeline’ function with ‘trim-reads’ option in Methylpy (*41*). This pipeline allowed us to extract a list of gbM genes from the WGBS reads that were generated from a single sample.

To minimize the misidentification and influence from environmental factors or genotypes, we used *Z. marina* WGBS data from three data sources across the East Pacific Ocean to the North Atlantic Ocean. We denoted them by “data source 1”, “data source 2”, and “data source 3”. The data source 1 included 23 WGBS datasets, which were collected from a *Z. marina* natural clone from Ängsö Island, Finland, covering an area of at least 300m x 200m (data source 1). This clone was used in a previous study (*42*).

Data source 2 had 24 WGBS samples from 21 ramets. Three clones with a branching history of 4 years: Three small patches of *Z. marina* were collected from Kiel, Germany. Each patch was assumed to be derived from a single seedling based on visual observation after excavating them via SCUBA diving. Their size of between 20 and 30 leaf shoots and could be reached by rhizome branching (4-5 branching events) within one year. Hence, we assumed that the branching history before sample collection was 1 year. They were then transferred to GEOMAR Helmholtz Center for Ocean Research Kiel and cultured in in ambient flowing seawater from Kiel Fjord in the indoor "Zosteratron" set up under a natural time course of water temperatures and light conditions, in ambient sediment collected close to the plants. All tanks had a wave generator at a frequency of about 0.5 Hz. Leaf shoots including the meristematic regions were taken for tissue samples after they had been cultured for 3 years. The total branching history is thus 4 years.

In data source 3, two ramets have been sampled from Bodega Harbor, CA, United States in 2004. They had been used as the founder of two clones. After 17 years of growth in lab conditions, a total of 40 WGBS samples were collected and sequenced.

We identified a list of possible gbM genes for every single sample using the pipeline mentioned above (Fig. S6A). The gbM lists from the same data source show substantial overlap. In samples from Ängsö Island, Finland (data source 1), 3,290 genes were identified as gbM genes in 90% of the samples (i.e., in at least 21 samples). With the same cut-off, 3,474 genes and 3,573 genes were identified in 90% of samples within data source 2 and data source 3, respectively (Fig. S6B). We defined the intersection of these three lists of gbM genes as "core gbM gene list" and used them in the following analyses (Fig. S6C, Table S14).

Experimental calibration of an epigenetic evolutionary clock in *Z. marina*

We studied two clones with a branching history of 17 years (referred to as clone R and clone G, Fig. 4A): The two clones with a branching history of 17 years were the same as those previously studied by Yu et al. (*43*). Each clone was initiated by a single leaf shoot (*44*). WGBS was performed for 15 ramets of clone R, as one ramet (including 5 technical replicates of WGBS samples, which are R10_01, R10_02, R10_03, R10_04, and R10_05) did not belong to this clone, it was removed from later analysis (*43*). They were cultured at Bodega Bay Marine lab in outdoor tanks in ambient sediment and under flowing seawater under a natural time course of water temperatures and light conditions. Any flowering shoots observed were removed to maintain the clonal lineage. Leaf shoots including the meristematic regions were taken for tissue samples after they had been cultured for 17 years.

The protocol for pre-processing of samples and DNA extraction was same with the previous study (*42*). WGBS libraries were constructed by BGI (Beijing Genomics Institute) and then sequenced on a NovaSeq 6000 platform (Illumina) in a paired-end 150bp format. All *Z. marina* WGBS data were processed using the MethylStar pipeline (*34*), using *Z. marina* (version 3.1, (*45*) as a reference. Summary statistics, including sequencing depth, mapping efficiency, bisulfite conversion using the lambda spike-in control and coverage for each sample are found in Table S13. For all samples, cytosine-level methylation states were called either methylated ("M"), unmethylated ("U"), or intermediate methylated ("I") using a three-state Hidden Markov Model (Taudt et al. 2018). All *Z. marina* WGBS sequence data have been submitted to the SRA repository under BioProject numbers PRJNA933356, PRJNA943354, and PRJNA943356.

As the "baseline model" was not well-supported by software based on maximum likelihood or Bayesian methods, we used the distance-based approach, which we have applied to *A. thaliana* epimutation data, to construct the evolutionary tree of *Z. marina*. The first step of this method is estimating the rate heterogeneity parameters (such as gamma parameter α and invariable sites p_0). The second step is using the estimated parameters and sequence data as input and calculate the pairwise distances, which can further be used in neighbor-joining tree construction.

For the first step, IQtree2 can estimate the rate heterogeneity parameters with the naïve Bayesian method by using the "--rate" option (*14*). We finally decided to use only the gamma corrected baseline model (baseline $+\Gamma$) that we introduced in previous sections (eq. 2-8, eq. 18, Materials and Methods). We concerned about if the proportion of invariable sites p_0 in the gamma invariable sites model (i.e., $+\mathrm{I}+\Gamma$) can be accurately estimated from molecular data. This concern has been raised since it the model was proposed (*5*), and this question seems remained open after decades of debating (*1*, *46–49*). Also, in *A. thaliana* MA1_1 and MA1_2, the estimated number of epimutations per lineages given by gamma invariable sites models and gamma models didn't show large differences (Supplementary Text). For the second step, we generated 500 bootstrap datasets for each methylome from two 17-years old clones. Each

bootstrap dataset contained the same number of CpG sites as the original dataset. We inferred the pairwise divergence for every pair of methylomes in each bootstrap dataset based on the estimated parameters and the rate-variation-among-sites models (baseline +Γ) introduced above (eq. 2-8, eq. 18, Materials and Methods). We obtained the neighbor-joining trees for the bootstrap samples as well as the summarized phylogenies (see previous section for details, Materials and Methods). We extracted the depth of each taxon from the summarized phylogenies (for details of methods please refer to the previous sections, Materials and Methods, Fig 4F). We further evaluated the variations of the depth (Fig. 4G) and calibrated the time between each taxon and its clone (Fig. 4H, Table S12).

Inferring evolutionary histories of *Z. marina* clones from segregating SNPs

The genomic data available for the two 17-year-old clones were also analyzed (*43*), including 6 ramets for clone G and 5 ramets for clone R. We used a two-step strategy to generate a multi-sample alignment file. 1. Identifying fixed SNPs using Mutect2 (*50*). Mutect2 requires a "normal" sample and a "tumor" sample for each run, and detects the somatic mutations in the "tumor" sample. Our goal was to detect the fixed somatic mutations, which was visualized by a histogram of variant read frequency. Since mosaic mutations and fixed mutations overlapped at the low-frequency region, we only focused on the somatic mutation with variant read frequency >= 0.5, which represented around half of the total number. The collection of the SNPs across all Mutect2 runs was used for further analysis. 2. Obtaining multi-sample genotypes using GATK4. We conducted joint SNP calling on all samples using GATK4. The multi-sample genotypes for the target SNPs were extracted, based on which a multi-sample alignment file was constructed.

Supplementary Text

Relationship between GTR2 model and baseline model

Here we prove that (eq. 18) can be obtained from the GTR2 model (*14*). On a haploid methylome, each CpG site can either be methylated (M) or unmethylated (U). The GTR2 model is sufficient for describing the evolution of a haploid methylome with transition probability matrix:

$$P(t) = e^{Q_{GTR2}t}, \text{ where } Q_{GTR2} = \begin{bmatrix} -a & a \\ b & -b \end{bmatrix} \quad \textbf{(eq. 31)}$$

, where a and b have the same definition as in (eq. 18). By solving $[\pi_U, \pi_M] \cdot Q_{GTR2} = 0$, the equilibrium frequencies of unmethylated CG sites and methylated CG sites are

$$[\pi_U, \pi_M] = [\frac{b}{a+b}, \frac{a}{a+b}] \quad \textbf{(eq. 32)}$$

Also, with (eq. 3) we have the expected total number of methylation gain and methylation loss epimutations per year per CG site is

$$\mu_{GTR2} = \frac{2ab}{a+b} \quad \textbf{(eq. 33)}$$

In a diploid genome, we can have the following possible epigenotypes at a locus: unmethylated homozygous (UU), two kinds of epiheterozygous (UM and MU) and methylated homozygous

(MM). In a diploid methylome from a clonal plant, and assuming independence between epialleles, the transition among these four epigenotypes can be represented with a continuous-time Markov chain model too. Let this continuous-time Markov chain be $X = \{X_n\}$ with state space $s = \{UU, UM, MU, MM\}$ and transition probability matrix $P(t) = e^{Qt}$. The Q can be derived from Q_{GTR2} (eq. 31).

$$Q = \begin{bmatrix} -2a & a & a & 0 \\ b & -a-b & 0 & a \\ b & 0 & -a-b & a \\ 0 & b & b & -2b \end{bmatrix} = Q_{GTR2} \oplus Q_{GTR2}$$

(eq. 34)

The $\oplus$ here indicates the Kronecker sum. For epigenotypes I and $J \in \{UU, UM, MU, MM\}$, the number of transitions in unit time from I to J is the off-diagonal element $Q_{I,J}$ in Q. For four possible states (four epigenotypes), the vector of equilibrium frequencies is:

$$[\pi_{UU}, \pi_{UM}, \pi_{MU}, \pi_{MM}] = \left[\frac{b}{a+b}, \frac{a}{a+b}\right] \otimes \left[\frac{b}{a+b}, \frac{a}{a+b}\right] = \left[\frac{b^2}{(a+b)^2}, \frac{ab}{(a+b)^2}, \frac{ab}{(a+b)^2}, \frac{a^2}{(a+b)^2}\right]$$

(eq. 35)

The $\otimes$ here indicates the Kronecker product. With (eq. 3), the expected total number of methylation gain and methylation loss epimutations per diploid methylome per locus per year is

$$\mu = -\sum_{I \in s} \pi_I Q_{I,I} = \frac{4ab}{a+b}$$

(eq. 36)

The model above cannot be applied to unphased methylomes, as the states UM and MU are indistinguishable from each other. To solve this problem, we can re-partition the state space s into three subsets $E_1 = \{UU\}, E_2 = \{UM, MU\}$ and $E_3 = \{MM\}$. They correspond to three epigenotypes: methylated homozygous, epi-heterozygous, and the unmethylated homozygous. Then we have a new set:

$$\tilde{s} = \{E_1, E_2, E_3\} = \{\{UU\}, \{UM, MU\}, \{MM\}\}$$

(eq. 37)

Let $\tilde{X} = \{\widetilde{X_n}\}$ be a continuous-time Markov chain with state space $\tilde{s} = \{E_1, E_2, E_3\}$. The $\tilde{X}$ is called a "lumped chain" of the Markov chain that we defined in (eq. 34). The $\tilde{s}$ is called a "partition" of the states in s (*51*, *52*). Let U be the 3×4 matrix whose i-th row is the probability vector having equal components for states in E_i and 0 for remaining states. Let V be the 4×3 matrix whose j-th column is a vector with 1's in the corresponding to states in E_j and 0's otherwise (*51*). Then we have

$$U = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, V = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(eq. 38)

Let the transition probability function of $\tilde{X}$ be $P(t) = e^{\tilde{Q}t}$. The transition rate matrix $\tilde{Q}$ of $\tilde{X}$ satisfies (*51*, *52*):

$$\tilde{Q} = UQV = \begin{bmatrix} -2a & 2a & 0 \\ b & -a-b & a \\ 0 & 2b & -2b \end{bmatrix}$$

(eq. 39)

It finally leads us to the transition rate matrix that we defined for unphased diploid methylomes from clonal plants in (eq. 18). Let $\tilde{\pi}$ be the vector of equilibrium frequencies of E_1, E_2 and E_3. It can be obtained by adding the from the equilibrium frequencies of four states in state space $s = \{UU, UM, MU, MM\}$ by adding corresponding components in the same subset of partition $\tilde{s}$ (*52*). Thus, we have

$$\tilde{\pi} = [\pi_{UU}, (\pi_{UM} + \pi_{MU}), \pi_{MM}] = [\frac{b^2}{(a+b)^2}, \frac{2ab}{(a+b)^2}, \frac{a^2}{(a+b)^2}]$$

(eq. 40)

It is easy to check that $\tilde{\pi} \cdot \tilde{Q} = 0$. The per year evolutionary rate of this model can be obtained with (eq. 3). Namely,

$$\tilde{\mu} = -\sum_{K \in \tilde{s}} \tilde{\pi}_K \tilde{Q}_{K,K} = \frac{4ab}{a+b}$$

(eq. 41)

Please note that the unit of $\tilde{\mu}$ and μ is the number of methylation gain and methylation loss epimutations per *diploid methylome* per locus per year. However, for the GTR2 model, the unit of μ_{GTR2} is the number of methylation gain and methylation loss epimutations per *haploid methylome* per CpG site (locus) per year.

Continuous-time Markov chain model and discrete-time Markov chain model for epimutations

The accumulation of mutations over time is a key concept in evolutionary genetics. It can be described by a Markov chain model. In the first section of Materials and Methods, we introduced modeling substitution with continuous time Markov chain models (CTMC models, eq. 1-8), which are more frequently used for phylogenetics. However, in some cases, such as generating forward-time simulation data, discrete-time Markov chain models (DTMC models) are widely used too (*18*, *53*). Here, we will show the DTMC models that we used in simulation have equivalent biological meaning with our CTMC models that we used in phylogenetic analysis. For the same divergence time, their transition probability matrices have the same values. Also, they have the same equilibrium frequencies for all states.

Please note again the following assumptions: 1) all methylation gain and loss epimutations are spontaneous and independent, 2) the rate of methylation gain and loss epimutations are different. For an unmethylated CG site, denote the probability of methylation gain epimutations in consecutive generations by α. For a methylated CG, denote the probability of a methylation loss epimutation in consecutive generations by β. The two assumptions above lead us to a DTMC model for haploid methylomes. This model and GTR2 model (eq. 31) should have the same biological meaning. Let

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

(eq. 42)

be the homogeneous one-step transition probability matrix. The n-step transition probability matrix $P^{(n)}$ of this new model is: $P^{(n)} = P^n$ with $n = 0, 1, 2$ ....The equilibrium frequencies $\pi' = [\pi_U', \pi_M']$ should satisfy $\pi' \cdot P = \pi'$ and $\pi_U' + \pi_M' = 1$. As the result,

$$\pi' = [\pi_U', \pi_M'] = \left[\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right] \quad \textbf{(eq. 43)}$$

If and only if

$$\frac{\alpha}{\beta} = \frac{a}{b} = k \quad \textbf{(eq. 44)}$$

We have $\pi' = [\pi_U', \pi_M'] = \pi = [\pi_U, \pi_M]$, where $\pi = [\pi_U, \pi_M]$ are equilibrium frequencies of GTR2 model in (eq. 32). Also, if

$$a + b = -log\,(1 - \alpha - \beta) \quad \textbf{(eq. 45)}$$

we have $P^n = e^{Q_{GTR2}t}$ when $t = n$. For two haploid methylomes with divergence time T, let be $S_{I,J} = \pi_I \cdot P_{I,J}(T)$ be the probability of observing a substitution from I to J at a site. When (eq. 44) and (eq. 45) hold, (eq. 31) and (eq. 42) always can lead us to the same value of $S_{I,J}$.

In any diploid methylome from a clonal plant, we further assume that spontaneous epimutations occurring on homologous chromosomes are independent from each other. This leads us to a DTMC model for phased diploid methylomes whose 1-step transition probability matrix is

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \otimes \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

The $\otimes$ here indicates the Kronecker product, i.e.

$$P = \begin{bmatrix} (1-\alpha)^2 & (1-\alpha)\alpha & (1-\alpha)\alpha & \alpha^2 \\ (1-\alpha)\beta & (1-\alpha)(1-\beta) & \alpha\beta & (1-\beta)\alpha \\ (1-\alpha)\beta & \alpha\beta & (1-\alpha)(1-\beta) & (1-\beta)\alpha \\ \beta^2 & (1-\beta)\beta & (1-\beta)\beta & (1-\beta)^2 \end{bmatrix} \quad \textbf{(eq. 46)}$$

The equilibrium frequencies are

$$\pi' = [\pi_{UU}', \pi_{UM}', \pi_{MU}', \pi_{MM}'] = \left[\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right] \otimes \left[\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right]$$

Namely,

$$\pi' = \left[\frac{\beta^2}{(\alpha+\beta)^2}, \frac{\alpha\beta}{(\alpha+\beta)^2}, \frac{\alpha\beta}{(\alpha+\beta)^2}, \frac{\alpha^2}{(\alpha+\beta)^2}\right] \quad \textbf{(eq. 47)}$$

We have a CTMC model defined in (eq. 34), whose transition rate matrix was denoted by Q. With (eq. 44) and (eq. 45), we can make $P^n = e^{Qt}$ for any $t = n$. It means the DTMC model defined in (eq. 46) has the same biological meaning as the CTMC model defined in (eq. 34), when (eq. 44) and (eq. 45) hold. Again, if the two epi-heterozygous states are not differentiated (unphased diploid methylome), the state space of (eq. 46) can be divided into three subsets. The new space state is $\bar{s} = \{E_1, E_2, E_3\} = \{\{UU\}, \{UM, MU\}, \{MM\}\}$. According to the definition of

lumping (*51*), the $\tilde{P}$ is a lumped 1-step transition probability matrix of P defined in (eq. 46) with partition $\bar{s}$. Thus, we have:

$$\tilde{P} = UPV = \begin{bmatrix} (1-\alpha)^2 & 2(1-\alpha)\alpha & \alpha^2 \\ (1-\alpha)\beta & (1-\alpha)(1-\beta)+\alpha\beta & (1-\beta)\alpha \\ \beta^2 & 2(1-\beta)\beta & (1-\beta)^2 \end{bmatrix},$$

(eq. 48)

where U and V are same with what are in (eq. 39). The equilibrium frequencies are

$$\widetilde{\pi}' = [\pi_{UU}', (\pi_{UM}' + \pi_{MU}'), \pi_{MM}'] = [\tfrac{\beta^2}{(\alpha+\beta)^2}, \tfrac{2\alpha\beta}{(\alpha+\beta)^2}, \tfrac{\beta^2}{(\alpha+\beta)^2}].$$

When we have (eq. 44) and (eq. 45), the DTMC model defined by (eq. 48) corresponds to the "baseline model" we proposed in (eq. 39). The model in (eq. 48) has recently been used for epimutation rates estimation in the software package AlphaBeta (*7*). This establishes a connection between our current work and previous studies.

Accuracies of estimating pairwise divergence on genomes and methylomes from selfing plants

In this section, we illustrate that estimating pairwise divergence from CG methylation is more robust than using DNA point mutations for recent evolutionary histories of selfing-plants. We further show that estimating pairwise divergence from DNA sequence data can be variable over short timescales, as too few SNPs will have accumulated. To see this, assume SNPs on a pair of DNA sequences follows the JC69 model (*54*). Let $\hat{p}$ be the observed divergence (P-distance) between them. The estimated genetic distance under JC69 model $\hat{d}_{JC69}$ (eq 21) is a function of $\hat{p}$ (*54*):

$$\hat{d}_{JC69} = -\frac{3}{4}\log\left(1-\frac{4}{3}\hat{p}\right)$$

(eq. 49)

The sequence length is denoted by n. The large sample variance of $\hat{d}_{JC69}$ (eq. 50) was derived by Kimura and Ohta in 1972 (*55*):

$$Var\left(\hat{d}_{JC69}\right) = \left[\frac{\mathrm{d}\hat{d}_{JC69}}{\mathrm{d}\hat{p}}\right]^2 \cdot Var(\hat{p}) = \frac{9\hat{p}(1-\hat{p})}{(3-4\hat{p})^2 n} = SE\left(\hat{d}_{JC69}\right)^2$$

(eq. 50)

where the variance of $\hat{p}$, $Var(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}$, and $SE\left(\hat{d}_{JC69}\right)$ is the standard error of $\hat{d}_{JC69}$.

We can reduce the GTR2 model to the MK model for binary traits (equivalent to JC69 model for binary traits, (*21*). We have the estimates of pairwise divergence under the MK model

$$\hat{d}_{MK} = -\frac{1}{2}\cdot\log\left(1-2S\right)$$

(eq. 51)

Let the number of sites be n. The large sample variance of $\hat{d}_{MK}$ is

$$Var(\hat{d}_{MK}) = \frac{S(1-S)}{n(1-2S)^2} = SE(\hat{d}_{MK})^2$$

(eq. 52)

The $SE(\hat{d}_{MK})$ is the standard error of $\hat{d}_{MK}$. Since epimutation rate is roughly 10,000 times higher than DNA point mutation rate, for the same divergence time, $\hat{d}_{GTR2}$ and $\hat{d}_{JC69}$ are not in the same scale. Comparing them with their standard errors or variances is unreasonable. The coefficient of variation (CV) of $\hat{d}_{JC69}$ is

$$CV(\hat{d}_{JC69}) = \frac{SE(\hat{d}_{JC69})}{\hat{d}_{JC69}} = -\frac{4}{3}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \cdot \frac{1}{(1-\frac{4}{3}\hat{p})\log(1-\frac{4}{3}\hat{p})}$$

(eq. 53)

Similarly, we have

$$CV(\hat{d}_{MK}) = \frac{SE(\hat{d}_{MK})}{\hat{d}_{MK}} = \sqrt{\frac{S(1-S)}{n(1-2S)^2}} \cdot \frac{-2}{log(1-2S)}$$

(eq. 54)

From the (eq. 55) and (eq. 56), we can see both the observed divergence and the length of the sequences can influence the accuracy of estimation.

To make this more concrete, consider the following example: For two individuals that diverged from each other 500 generations ago, if we collected SNPs from 100MB DNA sequences and 300,000 CG sites from each individual. Let the observed divergence among the DNA sequences be $10^{-8} \times 1000 = 10^{-5}$ and observed divergence on methylome be $10^{-4} \times 1000 = 10^{-1}$. The CV of $\hat{d}_{mk}$ is about 0.685%. However, the CV of $\hat{d}_{JC69}$ is over 3.16%. In our example, for $\hat{d}_{mk}$, the upper bound of 95% confidence intervals is only 1.34% higher than its mean (for the lower bound, it's 1.34% lower than the mean). However, the upper bound (95% CI) of $\hat{d}_{JC69}$ is about $1.96 \times 3.16\% = 6.19\%$ higher than the mean of $\hat{d}_{JC69}$. Hence, inferences of evolutionary histories on a scale of about 500 years will be more accurate when using CpG methylation data than with SNP data.

Accuracies of estimating pairwise divergence on genomes and methylomes from clonal plants

We will show that using CpG methylation data from clonal plants to infer pairwise divergence is more robust than the traditional SNP-based method. In this section, we denote three kinds of epigenotypes UU, $\{UM, MU\}$, and MM by 0, 1 and 2. For $i, j \in \{0,1,2\}$, let $S_{i,j}$ be observed transitions from i to j. For the substitution model we defined for unphased diploid methylome from clonal plants in (eq 2), we have $S_{i,j} = \pi_i P_{i,j}(t)$, where π_i is equilibrium frequency of epigenotype i, $P_{i,j}(t)$ is the transition probability function from epigenotype i to j. Let $S_1 = S_{0,1} + S_{1,0}$ be the observed transitions between UU and $\{UM, MU\}$. Similarly, let $S_2 = S_{0,2} + S_{2,0}$ and $S_3 = S_{1,2} + S_{2,1}$ be observed proportion of transitions for $UU \leftrightarrow MM$, and $\{UM, MU\} \leftrightarrow MM$.

Thus, we have $\begin{cases} S_1 = c(1-x)(x+k) \\ S_2 = \frac{c}{2}(1-x)^2 \\ S_3 = c(1-x)(x+\frac{1}{k}) \end{cases}$, where $c = \frac{4k^2}{(1+k)^4}$, $x = e^{-(1+k)at}$. Same with definitions in previous sections, the number of methylation events per site per haploid methylome per generation is denoted by a. The ratio $k = \frac{b}{a}$, where b is the number of demethylation events per site per haploid methylome per generation. Again, to simplify the proof and demonstrating our basic ideas, we assume rates of methylation and demethylation are the same, i.e., $k = \frac{b}{a} = 1$.

Now, the expected observed transitions should follow $\begin{cases} S_1 = S_3 = \frac{1}{4}(1-e^{-4at}) \\ S_2 = \frac{1}{8}(1-e^{-2at})^2 \end{cases}$. Also, we have the expected divergence per locus per diploid methylome $d = \frac{4ab}{a+b}t = 2at$. Thus, the expected divergence can be either estimated from S_1 or S_2 with

$$\hat{d}(S_1) = -\frac{1}{2}\log(1-4S_1) \quad \textbf{(eq. 55)}$$

$$\hat{d}(S_2) = -\log(1-(8S_2)^{\frac{1}{2}}) \quad \textbf{(eq. 56)}$$

Like last section, the coefficient of variation of $\hat{d}(S_1)$ is

$$CV\left(\hat{d}(S_1)\right) = \frac{\mathrm{d}(\hat{d}(S_1))}{\mathrm{d}(S_1)} \cdot \sqrt{Var(S_1)} \cdot \frac{1}{\hat{d}(S_1)} = \frac{2}{1-4S_1} \cdot \sqrt{\frac{(1-S_1)S_1}{n}} \cdot \frac{-2}{\log(1-4S_1)} \quad \textbf{(eq. 57)}$$

For $\hat{d}(S_2)$, the coefficient of variation is

$$CV\left(\hat{d}(S_2)\right) = \frac{d(\hat{d}(S_2))}{d(S_2)} \cdot \sqrt{Var(S_2)} \cdot \frac{1}{\hat{d}(S_2)} = \frac{2}{\left(-4S_2 + (2S_2)^{\frac{1}{2}}\right)} \cdot \sqrt{\frac{(1-S_2)S_2}{n}} \cdot \frac{-1}{\log(1-(8S_2)^{\frac{1}{2}})} \quad \textbf{(eq. 58)}$$

Like the example in the last section, we assume there are two individuals that diverged from each other 500 years ago. From each individual, 100 MB DNA sequence and 300,000 CG sites per haploid methylome (300,000 epigenotypes) were collected. When the DNA mutation rate is 10^{-8} per site per year, the observed DNA sequence divergence is approximately equal to $10^{-8} \times 500 \times 2 = 10^{-5}$. With the relative standard error function of JC69 model, the CV of $\hat{d}_{JC69}$ is over 3.16%. When the epimutation rate is 10^{-4} per CG site per year, the expected S_1 should be 0.0824, and the expected S_2 should be 0.00411. The $CV\left(\hat{d}(S_1)\right) = 0.749\%$ and $CV\left(\hat{d}(S_2)\right) = 1.573\%$. Thus, over short timescales, estimating divergence based on CG methylation data is far more robust than using DNA sequence data.

Simulation results

As we have derived in the previous section (eq. 14, eq. 17, eq. 55-56, eq. 59-60), the coefficient of variation (CV) of pairwise divergence was approximately inversely proportional to

the square root of sequence length. When the data type and substitution rates were constant, longer sequences would always lead to smaller variations. However, even when the simulated genomes had longer sequences (over 300 times longer), in our simulation results (Table S9-S10, Fig. S5-S7), regardless of the propagation types, the CV of depth observed from epimutations was smaller than that observed from SNPs.

Epimutation accumulation in *A. thaliana* MA1_1 and MA1_2

In order to accurately infer the evolutionary histories of MA1_1 and MA1_2 from the epimutations, we utilized maximum likelihood and distance-based methods, as well as a range of different models to construct the phylogenetic trees. We found that although the distance-based method appears to be more straightforward, the epimutation number per lineage estimated by both methods were very similar in terms of mean and standard error. The mutual corroboration between these two methods gave us increased confidence in the results. Moreover, it suggested the reliability of the distance-based method, which ultimately led us to choose it for the data analysis of epimutation data from *Z. marina*.

Furthermore, when we took into account the variation in substitution rate among sites, the depth and epimutation rates estimated by different models were very similar (Table S6). This implied that although the $\mathrm{I} + \Gamma$ model had one more parameter than the gamma model (i.e., the proportion of invariable sites, p_0), this additional parameter did not seem to produce different results. Given the debate about the $\mathrm{I} + \Gamma$ model mentioned earlier (Materials and Methods), we finally chose to use the gamma model. As with most maximum likelihood-based software, IQtree2 utilized a discrete gamma distribution instead of a continuous gamma distribution. This well-known approximation method divides all sites into n categories with equal proportions based on their evolutionary rates, and further estimates the gamma distribution parameter (α(*4*). When the number of categories (n) is sufficiently large, the discrete gamma distribution will converge the continuous gamma distribution. Thus, we chose $n = 12$ and used the corresponding estimated value of α in the continuous gamma distance to reconstruct the distance-based phylogenetic tree of the G31 individuals. The depth extracted from this tree were presented in the main text (Fig. 2C). Additionally, the epimutation rates of the G31 individuals were also estimated from this phylogenetic tree.

Divergence time of 12 non-recombinant *A. thaliana* North American accessions

To determine the time of divergence event of 12 non-recombinant taxa from Hagmann et al. (*35*), we used three different strategies: 1) Dating divergence event with average depth and epimutation rates from MA lines (Fig. 3D-F). 2) Dating divergence event with average depth and the substitution rates of SNPs from MA lines (Fig. 3D-F). 3) Dating the divergence event with Bayesian tip-calibration and herbarium samples (Fig 3G).

For the first strategy, we also observed that both the maximum likelihood and distance-based methods, as well as different rate-variation-among-sites models, provided us with very similar depth estimates (Table S11). For the same reasons, we selected the estimated α and phylogeny generated by the gamma model with the largest number of rate categories ($n = 12$) from all outputs of IQtree2. Then, we showed the subsequent results in main text (Fig. 3D-F).

For the second strategy, we mentioned it gave us wider CIs of time to the common ancestor of 12 taxa. We also found that the divergence times obtained by this strategy appeared

to be larger than those obtained by the first strategy. The previous study (*35*) has proposed a hypothesis that positive selection leads to higher substitution rates in the genomes of wild populations than those measured in the laboratory (*20*). Therefore, the second strategy could lead to an overestimation of divergence times. However, as claimed in the original paper (*35*), this hypothesis was difficult to verify.

Regarding the third strategy, firstly, performing Bayesian tip-calibration is never easy. This was because this method relies on the collection time of the samples and sequence divergence between the samples and the common ancestor (*38*, *39*). When the strict molecular clock is still valid, the earlier collected samples should be less diverged from their common ancestor on the sequences. However, our previous theory showed that this correlation might become blurred when the timescale was too small (fewer mutations led to a larger coefficient of variation). We found that when the sample size was very large, the 95% highest posterior density (95% HPD) intervals of divergence times obtained were relatively small (Table S12). However, when we used different DNA samples from the specimens as outgroups for the 12 taxa, the mean divergence times of the 12 taxa shifted. Moreover, the HPD intervals of divergence times also increased. This further revealed the complexity of molecular dating (Table S12). We still needed more molecular evidence and advanced methods to infer the divergence times of these 12 taxa.

gbM genes as a proxy for clock-like regions on *A. thaliana* methylomes

We demonstrated epimutation rate estimation and divergence time estimation using CpG sites on gbM genes from *A. thaliana*. We used a similar strategy, which had been applied to chromatin-state clock regions. We constructed the maximum likelihood trees with IQtree2 and "GTR2+G12" model for G31 (line 79 was excluded as an outlier), and G32 samples from *A. thaliana* MA1_1 and MA1_2. Based on the depth from G31 samples, the average epimutation rates in MA1_1 and MA1_2 are 2.78×10^{-4} and 2.71×10^{-4}, respectively (Fig. S4, Table S8). With two samples T-test, epimutation rates on these two data sources are not significantly different from each other (P-value = 0.41). We applied the rate from MA1_2 gbM genes to MA1_1. For G31 samples from MA1_1, their estimated time to MRCA is 31.9 ± 3.43 generations (95% CI), which agrees well with actual value of 31 generations.

Using the average epimutation rate (2.76×10^{-4}) from all G31 samples, we estimated the divergence time of 12 non-recombining taxa from North American accessions (Fig. S4, Table S8). The MRCA for these taxa was inferred to date back to the year 1872±10.71 (95% CI), which is close to the CS-clock estimate obtained in our manuscript.

To investigate if the epimutation accumulation on each CpG site was influenced by the nearby CpG sites, we took 1 CG site in every 10 CG sites (size of neighborhood = 10) to generate subsets of the original dataset. This operation was replicated thrice, focusing on the 1st, 5th, and 10th CpG sites within each identified neighborhood (refer to Figure S4, Table S8). Assuming that the epimutations occurring in the clock-like regions are independent and follow the same distribution, we would expect the number of accumulated epimutations to be similar across these subsets (i.e. not sensitive to the location in each neighborhood). Intriguingly, our new findings based on these subsets align precisely with this expectation, revealing no significant differences in the estimated epimutation rates.

1. Z. Yang, *Molecular Evolution: A Statistical Approach* (Oxford University Press, 2014).

2. J. Sullivan, K. E. Holsinger, C. Simon, Among-site rate variation and phylogenetic analysis of 12S rRNA in sigmodontine rodents. *Mol. Biol. Evol.* **12**, 988–1001 (1995).

3. Z. Yang, Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994).

4. Z. Yang, S. Kumar, Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. *Mol. Biol. Evol.* **13**, 650–659 (1996).

5. X. Gu, Y. X. Fu, W. H. Li, Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol.* **12**, 546–557 (1995).

6. T. Ohta, M. Kimura, On the constancy of the evolutionary rate of cistrons. *J. Mol. Evol.* **1**, 18–25 (1971).

7. Y. Shahryary, A. Symeonidi, R. R. Hazarika, J. Denkena, T. Mubeen, B. Hofmeister, T. van Gurp, M. Colomé-Tatché, K. J. F. Verhoeven, G. Tuskan, R. J. Schmitz, F. Johannes, AlphaBeta: computational inference of epimutation rates and spectra from high-throughput DNA methylation data in plants. *Genome Biol.* **21**, 260 (2020).

8. N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).

9. J. Felsenstein, Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*. **39**, 783–791 (1985).

10. P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M. J. L. de Hoon, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. **25**, 1422–1423 (2009).

11. T. S.-B. D. Team, *scikit-bio: A Bioinformatics Library for Data Scientists, Students, and Developers* (2020; http://scikit-bio.org).

12. J. Sukumaran, M. T. Holder, DendroPy: a Python library for phylogenetic computing. *Bioinformatics*. **26**, 1569–1571 (2010).

13. J. Sukumaran, M. T. Holder, *SumTrees: phylogenetic tree summarization* (2015; https://github.com/jeetsukumaran/DendroPy).

14. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

15. B. Q. Minh, M. A. T. Nguyen, A. von Haeseler, Ultrafast Approximation for Phylogenetic Bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).

16. E. Zuckerkandl, L. Pauling, Evolutionary Divergence and Convergence in Proteins. *Evolving Genes and Proteins* (1965), pp. 97–166.

17. W. H. Li, M. Tanimura, P. M. Sharp, Rates and dates of divergence between AIDS virus nucleotide sequences. *Mol. Biol. Evol.* **5**, 313–330 (1988).

18. B. Peng, M. Kimmel, simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*. **21**, 3686–3687 (2005).

19. M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).

20. S. Ossowski, K. Schneeberger, J. I. Lucas-Lledó, N. Warthmann, R. M. Clark, R. G. Shaw, D. Weigel, M. Lynch, The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science*. **327**, 92–94 (2010).

21. P. O. Lewis, A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925 (2001).

22. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*. **17**, 261–272 (2020).

23. C. Jiang, A. Mithani, E. J. Belfield, R. Mott, L. D. Hurst, N. P. Harberd, Environmentally responsive genome-wide accumulation of de novo Arabidopsis thaliana mutations and epimutations. *Genome Res.* **24**, 1821–1829 (2014).

24. D. R. Ganguly, P. A. Crisp, S. R. Eichten, B. J. Pogson, The Arabidopsis DNA Methylome Is Stable under Transgenerational Drought Stress. *Plant Physiol.* **175**, 1893–1912 (2017).

25. M. A. Urich, J. R. Nery, R. Lister, R. J. Schmitz, J. R. Ecker, MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat. Protoc.* **10**, 475–483 (2015).

26. P. Lamesch, T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez, A. S. Karthikeyan, C. H. Lee, W. D. Nelson, L. Ploetz, S. Singh, A. Wensel, E. Huala, The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2012).

27. A. Taudt, D. Roquis, A. Vidalis, R. Wardenaar, F. Johannes, M. Colomé-Tatché, METHimpute: imputation-guided construction of complete methylomes from WGBS data. *BMC Genomics*. **19**, 1–14 (2018).

28. Y. Liu, T. Tian, K. Zhang, Q. You, H. Yan, N. Zhao, X. Yi, W. Xu, Z. Su, PCSD: a plant chromatin state database. *Nucleic Acids Res.* **46**, D1157–D1167 (2018).

29. R. R. Hazarika, M. Serra, Z. Zhang, Y. Zhang, R. J. Schmitz, F. Johannes, Molecular properties of epimutation hotspots. *Nat Plants*. **8**, 146–156 (2022).

30. Y. Zhang, H. Jang, R. Xiao, I. Kakoulidou, R. S. Piecyk, F. Johannes, R. J. Schmitz, Heterochromatin is a quantitative trait associated with spontaneous epiallele formation. *Nat. Commun.* **12**, 6958 (2021).

31. C. Becker, J. Hagmann, J. Müller, D. Koenig, O. Stegle, K. Borgwardt, D. Weigel, Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. *Nature*. **480**, 245–249 (2011).

32. R. J. Schmitz, M. D. Schultz, M. G. Lewsey, R. C. O'Malley, M. A. Urich, O. Libiger, N. J. Schork, J. R. Ecker, Transgenerational epigenetic instability is a source of novel methylation variants. *Science*. **334**, 369–373 (2011).

33. R. G. Shaw, D. L. Byers, E. Darmo, Spontaneous mutational effects on reproductive traits of arabidopsis thaliana. *Genetics*. **155**, 369–378 (2000).

34. Y. Shahryary, R. R. Hazarika, F. Johannes, MethylStar: A fast and robust pre-processing pipeline for bulk or single-cell whole-genome bisulfite sequencing data. *BMC Genomics*. **21**, 1–8 (2020).

35. J. Hagmann, C. Becker, J. Müller, O. Stegle, R. C. Meyer, G. Wang, K. Schneeberger, J. Fitz, T. Altmann, J. Bergelson, K. Borgwardt, D. Weigel, Century-scale methylome stability in a recently diverged Arabidopsis thaliana lineage. *PLoS Genet.* **11**, e1004920 (2015).

36. M. Exposito-Alonso, C. Becker, V. J. Schuenemann, E. Reiter, C. Setzer, R. Slovak, B. Brachi, J. Hagmann, D. G. Grimm, J. Chen, W. Busch, J. Bergelson, R. W. Ness, J. Krause, H. A. Burbano, D. Weigel, The rate and potential relevance of new mutations in a colonizing plant lineage. *PLoS Genet.* **14**, e1007155 (2018).

37. K. Tamura, G. Stecher, S. Kumar, MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.* **38**, 3022–3027 (2021).

38. A. Rieux, F. Balloux, Inferences from tip-calibrated phylogenies: a review and a practical guide. *Mol. Ecol.* **25**, 1911–1924 (2016).

39. M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, A. Rambaut, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol*. **4** (2018), doi:10.1093/ve/vey016.

40. V. Hill, G. Baele, Bayesian estimation of past population dynamics in BEAST 1.10 using the Skygrid coalescent model. *Mol. Biol. Evol.* **36**, 2620–2628 (2019).

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

42. L. Yu, C. Boström, S. Franzenburg, T. Bayer, T. Dagan, T. B. H. Reusch, Somatic genetic drift and multilevel selection in a clonal seagrass. *Nat Ecol Evol*. **4**, 952–962 (2020).

43. L. Yu, J. J. Stachowicz, K. DuBois, T. B. H. Reusch, Detecting clonemate pairs in multicellular diploid clonal species based on a shared heterozygosity index. *Mol. Ecol. Resour.* (2022), doi:10.1111/1755-0998.13736.

44. A. R. Hughes, J. J. Stachowicz, S. L. Williams, Morphological and physiological variation among seagrass (Zostera marina) genotypes. *Oecologia*. **159**, 725–733 (2009).

45. X. Ma, J. L. Olsen, T. B. H. Reusch, G. Procaccini, D. Kudrna, M. Williams, J. Grimwood, S. Rajasekar, J. Jenkins, J. Schmutz, Y. Van de Peer, Improved chromosome-level genome assembly and annotation of the seagrass, Zostera marina (eelgrass). *F1000Res.* **10**, 289 (2021).

46. J. S. Rogers, Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. *Syst. Biol.* **50**, 713–722 (2001).

47. J. Chai, E. A. Housworth, On Rogers' proof of identifiability for the GTR + Γ + I model. *Syst. Biol.* **60**, 713–718 (2011).

48. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*. **14**, 587–589 (2017).

49. L.-T. Nguyen, A. von Haeseler, B. Q. Minh, Complex Models of Sequence Evolution Require Accurate Estimators as Exemplified with the Invariable Site Plus Gamma Model. *Syst. Biol.* **67**, 552–558 (2018).

50. D. Benjamin, T. Sato, K. Cibulskis, G. Getz, C. Stewart, L. Lichtenstein, Calling Somatic SNVs and Indels with Mutect2. *bioRxiv* (2019), p. 861054.

51. J. G. Kemeny, J. L. Snell, *Finite Markov chains* (Springer, 1976).

52. J. P. Tian, D. Kannan, Lumpability and Commutativity of Markov Processes. *Stoch. Anal. Appl.* **24**, 685–702 (2006).

53. X. Yuan, D. J. Miller, J. Zhang, D. Herrington, Y. Wang, An overview of population genetic data simulation. *J. Comput. Biol.* **19**, 42–54 (2012).

54. T. H. Jukes, C. R. Cantor, "Evolution of Protein Molecules" in *Mammalian Protein Metabolism*, H. N. Munro, Ed. (Academic Press, 1969), pp. 21–132.

55. M. Kimura, T. Ota, On the stochastic model for estimation of mutational distance between homologous proteins. *J. Mol. Evol.* **2**, 87–90 (1972).

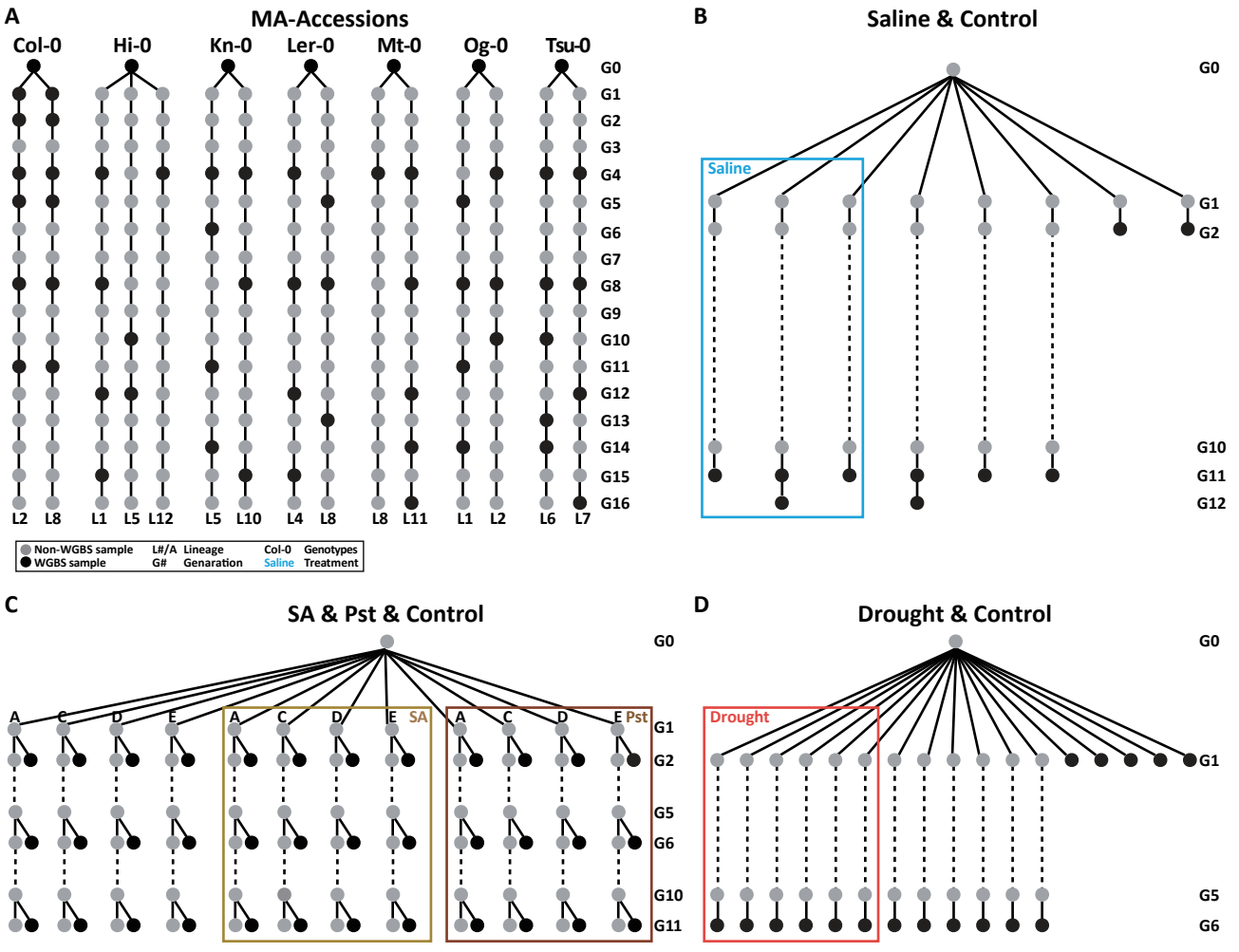


Fig. S1. Pedigrees of *A. thaliana* MA-lines. (A) The pedigrees of MA-Accessions from seven different natural accessions (i.e. genotypes) **(B-D)** MA-lines under multi-generational biotic (C) or abiotic (B and D) stress and their corresponding controls. The non-WGBS samples (grey circle) inside the rectangles are treated with stress, but the WGBS samples (black circle) are not, other samples outside the rectangles are treated as controls. A and C are newly produced data, whereas B and D are from previously published studies (B(23); D(24). B, C, and D were derived from a single Columbia (Col-0) genotype.

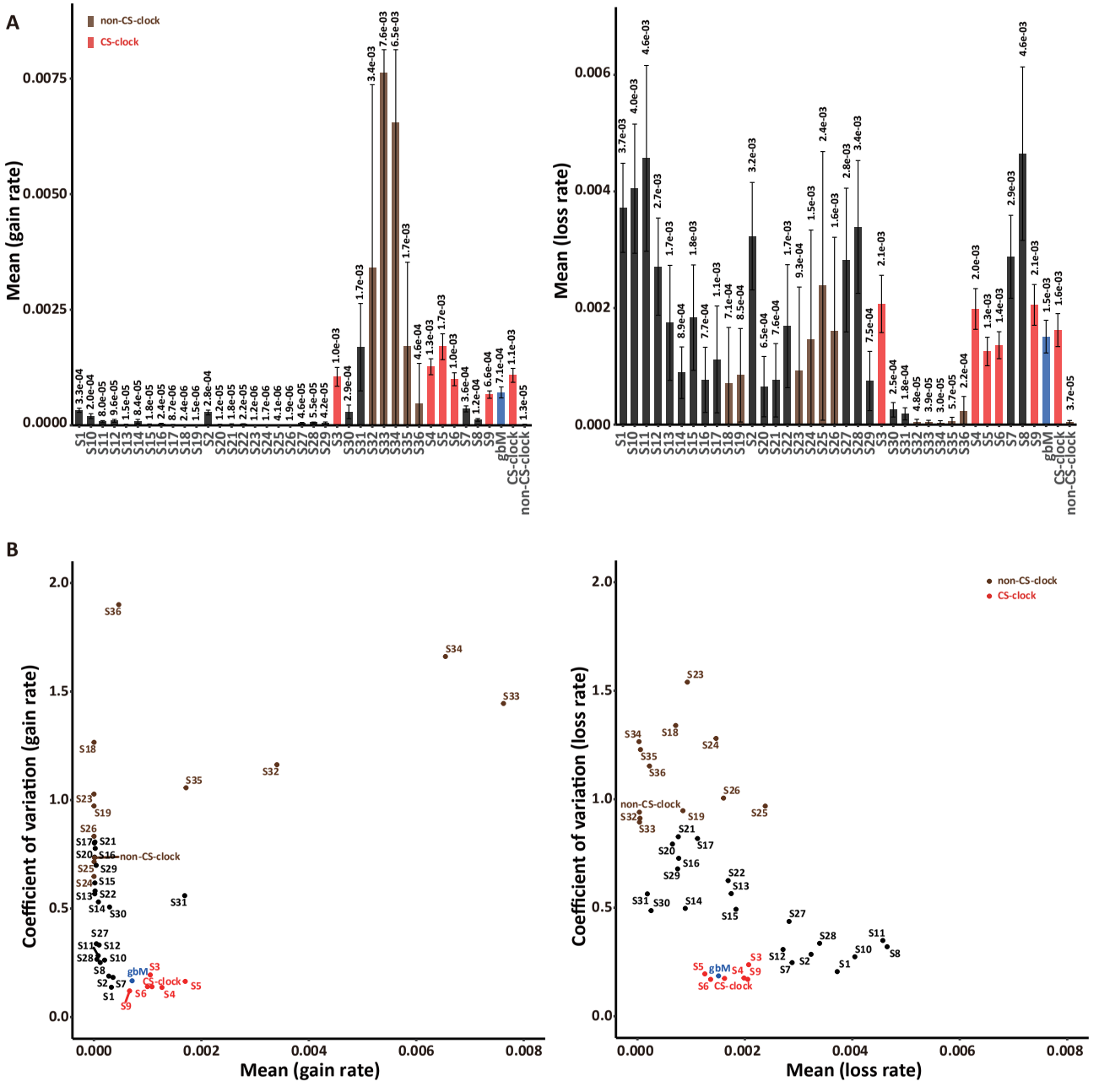


Fig. S2. Epimutation rate of each chromatin state across the 14 different genetic and environmental MA lines. (A) Mean of gain or loss rate. **(B)** Coefficient of variation versus the mean for gain or loss rate. This shows that the states with the lowest coefficient of variation (CS-clock states, red color) are not because they have the largest gain or loss rate.

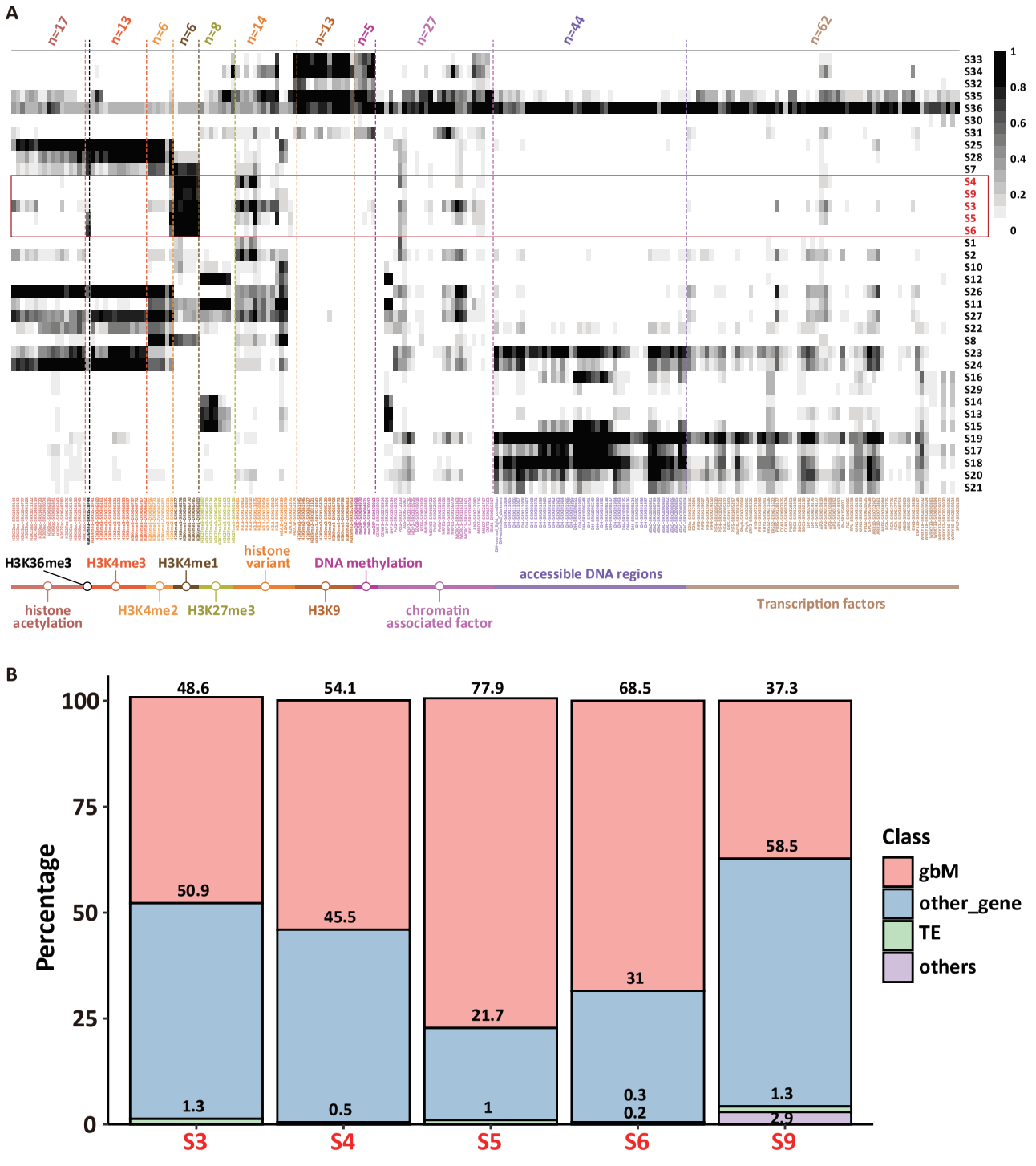


Fig. S3. Chromatin modifications and genomic annotations within defined chromatin states. (A) The enrichment of different chromatin modifications within each of the 36 chromatin states. **(B)** The percentage of genomic annotations of the five chromatin states that define the clock-like regions. "other_gene" was defined by removing gbM genes from the whole gene sets. "others" was defined by removing genes and TE from the whole genome.

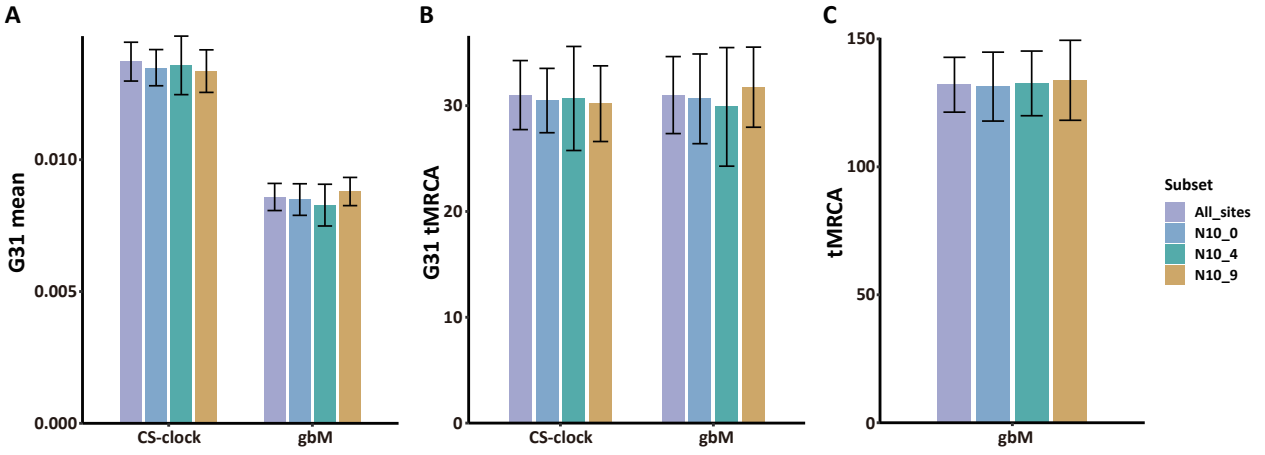


Fig. S4. Comparative analysis of depth and tMRCA in gbM and CS-clock subsets. This figure illustrates the measurement of depth and time to the most recent common ancestor (tMRCA) across subsets of CS-clock regions and gbM genes in *A. thaliana* MA1_1, MA1_2, and North American accessions. We created subsets by selecting one CG site per every ten (neighbourhood size = 10) from the original data, with the process replicated thrice, targeting the 1st, 5th, and 10th CG sites within each neighbourhood (N10_0, N10_4, N10_9). **(A)** Depth of the G31 samples from MA1_1 and MA1_2. **(B)** Estimated tMRCA for G31 samples, drawn from both gbM genes and CS-clock regions (including their respective subsets). The average epimutation rate across all gbM CpG sites in G31 was used for each gbM gene subset. Similarly, for each subset of CS-clock regions, the average epimutation rate from all corresponding CpG sites in G31 was used. **(C)** Estimated tMRCA for 12 non-recombining taxa from the North American accessions. These estimations used the average epimutation rate of all gbM sites, sourced from G31.

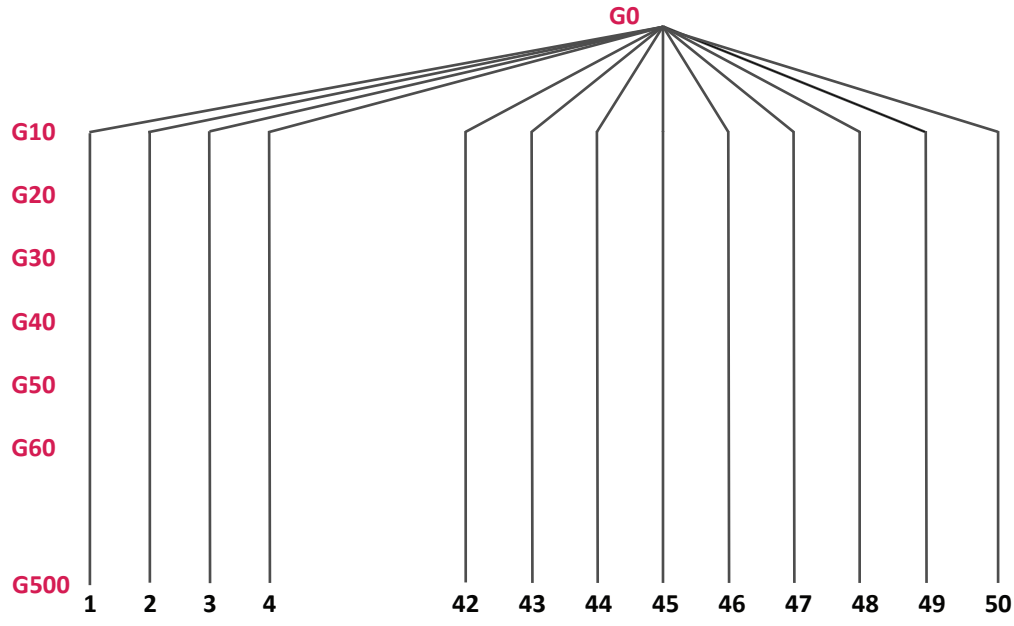


Fig. S5. Simulated MA lines. To study accumulation of SNPs and epimutations, we simulated 16 sets of MA lines. Each set of MA lines contained 50 MA lines that shared a single common ancestor (G0). The simulated MA lines were propagated with either selfing or clonal propagation for 500 generations. The substitution rates, generation time, and simulated sequence lengths were identical for all individuals in the same set of MA lines. Every 10 generations, the sequence divergence between the common ancestor and its descendants was measured and recorded.

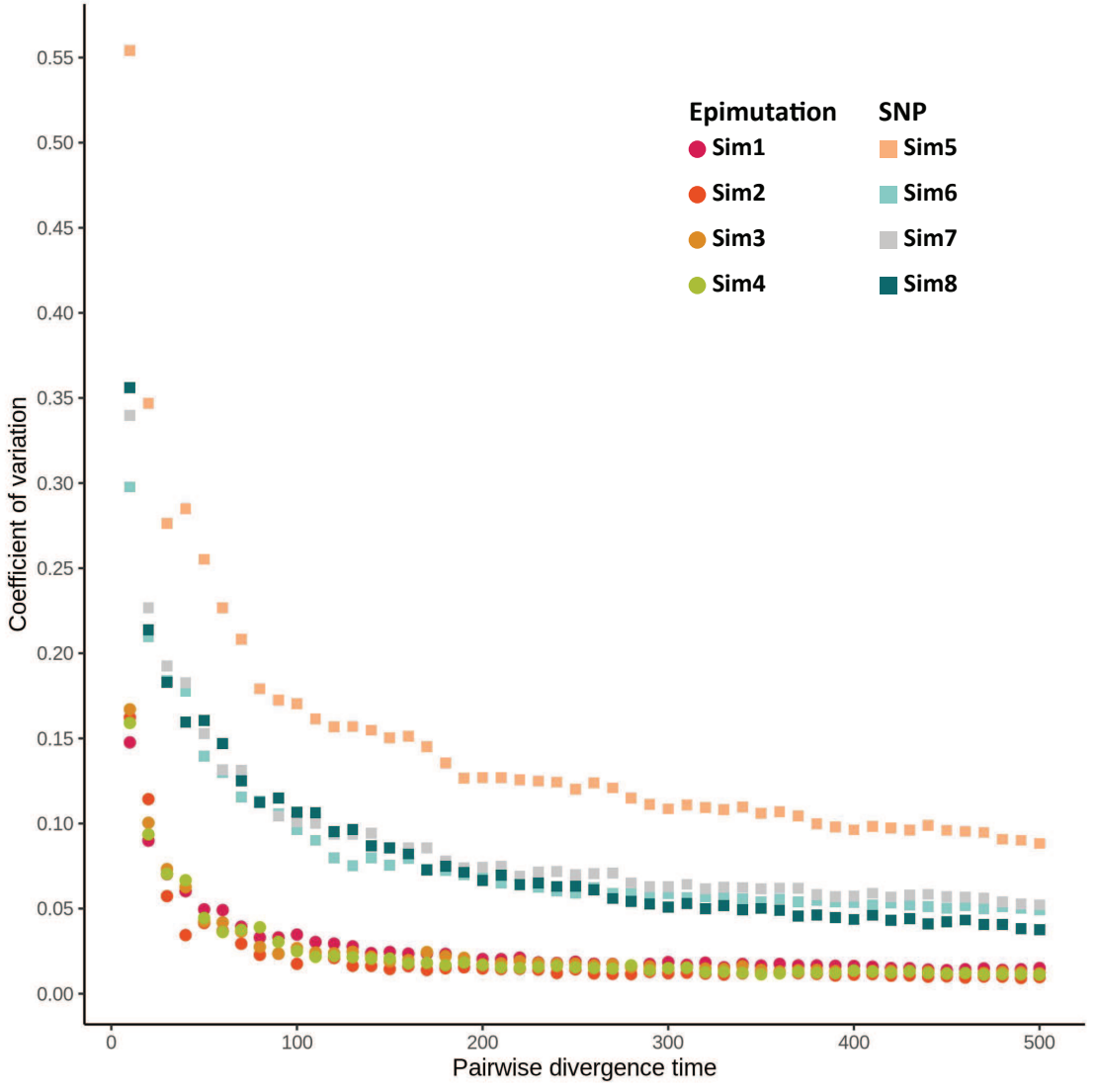


Fig. S6. Coefficient of variation of observed divergence in simulated selfing MA lines. We simulated accumulation of epimutations (Sim1-4, spots) and SNPs (Sim5-8, squares) on 8 sets of 50 simulated selfing MA lines. Within each set of MA lines, we measured the observed divergence every 10 generations and evaluated their dispersion via coefficient of variations (CVs). From G0 to G500, the CVs of observed epimutations are always ~50% lower than CVs of observed SNPs.

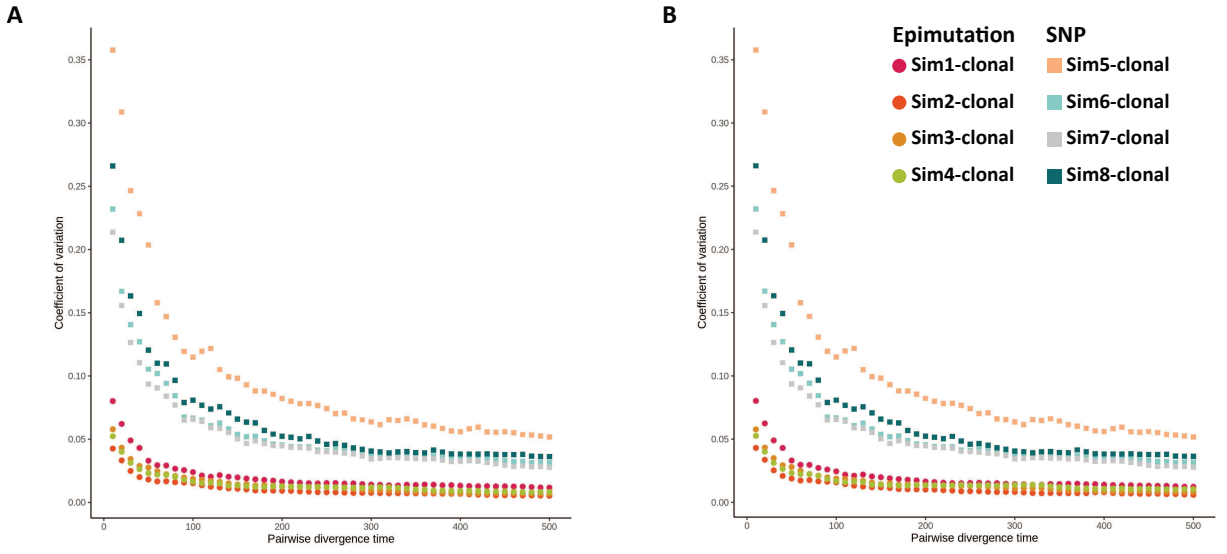


Fig. S7. Coefficient of variation of divergence in simulated clonal datasets. For clonal diploid plants, we also simulated the (epi)mutation accumulation on epigenetic clock regions (1×10^5 CpG sites, Sim1-clonal - Sim4-clonal) and genome (1×10^9 base pairs, Sim5-clonal - Sim8-clonal). There were eight sets of 50 simulated MA lines. Each set of the MA lines started from a single founder individual (G0) and reproduced through asexual reproduction for 500 generations. Every 10 generations, we measured the divergence between every single individual and G0. **(A)** Coefficient of variation of observed divergence (P-distance). **(B)** Coefficient of variation of estimated divergence. We used baseline distance to estimate the divergence on epigenetic clock regions. For simulated genomes, we used the K80 distance. Both observed and estimated divergence suggests the epigenetic clock has higher statistical robustness while inferring recent evolutionary histories.

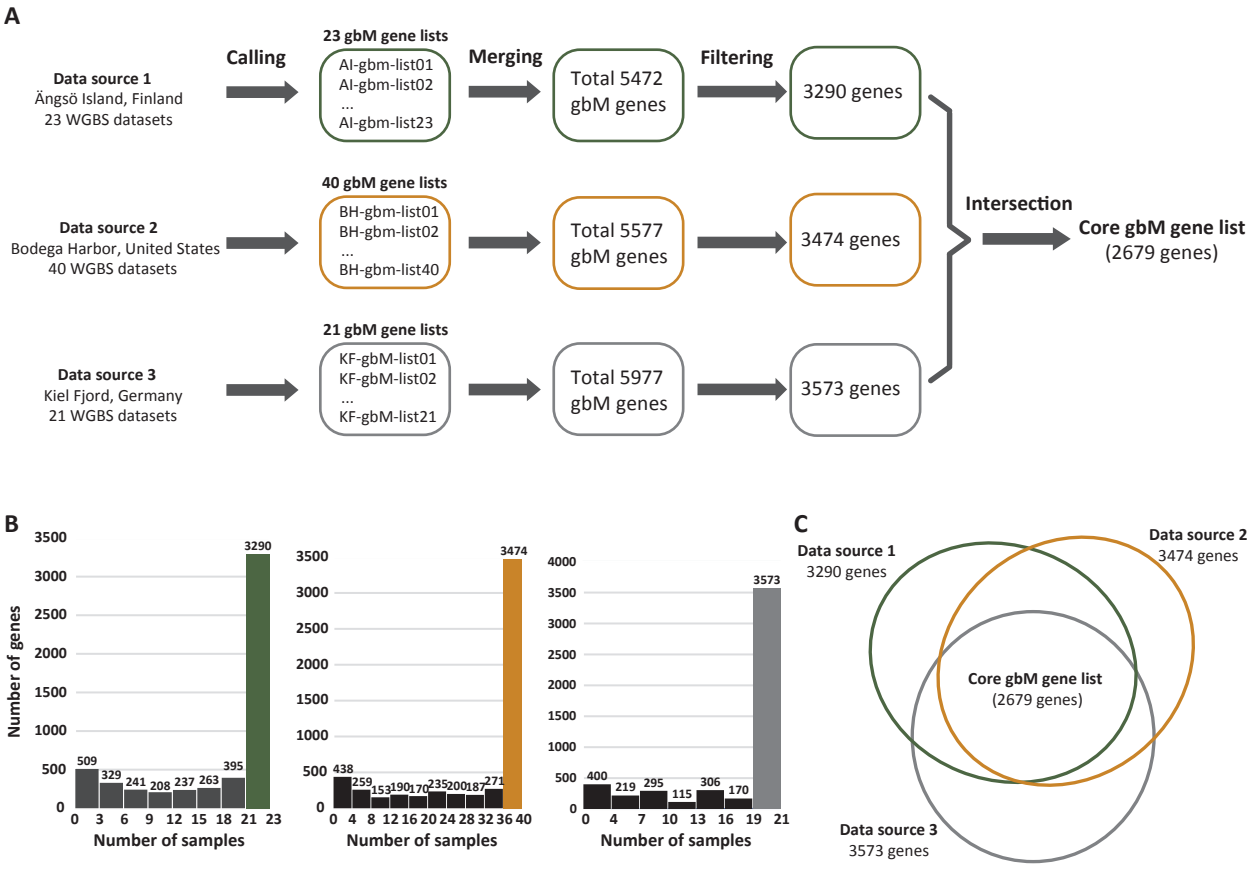


Fig. S8. Identifying gbM genes in *Z. marina* samples. (A) We identified gene-body methylation (gbM) genes from *Z. marina* samples from three data sources from various locations. From each sample, which corresponds to a WGBS dataset, a gbM gene list was generated with a published pipeline (30). **(B)** The gbM gene lists from the same data source showed high consistency. In samples from Ängsö Island, Finland (data source 1), 3,290 genes were identified as gbM genes in 90% samples (i.e., in at least 21 samples). With the same cut-off, 3,474 genes and 3,573 genes were identified in 90% of samples within each of the other two data sources. Thus, three lists of highly reliable gbM calls were obtained. **(C)** We defined the intersection of three lists in (B) as "core gbM gene list" and used them in for analyses.