Improved inference of population histories by integrating genomic and epigenomic data

Thibaut Sellinger[1,2], Frank Johannes[3], Aurélien Tellier[2*]

[1] Department of Environment and Biodiversity,
Paris Lodron University of Salzburg

[2] Professorship for Population Genetics,
Department of Life Science Systems,
Technical University of Munich

[3] Professorship for Plant Epigenomics,
Department of Molecular Life Sciences,
Technical University of Munich

[*] Corresponding author, aurelien.tellier@tum.de

Abstract

With the availability of high quality full genome polymorphism (SNPs) data, it becomes feasible to study the past demographic and selective history of populations in exquisite detail. However, such inferences still suffer from a lack of statistical resolution for recent, e.g. bottlenecks, events, and/or for populations with small nucleotide diversity. Additional heritable (epi)genetic markers, such as indels, transposable elements, microsatellites or cytosine methylation, may provide further, yet untapped, information on the recent past population history. We extend the Sequential Markovian Coalescent (SMC) framework to jointly use SNPs and other hyper-mutable markers. We are able to 1) improve the accuracy of demographic inference in recent times, 2) uncover past demographic events hidden to SNP-based inference methods, and 3) infer the hyper-mutable marker mutation rates under a finite site model. As a proof of principle, we focus on demographic inference in *A. thaliana* using DNA methylation diversity data from 10 European natural accessions. We demonstrate that segregating Single Methylated Polymorphisms (SMPs) satisfy the modelling assumptions of the SMC framework, while Differentially Methylated Regions (DMRs) are not suitable as their length exceeds that of the genomic distance between two recombination events. Combining SNPs and SMPs while accounting for site- and region-level epimutation processes, we provide new estimates of the glacial age bottleneck and post glacial population expansion of the European *A. thaliana* population. Our SMC framework readily accounts for a wide range of heritable genomic markers, thus paving the way for next generation inference of evolutionary history by combining information from several genetic and epigenetic markers.

Keywords— Kingman coalescent, Sequentially Markovian Coalescent, ancestral recombination graph, epigenetics, hidden markov model

Introduction

A central goal in population genetics is to reconstruct the evolutionary history of populations from patterns of genetic variation observed in the present. Relevant aspects of these histories include past demographic changes as well as signatures of selection. Inference methods based on Deep Learning (DL, [38]), Approximate Bayesian Computation (ABC, [9]) or Sequential Markovian Coalescent (SMC, [40, 58]) aim to infer this information directly from full genome sequencing data, which is becoming rapidly available for many (non-model) species due to decreasing costs. The SMC, in particular, offers an elegant theoretical framework that builds on the classical Wright-Fisher and the backward-in-time Kingman coalescent stochastic models (*e.g.* [36, 13, 75]). Both models conceptualize Mendelian inheritance as generating the genealogy of a population (or a sample), that is, the unique history of a fragment of DNA passing from parents to offspring. When this genealogy includes the effect of recombination, it is called the Ancestral Recombination Graph (ARG, [27, 79]).

Under the Kingmann coalescent model, the true genealogy of a population (or sample) is defined by its topology and branch length, and contains the information on past demographic changes and life history traits [50, 63, 68, 70] as well as selective events [13, 75]. The genealogical and the mutational processes of any heritable marker can therefore be disentangled, and the frequency of any given marker state is given by the shape of the genealogy in time (see Figure 1A). A central assumption about heritable genomic markers is that they are generated by two homogeneous Poisson mutation processes along the genome as well as through time. This entails that mutations in different genealogies are independent due to the effect of recombination [79, 47], and that there are no time periods with a large excess, or a severe lack, of mutations along a genealogy (mutations are independently distributed in time within a DNA fragment). In other words, the frequency of polymorphisms at DNA markers observed across a sample of sequences are constrained by, as well as inform on, the underlying genealogy at this locus (Figure 1A). To clarify these assumptions, we present a schematic representation of a marker 1 (yellow in Figure 1) which fulfills both homogeneous Poisson processes in time and along the genome. We also present cases applicable to a second genomic marker 2 that violates the model assumptions, namely by not being heritable (Figure 1B) or not following a non-homogeneous Poisson process in the genome (Figure 1C) or in time (Figure 1D).

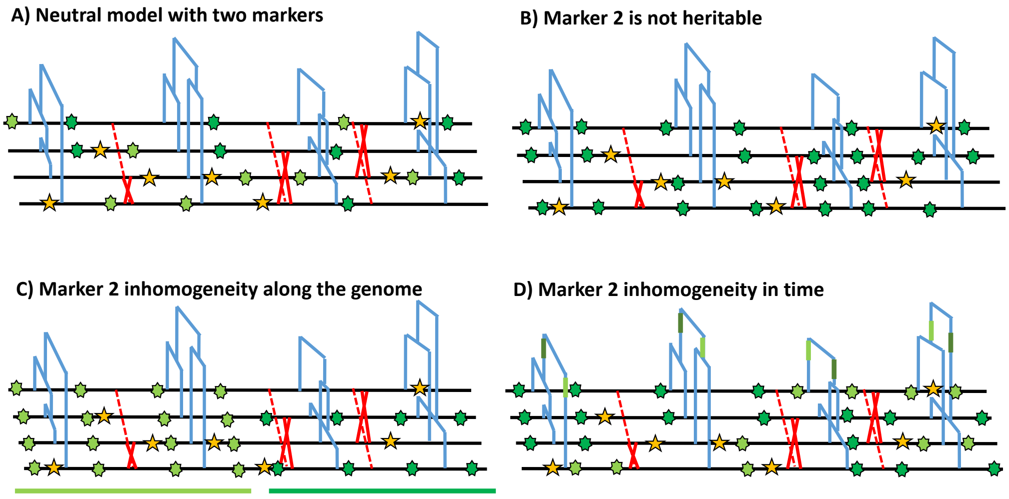
A) Neutral model with two markers
B) Marker 2 is not heritable
C) Marker 2 inhomogeneity along the genome
D) Marker 2 inhomogeneity in time

Fig. 1. Schematic distribution of two markers along the genealogy and four genomes. A) Schematic distribution of marker 1 (yellow star) and marker 2 (green star) along the genealogies in a sample of four genomes both following a homogeneous Poisson process. B) The green marker 2 is not heritable, so that its distribution is independent from the genealogy. C) The green marker 2 is spatially structured along the genome, violating the distribution of the Poisson process along the genome and conflicting with the genealogy. D) The green marker 2 does not follows Poisson process through time, *e.g.* burst of mutations at a specific time point represented by given branches of the genealogies in green. The yellow marker 1 has an identical Poisson process along the genome and the genealogy in all four panels, and for readability, marker 2 exhibits light and dark green states.

Despite the power of the SMC, well-known model violations such as variation in recombination and mutation rates along the genome [5, 4] or pervasive selection [61, 31, 30] can compromise the accuracy of demographic and selective inference [24, 64]. There are two other important issues that have received less attention in the literature. The first issue occurs when the population recombination rate (ρ) is higher than the population mutation rate (θ). In such cases, inferences can be biased if not erroneous [71, 64, 63], because several recombination events cannot be inferred due to the lack of Single Nucleotide Polymorphisms (SNPs for point mutations). This problem affects many species, though interestingly not humans which have a ratio $\rho/\theta \approx 1$. A second issue occurs when the mutational process along the genealogy is too slow be informative about sudden and strong variation in population size (*i.e.* population bottlenecks), such as during colonization events of novel habitats. The typical low mutation rate of 10^{-9} up to 10^{-8} (per base, per generation) found in most species therefore places strong limitations on SMC analysis of recent bottleneck events (up to ca. 10^{-4} generations ago) when inference is based solely on SNP data. Indeed, bottlenecks are often either not found, or when inferred, their timing and magnitude are not well estimated (inferred smoother than in reality, [31, 64]), even when a large number of samples is used. A typical example is the large uncertainty of the timing and magnitude of the population size bottleneck during the Last Glacial Maximum (LGM) and post-LGM expansion in *A. thaliana* European populations based on several studies using different accessions and SMC inference methods [2, 19].

Nonetheless, current SMC, DL or ABC inference methods making use of full genome sequence data rely almost exclusively on SNPs for inference [58, 71, 63, 9, 37]. There are both practical and theoretical reasons for using SNPs: They are easily detectable from short-read re-sequencing data and their mutational process is well approximated by the infinite site model [13, 75], simplifying the inference of the underlying genealogy. However, other heritable genomic markers exists whose mutation rates can be several orders of magnitude higher than that of SNPs, and could thus be more informative about recent demographic events. These include microsatellites, insertions, deletions and transposable elements (TEs). Although those heritable markers are not necessarily neutral (such as TEs which are likely to be under weak purifying selection) they contain information on the evolutionary history of the population. Current technological limitations still impede the easy detection and estimation of allele frequencies for many of these markers [81, 53, 76]. For example, identifying insertion/excision variation of transposable elements or copy number variation of microsatellites requires a high quality reference genome and ideally long-read sequencing approaches [53]. In addition to these genomic markers, DNA cytosine methylation is emerging as a potentially useful epigenetic marker for phylogenetic inference in plants [83, 84]. Stochastic gains and losses of DNA methylation at CG dinucleotides, in particular, arise at a rate of ca. 10^{-4} up to 10^{-5} per site per generation (that is 4 to 5 orders of magnitude faster than DNA point mutations, [73]), and can be inherited across generations [54, 78]. These so-called spontaneous epimutations are likely neutral at the genome-wide scale ([74, 29], but see [49, 54]), and can be easily detected from bisulphite converted short read sequencing data [41, 60]. Recent work suggests that CG methylation data can be used as a molecular clock for timing divergence between pairs of lineages over timescales ranging from years to decades [84].

However, theoretical integration of the above-mentioned (epi)genomic markers into a population genomics and SMC inference framework is not trivial. Because of the high mutation rate, the mutational process at these (hyper-mutable) markers is reversible and more consistent with a finite site, rather than infinite site, model, which can result in extensive homoplasy (as known for microsatellite markers, [20]). Indeed, classic expectations of population genetics diversity statistics, mostly build for SNPs, need to be revised for these hyper-mutable markers [14, 77]. Here we develop the theoretical and methodological inference framework named SMCtheo for the inclusion of additional (potentially hyper-mutable) markers into the SMC. We showcase our model using extensive simulations as well as application to published DNA cytosine methylation data (in genic regions) from local populations of *A. thaliana* [60, 74]. We demonstrate that integration of hyper-mutable genomic markers into SMC models significantly improves the inference accuracy of past variation of population size, or can even uncover demographic events not uncovered using SNPs alone. Our proof-of-principle approach opens up novel avenues for studying population genetic processes over time-scales that have been largely inaccessible using traditional SNP-based approaches. This may prove particularly useful when exploring recent demographic changes of endangered species as a way to assess their potential for extinction in the context of biodiversity loss and global change.

Results

Theoretical results with two markers underlying the SMC computations

We study polymorphic sites across genomes of several sampled individuals which exhibit several possible markers (DNA nucleotides, methylation, TEs, indels, microsatellites,...). We define any marker by 1) its maximum number of possible states (nb_s), for example nucleotide sites have four states (A, T, C and G) while a methylation site has two states (methylated or unmethylated), and 2) its mutation rate μ, *i.e.* the rate at which the state of a marker changes into another state per position and per generation [3] (for simplicity we assume an equal mutation rates between all bases, known as the Jukes-Cantor model). More specifically, we are interested in two rates: the DNA mutation rate for changes in DNA nucleotides, and epimutation rate for change in methylation state. Furthermore, we assume that at each position on the genome only one type of marker can occur and be observed. We obtain as a first theoretical result the probability for a given site in the genome to be identical ($P(id)$) or segregating ($P(seg)$) (*i.e.* polymorphic) in a sample of size two ($n = 2$, two sampled chromosomes are compared):

$$\begin{aligned} P(id, n = 2) &= \frac{1}{nb_s} + \frac{(nb_s - 1)}{nb_s} e^{-2\mu t_M \frac{(nb_s)}{(nb_s - 1)}} \\ P(seg, n = 2) &= \frac{(nb_s - 1)}{nb_s} - \frac{(nb_s - 1)}{nb_s} e^{-2\mu t_M \frac{(nb_s)}{(nb_s - 1)}} \end{aligned} \tag{1}$$

This probability is a function of the time to the most recent common ancestor (TMRCA in text and t_M in equation 1, details in Supplementary Text). The probability for a mutation to occur for a given marker increases with an increased

TMRCA [13, 75], but under high mutation rates (and high effective population size) the marker may not be polymorphic in the sample as mutations may be reversed (so-called homoplasy, [20, 14]). In Supplementary Figure S1 we illustrate these properties by computing the probability 1 for different mutation rates. The inference of recent demographic events and bottlenecks relies on the presence of polymorphic sites to detect recent coalescent events (TMRCA), and should be improved by using markers with high (or fast) mutation rate (*e.g.* hyper mutable).

In the following, we simulate data under different demographic scenarios using the sequence simulator program *msprime* [6, 33], which generates the ARG of n sampled diploid individuals (set to $n = 5$ throughout this study, leading to 10 haploid genomes). This ARG contains the genealogy of a given sample at each position of the simulated chromosomes. We then process the ARG to create DNA sequences according to the model parameters and the type of marker considered. We first assume a set of genomic markers obtained for a sample size n, and mutating according an homogeneous Poisson process along the genome and in time (along the genealogy) as in Figure 1A. To simulate the sequence data, we define the number of marker types (any number between 1 and the sequence length) and the proportion of sites of each marker type in the sequence. Each marker is characterized by both parameters nb_s and μ. For simplicity, we simulate sequences with two markers, but note that the method can be easily extent to additional markers. Marker 1 represents 98% of the sequence, and has a per site mutation rate $\mu_1 = 10^{-8}$ mimicking nucleotide SNP markers under an infinite site model (thus considered as bi-allelic at a given DNA site, [82]). By contrast, marker 2 composes the complementary 2% of the sequence length, with a per site mutation rate of $\mu_2 = 10^{-4}$ per generation between two possible states. Marker 2 is thus hyper-mutable compared to marker 1 and mimics methylation/epimutation sites. Note, that mutation events in Marker 1 and 2 are simulated under a finite site model.

We use different SMC-based methods throughout this study. These methods include: 1) MSMC2 used as a reference method [45], 2) SMCtheo is an extension of the PSMC' [40, 58] accounting for any number of heritable theoretical markers, and 3) eSMC2 which is equivalent to SMCtheo but accounting only for SNPs markers [64] (to avoid any bias in implementation differences between SMCtheo and MSMC2). All methods are Hidden Markov Models (HMM) derived from the Pairwise Sequentially Markovian Coalescent (PSMC') [58] and assume neutral evolution and a panmictic population. The hidden states of these methods are the coalescence time of a sample of size two at a position on the sequence. From the distribution of the hidden states along the genome, all methods can infer population size variation through time as well as the recombination rate [58, 45, 64].

The inclusions of hyper-mutable genomic markers improves demographic inference

We assume that the mutation rate of marker 1 is $\mu_1 = 10^{-8}$ per generation per bp. We use this information to estimate the mutation rate of marker 2, which we vary from $\mu_2 = 10^{-8}$ to $\mu_2 = 10^{-2}$ per generation per bp. The estimation results based on simulated data under a constant population size of $N = 10,000$ are displayed in Table 1. We find that our approach is capable of inferring μ_2 with high accuracy for rates up to $\mu_2 = 10^{-4}$. However, when the mutation rate μ_2 is 10^{-2}, our approach underestimates it by a factor three, suggesting the existence of an accuracy limit. To demonstrate that information can be gained by integrating marker 2 (with $\mu_2 = 10^{-4}$), we compared the ability of several inference methods to recover a recent bottleneck (Figure 2A). All methods correctly infer the amplitude of population size variation. When accounting only for marker 1 (with $\mu_1 = 10^{-8}$, MSMC2 and eSMC2 fail to infer accurately the sudden variation of population size. However, with the inclusion of hyper-mutable marker 2, our SMCtheo approach correctly infers the rapid change of population size of the bottleneck (Figure 2A, green). It is encouraging that an accurate estimation of the demography is obtained, even when the mutation rate of marker 2 is unknown (Figure 2A, blue).

True μ_2 value	Estimated value of μ_2
10^{-8}	9.9×10^{-9} (0.02)
10^{-6}	1.0×10^{-6} (0.008)
10^{-4}	1.4×10^{-4} (0.01)
10^{-2}	3.05×10^{-3} (0.41)

Table 1: Average estimated values of the mutation rate of marker 2 (μ_2), knowing that of marker 1. We use 10 sequences (5 diploid individuals) of 100 Mb ($r = \mu_1 = 10^{-8}$ per generation per bp) under a constant population size fixed at $N = 10,000$. The coefficient of variation over 10 repetitions is indicated in parentheses.

Furthermore, some species or populations might feature small effective population sizes (ca. $N = 1,000$), potentially resulting in reduced genomic diversity. In such cases the inclusion of hyper-mutable markers should also improve demographic inference. We present the results of such a scenario in Figure 2B, where the population size was divided by a factor 10 compared to the previous scenario in Figure 2A. We find that in the absence of the hyper-mutable marker 2, no approach can correctly infer the variation of population size. From the shape of the inferred demography, methods using only marker 1 do not suggest the existence of a bottleneck followed by recovery (the "U-shaped" demographic scenario is not apparent with the orange and red lines, Figure 2 B). Yet, when integrating both markers, the population size can be recovered, even if the mutation rate of marker 2 is not *a priori* known. In both Figure 2A and B, we assume that the marker 2 occurs at a frequency of 2% in the genome. This percentage may be unrealistically high depending on the marker and the species. To test the impact of reducing marker 2 frequency, we repeat the simulations shown in Figure 2A, but set its frequency to as low as 0.1% (a 20-fold reduction). We find that the inclusion of the hyper-mutable marker 2 continues to improve inference accuracy in very recent times, albeit less pronounced than in Figure 2A (see Supplementary Figure 2). This suggests that a very small proportion of hyper-mutable genomic sites is sufficient to significantly improve the accuracy of inferences.

All full genome inference methods, especially SMC approaches, display lower accuracy when the population recombination rate ($\rho = 4Nr$) is larger than the population mutation rate of marker 1 ($\theta_1 = 4N\mu_1$). We simulate sequence data under a bottleneck scenario slightly more ancient than in Figure 2 A and assume that $\rho/\theta_1 = r/\mu_1 = 10$ and $\rho/\theta_2 = r/\mu_2 = 10^{-3}$. Our results show that by integrating the genomic marker 2 which mutation rate is larger than the recombination rate, estimates of the recombination rate as well as past population size variation are substantially improved (Table 2, Figure 2C). Indeed, analyzing only marker 1, eSMC2 and MSMC2 identify the bottleneck (albeit smoothed) and only slightly overestimate recent population size (Figure 2D). By integrating the hyper-mutable marker 2, our SMCtheo approach correctly infers the strength and time of the bottleneck when μ_1 and μ_2 are known (Figure 2D, green line), while the timing of the bottleneck is slightly shifted in the past when μ_2 is unknown and estimated by our method (Figure 2D, blue line). When μ_2 is unknown, SMCtheo additionally infers a spurious sudden variation of population size between 10,000 and 100,000 generations ago. Using only marker 1, the estimates of the recombination rate are inaccurate (Table 2). To complete the visual representation and provide a quantitative assessment of inference accuracy, we compute the root mean square error (RMSE) values for demographic inference (Supplementary Table 1). We further improve the accuracy of estimation by optimizing the likelihood (LH) to estimate the recombination rate and demography compared to the classically used Baum-Welch (BW) algorithm (Table 2 and Supplementary Figure S3). Our results demonstrate that SNPs are limiting and insufficient for accurate inferences in recent times and that the inclusion of an additional marker with mutation rate higher than the recombination rate generates significant improvements in demographic inference. However, by directly optimizing the likelihood the true recombination rate can be well recovered even with marker 1 only.

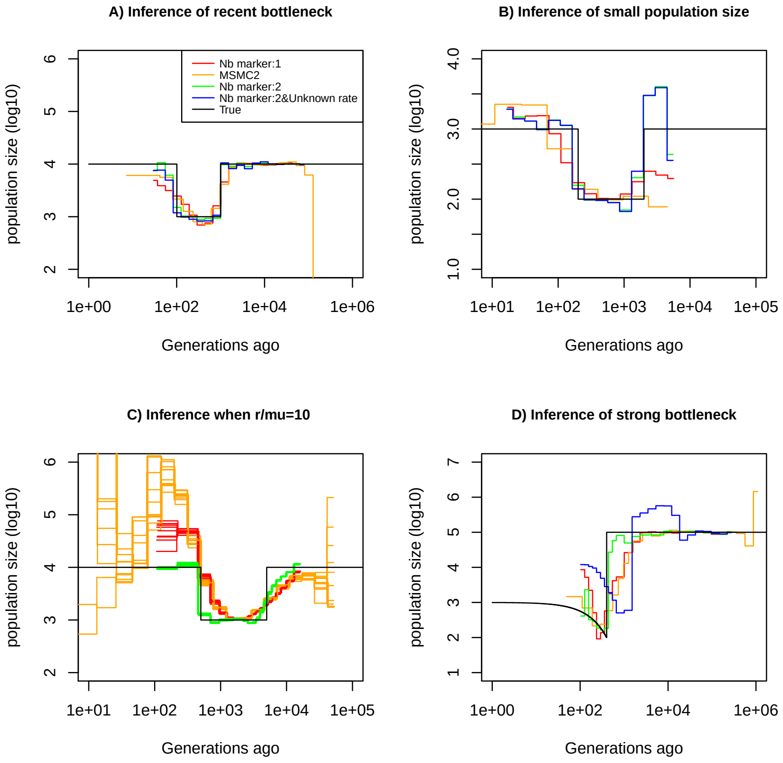

Fig. 2. Performance of SMC approaches using different markers. Estimated demographic history of a bottleneck (black line) by SMC approaches using two genomic markers. In orange and red, are the estimates by MSMC2 and eSMC2 based on only marker 1. Estimates from SMCtheo integrating both markers are in green (with known μ_2), and in blue with unknown μ_2. The demographic scenarios are A) 10-fold recent bottleneck with an ancestral population size $N = 10,000$, B) 10-fold recent bottleneck with an ancestral population size $N = 1,000$, C) 10-fold bottleneck with an ancestral population size $N = 10,000$, and D) a very severe (1,000 fold) and very recent bottleneck with incomplete size recovery. In A, B and D, we assume $r/\mu_1 = 1$ (with $r = \mu_1 = 10^{-8}$, $\mu_2 = 10^{-4}$ per generation per bp) and in C, $r/\mu_1 = 10$ (with $r = 10^{-7}$, $\mu_1 = 10^{-8}$, and $\mu_2 = 10^{-4}$ per generation per bp). In all cases (A, B, C and D) 10 sequences (5 diploid indivudals) of 100 Mb were used as input.

Method	True recombination rate	Average estimated recombination rate
MSMC2 (BW)	10^{-7}	0.23×10^{-7} (0.017)
1 Marker : BW	10^{-7}	0.25×10^{-7} (0.012)
2 Marker : BW	10^{-7}	0.90×10^{-7} (0.004)
1 Marker : LH	10^{-7}	0.84×10^{-7} (0.036)
2 Marker : LH	10^{-7}	0.94×10^{-7} (0.01)

Table 2: Estimates of recombination rates with one or both markers. For SMCtheo, BW stands for the use of the Baum-Welch algorithm to infer parameters, and LH to the use of the likelihood. We use 10 sequences of 100 Mb with $r = 10^{-7}$, $\mu_1 = 10^{-8}$ and $\mu_2 = 10^{-4}$ per generation per bp in a population with a past bottleneck event. The coefficient of variation over 10 repetitions is indicated in brackets.

Integrating DNA methylation improves the accuracy of inference

Definition of the theoretical model for DNA methylation

Following the previously encouraging results of demographic inference with SNPs and an hyper-mutable marker under the specific assumptions of Figure 1A, we develop a specific SMCm method to jointly analyse SNPs and CG methylation as an epigenetic hyper-mutable marker. Since our SMCm stems from the eSMC [63, 68] it corrects for the effect of self-fertilization when appliying to *A. thaliana*. We focus here on methylation located in CG contexts within genic regions as these have been found to evolve neutrally [74, 83, 84]. The methylation of individual CG dinucleotides produces a biallelic heritable marker with a finite number of (epi)mutable sites (Figure 3). In a sample of several sequences from a population, variation in the methylation status of individual CGs is known as single methylation polymorphism (SMP, Figure 3A) which could be used for demographic and divergence inference [73, 74]. However, CG methylation sites can also be organized in spatial clusters (of similar state) due to region level epimutation (Figure 3B, [78, 18, 49]). Region level epimutations can have different epimutation rates than individual CG sites. Population-level variation in the methylation status of these clusters is known as differentially methylated regions (DMRs). Furthermore, when integrating SMP and DMR epimutational processes (*i.e.* what we here call region level epimutation), the methylation status of CG sites is therefore affected by the superposition of both processes. Therefore the simulation and modeling of epimutational processes of SMPs is more complex than in our previous model as we need to account for the effect of region methylation as well as for methylation and demethylation epimutation rates to be different and asymmetrical [73, 18].

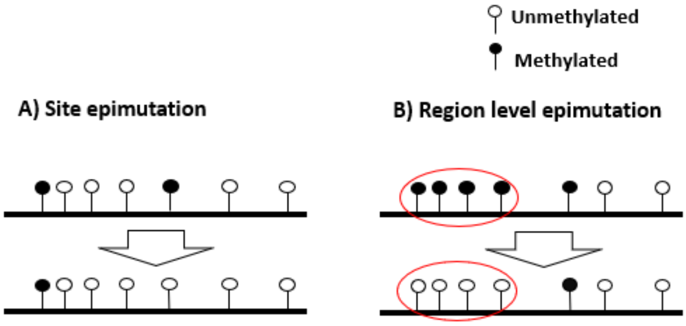

Fig. 3. Schematic representation of site and region epimutations Schematic representation of a sequence undergoing epimutation at A) the cytosine site level, and B) at the region level. A methylated cytosine in CG context is indicated in black and an unmethylated cytosine in white.

To make our simulations realistic, we use the *A. thaliana* genome sequence as a starting point, and focus on CG dinucleotides within genic regions. To that end, we selected random 1kb regions within genes and choose only those CG sites that are clearly methylated or unmethylated in *A. thaliana* natural populations based on whole genome bisulphite sequencing (WGBS) mesaurements from the 1001G project (SI text). Our simulator for CG methylation is built in a similar way as the one described above but the epimutation rates are allowed to be asymmetric with the per-site methylation rate (μ_{SM}) and demythylation (μ_{SU}). Region-level epimutations are also implemented, setting the region length to either 1kb [49] or 150 bp [18]. The region level methylation and demethylation rates are defined as μ_{RM} and μ_{RU}, respectively. We assume that site-level and region-level epimutation processes are independent. Making this assumption explicit later allow us to test if it is violated in comparisons with actual data. Our simulator also assumes that DNA mutations and epimutations are independent of one another. That is, for simplicity we ignore the fact that methylated cytosines are more likely to transition to thyamines as a result of spontaneous deamination [28]. We also ignore the possibility that new DNA mutations could act as CG methylation quantitative trait loci and affect CG methylation patterns in both cis and trans. Such events are extremely rare so that the above assumptions should hold reasonably well over short evolutionary time-scales. As the goal is to apply our approach to *A. thaliana*, we simulate sequence data for a sample size $n = 10$ (but considering *A. thaliana* haploid) from a population displaying 90% selfing [63?] under a recent severe population bottleneck demographic scenario. We simulate data assuming previously estimates of the rates of recombination [56], DNA mutation [52], and site- and region-level methylation [73, 18].

As guidance for future analyses of demographic inference using SNPs and DNA methylation data, the theoretical and empirical analysis of *A. thaliana* methylomes consist of the following five steps: 1) assessing the relevance of region-level methylation (DMRs) for inference, 2) inference of site and region epimutation rates, 3) comparing statistics for the SNPs, SMPs and DMRs distributions, 4) demographic inference using SNPs with SMPs or DMRS, and 5) demographic inference using SNPs with SMPs and DMRs.

Step 1: assessing the relevance of region-level methylation (DMRs) for inference

We determine our ability to detect the existence of spatial correlations between epimutations. That is, we asked if site-specific epimutations can lead to region-level methylation status changes across a range of epimutation rates (assuming two sequences of 100 Mb, $r = \mu_1 = 10^{-8}$ per generation per bp and a constant population size $N = 10,000$, results in Supplementary Table 2). If site-specific epimutations are independently distributed, the probability of a given site to be in a given (methylated or unmethylated) state should be independent from the state of nearby sites (knowing the epimutation rate per site). Conversely, if there is a region effect on epimutation (DMRs), two consecutive sites along the genome would exhibit a positive correlation in their methylated states. We therefore calculate from the per-site (de)methylation rates μ_{SM} and μ_{SD} the probability that two successive cytosine positions are identical in their methylation assuming they are independent. This probability can be compared to the one observed from methylation data (here simulated) so that we obtain a statistical test for the existence of a positive correlation in the methylation status of nearby sites, interpreted as a regional-level epimutation process (p-value = 0.05) according to Figure 1A. A small p-value of the test (<0.05) suggests the existence of a region effect for methylation/demethylation affecting neighbouring cytosines, contrary to a high p-value indicating no spatial structure of methylation distribution. We find that when region epimutation rates are higher than (or similar to) site-level epimutation rates, namely $\mu_{RM} \gtrapprox \mu_{SM}$ and $\mu_{RU} \gtrapprox \mu_{SU}$), the existence of regions of consecutive cytosines is detected with high accuracy. However, when site-level epimutation rates are higher ($\mu_{SU} > \mu_{RU}$ and $\mu_{SM} > \mu_{RM}$) than region-level epimutation rates, region-level changes cannot be readily detected (Supplementary Table 2). When methylated regions are detected, we can further determine their length using a specifically developed Hidden Markov Model (HMM) using all pairs of genomes (similarly to [65, 18, 69]). While the length of the methylated region is pre-determined in our simulations (1kb or 150bp), site-level epimutation occur which can change the distribution of methylation states in that region and across individuals, thus DMR regions can vary in length along the genome and between pairs of chromosomes.

Step 2: inference of site- and region-level epimutation rates

As the epimutation rates of most plant species remain unknown, we assess the accuracy of SMCm to infer epimutation rates at the site- and region-level directly from simulated data. We first assume that either only site- or only region epimutations can occur, and infer their respective rates (see Supplementary Table 3 and 4). Our SMCm approach can accurately recover these rates except when these are higher than 10^{-4}. Next, we assess the accuracy of our approach to simultaneously infer site- and region-level epimutation rates assuming that region and site epimutation rates are equal (Supplementary Table 5 and Supplementary Figure 4). Similar to our previous observation, we find that when the epimutation rates are very high (*e.g.* close to 10^{-2}), accuracy is lost compared to slower epimutation rates. Nonetheless, our average estimated rates are off from the true value by less than an factor 10. Hence, under our model assumptions, we are able to recover the correct order of magnitude for site- and region-level methylation and demethylation rates.

Step 3: distribution of statistics for SNPs, SMPs and DMRs

To gain insights on the distribution of epimutations under the described assumptions, we look at key statistics from our simulations: the distribution of distance between two recombination events versus the distribution of the length of estimated DMR regions (Figure 4A), and the LD decay for SMPs (in genic regions) and SNPs (in all contexts) (Figure 4C, D). In our simulations DMRs regions have a maximum fixed size, but their length depends on the interaction between the region- and site-level epimutation rates. As mentioned in step 1, the methylated/demethylated regions are detected using the binomial test and their length estimated by the HMM. Therefore, while variation exists for the length of these regions (Figure 4A), regions are on average shorter than the span of genealogies along the genome, which are defined by the frequency of recombination events along the genome ($r = 3.5 \times 10^{-8}$ as in *A. thaliana*). There is is virtually no linkage disequilibrium (LD) between epimutations due to the high epimutation rate (Figure 4C), while the LD between SNPs can range over few kbp (Figure 4D, as observed in *A. thaliana* [12, 60]). Note however, that the region methylation process in itself does not generate LD because this measure can only be computed if SMPs are present in frequency higher than $2/n$ in the sample, *i.e.* there is no LD measure defined for monomorphic methylated/unmethylated regions. In other words, our simulator generates SNPs, SMPs and DMRs which fulfill the three key assumptions of Figure 1A. We note that by using a constant population size $N = 10,000$, the LD decay for SNPs is higher than in the *A. thaliana* data which exhibit an effective population size of ca. $N = 250,000$ [12] and past changes in size.

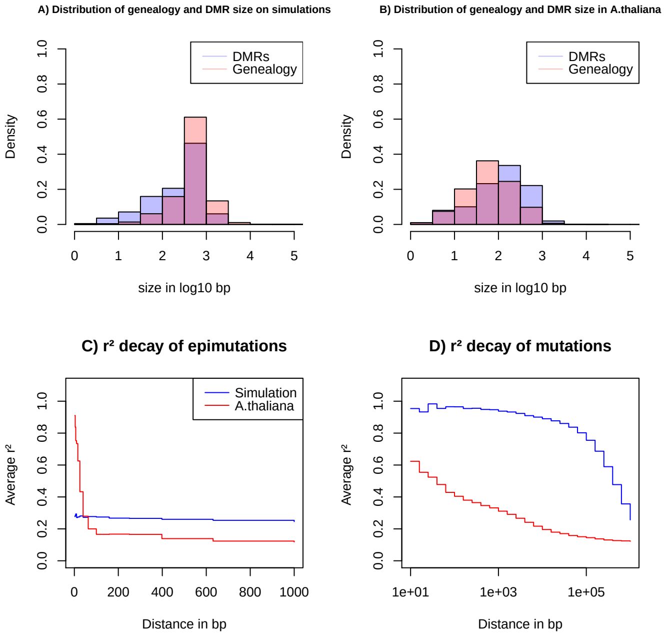

Fig. 4. Key statistics for epimutations and mutations. A) Histogram of the length between two recombination events (genomic span of a genealogy) and DMRs size in bp of the simulated data. B) Histogram of genealogy span and DMRs size in bp from the *A. thaliana* data (10 German accessions). C) Linkage desequilibrium decay of epimutations in our samples of *A. thaliana* (red) and simulated data (blue). D) Linkage desequilibrium decay of mutations in our *A. thaliana* samples (red) and simulated data (blue). The simulations reproduce the outcome of a recent bottleneck with sample size $n = 5$ diploid of 100 Mb, the rates per generation per bp are $r = 3.5 \times 10^{-8}$, $\mu_1 = 7 \times 10^{-9}$, $\mu_{SM} = 3.5 \times 10^{-4}$, $\mu_{SU} = 1.5 \times 10^{-3}$, and per 1kb region $\mu_{RM} = 2 \times 10^{-4}$ and $\mu_{RU} = 1 \times 10^{-3}$.

Step 4: demographic inference based on SNPs with SMPs or DMRs

We test the usefulness of either SMPs or DMRs for demographic inference. Simulations under the demographic model from steps 1-3 assume DNA mutations (SNPs) and only site epimutations (SMPs), *i.e.* no region-level methylation ($\mu_{RM} = \mu_{RU} = 0$). We perform inference of past demographic history under different amount of potentially methylated sites with and without *a priori* knowledge of the methylation/demythylation rates (Figure 5A, B). When the site epimutation rates are *a priori* known, the sharp decrease of population size can be accurately detected. When epimutation rates are unknown, the shape of the past demographic history is also well inferred except for a scaling issue (a shift along the x- and y-axes similar to that in Figure 5D). When we vary the amount of potentially methylated sites (2%, 10% and 20%) our inference results remain largely unchanged. This suggests that having methylation measurements for as low as 2% of all CG sites being epimutable in the genome is entirely sufficient to improved SNP-based demographic inference (eSMC2 in Figure 5A). The RMSE values for demographic inference are computed for all cases in Figure 5 to provide an additional quantitative understanding of our results (Supplementary Table 6).

The amount of sequence data used in Figure 5A and B is fairly large compared to real datasets (10 haploid genomes of length 100 Mb). We therefore ran the SMCm and eSMC2 on sequence data simulated under the same scenario but with a reduced sequence length of 10 Mb ($n = 5$ diploid, Figure 5C and D, only 3 repetitions are presented for visibility). In this case, we found that inference is significantly affected when using only SNPs (eSMC2 in blue), as we are unable to correctly recover the demographic scenario. However, incorporating SMPs with known site-level epimutations into the model leads to substantial inference improvements (Figure 5C and D, Supplementary Table 6).

We additionally quantify the accuracy gain in ARG inference by inferring the expected coalescent time (TMRCA) at each position in the genome by the three approaches (eSMC2, SMCm with unknown epimutation rates and SMCm with known epimutation rates) under the same scenario from Figure 5. The RMSE values of the TMRCA inference are presented in Supplementary Table 7. We confirm our intuition that integrating epimutations slightly improves the accuracy of TMRCA when the epimutation rates are known, but does not when the rates are unknown.

To quantify the effect of DMRs on inference, we simulate data under the same demographic scenario, but assume only region level epimutations (DMRs, $\mu_{SM} = \mu_{SU} = 0$). The results for DMR region sizes 1kb and 150bp are displayed in Supplementary Figure S5 and S6, respectively. As in Figure 5, we observed a gain of accuracy in inference when region-level epimutation rates are known, while the length of the region (1kb or 150bp) does not seem to affect the result. However, no significant gain of information is observed when integrating DMR data with unknown epimutation rates (Supplementary Figure 5 and 6). In summary, CG methylation SMPs and to a lesser extend DMRs, can be used jointly with SNPs to improve demographic inference (Supplementary Table 8 presents the corresponding RMSE values for demographic inference shown in Supplementary Figure 5 and 6), especially in recent times (Supplementary Table 6 and 8).

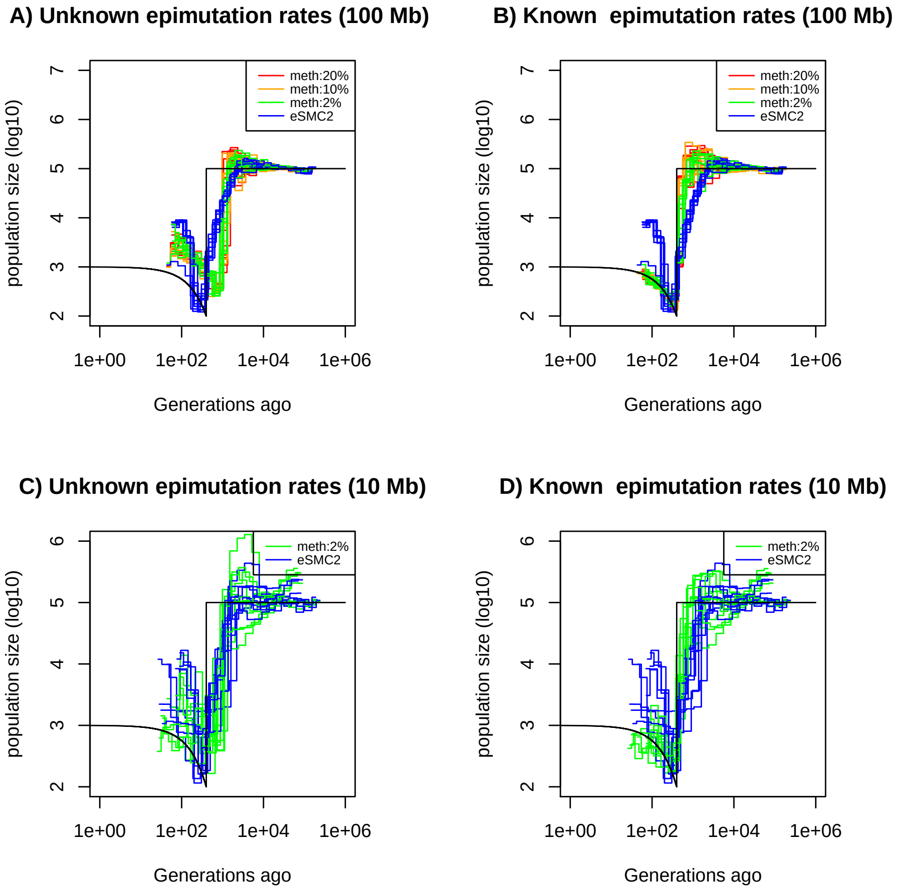

Fig. 5. Performance of SMC approaches using site epimutations (SMPs) and mutations (SNPs) under a bottleneck scenario. Estimated demographic history by eSMC2 (blue) and SMCm assuming the epimutation rate is known (B and D) or not (A and C) where the percentage of CG sites with methylated information varies between 20% (red), 10% (orange) and 2% (green) using 10 sequences of 100 Mb in A and B (with 10 repetitions) and 10 sequences of 10 Mb in C and D (three repetitions displayed) under a recent severe bottleneck (black). The parameters are: $r = 3.5 \times 10^{-8}$ per generation per bp, mutation rate $\mu_1 = 7 \times 10^{-9}$, methylation rate to $\mu_{SM} = 3.5 \times 10^{-4}$ and demethylation rate to $\mu_{SU} = 1.5 \times 10^{-3}$ per generation per bp.

Step 5: demographic inference based on SNPs with SMPs and DMRs

Since site- and region-level methylation processes can occur in real data, we run SMCm on simulated data under the same demographic scenario, but now using both site (SMPs) and region (DMRs) epimutations and accounting for both mutation processes (with rates similar to the one found in *Arabidopsis thaliana*). Inference results are displayed in Supplementary Figure 7 (RMSE values in Supplementary Table 9). When the epimutations rates are unknown, we observe a gain of accuracy when integrating epimutations, especially in the recent times. However when epimutation rates are *a priori* known we observe a loss of accuracy when accounting for epimutations. This loss of accuracy is due to the mislabeling of the methylation region status (in step 1) when site and region-level epimutations occur jointly at similar rates (as there will be methylated sites in unmethylated regions and unmethylated sites in methylated regions).

Finally, we assess the inference accuracy when using SNPs and SMPs but ignoring in SMCm the region methylation effect (DMRs), even though this latter process takes place (Supplementary Figure 8, RMSE values in Supplementary Table 10). The inference accuracy decreases compared to the previous results (Supplementary Figure 5-7), and while the sudden variation of population is somehow recovered, the estimates of the time and magnitude of size change are not well recovered in recent time. Hence those results demonstrate the importance of accounting for site and region level epimutations processes in steps 1 to 5.

We demonstrate that our SMCm can exhibit, to some extend, an improved statistical power for demographic inference using SNPs and SMPs while accounting for site and region-level methylation processes under the assumptions of Figure 1A. We show that 1) using SMPs we can unveil past demographic events hidden by limitations in SNPs, 2) the correct demography can be uncovered irrespective of knowing *a priori* the epimutation rates, 3) ignoring site or region-level processes can decrease the accuracy of inference, and 4) knowing the epimutation rates may improve the estimate of demography compared to simultaneously estimating them with SMCm.

Joint use of SNPs and SMPs improves the inference of recent demographic history in *A. thaliana*

Step 1: assessing the strength of region-level methylation process in *A. thaliana*

We apply our inference model to genome and methylome data from 10 *A. thaliana* plants from a German local population [12]. We start by assessing the strength of a region effect on the distribution of methylated CG sites along the genome. As expected from [18], for all 10 individual full methylomes we reject the hypothesis of a binomial distribution of methylated and unmethylated sites along the genomes, suggesting the existence of region effect methylation (yielding DMRs) meaning that CG are more likely to be methylated if in a highly methylated region, and conversely for unmethylated CG. This is consistent with the autocorrelations in mCG found in [18, 11, 43]. As a first measure of methylated region length, we test the independence between two annotated CG methylation given a minimum genomic distance between them (within one genome). We observe an average p-value smaller than 0.05 for distances up to 2,000bp but then the p-value rapidly increases ($>$0.4) (Supplementary Figure 9). As a second measure, our HMM (based on pairs of genomes) yields a DMR average length of 222 bp (distribution in Figure 4B).

We conclude that the minimum distance for epimutations to be independent along a genome is over 2kb and spans larger distance than the typically proposed DMR size (ca. 150 bp in [18] and 222bp in our analysis) and can therefore cover the size of a gene (see [49, 11]). The simulations and data from *A. thaliana* indicate that the epimutation processes that produces DMRs at the population level in plants cannot simply results from the cumulative action of single-site epimutations. This insights is consistent with recent analyses of epimutational processes in gene bodies, which seems to indicate that the autocorrelation in CG methylation is a function of cooperative methylation maintenance and the distribution of histone modifications [11, 43].

Step 2: site- and region-level epimutation rates

We use the rates empirically estimated in *A thaliana* and taken in the above simulations ($\mu_{SM} = 3.5 \times 10^{-4}$ and $\mu_{SU} = 1.5 \times 10^{-3}$ per bp per generation and $\mu_{RM} = 2 \times 10^{-4}$ and $\mu_{RU} = 1 \times 10^{-3}$ per region per generation, [73, 18]).

Step 3: distribution statistics for SNPs, SMPs and DMRs in *A. thaliana*

Since our SMC model assumes that DNA, SMP and DMR polymorphisms are determined by the underlying population/sample genealogy, DMR which span long genomic regions may spread across multiple genealogies and thus violates our modelling assumptions. We thus further investigate the potential discrepancies between the data and our model (Figure 4). We infer the DMR sizes from all 10 *A. thaliana* accessions using our *ad hoc* HMM, and measure the bp distance between a change in the expected hidden state (*i.e.* coalescent time) along the genome, which we interpret as recombination events (called the genomic span of a genealogy). The resulting distributions are found in Figure 4B. We observe that both distributions have a similar shape but DMRs are on average twice as large as the inferred genomic genealogy span: average length of 222 bp (DMR) vs 137 bp (genealogy) and median length of 134 bp (DMR) vs 62 bp (genealogy). This means that on average DMRs are larger than the average distance between two recombination events, thus violating the homogeneous distribution of epimutations along the genome (Figure

1C).

To further unveil potential non-homogeneity of the distribution of epimutations, we assess the decay of LD of mutations (SNPs) and epimutations (SMPs) (Figure 4C and D) confirming the results in [60]. We find the LD between SMPs in the data to be high (and higher than LD between SNPs) for distance smaller than 100 bp (red line in Figure 4C and D). The LD decay of SMPs is much faster than for SNPs (no linkage disequilibrium between epimutations for distances $>$ 100bp), likely stemming from 1) epimutation rates being much higher than the DNA mutation rate, and 2) the high per site recombination rate in *A. thaliana*. Moreover, the LD between SMPs at distance smaller than 100bp in *A. thaliana* being much higher compared to our simulations (Figure 4C), we suggest that additional local mechanisms of epimutation processes may not be accounted for in our model of the region-level methylation process.

Step 4: demographic inference for *A. thaliana* based only on SNPs and SMPs

Finally, we apply the SMCm approach to data from the German accessions of *A. thaliana*. When using SNP data only, the demographic results are similar to those previously found [63, 68] (Figure 6 purple lines), with no strong evidence for an expansion post-Last Glacial Maximum (LGM) [12]. We then sub-sample and analyze segregating SMPs, which exhibit both methylated and unmethylated states in our sample (as in [73]). Here we ignore DMRs and account only for SMPs. When we use as input the methylation and demethylation rates that have been inferred experimentally [73], a mild bottleneck post-LGM is followed by recent expansion (Figure 6 blue lines). By contrast, letting our SMCm estimate the epimutations rates, we find in recent times a somehow similar but stronger demographic change post-LGM. We find a strong bottleneck event occurring between ca. 5,000 and 10,000 generations ago followed by an expansion until today (Figure 6 green lines). The inferred site epimutation rates are 10,000 faster than the DNA mutation rate (Supplementary Table 11) which is close to the expected order of magnitude from experimental measures with and without DMR effects [73, 18]. Both estimates thus yield a post-LGM bottleneck followed by a recent population expansion.

These results indicate that the inclusion of DNA methylation data can aid in the accurate reconstruction of the evolutionary history of populations, particularly in the recent past where SNPs reach their resolution limit. This is made possible by the fact that the DNA methylation status at CG dinucleotide undergoes stochastic changes at rates that are several orders of magnitude higher than the DNA mutation rate, and can be inherited across generations similar to DNA mutations.

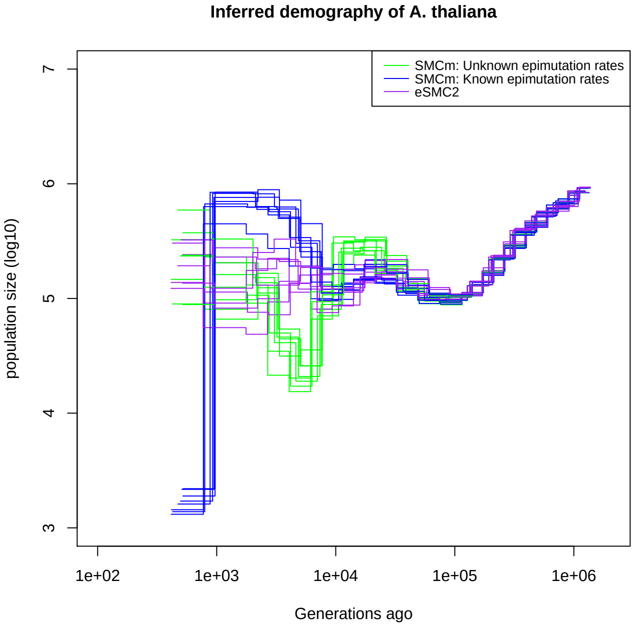

Fig. 6. Integrating epimutations and mutations on German accessions of *A. thaliana*. Estimated demographic history of the German population by eSMC2 (only SNPs, purple) and SMCm when keeping polymorphic methylation sites (SMPs) only: green with epimutation rates estimated by SMCm, blue with epimutation rates fixed to empirical values. The region epimutation effect is ignored. The parameters are $r = 3.6 \times 10^{-8}$, $\mu_1 = 6.95 \times 10^{-9}$, and when assumed known, the site methylation rate is $\mu_{SM} = 3.5 \times 10^{-4}$ and demethylation rate is $\mu_{SU} = 1.5 \times 10^{-3}$.

Step 5: demographic inference correcting for DMRs in *A. thaliana*

To assess the robustness of our inference results, we run SMCm using all cytosines (CG) sites with an annotated methylation status (segregating or not) while accounting or not for DMRs (Supplementary Figure 10). We fix epimutation rates to the empirically estimated values, and confirm the estimates from Figure 6. When the region-level methylation process is ignored the inferred demography (blue lines in Supplementary Figure 10) is similar to the estimates from SMPs with fixed rates in Figure 6 (blue lines). When the region-level methylation process is taken into account (orange lines in Supplementary Figure 10), the inferred demography is similar to that of the Figure 6 (green lines). In the case where we infer the epimutation rates (sites and region) the demographic history inference is not improved compared to that estimated using SNPs only (Supplementary Figure 10, green and red lines) while the inferred epimutation rates are smaller than expected (Supplementary Table 11 and 12), but the ratio of site to region epimutation rates is consistent with empirical estimates [18].

Discussion

Current approaches analyzing whole genome sequences rely on statistics derived from the distribution of ancestral recombination graphs [23, 64, 37, 68, 10, 80, 66, 34]. In this study we present a new SMC method that combines SNP data with other types of genomic (TEs, microsatallites) and epigenomic (DNA methylation) markers. We focus mainly on the inclusion of genomic markers whose mutation rates exceed the DNA point mutation rate, as such (hyper-mutable) markers can provide increased temporal resolution in the recent evolutionary past of populations, and aid in the identification of demographic changes (*e.g.* population bottlenecks). We demonstrate that by integrating multiple heritable genomic markers, the population size variation in very recent time can be more accurately recovered (outperforming any other methods given the amount of data used in this study [71, 66]). Our results indicate that correctly integrating multiple genomic marker can improve TRMCA inference, which is becoming a field of high interest [37, 26, 44]. Our simulations demonstrate that if the SNP mutation rate is known, the mutation rate of other markers can be recovered (under the condition that the marker follow all hypotheses described in Figure 1). Moreover, our method accounts for the finite site problem that arises at reversible (hyper-mutable) markers and/or where effective population size is high [70, 72]. Overall, the simulator and SMC methods presented here therefore pave the way for a rigorous statistical framework to test if a common ARG can explain the observed diversity patterns under the model hypotheses laid out in Figure 1. We find that comparisons of LD for different markers along the genome is a useful way to assess violations of our model assumptions.

As proof of principle, we apply our approach on data originating from whole genome and methylome data of *A. thaliana* natural accessions (focusing on CG context in genic regions, as in [74, 83, 84]). Indeed, *A. thaliana* presents the largest genetic and epigenetic data-set of high quality. Additionally the methylation states in CG context has been proven mainly heritable and is well documented [18, 25, 73]. We first investigate the distribution of epimuations along the genomes. Our model-based approach provides strong evidence that DMRs cannot simply emerge from spontaneous site-level epimutations that arise according to a Poisson processes along genome. Instead, stochastic changes in region-level methylation states must be the outcome of spontaneous methylation and demethylation events that operate at both the site- and region-level (as corroborated by [54, 11, 43]). Our epimutation model cannot fully describe the observed diversity of epimutations along the genome [18], meaning that the epimutation processes may indeed be more complex than expected [18, 25, 11, 43]. We observe non-independence between annotated methylation sites spanning genomic regions larger than the span of the underlying genealogy (determined by recombination events) which no model can currently describe. Additionally, we find high LD between SMPs over short distances which does not appear in our simulations (simulation performed under the current measures of epimutation rates). Thus, methylation probably violate the assumptions of a Poisson process distribution along the genome and in time (*i.e.* Figure 1), in line with recent functional studies [54, 25, 42, 43]. We thus further caution against conclusions on the role of natural (purifying) selection [49] or its absence [74] based on population epigenomic data due to the violation of the above mentioned assumptions. Additionally we suspect those model violations to explain the discrepancy between epimutation rates we inferred and the ones measured experimentally [73, 18]. To solve this discrepancy, one would need to develop a theoretical epimutation model capable of describing the observed diversity at the evolutionary time scale and then use this model to reanalyse the sequence data from the biological experiment to re-estimate the epimutation rates. We thus suggest a possible way forward for modeling epimutations through an Ising model [86] to account for the heterogeneous methylation process. However, our preliminary work and the simulation results in [11], indicate that such model generates non-homogeneous mutation process in space (*i.e* along the genome) and time, violating our current SMC assumptions (Figure 1C and D). Hence, there is a need to develop a more realistic methylation model for epimutations. A model accounting for heterogeneous rates would probably need to rely on a more sophisticated HMM (*e.g.* continuous time Markov chains [35] for SMC approaches) than what is presented here or to use other full genome inference methods (see [37]) which are not constrained by the SMC assumptions (Figure 1) but depends on simulations.

Interestingly, the distance of LD decay for SMPs matches quite well the estimated distance between recombination events (Figure 4). In addition to our theoretical results in Table 2, this observation reinforces the usefulness of using SMPs (or any hyper-mutable marker) to improve estimates of the recombination rate along the genome in species where the per site DNA mutation rate (μ) is smaller than the per site recombination rate (r) as in *A. thaliana*.

Nonetheless, we find that a restricted focus on segregating SMPs in genic regions could meet our model assumptions reasonably well, and thus provides a promising way forward. Using these segregating SMPs, we recover a past demographic bottleneck followed by an expansion which could fit the post- Last Glacial Maximum (LGM) colonization of Europe (although caution must be taken concerning the reliability of those results as pointed above), a hypothesized scenario [21] which could not be clearly identified using SNPs only from European (relic and non-relic) accessions [12]. Currently strong evidence from inference methods are lacking ([12], Figure 4 in [19]). Indeed, beyond the limits of using SNPs only, current results are limited by theoretical frameworks unable to simultaneously account (and disentangle) for extensive background selection (reinforced by very high selfing), population structure and variation in molecular rates (*e.g.* mutation rates, [48]), which are all known to be present in *A. thaliana*. Those various forces are known to bias inference results when non-accounted for [15, 55], and may explain the variance in our demographic estimates. We note also that using CG methylated sites in genic regions may be problematic as the typical genealogies at these loci could be shorter than the genome average due to the presence of background selection, thus making the inference of such short TMRCA more difficult (even with SMPs) than in non-coding regions (which do not harbour desirable CG methylation sites, [73, 74, 83]).

We suggest that simultaneously accounting for multiple heritable markers can help disentangle between different evolutionary forces, such as between selection and variation in mutation rate: selection has a local effect on the population genealogy, while the mutation rate variation would only locally affect that given marker but not the genealogy [15]. The absence of conflicting demography inferred from SNPs and from methylation confirms at the time scale of thousands of generations, CG methylation sites are mainly heritable and can be modeled using population genetics theory [14, 74] (but see [54]) and used to estimate divergence between lineages [84, 83]. In other words fast ecological local adaptation [59] and response to stresses [67] may likely not be prominent forces endlessly reshaping CG methylation patterns (non-heritability in Figure 1B).

Overall, our results demonstrate that our approach can be used in different cases. If the epimutations/genomic markers evolutionary mechanisms are not well understood [54, 11, 43], our approach provides inference tools to study the markers' rates and distribution process along the genome, without requiring additional experimental data. If the evolution of epimutations/genomic markers are well understood (including a measure of the mutation rates) and can be modeled to described the observed intra-population diversity, these can be integrated to improve the SMC performance. Hence when applying our approach to genome-wide genetic and epigenetic data, it is advisable to use accurately annotated markers with, if possible, information regarding their inheritance and mutational properties. Regarding methylation specifically, while the set of gene body methylated genes previously used [74, 84] are likely the optimal choice [83], these are too few and too scattered across the genome to maximize the statistical power of SMC methods. We therefore use methylation sites at all genic regions. Yet, despite the wealth of functional studies and data on methylation in *A. thaliana*, the distribution of epimutations is not fully understood [25, 54], but independent rates for sites and region-level have been estimated [73, 18, 84]. We note here the promising methylation modelling framework by [11, 43], albeit it does not yet consider evolutionary processes at the population level. Our results shed light on the inference accuracy in presence of site and region-level epimutations when occurring at similar rates (Supplementary Figure 7). When accounting for region-level epimutations, our algorithm requires to first infer via an HMM the methylation status of a region in order to later-on compute the epimutation probabilities (*i.e* the emission matrix of the SMC HMM). Hence, in presence of site and region-level epimutations occurring at similar rates, recovering the region methylation status becomes harder as methylated sites are observed in the unmethylated regions (and unmethylated sites observed in the methylated regions). The mislabelling of the region methylation status lead to accuracy loss due to the use of the wrong emission probability at the later steps of the SMC inference (Forward-Backward algorithm). In the case where epimutation rates are freely inferred, their values are based on the estimated methylation region status. Therefore, even if the inferred rates are incorrect, these are sufficiently consistent with the inferred region methylation status to contain information and slightly improve inference accuracy. Additionally, extra care must be taken when dealing with epigenomic data in other species as the SMP calling might not be as simple as for *Arabidopsis thaliana* due to potential difference of methylation between different tissues or pool of cells. Similarly, we ignore here the potential dependence between SNPs and SMPs, as more empirical evidence (and modelling) is required to quantify the potential interaction between both mutational processes.

On a brighter note, with the release of new sequencing technology [39], long and accurate reads are becoming accessible, leading to the availability of high quality reference genomes for model and non-model species alike [51, 7]. Additionally, the quality of re-sequencing (population sample) genome data and their annotations is enhanced so that additional markers such as transposable elements, insertion, deletion or microsatellites can be called with increasing confidence. These accurate genomes will provide access to new classes of genomic markers that span the entire mutational spectrum. We therefore suspect in the near future an improvement in our understanding of the heritability of many markers besides SNPs. Adding other genomic markers besides SNPs will improve full genome approaches, which are currently limited by the observed nucleotide diversity [34, 66, 62]. Additionally, the potential complexity resulting by integrating multiple independent markers could be tackled by the use of continuous time Markov chains for the emission matrix. We predict that our results pave the way to improve the inference of 1) biological traits or recombination rate through time [17, 68], 2) multiple merger events [37], and 3) recombination and mutation rate maps [5, 4]. Our method also should help to dissect the effect of evolutionary forces on genomic diversity [32, 31], and to improve the simultaneous detection, quantification and dating of selection events [1, 8, 30].

Hence, there is no doubt that extending our work, by simultaneously integrating diverse types of genomic markers into other theoretical framework (*e.g.* ABC approaches), likely represents the future of population genomics, especially to study species for which many thousands of samples cannot be obtained. We believe our approach helps to develop more general classes of models capable of leveraging information from any type and amount of diversity observed in sequencing data, and thus to challenge our current understanding of genome evolution.

Materials and Methods

Simulating two genomic markers

The sequence is written as a sequence of markers with a given state. Each site is annotated as $MXSY$, where X indicates the marker type and Y the current state of that marker: for example M1S1 indicate at this position a marker of type 1 in the state 1. To simulate sequence of theoretical marker we start by simulating an ARG which is then split in a series of genealogies (*i.e.* a sequence of coalescent trees) along the chromosome and create an ancestral sequence (based on equilibrium probability of marker states). Mutation events (nucleotides or epimutations for methylable cytosine) are then added when going along the sequence, *i.e.* along the series of genealogies. The ancestral sequence is thus modified by mutation event assuming a finite site model [82] conditioned to the branch length and topology of the genealogies. Each leaf of the genealogy is one of the n samples. Our model has thus two important features: 1) markers are independent from one another, and 2) a given marker has a polymorphism distribution between samples (frequencies of alleles) determined by one given genealogy. The simulator can be found in the latest version of eSMC2 R package (https://github.com/TPPSellinger/eSMC2).

Simulating methylome data

We now focus on methylation data located at cytosine in CG context within genic regions. Only, CG sites in those regions are considered "methylable", and CG sites outside those defined genic regions do not have a methylation status and are considered "unmethylable". We vary the percentage of CG site with methylation state annotated from 2 to 20% of the sequence length. The simulator can in principle simulate epimutations in different methylation context and different rates [41, 16, 87, 85]. We simulate epimutations as described above but with asymmetric rates: the methylation rate per site is $\mu_{SM} = 3.5 \times 10^{-4}$, and the demethylation rate per site is $\mu_{SM} = 1.5 \times 10^{-3}$ [73, 18]. For simplicity and computational tractability, we assume that when an epimutation occurs, it occurs on both DNA strands which then present the same information. In other words, for a haploid individual, a cytosine site can only be methylated or unmethylated (as in [69]). For region level epimutations, the region length is either 1kbp [49] or 150 bp [18]. The region level methylation and demethylation rates are set to $\mu_{RM} = 2 \times 10^{-4}$ and $\mu_{RU} = 10^{-3}$ respectively (similar to rates measured in *A. thaliana*, [18]). In addition to this, unlike for theoretical marker described above, mutations, site and region epimutations can occur at the same position of the sequence.

To simulate methylation data, we start with an ancestral sequence of random nucleotide and then randomly select regions in which CG sites have their methylation state annotated (representing the genic regions). Cytosine in CG context in those regions are either methylated or unmethylated (noted as M or U). Cytosine in other context or regions are considered as unmethylabe (and noted as C). The ancestral methylation state is then randomly attributed according to the equilibrium probabilities. Our simulator then introduces DNA mutations, site- and region-epimutations in a similar way as described above.

SMC Methods

All three methods (eSMC2, SMCtheo and SMCm) are based on the same mathematical foundations and implemented in a similar way within the eSMC2 R package (https://github.com/TPPSellinger) [68, 37, 64]. This allows to specifically quantify the accuracy gained by accounting for multiple genomic markers.

SMC optimization function

All current SMC approach rely on the Baum-Welch (BW) algorithm for parameter estimation in order to reduce computational load (as described in [71]). Yet, the Baum-Welch algorithm is an Expectation-Maximization algorithm, and can hence fall in local extrema when optimizing the likelihood. We alternatively extend SMCtheo to estimate parameters by directly optimizing the likelihood (LH) at the greater cost of computation time (even when using the speeding techniques described in [57]). We run this approach on a sub-sample of size six haploid genomes to limit the required computational time.

eSMC2 and MSMC2

SMC methods based on the PSMC' [58], such as eSMC2 and MSMC2, focus on the coalescent events between two individuals (*i.e.* two haploid genomes or one diploid genome). The algorithm moves along the sequence and estimates the coalescence time at each position by assessing whether the two sequences are similar or different at each position. If the two sequences are different, this indicates a mutation took place in the genealogy of the sample. The intuition being that the absence of mutations (*i.e.* the two sequences are identical) is likely due to a recent common ancestor between the sequences, and the presence of several mutations likely reflects that the most recent common ancestor of the two sequences is distant in the past. In the event of recombination, there is a break in the current genealogy and the coalescence time consequently takes a new value according to the model parameters [46, 58]. A detailed description of the algorithm can be found in [45, 63].

SMCtheo based on several genomic markers

Our SMCtheo approach is equivalent to PSMC' but take as input a sequence of several genomic markers. The algorithm goes along a pair of haploid genomes and checks at each position which marker is observed and then if both states of the marker are identical or not. The approach is identical to the one described above, except that the probability of both sequences to be identical at one site depends on the mutation rate of the marker at this site (equation 1). While the mutation rates for many heritable genomic markers are unknown, there is an increasing amount of measures of the DNA (SNP) mutation rate for many species. Our SMCtheo approach is able to leverage the information from the distribution of one theoretical marker (*e.g.* mutations for SNPs) to infer the mutation rate of the other marker 2 (assuming both mutation rates to be symmetrical). If more than 1% of sites are polymorphic in a sequence we use the finite site assumption. If not, then from the diversity observed, the different mutation rates can be recovered by simply comparing Waterson's theta (θ_W) between the reference marker (*i.e.* with known rate) and the marker with the unknown rates. For example, if the diversity (θ_W) at marker 2 is smaller by a factor ten than the reference marker 1 (and no marker violates the infinite site hypothesis), the mutation rate of marker 2 is inferred to be ten times smaller (corrected by the number of possible states). However, if the marker 2 violates the infinite site hypothesis, a Baum-Welch algorithm is run to infer the most likely mutation rates under the SMC to overcome this issue (the Baum-Welch algorithm description can be found in [63]).

SMCm

When integrating epimutations, the number of possible observations increases compare to eSMC2. As in eSMC2, if the two nucleotides (DNA mutation) at one position are identical at a non methylable site, we indicate this as 0. If the two nucleotides are different, it is indicated as 1 (*i.e.* a DNA mutation occurred). When assuming site-level epimutation only, three possible observations are possible at a given methylable posisiton: 1) if the two cytosines from the two chromosomes are unmethylated, it is indicated as a 2, 2) if the two cytosines are methylated, it is indicated as a 3, and 3) if at a position a cytosine is methylated and the other one unmethylated, it is indicated as a 4. Depending on the mutation, methylation and, demethylation rates, different frequencies of these states are possible in the sample of sequences, which provide information on the emission rate in the SMC method. When both site- and region-level methylation processes occur, the methylation state is conditioned by the region level methylation state (increasing the number of possible observation to 9)

To choose the appropriate settings for SMCm (*i.e.* if there are region level epimutations), we test if the methylation state are distributed independently from one another along one genome. In absence of region methylation effect, the probability at each site (position) to be methylated or unmethylated should be independent from the previous position (or any other position). Conversely, if there is a region effect on epimutation, two consecutive sites along one genome would exhibit a positive correlation in their methylated states (and across pairs of sequences). We therefore calculate the probability that two successive positions with an annotated methylation state would be identical under a binomial distribution of methylation along a given genome. We then compare theoretical expectations to the observed data and build the statistical test based on a binomial distribution of probabilities. If existence of region level epimutation is detected, the regions level methylation states are recovered through a hidden markov model (HMM) similarly to [65, 18, 69]. Note that this HMM model does not include information from epimutation rates known from empirical studies. The complete description of the mathematical models and probabilities are in the supplementary material Text S1.

We postulate that the epimutation rates remain unknown in most species, while the DNA mutation rate may be known (or approximated based on a closely related species). Hence, we develop an approach based on the SMC capable of leveraging information from the distribution of DNA mutations to infer the epimutation rates (similar to what is described above). Our approach first tests if epimutations violates or not the infinite site assumptions. If less than 1% of sites with their methymation state annotated are polymorphic in a sequence we use the infinite site assumption: the site and region level epimutation rates can be recovered straightforwardly from the observed diversity (θ_W, see above) . Otherwise, a Baum-Welch algorithm is run to infer the most likely epimutation rates (site rate for SMP, and region rates for DMRs) [73, 74, 69].

Calculation of the root mean square error (RMSE)

To quantify the accuracy of each demographic inference we evaluate the root mean square error (RMSE). To do so we choose a hundred points uniformly spread across the time window (in $\log_{10}$ scale), and compare the actual population size and the one estimated by a given method at each of these points. We thus have the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{10^2}(y_i - y_i^*)^2}{10^2}}, \tag{2}$$

where y_i is the true population size at the time point i, and y_i^* is the estimated population size at the time point i.

Inference of the Time to the Most Recent Common Ancestor (TMRCA)

To infer the TMRCA at each position of the genome we use an approach similar to the PSMC' described in [58]. We first run a forward and backward algorithm on our sequence data (see appendix of [63, 71] for computation details). From the output results we calculate the probability to be in each hidden state at each position of the genome (note that the output product of the forward backward algorithm is rescaled so that the sum of probability is one), which we use to compute the expected coalescent time at each position on the genome using the following formula:

$$TMRCA_i = \sum_{j=1}^{n} fo_{i,j} \times ba_{i,j} \times Tc_j, \tag{3}$$

with i is the position on the genome, j is the hidden state index, n is the number of hidden state, fo is the output from the forward algorithm, ba is the output from the backward algorithm, $\sum_{j=1}^{n} fo_{i,j} \times ba_{i,j} = 1$, and Tc is a vector containing all the hidden states (*i.e.* coalescent times).

Sequence data of *Arabidopsis thaliana*

We download genome and methylome data of *A. thaliana* from the 1001 genome project [12]. We select 10 individuals from the German accessions respectively corresponding to the accession numbers: 9783, 9794, 9808, 9809, 9810, 9811, 9812, 9816, 9813, 9814. We only keep methylome data in CG context and in genic regions [74, 18]. The genic regions are based on the current reference genome TAIR 10.1. The SNPs and epimutations are called according to previously published pipeline [69, 18]. As in previous studies [63, 22, 19], we assume *A. thaliana* data to be haploid due to high homozygosity (caused by high selfing rate). The resulting files are available on GitHub at https://github.com/TPPSellinger. To perform analysis we chose $\mu = 6.95 \times 10^{-9}$ per generation per bp as the DNA mutation rate [52] and $r = 3.6 \times 10^{-8}$ as the recombination rate [56] per generation per bp. In order to have the most realistic model, we assume that the methylome of *A. thaliana* undergoes both region (RMM) and site (SMM) level epimutations [18]. When fixed, we respectively set the site methylation and demethylation rate to $\mu_{SM} = 3.48 \times 10^{-4}$ and $\mu_{SU} = 1.47 \times 10^{-3}$ per generation per bp according to [73]. We additionally set the region level methylation and demethylation rate to $\mu_{RM} = 1.6 \times 10^{-4}$ and $\mu_{RU} = 9.5 \times 10^{-4}$ per generation per bp according to [18]. Because we do not account for the effect of variable mutation or recombination rate along the genome, we cut the five chromosome of *A. thaliana* into eight smaller scaffolds [4, 5]. By doing this we remove centromeric regions and limit the effect the variation of mutation and recombination rate along the genome. The selected regions and the SNP density (from the German accessions) are represented in Supplementary Figures 11 to 15.

Data Availability

eSMC2 R package can be found at : https://github.com/TPPSellinger/eSMC2 . The input files created from *Arabidopsis thaliana* sequence data are available on GitHub at : https://github.com/TPPSellinger/Arabidopsis_thaliana_methylation .

Acknowledgments

We thank Zhilin Zhang and Rashmi Hazarika for giving and processing the data of *Arabidopsis thaliana.* TS is supported by the Deutsche Forschungsgemeinschaft, project number 317616126 (TE809/7-1) to AT, and the Austrian Science Fund (project no. TAI 151-B) to Anja Hörger.

References

[1] P. K. Albers and G. McVean. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLOS BIOLOGY*, 18(1), JAN 2020. ISSN 1544-9173. doi: 10.1371/journal.pbio.3000586.

[2] C. Alonso-Blanco, J. Andrade, C. Becker, F. Bemm, J. Bergelson, K. M. Borgwardt, J. Cao, E. Chae, T. M. Dezwaan, W. Ding, et al. 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. *Cell*, 166 (2):481–491, 2016.

[3] T. Anzai, T. Shiina, N. Kimura, K. Yanagiya, S. Kohara, A. Shigenari, T. Yamagata, J. K. Kulski, T. K. Naruse, Y. Fujimori, et al. Comparative sequencing of human and chimpanzee mhc class i regions unveils insertions/deletions as the major path to genomic divergence. *Proceedings of the National Academy of Sciences*, 100(13):7708–7713, 2003.

[4] G. V. Barroso and J. Y. Dutheil. Mutation rate variation shapes genome-wide diversity in *Drosophila melanogaster.* Sept. 2021. doi: 10.1101/2021.09.16.460667. URL `http://biorxiv.org/lookup/doi/10.1101/2021.09.16.460667`.

[5] G. V. Barroso, N. Puzovic, and J. Y. Dutheil. Inference of recombination maps from a single pair of genomes and its application to ancient samples. *PLOS Genetics*, 15(11), NOV 2019. ISSN 1553-7404. doi: {10.1371/journal.pgen.1008449;10.1371/journal.pgen.1008449.r001;10.1371/journal.pgen.1008449.r002;10.1371/journal.pgen.1008449.r003;10.1371/journal.pgen.1008449.r004}.

[6] F. Baumdicker, G. Bisschop, D. Goldstein, G. Gower, A. P. Ragsdale, G. Tsambos, S. Zhu, B. Eldon, E. C. Ellerman, J. G. Galloway, A. L. Gladstein, G. Gorjanc, B. Guo, B. Jeffery, W. W. Kretzschumar, K. Lohse, M. Matschiner, D. Nelson, N. S. Pope, C. D. Quinto-Cortes, M. F. Rodrigues, K. Saunack, T. Sellinger, K. Thornton, H. van Kemenade, A. W. Wohns, Y. Wong, S. Gravel, A. D. Kern, J. Koskela, P. L. Ralph, and J. Kelleher. Efficient ancestry and mutation simulation with msprime 1.0. *GENETICS*, 220(3), MAR 3 2022. ISSN 0016-6731. doi: 10.1093/genetics/iyab229.

[7] A. C. Beichman, E. Huerta-Sanchez, and K. E. Lohmueller. Using Genomic Data to Infer Historic Population Dynamics of Nonmodel Organisms. In Futuyma, DJ, editor, *Annual Review of Ecology, Evolution, and Systematics, VOL 49*, volume 49 of *Annual Review of Ecology Evolution and Systematics*, pages 433–456. 2018. ISBN 978-0-8243-1449-1. doi: {10.1146/annurev-ecolsys-110617-062431}.

[8] G. Bisschop, K. Lohse, and D. Setter. Sweeps in time: leveraging the joint distribution of branch lengths. *GENETICS*, 219(2), OCT 2021. ISSN 0016-6731. doi: 10.1093/genetics/iyab119.

[9] S. Boitard, W. Rodríguez, F. Jay, S. Mona, and F. Austerlitz. Inferring population size history from large samples of genome-wide molecular data - an approximate bayesian computation approach. 12(3):e1005877. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005877. URL `https://dx.plos.org/10.1371/journal.pgen.1005877`.

[10] D. Y. C. Brandt, X. Wei, Y. Deng, A. H. Vaughn, and R. Nielsen. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *GENETICS*, 221(1), MAY 5 2022. ISSN 0016-6731. doi: 10.1093/genetics/iyac044.

[11] A. Briffa, E. Hollwey, Z. Shahzad, J. D. Moore, D. B. Lyons, M. Howard, and D. Zilberman. Millennia-long epigenetic fluctuations generate intragenic dna methylation variance in arabidopsis populations. *Cell Systems*, 2023.

[12] J. Cao, K. Schneeberger, S. Ossowski, T. Günther, S. Bender, J. Fitz, D. Koenig, C. Lanz, O. Stegle, C. Lippert, et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, 43(10): 956–U60, OCT 2011. ISSN 1061-4036. doi: {10.1038/ng.911}.

[13] B. Charlesworth and D. Charlesworth. Elements of evolutionary genetics. 2010.

[14] B. Charlesworth and K. Jain. Purifying Selection, Drift, and Reversible Mutation with Arbitrarily High Mutation Rates. *Genetics*, 198(4):1587+, DEC 2014. ISSN 0016-6731. doi: {10.1534/genetics.114.167973}.

[15] B. Charlesworth and J. D. Jensen. Population genetic considerations regarding evidence for biased mutation rates in arabidopsis thaliana. *Molecular Biology and Evolution*, 40(2):msac275, 2023.

[16] S. J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini, and S. E. Jacobsen. Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature*, 452(7184):215–219, 2008.

[17] Y. Deng, Y. S. Song, and R. Nielsen. The distribution of waiting distances in ancestral recombination graphs. *THEORETICAL POPULATION BIOLOGY*, 141:34–43, OCT 2021. ISSN 0040-5809. doi: {10.1016/j.tpb.2021.06.003}.

[18] J. Denkena, F. Johannes, and M. Colome-Tatche. Region-level epimutation rates in arabidopsis thaliana. *HEREDITY*, 127(2):190–202, AUG 2021. ISSN 0018-067X. doi: 10.1038/s41437-021-00441-w.

[19] A. Durvasula, A. Fulgione, R. M. Gutaker, S. I. Alacakaptan, P. J. Flood, C. Neto, T. Tsuchimatsu, H. A. Burbano, F. X. Picó, C. Alonso-Blanco, et al. African genomes illuminate the early history and transition to selfing in arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, 114(20): 5213–5218, 2017.

[20] A. Estoup, P. Jarne, and J.-M. Cornuet. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular ecology*, 11(9):1591–1604, 2002.

[21] O. François, M. G. B. Blum, M. Jakobsson, and N. A. Rosenberg. Demographic history of european populations of arabidopsis thaliana. *PLOS Genetics*, 4:1–15, 05 2008. URL `https://doi.org/10.1371/journal.pgen.1000075`.

[22] A. Fulgione, M. Koornneef, F. Roux, J. Hermisson, and A. M. Hancock. Madeiran Arabidopsis thaliana Reveals Ancient Long-Range Colonization and Clarifies Demography in Eurasia. *Molecular Biology and Evolution*, 35(3):564–574, MAR 2018. ISSN 0737-4038. doi: {10.1093/molbev/msx300}.

[23] L. Gattepaille, T. Guenther, and M. Jakobsson. Inferring Past Effective Population Size from Distributions of Coalescent Times. *Molecular Biology and Evolution*, 204(3):1191+, NOV 2016. ISSN 0016-6731. doi: {10.1534/genetics.115.185058}.

[24] L. M. Gattepaille, M. Jakobsson, and M. G. B. Blum. Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. *Heredity*, 110(5):409–419, MAY 2013. ISSN 0018-067X. doi: {10.1038/hdy.2012.120}.

[25] R. R. Hazarika, M. Serra, Z. Zhang, Y. Zhang, R. J. Schmitz, and F. Johannes. Molecular properties of epimutation hotspots. *Nature Plants*, 8(2):146–156, 2022.

[26] M. J. Hubisz, A. L. Williams, and A. Siepel. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLOS GENETICS*, 16(8), AUG 2020. ISSN 1553-7404. doi: {10.1371/journal.pgen.1008895}.

[27] R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, 1983. ISSN 0040-5809. doi: {10.1016/0040-5809(83)90013-8}.

[28] F. Johannes. DNA methylation makes mutational history. *Nature Plants*, 5 (8):772–773, AUG 2019. ISSN 2055-026X. doi: {10.1038/s41477-019-0491-z}.

[29] F. Johannes and R. J. Schmitz. Spontaneous epimutations in plants. *New Phytologist*, 221(3):1253–1259, FEB 2019. ISSN 0028-646X. doi: {10.1111/nph.15434}.

[30] P. Johri, B. Charlesworth, and J. D. Jensen. Toward an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. *GENETICS*, 215(1):173–192, MAY 2020. ISSN 0016-6731. doi: 10.1534/genetics.119.303002.

[31] P. Johri, K. Riall, H. Becher, L. Excoffier, B. Charlesworth, and J. D. Jensen. The impact of purifying and background selection on the inference of population history: Problems and prospects. *MOLECULAR BIOLOGY AND EVOLUTION*, 38(7):2986–3003, JUL 2021. ISSN 0737-4038. doi: 10.1093/molbev/msab050.

[32] P. Johri, C. F. Aquadro, M. Beaumont, B. Charlesworth, L. Excoffier, A. Eyre-Walker, P. D. Keightley, M. Lynch, G. McVean, B. A. Payseur, S. P. Pfeifer, W. Stephan, and J. D. Jensen. Recommendations for improving statistical inference in population genomics. *PLOS Biology*, 20(5):e3001669, May 2022. ISSN 1545-7885. doi: 10.1371/journal.pbio.3001669. URL `https://dx.plos.org/10.1371/journal.pbio.3001669`.

[33] J. Kelleher, A. M. Etheridge, and G. McVean. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology*, 12(5), MAY 2016. doi: {10.1371/journal.pcbi.1004842}.

[34] J. Kelleher, Y. Wong, A. W. Wohns, C. Fadil, P. K. Albers, and G. McVean. Inferring whole-genome histories in large population datasets (vol 51, pg 1330, 2019). *Nature Genetics*, 51(11):1660, NOV 2019. ISSN 1061-4036. doi: {10.1038/s41588-019-0523-7}.

[35] C. Ki and J. Terhorst. Exact decoding of the sequentially Markov coalescent. Sept. 2020. doi: 10.1101/2020.09.21.307355. URL http://biorxiv.org/lookup/doi/10.1101/2020.09.21.307355.

[36] J. Kingman. The Coalescent . *Stochastic Processes and their Applications*, 13, 1982.

[37] K. Korfmann, T. P. P. Sellinger, F. Freund, M. Fumagalli, and A. Tellier. Simultaneous inference of past demography and selection from the ancestral recombination graph under the beta coalescent. *bioRxiv*, 2022.

[38] K. Korfmann, O. E. Gaggiotti, and M. Fumagalli. Deep Learning in Population Genetics. *Genome Biology and Evolution*, 15(2), 01 2023. ISSN 1759-6653. doi: 10.1093/gbe/evad008. URL https://doi.org/10.1093/gbe/evad008. evad008.

[39] D. Lang, S. Zhang, P. Ren, F. Liang, Z. Sun, G. Meng, Y. Tan, X. Li, Q. Lai, L. Han, D. Wang, F. Hu, W. Wang, and S. Liu. Comparison of the two up-to-date sequencing technologies for genome assembly: Hifi reads of pacific biosciences sequel ii system and ultralong reads of oxford nanopore. *GIGASCIENCE*, 9(12), DEC 2020. ISSN 2047-217X. doi: 10.1093/gigascience/giaa123.

[40] H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–U84, JUL 28 2011. ISSN 0028-0836. doi: {10.1038/nature10231}.

[41] R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133(3):523–536, MAY 2 2008. ISSN 0092-8674. doi: {10.1016/j.cell.2008.03.029}.

[42] D. B. Lyons, A. Briffa, S. He, J. Choi, E. Hollwey, J. Colicchio, I. Anderson, X. Feng, M. Howard, and D. Zilberman. Extensive de novo activity stabilizes epigenetic inheritance of cg methylation in arabidopsis transposons. *bioRxiv*, 2022. doi: 10.1101/2022.04.19.488736. URL https://www.biorxiv.org/content/early/2022/04/19/2022.04.19.488736.

[43] D. B. Lyons, A. Briffa, S. He, J. Choi, E. Hollwey, J. Colicchio, I. Anderson, X. Feng, M. Howard, and D. Zilberman. Extensive de novo activity stabilizes epigenetic inheritance of cg methylation in arabidopsis transposons. *Cell Reports*, 42(3), 2023.

[44] A. Mahmoudi, J. Koskela, J. Kelleher, Y.-b. Chan, and D. Balding. Bayesian inference of ancestral recombination graphs. *PLOS Computational Biology*, 18 (3):e1009960, 2022.

[45] A.-S. Malaspinas, M. C. Westaway, C. Muller, V. C. Sousa, O. Lao, I. Alves, A. Bergström, G. Athanasiadis, J. Y. Cheng, J. E. Crawford, et al. A genomic history of aboriginal australia. *Nature*, 538(7624):207–214, 2016.

[46] P. Marjoram and J. Wall. Fast "coalescent" simulation. *BMC Genetics*, 7, MAR 15 2006. ISSN 1471-2156. doi: {10.1186/1471-2156-7-16}.

[47] G. McVean and N. Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 360(1459):1387–1393, JUL 29 2005. ISSN 0962-8436. doi: {10.1098/rstb.20053.1673}.

[48] J. G. Monroe, T. Srikant, P. Carbonell-Bejerano, C. Becker, M. Lensink, M. Exposito-Alonso, M. Klein, J. Hildebrandt, M. Neumann, D. Kliebenstein, M.-L. Weng, E. Imbert, J. Agren, M. T. Rutter, C. B. Fenster, and D. Weigel. Mutation bias reflects natural selection in arabidopsis thaliana. *NATURE*, 602 (7895):101+, FEB 3 2022. ISSN 0028-0836. doi: 10.1038/s41586-021-04269-6.

[49] A. Muyle, J. Ross-Ibarra, D. K. Seymour, and B. S. Gaut. Gene body methylation is under selection in arabidopsis thaliana. *Genetics*, 218(2):iyab061, 2021.

[50] M. Nordborg. Linkage disequilibrium, gene trees and selfing: An ancestral recombination graph with partial self-fertilization. *Molecular Biology and Evolution*, 154(2):923–929, FEB 2000. ISSN 0016-6731.

[51] S. Nurk, B. P. Walenz, A. Rhie, M. R. Vollger, G. A. Logsdon, R. Grothe, K. H. Miga, E. E. Eichler, A. M. Phillippy, and S. Koren. Hicanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *GENOME RESEARCH*, 30(9):1291–1305, SEP 2020. ISSN 1088-9051. doi: 10.1101/gr.263566.120.

[52] S. Ossowski, K. Schneeberger, J. I. Lucas-Lledo, N. Warthmann, R. M. Clark, R. G. Shaw, D. Weigel, and M. Lynch. The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis thaliana. *Science*, 327(5961):92–94, JAN 1 2010. ISSN 0036-8075. doi: {10.1126/science.1180677}.

[53] S. Ou, W. Su, Y. Liao, K. Chougule, J. R. A. Agda, A. J. Hellinga, C. S. B. Lugo, T. A. Elliott, D. Ware, T. Peterson, N. Jiang, C. N. Hirsch, and M. B. Hufford. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *GENOME BIOLOGY*, 20(1), DEC 16 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1905-y.

[54] R. Pisupati, V. Nizhynska, A. Mollá Morales, and M. Nordborg. On the causes of gene-body methylation variation in arabidopsis thaliana. *PLoS genetics*, 19 (5):e1010728, 2023.

[55] W. Rodriguez, O. Mazet, S. Grusea, A. Arredondo, J. M. Corujo, S. Boitard, and L. Chikhi. The IICR and the non-stationary structured coalescent: towards demographic inference with arbitrary changes in population structure. *Heredity*, 121(6):663–678, DEC 2018. ISSN 0018-067X. doi: {10.1038/s41437-018-0148-0}.

[56] P. A. Salome, K. Bomblies, J. Fitz, R. A. E. Laitinen, N. Warthmann, L. Yant, and D. Weigel. The recombination landscape in Arabidopsis thaliana F-2 populations. *Heredity*, 108(4):447–455, APR 2012. ISSN 0018-067X. doi: {10.1038/hdy.2011.95}.

[57] A. Sand, M. Kristiansen, C. N. S. Pedersen, and T. Mailund. zipHMMlib: a highly optimised HMM library exploiting repetitions in the input to speed up the forward algorithm. *BMC Bioinformatics*, 14, NOV 22 2013. ISSN 1471-2105. doi: {10.1186/1471-2105-14-339}.

[58] S. Schiffels and R. Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–925, AUG 2014. ISSN 1061-4036. doi: {10.1038/ng.3015}.

[59] M. W. Schmid, C. Heichinger, D. Coman Schmid, D. Guthörl, V. Gagliardini, R. Bruggmann, S. Aluri, C. Aquino, B. Schmid, L. A. Turnbull, et al. Contribution of epigenetic variation to adaptation in arabidopsis. *Nature Communications*, 9(1):1–12, 2018.

[60] 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, et al. Patterns of population epigenomic diversity. *Nature*, 495(7440):193–198, 2013.

[61] J. G. Schraiber and J. M. Akey. Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 16(12):727–740, DEC 2015. ISSN 1471-0056. doi: {10.1038/nrg4005}.

[62] R. Schweiger and R. Durbin. Ultra-fast genome-wide inference of pairwise coalescence times. *bioRxiv*, 2023.

[63] T. P. P. Sellinger, D. Abu Awad, M. Moest, and A. Tellier. Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data. *PLOS Genetics*, 16(4), APR 2020. ISSN 1553-7404. doi: {10.1371/journal.pgen.1008698;10.1371/journal.pgen.1008698.r001;10.1371/journal.pgen.1008698.r002;10.1371/journal.pgen.1008698.r003;10.1371/journal.pgen.1008698.r004;10.1371/journal.pgen.1008698.r005;10.1371/journal.pgen.1008698.r006}.

[64] T. P. P. Sellinger, D. Abu-Awad, and A. Tellier. Limits and convergence properties of the sequentially markovian coalescent. *MOLECULAR ECOLOGY RESOURCES*, 21(7):2231–2248, OCT 2021. ISSN 1755-098X. doi: {10.1111/1755-0998.13416}.

[65] Y. Shahryary, A. Symeonidi, R. R. Hazarika, J. Denkena, T. Mubeen, B. Hofmeister, T. van Gurp, M. Colome-Tatch, K. J. F. Verhoeven, G. Tuskan, R. J. Schmitz, and F. Johannes. Alphabeta: computational inference of epimutation rates and spectra from high-throughput dna methylation data in plants. *GENOME BIOLOGY*, 21(1), OCT 6 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02161-6.

[66] L. Speidel, M. Forest, S. Shi, and S. R. Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321+, SEP 2019. ISSN 1061-4036. doi: {10.1038/s41588-019-0484-x}.

[67] T. Srikant and H.-G. Drost. How stress facilitates phenotypic innovation through epigenetic diversity. *Frontiers in Plant Science*, 11:606800, 2021.

[68] S. Strütt, T. Sellinger, S. Glémin, A. Tellier, and S. Laurent. Joint inference of evolutionary transitions to self-fertilization and demographic history using whole-genome sequences. *Elife*, 12:e82384, 2023.

[69] A. Taudt, D. Roquis, A. Vidalis, R. Wardenaar, F. Johannes, and M. Colome-Tatche. Methimpute: imputation-guided construction of complete methylomes from wgbs data. *BMC GENOMICS*, 19, JUN 7 2018. ISSN 1471-2164. doi: 10.1186/s12864-018-4641-x.

[70] A. Tellier, S. J. Y. Laurent, H. Lainer, P. Pavlidis, and W. Stephan. Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proceedings of the National Academy of Sciences of the United States of America*, 108(41):17052–17057, OCT 11 2011. ISSN 0027-8424. doi: {10.1073/pnas.1111266108}.

[71] J. Terhorst, J. A. Kamm, and Y. S. Song. Robust and scalable inference of population history froth hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303–309, FEB 2017. ISSN 1061-4036. doi: {10.1038/ng.3748}.

[72] G. Upadhya and M. Steinrücken. Robust Inference of Population Size Histories from Genomic Sequencing Data. May 2021. doi: 10.1101/2021.05.22.445274. URL `http://biorxiv.org/lookup/doi/10.1101/2021.05.22.445274`.

[73] van der Graaf et al. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proceedings of the National Academy of Sciences of the United States of America*, 112(21):6676–6681, MAY 26 2015. ISSN 0027-8424. doi: {10.1073/pnas.1424254112}.

[74] A. Vidalis, D. Zivkovic, R. Wardenaar, D. Roquis, A. Tellier, and F. Johannes. Methylome evolution in plants. *Genome Biology*, 17, DEC 20 2016. ISSN 1474-760X. doi: {10.1186/s13059-016-1127-5}.

[75] J. Wakeley. Coalescent theory: an introduction. roberts and company. *Greenwood VillageWayne AF, Maxwell MA, Ward CG, Vellios CV, Wilson I, Wayne JC, Williams MR (2015) Sudden and rapid decline of the abundant marsupial Bettongia penicillata in Australia. Oryx*, 49:175185Webb, 2008.

[76] C. Wang and C. Liang. Msipred: a python package for tumor microsatellite instability classification from tumor mutation annotation data using a support vector machine. *SCIENTIFIC REPORTS*, 8, DEC 3 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-35682-z.

[77] J. Wang and C. Fan. A neutrality test for detecting selection on dna methylation using single methylation polymorphism frequency spectrum. *GENOME BIOLOGY AND EVOLUTION*, 7(1):154–171, JAN 2015. ISSN 1759-6653. doi: 10.1093/gbe/evu271.

[78] D. Weigel and V. Colot. Epialleles in plant evolution. *Genome biology*, 13 (10):1–6, 2012.

[79] C. Wiuf and J. Hein. Recombination as a point process along sequences. *Theoretical Population Biology*, 55(3):248–259, JUN 1999. ISSN 0040-5809. doi: {10.1006/tpbi.1998.1403}.

[80] A. W. Wohns, Y. Wong, B. Jeffery, A. Akbari, S. Mallick, R. Pinhasi, N. Patterson, D. Reich, J. Kelleher, and G. McVean. A unified genealogy of modern and ancient genomes. *SCIENCE*, 375(6583):836+, FEB 25 2022. ISSN 0036-8075. doi: 10.1126/science.abi8264.

[81] R. Yang, J. L. Van Etten, and S. M. Dehm. Indel detection from dna and rna sequencing data with transindel. *BMC GENOMICS*, 19, APR 19 2018. ISSN 1471-2164. doi: 10.1186/s12864-018-4671-4.

[82] Z. Yang. Statistical properties of a DNA sample under the finite-sites model. *Genetics*, 144(4):1941–1950, DEC 1996. ISSN 0016-6731.

[83] N. Yao, R. J. Schmitz, and F. Johannes. Epimutations define a fast-ticking molecular clock in plants. *Trends in Genetics*, 37(8):699–710, 2021.

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

[85] X. Zhang, J. Yazaki, A. Sundaresan, S. Cokus, S. W.-L. Chan, H. Chen, I. R. Henderson, P. Shinn, M. Pellegrini, S. E. Jacobsen, et al. Genome-wide high-resolution mapping and functional analysis of dna methylation in arabidopsis. *Cell*, 126(6):1189–1201, 2006.

[86] Y. Zhang, S. Wang, and X. Wang. Data-driven-based approach to identifying differentially methylated regions using modified 1d ising model. *BIOMED RESEARCH INTERNATIONAL*, 2018, 2018. ISSN 2314-6133. doi: 10.1155/2018/1070645.

[87] D. Zilberman, M. Gehring, R. K. Tran, T. Ballinger, and S. Henikoff. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genetics*, 39(1): 61–69, JAN 2007. ISSN 1061-4036. doi: {10.1038/ng1929}.