algorithms

MDPI

Article

Multiple Loci Selection with Multi-Way Epistasis in Coalescence with Recombination

Aritra Bose †, Filippo Utro †, Daniel E. Platt and Laxmi Parida *

Computational Genomics, IBM Thomas J. Watson Research Center, Ossining, NY 10562, USA;
A.Bose@ibm.com (A.B.); futro@us.ibm.com (F.U.); watplatt@us.ibm.com (D.E.P.)
* Correspondence: parida@us.ibm.com
† These authors contributed equally to this work.

Abstract: As studies move into deeper characterization of the impact of selection through non-neutral mutations in whole genome population genetics, modeling for selection becomes crucial. Moreover, epistasis has long been recognized as a significant component in understanding the evolution of complex genetic systems. We present a backward coalescent model, EpiSimRA, that accommodates multiple loci selection, with multi-way (k-way) epistasis for any arbitrary k. Starting from arbitrary extant populations with epistatic sites, we trace the Ancestral Recombination Graph (ARG), sampling relevant recombination and coalescent events. Our framework allows for studying different complex evolutionary scenarios in the presence of selective sweeps, positive and negative selection with multiway epistasis. We also present a forward counterpart of the coalescent model based on a Wright-Fisher (WF) process, which we use as a validation framework, comparing the hallmarks of the ARG between the two. We provide the first framework that allows a nose-to-nose comparison of multiway epistasis in a coalescent simulator with its forward counterpart with respect to the hallmarks of the ARG. We demonstrate, through extensive experiments, that EpiSimRA is consistently superior in terms of performance (seconds vs. hours) in comparison to the forward model without compromising on its accuracy.

Keywords: coalescent theory; natural selection; epistasis; recombination event; ancestral recombination graph

check for updates

Citation: Bose, A.; Utro, F.; Platt, D.E.; Parida, L. Multiple Loci Selection with Multi-Way Epistasis in Coalescence with Recombination. *Algorithms* **2021**, *14*, 136. https://doi.org/10.3390/a14050136

Academic Editor: Tatsuya Akutsu

Received: 31 March 2021
Accepted: 24 April 2021
Published: 25 April 2021

1. Introduction

EpiSimRA (both backward and forward) source, executable, user manuals are available at: https://github.com/ComputationalGenomics/SimRA (accessed on 24 April 2021). Nothing in Biology Makes Sense Except in the Light of Evolution [1], and simulating the evolution process, whether of multi-cellular humans, unicellular micro-organisms or even cancer-tumors, continues to be an important device in understanding the observed molecular profiles of populations. These profiles are an attribute of the genetic variability due to mutations and the change in frequency of alleles within populations over time. The selectively neutral infinite-sites model [2] is often used to analyze this variation [3]. Simulating random populations plays a significant role in investigating the effects of complex evolutionary processes on genetic diversity [4]. There are mainly two types of simulation algorithms: backward-time or coalescent and forward-time. The coalescent simulation [5] allows for fast approximation of the neutral Wright-Fisher (WF) model with natural selection, shaping patterns of variation in populations. The ARG [6] is a variant of Kingman's coalescent and is used to reconstruct the most recent common ancestor (MRCA), starting backwards from the extant populations or leaves, using coalescent and recombination events. Once it finds the MRCA or if it involves all the trees, the grand most recent common ancestor (GMRCA), it traverses the ARG to the extant populations and introduce mutations or other genetic information in the genealogy. Forward-time simulators are more precise than their backward (coalescent) counterparts in modeling selection along with other factors as it starts from an initial population and tracks its evolution under the influence of various factors such as recombination, mutations, varying effective population size, fitness effects, and so forth. It progresses over multiple sequential generations, usually drawing random samples from the last generation to construct an ARG and its hallmarks. However, coalescent processes are much faster than forward-time simulation algorithms [7].

The classical approach for coalescent simulation as defined by Hudson in the seminal *ms* tool [3], applied the effects of recombination and coalescence to the ancestors of the samples going back in time in the extant population. This was later computed more efficiently in *msprime* [8], which used a new encoding for correlated trees resulting from simulations of the coalescent with recombination. Some approximations to the coalescent algorithms, which are fast, also exist, such as *SMC* [9], *MaCS* [10] and *fastsimcoal* [11]. Many programs have been developed to simulate scenarios not captured by *ms* such as selection [12–15], demographic inference [12,16] and admixture [17] among others. Coalescent models tracking genealogies in the presence of selection can also build an Ancestral Selection Graph (ASG), which is a branching-coalescing random graph within which the genealogy of a sample is embedded [18] conditional on the frequencies of the selected allele of the sample [19]. However, none of these methods take into account epistasis, which has long been recognized as a significant component in understanding genealogies and the evolution of complex genetic systems [20]. Here, we present the first coalescent simulator EpiSimRA which captures multiway epistasis, that is, allowing for interaction between alleles in multiple loci under selection. EpiSimRA tracks the ARG from randomly sampled extant populations and unlike ASG, is not conditional on allele frequencies. It constructs the genealogy dependent on the time to the closest recombination and coalescent event going backwards. Along with this, we also present an alternative, simple forward-time algorithm fwd-EpiSimRA, which efficiently simulates epistatic scenarios, to provide a validation framework to the coalescent simulator.

Forward simulators usually track the complete ancestral information, that is, studying all the lineages that survived until the last generation as a result of recombination events. Although forward simulations have been around for decades, only a few forward simulators exist to provide a framework to model multi-way epistasis, such as *SELAM* [21], allowing for pairwise epistatic selection to model the process and consequences of admixture or *SLiM* [22,23], which constructs ecologically realistic scenarios while accounting for a host of complex biological processes beyond the WF framework. Specifically, its functionality of tree-sequence recording draws parallels to fwd-EpiSimRA in a WF framework, providing support for epistatic interactions. However, none of these packages [21,23] can be used to compare with EpiSimRA as it is not possible to reconstruct the ARG from random extant samples. As fwd-EpiSimRA traces the ARG to obtain the MRCA of the random extant samples and record its hallmarks, we use it for a nose-to-nose comparison with the coalescent simulator as a validation framework. In the remainder of the paper we introduce the coalescent simulator and explain how it tracks multiway epistasis in the presence of recombination, followed by an overview of the forward simulator and the ARG tracking algorithm. Thereafter, we show the concordance between the coalescent and forward models for complex evolutionary scenarios and finally conclude by discussing multiway epistasis in simulating real world scenarios of admixture, cryptic relatedness and viral phylodynamics.

2. Materials and Methods

2.1. The Coalescent Simulator

The algorithm works back-in-time starting from the present (time 0), moving back into the past. Here we focus on how EpiSimRA is able to simulate multiway epistasis (the interested reader is referred to [17] for the neutral scenario). Let the number of loci under selection be l, possibly with multiway epistasis. As an illustration let l be 3 with selection values s_1, s_2 and s_3. The algorithm will assign three random locations on the genetic segment, unless the locations are explicitly specified and we assume that one of the alleles (either major or minor) is under selection while the other is neutral. The possible multiway epistasis are e_{12}, e_{13}, e_{23} and e_{123}. If no value is specified then the epistasis is assumed to be neutral. Given this, we get 2^l possible types of lineages, each of them denoted as l_z. Let l_0 be the lineage type with no selection. For the example we ran, the other lineage types are l_1, l_2, l_3, l_{12}, l_{13}, l_{23} and l_{123}. For two lineage types z_a and z_b, let

$$l_{z_a} \prec l_{z_b} \text{ when } z_a \supset z_b.$$

For example, $l_{12} \prec l_1$ and $l_{12} \prec l_2$. Also, $l_{123} \prec l_{12}$. For the lineage type z, let N_z be the effective population size.

2.1.1. Selection Scenarios

Effective population size is the reciprocal of the probability that two individuals will have the same parent—or that two chromosomes in the next generation will share the same parent chromosome. Fitness, in this case, would just be the ratio of the probability that the parent chromosome lineage with the allele will have an offspring chromosome to the probability that a parent chromosome lineage without the allele (neutral) will have an offspring chromosome [24]. The fitness, $1+s$ is thus the expected number of copies that a copy of the allele gives rise to in generation $t+1$, relative to the expected number that a neutral allele will give rise to. Thus, in infinitely large populations, the proportion of alleles under selection in a generation is related to the effective population size. Let N_s be the partially effective population size with the allele under selection and $N_{\bar{s}}$ $(= N - N_s)$ is the partially effective population size with the reference or ancestral allele which is not under selection, giving:

$$\frac{N_s}{N_s + N_{\bar{s}}} : \frac{N_{\bar{s}}}{N_s + N_{\bar{s}}} = 1 + s : 1 \implies N_s = \frac{1+s}{2+s} N = f_s N. \tag{1}$$

Thus $-1 < s$, extendable to multiple loci with or without epistasis and the fitness defined as

$$f_s = \frac{1+s}{2+s}. \tag{2}$$

The fitness coefficient is a representative average of the allele frequency of the selected alleles in a generation, p. With the allele frequency, the effective population size with selection at a single locus would be $\mathbb{E}[N_s] = \frac{(1+s)pN}{(1+ps)}$ where the fitness $f_s = \frac{p(1+s)}{(1-p)}$. The effective population size is defined as $N_z = 2Nf'$ where f' is the fitness for l_z-coalescence in the coalescent simulator and N_z is the effective population size for l_z lineage coalescence with alleles under selection.

The f' varies with neutral or epistatic scenarios for the loci under selection. For a neutral scenario with no selection f' is defined as follows

$$f' = 1 - \sum_i f_{s_i} + \sum_{i,j} f_{s_{ij}} - f_{s_{\alpha\beta\gamma}}, \tag{3}$$

where we remove the fitness effects of odd sites under selection and add the effects from even sites in a simulation scenario with three Single Nucleotide Polymorphisms or SNPs (α, β and γ) are considered to be under selection. Alternatively, for a single locus (α) under selection with no epistasis in effect f' will be defined as follows with the signs reversed for even and odd sites under selection

$$f' = f_{s_\alpha} - \sum_i f_{s_{\alpha i}} + \sum_{i,j} f_{s_{\alpha ij}}. \tag{4}$$

For two or multiple loci under selection there can be two cases with differing f_s. We define it as follows:

$$f_s = \begin{cases} f_{e_{s_{\alpha\beta..\omega}}} & \text{with epistasis} \\ \prod_i^{\omega} f_{s_i} & \text{without epistasis,} \end{cases}$$

where e_s is the user defined epistatic coefficient when epistasis is in effect across all ω sites. For a scenario with all three sites are under selection $f' = f_{s_{\alpha\beta\gamma}}$.

2.1.2. EpiSimRA: Multiple Loci Selection & Multiway Epistasis

If s_i and s_j are two locations with the minimum (or derived) allele under selection at locus i and j respectively, then e_{ij} denotes the epistasis between the two. The fitness coefficients in Equation (2) takes into account the e_{ij} with respect to the s. If it is not explicitly specified then a neutral case (without epistasis) is assumed. The algorithm randomly chooses the location of the SNPs on the genetic segment being simulated.

We assume that no more than one event, coalescent or recombination, occurs at a generation and there is no back mutation, that is, a base undergoes no more than one mutation in the entire ARG. The mutation and recombination rates are uniform over the segment being simulated. If there is recombination, the lineages are randomly assigned but if $r = 0$, the lineages are so assigned that no pair of types of lineages straddle (either they are disjoint or one is contained in the other). Lineage l_0 corresponds to lineage with no alleles under selection. For each lineage l_z, the algorithm only appends each node to a list when a recombination occurs with time $t > T_z$ where T_z keeps track of the time to GMRCA. The recombination rate for l_z, r'_l is defined as

$$r'_l = N_z g r I.$$

For nodes which are not leaf nodes the length of the genetic material, s is proportional to the recombination rate as the rate is governed by the effective population size N_z. The stochastic nature of the method allows for a loop which pools lineages together at each iteration to find the event closest to the time t_z, over all lineages l_z. For each lineage $t_z = N_z \times t$ is computed where t is the time to next event using

$$\begin{aligned} t &= \min\left(\overbrace{\min_{1\le a<b\le L_z}(t_{ab}^{\text{coal}})}, \underbrace{\min_{1\le i\le L_z}(t_i^{\text{rcmb}})}\right) \\ &= \text{Exp}\left(\overbrace{1+1+...+1}+\underbrace{r'_1+r'_2+..+r'_{L_z}}\right) \\ &= \text{Exp}\left(\overbrace{L_z}+\underbrace{r'_1+r'_2+..+r'_{L_z}}\right). \end{aligned} \tag{5}$$

t^{coal} is the time to coalescence and similarly t^{rcmb} is the time to the next recombination event. Equation (5) computes the closest event to this time (coalescent or recombination) where the overbraces capture the $\binom{L_z}{2}$ coalescent events and the underbraces capture the L_Z recombination events. When there is only one lineage in the pool, only recombination event can occur. Otherwise, we use the three properties outlined in the Appendix A to find the next event closest to the time computed in Equation (5). The time T, aggregated over t is thus the time to GMRCA as outlined in Algorithm 1. The event is a coalescence is chosen with the probability

$$\frac{\binom{L_z}{2}}{\binom{L_z}{2} + \sum_l r'_l} \tag{6}$$

and recombination at lineage $1 \le k \le L_z$ with probability,

$$\frac{r'_k}{\binom{L}{2} + \sum_l r'_l}. \tag{7}$$

Equations (6) and (7) are used in a single draw of a random number such as in unit interval [0,1] broken up into $1 + L_z$ sub intervals with cumulative ratio. The first interval implies a coalescent event and k_{th} interval ($k > 1$) implies a recombination at the lineage l_{k-1}. Since the events are randomly selected, t is estimated first and then the lineages are picked at random from L_z active lineages. If the T_z falls in the respective interval for a coalescent event, then the next event is coalescence and otherwise, if $r > 0$, then the next event is recombination.

```
Algorithm 1: EpiSimRA
Input  : Parameters from Table 1
Output: T, time to GMRCA
$L = \bigcup_z l_z$
foreach lineage $l_z$ until $L_0 = 1$ do
    $T_z = 0, C_z = \{\}$
    $r'_l = N_z g r I$
    $t_z = N_z \times t$
    Compute $t$ as per Equation (5)
    $T_z = T_z + t_z$.
end
```

Coalescence Event

In a coalescence event L_z is decremented by 1 as two random lineages of type l_z are coalesced into one at time T_z and the outgoing edge of the coalesced node is labeled by lineage l_z. If $|L| = 1$, z is a singleton label (such as s_1 but not $s_1 s_2$ or $s_1 s_2 s_3$), and, there exist no active lineage l'_z such that $z' \prec z$, then the mutation(s) corresponding to lineage l_z is assigned to this edge (using an approach in [17]) and the label of the outgoing edge of the new node is changed to l_0 and L is incremented by 1. Next, L is set to 0 and thus the lineage l_z is made inactive.

Recombination Event

In a recombination event a lineage of type l_z is randomly picked and a node v is created at T_z. The label of z is randomly split it into two lineage labels that is compatible with the location of the SNPs on the genetic segment I carried by the node v. Thereafter, L is incremented by 1.

The algorithm for EpiSimRA is described in Algorithm 1 and see Appendix A for an illustrative example of the algorithm for a three-way epistatic scenario.

Table 1. Input parameters of the coalescent simulator.

	Parameters	Example Values	User-Specified Units	Units in bp for the Algorithm	Scaling Factor
g	seqment length	25; 75	Kb	$\times 10^3$ bp	$\times 10^3$
m	extant units	10; 20; 30; 40	-	-	$\times$ 1
N	population size	100; 200; 500; 1000	-	-	$\times$ 1
I	length of genetic material	1000	bp	1 bp	$\times$ 1
			rates/generation		
r	recombination rate	1	bp/gen $\times 10^{-7}$	bp/gen	$\times 10^{-7}$
μ	SNP mutation rate	1.5	mut/bp/gen $\times 10^{-8}$	$\times$1 mut/bp/gen	$\times 10^{-8}$
			selection, epistasis parameters		
s_i	fitness	0.3	-	$\times$ 1	
e_{ij}	epistasis	0.1, 0.15	-	-	$\times$ 1

2.2. *The Forward Simulator*

The model simulates evolution for a full population, forward in time with each generation containing N individuals with equal number of males and females, each carrying two chromosomes (see Appendix A for a detailed discussion and extension to selection on multiple loci). The complex evolutionary relationships between generations yields a number of mutations, recombinations, selected allele inheritance, linkage disequilibrium, and so forth, along the length of chromosome for each individual. These data are recorded in a data structure, which we call the "book of populations", keeping a record of the past genealogy of the population. We trace the lineage of each site along the chromosome while tracing the 'book' and constructing the ARG. Inheritance follows the convention of a standard WF model applied to diploid organisms [3], with children randomly picking their parents weighted by the fitness coefficients when selection is in effect.

2.2.1. Simulating the "Book of Populations"

Each chromosome is represented by the alleles at each locus $l \in [1, g]$, which is randomly assigned initially. We use same notations as defined in Table 1 to describe fwd-EpiSimRA. The model assumes that each locus l has a fitness function $s_l(a) \in \mathbb{R}$, where a is an allele comprising the genotype. An individual i with allele a_{il} at locus l is assigned a selection coefficient $s_{il} = s(a_{il})$ which is user-defined, similar to EpiSimRA. The function $s(.)$ denotes the selective pressure and can be varied by intentional specification of recessive, dominant, additive, and other configurations, including homozygous advantage. This function encompasses selection at both single and multiple loci allowing flexible user-defined variations. When selection is not present, we set $s_{il} = 0$.

For an individual i, the probability that it has children is given by

$$p_i = \frac{\prod_l (1 + s_{il})}{\sum_i \prod_l (1 + s_{il})}. \tag{8}$$

(See Appendix A for derivation). In each new generation, as in the WF model, the N children pick their parents with replacement according to the parent probabilities p_i. The simulation is run for $t = \{0, 1, \ldots, G\}$ discrete generations with the $t = 0$ being the base generation, outlined in Figure 1.

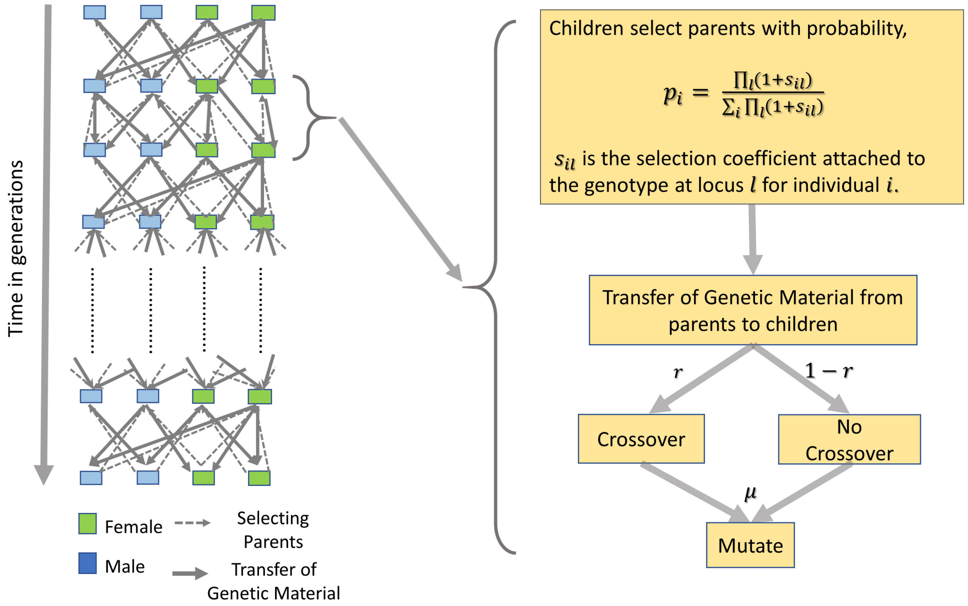
Time in generations
🟩 Female
🟦 Male
Selecting Parents
Transfer of Genetic Material
Children select parents with probability,
$p_i = \frac{\prod_l (1+s_{il})}{\sum_i \prod_l (1+s_{il})}$
s_{il} is the selection coefficient attached to the genotype at locus l for individual i.
Transfer of Genetic Material from parents to children
r
$1 - r$
Crossover
No Crossover
μ
Mutate

Figure 1. Schematic diagram for simulating the "book of populations" which closely resembles the biological process of evolution.

In each new generation, as in the WF model, the N children pick their parents with replacement according to the parent probabilities p_i. The simulation is run for $t = \{0, 1, \ldots, G\}$ discrete generations with the $t = 0$ being the base generation, outlined in Figure 1.

2.2.2. Modeling Multiway Epistasis

Multiway epistasis requires multiple interacting loci with similar selection effects. We assign selection coefficients to interacting sites for k-way epistasis, where k is the maximum number of interacting sites. Let there be q groups of loci, each containing at most k elements and we re-compute Equation (8) accounting for fitness related to interacting sites as,

$$p_i = \frac{\prod_q (1 + S_{iq})}{\sum_i \prod_q (1 + S_{iq})}. \tag{9}$$

If a group only has one element, that is if the selected locus is non-interacting, then we allow $S = s$, the user defined selection coefficient. For all other cases, we select S from a matrix or tensor of all possible allele combinations with respect to the number of interacting sites. S, the combined fitness coefficient is calculated by taking the fitness product of each interacting site as,

$$S = \prod_j \left(1 + \sum_i s_i^{(j)}\right) + e_q. \tag{10}$$

e_q is the epistatic interaction coefficient for each combination of interacting sites and $s_i^{(j)}$ is the selection coefficient at allele j in individual i's chromosome.

2.2.3. Tracing the ARG

Detecting the past recombination events from extant sequences and specifying the place of each recombination is well studied [25–27]. The ARGs define a genealogical graph for all of the chromosomes in a population. Recent advances in population genetics simulators have resulted in tree-sequence recordings, which obtains the genealogical history of all genomes in a simulated population [28]. However, no natural ARG is recorded for the interacting loci with epistasis in effect and randomly sampling populations from extant generation, in forward simulators. It is traced from the "book of populations" from a number of extant haplotypes. We start from m randomly selected extant populations and trace the recombination and coalescent events back each generation. We keep track of each lineage corresponding to every site along the chromosome and stop when we have found a convergence for all lineages. This final coalescent event along the entire "book" is known as MRCA and we output the corresponding ARG.

3. Results

3.1. Comparison Study

Comparing the two models under selection calls for an assessment of the values. In both the models, common phenomena such as faster coalescence, decreasing diversity, decreasing number of recombination events occur when we study the individuals under selection. Hence, we compare the H, the height of the ARG or the time to MRCA, as it is a significant hallmark of the common history of a sample.

We run simulations for different parameter set-ups for the forward and backward model by running each experiment 100 times. We demonstrate the accuracy of the two algorithms by comparing H under different simulation scenarios allowing at most three interacting loci. The simple scenarios in this case is when there is no selection in effect, that is, the neutral coalescent model and selection at a single locus. We show that the two proposed models EpiSimRA and fwd-EpiSimRA show agreement in all of the different epistatic scenarios including selection in single locus.

The results for the complex scenario in this setting, accounting for epistasis with three loci, are shown in Figure 2, where we show the concordance for the forward and backward simulation with box-whisker plots, QQ-plots, CDF plots and PP plots (Figure A5). To obtain further validation we observed similar agreement in the Kolmogorov-Smirnov (KS) test on the distributions of H as returned by fwd-EpiSimRA and EpiSimRA for all scenarios. We found that for each, the null hypothesis that the two samples are drawn from the same distribution is never rejected and the test statistic is very small (Table A1).

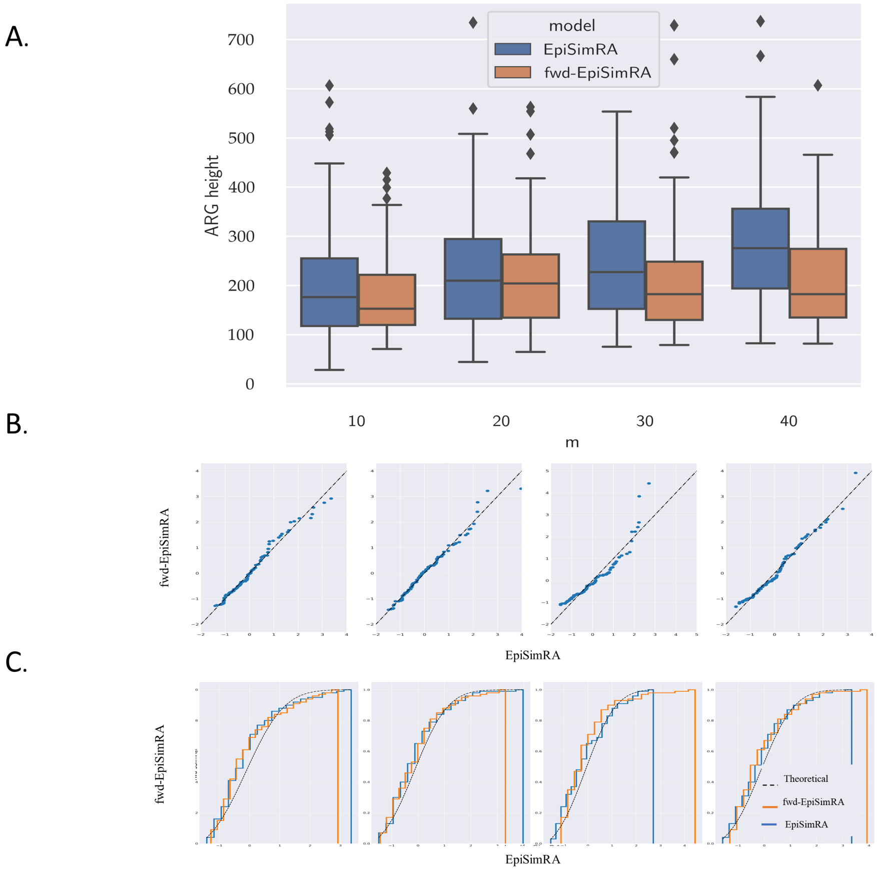

Figure 2. Comparison of the height of the ARG (H) between fwd-EpiSimRA and EpiSimRA with and without epistatis with recombination for $N = 100$, $g = 250$ kbp, $r = 1.0 \times 10^{-8}$, $m = \{10, 20, 30, 40\}$, $s = \{0.3, 0.3, 0.3\}$ with epistastic parameters for $s_i s_j = 0.15$ for $i, j \in [1, 3]$ and $s_1 s_2 s_3 = 0.125$. (**A**) The box-and-whisker diagram summarizes the result for each. On each box, the central mark is the mean, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points and outliers are plotted individually. (**B**) QQ plot and (**C**) Cumulative Distribution Function (CDF) plot of the backward and forward models show similar distributions with further agreement in Kolmogorov-Smirnov tests (Table A1).

3.2. *Evaluating Epistatic Scenarios*

We compare the H under selection in EpiSimRA and show how different scenarios impact the height of the ARG (Figure A1). We find that positive selection affects the time to coalescence inversely with more selective pressure results in less time to coalescence when simulated with $N = 1000$ samples and genome length of $g = 250$ kbp. Epistasis in two and three interacting loci results in lower time to MRCA (MRCA) than single locus selection and the neutral case amounting to higher selective pressure. We further studied effects of epistasis by simulating populations of $N = 10{,}000$ with three-way epistasis. We find epistasis leads to a more complex evolutionary history resulting in longer time to coalescence in MRCA (Figure A2 in Appendix A). When epistasis is not in effect, that is, when $f_{s^*} = \prod_i f_{s_i}$, when $i \in [1,3]$ results in a shorter TMRCA with simpler evolutionary history. In addition, an exhaustive comparison between the two simulators for all scenarios with or without epistasis is included in the Appendix A.

4. Discussion

Selection in a diploid organism impacts heritability two ways: (1) heterozygosity can mediate the impact of selection on the transmission of a selected haploid lineage, and (2) recombination can hide the impact selection may have had on the ARG. This model focuses more on the impact of the latter. Selection in a diploid heterozygous sample can boost, for one generation, the non-selected chromosome. This can complicate the impact of selection on lineages in the diploid forward model, but not the haploid. We expected the impact of boosted preference to be minimal along any given lineage since such a boost only occurs for dominant or additive alleles, and then for only one generation, with combinations in a population over time, this effect could be more significant.

The coalescent model and its forward counterpart under epistatic selection scenarios were concordant in the simulation studies. Epistasis makes the evolutionary history of extant populations more complex, but with selective pressure on certain alleles, the TMRCA appears to be shorter than the single locus selection or the neutral scenarios, respectively. It is expected that the H would be shorter for any selection scenario when compared to the neutral case, however, it is particularly intriguing to observe how the epistatic two and three-way scenarios have more selective pressure with a cumulative effect resulting in a decreased TMRCA. Although, the mean of fwd-EpiSimRA distribution (Figure 2) is a bit lower than EpiSimRA, we see concordance in the overall distribution, including the QQ and CDF plots and as well as in the KS tests. We posit that the difference in mean may be due to underlying differences in model assumptions such as diploid mechanisms for fwd-EpiSimRA in comparison to a haploid structure in EpiSimRA.

Computational complexity of EpiSimRA is directly proportional to N, the number of individuals per generation; g, length of the genome under simulation and k for k-way epistatic interactions. As we increase these parameters, we obtain a more complex evolutionary history leading to longer running time due to complex interactions between the inherited loci from one generation to the other, for randomly sampled extant populations. As we observe concordance in the observed TMRCA for the coalescent simulator EpiSimRA as well as in fwd-EpiSimRA when multiway epistasis is in effect, we obtain validation about the empirical correctness of the coalescent simulator. As in the coalescent simulator we cannot assume correctness until after the ARG has been established, we used the forward model to show the correctness of EpiSimRA, under varying values of selection coefficients and epistatic scenarios. The coalescent simulator, EpiSimRA is extremely fast in finding approximations to TMRCA, in comparison to fwd-EpiSimRA, as the latter has to build the entire "book of populations" and trace it. This leads to a difference in running time of hours for the forward model vs. seconds for its coalescent counterpart with varying input dimensions.

5. Conclusions

We present an algorithm that builds multi-locus selection and multiway epistasis into the backward coalescent model with recombinations, as well as in a forward scheme. Moreover, to the best of our knowledge, this is the first model that has taken a backward simulator with multiway epistasis and compared it nose-to-nose with its forward counterpart. Through extensive empirical comparison studies, albeit for small populations due to the time constraint of the forward model, we show that for complex scenarios with selection and epistasis (or even under neutral scenarios) the hallmark values by the backward and the forward schemes approximate each other. As the distributions of both the schemes are concordant, we conclude that either of the simulators (EpiSimRA or fwd-EpiSimRA) can be used to understand the effects of negative and positive selection, with multiway epistasis, along with selective sweeps across generations. Due to the lack of similar assumptions, parameters and hallmarks of ARGs returned, we did not compare EpiSimRA with present coalescent simulators for selection at a single locus. As fwd-EpiSimRA is based on the Wright-Fisher model and allows for epistatic interactions, we used it as a validation framework.

Multiway interaction across multiple loci leads to complex population genetic history but has a shorter height of the ARG relative to non-epistatic interactions. EpiSimRA encompasses all such scenarios with the potential for further exploration for viral phylodynamics with random sampling of a bacteria or virus populations. The time to coalescence when reconstructing its phylogeny under selection and epistasis allows us to study important epidemiological, immunological and evolutionary processes of viruses [29] such as the recent SARS-CoV-2 or similar Coronaviridae. This allows a validation framework for including selection and epistasis into standard population genetic models where we can now study the different scenarios when all the diploids associated with mutated sites along the chromosome with differing fitness values corresponding to the alleles.

6. Patents

There is no patent resulting from this work.

Author Contributions: Conceptualization, L.P. and D.E.P.; methodology, A.B., F.U., D.E.P., L.P.; software, A.B. and F.U.; validation A.B. and D.E.P.; writing—original draft preparation, A.B., D.E.P., L.P.; writing—review and editing, A.B., F.U., D.E.P. and L.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by International Business Machines Corporation.

Data Availability Statement: EpiSimRA (both backward and forward) source, executable, user manuals are available at: https://github.com/ComputationalGenomics/SimRA (accessed on 24 April 2021).

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ARG	Ancestral Recombination Graph
CDF	Cumulative Distribution Function
EpiSimRA	Epistatic simulations based on Random Graph Algorithms
fwd-EpiSimRA	forward EpiSimRA
KS	Kolmogorov-Smirnov
MRCA	Most Recent Common Ancestor
SNP	Single Nucleotide Polymorphisms
TMRCA	Time to Most Recent Common Ancestor
WF	Wright-Fisher

Appendix A. The Forward Simulator

For two individuals i and i', the ratio of the probabilities that a locus l contributes to whether an individual will have an offspring is the relative fitness

$$\frac{p_{il}}{p_{i'l}} = \frac{1+s_{il}}{1+s_{i'l}}.$$

The total ratio of probability i will have children to i' having children is

$$\frac{p_i}{p_{i'}} = \prod_l \frac{1+s_{il}}{1+s_{i'l}} = \frac{\prod_l (1+s_{il})}{\prod_l (1+s_{i'l})}.$$

From this, it follows that

$$\frac{p_i}{\prod_l (1+s_{il})} = \frac{p_{i'}}{\prod_l (1+s_{i'l})} = \delta,$$

where δ has the same for all i, i' and all other individuals.

Given this,

$$p_i = \delta \prod_l (1+s_{il}).$$

Since $\sum_i p_i = 1 = \delta \sum_i \prod_l (1+s_{il})$, it follows that

$$r = \frac{1}{\sum_i \prod_l (1+s_{il})}.$$

Therefore, the probability that i has children is

$$p_i = \frac{\prod_l (1+s_{il})}{\sum_i \prod_l (1+s_{il})}.$$

For multiway epistasis we include the conditions of k loci being linked with each other such that a combined fitness coefficient is calculated by taking the fitness product at each interacting site as defined in Equation (10).

Appendix A.1. Choosing Parents

The probability that two children will pick the same parent is operationally, the reciprocal of the effective population size [30,31]. Likewise, the same interpretation was made by [5] in the construction of the coalescent. Given a set of p_i's in a given generation t, the probability that two children will pick the same parent is $\frac{1}{N} = \sum_i p_i^2$. While the p_i's define the probability that children pick their parents, N does not play a direct role in determining the course of the algorithm in constructing the book of populations but will affect the shape of the ARG that is traced in the second stage.

With selection on a single locus in effect, each generation will have N_s individuals that contains the allele under selection, yielding $\prod_l (1+s_{il}) = 1+s$, and $(N - N_s)$ individuals without the allele with $\prod_l (1+s_{il}) = 1$.

Transfer of Genetic Material

After the children have randomly selected their parents, the child requests one chromosome from each of the parents. The parents randomly select whether to pass one of their two chromosomes, or to construct a new chromosome via a recombination event involving a crossover between its two chromosomes with respect to the recombination rate, r. If a crossover is generated, the parent randomly selects a location and transfers the genetic material up to that location from one chromosome and the rest from the its other copy. This is done in part to reconstruct the ARG, and to characterize genetic variation along chromosomes yielding the final recombinations [32]. In case of no recombination, the parent randomly decides which chromosome's genetic material should be passed over to the child (see Figure 1 of the main manuscript).

Each newly constructed chromosome is painted with new SNP mutations randomly generated according to a mutation rate probability μ, a randomly selected location, and allele value. With probability of mutation on each polymorphic site, the resultant mutated chromosomes are finally passed to the child from the parents along with the sites of mutations and recombinations.

Throughout the generation, forward in time, we keep track of the sites of recombinations and mutations to efficiently trace the ARG from extant individuals to its GMRCA.

Appendix A.2. Tracing the ARG from the Book of Populations

Detecting the past recombination events from extant sequences and specifying the place of each recombination and recombinant sequences has been well studied [25–27]. The ARGs define a genealogical graph for all of the chromosomes in a population. For each locus, the ARG for any given segment between recombination crossovers will form a tree. When the sequences are non-recombining, we only need to use coalescences and mutations to describe their genealogy to find a most recent common ancestor (MRCA). Traversing back through an ARG, coalescent events are very common in occurrence, but, in case of a recombination, the history of lineages not only show bifurcations, but also recombinations resulting in cycles. Our algorithm looks for recombination events going back every generation and traces them until convergence to a GMRCA.

Appendix A.3. Simulating the Book of Populations with Selection and Two-Way Epistasis

ALGORITHM:

```
1.  Initialization:
      (a) $N$ individuals ($\frac{N}{2}$ males and $\frac{N}{2}$ females) in the base generation, which remains constant throughout the simulation.
      (b) Number of Generations, $G = c * N$, where $c$ is a constant.
      (c) Randomly allocate genetic material along the length of chromosome, $g$.
      (d) Assign selection coefficients for interacting sites for two-way epistasis (0 for neutral).
      (e) Set flag, $f$, for allele(s) under selection on a mutated site (0 for neutral).
2.  If $f$ is set, randomly select an individual among $N$ and a site, $g_s$ along $g$ which underwent mutation. Select an allele randomly in $g_s$ and set $f$ to 1.
3.  Loop For each generation, $t \in \{1, \cdots, G\}$
4.    Loop For each individual $i$ in $\{1, \cdots, N\}$, in $(t-1)$th generation.
5.      Compute $p_i = \frac{\prod_k (1+S_{ik})}{\sum_i \prod_k (1+S_{ik})}$, where any group $k$ of loci could contain a single locus under selection, for which $S = s$ is defined as the user input. It can also contain a locus interacting with another locus, in a two-way epistasis. In this case $s$ is populated from a matrix formed by the all possible alleles at each loci, from the following form, $S = \prod_j \left(1 + \sum_i s_i^{(j)}\right)$. $s_i^{(j)}$ is the selection cofficient at allele $j$ in individual $i$'s chromosome.
6.      Select parents for each child in $t$th generation based on $p_i$ from $(t-1)$th generation.
7.    End
8.    For each child $i$ in $t$th generation, compute scaled recombination rate $r' = r * g$ and select a value, $r_{val} \in [0, 1]$.
9.    If $r_{val} = \begin{cases} [0, (1-r')), & \text{No recombination event} \\ [(1-r'), 1], & \text{recombination event} \end{cases}$
10.   If No recombination event: Randomly pick a chromosome from the parent and assign its genetic material to the child.
11.   Else Randomly pick a crossover index $z \in [1, g]$. Get the genetic material from $[1, z]$ in the first chromosome of the parent and $[(z+1), g]$ in the second, combine them and assign it to the child.
12.   In the child's genetic material, randomly select locations along the chromosome length, $g$ for mutation according to the Poisson distribution and the scaled mutation rate $\mu' = \mu * g$. Assign the alleles randomly to other bases. For example, if the allele was $A$, change it randomly to one of the other bases $\{G, T, C\}$.
13.   Update the Chromosomes of the current generation with the new genetic information obtained from the previous generation and continue until the last generation, $G$.
14. End
```

Appendix A.4. Tracing the ARG from the Book of Populations

```
ALGORITHM:
1.  Initialization:
    (a) Randomly select $m$ number of extant individuals from $N$ in the last generation.
    (b) Select one chromosome out of the two in these $m$ extant samples, randomly.
        Compute the active lineages, $j$ by comparing the genetic material $g$ in each of
        the $m$ chromosomes selected.
2.  Loop for each generation, $t$ going backwards from $\{G, \cdots, 1\}$
3.      Identify each chromosome from the previous generation $(t-1)$ which contributed
    to each chromosome in the current generation, following the book of populations.
4.      Check to see if multiple children in the $g$th generation share the same parent in the
    previous generation.
5.      Iterate and Count the number of active samples, $m'$ in each generation.
6.  Until $m' = 1$
7.  Compute the Height of the GMRCA from the height of convergence.
```

Appendix A.5. Experiments and Comparison Study

Here we exhaustively list all the box-whisker diagrams, Q-Q and CDF plots for two-way epistasis and P-P plots for all experiments conducted while comparing the two simulators fwd-EpiSimRA and EpiSimRA (Figures A3–A5).

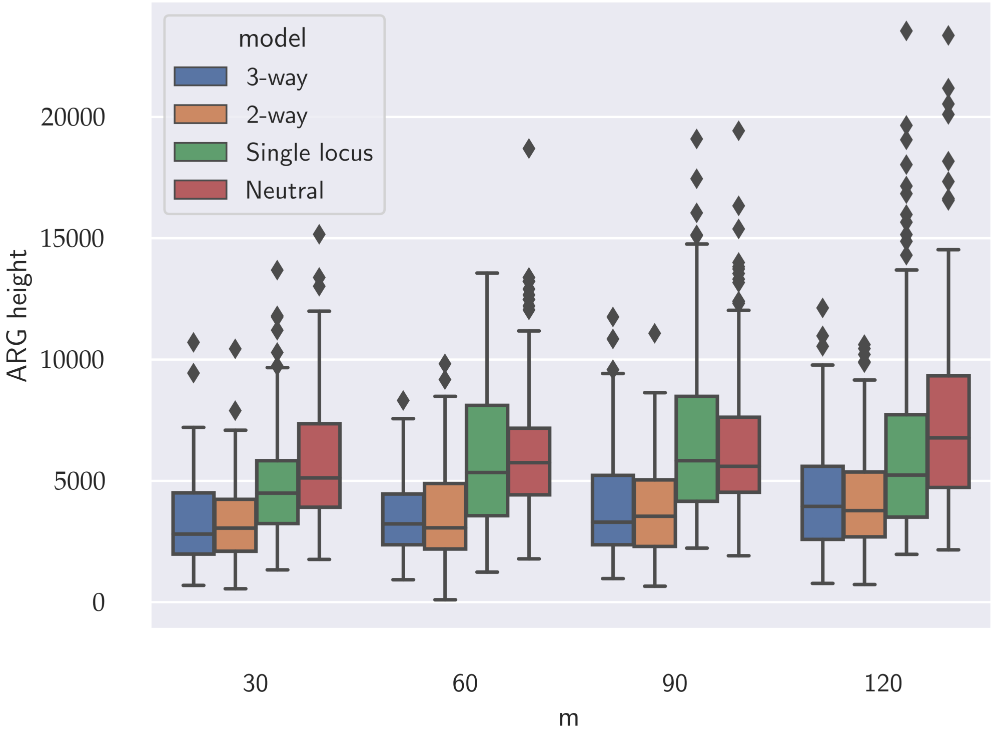

Figure A1. Comparing the height of the ARG (H) for different scenarios of selection in EpiSimRA with epistatis and recombination for $N = 1000$, $g = 250$ kbp, $r = 1.0 \times 10^{-8}$, $m = \{30, 60, 90, 120\}$, $s = \{0.3, 0.3, 0.3\}$ with epistastic parameters for $s_i s_j = 0.15$ for $i, j \in [1, 3]$ and $s_1 s_2 s_3 = 0.125$. The box-and-whisker diagram summarizes the result for each m and selection scenarios such as neutral ($s = 0$), single locus ($s = 0.3$), epistatic interaction at two loci and three loci respectively.

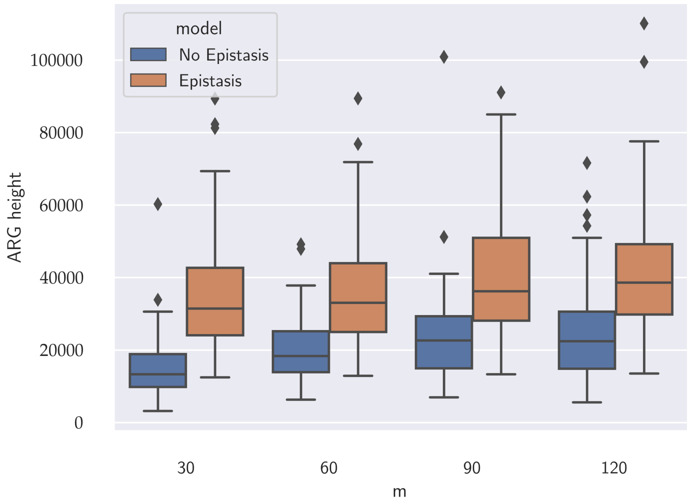

Figure A2. Comparing the height of the ARG (H) for different scenarios of selection in EpiSimRA with and without epistatis in recombination for N = 10,000, $g = 250$ kbp, $r = 1.0 \times 10^{-8}$, $m = \{30, 60, 90, 120\}$, $s = \{0.3, 0.3, 0.3\}$ with epistastic parameters for $s_i s_j = 0.15$ for $i, j \in [1, 3]$ and $s_1 s_2 s_3 = 0.125$. The box-and-whisker diagram summarizes the result for each m and selection scenarios with and without epistatic interaction at three loci.

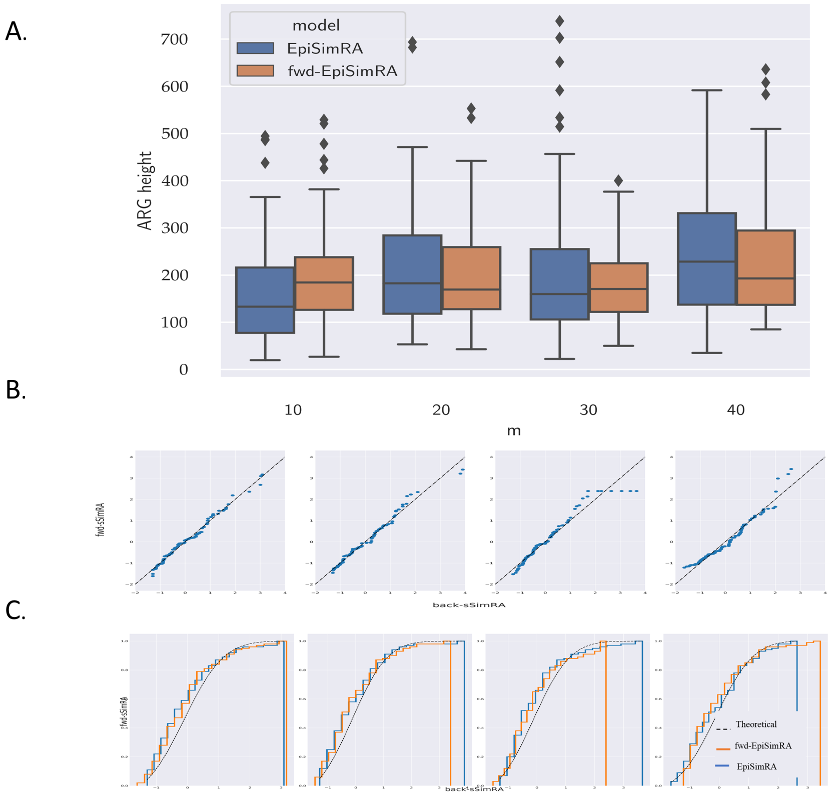

Figure A3. Comparing the height of the ARG (H) between fwd-EpiSimRA and EpiSimRA with and without epistatis in two loci with recombination for $N = 100$, $g = 250$ kbp, $r = 1.0 \times 10^{-8}$, $m = \{10.20, 30, 40\}$, $s = \{0.3, 0.3\}$ with epistastic parameters for $s_0 s_1 = 0.15$. (**A**) The box-and-whisker diagram summarizes the result for each. On each box, the central mark is the mean, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. (**B**) QQ plot and (**C**) CDF plot of the backward and forward models show similar distributions.

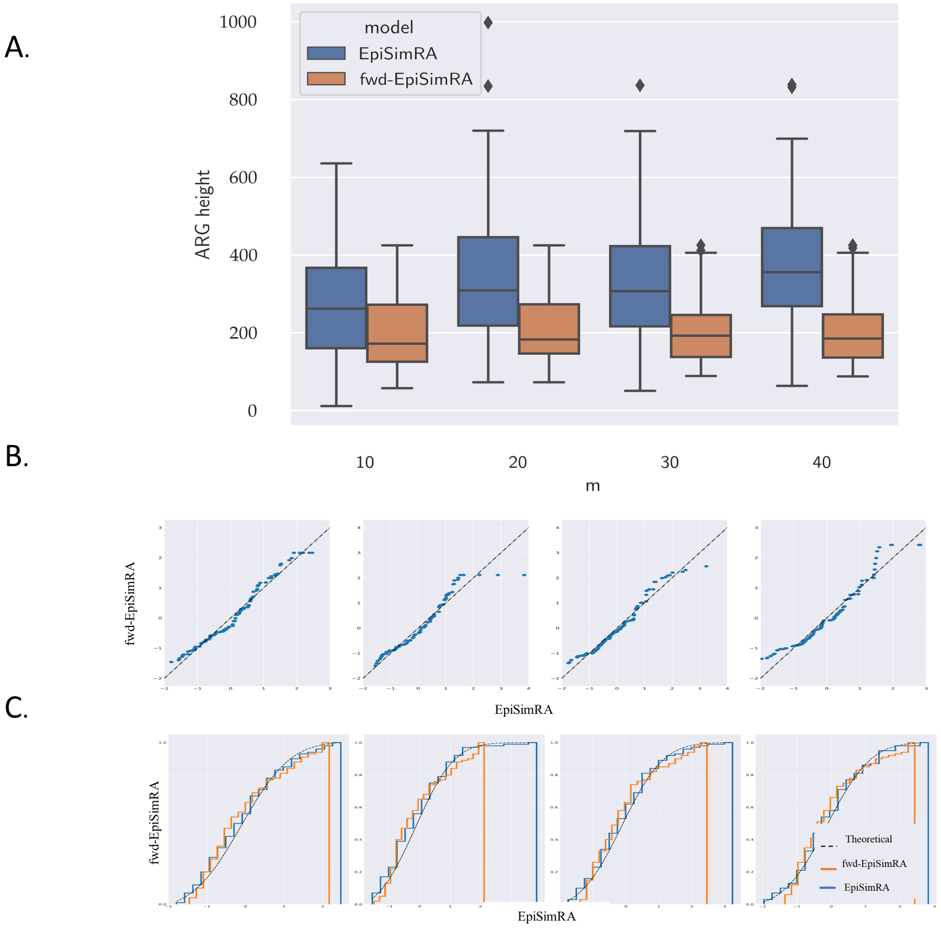

Figure A4. Comparing the height of the ARG (H) between fwd-EpiSimRA and EpiSimRA for selection in single locus with recombination for $N = 100$, $g = 250$ kbp, $r = 1.0 \times 10^{-8}$, $m = \{10.20, 30, 40\}$, $s = 0.3$. (**A**) The box-and-whisker diagram summarizes the result for each. On each box, the central mark is the mean, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. (**B**) QQ plot and (**C**) CDF plot of the backward and forward models show similar distributions.

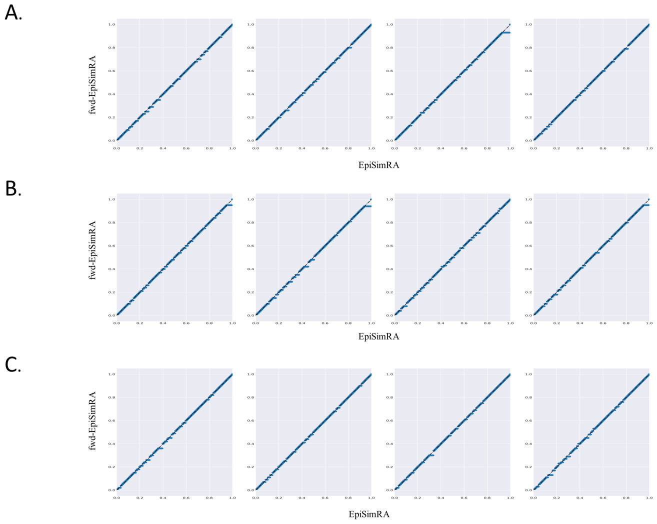

Figure A5. P-P plots of distributions of the height of the ARG (H) between fwd-EpiSimRa and EpiSimRA for (**A**) single locus selection, (**B**) epistatic interaction at two loci and (**C**) epistatic interaction at three loci $g = 250K$, $r = 1.0 \times 10^{-8}$, $N = 100$, $s = 0.3$, $e_s = \{0, 0.1\}$ and $m = \{10, 20, 30, 40\}$.

We also provide the test statistics and p-values obtained by running K-S test which does not reject the null hypothesis that the samples of H as returned by the two simulators are indeed drawn from the same distribution as shown in Table A1.

Table A1. K-S test statistics with corresponding p-values showing that the probability distributions of H as returned by *fwd-sSimRA* and *back-sSimRA* abstracts each other very closely.

3 Interacting loci			e_s	**m**	p-**Value**	**Test Statistic**
s_1	s_2	s_3				
×	×	×	×	10	0.1400	0.16
				20	0.4431	0.12
				30	0.3439	0.13
				40	0.9995	0.05
s_1	×	×	×	10	0.6766	0.08
				20	0.7942	0.08
				30	0.6766	0.10
				40	0.5750	0.11
s_1	s_2	×	×	10	0.9921	0.06
				20	0.5560	0.11
				30	0.7942	0.09
				40	0.8938	0.08
s_1	s_2	×	0.1	10	0.8938	0.08
				20	0.9995	0.05
				30	0.9710	0.06
				40	0.7942	0.09
s_1	s_2	s_3	×	10	0.3439	0.13
				20	0.7942	0.08
				30	0.6766	0.10
				40	0.5576	0.11
s_1	s_2	s_3	0.1	10	0.9610	0.07
				20	0.9610	0.07
				30	0.3556	0.13
				40	0.6766	0.10

References

1. Dobzhansky, T. Nothing in biology makes sense except in the light of evolution. *Am. Biol. Teach.* **2013**, *75*, 87–91. [CrossRef]
2. Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **1969**, *61*, 893. [CrossRef] [PubMed]
3. Hudson, R.R. Estimating the recombination parameter of a finite population model without selection. *Genet. Res.* **1987**, *50*, 245–250. [CrossRef] [PubMed]
4. Calafell, F.; Grigorenko, E.L.; Chikanian, A.A.; Kidd, K.K. Haplotype evolution and linkage disequilibrium: A simulation study. *Hum. Hered.* **2001**, *51*, 85–96. [CrossRef]
5. Kingman, J.F.C. On the Geneaology of Large Populations. *J. Appl. Probab.* **1982**, *19*, 27–43. [CrossRef]
6. Griffiths, R.; Marjoram, P. An ancestral recombination graph. In *Progress in Population Genetics and Human Evolution, IMA Vols in Mathematics and Its Applications*; Donnely, P., Tavare, S., Eds.; Springer: New York, NY, USA, 1997; Volume 87, pp. 257–270.
7. Carvajal-Rodríguez, A. GENOMEPOP: A program to simulate genomes in populations. *BMC Bioinform.* **2008**, *9*, 223. [CrossRef]
8. Kelleher, J.; Etheridge, A.M.; McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.* **2016**, *12*, e1004842. [CrossRef] [PubMed]
9. McVean, G.A.; Cardin, N.J. Approximating the coalescent with recombination. *Philos. Trans. R. Soc. B Biol. Sci.* **2005**, *360*, 1387–1393. [CrossRef]
10. Chen, G.K.; Marjoram, P.; Wall, J.D. Fast and flexible simulation of DNA sequence data. *Genome Res.* **2009**, *19*, 136–142. [CrossRef]
11. Excoffier, L.; Foll, M. fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **2011**, *27*, 1332–1334. [CrossRef]

12. Ewing, G.; Hermisson, J. MSMS: A coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **2010**, *26*, 2064–2065. [CrossRef]
13. Shlyakhter, I.; Sabeti, P.C.; Schaffner, S.F. Cosi2: An efficient simulator of exact and approximate coalescent with selection. *Bioinformatics* **2014**, *30*, 3427–3429. [CrossRef]
14. Spencer, C.C.A.; Coop, G. SelSim: A program to simulate population genetic data with natural selection and recombination. *Bioinformatics* **2004**, *20*, 3673–3675. [CrossRef]
15. Teshima, K.M.; Innan, H. mbs: Modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection. *BMC Bioinform.* **2009**, *10*, 166. [CrossRef] [PubMed]
16. Excoffier, L.; Dupanloup, I.; Huerta-Sánchez, E.; Sousa, V.C.; Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **2013**, *9*, e1003905. [CrossRef] [PubMed]
17. Carrieri, A.P.; Utro, F.; Parida, L. Sampling ARG of multiple populations under complex configurations of subdivision and admixture. *Bioinformatics* **2016**, *32*, 1048–1056. [CrossRef] [PubMed]
18. Neuhauser, C.; Krone, S.M. The Genealogy of Samples in Models with Selection. *Genetics* **1997**, *145*, 519–534. [CrossRef] [PubMed]
19. Stephens, M.; Donnelly, P. Ancestral inference in population genetics models with selection (with discussion). *Aust. N. Z. J. Stat.* **2003**, *45*, 395–430. [CrossRef]
20. Barton, N.H. How does epistasis influence the response to selection? *Heredity* **2016**, *118*, 96–109. [CrossRef] [PubMed]
21. Corbett-Detig, R.; Jones, M. SELAM: Simulation of epistasis and local adaptation during admixture with mate choice. *Bioinformatics* **2016**, *32*, 3035–3037. [CrossRef]
22. Messer, P.W. SLiM: Simulating Evolution with Selection and Linkage. *Genetics* **2013**, *194*, 1037–1039. [CrossRef]
23. Haller, B.C.; Messer, P.W. SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. *Mol. Biol. Evol.* **2019**, *36*, 632–637. [CrossRef]
24. Felsenstein, J. Theoretical Evolutionary Genetics. Available online: https://evolution.gs.washington.edu/pgbook/pgbook.pdf (accessed on 24 April 2021).
25. Javed, A.; Pybus, M.; Melé, M.; Utro, F.; Bertranpetit, J.; Calafell, F.; Parida, L. IRiS: Construction of ARG networks at genomic scales. *Bioinformatics* **2011**, *27*, 2448–2450. [CrossRef]
26. Melé, M.; Javed, A.; Pybus, M.; Calafell, F.; Parida, L.; Bertranpetit, J.; Consortium, T.G. A New Method to Reconstruct Recombination Events at a Genomic Scale. *PLOS Comput. Biol.* **2010**, *6*, 1–13. [CrossRef] [PubMed]
27. Parida, L.; Melé, M.; Calafell, F.; Bertranpetit, J. Estimating the Ancestral Recombinations Graph (ARG) as Compatible Networks of SNP Patterns. *J. Comput. Biol.* **2008**, *15*, 1133–1153. [CrossRef]
28. Kelleher, J.; Thornton, K.R.; Ashander, J.; Ralph, P.L. Efficient pedigree recording for fast population genetics simulation. *PLoS Comput. Biol.* **2018**, *14*, e1006581. [CrossRef] [PubMed]
29. Volz, E.M.; Koelle, K.; Bedford, T. Viral phylodynamics. *PLoS Comput. Biol.* **2013**, *9*, 11. [CrossRef] [PubMed]
30. Crow, J.F. Breeding structure of populations II. Effective population number. In *Statistics and Mathematics in Biology*; Kempthorne, O., Bancroft, T.A., Lush, J.L., Eds.; Iowa State College Press: Ames, IA, USA, 1954; pp. 543–556.
31. Kimura, M.; Crow, J.F. The number of alleles that can be maintained in a finite population. *Genetics* **1964**, *49*, 725–738. [CrossRef]
32. Stephens, M.; Donnelly, P. Inference in molecular population genetics. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2000**, *62*, 605–635. [CrossRef]