DISTINGUISHING COALESCENT MODELS - WHICH STATISTICS MATTER MOST?

FABIAN FREUND AND ARNO SIRI-JÉGOUSSE

Abstract. Modelling genetic diversity needs an underlying genealogy model. To choose a fitting model based on genetic data, one can perform model selection between classes of genealogical trees, e.g. Kingman's coalescent with exponential growth or multiple merger coalescents. Such selection can be based on many different statistics measuring genetic diversity. We use a random forest based Approximate Bayesian Computation to disentangle the effects of different statistics on distinguishing between various classes of genealogy models. For the specific question of inferring whether genealogies feature multiple mergers, we introduce a new statistic, the minimal observable clade size, which corresponds to the minimal allele count of non-private mutations in an individual.

1. Introduction

Modelling genetic diversity of a sample of present-day individuals is a key tool when reconstructing the evolutionary history of populations or scanning the genome for regions under selection, see e.g. [CB17, Section "Nonequilibrium theory", p. 1026]. For a genetic region without recombination, a standard approach is to model the genealogical tree and the mutations along its branches as random processes. The genealogy thus is given by a random tree with n leaves. On this, mutations are modelled as an independent Poisson process along the edges of the coalescent tree.

Several genealogical tree models have been proposed. The most widely used model, Kingman's n-coalescent, is a strictly bifurcating random tree [Kin82]. It approximately describes the genealogy of a fixed-size population under neutral evolution if offspring numbers per parent are not too variable, e.g. if the population is described by a (haploid or diploid) Wright-Fisher model or a Moran model with fixed size N, see [Kin82]. More precisely, the n-coalescent is the limit (in distribution) of the genealogies in the discrete populations as $N \to \infty$, if one treats $\lceil c_N^{-1} t \rceil$ generations as one unit of time in the coalescent tree (which has a continuous time scale). Here, c_N is the coalescence probability, i.e. the probability that two individuals, picked in any generation, have a common parent. Variants of Kingman's n-coalescent as limits of genealogies from Wright-Fisher models with varying population

Date: December 20, 2019.

Key words and phrases. Multiple merger, Approximate Bayesian Computation, Diversity statistics, Clade size, Genealogical model selection.

size and/or population structure are available, see e.g. [GT94], [WH98]. These variants still are strictly bifurcating trees. Barring recombination, both haploid and diploid populations are well-captured by Kingman's n-coalescent and its variants.

However, some populations will feature genealogies not well approximated by Kingman's n-coalescent and its bifucating variants. Certain modes of rapid selection, affecting the whole genome, should lead to the Bolthausen-Sznitman n-coalescent, e.g. see [NH13], [DWF13], [Sch17]. This genealogy model likely produces multifurcating genealogical trees. Even a succession of hard selective sweeps may cause a multifurcating genealogy (though not the same model), see [DS05].

Generally, if one considers the genealogy in an arbitrary Cannings model, a fixed-size discrete-time population model where offspring sizes are exchangeable among parents, a whole class of multifurcating trees, the Λ- or Ξ-n-coalescents, emerges as possible weak limits of the genealogies for $N \to \infty$, again with the analogous rescaling as for Kingman's n-coalescent, see [MS01]. Non-Kingman n-coalescents arise if the distribution of offspring size per parent is variable enough, see [Möh98]. They have been shown to explain the genetic diversity of several maritime species better than Kingman's n-coalescent and its variants, see e.g. [EW06], [NNY16] or [HSJB19]. They seem to capture well the concept of sweepstake reproduction, where one individual may produce a considerable part of the population by chance. Two specific haploid models proposed for this are leading to the Eldon-Wakeley n-coalescents, see [EW06], and to the Beta-n-coalescents, see [Sch03]. However, for diploid populations, the multiple merger mechanism is modified from its haploid version, see [BLS18]. Again, variants for varying population sizes in the pre-limit models are available, see [MHAJ18], [Fre19], [KWB19], [GCMPSJ19]. For further discussion where multiple merger genealogies may arise, see the reviews [TL14] and [ILM$^+$16].

How can one distinguish these genealogical models based on a sample of n genomic sequences? Available methods using the full sequence information do not scale well with data sets with many individuals and/or regions with many mutations [SBB13], [Ste09], [Kos18]. Thus, less costly methods based on a summary statistic of the genomic information, the site-frequency spectrum (SFS), have been proposed, see e.g. [EBBF15], [Kos18]. The SFS records how many mutations appear exactly i times in the sample for each $i \in \{1, \dots, n\}$. These methods are noisy for one or few regions, but have fairly low error rates for data taken across the genome for many unlinked loci, see [Kos18]. However, even whole genomes may not contain many unlinked loci. For instance, some bacterial genomes consist of a single chromosome with neglectable recombination, e.g. in *Mycobacterium tuberculosis*, also several fungi have comparably small genomes with large linkage blocks. For cancer cells, (one-locus) Beta-n-coalescents with exponential growth have been proposed as a genealogy model for copy numbers (specific mutations) in cancer cells [KVS$^+$17] within multiple regions. In this study,

quantiles of the SFS and several additional statistics were used as summary statistics for inference based on Approximate Bayesian Computation (ABC). Low classification errors ($\leq 2\%$) were reported, which were much lower than the SFS-based classification errors from [EBBF15]. The discrepance of error rates may come from slightly different hypotheses, a different approach to mutation rates, a different ABC approach and/or the use of additional summary statistics.

Motivated by this drop in error rates, our goal is to investigate which statistics are best suited to distinguish different classes of n-coalescents. Additionally to the statistics from [KVS+17], we consider some further common statistics as well as a new statistic, the smallest allele frequency among non-private mutations observed in one individual.

These statistics are used in an ABC framework. We use a slight modification of the ABC approach based on random forests from [PME+15] that performs well with many, potentially uninformative statistics and that measures the importance of each statistic to distinguish between the hypotheses.

1.1. **Methods.** The genealogical trees are all given by Λ- or Ξ-n-coalescents, which are random trees with a Markovian structure. We start with n leaves at time 0 and build the tree until its root is reached. If b branches are present at a time t, k of these, chosen at random, will be merged with rate $\lambda_{b,k}$. In the case of Λ-n-coalescents, k lines are merged into one new line, in other words they are joined in a node of the tree and a new branch towards the root is started. These rates are characterized by

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\Lambda(dx)$$

with a finite measure Λ on $[0,1]$, see [Pit99]. For Ξ-n-coalescents, $k = (k_1, \ldots, k_m)$ sets of lineages are each merged into one new line, so m new nodes are added from which m new branches extend (the new node i has thus in-degree 1 and outdegree k_i). The rate $\lambda_{b,(k_1,\ldots,k_m)}$ of such a merger can be expressed in terms of a finite measure Ξ on the simplex, as introduced in [Sch00].

As introduced above, all n-coalescents can be seen as limit trees of genealogies of a sample size n in discrete populations of fixed size $N \to \infty$ across all generations. If population size may change, but stays of order N across generations, the genealogy tree is then given by a n-coalescent characterized by the same measure as in the fixed size case, but with time changed by a deterministic function, see [GT94], [MHAJ18], [Fre19], [KWB19]. As Λ-coalescent model classes, we consider Kingman's n-coalescent, Dirac n-coalescents with $\Lambda = \delta_p$ for $p \in (0,1]$, which are a subclass of Eldon-Wakeley n-coalescents, and Beta-n-coalescents with $\Lambda = Beta(2-\alpha, \alpha)$ for $1 \leq \alpha \leq 2$. For the first two classes, we also consider the coalescent models within exponentially growing populations with growth rate ρ, which is reflected by a time change $t \mapsto (\gamma\rho)^{-1}\left(e^{\gamma\rho t} - 1\right)$ of the coalescent model. The parameter γ is

determined by the underlying discrete population model, see [Fre19]. For Kingman's n-coalescent, we use $\gamma = 1$, which corresponds to seeing Kingman's n-coalescent as the limit genealogy of genealogies in Wright-Fisher models, see [GT94]. For the Dirac n-coalescent, γ reflects the rarity of large families in the discrete population model, see [MHAJ18], we choose $\gamma = 1.5$. We also consider one class of (non-Λ) Ξ-n-coalescents, the Beta-Ξ-n-coalescents from [BCEH16] and [BLS18]. To assess population structure as a confounding factor, we also consider the structured (Kingman's) n-coalescent with exponential growth. We assume 2 subpopulations with equal size, symmetric migration with scaled migration rate m (in units of $4N$) and samples of sizes n_1, n_2 from each subpopulation summing up to n. See Table 1 for the short names for each coalescent model class and its (theoretical) parameter ranges.

Symbol	Model	Model parameter(s)
$\mathbb{K}$	Kingman's n-coalescent	-
$\mathbb{K}$-*exp*	$\mathbb{K}$ with exponential growth	$\rho \in [0, \infty)$
$\mathbb{B}$	Beta-n-coalescent	$\alpha \in [1, 2]$
$\mathbb{D}$	Dirac-n-coalescents	$p \in (0, 1]$
$\mathbb{D}$-*exp*	$\mathbb{D}$ with exponential growth	$p \in (0, 1]$, $\rho \in [0, \infty)$
Ξ-$\mathbb{B}$	Beta-Ξ-n-coalescent	$\alpha \in [1, 2]$
$2p.5$-$\mathbb{K}$-*exp*	$\mathbb{K}$-exp with 2 subpopulations, equal population sizes, $n_1 = n_2 = 50$	$m = 0.5$, $\rho \in [0, \infty)$
$2p.9$-$\mathbb{K}$-*exp*	$\mathbb{K}$-exp with 2 subpopulations, equal population sizes, $n_1 = 90, n_2 = 10$	$m = 0.5$, $\rho \in [0, \infty)$

TABLE 1. Coalescent classes and parameters ranges

1.2. **Simulation of coalescent models.** We use discretized parameter ranges for simulation of each model class, see Table 2. For setting the mutation rates, we use several approaches. To be consistent with [KVS+17], for a scenario comparable to their setup we use a predefined mesh of mutation rates θ, identical for all model classes considered. However, the distribution of total length of coalescent processes differs strongly between (and within) model classes, and thus does the total number of mutations. If statistics are strongly affected by the mutation rate, this may either lead to a bad fit to the observed genetic diversity (since the parameter range used simply produces too many or too few mutations for specific models), or parameter ranges have to be very large to counteract these effects. In our case, most statistics are relatively sensitive to changes in the mutation rate. To make models comparable, we start with setting the mutation rate in each model such that the expected number of mutations observed is identical. In other words, we assume, as e.g. in [EBBF15], an (integer) number s of observed mutations and set θ as the generalized Watterson estimator $\theta_w(s) = \frac{2s}{E(L_n)}$,

where L_n is the total branch length in the specific model we are looking at. Choosing different coalescent parameters will cause different θ values within one model class. Naturally, we do not know the true mutation rate, so we should allow some fluctuation of θ for each model. We do this either by picking the number s of observed mutations uniformly from a set of integers, or by 'blurring' θ by drawing the mutation rate from $[b^{-1}\theta_w(s), b\theta_w(s)]$ for one fixed value of s for an integer b (we choose 2 or 10). We do the latter by a discrete prior symmetric on a log scale. We first draw X from a binomial distribution $B(b, 0.5)$ and set the mutation rate to $\theta_w(s)b^{(2X-b)/b}$. The specific scenarios for all three choices for mutation rates are shown in Table 3. All scenarios 1-5, consisting of specific choices of *coalescent model* $\times$ *sample size* n $\times$ *mutation model* are simulated twice with 175,000 times with an uniform prior on the model class parameter. For the mutation rate, we either draw the mutation parameter θ, s uniformly from a discrete set, or we draw with binomial probabilities from a log-equidistant discrete set around $\theta_w(s)$ (as described above). To assess whether changes in mutation prior and range or parameter range affect which statistics distinguish well, we also consider two models who are slightly modificated versions of Scenario 2, named Scenarios 2a and 2b.

For simulation of $\mathbb{K}$, $\mathbb{K}$-*exp*, $2p.5$-$\mathbb{K}$-*exp* and $2p.9$-$\mathbb{K}$-*exp*, we perform simulations with Hudson's `ms`, implemented in the `R` package `phyclust` [Che11]. $\mathbb{D}$, $\mathbb{B}$ and Ξ-$\mathbb{B}$ were simulated with `R`, see Section 4 for code availability. $\mathbb{D}$-*exp* was simulated as described in [MHAJ18], using the implementation available at
`https://github.com/Matu2083/MultipleMergers-PopulationGrowth`.

To compute the (generalized) Watterson estimator $\theta_w(s) = \frac{2s}{E(L_n)}$, where s is the number of observed segregating sites, we employ different methods to compute (resp. approximate) the expected total length $E(L_n)$ of the genealogy. For $\mathbb{K}$, we simply compute $\theta_w(s) = \frac{s}{\sum_{i\in[n-1]} i^{-1}}$. For $\mathbb{K}$-*exp*, we use a recursive computation approach for $E(L_n)$ as described in [EBBF15], available from `http://page.math.tu-berlin.de/~eldon/programs.html`. For $\mathbb{D}$-*exp*, we again use a recursive approach for computing $E(L_n)$ as described in [MHAJ18] (url as above). For $\mathbb{B}$, $\mathbb{D}$ and Ξ-$\mathbb{B}$, we implemented the recursion for $E(L_n)$ from [Möh06, Eq. 2.3] in `R` (see Section 4 for availability). For $2p.5$-$\mathbb{K}$-*exp* and $2p.9$-$\mathbb{K}$-*exp*, we approximate $E(L_n)$ by averaging the total length of 10,000 simulated n-coalescents for each parameter set (again via `ms`).

1.3. **Summary statistics.** Each simulation produces a matrix with elements in $\{0, 1\}$ with n rows and a variable number s of columns. The rows and columns correspond to the number of simulated sequences and to the number of mutations therein. We call it a (simulated) SNP matrix. We use a variety of standard summary statistics from population genetics, mostly as in [KVS+17], but also introduce a new statistic. All statistics depend on the coalescent model, θ and n. For sake of readability, we omit this dependence

Scenario	Models	sample size	coalescent parameters
1	$\mathbb{K}/\mathbb{K}$-exp	$n = 25$	-/$\rho \in \{10, 20, \ldots, 100\}$
	$\mathbb{B}$		$\alpha \in \{1, 1.1, 1.2, \ldots, 1.9\}$
2	$\mathbb{K}$-exp	$n \in \{25, 100, 200\}$	$\rho \in \{0, 0.5, 1, 2.5, 4, 7, 10, 25,$ $50, 75, 100, 500, 1000\}$
	$\mathbb{B}$		$\alpha \in \{1, 1.1, 1.2, \ldots, 1.9\}$
	$\mathbb{D}$		$p \in \{0.05, 0.15, \ldots, 0.95\}$
	Ξ-$\mathbb{B}$		$\alpha \in \{1, 1.2, \ldots, 1.8, 1.9\}$
2a	$\mathbb{K}$-exp	$n = 100$	$\rho \in \{0, 0.5, 1, 2.5, 4, 7, 10, 25,$ $50, 75, 100, 500, 1000\}$
	$\mathbb{B}$		$\alpha \in \{1, 1.1, 1.2, \ldots, 1.9\}$
2b	$\mathbb{K}$-exp	$n = 100$	$\rho \in \{0, 0.5, 1.5, 3, 5, 8, 11, 15, 25\}$
	$\mathbb{B}$		$\alpha \in \{1, 1.1, 1.2, \ldots, 1.9\}$
3	$\mathbb{K}$-exp	$n = 147$	$\rho \in \{0, 0.5, 1, 2.5, 4, 7, 10, 25,$ $50, 75, 100, 500, 1000\}$
	$\mathbb{B}$		$\alpha \in \{1, 1.2, \ldots, 1.8, 1.9\}$
	$\mathbb{D}$		$p \in \{0.05, 0.15, \ldots, 0.95\}$
4	$\mathbb{K}$-exp	$n = 100$	$\rho \in \{0, 0.5, 1, 2.5, 4, 7, 10, 25,$ $50, 75, 100, 500, 1000\}$
	$\mathbb{B}$		$\alpha \in \{1, 1.2, \ldots, 1.8, 1.9\}$
	$2p.5$-$\mathbb{K}$-exp		ρ as for $\mathbb{K}$-exp
	$2p.9$-$\mathbb{K}$-exp		ρ as for $\mathbb{K}$-exp
5	$\mathbb{K}/\mathbb{K}$-exp	$n = 100$	$\rho \in -/\{0.5, 1, 2.5, 4, 7, 10, 25,$ $50, 75, 100, 500, 1000\}$

TABLE 2. Simulation settings - model classes

Scenario	mutation model
1	$\theta \in \{10, 20, \ldots, 100\}$
2	$s \in \{15, 20, 30, 40, 60, 75\}$
2a	$s = 60$
2b	$\theta \in [2^{-1}\hat{\theta}_w(100), 2\hat{\theta}_w(100)]$
3	$\theta \in [10^{-1}\hat{\theta}_w(454), 10\hat{\theta}_w(454)]$
4	$s = 75$
5	$s \in \{100, 150, 200\}$

TABLE 3. Simulation settings - mutation models

in the notations.
The standard statistics are all based on the following four sets of statistics
Let $(c_{i,j})_{i\in[n],j\in[s]}$ be a SNP matrix.

Ind.	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	O_i
1	1	0	0	0	0	0	3
2	1	0	0	0	0	1	2
3	1	0	0	0	0	1	2
4	0	0	1	1	0	0	2
5	0	0	1	1	1	0	2
6	0	1	0	0	0	0	6

TABLE 4. Example computation for the minimal observable clade sizes $(O_i)_{i\in[n]}$. The columns between the vertical lines are a SNP matrix, where 0 denotes the ancestral type and 1 denotes the derived type.

- The site frequency spectrum $(S_i)_{i\in[n-1]}$ (SFS). The r.v. S_i counts how many columns sum up to i, corresponding to how many mutations have a mutant allele count of i. These are $n-1$ statistics. Alternatively, we can also summarize this information by recording the (quantiles of the mutant) allele frequencies $(\sum_{i=1}^n c_{i,j})_{j\in[s]}$.
- The Hamming distances for each pair of rows of the SNP matrix, i.e. the number of pairwise differences between two sequences. These are $\binom{n}{2}$ statistics.
- All tree lengths of the phylogenetic tree reconstructed by the neighbor-joining method based on the set of Hamming distances. These are $2n-2$ statistics
- The squared pairwise correlation between sums for each pair of columns, i.e. the pairwise linkage disequilibrium (LD) measure r^2. These are $\binom{s}{2}$ statistics.

We also introduce a set of n new statistics, the minimal observable clade size $(O_i)_{i\in[n]}$. Again, let $(c_{i,j})_{i\in[n],j\in[s]}$ be a SNP matrix. Then,

$$O_i = \min\left\{\sum_{k\in[n]} c_{k,j} : j\in[s] \text{ with } c_{i,j}=1, \sum_{k\in[n]} c_{k,j}\geq 2\right\} \tag{1}$$

with the convention that $O_i = n$ if the condition on j is not fulfilled by any j. Interpreted in genetic terms, this is the minimal allele count of the mutations of i that are shared by at least one other sequence, see Table 4 for an example. We call O the minimal observable clade size, since it corresponds to the size of the minimal clade including i and at least one other sequence that can be distinguished from the data. See Section A.1 for more details and [FSJ19] for further mathematical properties of O.

To reduce the number of statistics, we follow [KVS+17] and use both low dimensional summaries of the statistics or quantiles of each of the five sets of statistics. Additionally to quantiles of the different sets of statistics presented above, we use the following summaries of the

Short names	Setup
Ham	Quantiles .1,.3,.5,.7,.9 of the Hamming distances
Phy	Quantiles .1,.3,.5,.7,.9 of the phylogenetic branch lengths
$\mathbf{r^2}$	Quantiles .1,.3,.5,.7,.9 of LD measure r^2
O	Quantiles .1,.3,.5,.7,.9, sample mean, sample variance, harmonic mean of $(O_i)_{i\in[n]}$
nO	Quantiles .1,.3,.5,.7,.9 of $(O_i/\max_i\{O_i\})_{i\in[n]}$
nHam	Quantiles .1,.3,.5,.7,.9 of (Hamming distances)/π
AF	Quantiles .1,.3,.5,.7,.9 of the SFS
SFS, **fSFS**,	SFS, folded SFS
nSFS	scaled SFS
SFSl, **fSFSl**	lumped **SFS**, **fSFS**
nSFSl	lumped **nSFSl**
$\ldots+$	$\ldots$, S, π
$\ldots*$	$\ldots$, S, π, Tajima's D, Fay and Wu's H
Si.t	$(S_1, S_{15}+)$

TABLE 5. Short names for sets of statistics used for model comparisons. For the sets beneath the line, adding + ($*$) denotes adding the two (four) additional statistics.

- SFS: Tajima's D [Taj89], Fay and Wu's H [FW00], the lumped tail $S_{15+} = \sum_{i=15}^{n-1} S_i$, e.g. [Kos18], [EBBF15], and the total number $S := \sum_{i=1}^{n-1} S_i$ of segregating sites (the number of columns of the SNP matrix). The lumped SFS is $(S_1, S_2, \ldots, S_{15+})$
- Hamming distances: sample mean π, i.e. the nucleotide diversity (which can also be computed based on the SFS, see e.g. [Dur08, Example 1.4])
- The sample mean, sample variance and harmonic mean of the observable clade sizes $(O(1), \ldots, O(n))$.

We will also use the folded version of the SFS. The folded site frequency spectrum is defined as $fS_i := \frac{S_i+S_{n-i}}{1+\delta_{i,n-i}}$ for $1 \leq i \leq \lfloor \frac{n}{2} \rfloor$, where $\delta_{i,n-i}$ is the Kronecker symbol which is 0 unless $i = n-i$ where it equals 1. The lumped fSFS, analogously to the lumped SFS, is $\left(fS_1, fS_2, \ldots, \sum_{i\geq 15} fS_i\right)$.
Since our statistics are rather sensitive to changes in the mutation rate, we consider the following statistics that are transformed to be more robust to such changes

- The scaled SFS $\left(\frac{S_1}{S}, \frac{S_2}{S}, \ldots, \frac{S_{n-1}}{S}\right)$ and its lumped version $\left(\frac{S_1}{S}, \frac{S_2}{S}, \ldots, \frac{S_{15+}}{S}\right)$
- Quantiles $(.1, .3, \ldots, .9)$ of: the Hamming distances divided by the nucleotide diversity π

- Quantiles $(.1, .3, \ldots, .9)$ of $\left(\frac{O_i}{\max_i\{O_i\}}\right)_{i\in[n]}$

For the reduced sets of statistics mostly used in this study, see Table 5 for short names and setup. For Tajima's D and Fay and Wu's H, we used the function `sfsR` from `https://github.com/sdwfrost/popseq`. All statistics have been computed with `R`, see Section 4 for availability.

1.4. **Model selection via ABC.** We perform model selection with the ABC approach introduced by [PME+15], based on random forests. For this approach, we simulate each model class used with parameters drawn from the prior distribution. Then, bootstrap samples are drawn from the simulations, i.e. we draw with replacement n_{trees} times a bootstrap sample of size n_{boot} from n_{sim} simulations. Then, for each bootstrap sample a decision tree (randomized CART tree) is constructed, whose decision nodes are of the form $\{T > t\}$ or $\{T < t\}$ for any of the summary statistics used and for $t \in \mathbb{R}$. For each node, a random subset of all statistics is considered and the statistic T is picked such that it minimizes misclassification for the bootstrap sample as measured by the Gini index. Nodes are added until the bootstrap sample can be perfectly allocated to the model classes. This means sorting the simulations via the decision rules to distinct sets of simulations, so that each set only includes simulations from the same model class A. This leaf of the decision tree is then classifying other simulations, resp. real observed data, which are guided to it by the decision tree (also) to model class A. The (mis)classification probability for each model class A_1 is estimated by the out-of-bag (OOB) error. For each simulation from a model A_1, the (mis)classification probability can be assessed by recording the proportion of the trees built without this simulation that sort it into model class $A_1, A_2, \ldots$. The probability of a misclassification (classification error) is then the sum of these proportions who sort the simulation into a different class than the one it comes from. The out-of-bag (classification) error for model class A_1 is given by the mean of the misclassification probabilities over all simulations of A_1. Finally, the mean out-of-bag error is the mean classification error across all model classes.
Variable importance of a statistic T is measured as the decrease in misclassification by all nodes using T as decision statistic, averaged over the whole random forest (Gini VI measure). However, this measure is known to be biased if statistics with different scales of measurement are used [SBZH07]. Since we use a range of statistics with different ranges, we account for this bias by correcting for uninformative splits using the approach from [SZ08]. This ABC approach should cope well with high numbers of potentially uninformative summary statistics (curse of dimensionality) and should perform at least comparable to standard ABC methods. See Section A.6 for a comparison with other ABC methods for Scenario 2. Since we are interested in the variable importances for the presented summary statistics, we omitted the option to include the linear discriminant(s) of the statistics in the ABC

Black bar	Range of mean misclassification probabilities (resp. variable importances) over 10 ABC runs in the first set of simulations
Gray bar	Range of mean misclassification probabilities (resp. variable importances) over 10 ABC runs in the second set of simulations
Asterisk (on a range bar)	shows the mean of all objects that contribute to the range bar

TABLE 6. Common legend for figures showing classification errors and/or variable importances

analysis, i.e. the linear combination of variables that maximizes difference between model classes for the training data.
For this ABC approach, we used the `R` package `abcrf`. We modified the default use of variable importance as described above in the underlying command `ranger` from the `ranger` package [WZ17].
For any scenario, for each of 2 replications, we perform 10× an ABC model selection with a random forest of $n_{trees} = 500$ trees, each constructed from a bootstrap sample with (default) size $n_{boot} = 100,000$. Thus, the mean misclassification probabilities across the 10 runs can also be interpreted as a single run with a random forest of 5,000 trees. We can thus aggrevate both variable importance and mean classification errors by further averaging over the 10 runs. We will report the range across the 10 runs as well as the overall mean. Table 6 provides a common plot legend for all ABC result plots. For code availability, see Section 4.

2. Results

We assessed the effects of different sets of diversity statistics on the classification errors for distinguishing different coalescent models. We report changes in classification errors, e.g. if we report a decrease of 2% in error, this is an absolute decrease, e.g. from 12% to 10% misidentification chance. If a statistic's variable importance is not reported, it showed no variation among simulations and was omitted from the ABC analysis. Results are available as `R` objects, see Section 4.

2.1. Beta vs. Kingman with exponential growth.

In the first scenario, we partially reproduce the approach in [KVS+17], ignoring sequencing and model identification errors and only including a subset of models to select from, see Tables 2 and 3. We consider the statistics as in [KVS+17], but adding π and **O**. Results are shown in Figure 1.
In Scenario 1, we clearly see that the strategy of [KVS+17] to add **Phy**, **Ham** and $\mathbf{r^2}$ to the (summaries S, π and quantiles of) the allele frequencies is beneficial (mean out-of-bag error reduced by $> 6\%$ in both replications). However, further adding the observable clade sizes only leads to marginally lower

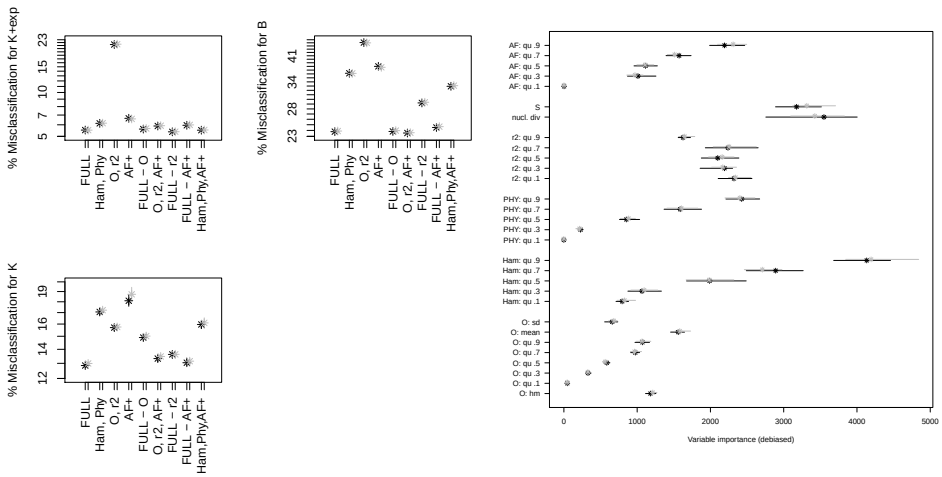
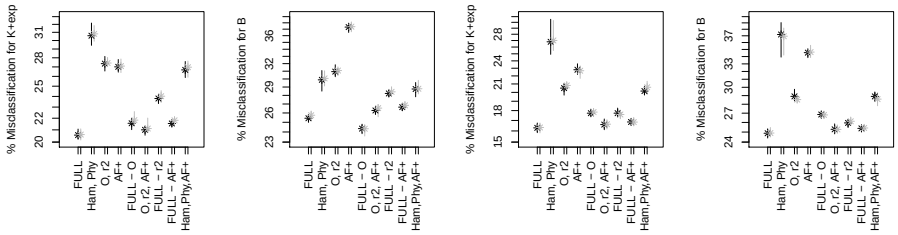
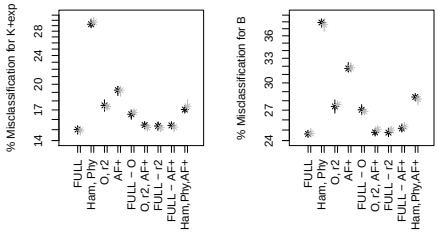


FIGURE 1. Classification errors and variable importances of model comparison between $\mathbb{K}$-*exp*, $\mathbb{K}$ and $\mathbb{B}$ (Scenario 1). The full set of statistics consists of **AF**, $\mathbf{r^2}$, **Phy**, **Ham**, **O**, S and π. For the legend, see Table 6.



(A) n=25 (B) n=100



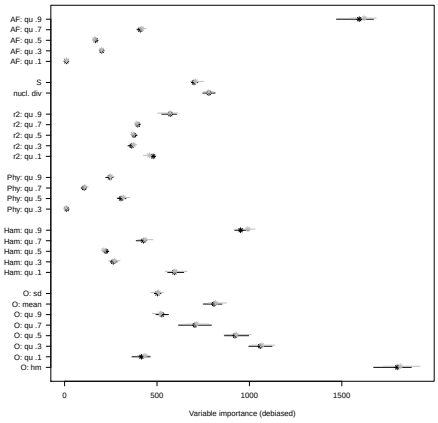
(C) n=200

FIGURE 2. Classification errors of model comparison between $\mathbb{K}$-*exp* and $\mathbb{B}$ (Scenario 2) with $n = 100$. See Table 6 for the legend. The full set of statistics consists of $\mathbf{r^2}$, **Phy**, **Ham**, **O**, S, π and **AF**

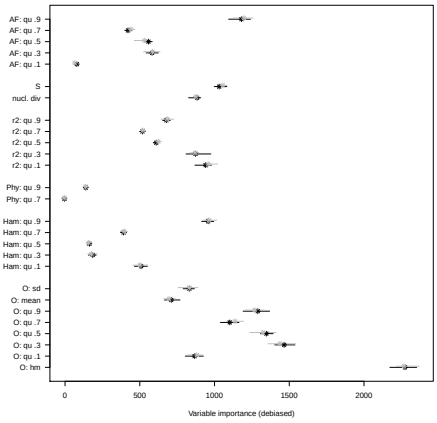
classification errors (further mean out-of-bag error reduction by $< 1\%$), see Figure 1. The best addition to the site frequency spectrum-based statistics in this scenario is $\mathbf{r^2}$. This is reflected by a mean OOB increase of $\approx 2\%$ when only omitting $\mathbf{r^2}$, which is far stronger than the effect when omitting $\mathbf{Phy}, \mathbf{Ham}$, $\mathbf{O}$ or $\mathbf{AF+}$. Compared to later scenarios, misclassification probabilities are rather low (especially given the small size of the sample). However, classification errors are strongly increased compared to the original study of [KVS$^+$17], which we attribute to us classifying only part of their range of parameters in each model (see discussion in Section 3). Scenario 1 uses common mutation rates for each model, leading to quite different expectations of how many segregating sites appear for coalescents with different coalescent parameters (and models). This is due to the differences in total genealogical tree length between models. Thus, the ABC approach should also use this information to select models, i.e. pick statistics for decision nodes that rely on the number of mutations observed. The three most important statistics in Scenario 1 are the 0.9 quantiles of the Hamming distances as well as the nucleotide diversity π and the number of segregating sites S, which are all strongly influenced by the number of mutations observed.

In Scenario 2, exponential growth and Beta coalescents are harder to distinguish. Again, adding statistics to the quantiles of the allele frequency spectrum helps distinguishing the model classes, see Figure 2 and leads to moderate to strong error decrease (mean OOB error reduced by $\geq 8\%$). For this scenario, the best additions to the allele frequency quantiles are $\mathbf{r^2}$ and $\mathbf{O}$. While for $n \in \{100, 200\}$, the positive effect of $\mathbf{O}$ is clearly visible, e.g. if we leave out one set of statistics $\mathbf{O}$, $\mathbf{Ham}$, $\mathbf{Phy}$ or $\mathbf{AF+}$ and perform the ABC with all other statistics, mean OOB is highest when leaving out $\mathbf{O}$. For $n = 25$, $\mathbf{r^2}$ masks the effect of $\mathbf{O}$, if $\mathbf{r^2}$ is left out, adding $\mathbf{O}$ to $\mathbf{Ham}, \mathbf{Phy}, \mathbf{AF+}$ is again beneficial (mean OOB reduction of $\approx 1.7\%$).

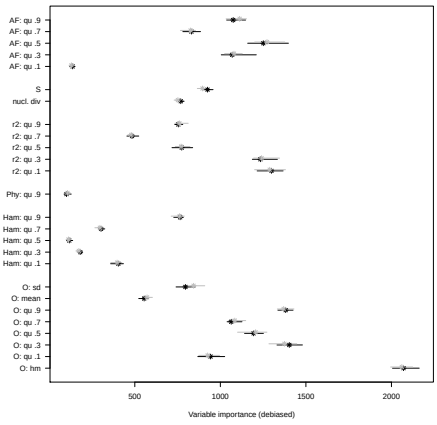
For all n, pooling $\mathbf{AF+}$ with $\mathbf{Phy}$ and $\mathbf{Ham}$ is inferior to using $\mathbf{O}, \mathbf{r^2}$ together with $\mathbf{AF+}$ (at least a decrease of mean OOB of 3.5 % when using the latter). Moreover, the combination $\mathbf{O}, \mathbf{r^2}, \mathbf{AF+}$ leads to only slightly increased errors compared to using all statistics (mean OOB difference of $\leq 1\%$, decreasing with increasing sample size). The variable importance scores in Figure 3 show that especially the harmonic mean of $\mathbf{O}$ is a statistic that distinguishes the model classes well. Moreover, for $n \in \{100, 200\}$ the variable importances nicely mirror the eligibility of $\mathbf{O}, \mathbf{r^2}, \mathbf{AF+}$ as summary statistics for distinguishing $\mathbb{K}$-*exp* and $\mathbb{B}$. Apart from the .9 quantile of the Hamming distances for $n = 100$, all higher importance scores belong to $\mathbf{O}, \mathbf{r^2}, \mathbf{AF+}$. Adding the linear discriminant discriminating the two models at hand did not change the errors ($\leq 0.15\%$ change in mean OOB error for $n = 100$). While including $\mathbf{r^2}$ for model selection is always decreasing errors considerably, whether and to which degree $\mathbf{O}$ does this is also sensitive to the specific choice of hypotheses and prior distributions, e.g. for Scenarios 2a and 2b its effect is clearly reduced, see Figure A8. However,



(A) n=25



(B) n=100

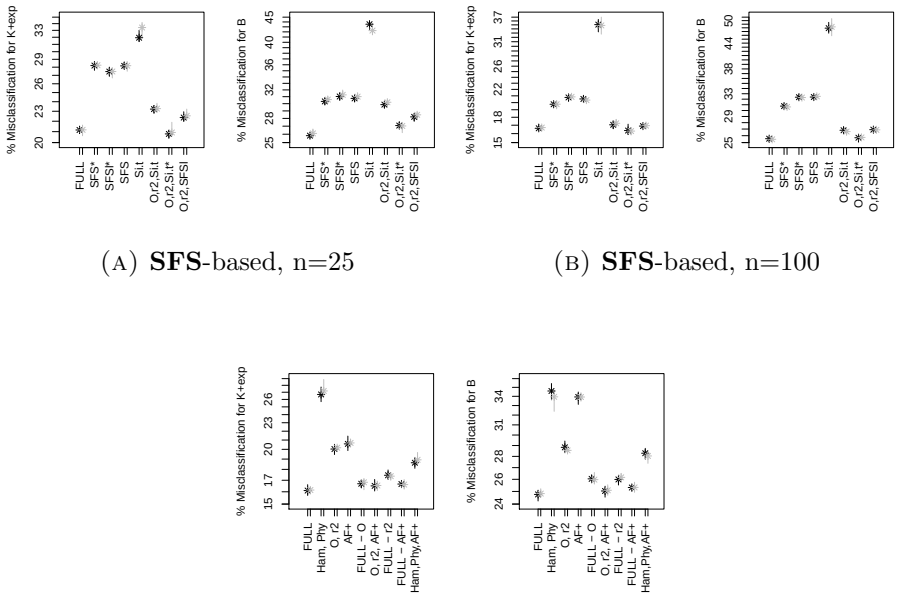


(C) n=200

FIGURE 3. Variable importances for the model comparison between $\mathbb{K}$-*exp* and $\mathbb{B}$ (Scenario 2) using **O**, **Ham**, **Phy**, $\mathbf{r^2}$ and **AF**+. See Table 6 for the legend.

statistics from **O** still have comparatively high varable importances, see Figure A9, and $\mathbf{O}, \mathbf{r^2}, \mathbf{AF}+$ is still clearly reducing errors when compared to $\mathbf{Ham}, \mathbf{Phy}, \mathbf{AF}+$.

Above, we followed [KVS$^+$17] and used relatively few quantiles of each statistic. Does adding further quantiles decrease errors meaningfully? Especially for **AF**, an alternative is to directly use the full site frequency spectrum **SFS** and/or even further summaries thereof. We assessed the effect by both increasing the number of quantiles for all statistics used or to use the full site frequency spectrum as well as two further summaries, Tajima's D and

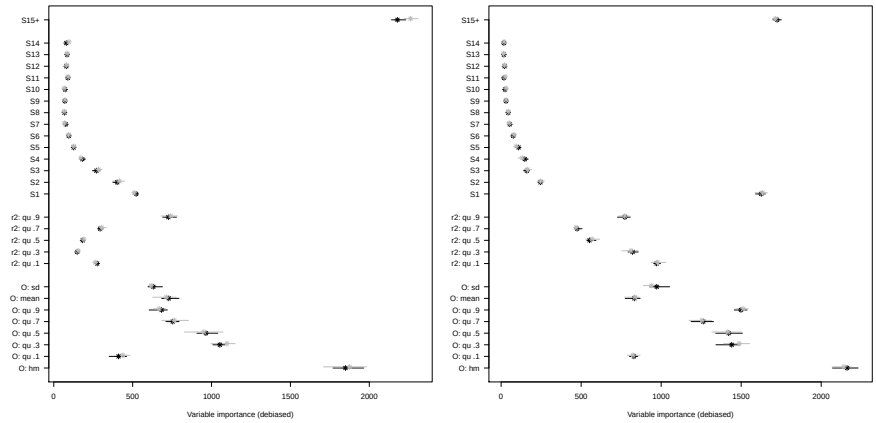


(A) **SFS**-based, n=25 (B) **SFS**-based, n=100

(C) Quantiles $\{.1, .2, .3, \ldots, .9\}$

FIGURE 4. Classification errors of model comparison between $\mathbb{K}$-*exp* and $\mathbb{B}$ (Scenario 2) with $n = 100$. See Table 6 for the legend. The full set of statistics consists of (A), (B) $\mathbf{r^2}$, **Phy**, **Ham**, **O**, S, π, D, H and **SFS** and (C) $\mathbf{r^2}$, **Phy**, **Ham**, **O**, S, π and **AF**, but represented by more quantiles. We denote $Si.t = (S_1, S_{15+})$.

Fay and Wu's H, instead of only quantiles of the allele frequencies. If the model selection is only based on allele frequencies, errors are decreased considerably when using more quantiles of **AF**, even more so when using the full **SFS** and related statistics instead. However, this additional information does not decrease the errors when all statistics are used, it even leads to a slight increase of the (mean) OOB error, see Figure 4. In more detail, if only allele frequency statistics are used, we find an approximate 3% reduction of mean OOB error when comparing inference using the SFS, π, S, Tajima's D and Fay and Wu's H compared to inference using **AF**+. We also find that not all entries of the SFS are highly informative for model selection (for n large enough), which was also reported in previous studies. However, when combining with **O** and $\mathbf{r^2}$, adding SFS summaries π, S, S_1, S_{15+}, D and H produces virtually identical errors as using $\mathbf{O}, \mathbf{r^2}, \mathbf{Ham}, \mathbf{Phy}, \mathbf{SFS}*$. Errors are also very close when inferring models using $\mathbf{O}, \mathbf{r^2}, \mathbf{S_1}, \mathbf{S_{15+}}$ opposed to combining $\mathbf{O}, \mathbf{r^2}$ with the complete 15+-lumped SFS **SFSl**. This is also reflected in the variable importance scores, where, when using the lumped SFS with $\mathbf{O}, \mathbf{r^2}$ for model selection, the statistics S_1 and S_{15+} from the

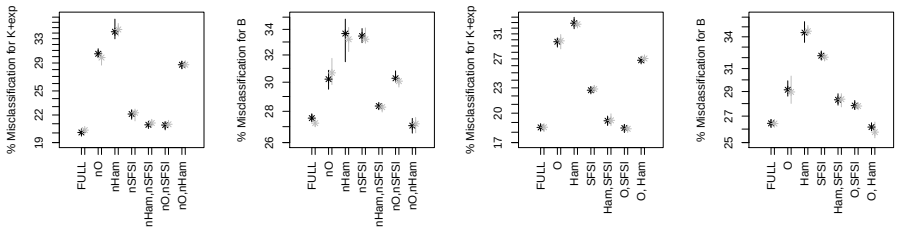


(A) n=25 (B) n=100

FIGURE 5. Variable importances for the model comparison between $\mathbb{K}$-*exp* and $\mathbb{B}$ (Scenario 2) using **O**, **Ham**, **Phy** and and **SFSl**+. See Table 6 for the legend.

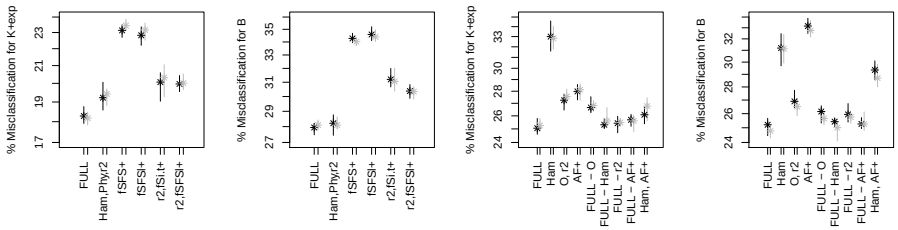
lumped SFS have highest importance apart from the harmonic mean from **O** (Figure 5). Combining S_1 and S_{15+} is the idea behind the singleton-tail statistic of [Kos18]. Not decreasing model selection errors extends to adding more quantiles to other statistics, as well. The error decrease when using more quantiles of $\mathbf{O}, \mathbf{r^2}, \mathbf{AF}+$ or $\mathbf{O}, \mathbf{r^2}, \mathbf{Ham}, \mathbf{Phy}, \mathbf{AF}+$ can be neglected (mean OOB error reduction $\leq 0.2\%$).

The mutation rate strongly affects many of the summary statistics used and thus likely also the model selection, which we also see from the comparison between scenarios 1 and 2. For SFS-based model selection, previous publications scaled the SFS by dividing it by S which weakens the effect of mutation rate differences, see e.g. [EBBF15] or the singleton-tail statistics from [Kos18]. Thus, can we also decrease model selection errors by scaling the statistics in our setting? We used scaled versions of the SFS, the Hamming distances and **O** to assess whether using robust inference statistics helps decreasing errors. We compared model inference in Scenario 2 with both the scaled and unscaled statistics, see Figure 6. Being not influenced strongly by mutation rate changes works if mutation rates are high enough, see Figures A2-A5. However, the information lost by scaling increases error rates slightly, see Figure 6 (mean OOB error increase of $\approx 1.3\%$ for inferring via **nO**, **nHam**, **nSFS** instead of their unscaled versions). While inferring model classes using either **nHam** or **nSFSl** does yield error rates that are essentially unchanged by scaling (mean OOB error increase $\leq 0.35\%$), there is still some information lost when their combination is used (mean OOB error increase of $\approx 0.9\%$). The scaling of **O** by the maximum of the minimal observable clade sizes is relatively crude, which is reflected by an elevated



(A) Scaled statistics (B) Unscaled statistics

FIGURE 6. Classification errors of model comparison between $\mathbb{K}$-*exp* and $\mathbb{B}$ (Scenario 2) with $n = 100$. See Table 6 for the legend. Full set of statistics used: (A) $\mathbf{nO}, \mathbf{nHam}, \mathbf{nSFSl}$ (B) $\mathbf{O}, \mathbf{Ham}, \mathbf{SFSl}$.

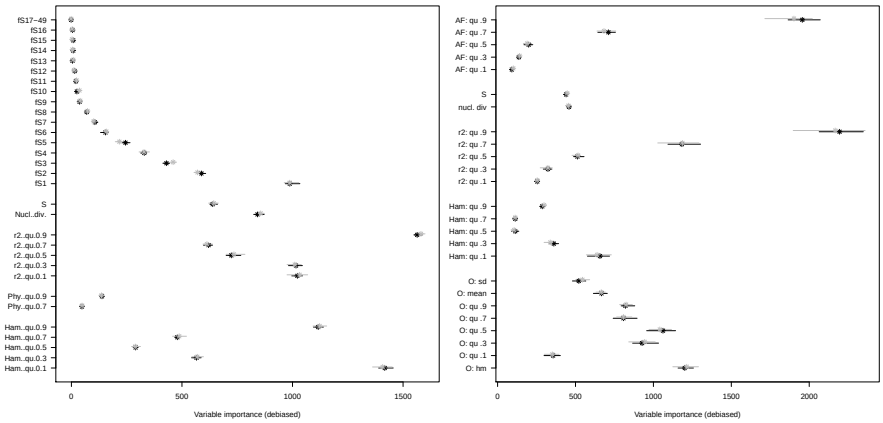


(A) Folded (B) No singletons

FIGURE 7. Classification errors of model comparison between $\mathbb{K}$-*exp* and $\mathbb{B}$ (Scenario 2) with $n = 100$. See Table 6 for the legend.

error increase for inferring using **O** compared to the other sets of statistics (mean OOB error increase of $\approx 0.9\%$ when using **O** and $\geq 1.3\%$ when added to other sets).

So far, we assumed that one can polarize the data, i.e. that the ancestral and derived allele can be distinguished with certainty. However, this is usually done by comparison with one or more outgroups and is not error-free, see e.g. [KJ18]. Thus, this depends on the availability of suitable outgroups and may not always be possible or precise enough. If mutant and ancestral alleles cannot be distinguished, **O** cannot be computed and, for allele/site frequencies, one can only compute the folded site frequency spectrum **fSFS** (resp. the minor allele frequencies). The other statistics S, π, **Phy**, **Ham** and $\mathbf{r^2}$ are still computable. In this case, while error rates are increased compared to when polarized data is available (for mean OOB by $\approx 2\%$ when using $\mathbf{fSFS+}, \mathbf{r^2}, \mathbf{Phy}, \mathbf{Ham}$ instead of e.g. $\mathbf{O}, \mathbf{r2}, \mathbf{AF+}$), there is still a clear benefit of adding statistics to the **fSFS**, see Figure 7. Adding

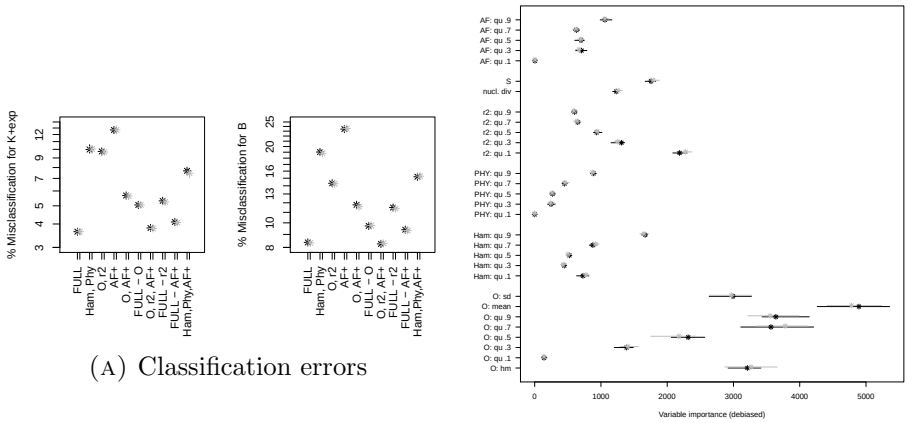


(A) Using $\mathbf{Ham}, \mathbf{Phy}, \mathbf{r^2}, \mathbf{fSFS}$ (B) No singletons

FIGURE 8. Variable importances for the model comparison between $\mathbb{K}$-*exp* and $\mathbb{B}$ (Scenario 2). See Table 6 for the legend. (A) 'fS17-49' shows the range (and means) of the importance scores of the variables $fS17, \ldots, fS49$ combined.

$\mathbf{r^2}, \mathbf{Phy}, \mathbf{Ham}$ to **fSFS**+ reduces model misclassification strongly ($\geq 5\%$ mean OOB reduction). Interestingly, adding $\mathbf{Phy}, \mathbf{Ham}$ on top of $\mathbf{r^2}$ does help, while again not all entries of the fSFS are well suited to distinguish the model classes. This is also reflected by the variable importances, which are well distributed across the different sets of statistics, while many entries of the fSFS have very low importance, see Figure 8(A).

Other factors may also influence whether all statistics can be measured (well) from SNP data. For instance, if the data are not sequenced with high quality (e.g. low coverage data), singleton mutations, i.e. mutations contributing to S_1, may not be highly reliable due to sequencing errors (or overcompensating for sequencing errors). Thus, we assessed what happens if we remove all singleton mutations before computing the statistics and perform inference with $\mathbf{O}, \mathbf{Ham}, \mathbf{r^2}, \mathbf{AF}+$. The information of the singleton mutations is important for the model selection between $\mathbb{K}$-*exp* and $\mathbb{B}$, reflected by a strong increase in error rate when comparing to results without removal of them (OOB error increase of $\geq 4\%$ when comparing inferring via $\mathbf{O}, \mathbf{Ham}, \mathbf{r^2}, \mathbf{AF}+$ without singletons). In this case, including **O** is important, reflected by the highest increase in mean OOB error for **O** when excluding exactly one set of statistics (followed by $\mathbf{r^2}$). That the eligibility of **O** for the model selection is not diminished compared to the scenario with singleton mutations is not very suprising, since **O** does not take into account singleton mutations and thus is identical in both scenarios. However, the information available changes by the exclusion of singleton mutations, which strongly affects the ranking of statistics by their importance, see Figure 8. Interestingly, the harmonic



(A) Classification errors

(B) Variable importances

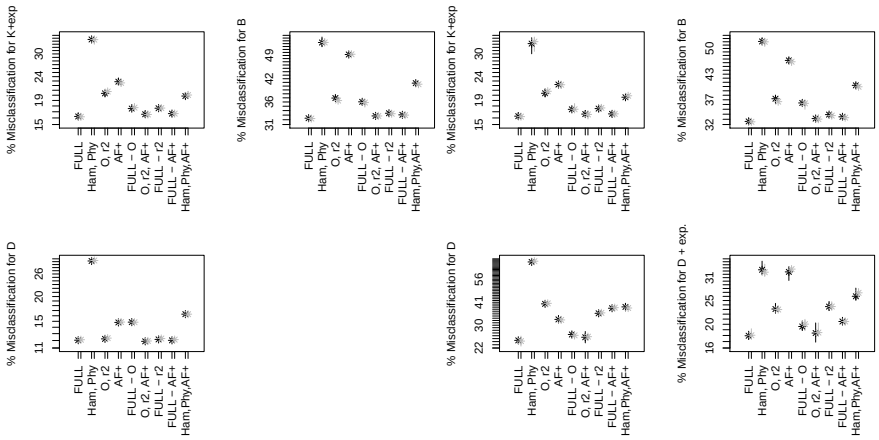
FIGURE 9. Classification errors and variable importances of model comparison between $\mathbb{K}$-*exp* and $\mathbb{B}$ (Scenario 3). See Table 6 for the legend. Full set of statistics **Phy**, **Ham**, **O**, $\mathbf{r^2}$ and **AF**+.

mean from **O** is not the most important statistic any more (high quantiles of $\mathbf{AF}+, \mathbf{r^2}$ have higher variable importance).

How realistic is Scenario 2? Some bacterial pathogens are haploid species with a single-locus genome, neglectible amount of recombination, and potential for multiple mergers, due to rapid selection or quick emergence of specific strains. Here, we look at a scenario, which we denote by Scenario 3, matching a sample of *Mycobacterium tuberculosis* from [LRP+15] regarding sample size and number of mutations. Since the Watterson estimator $\hat{\theta}_w(s)$ will be not estimating the true mutation rate perfectly, we allow θ to fluctuate around the Watterson estimate in the models, see Section 1.2. Misclassification is still common in Scenarios 1 and 2. In Scenario 3 however, classification errors drop considerably, while adding statistics to the SFS strongly decreases misclassification, see Figure 9. As in Scenario 2, $\mathbf{O}, \mathbf{r^2}$ are the meaningful additions to **AF**+ to decrease errors and adding further statistics has virtually no effect (adding $\mathbf{O}, \mathbf{r^2}$ reduces mean OOB error from $\approx 18\%$ to 6%, other combinations have increased errors).

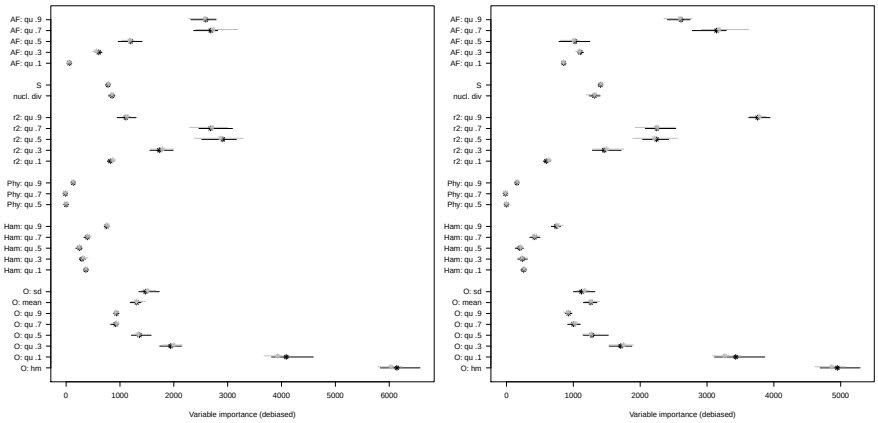
2.2. **Comparison with Dirac and Beta-Ξ-n-coalescents.** We consider again Scenario 2 with $n = 100$, where we subsequently add $\mathbb{D}$ and $\mathbb{D}$-*exp* as additional model classes. First, we perform model selection between all models, the results are shown in Figures 10 and 11.

If we compare with the (pairwise) model selection between $\mathbb{K}$-*exp* and $\mathbb{B}$ (Figure 2(B)), we see that adding $\mathbb{D}$ leads to more frequent misidentification of $\mathbb{B}$, but not of $\mathbb{K}$-*exp*. The model class $\mathbb{D}$ is easier to identify than the



(A) $\mathbb{K}$-*exp* vs. $\mathbb{B}$ vs. $\mathbb{D}$ (B) $\mathbb{K}$-*exp* vs. $\mathbb{B}$ vs. $\mathbb{D}$ vs. $\mathbb{D}$-*exp*

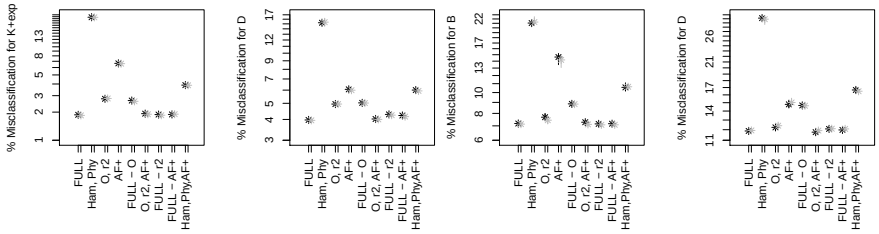
FIGURE 10. Classification errors of model comparisons in Scenario 2 with $n = 100$. See Table 6 for the legend. The full set of statistics consists of $\mathbf{r^2}$, **Phy**, **Ham**, **O** and **AF+**



(A) $\mathbb{K}$-*exp* vs. $\mathbb{B}$ vs. $\mathbb{D}$ (B) $\mathbb{K}$-*exp* vs. $\mathbb{B}$ vs. $\mathbb{D}$ vs. $\mathbb{D}$-*exp*

FIGURE 11. Variable importances in Scenario 2, $n = 100$. See Table 6 for the legend.

other two classes, having a far smaller misidentification probability. When we further add $\mathbb{D}$-*exp*, we see that misidentification between for $\mathbb{D}$-*exp* and $\mathbb{D}$ is quite common, while misidentification of $\mathbb{K}$-exp and $\mathbb{B}$ is essentially unchanged. This is due to the misidentification of $\mathbb{D}$-*exp* as $\mathbb{D}$ and vice versa, see Table A4. For the threefold model comparison, the statistics in **O** have the strongest effect when coupled with other statistics, i.e. they cause



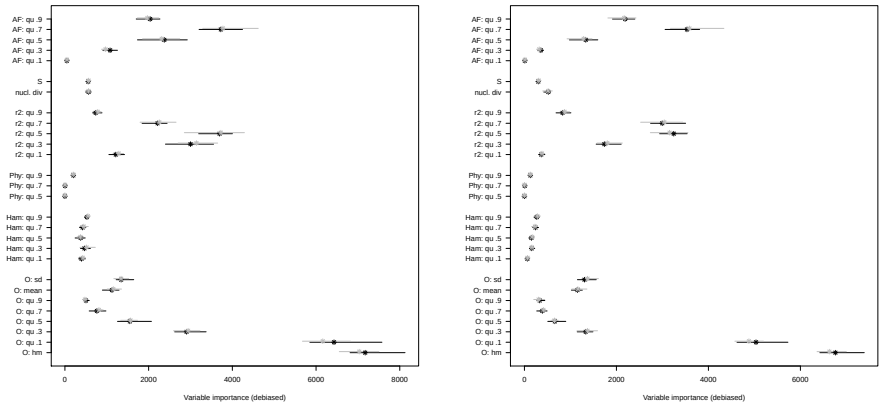
(A) $\mathbb{K}$-*exp* vs. $\mathbb{D}$ (B) $\mathbb{B}$ vs. $\mathbb{D}$

FIGURE 12. Classification errors of model comparisons in Scenario 2 with $n = 100$. See Table 6 for the legend. The full set of statistics consists of $\mathbf{r^2}$, **Phy**, **Ham**, **O** and **AF+**

the highest error increase of all sets of statistics when left out (followed by $\mathbf{r^2}$). The positive effect of **O** is weakened for the fourfold comparison. In both cases, errors are also only slightly increased when using $\mathbf{O}, \mathbf{r^2}, \mathbf{AF+}$ instead of all statistics (mean OOB error differences of $\leq 0.3\%$,$\leq 0.75\%$ for the threefold, fourfold comparisons). These findings are also mirrored by the variable importances. The variable with highest importance is again the harmonic mean of the minimal observable clade sizes (while its relative importance compared to other statistics is greater for the threefold comparison), and statistics from $\mathbf{Phy}, \mathbf{Ham}$ are not suited well to distinguish the hypotheses (at least if other statistics are available).

How do model selection errors and variable importances change if we consider other pairwise model selections? We additionally performed model selection for $\mathbb{K}$-*exp* vs. $\mathbb{D}$ and for $\mathbb{B}$ vs. $\mathbb{D}$. The results are shown in Figures 12 and 13. While again $\mathbf{O}, \mathbf{r^2}, \mathbf{AF+}$ has essentially the same classification errors as the full set of statistics, for distinguishing the Dirac coalescent **O** is a very suitable addition, both reflected by classification errors and variable importances, where for the latter the harmonic mean from **O** is again the statistic with highest importance. As already known from earlier works, e.g. [EBBF15], the Dirac coalescent is easy to distinguish both from $\mathbb{K}$-*exp* and $\mathbb{B}$, reflected by the low misclassification rates compared to the model selection between $\mathbb{K}$-*exp* and $\mathbb{B}$.

On the other hand, if we instead add Ξ-$\mathbb{B}$ as a model in Scenario 2, we see a very strong increase in misclassification rates, see Figure 14. The high chance of misclassifying $\mathbb{B}$ as Ξ-$\mathbb{B}$ stands out, see the results of the pairwise model selection approach in Figure 15. Again, performing the model comparison based on $\mathbf{O}, \mathbf{r^2}, \mathbf{AF+}$ does not increase misclassification meaningfully compared to using all statistics (if one does not simply choose not to compare models due to very high errors). While for the threefold comparison, the most important statistic still is the harmonic mean from **O**,



(A) $\mathbb{K}$-*exp* vs. $\mathbb{D}$ (B) $\mathbb{B}$ vs. $\mathbb{D}$

FIGURE 13. Variable importances for model comparisons in Scenario 2 with $n = 100$. See Table 6 for the legend.

followed by statistics from $\mathbf{AF}, \mathbf{r^2}, \mathbf{O}$, this changes if one looks at the pairwise comparisons. There, $\mathbf{O}$ only plays an important, yet diminished role for distinguishing Ξ-$\mathbb{B}$ from $\mathbb{K}$-*exp*, it seems to play much less of a role for distinguishing Ξ-$\mathbb{B}$ from $\mathbb{B}$ (Figure 16). This is not entirely mirrored by the error rate changes between different sets of summary statistics, see Figure 15, but the differences between most sets of statistics for these scenarios are small to begin with. Again, model selection based on $\mathbf{O}, \mathbf{r^2}, \mathbf{AF}+$ is also very close error-wise to using all statistics in the pairwise comparisons (mean OOB difference of $\leq 0.2\%$).

Our results imply that the distribution of diversity statistics within different (parametric) subclasses of Ξ-n-coalescents can both be rather similar or very different. This also raises the question whether misspecifying the exact multiple merger model as a different multiple merger model increases model selection errors. We analyze this by allocating simulations from one Ξ-n-coalescent class to either $\mathbb{K}$-*exp* or another Ξ-n-coalescent, see Section A.2 for details. For instance, we build a random forest from simulations of $\mathbb{K}$-*exp* and $\mathbb{B}$, but sort simulations from $\mathbb{D}$ into these two classes via ABC. The results are shown in Table 7. They highlight that multiple merger n-coalescents do not generally differ in the same way from $\mathbb{K}$-*exp*. While $\mathbb{D}$ gets nearly always identified as $\mathbb{B}$ instead of $\mathbb{K}$-*exp*, this is not symmetric: $\mathbb{B}$ gets considerably more often identified as $\mathbb{K}$-*exp*. Unsurprisingly, given the high classification errors between $\mathbb{B}$ and Ξ-$\mathbb{B}$, both models get more often identified as the other multiple merger model, with chance around 0.7, in contrast to as being identified as $\mathbb{K}$-*exp*. Still, model misspecification comes with a cost: if we use $\mathbb{B}$ as a proxy model for Ξ-$\mathbb{B}$ and have to allocate a true realization of Ξ-$\mathbb{B}$ to either $\mathbb{B}$ or $\mathbb{K}$-*exp*, this leads to considerably higher

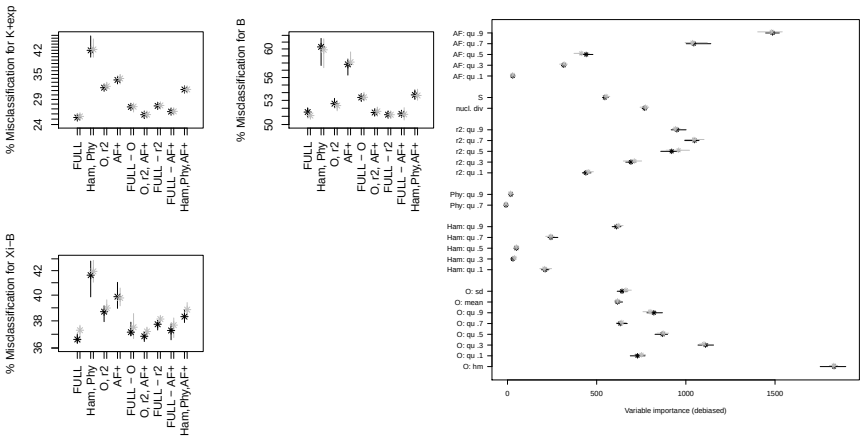
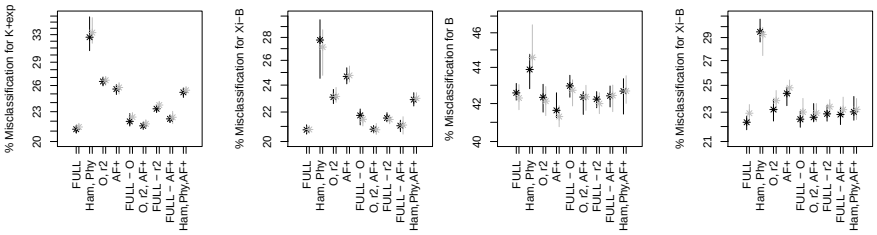


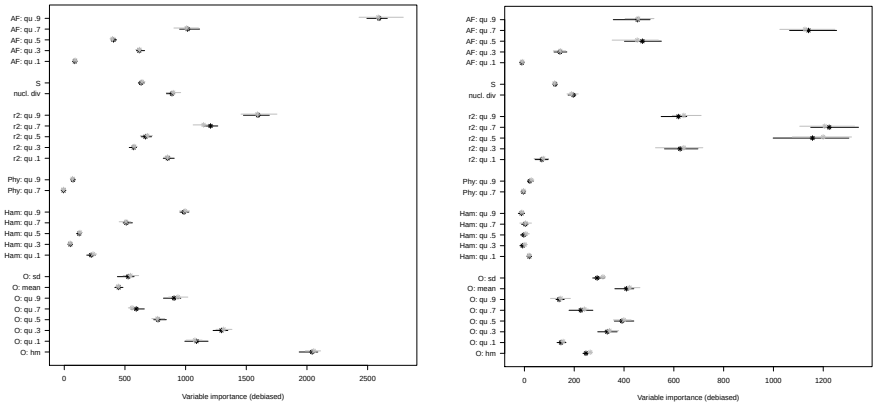
FIGURE 14. Classification errors and variable importances for model comparison $\mathbb{K}$-*exp* vs. $\mathbb{B}$ vs. Ξ-$\mathbb{B}$ in Scenario 2 with $n = 100$. See Table 6 for the legend. The full set of statistics consists of $\mathbf{r^2}$, **Phy**, **Ham**, **O** and **AF**+



(A) $\mathbb{K}$-*exp* vs. Ξ-$\mathbb{B}$ (B) $\mathbb{B}$ vs. Ξ-$\mathbb{B}$

FIGURE 15. Classification errors of model comparisons in Scenario 2 with $n = 100$. See Table 6 for the legend. The full set of statistics consists of $\mathbf{r^2}$, **Phy**, **Ham**, **O** and **AF**+

classification errors (though they should still be manageable) than when performing the model selection with the correct model Ξ-$\mathbb{B}$, see Figure 15. However, in this case only considering simulations that can be allocated to a model class with high posterior probability lowers errors, but at the (different) cost of being unable to allocate a high proportion of observed data. To a degree, one can try to balance cost and benefit here, e.g. retaining only simulations from Ξ-$\mathbb{B}$ with posterior > 0.72 for either $\mathbb{B}$ or $\mathbb{K}$-*exp*, which is fulfilled by a proportion of 0.54 of simulations, leading to a decreased identification chance as $\mathbb{K}$-*exp* of 0.18, which is comparable to the misidentification rate in the model selection using the correct model Ξ-$\mathbb{B}$.



(A) $\mathbb{K}$-*exp* vs. Ξ-$\mathbb{B}$ (B) $\mathbb{B}$ vs. Ξ-$\mathbb{B}$

FIGURE 16. Variable importances for model selection in Scenario 2, $n = 100$. See Table 6 for the legend.

A	B	C	$\to B$	$\to C$	High posterior
$\mathbb{B}$	$\mathbb{K}$-*exp*	$\mathbb{D}$	0.64 (0.72)	0.36 (0.28)	0.58
$\mathbb{D}$	$\mathbb{K}$-*exp*	$\mathbb{B}$	0.01 (0.00)	0.99 (1)	0.87
$\mathbb{B}$	$\mathbb{K}$-*exp*	Ξ-$\mathbb{B}$	0.27 (0.11)	0.73 (0.89)	0.27
Ξ-$\mathbb{B}$	$\mathbb{K}$-*exp*	$\mathbb{B}$	0.31 (0.03)	0.69 (0.97)	0.19
$2p.5$-$\mathbb{K}$-*exp*	$\mathbb{K}$-*exp*	$\mathbb{B}$	0.53 (0.64)	0.47 (0.36)	0.10
$2p.9$-$\mathbb{K}$-*exp*	$\mathbb{K}$-*exp*	$\mathbb{B}$	0.43 (0.28)	0.57 (0.72)	0.06

TABLE 7. $\to B$, $\to C$: Proportion of simulations from model class A allocated to two other model classes B and$C$ in Scenarios 2 and 4. In parentheses: same proportion only among high posterior allocations. High posterior allocations are those with posterior probability > 0.9, the column 'high posterior' shows their proportion among all simulations from A. Random forest based on $\mathbf{r^2}$, **Phy**, **Ham**, **O**, **AF**+ for the first four rows, $\mathbf{r^2}$, **Phy**, **SFS**$*$, **AF**, **O** and **Ham** for rows 5,6 (quantiles used: $.1, .2, .3, .4. \ldots, .9$). Results are rounded to two digits.

2.3. **The influence of exponential growth.** We further consider Scenario 5, focusing only on comparisons between scenarios with growth and without, see Figure 17. Even for $\mathbb{K}$ vs. $\mathbb{K}$-*exp* and high mutation counts, the model with growth is relatively often misidentified, which partly comes from our focus on low to medium growth rates, compared e.g. to [EBBF15]. For instance, subsetting for only growth rates above 20 (and drawing the same number of simulations from $\mathbb{K}$) leads to a probability of $< 1\%$ to misidentify

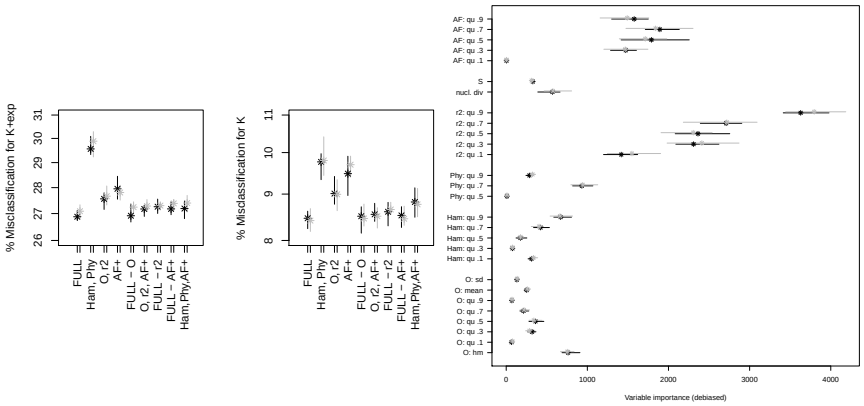


FIGURE 17. Classification errors and variable importances for model comparison $\mathbb{K}$ vs. $\mathbb{K}$-*exp* in Scenario 5 ($n = 100$). See Table 6 for the legend. The full set of statistics consists of $\mathbf{r^2}$, **Phy**, **Ham**, **O** and **AF**+

for both comparisons. Additionally, the larger error for identifying $\mathbb{K}$-*exp* is also affected by the prior: we have only a relatively small number of simulations from models close to $\mathbb{K}$ (i.e. with very small growth rates) which are very hard to distinguish from $\mathbb{K}$. Such simulations, if left out of a bootstrap sample, then get easily sorted into a leaf associated to the more numerous simulations from $\mathbb{K}$. Generally, for distinguishing these two model classes, adding statistics to **AF** decreases misclassification in the random forest-based ABC model selection only slightly (at most by 1%) and there is no specific set of statistics that has information which is non-redundant for all other sets.

If one looks at distinguishing $\mathbb{D}$ from $\mathbb{D}$-*exp*, now again with the lower mutation rates of Scenario 2, we see increased error probabilities (Figure 18). While $\mathbb{D}$ could be well distinguished from $\mathbb{B}$ and $\mathbb{K}$-*exp*, see Figure 12, it is much more similar in its summary statistics to $\mathbb{D}$-*exp*. In contrast to model selection between $\mathbb{K}$-*exp* and $\mathbb{K}$, adding statistics to **AF** does strongly decrease misclassification, which is directly attributable to adding $\mathbf{r^2}$.

For both scenarios distinguishing exponential growth from no growth, $\mathbf{r^2}$ (especially the .9-quantile) and **AF**+ show high importance, while the statistics from **O** are non-important (Figures 17 and 18). This is in accordance with the misclassification rates.

2.4. **The influence of population structure.** For two very simplistic scenarios with two subpopulations and low symmetric migration rates, where either equal samples or skewed samples are taken from the subpopulations, we find that model classes $\mathbb{K}$-*exp*, $\mathbb{B}$ and $2p5/2p9$-$\mathbb{K}$-*exp* with population

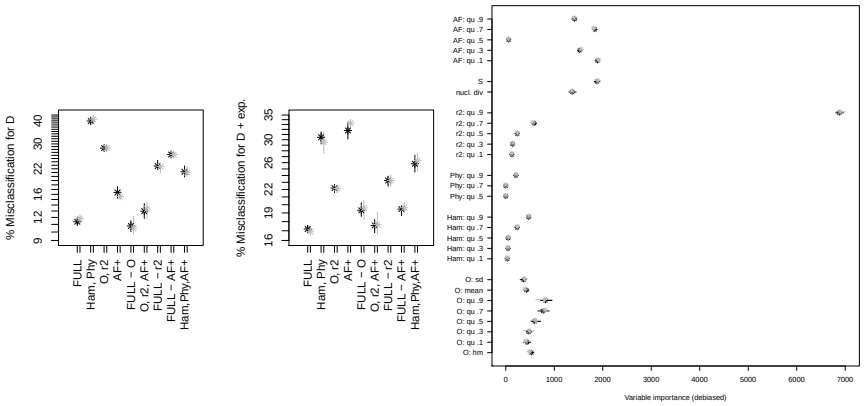


FIGURE 18. Classification errors and variable importances of model comparison $\mathbb{D}$ vs. $\mathbb{D}$-*exp* in Scenario 2 for $n = 100$. See Table 6 for the legend. The full set of statistics consists of $\mathbf{r^2}$, **Phy**, **Ham**, **O** and **AF**+.

structure can be distinguished quite well compared to other sets of models, see Figure 19. As statistics, we use $\mathbf{AF}+, \mathbf{Ham}, \mathbf{Phy}, \mathbf{r^2}, \mathbf{O}$. Again, we see the highest error reduction when coupling $\mathbf{O}, \mathbf{AF}+, \mathbf{r^2}$, essentialy with the same error rates as when using all statistics ($\leq 0.5\%$ mean OOB error difference), while omitting either **O** or $\mathbf{r^2}$ increases errors slightly ($\geq 1\%$ mean OOB error difference). Due to their known potential to distinguish demography from population structure ability, we also checked whether performing model selection using Tajima's D and Fay and Wu's H together with the SFS, π, S, instead of **AF**+ lowers misclassification. For both population structure choices, combining either **AF**+ or $\{\mathbf{SFS}, \pi, S, D, H\}$ with $\mathbf{Ham}, \mathbf{Phy}, \mathbf{O}, \mathbf{r^2}$ essentially produces identical errors (using **AF**+ decreases mean OOB error further by $\leq .2\%$). The same holds true when using $\mathbf{O}, \mathbf{r^2}, \mathbf{AF}+$ or $\mathbf{O}, \mathbf{r^2}, \mathbf{Si.t}*$ (the latter further decreases mean OOB error by $\leq 0.05\%$).

When looking at variable importances, we see that the importance of the harmonic mean of the minimal observable clade sizes is reduced compared to other scenarios, while high allele frequency and Hamming distance quantiles are elevated, see Figure 20. This may reflect that strong population structure tends to produce long branches in the genealogy connecting the root to the samples from different subpopulations, causing bumps in the allele frequency distribution and an abundance of high Hamming distances.

The error rates strongly drop for pairwise comparisons between model classes, see Figures 21 and A11. First, have a look at selecting between $\mathbb{K}$-*exp* with and without population structure. While for equal-sized samples from the subpopulations, adding statistics to **AF**+ does not considerably lower misidentification ($< 1\%$ mean OOB error decrease for any addition), this

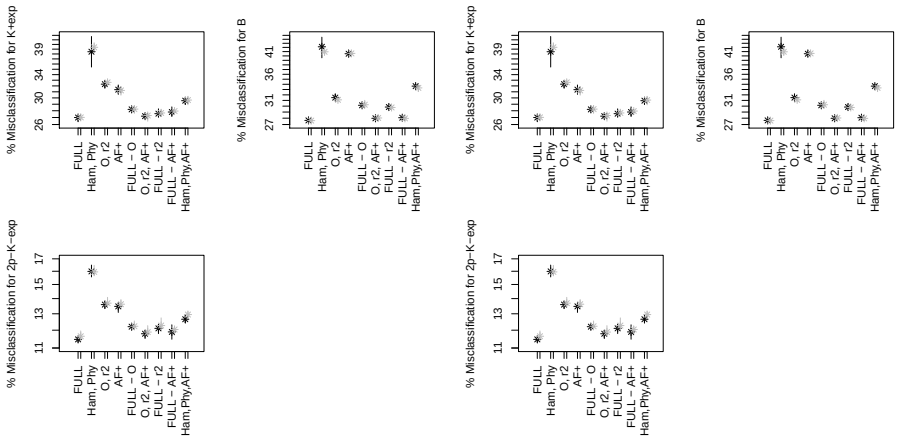
can be the case for unbalanced sampling. For subsample sizes 9:1, we see a drop-off in mean OOB error of $\approx 3.3\%$. In both cases, there is not a strong difference which set of statistics is added, all tested sets added to $\mathbf{AF}+$ only differ by $\leq 1\%$ in mean OOB error for model inference, with $\mathbf{O}, \mathbf{r^2}, \mathbf{AF}+$ being again very close ($\leq 0.3\%$ mean OOB error difference). For single variables, importance scores are highest for statistics from $\mathbf{AF}+$ and **Ham**, see Figures 22(A) and A12(A). The high importance of statistics from **Ham** is in contrast to most other scenarios. For both equal and unequal sampling from the sub-populations, the .9-quantile statistics from the two sets are among the most important statistics for distinguishing $\mathbb{K}$-*exp* with and without population structure.

For selecting between $\mathbb{K}$-*exp* with population structure and $\mathbb{B}$, the additional error decrease when adding statistics to $\mathbf{AF}+$ is slightly stronger (4% or higher decrease of mean OOB error), which is due to the addition of $\mathbf{O}, \mathbf{r^2}$, with a stronger effect of the former. For equal-sized supsamples, all statistics but **Phy** have at least some member with reasonable importance (Figure 22(B)), again (quantiles of the) Hamming distances are among the variables with highest importance. For a sampling ratio of 9:1 from the two subpopulations, again the .9 quantiles of allele frequencies and Hamming distances show high importance, as does $\mathbf{r^2}$, see Figure A12(B). Surprisingly, given the stronger effect of **O** on misclassification than the other statistics when added to $\mathbf{AF}+$, no single statistic of **O** shows high importance.

While the simple scenario of strong population structure and two subpopulations can be quite well distinguished from $\mathbb{K}$-*exp* and $\mathbb{B}$, this only works when population structure is known *a priori*. Indeed, if we allocate simulations from $2p.5$-$\mathbb{K}$-*exp* to $\mathbb{K}$-*exp* or $\mathbb{B}$, i.e. perform model selection between the latter two models based on data that truly follows $2p.5$-$\mathbb{K}$-*exp*, Table 7 reveals that these simulations are just randomly assigned to one of the model classes. To be allowed to ignore population structure for inference, such simulations should be sorted into $\mathbb{K}$-*exp* with high probability, which is clearly not the case. This effect persists even if one only regards simulations that are allocated to one model class with high posterior probability. While for equal-sized samples from the subpopulations, the effect is weakened, it is increased for sampling ratio 9:1, even allocating simulations from a model with exponential growth and population structure ≈ 2.5 times more often to the multiple-merger model than to $\mathbb{K}$-*exp* without population structure.

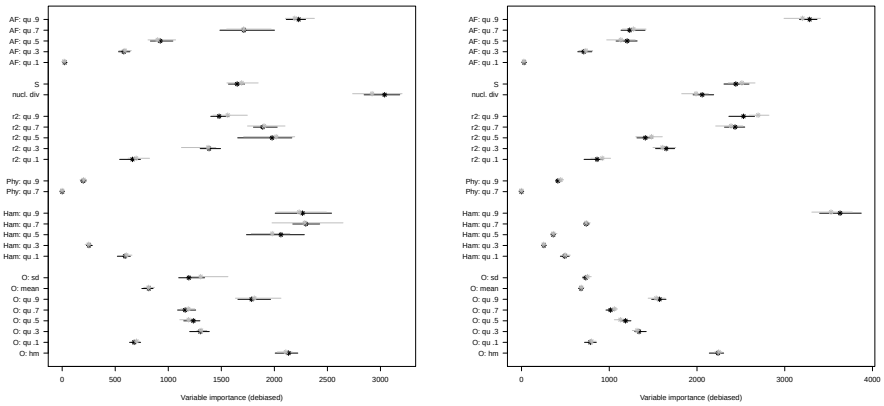
3. Discussion

Our results highlight that the information not contained in the site frequency spectrum is still highly valuable for distinguishing between different classes of n-coalescents, e.g. between exponential growth and multiple merger coalescents. This beneficial effect of using site frequency spectra with other statistics of genetic diversity has been also recently reported for other coalescent models (also incorporating recombination) in [JBA19]. In addition



(A) $\mathbb{K}$-*exp* vs. $\mathbb{B}$ vs. $2p.5$-$\mathbb{K}$-*exp* (B) $\mathbb{K}$-*exp* vs. $\mathbb{B}$ vs.$2p.9$-$\mathbb{K}$-*exp*

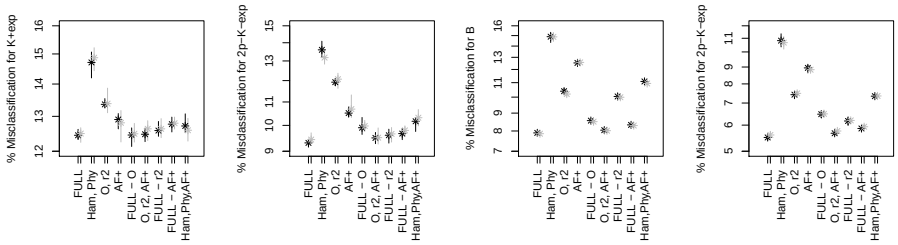
FIGURE 19. Classification errors in Scenario 4, $n = 100$. See Table 6 for the legend. Full set of statistics includes $\mathbf{r^2}$, **Phy**, **AF+**, **O** and **Ham**.



(A) $\mathbb{K}$-*exp* vs. $\mathbb{B}$ vs. $2p.5$-$\mathbb{K}$-*exp* (B) $\mathbb{K}$-*exp* vs. $\mathbb{B}$ vs. $2p.9$-$\mathbb{K}$-*exp*

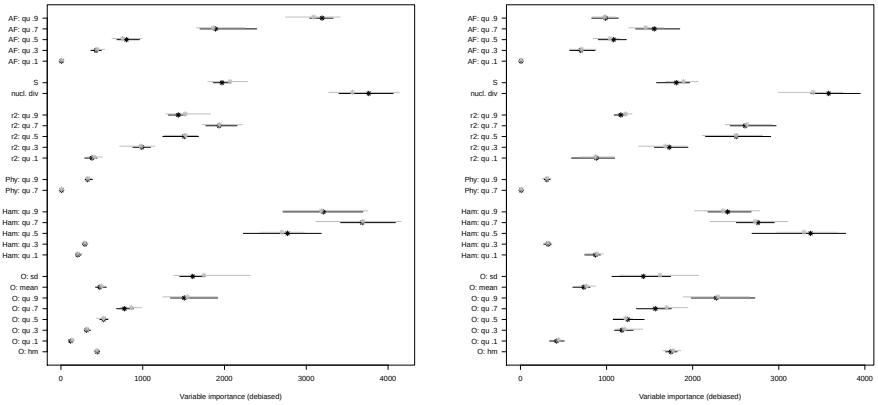
FIGURE 20. Variable importances in Scenario 4, $n = 100$. See Table 6 for the legend.

to scenarios with fixed mutation rate ranges across models as in [KVS$^+$17], this also holds true for scenarios where mutation rates are chosen so that, on average, each model produces the same number of mutations. While we concentrated on full, polarized genetic data, there is still a strong beneficial effect of using more statistics than allele frequencies spectra when performing model selection, i.e. where data cannot be polarized or if one ignores singleton mutations (due to their susceptibility to being mixed with



(A) $\mathbb{K}$-*exp* vs. $2p.5$-$\mathbb{K}$-*exp* (B) $\mathbb{B}$ vs. $2p.5$-$\mathbb{K}$-*exp*

FIGURE 21. Classification errors for model selection in Scenario 4, $n = 100$. See Table 6 for the legend. Full set of statistics includes $\mathbf{r^2}$, **Phy**, **AF+**, **O** and **Ham**.



(A) $\mathbb{K}$-*exp* vs. $2p.5$-$\mathbb{K}$-*exp* (B) $\mathbb{B}$ vs. $2p.5$-$\mathbb{K}$-*exp*

FIGURE 22. Variable importances in Scenario 4, $n = 100$. See Table 6 for the legend.

sequencing errors). The benefit of including statistics not based on the SFS is especially relevant for species with small genomes and few unlinked loci, e.g. for some bacteria, but also likely for some selfing plants and clonally propagating fungi. Giving high enough mutation rates and reasonable sample sizes, as in our Scenario 3, one can distinguish exponential growth from multiple mergers with ease. We emphasize that seeing hundreds of mutations at a single locus in a species with low recombination rates is not unrealistic (though clearly not always the case), we based our setting on real data sets in *Mycobacterium tuberculosis*. While model misidentification becomes less of an issue if the genome consists of many unlinked loci, see [Kos18], this may still be relevant if further evolutionary processes are acting, as discussed in [KWB19].

The computational burden of obtaining statistics can be considerable. Usually, one has many more SNPs than individuals, thus punishing more strongly statistics whose computational load scales with the number of mutations, most notably $\mathbf{r^2}$ (see Table A1). The statistics based on the allele frequencies are often rather inexpensive to compute, so we focused on adding further statistics to these.

In most model comparisons, especially the quantiles of the computationally more costly LD statistic r^2 were helpful to distinguish the model classes when added to allele-frequency based statistics. This reflects nicely that combining allele frequencies with linkage information is a good strategy to distinguish multiple-merger n-coalescents from other genealogy models, as exploited in [RND18]. For many scenarios involving multiple-merger coalescents, the statistics based on the minimal observable clade sizes are a further meaningful addition to the information summarized in the allele frequency spectrum, if sample sizes are not too low.

Quantiles of the Hamming distances between sequences seem not to be helpful for these model selections as additions to $\mathbf{AF+}, \mathbf{O}, \mathbf{r^2}$ (we see nucleotide diversity, the mean of the Hamming distances, as an allele frequency based statistic). Still, at least when distinguishing models from Kingman's n-coalescent with population structure, some of these quantiles feature high variable importances. The quantiles of the branch lengths of a phylogenetic reconstruction of the genealogy seem to have even less of an effect than the quantiles of the Hamming distances. They were not relevant in any scenario, the lower quantiles often even not showing variation (and quite costly to compute for larger samples or genetic regions, i.e. higher number of mutations).

Three observations lead us to believe that the essential null effect of the two sets $\mathbf{Ham}, \mathbf{Phy}$ of statistics on model selection when added to the other statistics is less that these statistics carry no or few information about the differences between the model classes, but that this or similar information is also coded in other statistics. First, there are scenarios where $\mathbf{Ham}, \mathbf{Phy}$ allows to distinguish the model classes better than other sets of statistics (but worse than their combination): If coalescent models feature similar mutation rates (Scenario 1) or small sample sizes ($n = 25$), we found that inferring model classes via $\mathbf{Ham}, \mathbf{Phy}$ alone is better than inference based on either $\mathbf{O}, \mathbf{r^2}$ or $\mathbf{AF+}$ alone (we also find, in Scenario 2, that combining $\mathbf{Ham}, \mathbf{Phy}$ decreases errors slightly, about 1% decrease in mean OOB error, compared to inference with only **Ham**, see Figures 2 and 6). In a high sample size, high mutation rate scenario (Scenario 3), $\mathbf{Ham}, \mathbf{Phy}$ performs in between the misclassification of the two other sets. Second, if one wants to distinguish Kingman's n-coalescent with exponential growth and population structure from other coalescent models, high quantiles of the Hamming distances show high importance. Third, for non-polarized data, adding $\mathbf{Ham}, \mathbf{Phy}$ does meaningfully reduce misclassification.

Generally, in all scenarios (when data can be polarized), combining the allele frequency spectrum with r^2-statistics and the minimal observable clade sizes produces essentially the same classification errors for model selection as using all statistics we analyzed. Our recommendation is thus to use this combination for model selection between genealogy models, especially if multiple-merger n-coalescents are involved (for polarized data). Interestingly, at least when adding statistics to the allele frequency-based statistics, the information coded by the quantiles of the statistics that helps to distinguish the model classes seems to be already contained in rather few quantiles, i.e. already 5 quantiles of the statistics considered allow model selection with comparable errors as a finer mesh of quantiles. Additionally, our analysis shows that the information of the combination of the different statistics allows one to cope with not knowing the mutation rate precisely and that, when statistics are combined, scaling statistics to be non-sensitive to changes in the mutation rate is not necessary (while it still is reasonable when e.g. only using allele frequency spectra).

Our results can be seen as emphasizing the potential for constructing computationally efficient statistical methods for genealogy model selection based on extracting only relevant features of the data, e.g. as done in [PVC$^+$19]. While we have included a wide range of non-SFS-based statistics of genetic diversity, further interesting statistics to add to the SFS are available, most notably including linkage information via the two-site frequency spectrum as in [RND18].

Compared with the results in [KVS$^+$17], our results show highly increased error rates. This is likely not a consequence of our ABC approach, but of our focus on distinguishing Beta$(2-\alpha,\alpha)$-n-coalescents for $\alpha \in [1,2)$ from Kingman's n-coalescent with growth, while [KVS$^+$17] also add further Beta coalescents that are easier to distinguish from the bifurcating genealogies. Nevertheless, this specific set of Beta$(2-\alpha,\alpha)$-n-coalescents is chosen here because it arises as limit genealogies in an explicit evolutionary model, see [Sch03].

As reported in [KWB19], considering Kingman's n-coalescent with both exponential growth and population structure often causes dramatic misidentification of whether multiple merger coalescents are the correct genealogy models or not. Our results confirm that this is not only a consequence of the use of the singletons and the tail of the site frequency spectrum as the statistics used to perform model selection. Even when adding non-SFS-based statistics, non-accounted population structure makes correct inference of multiple merger genealogies non-feasible. For data sets containing many unlinked loci, assessing population structure may be, at least approximately, possible with standard model-free methods like DAPC [JDB10] or principal component analysis. On the plus side, both our results and the results from [KWB19] show that the genetic diversity between model classes with and without population structure (and multiple mergers) is different enough so that additional population structure should be detectable, the open question

is how to do it. One potential avenue based on the ABC with random forest approach would be to not estimate population structure at all, but instead to include a further model class that combines Kingman's n-coalescent with exponential growth and not one specific model of population structure but the complete range of reasonable populations structure setups. We tested this in very limited fashion for populations with two subpopulations of equal size, variable migration rates and subsample sizes, see Section A.5. At least in this scenario, Table A2 shows that Kingman's n-coalescent with exponential growth, Beta-n-coalescents or Kingman's n-coalescent with exponential growth and specific models of population structure can be allocated to the true model class with much lower error rates than when just allocating to the models without accounting for population structure. This works both for adding population structure with parameters already featuring in the model class used for training the random forest or not, and does not need a precise *a priori* knowledge or independent (precise) estimation of the population structure.

We introduced a new statistic, the minimal observable clade size, defined in Equation (1). This statistic, for a non-recombining locus, gives the clade size of the smallest genealogical clade each individual is a member of, and that can be distinguished from data, see Section A.1. It is not very costly to compute, its computational cost scales well with both population size and mutation rate (see Table A1) and, for many scenarios, decreases error rates considerably if (and only if) coupled with other summary statistics. While it needs polarized data, i.e. one needs to be able to tell apart ancestral and derived states of segregating sites, it is not taking into account any singleton mutations, which makes it relatively robust towards sequencing errors, as discussed in [Ach08]. Intuitively, the statistic should be suited well for distinguishing between multiple merger genealogies and Kingman's n-coalescent with exponential growth since it is an upper bound of the size of the minimal clade, i.e. the size of the clade spanned by the most recent ancestor node in the genealogy of each individual. For high enough mutation rates, the minimal observable clade size should equal it or be close to it. The minimal clade is known to be typically small for Kingman's n-coalescent, while larger for Beta-n-coalescents [BF05], [FSJ14] and [SJY16]. Thus, it would be a well-suited statistic to distinguish the different coalescent models, see Figure A1, but it is not directly observable from SNP data (in contrast to the minimal observable clade size). Additionally, the distribution of the minimal clade size is, by definition, invariant under time changes of the coalescent model that do not change its tree topology. Such time-changes of coalescents are the result of non-extreme population size changes in the underlying discrete population models, as mentioned in Section 1.1. This should translate to a weak effect of population size changes on the minimal observable clade sizes, which is suggested by our results.

Interestingly, the harmonic mean of all minimal observable clade sizes often

shows one of the highest variable importance among all statistics. A heuristic argument for this is that some of the minimal observable clades may be also minimal clades (i.e. if there is a mutation on the branch adjacent to the ancestor of the minimal clade on its path to the root of the genealogy), thus inheriting the potential of the minimal clade to distinguish between coalescent models. Additionally, not only the information about small clades is included in the harmonic mean, which may prove beneficial. The sizes of small minimal observable clades is likely not very informative to distinguish different coalescent models. For instance, the minimal size of all minimal observable clades, at least for large n and θ, should be 2 for many different coalescent models. For Kingman's n-coalescent, this follows from [DK15, Corollary p.5] and for Beta coalescents from [BBL14, Thm. 8].

4. Code availability

Please see the public R code repository `https://github.com/fabianHOH/mmc_R_gendiv` for the code used for coalescent simulation, computation of the diversity statistics and performing the ABC analysis. The repository also contains the ABC results for each model comparison (as `R` objects).

FF was funded by DFG grant FR 3633/2-1 through Priority Program 1590: Probabilistic Structures in Evolution. The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

References

[Ach08] Guillaume Achaz. Testing for neutrality in samples with sequencing errors. *Genetics*, 179(3):1409–1424, 2008.

[BBL14] Julien Berestycki, Nathanaël Berestycki, and Vlada Limic. Asymptotic sampling formulae for Λ-coalescents. *Ann. Inst. H. Poincare Probab. Statist.*, (3):715–731, 2014.

[BCEH16] Jochen Blath, Mathias Christensen Cronjäger, Bjarki Eldon, and Matthias Hammer. The site-frequency spectrum associated with Ξ-coalescents. *Theor. Pop. Biol.*, 110:36–50, 2016.

[BF05] Michael G. B. Blum and Olivier François. Minimal clade size and external branch length under the neutral coalescent. *Adv. in Appl. Probab.*, 37(3):647–662, 06 2005.

[BLS18] Matthias Birkner, Huili Liu, and Anja Sturm. Coalescent results for diploid exchangeable population models. *Electron. J. Probab.*, 23(49):44, 2018.

[CB17] Sònia Casillas and Antonio Barbadilla. Molecular population genetics. *Genetics*, 205(3):1003–1035, 2017.

[Che11] Wei-Chen Chen. *Overlapping codon model, phylogenetic clustering, and alternative partial expectation conditional maximization algorithm.* PhD thesis, Iowa State University, 2011.

[DK15] Iulia Dahmer and Götz Kersting. The internal branch lengths of the Kingman coalescent. *Ann. Appl. Probab.*, 25(3):1325–1348, 2015.

[DS05] Rick Durrett and Jason Schweinsberg. A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stochastic Proc. Appl.*, 115(10):1628 – 1657, 2005.

[Dur08] Richard Durrett. *Probability models for DNA sequence evolution*. Springer Science & Business Media, 2008.

[DWF13] Michael M. Desai, Aleksandra M. Walczak, and Daniel S. Fisher. Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics*, 193(2):565–585, 2013.

[EBBF15] Bjarki Eldon, Matthias Birkner, Jochen Blath, and Fabian Freund. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics*, 199(3):841–856, 2015.

[EW06] Bjarki Eldon and John Wakeley. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172(4):2621–2633, 2006.

[Fre19] Fabian Freund. Cannings models, populations size changes and multiple-merger coalescents. *Preprint on Arxiv*, 2019.

[FSJ14] Fabian Freund and Arno Siri-Jégousse. Minimal clade size in the Bolthausen-Sznitman coalescent. *J. Appl. Probab.*, 51(3):657–668, 2014.

[FSJ19] Fabian Freund and Arno Siri-Jégousse. The minimal observable clade size of exchangeable coalescents. *Preprint on Arxiv*, 2019.

[FW00] Justin C. Fay and Chung-I Wu. Hitchhiking under positive darwinian selection. *Genetics*, 155(3):1405–1413, 2000.

[GCMPSJ19] Adrián González-Casanova, Verónica Miró Pina, and Arno Siri-Jégousse. The symmetric coalescent and Wright-Fisher models with bottlenecks. *Preprint on Arxiv*, 2019.

[GT94] Robert C. Griffiths and Simon Tavare. Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 344(1310):403–410, 1994.

[HSJB19] Asger Hobolth, Arno Siri-Jegousse, and Mogens Bladt. Phase-type distributions in population genetics. *Theor. Pop. Biol.*, 127:16–32, 2019.

[ILM+16] Kristen K. Irwin, Stefan Laurent, Sebastian Matuszewski, Severine Vuilleumier, Louise Ormond, Hyunjin Shim, Claudia Bank, and Jeffrey D. Jensen. On the importance of skewed offspring distributions and background selection in virus population genetics. *Heredity*, 2016.

[JBA19] Flora Jay, Simon Boitard, and Frédéric Austerlitz. An abc method for whole-genome sequence data: inferring paleolithic

and neolithic human expansions. *Molecular biology and evolution*, 36(7):1565–1579, 2019.

[JDB10] Thibaut Jombart, Sébastien Devillard, and François Balloux. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics*, 11(1):94, 2010.

[Kin82] John Frank Charles Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.

[KJ18] Peter D. Keightley and Benjamin C. Jackson. Inferring the probability of the derived versus the ancestral allelic state at a polymorphic site. *Genetics*, 2018.

[Kos18] Jere Koskela. Multi-locus data distinguishes between population growth and multiple merger coalescents. *Stat. Appl. Genet. Mol. Biol.*, 17(3), 2018.

[KVS+17] Mamoru Kato, Daniel A. Vasco, Ryuichi Sugino, Daichi Narushima, and Alexander Krasnitz. Sweepstake evolution revealed by population-genetic analysis of copy-number alterations in single genomes of breast cancer. *Royal Sic. Open Sci.*, 4(9), 2017.

[KWB19] Jere Koskela and Maite Wilke Berenguer. Robust model selection between population growth and multiple merger coalescents. *Math. Biosci.*, 311:1–12, 2019.

[LRP+15] Robyn S. Lee, Nicolas Radomski, Jean-Francois Proulx, Ines Levade, Jesse B. Shapiro, Fiona McIntosh, Hafid Soualhine, Dick Menzies, and Marcel A. Behr. Population genomics of Mycobacterium tuberculosis in the inuit. *Proc. Natl. Acad. Sci. USA*, 112(44):13609–13614, 2015.

[MHAJ18] Sebastian Matuszewski, Marcel E. Hildebrandt, Guillaume Achaz, and Jeffrey D. Jensen. Coalescent processes with skewed offspring distributions and non-equilibrium demography. *Genetics*, 208(1):323–338, 2018.

[Möh98] Martin Möhle. Robustness results for the coalescent. *J. Appl. Probab.*, 35(2):438–447, 1998.

[Möh06] Martin Möhle. On the number of segregating sites for populations with large family sizes. *Adv. in Appl. Probab.*, 38(03):750–767, 2006.

[MS01] Martin Möhle and Serik Sagitov. A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.*, 29(4):1547–1562, 2001.

[NH13] Richard A. Neher and Oskar Hallatschek. Genealogies of rapidly adapting populations. *Proc. Natl. Acad. Sci. USA*, 110(2):437–442, 2013.

[NNY16] Hiro-Sato Niwa, Kazuya Nashida, and Takashi Yanagimoto. Reproductive skew in japanese sardine inferred from dna sequences. *ICES J. Mar. Sci.*, 73(9):2181–2189, 2016.

[Pit99] Jim Pitman. Coalescents with multiple collisions. *Ann. Probab.*, 27(4):1870–1902, 1999.

[PME+15] Pierre Pudlo, Jean-Michel Marin, Arnaud Estoup, Jean-Marie Cornuet, Mathieu Gautier, and Christian P. Robert. Reliable ABC model choice via random forests. *Bioinformatics*, 32(6):859–866, 2015.

[PVC+19] Julia A. Palacios, Amandine Véber, Lorenzo Cappello, Zhangyuan Wang, John Wakeley, and Sohini Ramachandran. Bayesian estimation of population size changes by sampling tajima's trees. *Genetics*, 213(3):967–986, 2019.

[RND18] Daniel P. Rice, John Novembre, and Michael M. Desai. Distinguishing multiple-merger from Kingman coalescence using two-site frequency spectra. *bioRxiv*, 2018.

[SBB13] Matthias Steinrücken, Matthias Birkner, and Jochen Blath. Analysis of DNA sequence variation within marine species using Beta-coalescents. *Theor. Pop. Biol.*, 87:15–24, 2013.

[SBZH07] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

[Sch00] Jason Schweinsberg. Coalescents with simultaneous multiple collisions. *Electron. J. Probab.*, 5:1–50, 2000.

[Sch03] Jason Schweinsberg. Coalescent processes obtained from supercritical Galton–Watson processes. *Stochastic Proc. Appl.*, 106(1):107–139, 2003.

[Sch17] Jason Schweinsberg. Rigorous results for a population model with selection II: genealogy of the population. *Electron. J. Probab.*, 22, 2017.

[SJY16] Arno Siri-Jégousse and Linglong Yuan. Asymptotics of the minimal clade size and related functionals of certain Beta-coalescents. *Acta Appl. Math.*, 142(1):127–148, 2016.

[Ste09] Matthias Steinrücken. *Multiple merger coalescents and population genetic inference*. PhD thesis, Berlin, Techn. Univ., Diss., 2009, 2009.

[SZ08] Marco Sandri and Paola Zuccolotto. A bias correction algorithm for the Gini variable importance measure in classification trees. *J. Comput. Graph. Statist.*, 17(3):611–628, 2008.

[Taj89] Fumio Tajima. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3):585–595, 1989.

[TL14] Aurelien Tellier and Christophe Lemaire. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Mol. Ecol.*, 23(11):2637–2652, 2014.

[WH98] Hilde M. Wilkinson-Herbots. Genealogy and subpopulation differentiation under various models of population structure.

J. Math. Biol., 37(6):535–585, Dec 1998.

[WZ17] Marvin N. Wright and Andreas Ziegler. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.*, 77(i01), 2017.

Appendix A. Supplementary information

A.1. **The minimal observable clade [FSJ19].** Consider any (genealogical) tree with n leaves with s mutations on its branches interpreted under the infinite sites model and its corresponding SNP matrix. A clade of the genealogy is a set of leaves that have a common ancestor. Consider all clades that include i. In a tree, these need to be nested. Any mutation describes a clade in the tree, the set of all sequences sharing this mutation. A clade can only be distinguished from the information contained in the SNP matrix if there is a mutation directly on the branch connecting the ancestor of the clade to the rest of the tree in direction towards the root. Implicitly, this requires that we can distinguish the mutation from the ancestral type. Each non-private mutation of individual i, i.e. which appears in i and at least one other individual, that has the minimal allele count among all mutations of i, defines the same clade, the minimal observable clade of i. Since the minimal clade of i is the smallest clade containing i, it has to be nested within the minimal observable clade of i. Thus, the size M_i of the minimal clade satisfies $M_i \leq O_i$. The minimal observable clade is the minimal clade ($M_i = O_i$) if and only if there is one or more mutations on the branch connecting the ancestor of the minimal clade to its further path to the root of the genealogy. The minimal clade is only observable in the data if it equals the observable clade.
The distribution of the minimal clade size of an individual i is known for fixed sample size n for both Kingman's n-coalescent [BF05] and the Bolthausen-Sznitman n-coalescent $\mathbb{B}$ with parameter $\alpha = 1$ [FSJ14]. Thus, we can compute exact likelihood ratios, which we show for $n = 100$ in Figure A1.

A.2. **Model selection for observed/pseudo-observed data.** Given a random forest trained to distinguish between model classes, the approach from [PME+15] performs model selection for a set of observed summary statistics by a majority vote: The model that the data is allocated to by the decision trees in the (simple) majority of cases is the selected one. A second approach from [PME+15] also allows to give an estimate of the posterior probability for this choice of model M: First, for each simulation, the (out-of-bag) misclassifier $1_{\bar{M}}$ from the majority vote all trees built without them, i.e. 0 if the majority vote is model M and 1 otherwise, is recorded. Then, a second random forest from the same simulations is trained to perform a regression of these values $1_{\bar{M}}$ on the summary statistics, i.e. to predict $1_{\bar{M}}$ using a random forest of decision trees on the summary statistics (the trees are built slightly different as for model selection, see [PME+15] for details), the estimate is then the average over the forest. Then, for the

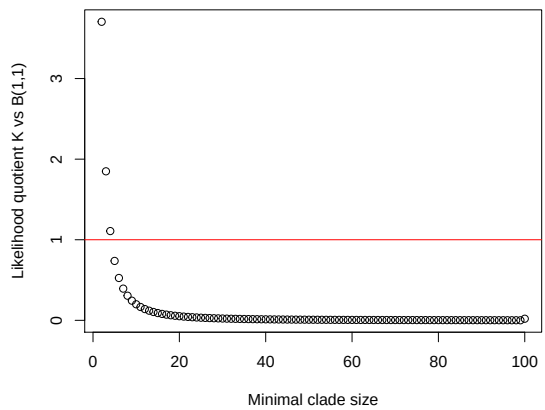


FIGURE A1. Likelihood ratios for the minimal clade size between Kingman's 100-coalescent and the Bolthausen-Sznitman 100-coalescent

actual observed summary statistic, the posterior probability of classifying as M is reported as the average estimate of $1 - 1_{\bar{M}}$ over the second random forest.

We use this is approach to assess the effects of model misspecification, i.e. we perform model selection between two or more model classes for simulations from a further, different class. The simulations from the further model class are treated as (pseudo-)observed data: For each simulation, the best fitting model class from the other model classes is recorded (so the winner of the majority vote), alongside its posterior probability.

A.3. **Computation times for summary statistics.** We measured the time necessary to compute the statistics for 1,000 simulations for $\mathbb{K}$-*exp* and $\mathbb{B}$ in 3 scenarios (with ten replications): Scenario 2, Scenario 3 with mutation rate θ fixed as the generalized Watterson estimate for each model parameter choice and a new scenarion which is as Scenario 3, but with increased sample size $n = 500$ and increased mutation rates given by the Watterson estimate based on $s = 500$ mutations assumed to be observed. This scenario is referred to as Scenario 3+. The computation times are given for our implementation of the statistics. Times were measured via the `R` commmand `system.time` and the total elapsed time is reported. For all sets, quantiles $(.1, .3, .5, .7, .9)$ were used. For $\mathbf{O}$, we compute quantiles and the harmonic mean, and for $\mathbf{AF}+$, we compute the quantiles and separately π and S.

A.4. **Simulations of robust statistics.** For the simulation scheme for Scenario 2 for model classes $\mathbb{K}$-*exp* and , we performed 350,000 simulations

Set	Scenario 2	Scenario 3	Scenario 3+
AF+	$[0.915, 1.036]$	$[3.918, 5.011]$	$[6.990, 8.632]$
Ham	$[2.976, 3.707]$	$[42.347, 47.001]$	$[138.085, 147.812]$
O	$[5.826, 7.444]$	$[33.419, 40.811]$	$[78.437, 91.581]$
Phy	$[5.043, 6.038]$	$[48.566, 52.441]$	$[168.210, 175.143]$
$\mathbf{r^2}$	$[1.674, 2.018]$	$[91.630, 100.110]$	$[183.471, 197.454]$

TABLE A1. Range $[\min, \max]$ of computation times for 1000 simulations in seconds for different sets of summary statistics across 10 replications

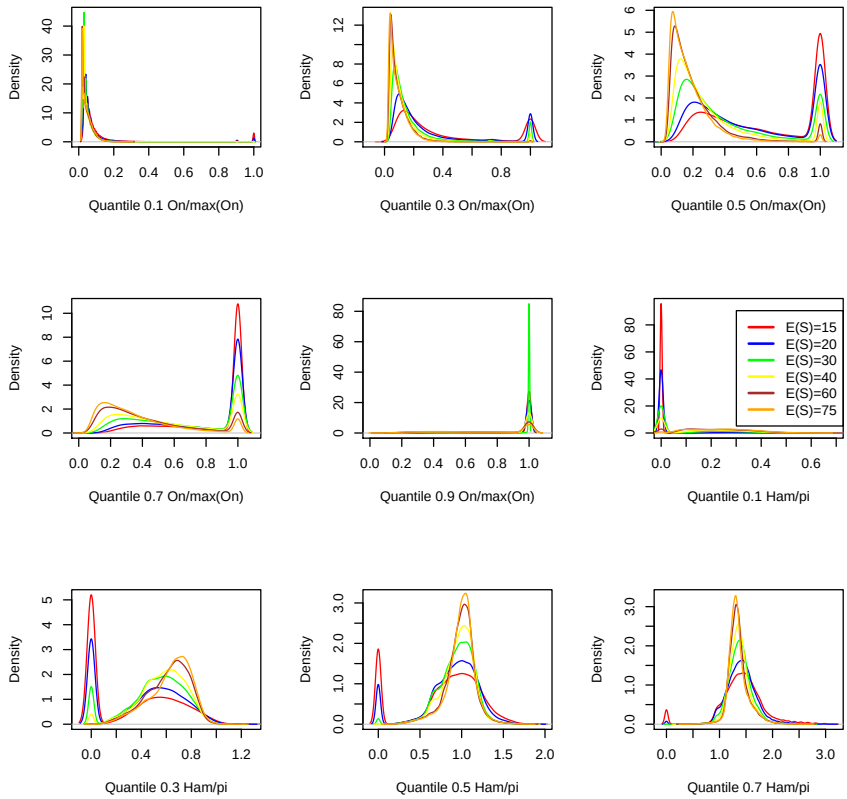


FIGURE A2. Distributions of statistics **nO** and **nHam** in Scenario 2 for $\mathbb{K}$-*exp*

of the sets **nSFS**, **nHam** and **nO** of statistics. Figures A2 to A7 show kernel density estimates of their distributions.

A.5. **Using *a priori* arbitrary structured populations to assess multiple mergers vs. exponential growth in structured populations.**

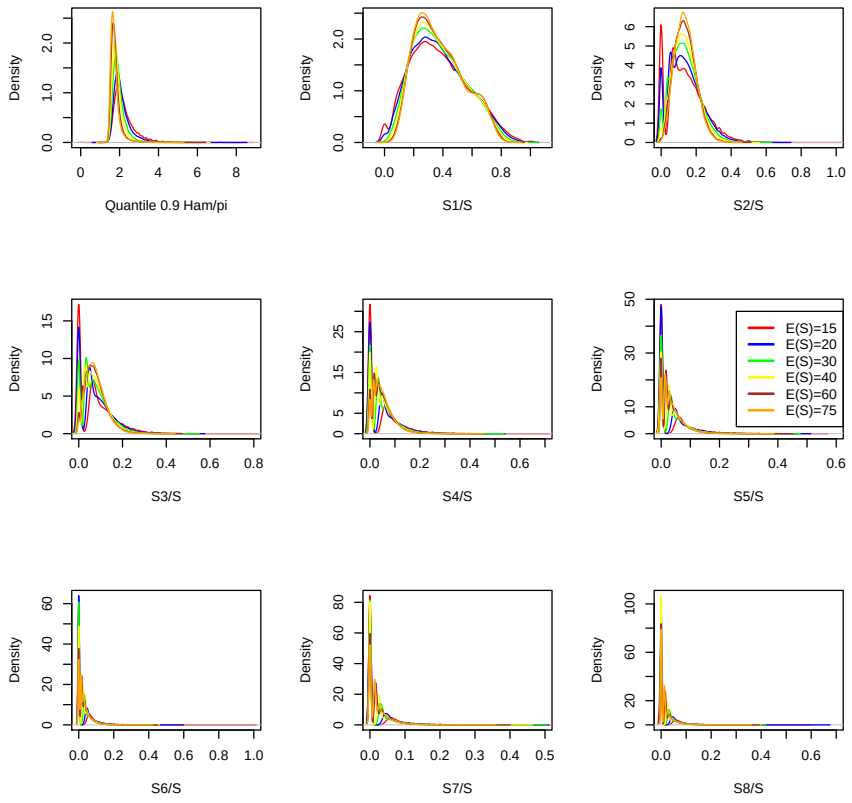


FIGURE A3. Distributions of statistics **nHam** and **nSFS** in Scenario 2 for $\mathbb{K}$-*exp*

Assume Scenario 4 with $n = 100$, but replace the fixed migration rate as well as the fixed subsample sizes with scaled migration rates m picked uniformly from $\{0.1, 0.6, \ldots, 2.1\}$ and subsample sizes picked uniformly from $\{(10, 90), (20, 80), (30, 70), (60, 40), (50, 50)\}$. We used the same summary statistics as in the analysis for Scenario 4 in Section 2.4. We generated 175,000 simulations under each scenario twice. Figure A13 shows the result of the three-fold model selection in this scenario, where we denote the model class with population structure and growth by $2p$-$\mathbb{K}$-*exp*. We then took the random forest trained on the first 175,000 simulations and allocated the set of simulations for model classes $2p.5$-$\mathbb{K}$-*exp* and $2p.9$-$\mathbb{K}$-*exp* using ABC based on the constructed random forest and recorded the proportion of misclassification for all simulations and for the ones allocated with high posterior probability as described in Secyion A.2. We then generated a third set of simulations, where we changed migration rates to be uniformly picked from $\{0.5, 1\}$ and subsample sizes from $(5, 95), (25, 75), (45, 55)$. For

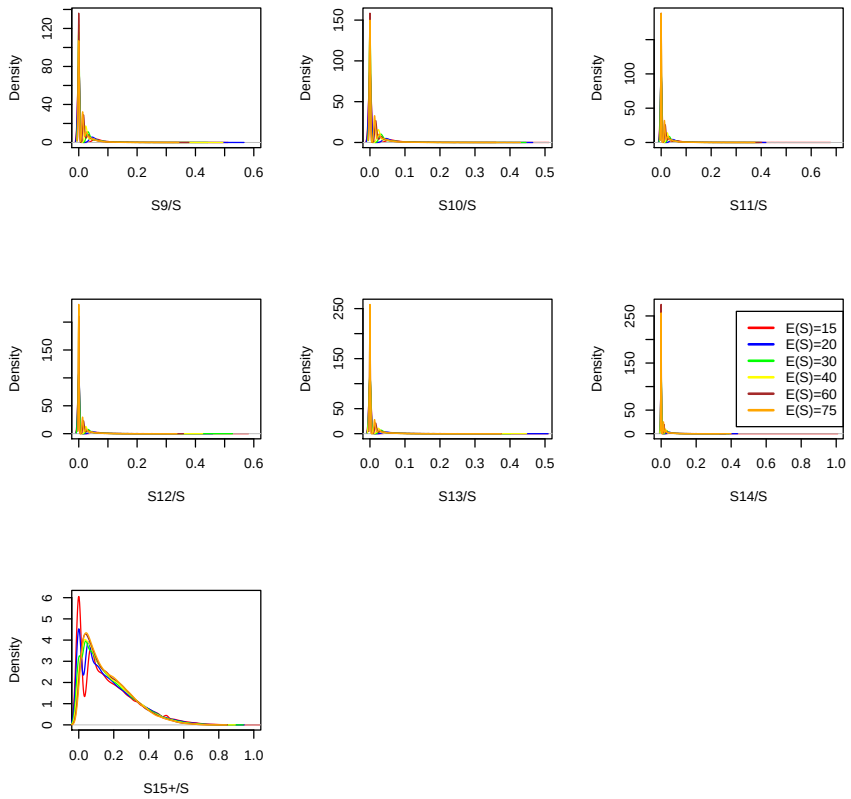


FIGURE A4. Distributions of statistics **nSFS** in Scenario 2 for $\mathbb{K}$-*exp*

these simulations, we denote the model class with population structure and growth by $2p^*$-$\mathbb{K}$-*exp*. We used the random forest constructed above, allocated also these simulations and recorded the errors as for the second set of simulations. Results are summarized in Table A2.

A.6. **Comparison with other ABC approaches.** For Scenario 2 with $n = 100$ and model classes $\mathbb{K}$-*exp*, $\mathbb{B}$ and $\mathbb{D}$, we performed ABC model selection. We used **O**, **Ham**, **Phy**, $\mathbf{r^2}$, **AF**+ as summary statistics with quantiles $(.1, .3, \ldots, .9)$. We compared the ABC with random forest as described in the main manuscript (without bias correction for the variable importances) with three standard ABC approaches as implemented in the R package `abc`: the simple rejection method (REJ), using multinomial logistic regression (MNL) and the neural net approach (NN). For the latter three methods, we used a relatively strict tolerance level of 0.005. All four methods were based on 175,000 simulations per model class as described in the main manuscript.

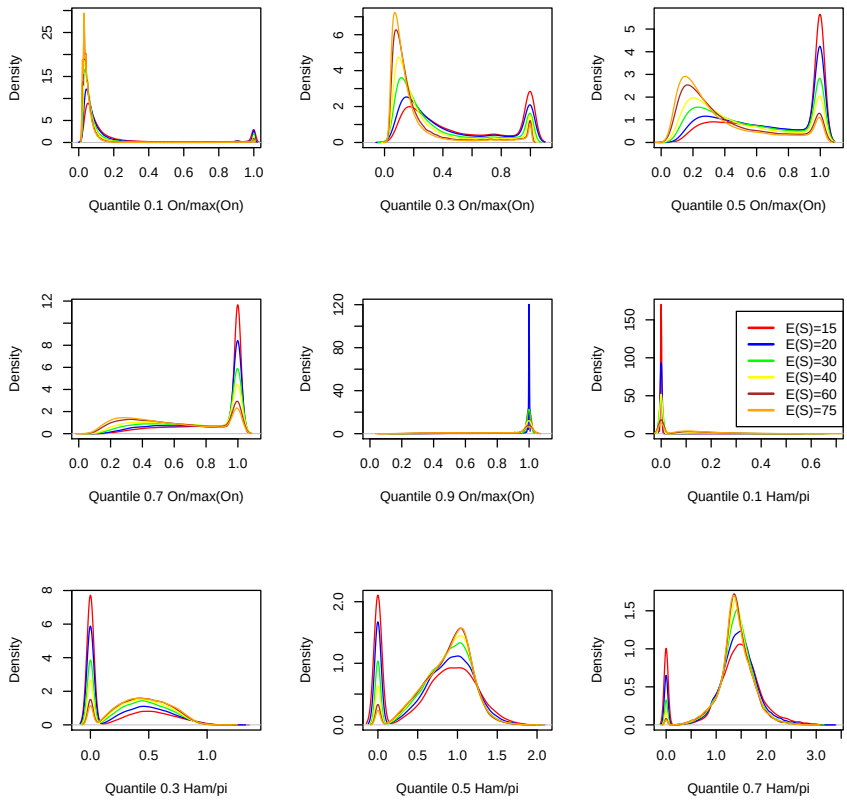


FIGURE A5. Distributions of statistics **nO** and **nHam** in Scenario 2 for $\mathbb{B}$

A	B	C	D	$\rightarrow B$	$\rightarrow C$	$\rightarrow D$	HP
$2p.5$-$\mathbb{K}$-exp	$\mathbb{K}$-exp	$\mathbb{B}$	$2p$-$\mathbb{K}$-exp	0.09 (0)	0.04 (0)	0.87 (1)	0.37
$2p.9$-$\mathbb{K}$-exp	$\mathbb{K}$-exp	$\mathbb{B}$	$2p$-$\mathbb{K}$-exp	0.09 (0)	0.16 (0.09)	0.75 (0.91)	0.26
$2p^*$-$\mathbb{K}$-exp	$\mathbb{K}$-exp	$\mathbb{B}$	$2p$-$\mathbb{K}$-exp	0.08 (0)	0.09 (0.02)	0.83 (0.97)	0.25
$\mathbb{K}$-exp	$\mathbb{K}$-exp	$\mathbb{B}$	$2p$-$\mathbb{K}$-exp	0.7 (0.95)	0.14 (0.03)	0.16 (0.02)	0.06
$\mathbb{B}$	$\mathbb{K}$-exp	$\mathbb{B}$	$2p$-$\mathbb{K}$-exp	0.21 (0.01)	0.71 (0.99)	0.08 (0.01)	0.22

TABLE A2. $\rightarrow B$, $\rightarrow C$: Proportion of simulations from model class A allocated to two other model classes B,C in Scenarios 2 and 4. In parentheses: same proportion only among high posterior allocations. HP: High posterior allocations are those with posterior probability > 0.9, the column 'high posterior' shows their proportion among all simulations from A. Random forest based on summary statistics $\mathbf{r^2}$, **Phy**, **SFS**$*$, **AF**, **O** and **Ham** (quantiles used: $.1, .2, .3, .4. \ldots, .9$). Results are rounded to two digits.

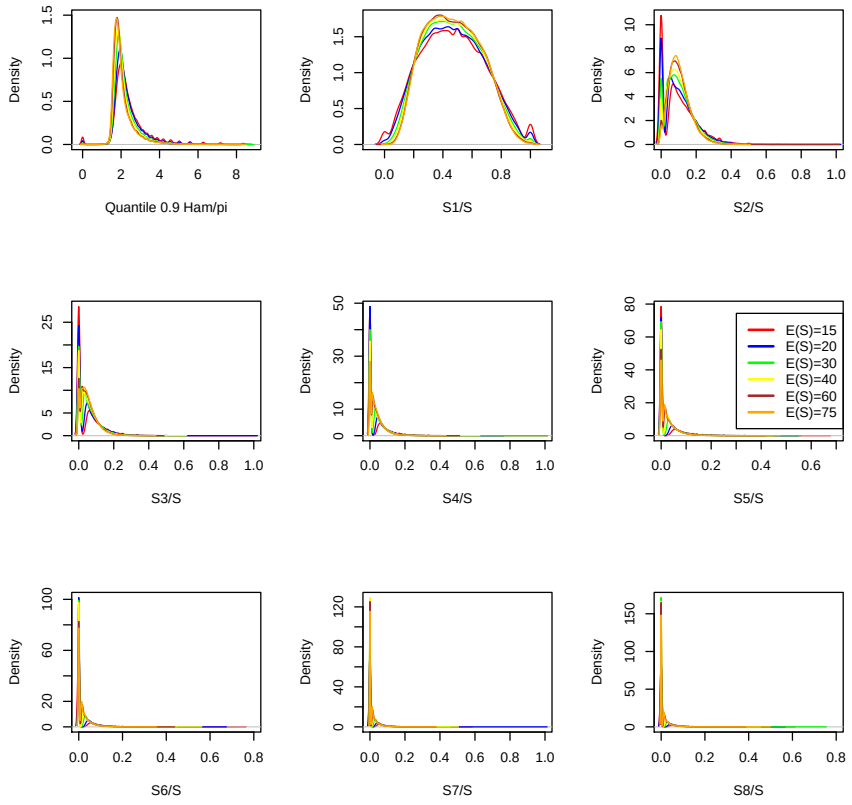


FIGURE A6. Distributions of statistics **nHam** and **nSFS** in Scenario 2 for $\mathbb{B}$

To assess errors, we then performed model selection for two independent test sets each consisting of 2,500 simulations per model class that then were allocated to the three model classes. For the random forest approach, we proceeded as described in Section A.2, for the other methods we set the model selected to the model with the highest posterior probability. Misclassification is reported in Table A3.

CROP PLANT BIODIVERSITY AND BREEDING INFORMATICS GROUP (350B), INSTITUTE OF PLANT BREEDING, SEED SCIENCE AND POPULATION GENETICS, UNIVERSITY OF HOHENHEIM, FRUWIRTHSTRASSE 21, 70599 STUTTGART, GERMANY
Email address: `fabian.freund@uni-hohenheim.de`

DEPARTAMENTO DE PROBABILIDAD Y ESTADÍSTICA, IIMAS, UNAM. APARTADO POSTAL 126 01000 MÉXICO, D.F. MEXICO
Email address: `arno@sigma.iimas.unam.mx`

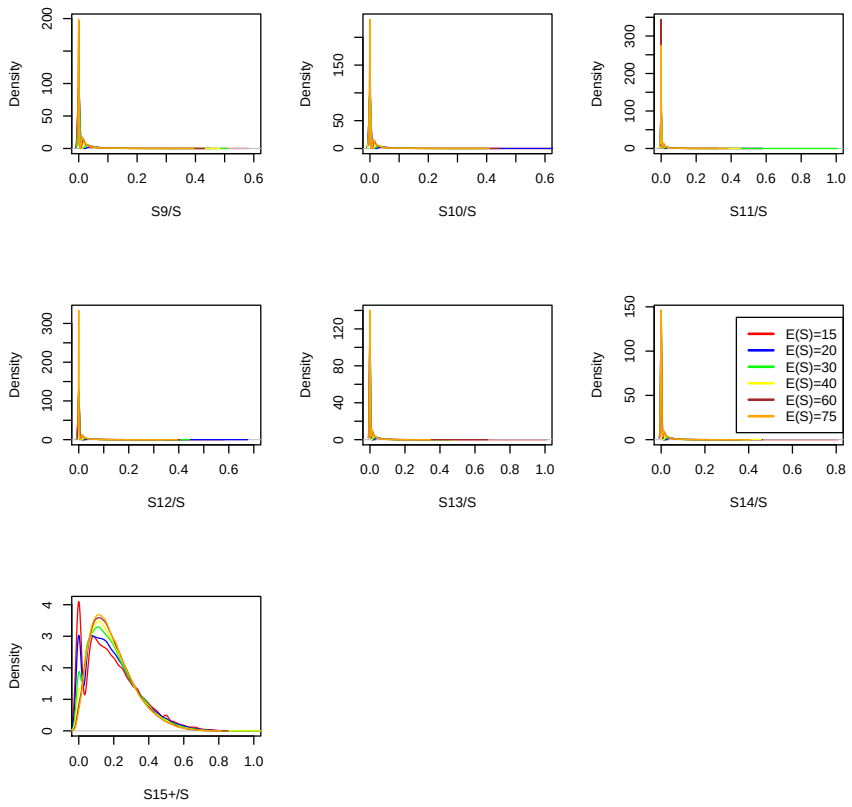
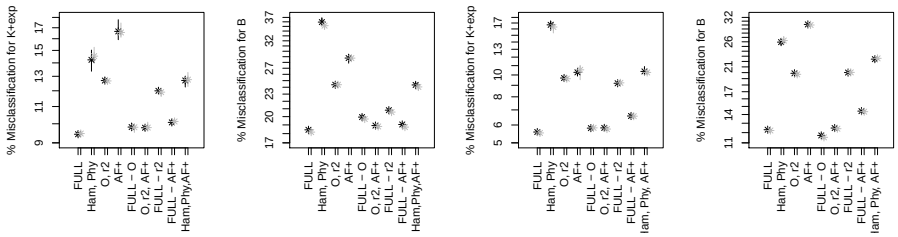
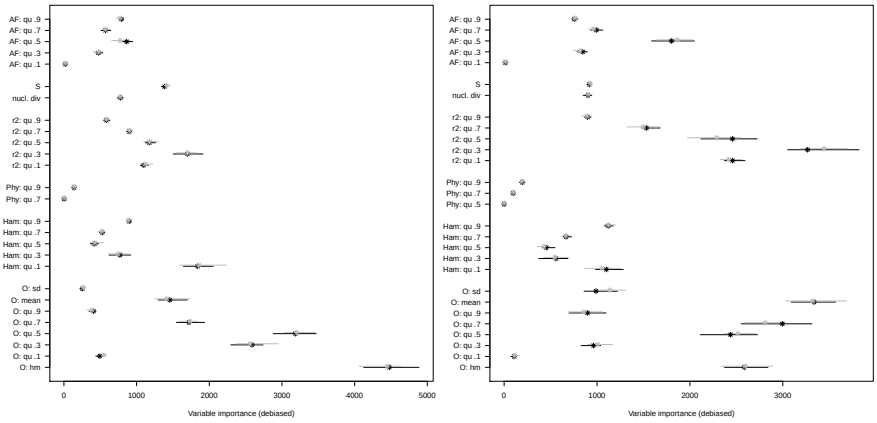


FIGURE A7. Distributions of statistics **nSFS** in Scenario 2 for $\mathbb{B}$



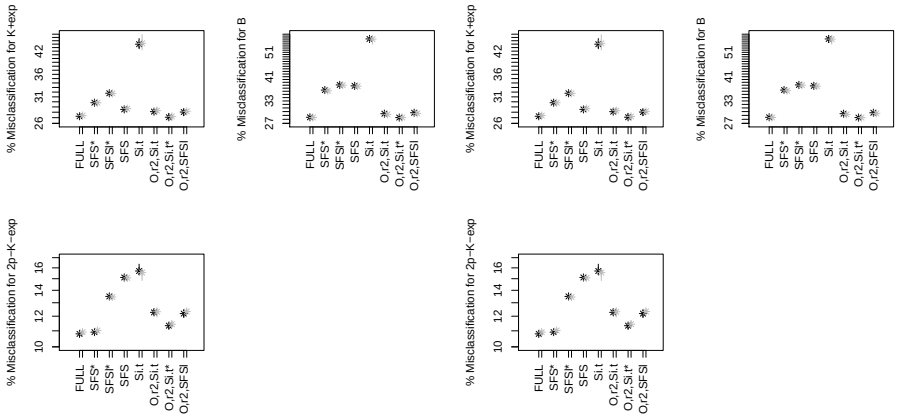
(A) Scenario 2a (B) Scenario 2b

FIGURE A8. Classification errors of model comparison between $\mathbb{K}$-*exp* and $\mathbb{B}$ with $n = 100$ for Scenarios 2a and 2b. See Table 6 for the legend. The sets of statistics used are $\mathbf{O}, \mathbf{r^2}, \mathbf{Ham}, \mathbf{Phy}, \mathbf{AF}+$.



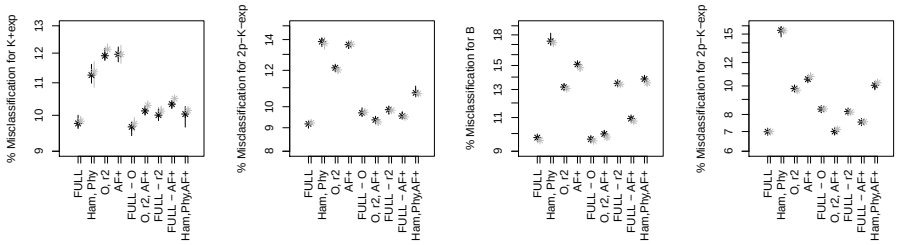
(A) Scenario 2a (B) Scenario 2b

FIGURE A9. Variable importances for the model comparison between $\mathbb{K}$-*exp* and $\mathbb{B}$. See Table 6 for the legend.



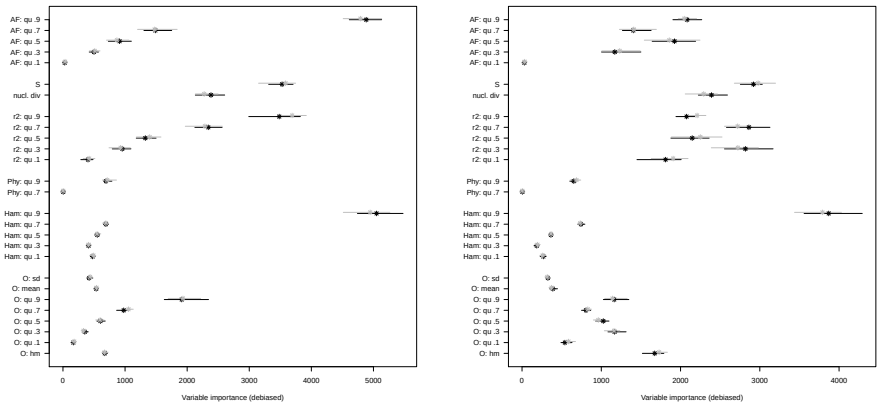
(A) $\mathbb{K}$-*exp* vs. $\mathbb{B}$ vs. $2p.5$-$\mathbb{K}$-*exp* (B) $\mathbb{K}$-*exp* vs. $\mathbb{B}$ vs.$2p.9$-$\mathbb{K}$-*exp*

FIGURE A10. Classification errors in Scenario 4, $n = 100$. See Table 6 for the legend. Full set of statistics includes $\mathbf{r^2}$, **Phy**, **O**, **Ham**, **SFS**, π,S, Tajima's D and Fay and Wu's H.



(A) $\mathbb{K}$-*exp* vs. $2p.9$-$\mathbb{K}$-*exp* (B) $\mathbb{B}$ vs. $2p.9$-$\mathbb{K}$-*exp*

FIGURE A11. Classification errors for model selection in Scenario 4, $n = 100$. See Table 6 for the legend. Full set of statistics includes $\mathbf{r^2}$, **Phy**, **AF**+, **O** and **Ham**.



(A) $\mathbb{K}$-*exp* vs. $2p.9$-$\mathbb{K}$-*exp* (B) $\mathbb{B}$ vs. $2p.9$-$\mathbb{K}$-*exp*

FIGURE A12. Variable importances in Scenario 4, $n = 100$. See Table 6 for the legend.

	REJ	MNL	NN	RF
Miscl. $\mathbb{K}$-*exp* Sc.2	0.15 (0.16)	0.173 (0.186)	0.173 (0.196)	0.157 (0.176)
Miscl. $\mathbb{B}$ Sc.2	0.551 (0.554)	0.295 (0.296)	0.31 (0.304)	0.312 (0.312)
Miscl. $\mathbb{D}$ Sc. 2	0.222 (0.22)	0.127 (0.124)	0.114 (0.116)	0.116 (0.117)

TABLE A3. Misclassification frequencies of 2,500 simulations/model class (once replicated, replication result in brackets). ABCs based on MNL=multinomial logistic regression, NN=Neural net, REJ = rejection, RF = random forests.

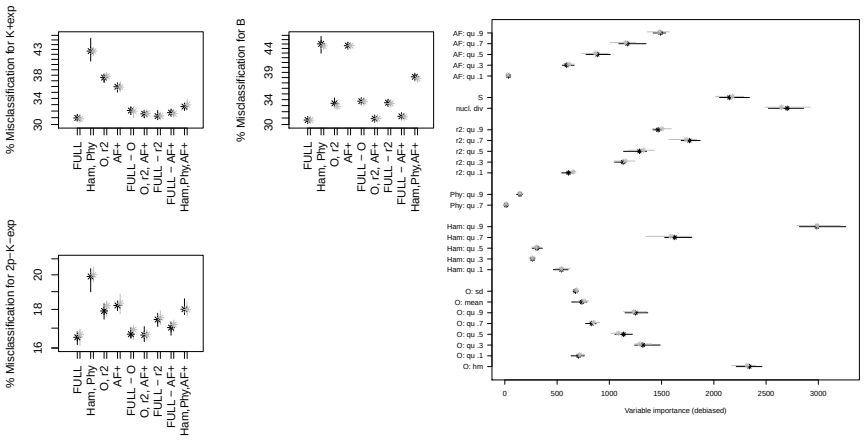


FIGURE A13. Classification errors and variable importances for model selection $\mathbb{K}$-*exp* vs. $\mathbb{B}$ vs. $2p$-$\mathbb{K}$-*exp* (variable migration rates and subsample sizes) in Scenario 4 ($n = 100$). See Table 6 for the legend. Full set of statistics includes $\mathbf{r^2}$, **Phy**, **AF+**, **O** and **Ham**.

	as $\mathbb{K}$-*exp*	as $\mathbb{B}$	as $\mathbb{D}$	as $\mathbb{D}$-*exp*
from $\mathbb{K}$-*exp*	0.838	0.159	0.003	0.001
from $\mathbb{B}$	0.253	0.674	0.066	0.007
from $\mathbb{D}$	0.011	0.110	0.757	0.122
from $\mathbb{D}$-*exp*	0.001	0.015	0.164	0.820

TABLE A4. Complete mean misidentification matrix (classification errors of model class A as model class B) of model comparison $\mathbb{K}$-*exp* vs. $\mathbb{B}$ vs. $\mathbb{D}$ vs. $\mathbb{D}$-*exp* in Scenario 2 with $n = 100$. Result shown for the full set of statistics $\mathbf{r^2}$, **Phy**, **Ham**, **O** and **AF+**. Results averaged over 2x 10 ABC runs from each of two simulation sets and rounded to three digits.