Multiple merger genealogies in outbreaks of *Mycobacterium tuberculosis*

F. Menardo[1,2*], S. Gagneux[1,2] and F. Freund[3*]

[1] Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Basel, Switzerland

[2] University of Basel, Basel, Switzerland

[3] Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany

* Equal contribution, corresponding authors

Abstract

The Kingman coalescent and its developments are often considered among the most important advances in population genetics of the last decades. Demographic inference based on the coalescent theory has been used to reconstruct the population dynamics and evolutionary history of several species, including *Mycobacterium tuberculosis* (MTB), an important human pathogen causing tuberculosis. One key assumption of Kingman's coalescent is that the number of descendants of different individuals does not vary strongly, and violating this assumption could lead to severe biases caused by model misspecification. Individual lineages of MTB are expected to vary strongly in reproductive success because 1) MTB is potentially under constant selection due to the pressure of the host immune system, 2) MTB undergoes repeated population bottlenecks when it transmits from one host to another, and 3) some hosts show much higher transmission rates compared to the average ("super-spreaders").

Here we used an Approximate Bayesian Computation approach to test whether multiple merger coalescents (MMC), a class of models that allow for large variation in offspring sizes, are more adequate models to study MTB populations. We considered eleven publicly available whole genome sequence data sets sampled from MTB local populations and outbreaks and found that MMC had a better fit compared to the Kingman coalescent for nine of the eleven data sets. These results indicate that the neutral model for analyzing MTB outbreaks, and potentially the outbreaks of other pathogens, should be reassessed, and that past findings based on the Kingman coalescent need to be revisited.

Keywords

Mycobacterium tuberculosis, demographic inference, multiple merger coalescent, Approximate Bayesian Computation, Random Forest.

Introduction

The coalescent is a stochastic mathematical model that formally describes the shapes of the expected genealogies in a population (Kingman 1982). The original formulation of Kingman has been extended to include different evolutionary processes such as fluctuations in population size (Griffith and Tavare 1994), population subdivision and migration (Wilkinson-Herbots 1998), recombination (Hudson 1983), and selection (Kaplan et al. 1988, Neuhauser and Krone 1997). Although the genealogy of a sample is typically unknown, mutational models can be superimposed onto the coalescent to describe DNA sequence polymorphisms. These are generally easy to obtain from natural populations, thus opening the possibility of data-based statistical inference. Applications of the coalescent include the study of the evolutionary histories and population dynamics of a variety of taxa (Kuhner 2009), including humans (Li and Durban 2001, Excoffier et al. 2013) and pathogens (Pybus et al. 2001, Joy et al. 2003), and the identification of genetic loci under selection (Biswas and Akey 2006, Hernandez et al. 2011).

One of the assumptions of the Kingman coalescent is that the variance in the number of offspring of each individual is small, such as at most one pair of sampled lineages can find a common ancestor for any time interval corresponding to a single time point on the coalescent time scale (short evolutionary time interval, SETI). Therefore, under the Kingman coalescent, the genealogies are strictly bifurcating. While this is a realistic assumption for many species, it has been shown that organisms with a life cycle characterized by high fecundity and high early mortality (sweepstake reproduction) have a very large variance in the offspring size (Eldon and Wakeley 2006, Sargsyan and Wakeley 2008). Additionally, in populations under constant selective pressure, the fittest individuals are expected to produce many more descendants compared to the less fit, thus also resulting in a skewed offspring distribution (Neher and Hallatschek 2012).

A more general class of models, of which the Kingman coalescent is a special case, has been developed to accommodate large offspring variation between individuals. These models are known as multiple merger coalescents (MMC), because unlike Kingman's coalescent, they allow more than two lineages to coalesce during a SETI, thus resulting in multifurcating genealogies (Tellier and Lemaire 2014). MMC have been proposed to be more adequate models to investigate marine organisms with sweepstakes reproduction (Sargsyan an Wakeley 2008), agricultural pathogens with recurrent seasonal bottlenecks (Tellier and Lemaire 2014), loci under positive selection (Durrett and Schweinsberg 2005), and rapidly adapting pathogens (Neher and Hallatschek 2012).

Despite a growing interest in MMC, there are few studies that used genetic polymorphisms to test whether MMC are indeed a better fitting model compared to the Kingman coalescent. Signatures of MMC have been detected at the creatin kinase muscle type A locus of the Atlantic cod (*Gadus morhua*; Árnason and Halldórsdóttir 2015), in the mitochondrial genome of Japanese sardines

(*Sardinops melanostictus*; Niwa et al. 2016), in populations of breast cancer cells (Kato et al. 2017), and in the B-cell repertoire response to viruses such as HIV-1 and influenza (Nourmmohammad et al. 2019, Horns et al. 2019). While MMC are theoretically appealing genealogy models for pathogen samples (Irwin et al. 2016, Rocha 2018, Neher and Walczak 2018), their fit to observed data in pathogens populations has not been investigated so far. Only very recently, MMC have been used to study the within-host genetic diversity of *Mycobacterium tuberculosis* (MTB), a major human pathogen causing tuberculosis (Morales-Arce at al. 2019).

Here we look for evidence of MMC in between-host populations of *Mycobacterium tuberculosis.* Between-host populations of MTB are expected to have a skewed offspring distribution because of three reasons: 1) MTB is an obligate pathogen, and therefore potentially constantly adapting under the pressure of the immune system (Gagneux 2018). 2) Super-spreaders; these are patients responsible for a very large number of secondary infections compared to the average (Gardy et al. 2011, Walker et al. 2013, Ypma et al. 2013, Stucki et al. 2015, Lee et al. 2019), thus causing a large variance of the pathogen's offspring size. 3) MTB undergoes repeated bottlenecks when transmitting from one host to another, with a few bacteria, and potentially as few as one, producing the entire population infecting the new host (Lin et al. 2014).

Additionally, a low genetic diversity and an excess of rare variants (singletons) have been reported in MTB (Hershberg et al. 2008, Pepperel et al. 2013), and both are known signatures of MMC genealogies (Tellier and Lemaire 2014).

Methods based on the Kingman coalescent are often used in population genetic analyses of MTB. For example: 1) The Bayesian Skyline Plot (Drummond et al. 2005) has been used to infer past population dynamics in tuberculosis outbreaks, finding evidence for constant population size (Bainomugisa et al. 2018), rapid population growth (Eldholm et al. 2015, Folkvardsen et al. 2017) or slow population decline (Lee et al. 2015). 2) Different methods have been used to infer the demographic history of the global MTB population (Pepperell et al. 2013, Comas et al. 2013, Bos et al. 2014) and of single MTB lineages (Kay et al. 2015, Luo et al. 2015, Merker et al. 2015, Merker et al. 2018, Liu et al. 2018, O'Neill et al. 2019), finding evidence for population growth or for complex fluctuations that have been correlated with major events in human history such as the introduction of antibiotic treatment. 3) The strength of purifying selection was estimated with a simulation based approach, finding a genome-wide selection coefficient several order of magnitude higher compared to other prokaryotes and eukaryotes (Pepperell et al. 2013).

While some of these results might be biased by unaccounted population structure (Heller et al. 2013) or sampling biases (Lapierre et al. 2016), potentially they are all impacted by the violation of the Kingman's assumption described above, and their conclusions could be affected by model misspecification (Tellier and Lemaire 2014).

Given the undergoing efforts in controlling and stopping the spread of tuberculosis, and the global impact of this pathogen that causes more than 1.4 million deaths each year (WHO 2019), it is important to evaluate the adequacy of the population genetic models used to study tuberculosis epidemics. To this end, we considered eleven MTB whole genome sequence (WGS) data sets, we used an approximate Bayesian computation (ABC) approach based on simulations to find the best fitting model among Kingman's coalescent, with and without accounting for population sizes changes, and two MMC models, the Beta coalescent (Schweinsberg 2003) and the Dirac coalescent (Eldon and Wakeley 2006). We found that MMC were the best fitting model for nine of the eleven data sets (eight fitted best to the Beta, one to the Dirac coalescent). Our results indicate that the shape of the genealogies of MTB samples, and therefore their sequences and genetic diversity, are influenced by the skewed offspring distribution of MTB. Consequently, demographic inference based on models assuming non-skewed offspring distribution (i.e. Kingman's coalescent) could lead to inaccurate results when applied to MTB epidemics, and potentially to the epidemics of other pathogens with similar life histories.

Results

Models and data sets

MTB is thought to be strictly clonal, lateral gene flow is completely absent, or very rare (Hershberg et al. 2008, Gagneux 2018, Chiner-Oms et al. 2019). Therefore, the MTB genome can be considered as a single genetic locus, and one single genealogy describes the relationships among all MTB strains in any data set. The shape of the genealogy of a sample is influenced by many factors, such as the underlying offspring distribution, sampling scheme, population subdivision, geographic population structure, migration and changes in population size. To avoid these confounding effects, we considered only populations that were unlikely to be affected by population structure, sampling biases, population subdivision and migration. We searched the literature for WGS data sets of MTB where all strains were sampled from a single phylogenetic clade that is restricted to a particular geographic region, and identified eleven studies. Most of these data sets represent single outbreaks (Methods). For each data set, we downloaded the raw Illumina sequences (Sup. Table 1) and used a bioinformatic pipeline described in the Methods to identify high confidence SNPs (Table 1). Excluding population structure, two factors that can shape the diversity of these data sets are changes in population size, and whether effective offspring distributions are skewed. We modeled changes in population size assuming exponential population growth, as has often been done in previous studies (Eldholm et al. 2015, Merker et al. 2015, Eldholm et al. 2016, O'Neill et al. 2019). We modeled skewed offspring distributions with two MMC models: 1) the Beta coalescent, in which the probability of each individual to coalesce in a multiple merger event is regulated by a

Beta distribution with parameters α (between 1 and 2) and 2- α. The Beta coalescent explicitly models populations with skewed offspring distributions (Schweinsberg 2003). Additionally, it was also proposed to capture the genealogies of populations undergoing recurrent bottlenecks and of epidemics characterized by super-spreaders (Tellier and Lemaire 2014, Hoscheit and Pybus 2019). Lower values of α (closer to one) correspond to larger multiple mergers events, and for α=1 the Beta coalescent corresponds to the Bolthausen-Sznitman (BSZ) coalescent. The BSZ coalescent is an explicit model for genealogies of populations evolving under rapid selection, which lead certain families of selected genotypes to have strongly increased sizes compared to the average (Bolthausen and Sznitman 1998, Neher and Hallatschek 2012, Desai et al. 2013).

2) The Dirac coalescent, also known as psi coalescent, is defined by a single parameter (ψ). The parameter ψ represents the average proportion of sampled lineages that coalesce in a single multiple merger event, and was also proposed for populations with skewed offspring distributions (Eldon and Wakeley 2006).

Our goal is to test whether modeling skewed offspring distributions alone explained the observed genetic diversity better than modeling variable population sizes (with an exponential growth model) and standard offspring distributions. Therefore, we considered MMC models with constant population sizes. It was previously shown that even for a single locus, these hypotheses can be distinguished for moderate sample sizes and high enough mutation rates (Eldon et al. 2015, Freund and Siri-Jégousse 2019)

Table 1. Data sets used in this study

Data set[1]	Number of strains	Number of polymorphic positions	Locality of sampling
Eldholm 2015	248	497	Buenos Aires (Argentina)
Lee 2015	147	454	Nunavit (Canada)
Stucki 2016	175	6264	Central African countries
Shitikov 2017	176	1164	Russia and Belarus
Roetzer 2013	61	74	Hamburg (Germany)
Comas 2015	21	1334	Ethiopia
Bainomugisa 2018	81	401	Daru Island (PNG)
Bjorn-Mortensen 2016	121	128	East Greenland
Folkvardsen 2017	702	214	Copenhagen (Denmark)
Stucki 2014	60	128	Bern (Switzerland)
Eldholm 2016	25	17	Oslo (Norway)

[1] We identified the data sets with the first author's name and year of the original publication

Model selection and parameter estimation with Approximate Bayesian Computation

For model selection and parameter estimation, we used an ABC approach based on random forests (RF), as reported in detail in the Methods section and represented in Figure 1. We considered four models, Kingman's coalescent with constant population size (KM), Kingman's coalescent with exponential population growth (KM+exp), Beta coalescent with constant population size (BETA), and Dirac coalescent with constant population size (Dirac). Briefly, for each data set, we collected the SNPs identified with the bioinformatic analysis, reconstructed the genotype of the most recent ancestor (MRCA) and used it to polarize the SNPs. We then calculated a set of 20 summary statistics measuring genetic diversity, linkage disequilibrium and phylogenetic properties. The only exception was the data set Stucki 2016, where for computational reasons we omitted the statistics measuring linkage disequilibrium (see Methods for details). For each model, we performed 125,000 simulations of a sample of size n, where n is the number of individuals in the data set, drawing the scaled population size from a prior distribution spanning one order of magnitude around the Watterson estimator (θ_{obs}).

As described in Pudlo et al. (2015), we performed model selection via ABC using a random forest of 1,000 decision trees. For parameter estimation within a model class, we followed the approach of Raynal et al. (2018). To control for stochastic effects, we repeated model selection and parameter estimation three times for each data set. We found consistent results across replicates, and we report

here the consensus of the three analyses (see Methods for details), the results of all individual analyses are available in Sup. Table 2.

We found that for most data sets, the ABC approach had overall good discriminatory power, with out-of-bag (OOB) error rates (the misclassification probabilities) ranging from 1.9% to 12.2% (Table 2). The only exception was the data set Eldholm 2016 (OOB error rate = 29.8%.), which was the data set with the lowest genetic diversity. Most importantly for our study, the probability that data generated under a model with standard offspring distribution (KM and KM+exp) was misclassified as multiple merger was low (0.7% - 5.5%), again the only exception was the data set Eldholm 2016 (16.6%).

We found that BETA was the best fitting model for eight of the eleven data sets, KM+exp was the best model for two data sets, and Dirac was the best model for one data set (Table 2). For all data sets, the posterior probability of the selected model was higher than 80% and therefore more than four times more likely than the second best fitting model.

One potential problem when performing model selection, is that none of the considered models captures the observed data. To exclude this possibility, we performed posterior predictive checks, in which for each data set, we simulated data under the best fitting model using the median of the posterior distribution of the relative parameter (averaged over the three replicates). We then compared the observed data with the simulated data. If the selected model fits the data well, we expect the simulated and observed data to be similar. Conversely, if the selected model does not fit the data well, we expect simulated and observed data to be different. We found that for all but one data set, the observed values of 20 summary statistics was within the range of values obtained from the simulations, indicating that the best model can reproduce the observed data (Sup. Figs. 1-11). The only exception was the data set Shitikov 2017, for which the mean and the standard deviation of the minimal observable clade size statistic were not overlapping with the simulated values (Sup. Fig. 11). This indicates that the best fitting model (KM+exp) cannot reproduce the observed data, and that none of the considered models is adequate for this data set.

Table 2. Results of model selection and parameter estimation

Data set	Selected Model	OOB error rate (misclassification % as MMC)[1]	Posterior Probability	Median and 95% posterior credibility interval of coalescent parameters[2]
Eldholm 2015	BETA	2.4% (1.5%.)	98.0%	α: 1.167 (1 – 1.4)
Lee 2015	BETA	4.3% (2.0%)	96.6%	α: 1.3 (1.075 – 1.5)
Stucki 2016	KM+exp	3.4% (1.1%)	99.9%	g: 2325 (815 - 6795)
Shitikov 2017	KM+exp	3.8% (1.6%)	99.4%	g: 2953 (1198 – 4888)
Roetzer 2013	BETA	12.2% (5.5%)	93.4%	α: 1.217 (1 – 1.8)
Comas 2015	BETA	7.3% (0.7%)	87.8%	α: 1.517 (1.025 – 1.95)
Bainomugisa 2018	BETA	7.2% (3.1%)	97.9%	α: 1.175 (1 – 1.5)
Bjorn-Mortensen 2016	BETA	7.1% (3.3%)	86.1%	α: 1.042 (1 – 1.25)
Folkvardsen 2017	BETA	1.9% (0.9%)	98.9%	α: 1.142 (1.-1.325)
Stucki 2015	BETA	10.5% (4.8%)	85.3%	α: 1.05 (1 – 1.3)
Eldholm 2016	Dirac	29.8% (16.6%)	80.8%	ψ: 0.350 (0.1 – 0.675)

[1] The out-of-bag error rate is the probability that a simulation is misclassified as coming from any other model class, between parentheses we report the probability that a simulation generated with KM or KM+exp is miss-classified as a MMC (BETA or Dirac).

[2] The interval between the 0.025 quantile and the 0.975 quantile of the parameter of the selected model (g for KM+exp, ψ for Dirac and α for BETA). The growth rate g is reported as used in Hudson's ms (for diploid genealogies), thus all growth estimates have to be halved to be interpreted for MTB.

Hidden population structure and population decline in the data set Lee 2015

In our analysis, we focused on local data sets to control for the confounding effect of complex population dynamics and population structure. However, in one case (Lee 2015), it is possible that some degree of population structure is still present. Lee 2015 is a data set sampled from an epidemic in Inuit villages in Nunavik, Quebec, Canada (Lee et al. 2015). Lee et al. (2015) showed that transmission of MTB among patients was more frequent within a village than between villages, and that related strains tended to be present in the same village. This was supported by the reconstructed phylogenetic tree, which showed three clades separating at the root that could represent well separated sub-populations (Fig. 2; see also Fig. 2 in Lee et al. 2015). These data suggest the existence of some degree of geographic population structure, therefore we tested whether this might influence the results of our model selection. To do this, we ran two analyses: 1) we repeated the ABC-RF analysis on three subsets of Lee 2015, which represent the three main

clades described above (Fig. 2). Under the assumptions that the separate branches of the phylogeny reflect different sub-populations, and that migration does not alter the coalescent rates within the subpopulations, the genealogy of each sub-clade should then follow one of the coalescent models that we are fitting. We found that BETA was the best fitting model for two of the sub-clades, while Dirac was the best fitting model for the third (Table 3). The posterior predictive checks showed that the best model could reproduce the data of these three subsets (Sup. Figs. 12-14). However, the posterior probabilities were low compared to the complete data set, and the misclassification probabilities were larger. This was probably due to the smaller sample size of the individual subsets compared to the full data set (Table 3). 2) We performed an additional model selection analysis between three competing models, BETA, Dirac and a third scenario, in which we modeled a structured population with migration and with standard offspring distribution and exponential growth (KM+exp). Also in this case, BETA resulted to be the most likely model (Table 3, see Methods for details). Overall, our findings indicate that it is unlikely that the MMC signal in the Nunavik MTB population is an artifact caused by population structure.

Structured populations have similar genealogies to populations that are shrinking in size, with many lineages coalescing close to the root. In their original publication, Lee et al. (2015) used the Bayesian Skyline Plot (Drummond et al. 2005) to reconstruct the fluctuations in population size of the Nunavik population, and found evidence for a slow population decline. Here, we are not interested in whether the inferred population decline is genuine or caused by unaccounted population structure, we only want to assess whether a decline in population size could bias our analysis. To do this, we repeated the ABC-RF model selection among two models: BETA and KM with population decline (see Methods for details). Again, we found that BETA was the best fitting model (Table 3), thus indicating that our results for this data set are unlikely to be an artifact caused by population decline.

Table 3. Results of model selection for the complete Lee 2015 data set, and for the three major sub-clades separately. The shaded row represents the results of the standard analysis on the full data set.

Data set	N. of strains	Selected Model	OOB error rate (misclassification % as MMC)[1]	Posterior Probability	Second best fitting model
Lee 2015	147	BETA	4.3% (2.0%)	96.6%	Dirac
Lee 2015 Clade A	61	Dirac	18.0% (8.7%)	74.8%	BETA
Lee 2015 Clade B	36	BETA	14.3% (6.6%)	75.2%	KM
Lee 2015 Clade C	49	BETA	10.8% (4.7%)	62.5%	Dirac
Lee 2015 Pop. Structure[2]	147	BETA	5.8% (4.6%)	96.3%	Dirac
Lee 2015 Pop. Decline[3]	147	BETA	3.2 % (2.7 %)	100%	Pop. decline

[1] The out-of-bag error rate is the probability that a simulation is misclassified as coming from any other model class, between parentheses we report the probability that a simulation generated with KM or KM+exp is missclassified as MMC (BETA or Dirac)

[2] Model selection among BETA, Dirac, and KM with structure

[3] Model selection among BETA and KM with population decline. For computational reasons we used a reduced set of statistics compared to the other data sets (see Methods)

Serial sampling

One limitation of our analysis is that it assumes that all samples are collected at the same time (synchronous sampling). Generally, MTB strains are sampled from the sputum of patients, which is collected when they first present for diagnosis. All data sets that resulted in a MMC as best fitting model included samples obtained over extended periods of time (serial sampling), corresponding to between ~ 8% and ~100% of the estimated tree age (Sup. Table 3).

We investigated whether, at least in principle, the violation of the assumption of synchronous sampling could bias the results of the ABC analysis performed above, and whether the better fit of MMC could be an artifact due to such violation. To do this, we ran simulations assuming serial sampling, followed by model selection on the simulated data assuming synchronous sampling (see Methods). Since this analysis depends on assumptions about the sample size, the genetic diversity, and the sampling times, we used the settings (sample size, observed generalized Watterson's estimator as scaled mutation rate, and the real years of isolation) of three of the observed data sets, which differed in these characteristics (Eldholm 2015, Lee 2015 and Roetzer 2013).

We found that data simulated under KM+exp can be misclassified as BETA or Dirac if we do not account for serial sampling. Specifically, this was true for extended sampling periods compared to

the expected height of the genealogy (on the coalescent time scale), and for low growth rates (Fig. 3). Similarly to model selection, not accounting for serial sampling affected the estimation of the growth rate parameter, and this effect was greater for large sampling periods and low growth rates (Sup. Fig. 15).

It is difficult to relate these results to the observed data sets because we do not know the scaling factor between coalescent time and real time (and therefore cannot estimate the value of c in Fig. 3, see Methods). However, for seven of the nine data sets that resulted in a MMC as best fitting model, we estimated large growth rates ($g \geq 1{,}000$) under the KM+exp model (Sup. Table 2), indicating that serial sampling is unlikely to affect the results of model selection in these cases (Fig. 3, Sup. Fig. 15).

Nevertheless, we adopted a complementary approach, in which we virtually eliminated serial sampling by sub-sampling only strains that were isolated in a single year. Since small data sets have lower discriminatory power, for this analysis, we selected the four data sets with the highest genetic diversity (data sets with more than 200 polymorphic positions: Eldholm 2015, Lee 2015, Folkvardsen 2017 and Bainomugisa 2018) among the ones for which the sampling times were available. For each data set, we repeated the ABC analysis on the largest possible subset of strains that were sampled in a single year (Sup. Table 1, Table 4). We found that all subsets had lower posterior probabilities and higher misclassification errors compared to the full data sets, most likely because of the smaller sample size (Table 4). BETA was the best fitting model for two subsets, Dirac and KM+exp were the best fitting model for one subset each. For the data sets Eldholm 2015 and Folkvardsen 2017 the second and third most sampled years had a similar number of strains compared to the most sampled year. Therefore, we extended the analysis to these additional four subsets, which all resulted in BETA as the best fitting model (Table 4). We performed posterior predictive checks for all subsets and found that in all cases but one, the best fitting model could reproduce the observed data (Sup. Figs. 16-23). The single exception was the subset of Lee 2015 for which two observed quantiles of the r^2 statistic were not overlapping with the values obtained from the simulations. In this subset, all strains but one belonged to one of the three clades discussed above (clade A; Sup. Table 1). We suspected that this analysis was influenced by population structure and we repeated it excluding the single strain not belonging to clade A. Again we found that Dirac was the best fitting model (Table 4, Sup. Table 2), and this time the posterior predictive check could reproduce the data (Sup. Fig. 24).

Overall, these findings indicate that not accounting for serial sampling can indeed bias the results of model selection in favor of MMC models. However, this was unlikely to affects data sets with large growth rates (seven out of nine). Additionally, eight of the nine subsets in which we minimized the serial sampling to one single year resulted in a MMC as best fitting model.

Table 4. Results of model selection for the temporal subsets. Shaded rows contain the results for the full data sets.

Data set[1]	N. of strains	Selected Model	OOB error rate (misclassification % as MMC)[2]	Posterior Probability	Second Best fitting model
Eldholm 2015	248	BETA	2.4% (1.5%)	98.0%	KM+exp
Eldholm 2015 (1998)	34	BETA	17.1% (8.3%)	78.9%	KM+exp
Eldholm 2015 (2001)	31	BETA	15.5% (7.1%)	79.5%	Dirac
Eldholm 2015 (2003)	32	BETA	15.4% (7.1%)	93.7%	Dirac/KM+exp[3]
Lee 2015	147	BETA	4.3% (2.0%)	96.6%	Dirac
Lee 2015 (2012)	45	Dirac	13.2% (6.1%)	66.5%	BETA
Lee 2015 (2012) Clade A	44	Dirac	21.1% (10.5%)	91.3%	BETA
Bainomugisa 2018	81	BETA	7.2% (3.1%)	97.9%	KM+exp
Bainomugisa 2018 (2014)	56	BETA	9.5% (4.1%)	94.9%	KM+exp
Folkvardsen 2017	702	BETA	1.9% (0.9%)	98.9%	KM +exp
Folkvardsen 2017 (2009)	53	BETA	10.8% (4.8%)	92.6%	KM+exp
Folkvardsen 2017 (2010)	64	KM+exp	10.2% (4.7%)	84.5%	BETA
Folkvardsen 2017 (2012)	52	BETA	11.7% (5.2%)	91.0%	KM

[1] Between parentheses we report the year in which the strains were sampled (only for temporal subsets)

[2] The out-of-bag error rate is the probability that a simulation is misclassified as coming from any other model class, between parenthesis we report the probability that a simulation generated with KM or KM+exp is miss-classified as MMC (BETA or Dirac)

[3] Two replications resulted in Dirac as second best fitting model, one in KM+exp

Discussion

The main goal of this study was to test whether MMC models are more adequate than the Kingman coalescent to study MTB local populations and outbreaks, and whether assuming non-skewed offspring distribution (Kingman) as null model could lead to biased results.

For nine of the eleven full data sets, we found a better fit if the genetic diversity was described by a MMC as genealogy model, compared to the Kingman coalescent with exponential population growth. Additionally, the posterior predictive checks showed that the best fitting models captured the genetic diversity in the data sufficiently well (Sup. Figs. 1-11).

Our results are robust towards two possible confounders: population structure and serial sampling. To avoid the effect of population structure, we chose data sets from single outbreaks and local populations in restricted geographic regions. For one data set where a prior analysis suggested some

degree of population structure (Lee et al. 2015), we found that including a model with population subdivision and migration, or sub-sampling the potential sub-populations resulted again in a MMC as the best fitting model, indicating that population structure is unlikely to bias the results of this analysis

Serial sampling could also affect the results of model selection. All the considered models assume a common sampling time for all strains, which is almost never the case for MTB data sets. Due to the relatively short generation time of MTB compared to the sampling period, this may likely correspond to serial sampling on the coalescent time scale. Under serial sampling on the coalescent time scale, our simulations of several scenarios mimicking three of our data sets revealed that Kingman's genealogies under exponential growth can be misidentified as MMC, while inferring a true MMC is not affected. The misclassification probability was higher when the sampling window spanned a large part of the genealogical history of the sample, but dropped considerably under strong exponential growth (Fig. 3).

For seven of the nine full data sets that resulted in a MMC as best model, the fitted growth parameter under KM+exp was 1,000 or higher, and therefore, it is unlikely that the serial sampling influenced the results of model selection for these data sets (Fig. 3, Sup. Fig. 15). Additionally, when we sub-sampled strains from a single year from four data sets, thus minimizing the effect of serial sampling, eight of the nine subsets resulted in a MMC model (Table 4).

To overcome the limitation of assuming synchronous sampling, we encourage future studies to develop MMC models that explicitly consider the time of sampling. Such a model is proposed in Hoscheit and Pybus (2019), but without an explicit mechanism to convert real time units in coalescent time units.

Overall, among the 23 data sets considered here (including subsets), 20 supported a MMC as the best fitting model, for 14 of these, the fitted model had a posterior probability higher than 80% and could reproduce the observed data. These results provide compelling evidence that, in most cases, MMC models have a better fit to data from MTB outbreaks compared to standard models based on the Kingman coalescent. These findings have deep implications for population genetic studies of MTB (also discussed in Morales-Arce et al. 2019): 1) in the last five years, at least a dozen studies inferred the demographic history of different MTB populations (Pepperell et al. 2013, Comas et al. 2013, Bos et al. 2014, Lee et al. 2015 Eldholm et al. 2015, Kay et al. 2015, Luo et al. 2015, Merker et al. 2015, Folkvardsen et al. 2017, Merker et al. 2018, Liu et al. 2018, Bainomugisa et al. 2018, O'Neill et al. 2019). In the light of our results, it is not surprising that most of these studies found evidence for population growth, as it is known that performing demographic inference on MMC genealogies assuming the Kingman coalescent fits high growth rates (Eldon et al. 2015). Moreover, a recent study fitted an exponential growth (or shrinkage) model to 21 MTB data sets, and found

evidence for population growth for 14 of them, for seven data sets it was not possible to reject the hypothesis of constant population size, and none of them resulted in population shrinkage (Menardo et al. 2019). One additional factor that could bias the results of Kingman-based demographic inference towards population growth is sampling bias (Lapierre et al. 2016). However, the majority of the data sets considered in this study are composed by (nearly) all known TB cases caused by a certain phylogenetic clade (Roetzer et al. 2013, Lee et al. 2015, Stucki et al. 2015, Bjorn-Mortensen et al. 2016, Eldholm et al. 2016), or by a random subset of them (Folkvardsen et al. 2017). Therefore, these data sets should not be strongly affected by sampling bias, although we cannot exclude that some of them are.

2) Another implication of our results regards the interpretation of private mutations (singletons). It is known that MTB data sets show an excess of singletons (Hershberg et al. 2008, Pepperel et al. 2013, Gagneux 2018). While this could be the results of either skewed offspring distributions or population growth, in the past, the excess of singleton has been interpreted as evidence for strong selection. In particular Pepperell et al. (2013) fitted a model including both population growth and selection to the global MTB population, and found pervading strong purifying selection across 95% of the genome, with a genome wide selection coefficient several orders of magnitude higher than what estimated for other organisms. Again, this analysis assumed non-skewed offspring distribution, and its outcome was potentially biased by model misspecification.

While MMC genealogies were fitting better to most data sets, for three of them, KM+exp was best fitting model (Sup. Table 2). In one case (Shitikov 2017), the posterior predictive check could not reproduce the genetic diversity of the data, indicating that other factors such as complex population dynamics or sampling biases are likely to influence the shape of the genealogies (Lapierre et al. 2016). Conversely, for the remaining two data sets (Stucki 2016 and Folkvardsen 2017 sampled in 2010), the posterior predictive checks could reproduce the observed genetic diversity. We therefore cannot exclude that in these populations the variance of the offspring distribution was small, and that the Kingman coalescent is an adequate model. However, it is difficult to reconcile this with the fact that BETA was the best fitting model for the complete data set and for the other two temporal subsets of Folkvardsen 2017 (sampled in 2009 and 2012; Table 4).

Can we say more on the type and size of multiple mergers? The majority of MMC signals that we found were Beta genealogies, pointing to a moderate size of ancestral lineages merged in a SETI (short evolutionary time interval, corresponding to a single time point on the coalescent time scale). However, we also found four data sets fitting best to a Dirac coalescent genealogy, pointing to large groups of lineages merged during a SETI. This corresponds to a single strain having the potential to found large families over a SETI compared to other strains. The Dirac signals come from the

smallest data set with the highest errors of misclassification between the model classes (Eldholm 2016), and from three subsets of data that featured high OOB error rate or decreased posterior probabilities for the inferred Dirac genealogy. Therefore, much caution should be taken with proposing the Dirac coalescent as a general model, especially since in all cases the Beta coalescent was the second best fitting model. These results are in contrast with the choice of the Dirac coalescent as underlying MMC model for within-host MTB samples, as suggested in Morales-Arce et al. (2019). The choice of the Dirac coalescent in Morales-Arce et al. (2019) is, as our choices of MMC models, not informed by an explicit MTB population model, but by convenience that these models are commonly used. Thus, it would be interesting to assess whether the Beta coalescent would be a better fit also to within-host data. The relatively low ψ estimates obtained by Morale-Arce et al. (2019) suggest that this is quite possible (Morales-Arce et al. 2019; Figure 3).

If the Beta coalescent is the inferred genealogy, one question is whether rapid selection might be the reason for the inference of a MMC genealogy (i.e. the parameter α is equal to 1, corresponding to the Bolthausen-Sznitman coalescent; Bolthausen and Sznitman 1998). For eight of the 16 data (sub) sets with inferred Beta genealogy, the 95% posterior credible interval of the parameter α included 1 in all replicates (Sup. Table 2). However, due to the effect of serial sampling, the true variability of the estimates is likely bigger (Sup. Fig. 15). Thus, rapid selection may be a possible explanation for multiple merger. Other possible explanations are transmission bottlenecks, super-spreaders or further unknown processes.

Our study shows that genetic diversity in MTB outbreaks is modeled well by assuming MMC genealogies across many data sets. However, we stress that we did not use an explicit model for MTB, but two classes of MMC models that were employed in previous work, and span the strength of multiple mergers between strictly bifurcating (Kingman's coalescent) and star-shaped genealogies, where all sampled lineages merge at a single time point. Moreover, more elaborate multiple merger models including population size changes and/or serial sampling may improve the fit to the data. Both extensions are easily achieved (e.g. Spence et al. 2016, Hoscheit and Pybus 2019, Morales-Arce et al. 2019), and some tools for simulating multiple merger genealogies with changes in population sizes are available (Matuszewski et al. 2018, Hoscheit and Pybus 2019). However, we refrained from adding multiple merger models with varying population size to our model comparison. The main reason is that with this work, we wanted to investigate multiple merger models as possible alternative to bifurcating genealogies to study between-host populations of MTB. We found that, even when assuming a constant population size, MMC models fitted better compared to Kingman with changes in population size (modeled as exponential growth), and that the fitted MMC models could reproduce the observed data. Therefore, adding MMC models with varying population size to our model comparison would not alter our main results: that skewed

offspring distribution (over several generations) shapes the genealogies and the genetic diversity of MTB populations, and that ignoring this can bias the results of demographic inference. Moreover, it was reported before that at least for one MMC model, i.e. the Dirac coalescent, it is hard to distinguish between genealogies with and without population expansion (see the discussions of Figure 3 in Morales-Arce et al. 2019 and Table A4 in Freund and Siri-Jégousse 2019).
No discrete-generation population model for MTB is currently available that would allow to identify the corresponding coalescent model. Such population model should include host-to-host transmission, intra-host evolution, super-spreaders, serial sampling, latency, population size changes and the potential selective pressure caused by the host immune system. This might very well result in a different multiple merger model compared to the ones that we employed. For instance, while the Beta coalescent has been proposed as genealogy model for populations with recurrent strong bottlenecks (Tellier and Lemaire 2014), rigorous mathematical modeling predicted different coalescent processes for extreme bottlenecks (Tams et al. 2009, Casanova et al. 2019). Such a model could also close the following implicit modeling gap in applying MMC to MTB and to bacteria in general. Mathematically, MMC processes have been introduced as approximations (with changed time scale) of the genealogy in underlying discrete population reproduction models, so called Cannings models (e.g. Möhle and Sagitov 2001). The underlying population models feature many offspring of a single individual per generation (e.g. Schweinsberg 2003, Eldon and Wakeley 2006, Desai et al. 2013), however, bacteria replicate through binary fission. While such population models are not applicable directly to bacteria, the underlying mathematical theory only needs to guarantee that the mergers within a SETI follow a certain probability distribution, so one can define similar models where the large offspring number of one individual per generation is spread over multiple generations (Möhle and Sagitov 2001). However, the exact distribution and model for MTB multiple mergers is unclear, and our results only show evidence that moderate multiple merger events are likely to play a role in shaping the diversity. As mentioned above, to infer the exact nature of MTB multiple mergers, we encourage future studies to formulate an explicit population model for MTB.

In conclusion, our results show that, when studying MTB local population and outbreaks, models that do not allow for skewed offspring distribution on the coalescent timescale (Kingman), have consistently worse fit compared to MMC, and can lead to biased results. Further research is needed to extend MMC models to more realistic scenarios with complex population dynamics and serial sampling. These developments will be useful to study MTB and potentially other pathogens. Additionally, the formulation of an explicit population model for MTB, would help to identify the most appropriate genealogy model for demographic inference of MTB populations.

Methods

Data set selection

We searched the literature for WGS studies of outbreaks or local populations of *Mycobacterium tuberculosis.* We selected local data sets to avoid as much as possible geographic population structure and sampling biases that could influence the analysis. We identified 11 data sets: eight outbreaks and three clades with a restricted geographical range.

- *Roetzer et al. 2013:* lineage 4 outbreak in Hamburg, Germany (61 strains, 74 polymorphic positions).
- *Comas et al. 2015:* lineage 7 strains sampled Ethiopia. Lineage 7 is a rare human adapted lineage endemic to Ethiopia and perhaps also to neighboring countries, only few genomes are available and most of them are included in this data set (21 strains, 1334 polymorphic positions).
- *Eldholm et al. 2015:* lineage 4 multi-drug resistant outbreak in Buenos Aires, Argentina (248 strains, 497 polymorphic positions).
- *Lee et al. 2015:* lineage 4 outbreak in 11 Inuit villages in Nunavik, Québeq, Canada. We considered only the major sub-lineage Mj, a second smaller outbreaks of an unrelated sub-lineage (Mn) was excluded (147 strains, 454 polymorphic positions).
- *Stucki et al. 2015:* lineage 4 outbreak in Bern, Switzerland (60 strains, 128 polymorphic positions).
- *Bjorn-Mortensen et al. 2016:* lineage 4 outbreak in Greenland. To minimize the potential effect of population structure we considered only the major cluster GC4, because the other clusters represent independent outbreaks belonging to other sub-lineages (121 strains 128 polymorphic positions).
- *Stucki et al. 2016:* sub-lineage L4.6.1/Uganda, belonging to lineage 4. This sub-lineage is endemic to central African countries (175 strains, 6264 polymorphic positions).

Eldholm et al. 2016: lineage 2 outbreak in Oslo, Norway. From the data set of the original publication we excluded all strains that did not belong to the Oslo outbreak (25 strains, 17 polymorphic positions).

- *Folkvardsen et al. 2017:* large lineage 4 outbreak in Copenhagen, Denmark (702 strains 514 polymorphic positions).
- *Shitikov et al. 2017:* W148 outbreak belonging to lineage 2, this clade has also been named B, B0, CC2, East European 2 and ECDC0002 (176 strains, 1164 polymorphic positions).
- *Bainomugisa et al. 2018:* lineage 2 multi-drug resistant outbreak on a small island (Daru) in Papua New Guinea. From the data set of the original publication we excluded all the strains that did not belong to the Daru outbreak (81 strains, 401 polymorphic positions).

Bioinformatic pipeline

For all samples Illumina reads were trimmed with Trimmomatic v0.33 (SLIDINGWINDOW: 5:20,ILLUMINACLIP:{adapter}:2:30:10) (Bolger 2014). Reads shorter than 20 bp were excluded for the downstream analysis. Overlapping paired-end reads were then merged with SeqPrep (overlap size = 15; https://github.com/jstjohn/SeqPrep). The resulting reads were mapped to the reconstructed MTB complex ancestral sequence (Comas 2013) with BWA v0.7.12 (mem algorithm; Li and Durbin 2009). Duplicates reads were marked by the MarkDuplicates module of Picard v 2.1.1 (https://github.com/broadinstitute/picard). The RealignerTargetCreator and IndelRealigner modules of GATK v.3.4.0 (McKenna et al. 2010) were used to perform local realignment of reads around Indels. Reads with alignment score lower than (0.93*read_length)-(read_length*4*0.07)) were excluded: this corresponds to more than seven miss-matches per 100 bp.

SNPs were called with Samtools v1.2 mpileup (Li 2011) and VarScan v2.4.1 (Koboldt et al. 2012) using the following thresholds: minimum mapping quality of 20, minimum base quality at a position of 20, minimum read depth at a position of 7X, minimum percentage of reads supporting the call 90%.

Genomes were excluded if they had 1) an average coverage < 20x, 2) more than 50% of their SNPs excluded due to the strand bias filter, 3) more than 50% of their SNPs having a percentage of reads supporting the call between 10% and 90%, or 4) contained single nucleotide polymorphisms that belonged to different MTB lineages, as this indicates that a mix of genomes was sequenced.

Because missing data can significantly impact population genetic inference we further excluded all strains that had less SNP calls than (average - (2 * standard deviation)) of the respective data set (calculated after all previous filtering steps).

The filters described above were applied to all data sets with one exception: in the Comas 2015 data set most strains failed the strand bias filter, therefore this filter was not applied.

The single vcf were merged with the CombineVariant module of GATK v.3.4.0 (McKenna et al. 2010), the genotype field was edited to make it haploid (0/0 => 0; 1/1 => 1; 0/1 and 1/0 => .).

Vcftools 0.1.14 (Danecek et al. 2011) was used to extract variable positions excluding predefined repetitive regions (Comas et al. 2013) and excluding position with missing data.

The variable positions were converted in a multi fasta file including the reconstructed ancestral sequence on which the mapping was performed.

A phylogenetic tree based on the resulting variable positions was built with RaxML 8.2.11 (Stamatakis 2014) using a GTRCAT model and the -V option.

PAML (baseml) (Yang 2017) was used to reconstruct the ancestral sequence of each data set. To identify the MRCA of each data set the tree was rooted using the reconstructed ancestral sequence

of the MTB complex as published in Comas et al. (2013), which is also the genome reference sequence used for the mapping.
For each data set all polymorphic positions for all strains and their reconstructed ancestor were then collected in fasta files (the data is available together with the ABC pipeline at https://github.com/fabianHOH/mmc_R_gendiv/tree/master/MTB_MMC_repo).

Model selection and parameter estimation

For model selection and parameter estimation, we used a random forest based Approximate Bayesian Computation approach (Pudlo et al. 2015, Raynal et al. 2018).
We selected between Kingman's n-coalescent (KM), Kingman's n-coalescent with exponential growth (KM+exp), Beta coalescent (BETA) and Dirac coalescent (Dirac). For each data set we collected the genetic polymorphisms identified with the bioinformatic analysis and calculated a set of 20 summary statistics following the recommendations from Freund and Siri-Jégousse (2019), Scenario 3: the (.1,.3,.5,.7,.9) quantiles of the mutant allele frequency spectrum, the (.1,.3,.5,.7,.9) quantiles of the LD measure r^2 between pairs of segregating sites, the (.1,.3,.5,.7,.9) quantiles of the minimal observable clade sizes of each sequence, the number of segregating site, the nucleotide diversity and the mean, standard deviation and harmonic mean of the minimal observable clade sizes. For computational reasons, when we analyzed the data set Stucki 2016, we omitted the quantiles of r^2, but used more quantiles (.1,.2,.3,.4,.5,.6,.7,.8,.9) of the mutant allele frequencies. For each model we performed 125,000 simulations of a sample of size n where n is the number of individuals in the data set, drawing the scaled population size from a binomial distribution on log-equally spaced discrete θ spanning one order of magnitude around the Watterson estimator (θ_{obs}), as in Freund and Siri-Jégousse (2019). For KM+exp we drew the value of the exponential growth rate (g) from a uniform distribution [0,2,4,…,5000] except for the data sets Eldholm 2015, Stucki 2016 and Folkvardsen 2017, where we used a uniform distribution [0,5,10,…,20000]. Note that this is a growth rate for a coalescent within a diploid population, values should be halved for interpretation in a haploid setting. The choice of wider ranges were based on preliminary analyses of the data with narrower prior distributions that showed a posterior distribution of g skewed at the upper end. For BETA and Dirac we drew the value of the free parameters α and ψ from a uniform distribution, [1,1.975] and [0.025,0.975] respectively (discretized with equidistant steps of 0.025). Note that BSZ is included in BETA for α =1, while KM is additionally included in KM+exp for g = 0. Simulations were performed in R as described in Freund and Siri-Jégousse (2019), the code is available at https://github.com/fabianHOH/mmc_R_gendiv.
As described in Pudlo et al. (2015), we performed model selection via Approximate Bayesian Computation using a random forest of decision trees, using the R package abcrf (Pudlo et al. 2015).

We drew 1,000 bootstrap samples of size 100,000 from the simulations and then constructed decision trees based on decision nodes of the form S>t, where S is one of the summary statistics used. For each node, S and t are chosen so that the bootstrap sample is divided as well as possible in sets coming from the same of the four model classes (minimal Gini impurity). Nodes are added to the tree until all simulations of the bootstrap samples are sorted into sets from the same model class. Misclassification is measured by the out-of-bag (OOB) error, i.e. the proportion of decision trees for each simulation that sorts it into a wrong model class, averaged over simulations and, for the overall OOB error, model classes.

For parameter estimation within a model class, we followed Raynal et al. (2018). Here, the decision (regression) trees are constructed analogously, only S and t are chosen so that the parameters of the simulations have similar values in both sets divided by the node. This is achieved by minimizing the L^2 loss, i.e. minimizing, for the two sets divided by the node, the L^2 distances of the simulation parameter to the mean parameter in the set. Nodes are added until all simulations sorted into one leaf have the same parameter or there are less than 5 simulations allocated to the leaf.

The observed data is then assigned to the model class where the majority of decision trees for model selection assign it, and its posterior parameter distribution is given by the distribution of the weighted average parameter of the allocated leaf across all trees in the (regression) random forest (see Raynal et al. 2018, sections 2.3.2 and 2.3.3). The posterior probability for model selection is computed as a machine learning estimate of classifying the model class correctly, which includes another regression tree. See Pudlo et al. (2015) for details, a summary can be found in Appendix A.2 in Freund and Siri-Jégousse (2019).

All ABC analyses were repeated in triplicates to control for stochastic effects, the best fitting model did not change between replicates. For the OOB errors, posterior probabilities and median estimates of the parameters we report the mean among the three replicates, for the 95% credibility intervals of the parameter estimates we report the lowest and the highest values (resulting in the largest possible interval) among the three replicates. The individual results for each replicate are reported in Sup. Table 2.

Posterior predictive checks

To assess whether the best fitting model could reproduce the observed data, we performed posterior predictive checks. We simulated 10,000 sets of summary statistics under the best fitting model (using the median of the posterior growth rate or of the multiple merger coalescent parameter, averaged over the three replicates) and compared them graphically with the value of the statistics observed in each data set. As scaled mutation rate, we used the generalized Watterson estimate

2s/E(L), where s is the number of mutations observed in the data set and E(L) the expected length of all branches for the best fitting coalescent model.

Population structure and declining population size for the data set Lee 2015

To assess the effect of population structure in the data set Lee 2015, we simulated samples under Kingman's n-coalescent with population structure. From the phylogenetic tree (Fig. 2), we identified four different clades with sizes 61, 36, 49 and 1. We then assumed these to be sampled from different sub-populations of equal size in an island model with scaled symmetric migration. We performed coalescent simulations under a structured (Kingman) coalescent with exponential growth. We used the same prior for growth rate as in the approach not accounting for population structure and additionally drew the scaled migration rate m (in units of 4Nm*, where m* is the migration rate in the discrete island model) from the uniform discrete distribution {.25,.5,1,2,3}. We approximated Watterson's estimator for a specific choice of parameters by replacing the expected total length of the coalescent by the mean total length from 10,000 coalescent simulations with these parameters.

For generating samples under Kingman's n-coalescent with exponential decline, we had to slightly change the simulation procedure using ms. Since population decline may lead to coalescent times to large too simulate, we fixed the maximal population size in the past to 1,000 times the present population size. Then, given an exponential growth rate g<0, the decline starts at time log(1000)/(-g) (in coalescent time units backwards in time from time of sampling) and continues until the sampling time.

To compute Watterson's estimator in this scenario for any g, we need the expected total length of the coalescent tree. Instead of computing it analytically, we recorded the total coalescent tree length of 10,000 simulations under the model and used their mean as an approximation of the expected total branch length.

As parameters for exponential decline, we use exponential growth rates drawn uniformly from {-250,-200,-150,-100,-50,-25,-10}. As for Stucki 2016, we omitted the r^2 statistics (and added further quantiles of other statistics, see above) due to computational reasons (occasionally, trees with very long tree branches and thus many mutations are produced, which inflates computation time for r^2). For both exponential decline and population structure, we ran the ABC-RF analysis as for all other data sets. Simulations were produced with Hudson's ms as implemented in the R package phyclust.

Accounting for serial sampling

Following Hoscheit and Pybus (2019), we add serial sampling to the MMC and to Kingman's coalescent with exponential growth simply by stopping the coalescent at times (on the coalescent

time scale) where further individuals are sampled. Then, we start a new (independent) coalescent tree that has rates and waiting times as the non-serial coalescent (multiple merger or with growth) started in the last state of the stopped coalescent plus adding one block with a single individual for each individual sampled at this time. A R implementation is available at https://github.com/fabianHOH/mmc_R_gendiv/tree/master/MTB_MMC_repo.

A problem with this approach is that one needs the scaling factor between coalescent time and real time. While estimation procedures coming from phylogenetics are available in the case of Kingman's coalescent (e.g. Drummond and Rodrigo 2000), they cannot be applied directly to the case of multiple merger coalescents. Additionally, a brute force search for appropriate scaling on top of our models is computationally unfeasible with the ABC approach that we adopted in this study.

Hence, we assessed, for different fixed scaling factors, how strong the effect of ignoring serial sampling in the models is. We considered the setting of Eldholm 2015 (n=248, s=497), Lee 2013 (n=147, s=454) and Roetzer 2013 (n=61, s=74). We used the real dates of the serial sampling for these data sets and we performed serial coalescent simulations as described above. We used different time (re)scaling factors c, such as c determines the time ct at which an individual sampled at real time -t (0 corresponds to the latest sampling time) is added as a new lineage to the coalescent tree (so ct is in coalescent time units). Here, we assessed c by setting the earliest sampling time (highest t) to a fraction c'≥0 of the expected height of the coalescent tree if there was no serial sampling (so keeping all other parameters, but assuming c=0). For each c' in {0,0.1,0.2,0.3,0.4,0.5,0.75,1,1.5}, we simulated 1,000 simulations under each parameter set (g in {1,10,50,100,250,500,1000,2000}, α in {1,1.2,1.4,1.6,1.8,2}, ψ in {0.1,0.3,0.5,0.7,0.9]) and then performed ABC model selection for each simulation, recording how often the serial coalescent simulations were sorted to which non-serial model class. We also reported the quality of parameter estimation for the growth rate or coalescent parameter by measuring the (absolute) distances of the estimated parameter to the parameter used for the simulation (Sup. Fig. 15)

References

Árnason, E., & Halldórsdóttir, K. (2015). Nucleotide variation and balancing selection at the Ckma gene in Atlantic cod: analysis with multiple merger coalescent models. PeerJ, 3, e786.

Bainomugisa, A., Lavu, E., Hiashiri, S., Majumdar, S., Honjepari, A., Moke, R., ... & Coulter, C. (2018). Multi-clonal evolution of multi-drug-resistant/extensively drug-resistant Mycobacterium tuberculosis in a high-prevalence setting of Papua New Guinea for over three decades. Microbial genomics, 4(2).

Birkner, M., Blath, J., Möhle, M., Steinrücken, M., & Tams, J. (2009). A modified lookdown construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks. *Alea*, *6*, 25-61.

Biswas, S., & Akey, J. M. (2006). Genomic insights into positive selection. TRENDS in Genetics, 22(8), 437-446.

Bjorn-Mortensen, K., Soborg, B., Koch, A., Ladefoged, K., Merker, M., Lillebaek, T., ... & Kohl, T. A. (2016). Tracing Mycobacterium tuberculosis transmission by whole genome sequencing in a high incidence setting: a retrospective population-based study in East Greenland. Scientific reports, 6, 33180.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.

Bolthausen E, Sznitman AS (1998) On Ruelle's probability cascades and an abstract cavity method. Communications in Mathematical Physics, 197, 247–276.

Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., ... & Campbell, T. J. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature, 514(7523), 494.

Casanova, Adrián González, Verónica Miró Pina, and Arno Siri-Jégousse. "The symmetric coalescent and Wright-Fisher models with bottlenecks." *arXiv preprint arXiv:1903.05642* (2019).

Chiner-Oms, Á., Sánchez-Busó, L., Corander, J., Gagneux, S., Harris, S. R., Young, D., ... & Comas, I. (2019). Genomic determinants of speciation and spread of the Mycobacterium tuberculosis complex. *Science Advances*, *5*(6), eaaw3307.

Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K. E., Kato-Maeda, M., ... & Yeboah-Manu, D. (2013). Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nature genetics, 45(10), 1176.

Comas, I., Hailu, E., Kiros, T., Bekele, S., Mekonnen, W., Gumi, B., ... & Goig, G. A. (2015). Population genomics of Mycobacterium tuberculosis in Ethiopia contradicts the virgin soil hypothesis for human tuberculosis in Sub-Saharan Africa. Current Biology, 25(24), 3260-3266.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & McVean, G. (2011). The variant call format and VCFtools. Bioinformatics, 27(15), 2156-2158.

Desai, M. M., Walczak, A. M., & Fisher, D. S. (2013). Genetic diversity and the structure of genealogies in rapidly adapting populations. Genetics, 193(2), 565-585.

Drummond, A., & Rodrigo, A. G. (2000). Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. *Molecular Biology and Evolution*, *17*(12), 1807-1815.

Drummond, A. J., Rambaut, A., Shapiro, B., & Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular biology and evolution, 22(5), 1185-1192.

Durrett, R., & Schweinsberg, J. (2005). A coalescent model for the effect of advantageous mutations on the genealogy of a population. Stochastic processes and their applications, 115(10), 1628-1657.

Eldholm, V., Monteserin, J., Rieux, A., Lopez, B., Sobkowiak, B., Ritacco, V., & Balloux, F. (2015). Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. Nature communications, 6, 7119.

Eldholm, V., Pettersson, J. H. O., Brynildsrud, O. B., Kitchen, A., Rasmussen, E. M., Lillebaek, T., ... & Balloux, F.. (2016). Armed conflict and population displacement as drivers of the evolution and dispersal of Mycobacterium tuberculosis. Proceedings of the National Academy of Sciences, 113(48), 13881-13886.

Eldon, B., Birkner, M., Blath, J., & Freund, F. (2015). Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents?. *Genetics*, *199*(3), 841-856.

Eldon, B., & Wakeley, J. (2006). Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, *172*(4), 2621-2633.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. PLoS genetics, 9(10), e1003905.

Folkvardsen, D. B., Norman, A., Andersen, Å. B., Michael Rasmussen, E., Jelsbak, L., & Lillebaek, T. (2017). Genomic epidemiology of a major mycobacterium tuberculosis outbreak: retrospective cohort study in a low-incidence setting using sparse time-series sampling. The Journal of infectious diseases, 216(3), 366-374.

Freund, F., & Siri-Jégousse, A. (2019). Distinguishing coalescent models-which statistics matter most?. BioRxiv, 679498.

Gagneux, S. (2018). Ecology and evolution of Mycobacterium tuberculosis. *Nature Reviews Microbiology*, *16*(4), 202.

Gardy, J. L., Johnston, J. C., Sui, S. J. H., Cook, V. J., Shah, L., Brodkin, E., ... & Varhol, R. (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. New England Journal of Medicine, 364(8), 730-739.

Griffiths, R. C., & Tavare, S. (1994). Sampling theory for neutral alleles in a varying environment. Phil. Trans. R. Soc. Lond. B, 344(1310), 403-410.

Heller, R., Chikhi, L., & Siegismund, H. R. (2013). The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. PLoS One, 8(5), e62992.

Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G., ... & Przeworski, M. (2011). Classic selective sweeps were rare in recent human evolution. science, 331(6019), 920-924.

Hershberg, R., Lipatov, M., Small, P. M., Sheffer, H., Niemann, S., Homolka, S., ... & Gagneux, S. (2008). High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. PLoS biology, 6(12), e311.

Horns, F., Vollmers, C., Dekker, C. L., & Quake, S. R. (2019). Signatures of selection in the human antibody repertoire: Selective sweeps, competing subclones, and neutral drift. Proceedings of the National Academy of Sciences, 116(4), 1261-1266.

Hoscheit, P., & Pybus, O. G. (2019). The multifurcating skyline plot. Virus Evolution, 5(2), vez031.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. Theoretical population biology, 23(2), 183-201.

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics, 18(2), 337-338.

Irwin, K. K., Laurent, S., Matuszewski, S., Vuilleumier, S., Ormond, L., Shim, H., ... & Jensen, J. D. (2016). On the importance of skewed offspring distributions and background selection in virus population genetics. Heredity, 117(6), 393.

Joy, D. A., Feng, X., Mu, J., Furuya, T., Chotivanich, K., Krettli, A. U., ... & Beerli, P. (2003). Early origin and recent expansion of Plasmodium falciparum. Science, 300(5617), 318-321.

Kaplan, N. L., Darden, T., & Hudson, R. R. (1988). The coalescent process in models with selection. Genetics, 120(3), 819-829.

Kato, M., Vasco, D. A., Sugino, R., Narushima, D., & Krasnitz, A. (2017). Sweepstake evolution revealed by population-genetic analysis of copy-number alterations in single genomes of breast cancer. Royal Society open science, 4(9), 171060.

Kay, G. L., Sergeant, M. J., Zhou, Z., Chan, J. Z. M., Millard, A., Quick, J., ... & Achtman, M. (2015). Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. Nature communications, 6, 6717.

Kingman, J. F. C. (1982). The coalescent. Stochastic processes and their applications, 13(3), 235-248.

Koboldt, D., Zhang, Q., Larson, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, E., Ding, L., & Wilson, R. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing Genome Research DOI: 10.1101/gr.129684.111

Koskela, J., & Wilke Berenguer, M. (2019). Robust model selection between population growth and multiple merger coalescents. Mathematical biosciences, 311, 1-12.

Kuhner, M. K. (2009). Coalescent genealogy samplers: windows into population history. Trends in Ecology & Evolution, 24(2), 86-93.

Lapierre, M., Blin, C., Lambert, A., Achaz, G., & Rocha, E. P. (2016). The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. *Molecular biology and evolution*, *33*(7), 1711-1725.

Lee, R. S., Radomski, N., Proulx, J. F., Levade, I., Shapiro, B. J., McIntosh, F., ... & Behr, M. A. (2015). Population genomics of Mycobacterium tuberculosis in the Inuit. Proceedings of the National Academy of Sciences, 112(44), 13609-13614.

Lee, R. S., Proulx, J. F., McIntosh, F., Behr, M. A., & Hanage, W. P. (2019). Investigating within-host diversity of Mycobacterium tuberculosis reveals novel super-spreaders in the Canadian North. *bioRxiv*, 801308.

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.

Li H A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov 1;27(21):2987-93. Epub 2011 Sep 8. [PMID: 21903627]

Lin, P. L., Ford, C. B., Coleman, M. T., Myers, A. J., Gawande, R., Ioerger, T., ... & Flynn, J. L. (2014). Sterilization of granulomas is common in active and latent tuberculosis despite within-host variability in bacterial killing. Nature medicine, 20(1), 75.

Liu, Q., Ma, A., Wei, L., Pang, Y., Wu, B., Luo, T., ... & Zuo, T. (2018). China's tuberculosis epidemic stems from historical expansion of four strains of Mycobacterium tuberculosis. *Nature ecology & evolution*, *2*(12), 1982.

Luo, T., Comas, I., Luo, D., Lu, B., Wu, J., Wei, L., ... & Shen, X. (2015). Southern East Asian origin and coexpansion of Mycobacterium tuberculosis Beijing family with Han Chinese. Proceedings of the National Academy of Sciences, 201424063.

Matuszewski, S., Hildebrandt, M. E., Achaz, G., & Jensen, J. D. (2018). Coalescent processes with skewed offspring distributions and nonequilibrium demography. Genetics, 208(1), 323-338.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M. A., (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20:1297-303

Menardo, F., Duchêne, S., Brites, D. & Gagneux, S. (2019). The molecular clock of *Mycobacterium tuberculosis*. PLoS Pathogen 15(9): e1008067.

Merker, M., Barbier, M., Cox, H., Rasigade, J. P., Feuerriegel, S., Kohl, T. A., ... & Andres, S. (2018). Compensatory evolution drives multidrug-resistant tuberculosis in Central Asia. eLife, 7, e38200.

Merker, M., Blin, C., Mona, S., Duforet-Frebourg, N., Lecher, S., Willery, E., ... & Allix-Béguec, C. (2015). Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. Nature genetics, 47(3), 242.

Möhle, Martin, and Serik Sagitov. A classification of coalescent processes for haploid exchangeable population models. *The Annals of Probability* 29.4 (2001): 1547-1562.

Ana Y. Morales-Arce, Rebecca B. Harris, Anne C. Stone, Jeffrey D. Jensen. Within-host *Mycobacterium tuberculosis* evolution: a population genetics perspective. bioRxiv 863894; doi: https://doi.org/10.1101/863894

Neher, R. A., & Hallatschek, O. (2013). Genealogies of rapidly adapting populations. Proceedings of the National Academy of Sciences, 110(2), 437-442.

Neher, R. A., & Walczak, A. M. (2018). Progress and open problems in evolutionary dynamics. arXiv preprint arXiv:1804.07720.

Neuhauser, C., & Krone, S. M. (1997). The genealogy of samples in models with selection. Genetics, 145(2), 519-534.

Niwa, H. S., Nashida, K., Yanagimoto, T., & Handling editor: W. Stewart Grant. (2016). Reproductive skew in Japanese sardine inferred from DNA sequences. ICES Journal of Marine Science, 73(9), 2181-2189.

Nourmohammad, A., Otwinowski, J., Łuksza, M., Mora, T., & Walczak, A. M. (2019). Fierce selection and interference in B-cell repertoire response to chronic HIV-1. Molecular biology and evolution.

O'Neill, M. B., Shockey, A., Zarley, A., Aylward, W., Eldholm, V., Kitchen, A., & Pepperell, C. S. (2019). Lineage specific histories of Mycobacterium tuberculosis dispersal in Africa and Eurasia. Molecular ecology.

Pepperell, C. S., Casto, A. M., Kitchen, A., Granka, J. M., Cornejo, O. E., Holmes, E. C., ... & Feldman, M. W. (2013). The role of selection in shaping diversity of natural M. tuberculosis populations. PLoS pathogens, 9(8), e1003543.

Pudlo, P., Marin, J. M., Estoup, A., Cornuet, J. M., Gautier, M., & Robert, C. P. (2015). Reliable ABC model choice via random forests. Bioinformatics, 32(6), 859-866.

Pybus, O. G., Charleston, M. A., Gupta, S., Rambaut, A., Holmes, E. C., & Harvey, P. H. (2001). The epidemic behavior of the hepatitis C virus. Science, 292(5525), 2323-2325.

Raynal, L., Marin, J. M., Pudlo, P., Ribatet, M., Robert, C. P., & Estoup, A. (2018). ABC random forests for Bayesian parameter inference. Bioinformatics, 35(10), 1720-1728.

Rocha, E. P. (2018). Neutral theory, microbial practice: challenges in bacterial population genetics. Molecular biology and evolution, 35(6), 1338-1347.

Roetzer, A., Diel, R., Kohl, T. A., Rückert, C., Nübel, U., Blom, J., ... & Supply, P. (2013). Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. PLoS medicine, 10(2), e1001387.

Rosenberg, N. A., & Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews Genetics, 3(5), 380.

Sargsyan, O., & Wakeley, J. (2008). A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. Theoretical population biology, 74(1), 104-114.

Schweinsberg, J. (2003). Coalescent processes obtained from supercritical Galton–Watson processes. Stochastic Processes and their Applications, 106(1), 107-139.

Shitikov, E., Kolchenko, S., Mokrousov, I., Bespyatykh, J., Ischenko, D., Ilina, E., & Govorun, V. (2017). Evolutionary pathway analysis and unified classification of East Asian lineage of Mycobacterium tuberculosis. Scientific reports, 7(1), 9227.

Spence, J. P., Kamm, J. A., & Song, Y. S. (2016). The site frequency spectrum for general coalescents. *Genetics*, *202*(4), 1549-1561.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-1313. doi:10.1093/bioinformatics/btu033.

Stucki, D., Ballif, M., Bodmer, T., Coscolla, M., Maurer, A. M., Droz, S., ... & Fenner, L. (2015). Tracking a tuberculosis outbreak over 21 years: strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. The Journal of infectious diseases, 211(8), 1306-1316.

Stucki, D., Brites, D., Jeljeli, L., Coscolla, M., Liu, Q., Trauner, A., ... & Gagneux, S.. (2016). Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nature genetics, 48(12), 1535.

Tellier, A., & Lemaire, C. (2014). Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. Molecular ecology, 23(11), 2637-2652.

Walker, T. M., Ip, C. L., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., ... & Parkhill, J. (2013). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. The Lancet infectious diseases, 13(2), 137-146.

Wilkinson-Herbots, H. M. (1998). Genealogy and subpopulation differentiation under various models of population structure. Journal of Mathematical Biology, 37(6), 535-585.

World Health Organization. (2019) Global tuberculosis report 2019. (https://www.who.int/tb/publications/global_report/en/).

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution, 24(8), 1586-1591.

Ypma, R. J., Altes, H. K., van Soolingen, D., Wallinga, J., & van Ballegooijen, W. M. (2013). A sign of superspreading in tuberculosis: highly skewed distribution of genotypic cluster sizes. Epidemiology, 24(3), 395-400.

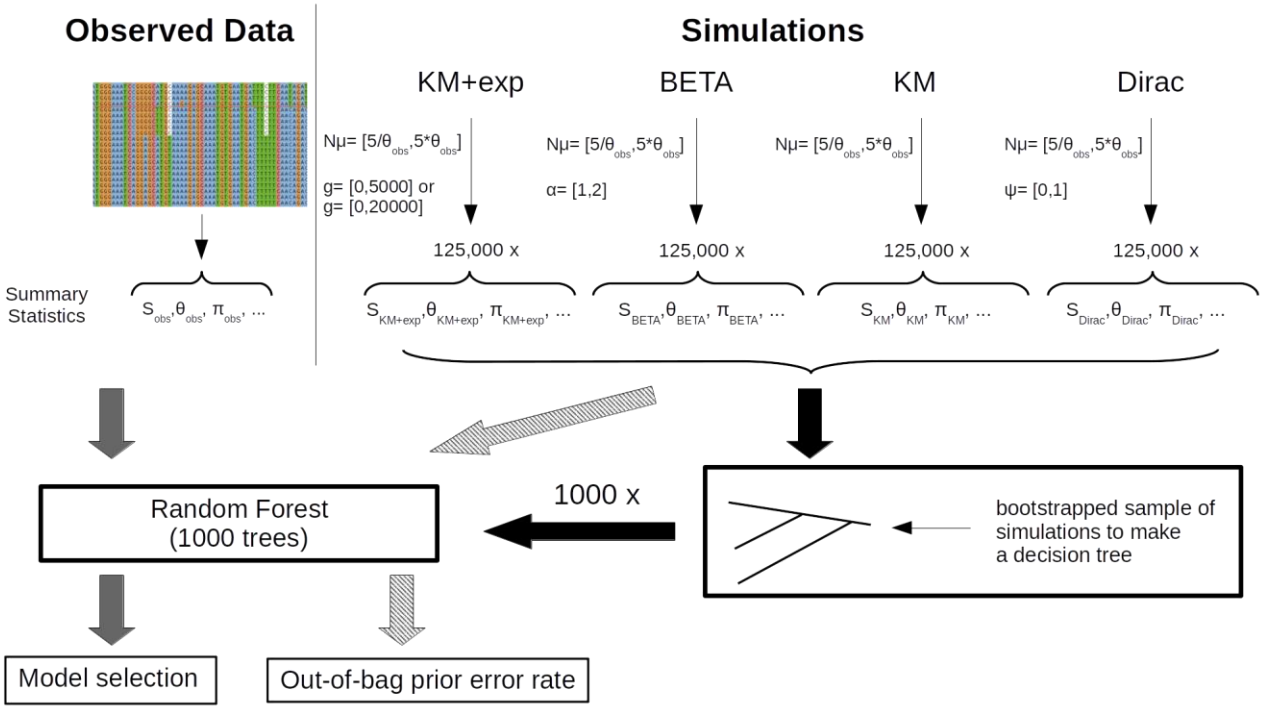


Figure 1. Workflow of the ABC-RF analysis (model selection), see text and Methods for details.

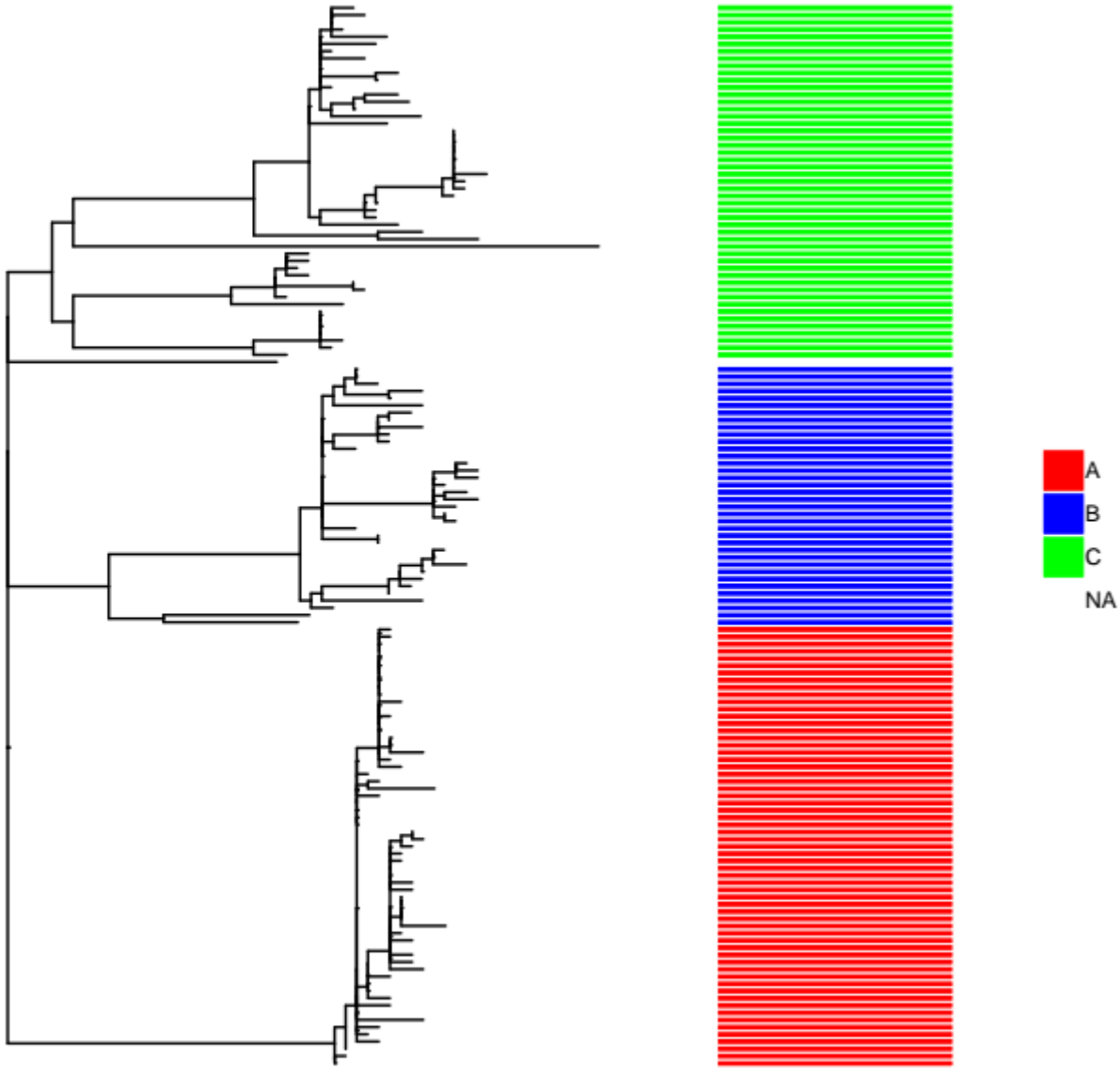


Figure 2. Phylogenetic tree of the data set Lee 2015 with the three sub-clades highlighted.

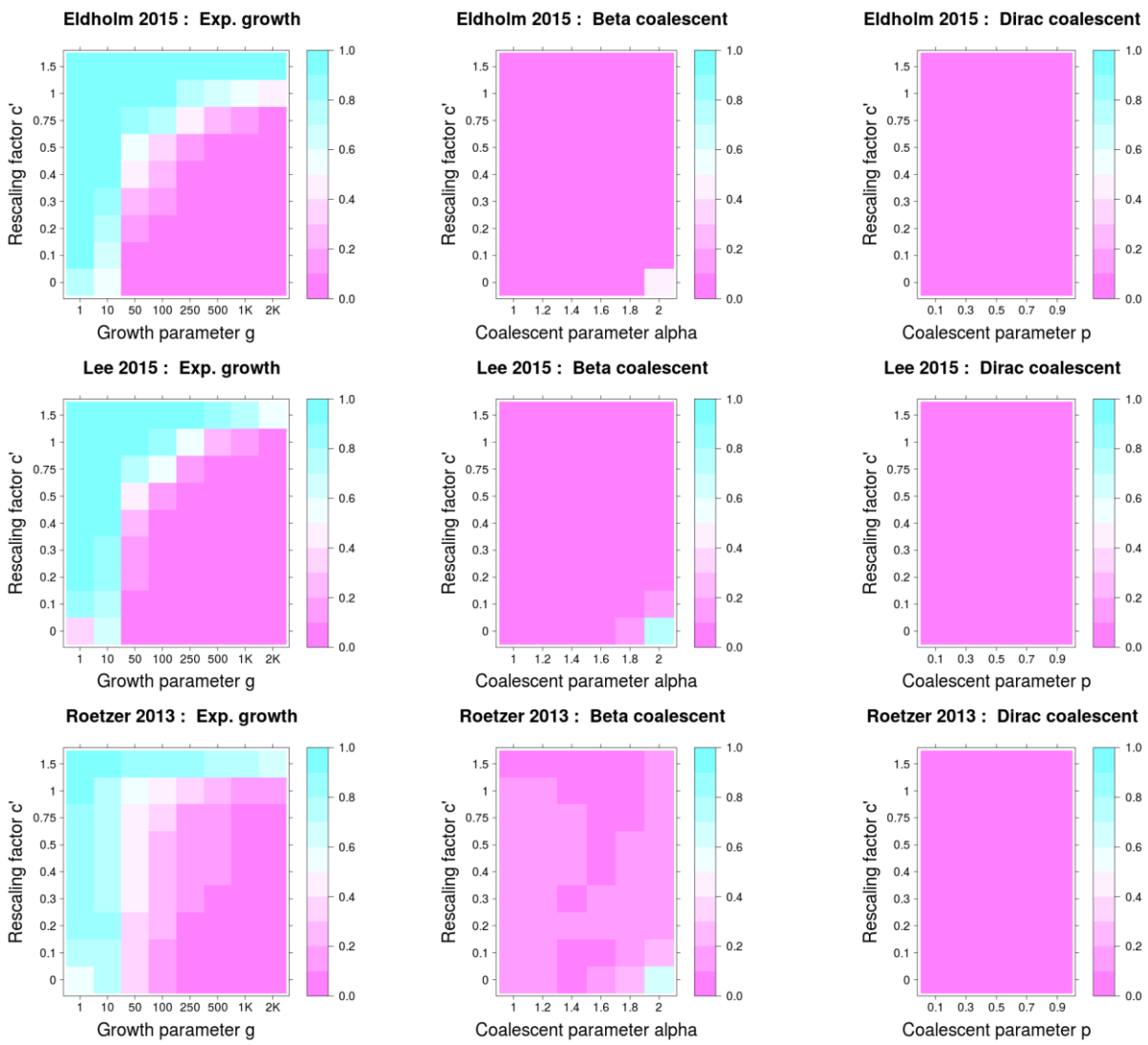
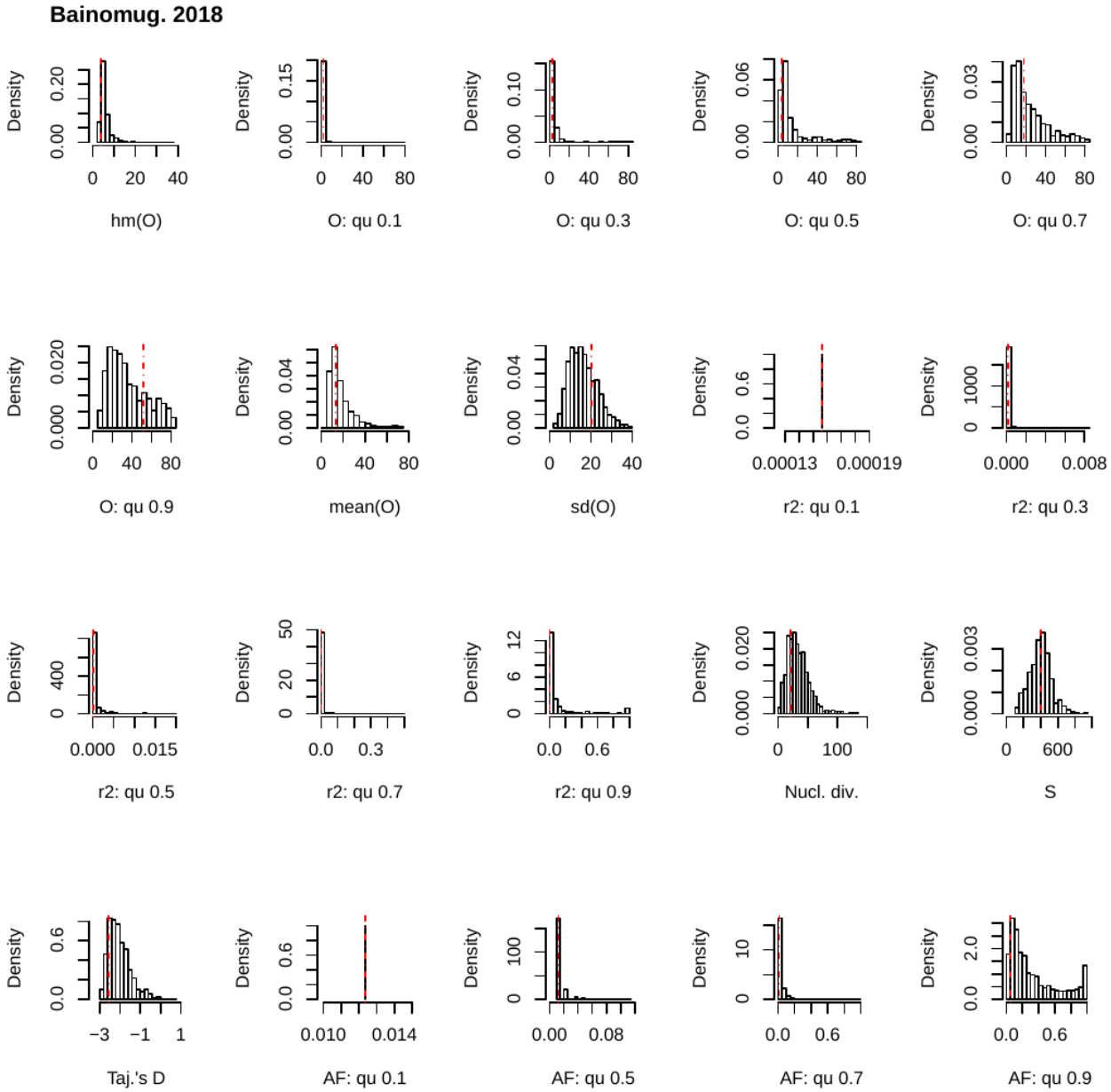
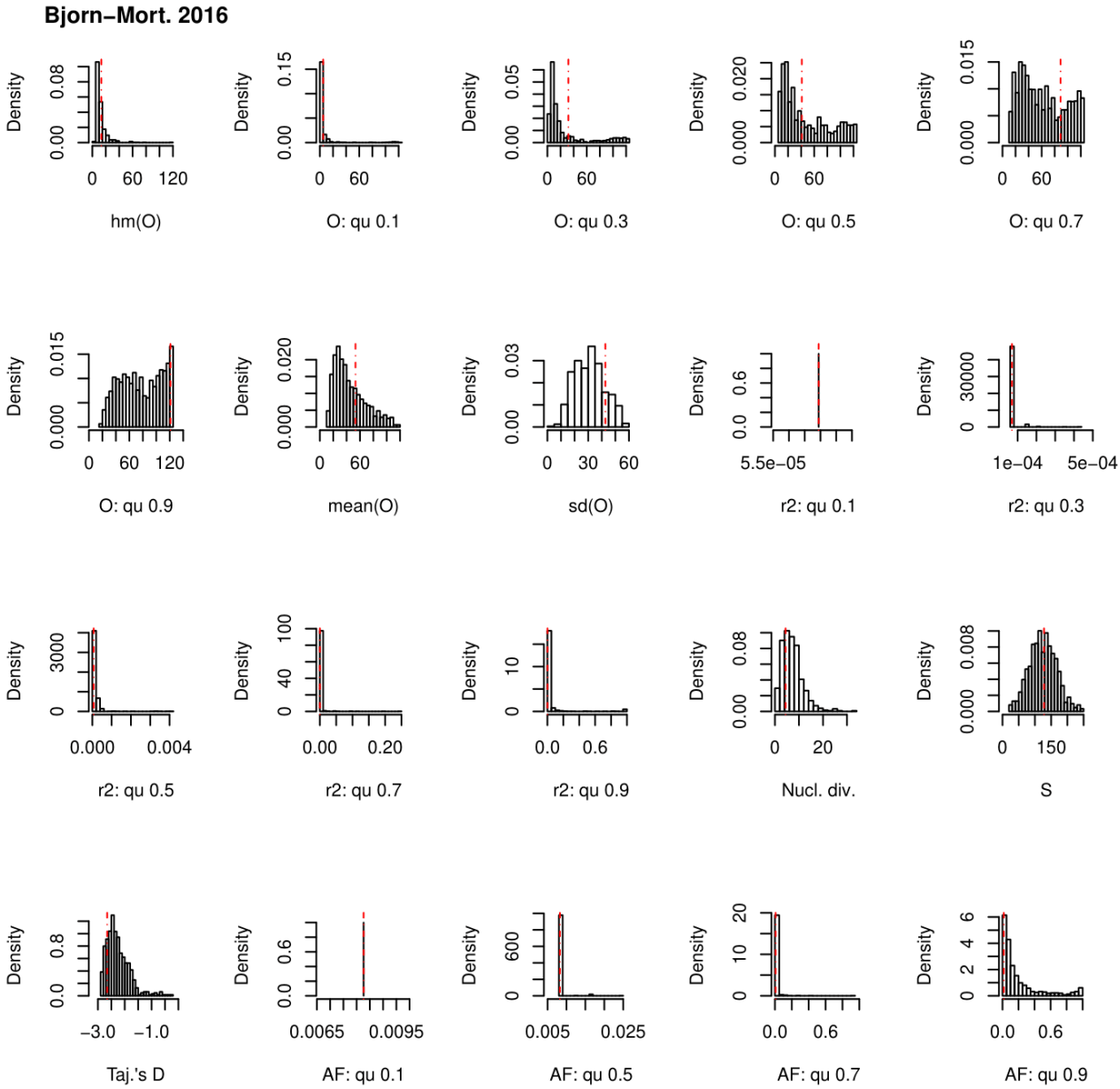


Figure 3: Proportion of model misidentification for serial simulations when model selection was performed via ABC using ultrametric tree models. Misclassification probabilities are shown as a function of c' (the proportion of the genealogy corresponding to the time period in which samples are collected. i,e the period of sampling spans a time period c'·h, where h is the expected height of the genealogy without serial sampling), and of the parameter of the coalescent models. Misclassification was measured as follow: i) for simulations from serially sampled Kingman's coalescent with exponential growth as being misidentified as either Beta or Dirac (first column) ii) for simulations from serially sampled Beta or Dirac coalescents as being misidentified as Kingman's coalescent with or without exponential growth (second and third columns).

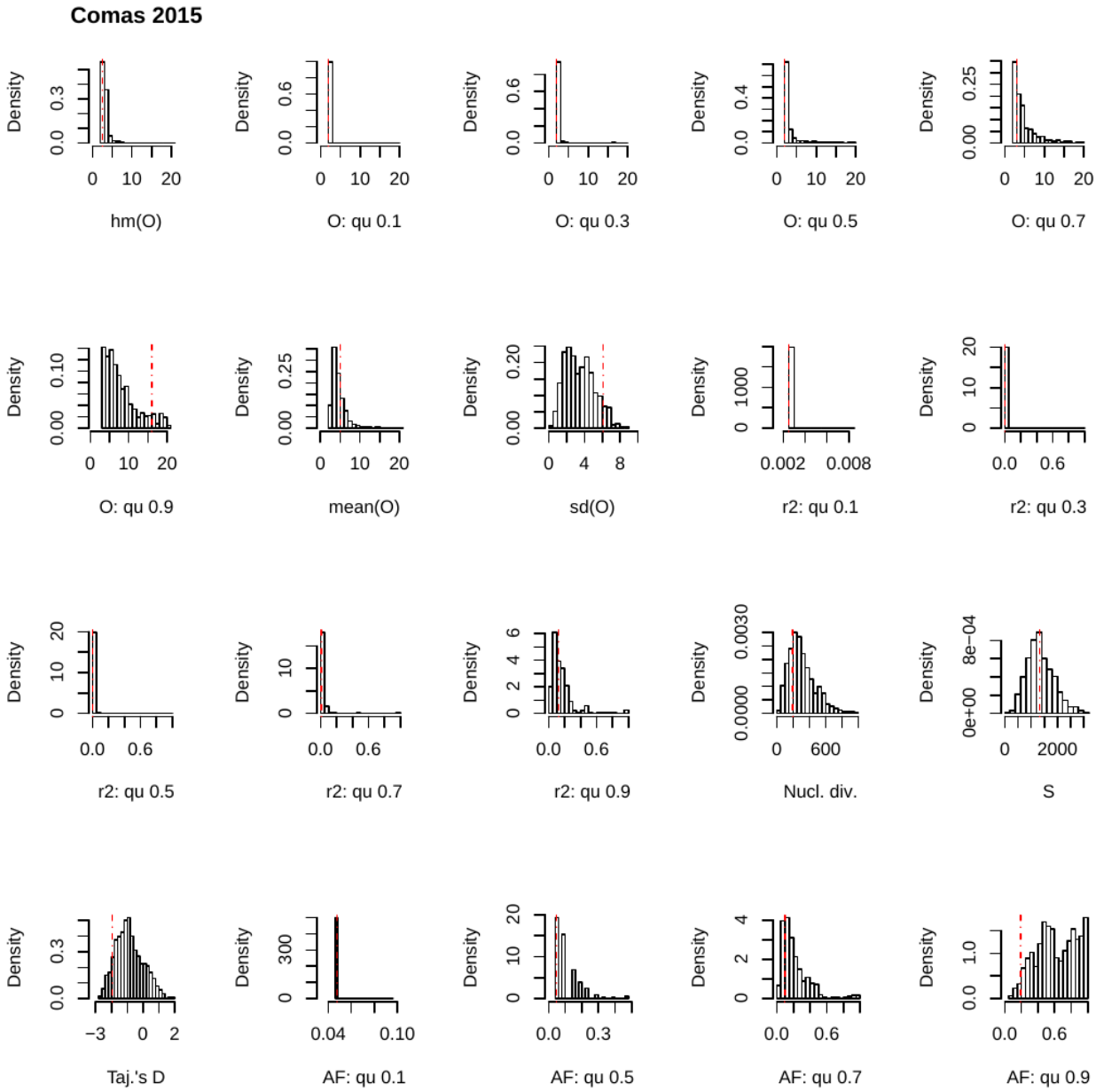
Supplementary Figures



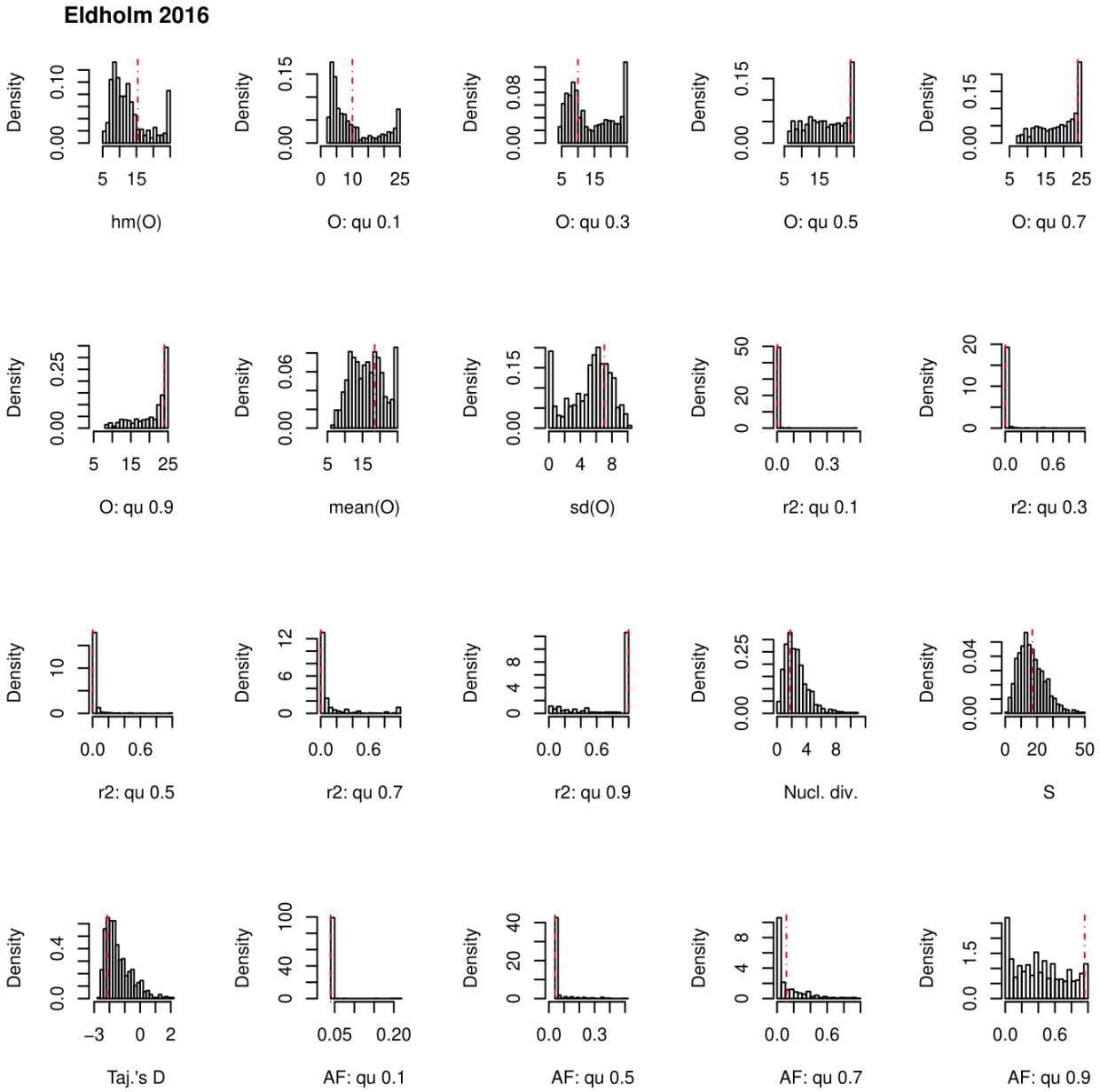
Supplementary Figure 1. Posterior predictive check for the data set Bainomugisa 2018. The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (BETA) using the median of the posterior distribution of the parameter α, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



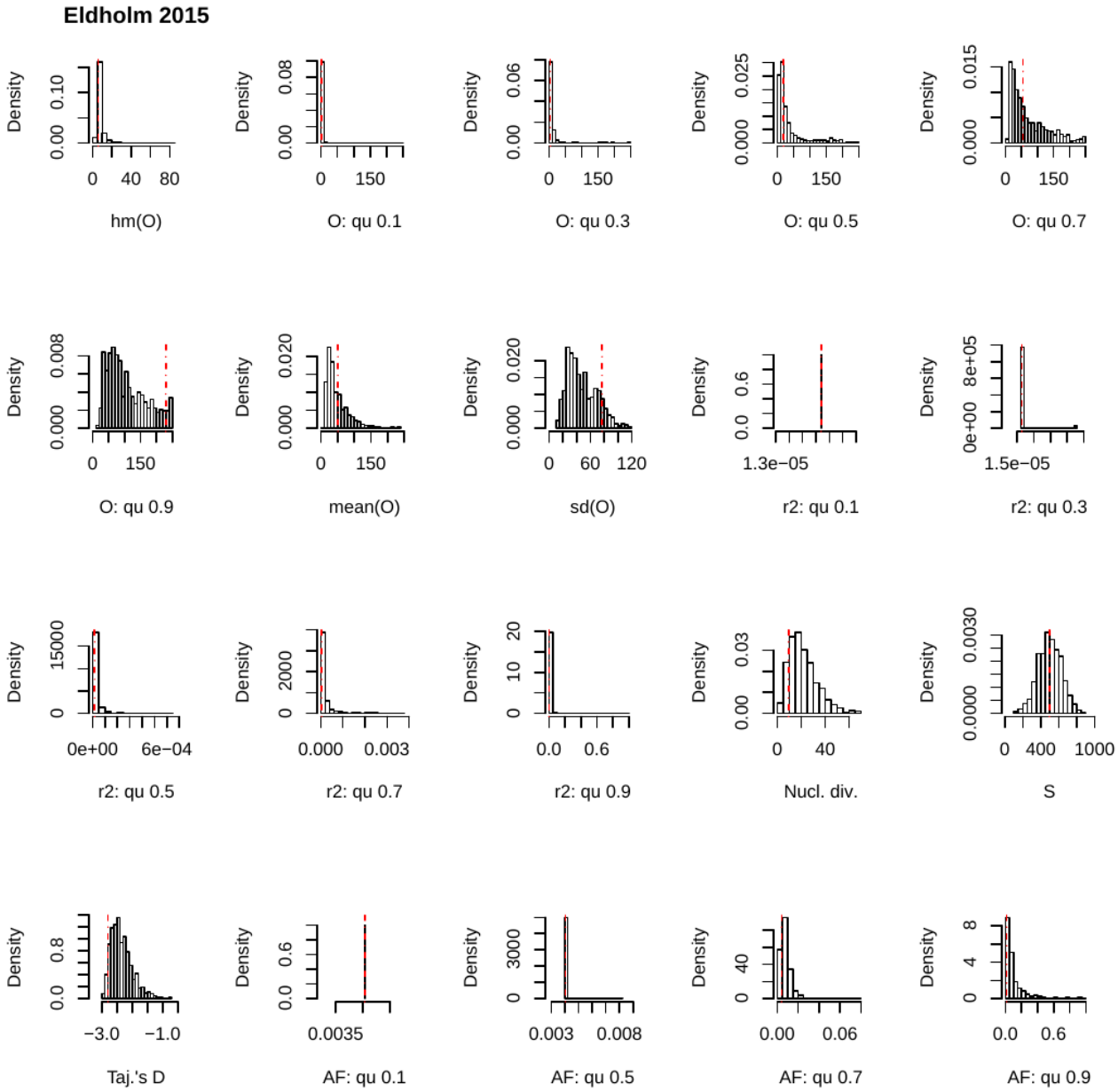
Supplementary Figure 2. Posterior predictive check for the data set Bjorn-Mortensen 2016. The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (BETA) using the median of the posterior distribution of the parameter α, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



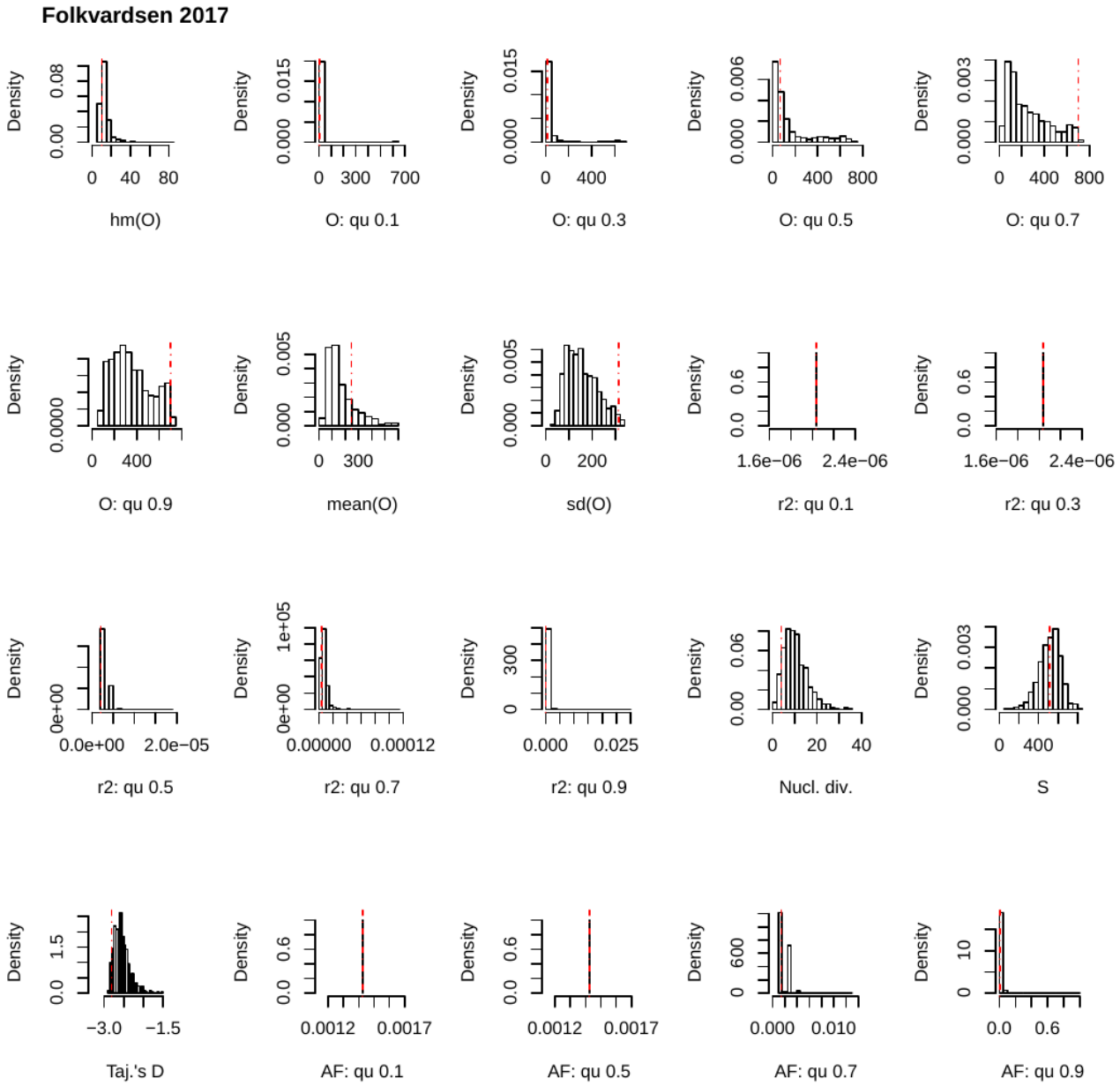
Supplementary Figure 3. Posterior predictive check for the data set Comas 2015. The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (BETA) using the median of the posterior distribution of the parameter α, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



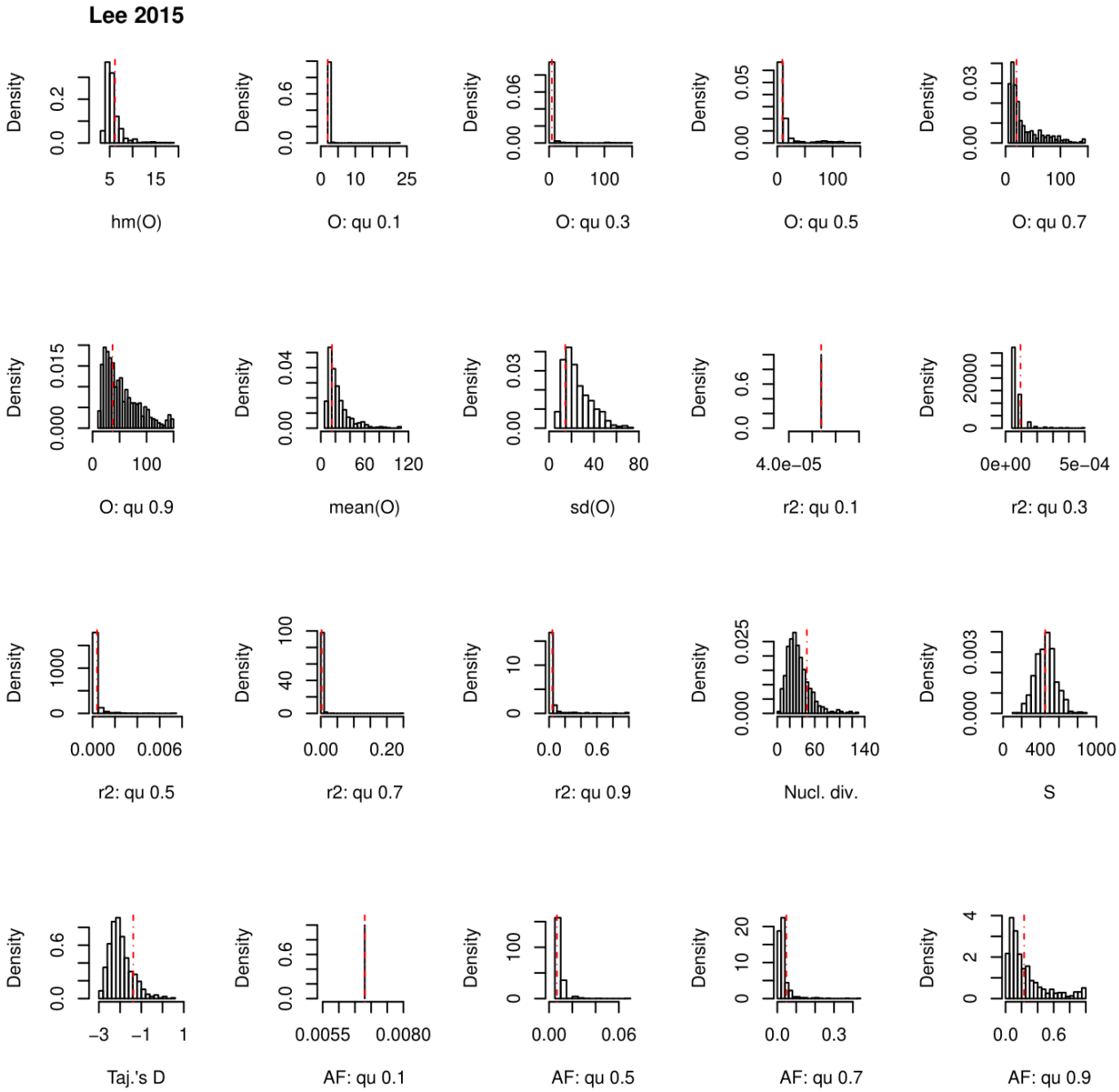
Supplementary Figure 4. Posterior predictive check for the data set Eldholm 2016. The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (Dirac) using the median of the posterior distribution of the parameter ψ, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



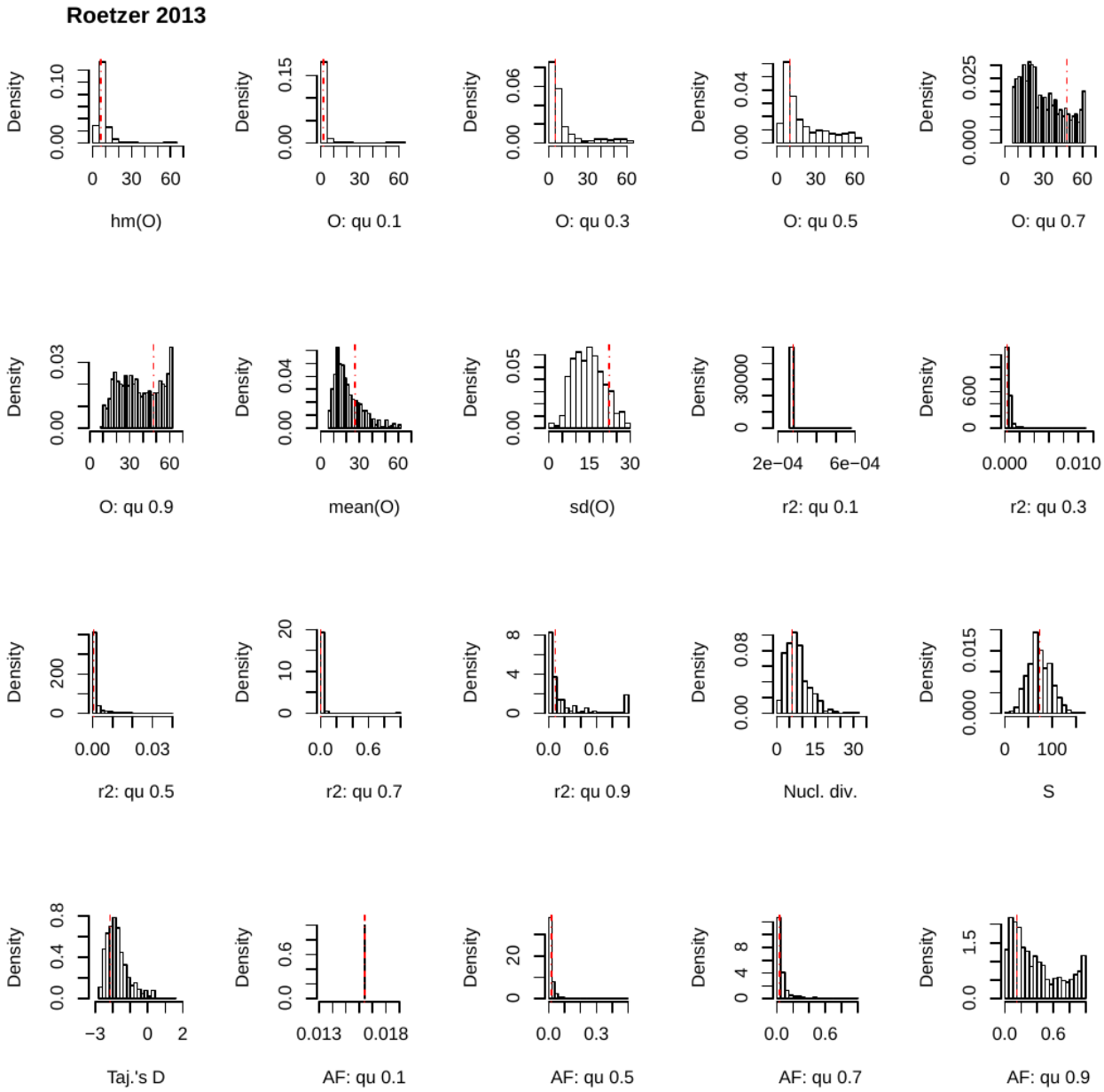
Supplementary Figure 5. Posterior predictive check for the data set Eldholm 2015. The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (BETA) using the median of the posterior distribution of the parameter α, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



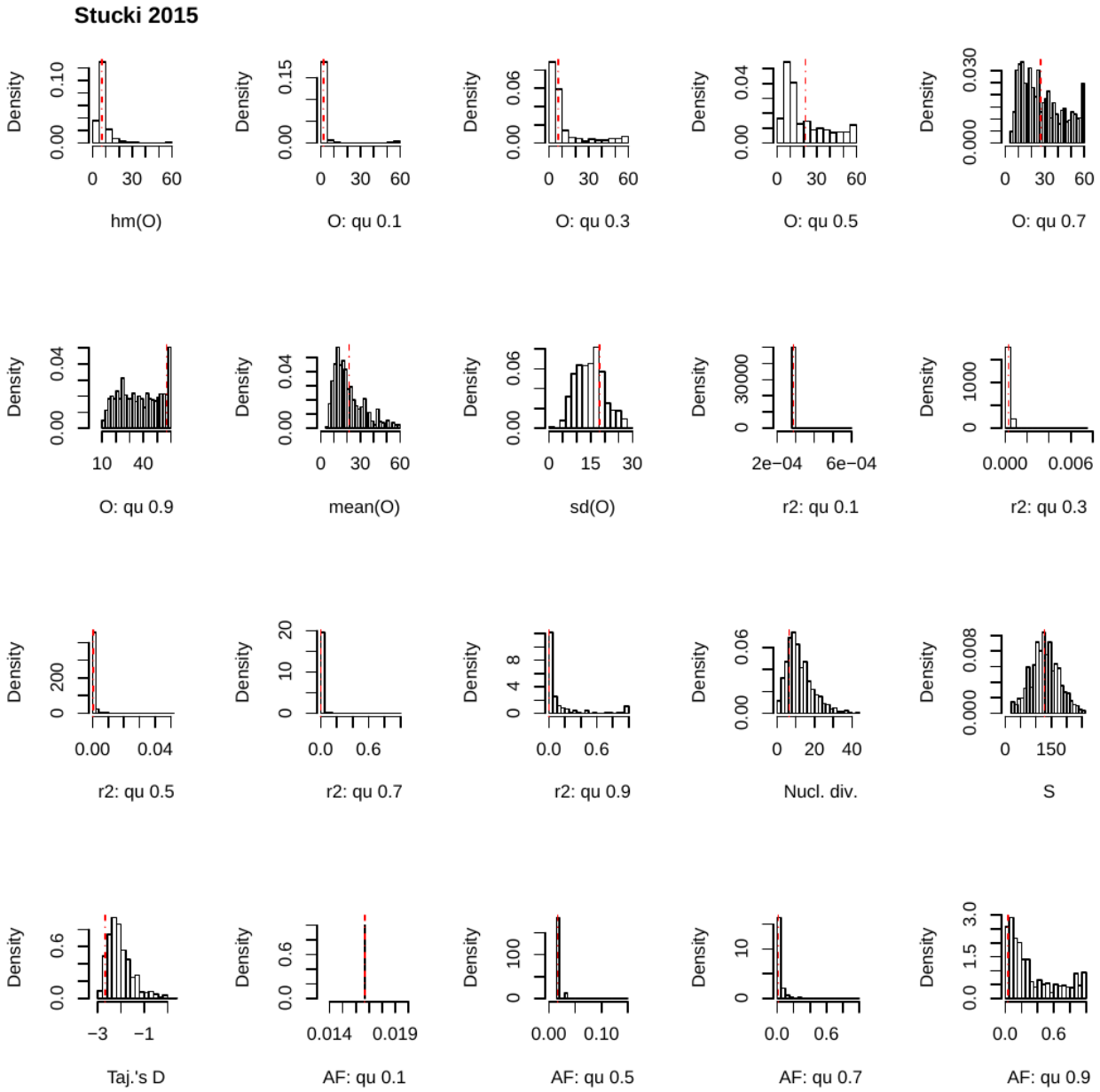
Supplementary Figure 6. Posterior predictive check for the data set Folkvardsen 2017. The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (BETA) using the median of the posterior distribution of the parameter α, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



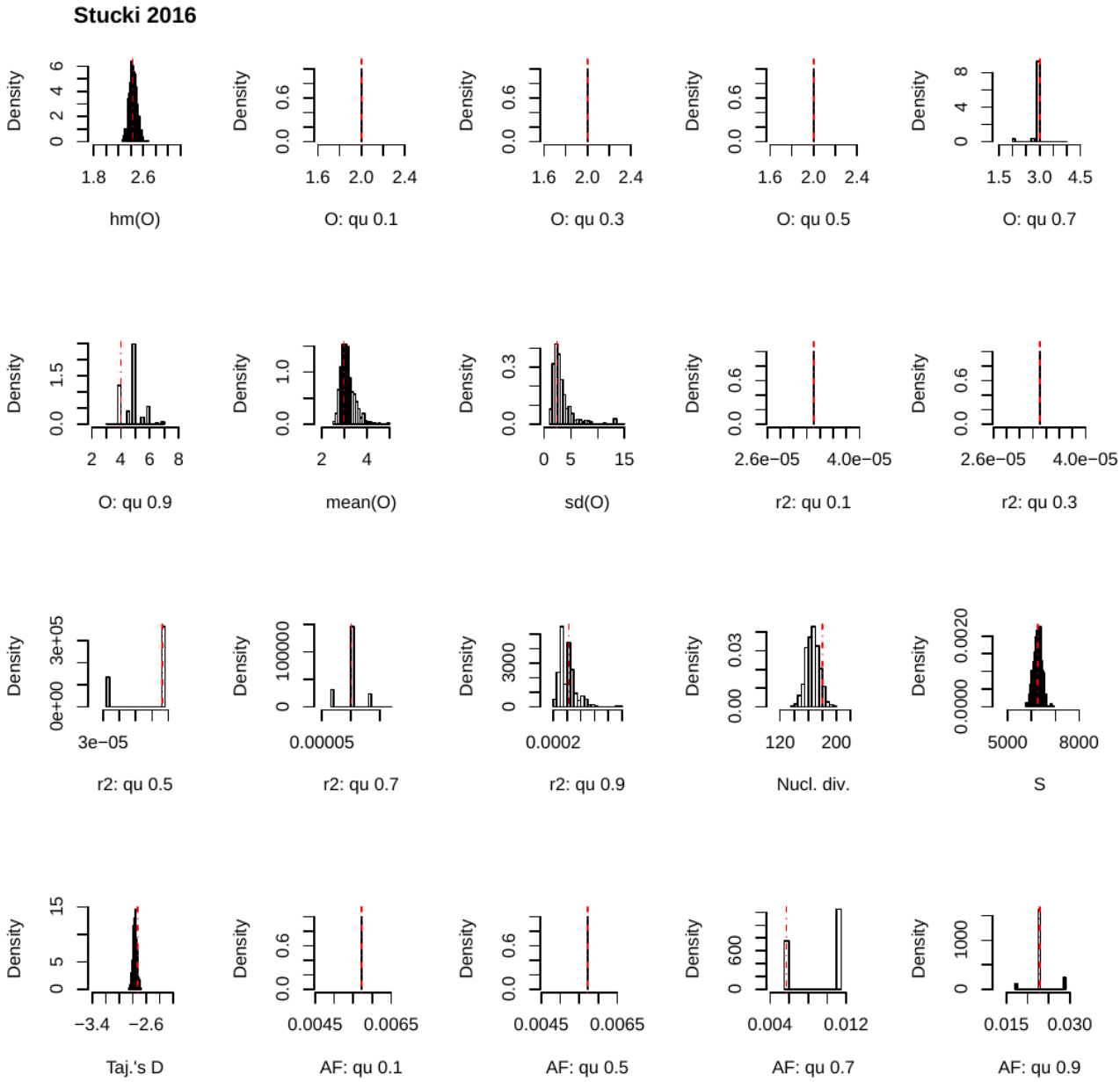
Supplementary Figure 7. Posterior predictive check for the data set Lee 2015. The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (BETA) using the median of the posterior distribution of the parameter α, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



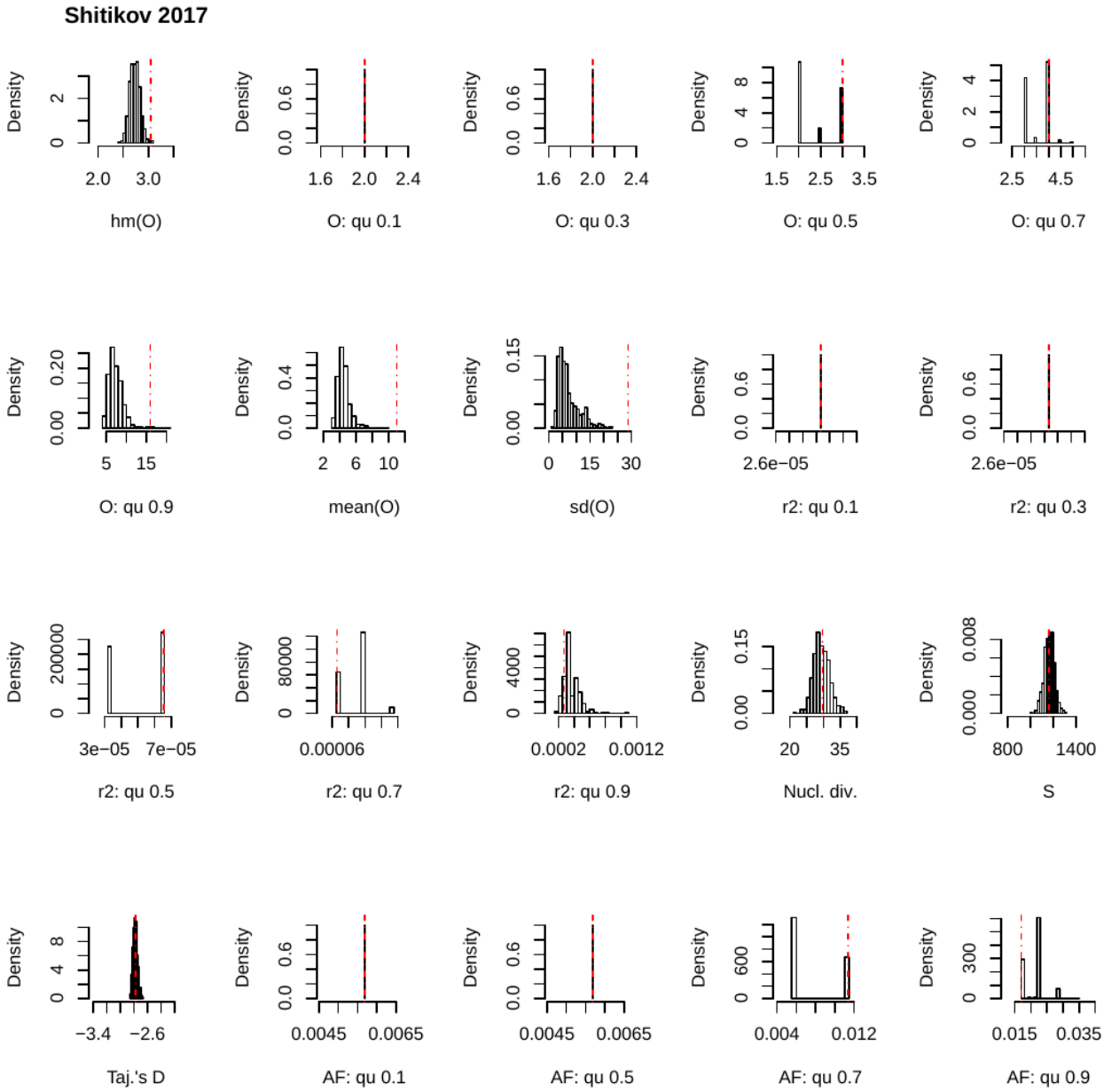
Supplementary Figure 8. Posterior predictive check for the data set Roetzer 2013. The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (BETA) using the median of the posterior distribution of the parameter α, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



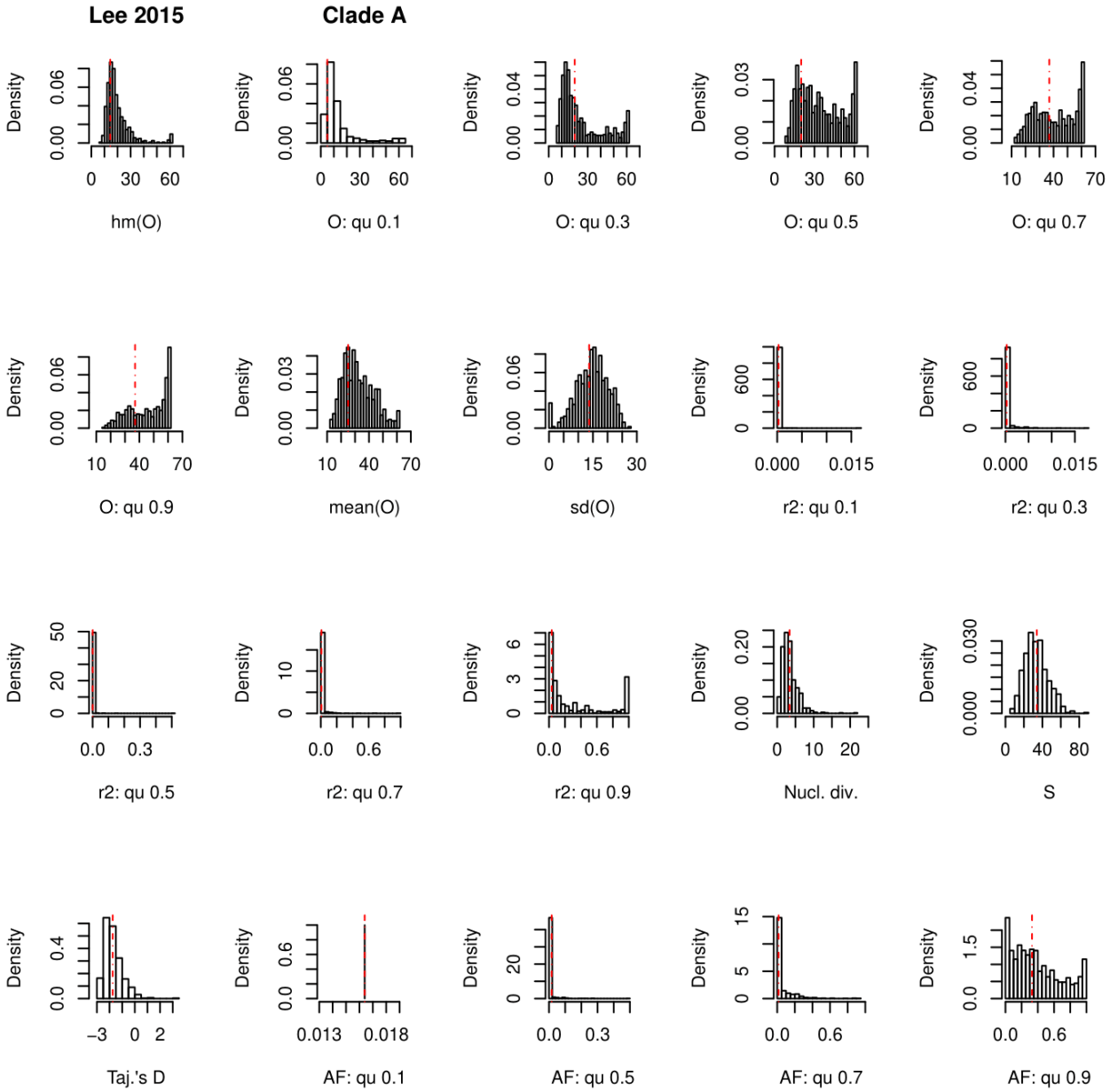
Supplementary Figure 9. Posterior predictive check for the data set Stucki 2015. The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (BETA) using the median of the posterior distribution of the parameter α, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



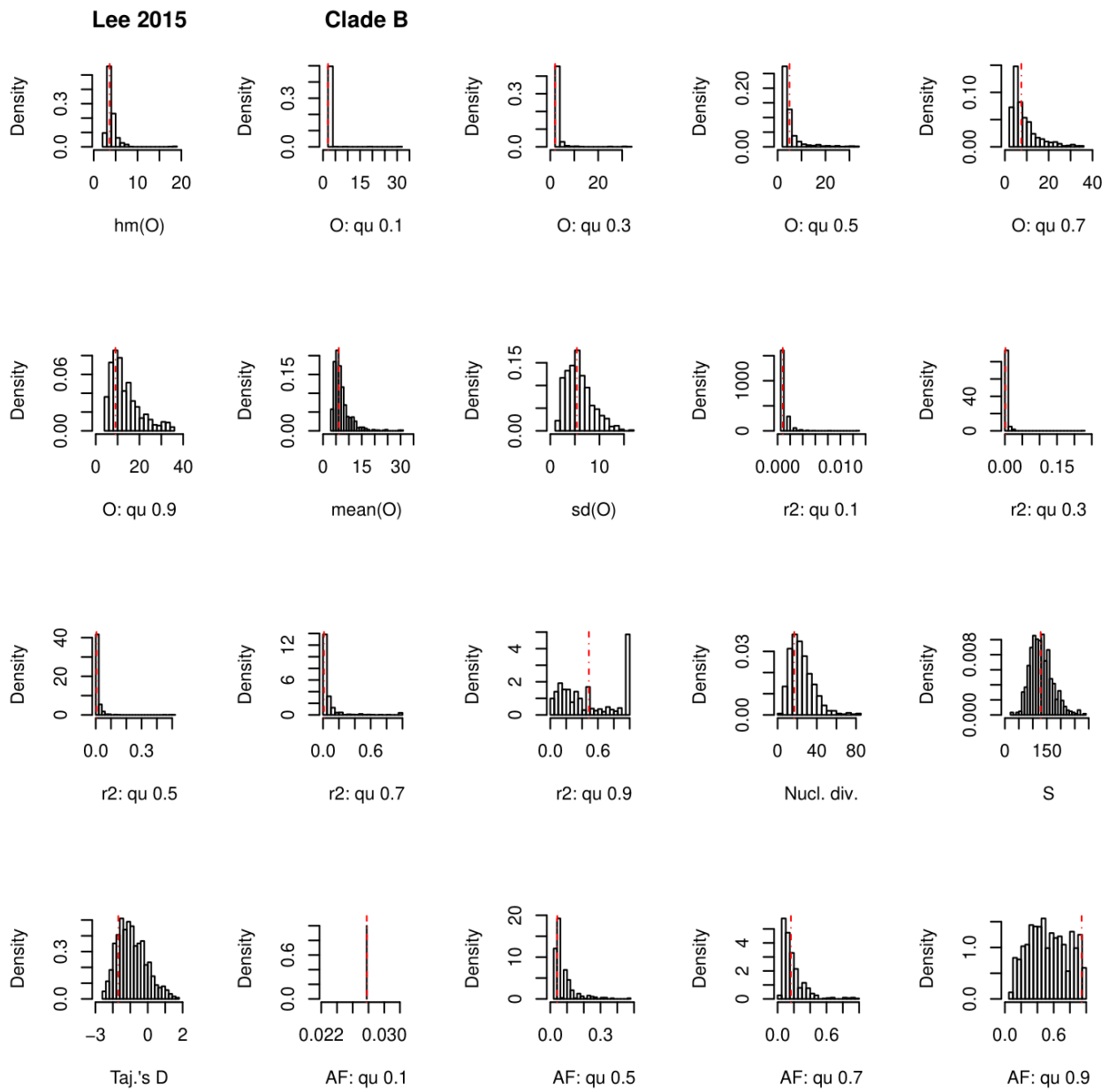
Supplementary Figure 10. Posterior predictive check for the data set Stucki 2016. The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (KM+exp) using the median of the posterior distribution of the parameter g, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectru



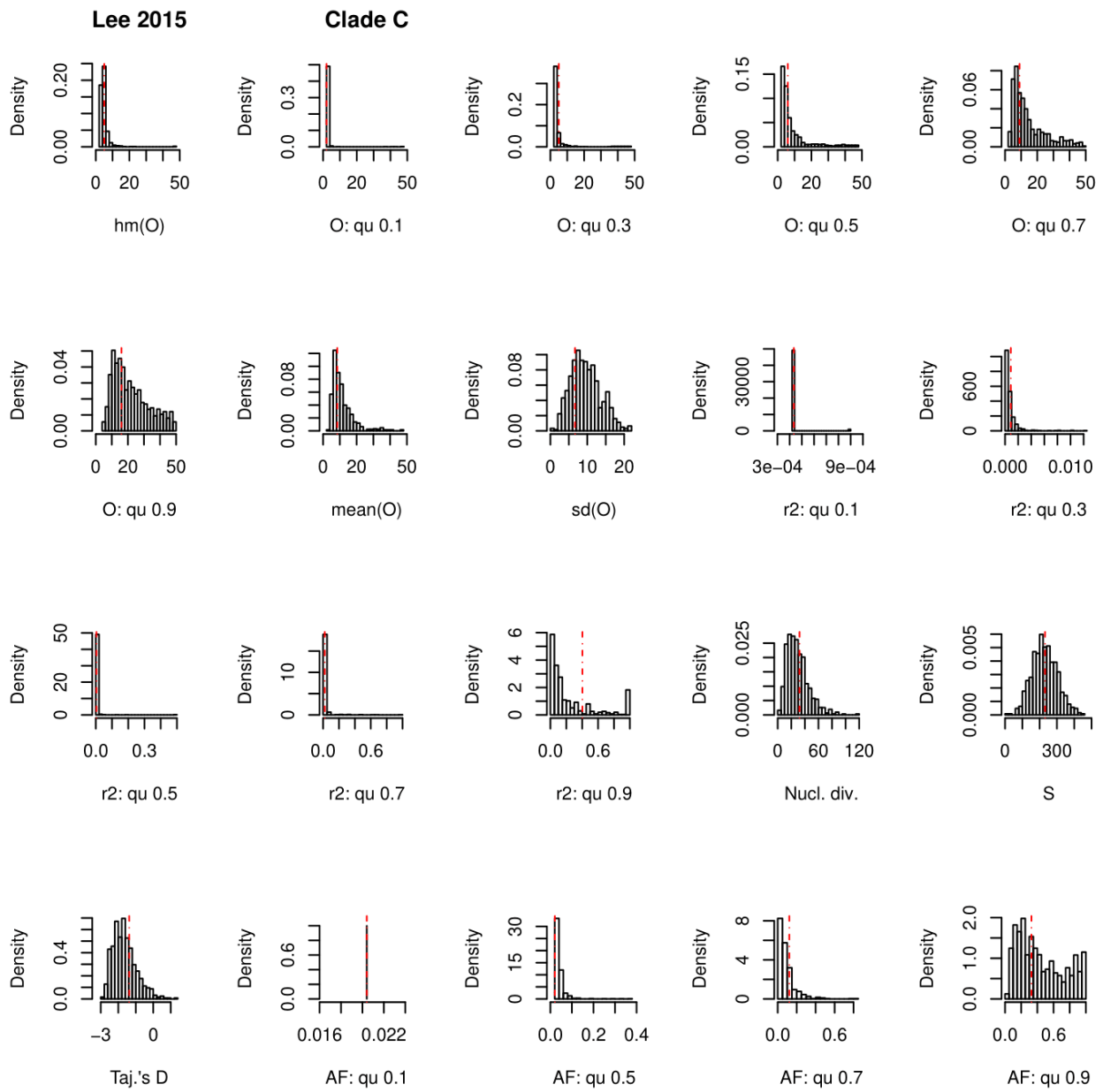
Supplementary Figure 11. Posterior predictive check for the data set Shitikov 2017. The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (KM+exp) using the median of the posterior distribution of the parameter g, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



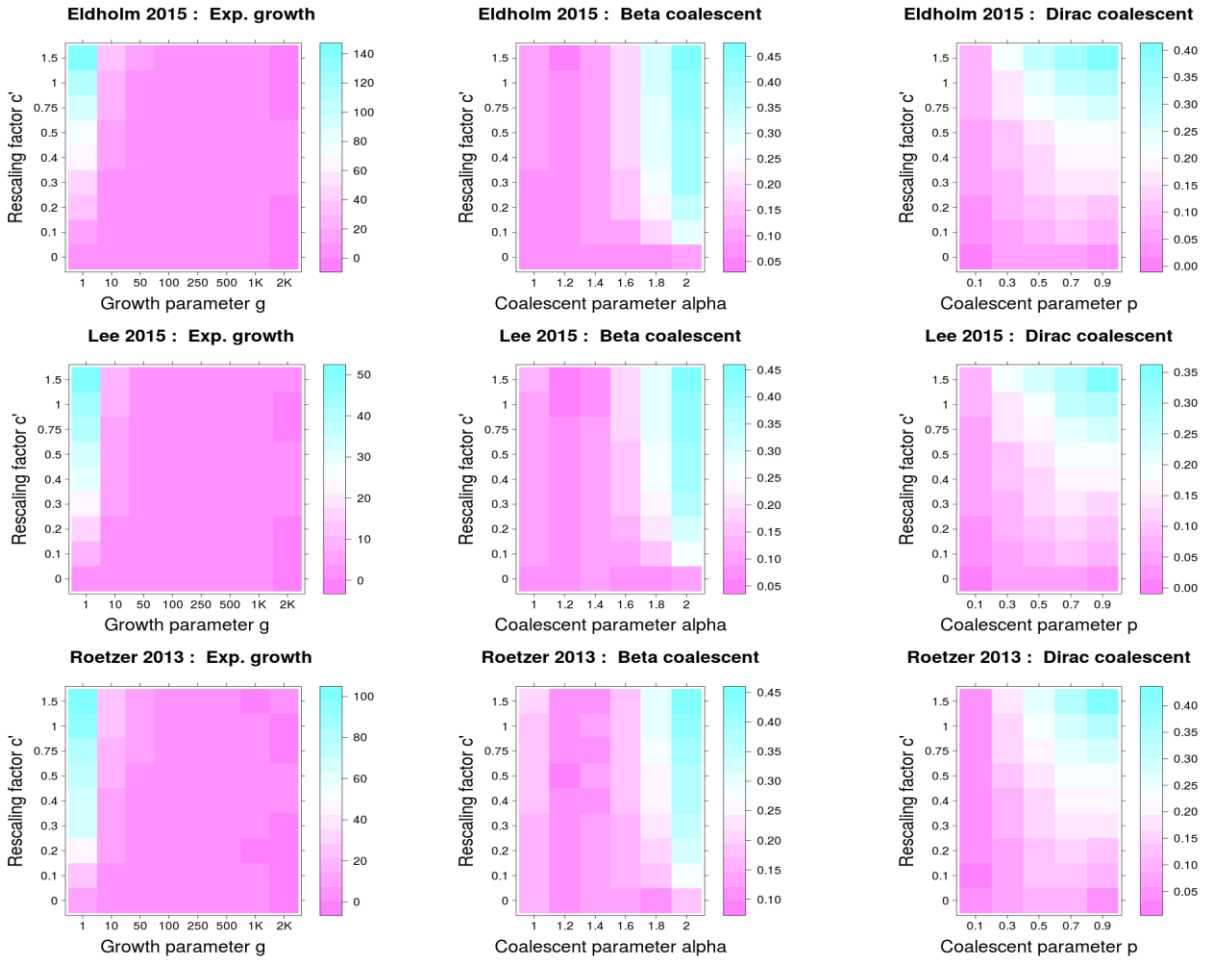
Supplementary Figure 12. Posterior predictive check for the data set Lee 2015 clade A. The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (Dirac) using the median of the posterior distribution of the parameter ψ, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



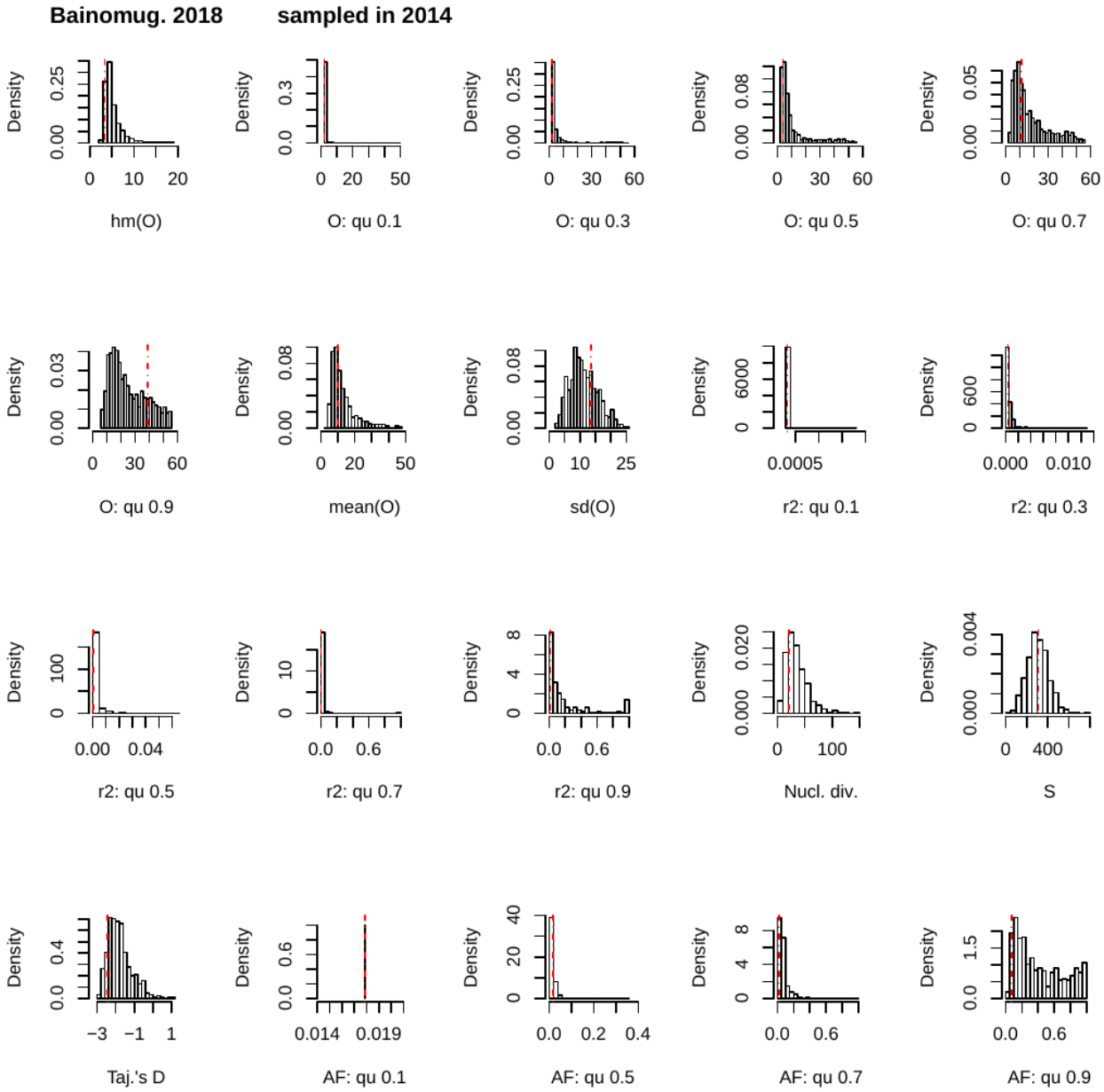
Supplementary Figure 13. Posterior predictive check for the data set Lee 2015 clade B. The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (BETA) using the median of the posterior distribution of the parameter α, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



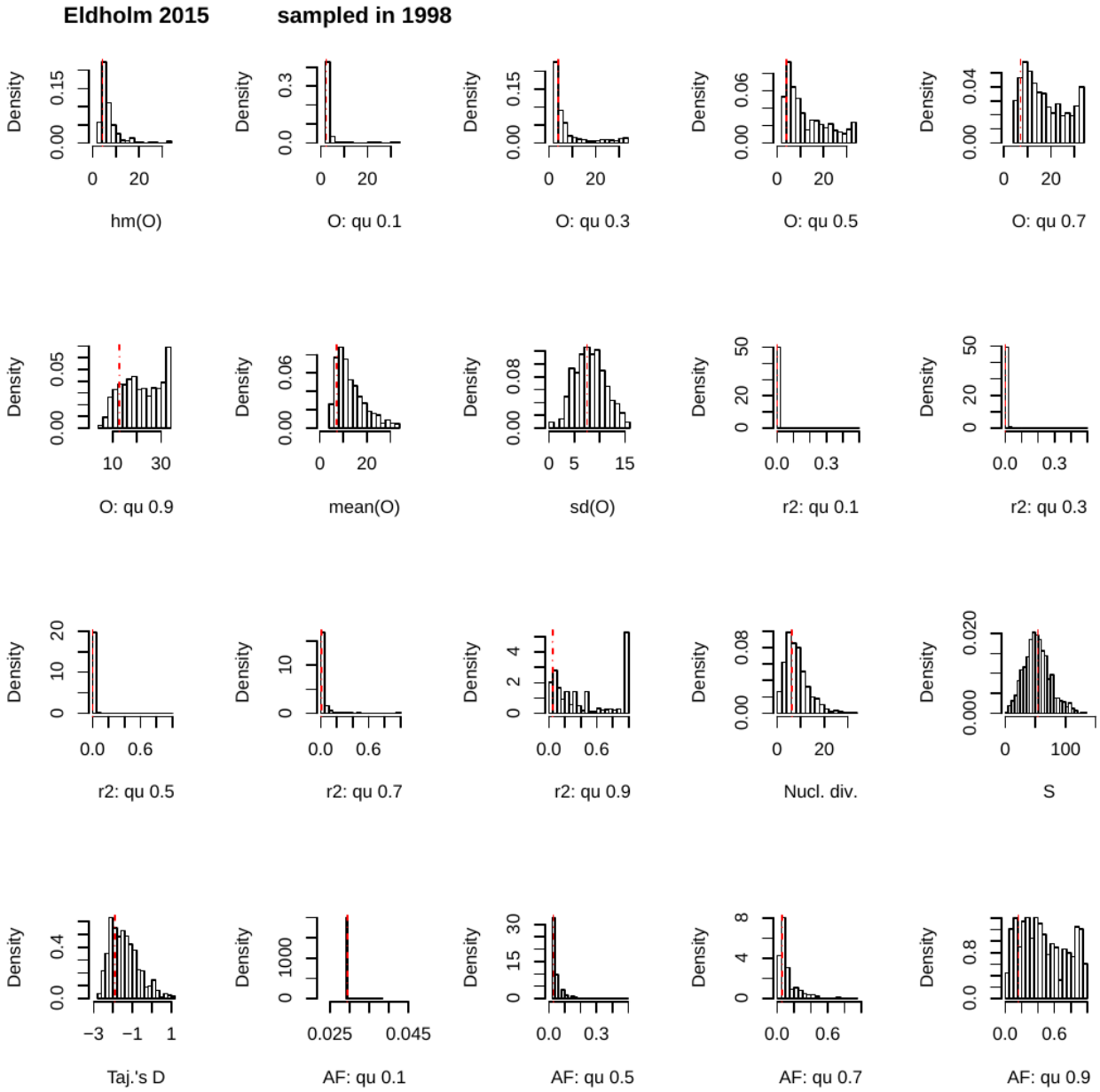
Supplementary Figure 14. Posterior predictive check for the data set Lee 2015 clade C. The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (BETA) using the median of the posterior distribution of the parameter α, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



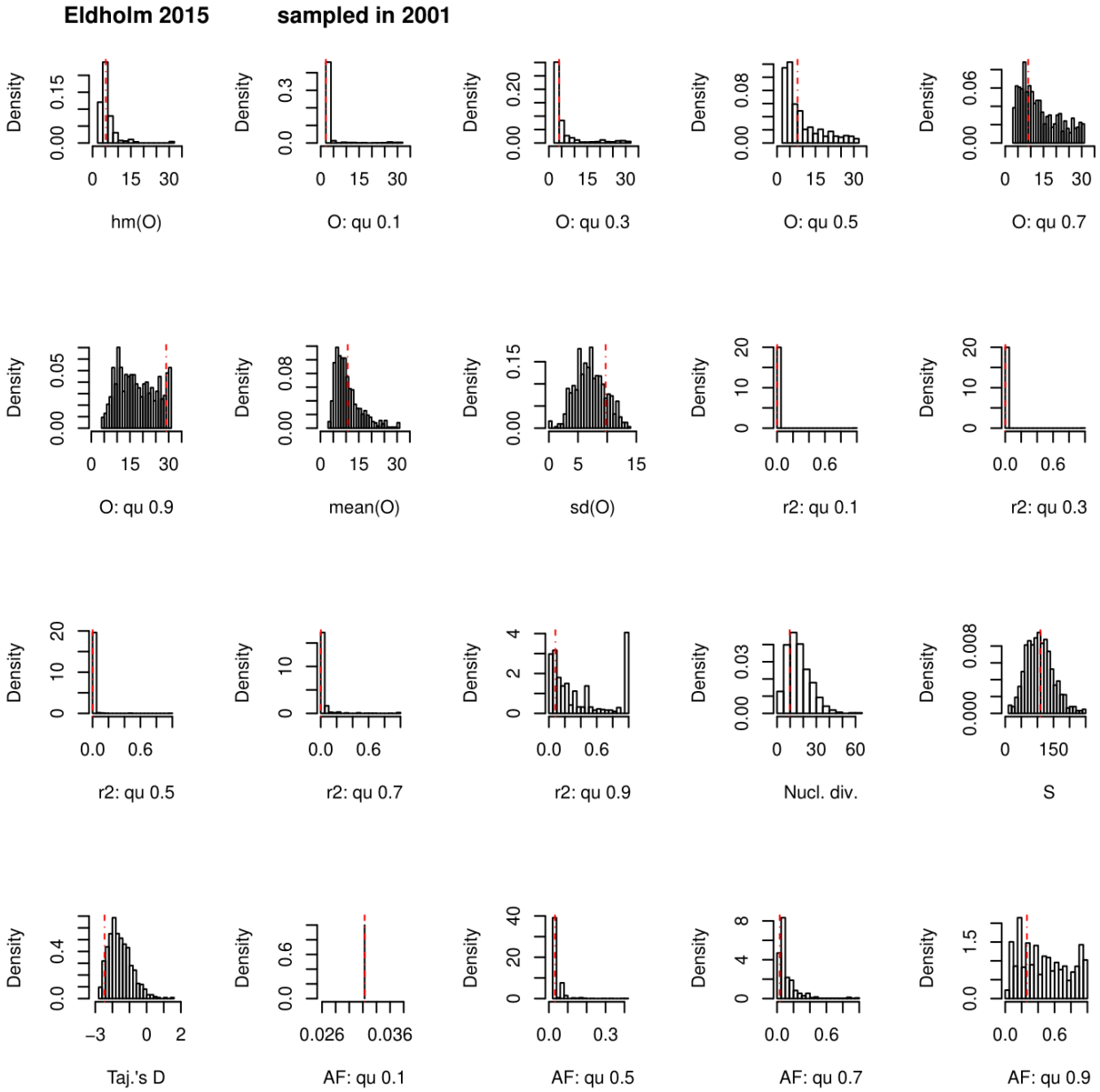
Supplementary Figure 15. Mean error of parameter estimation for serial simulations when model selection was performed via ABC using ultrametric tree models. First column (simulations under serially sampled Kingman's coalescent with exponential growth): colors show the absolute error in units of the true parameter, i.e. an error value of 10 corresponds to an average error of 10x the true parameter. Second and third column (serially sampled MMCs): colors show absolute error.



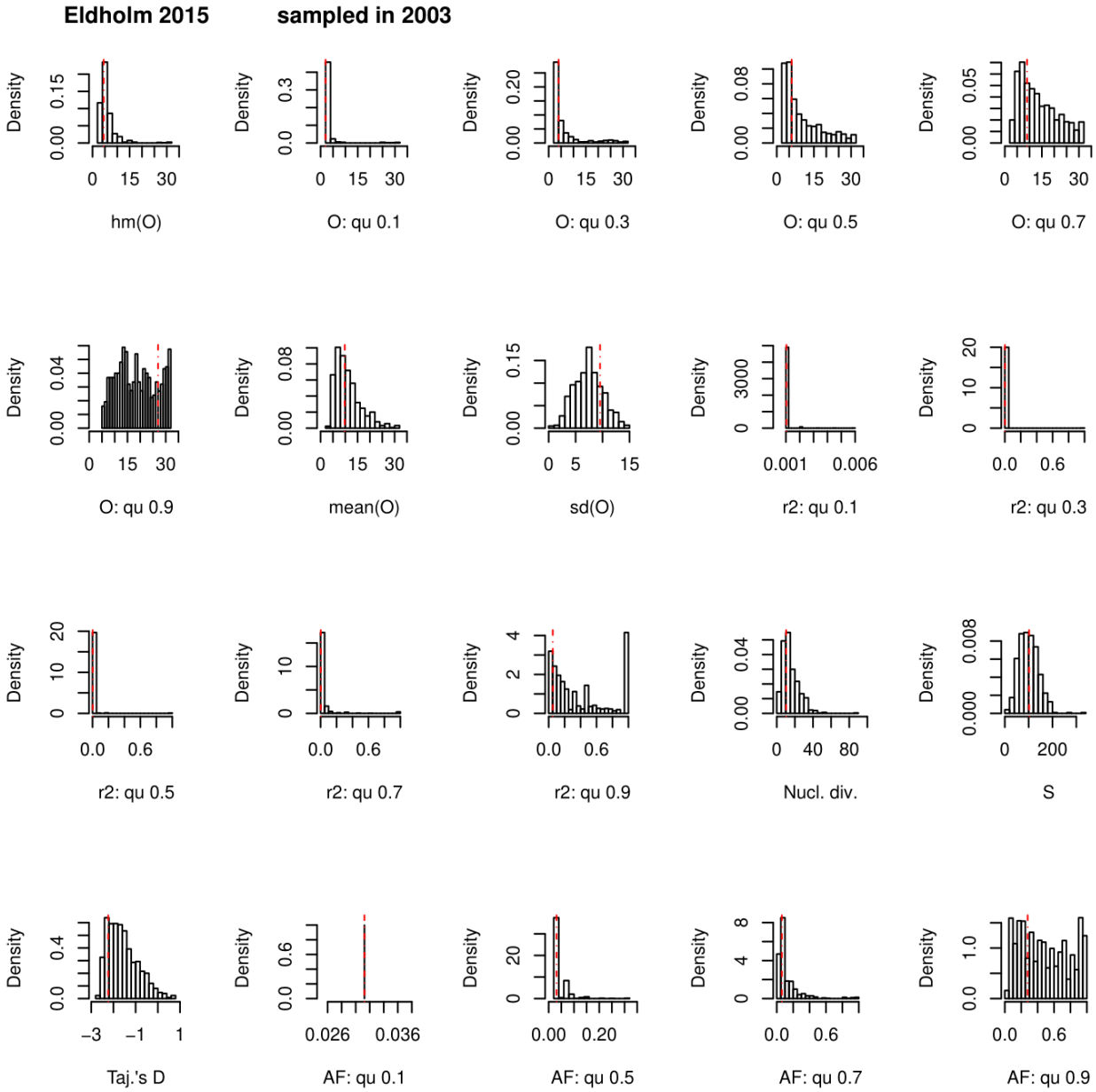
Supplementary Figure 16. Posterior predictive check for the data set Bainomugisa 2018 (2014). The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (BETA) using the median of the posterior distribution of the parameter α, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



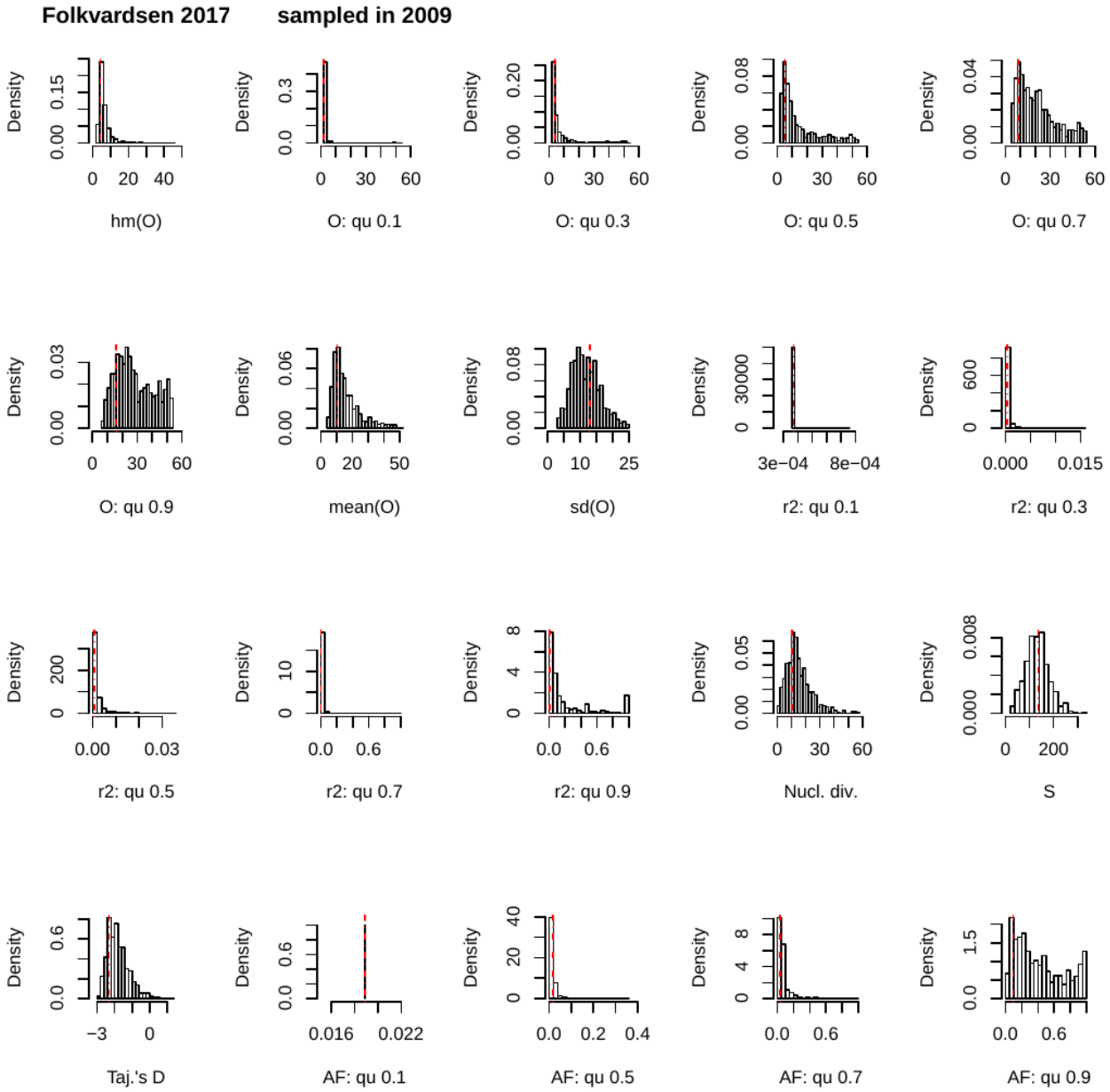
Supplementary Figure 17. Posterior predictive check for the data set Eldholm 2015 (1998). The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (BETA) using the median of the posterior distribution of the parameter α, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



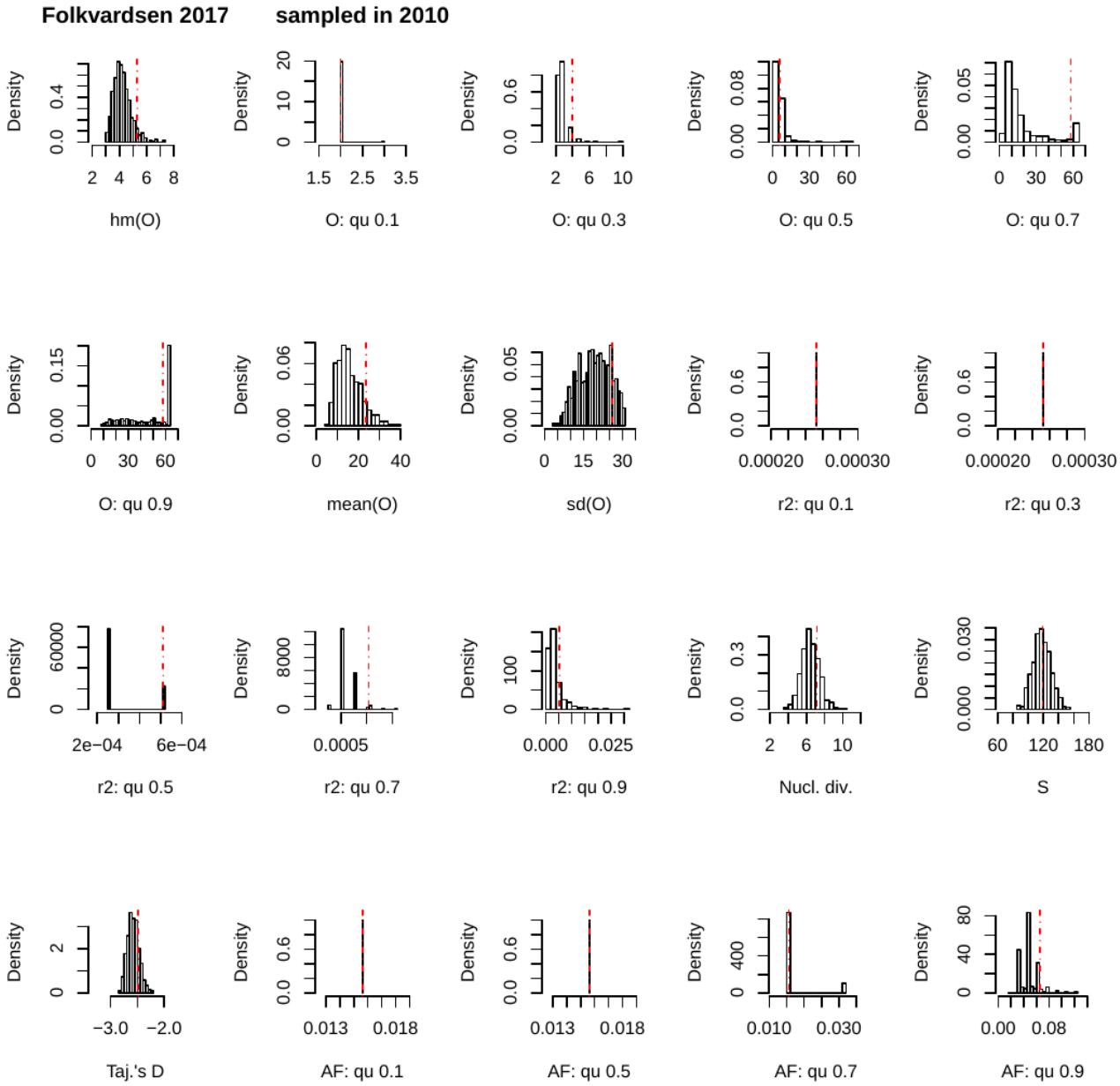
Supplementary Figure 18. Posterior predictive check for the data set Eldholm 2015 (2001). The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (BETA) using the median of the posterior distribution of the parameter α, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



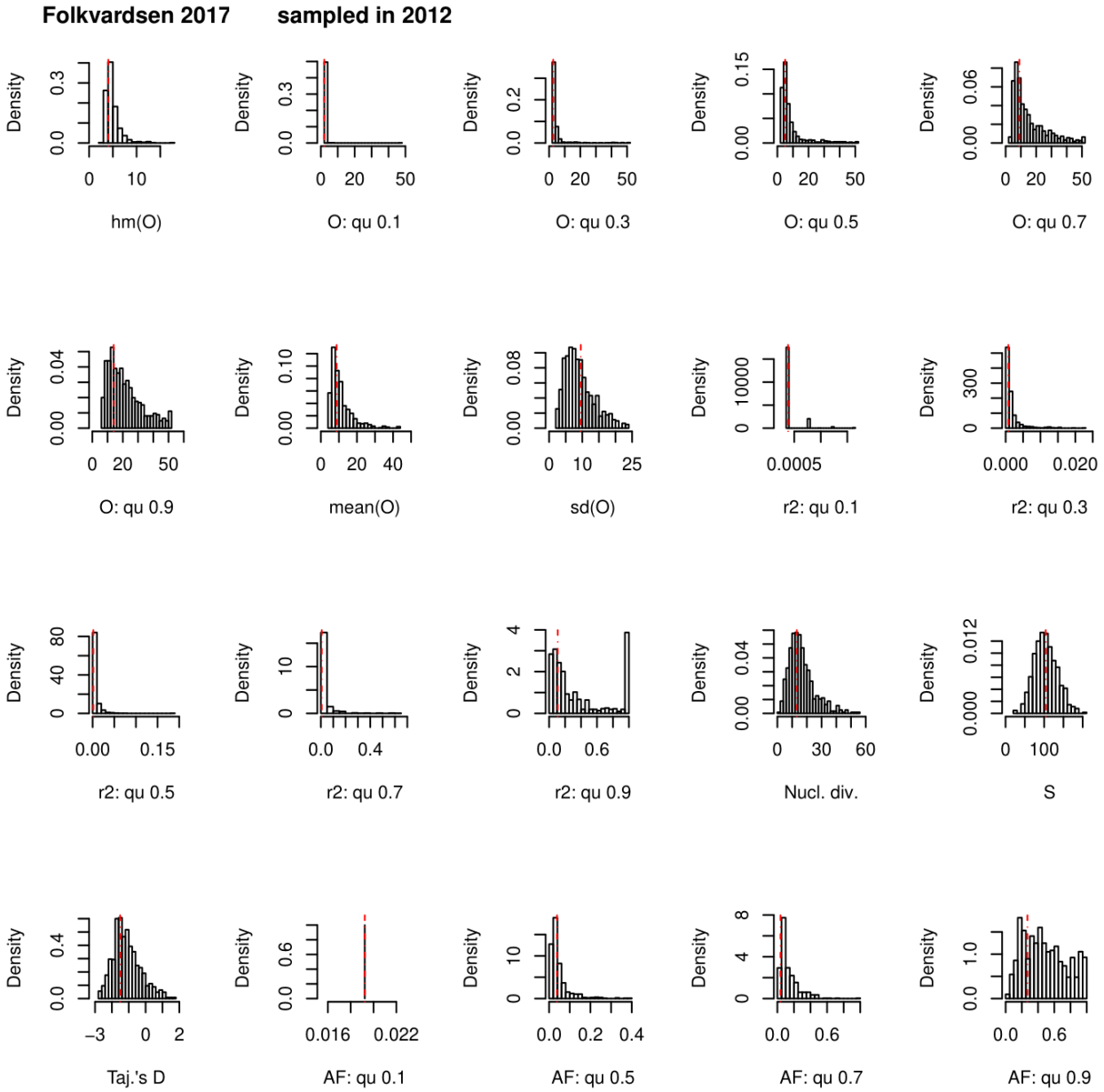
Supplementary Figure 19. Posterior predictive check for the data set Eldholm 2015 (2003). The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (BETA) using the median of the posterior distribution of the parameter α, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



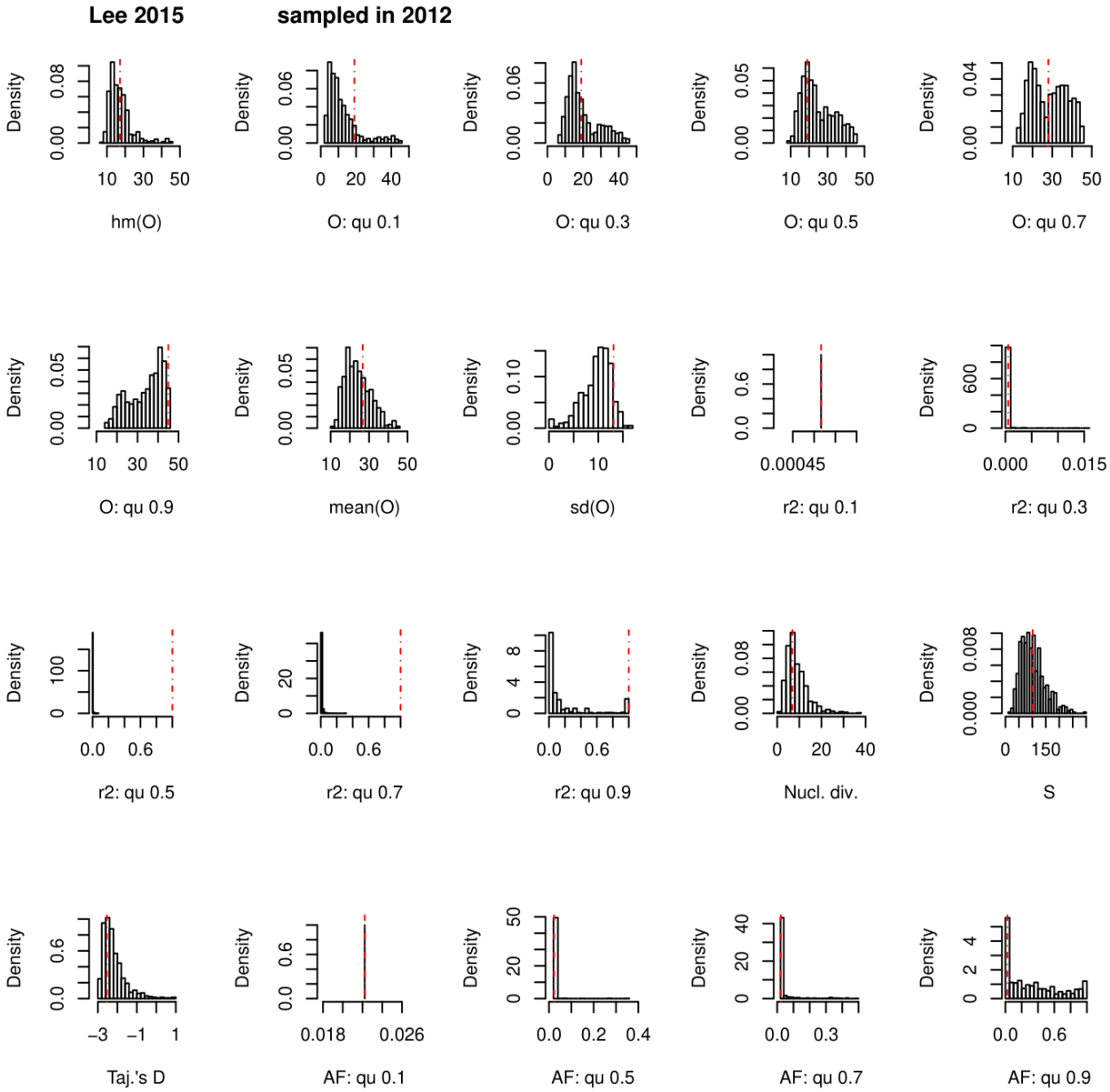
Supplementary Figure 20. Posterior predictive check for the data set Folkvardsen 2017 (2009). The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (BETA) using the median of the posterior distribution of the parameter α, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



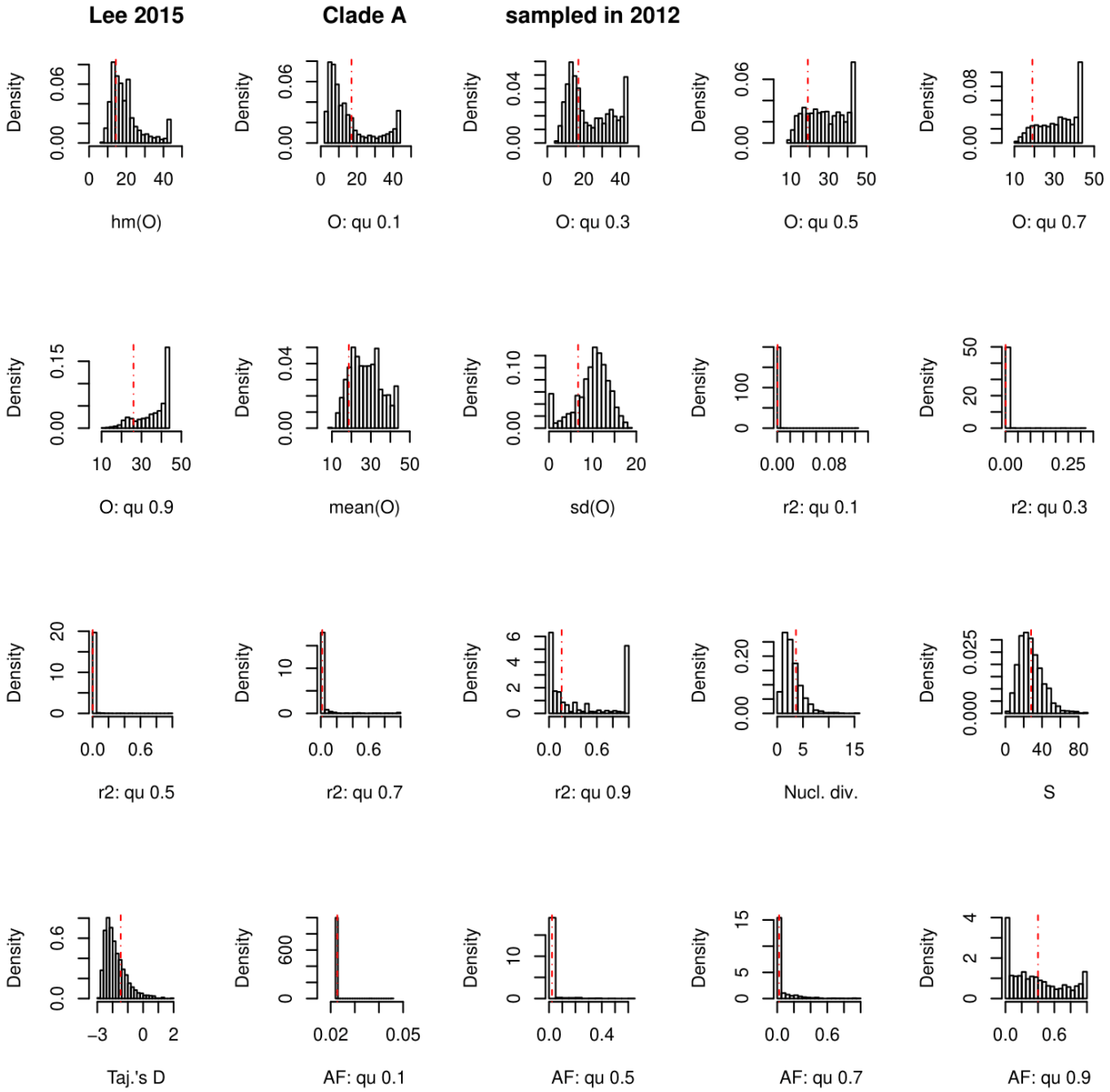
Supplementary Figure 21. Posterior predictive check for the data set Folvardsen 2017 (2010). The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (KM+exp) using the median of the posterior distribution of the parameter g, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



Supplementary Figure 22. Posterior predictive check for the data set Folkvardsen 2017 (2012). The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (BETA) using the median of the posterior distribution of the parameter α, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



Supplementary Figure 23. Posterior predictive check for the data set Lee 2015 (2012). The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (Dirac) using the median of the posterior distribution of the parameter ψ, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.



Supplementary Figure 24. Posterior predictive check for the data set Lee 2015 (2012) Clade A. The red line represents the observed data, the histograms represent the results of 10,000 simulations under the best fitting model (Dirac) using the median of the posterior distribution of the parameter ψ, averaged over the three replications. O: quantiles of the minimal observable clade size; r2: quantiles of the r-squared measure of linkage disequilibrium; Nucl. Div.: nucleotide diversity (π); S: number of polymorphic positions; Taj's D: Tajima's D; AF: quantiles of the mutant allele frequency spectrum.

Supplementary Tables

Supplementary Table 1:

List of accession numbers used in this study

File: Supplementary_table1.xslx

Supplementary Table 2:

Results of all analyses

File: Supplementary_table2.xslx

Supplementary Table 3.

Sampling period and estimated age of the most recent common ancestor for the data sets that resulted in BETA or Dirac as best fitting model. For Comas 2015 this data was not available.

Data set	**Sampling window in years**	**Age of the tree, in years before most recent sample[1]**
Eldholm 2015	14	~ 40
Lee 2015	22	~ 100
Roetzer 2013	14	~ 15
Bainomugisa 2018	4	~ 50
Bjorn-Mortensen 2016	21	~ 25
Folkvardsen 2017	23	~ 55
Stucki 2015	21	NA
Eldholm 2016	6	~ 10

[1] The estimated age of the most recent common ancestors were obtained from the original publications