TRAILS: tree reconstruction of ancestry using incomplete lineage sorting

IKER RIVAS-GONZÁLEZ[1,*], MIKKEL H SCHIERUP[1],
JOHN WAKELEY[2], ASGER HOBOLTH[3]

[1]Bioinformatics Research Center (BiRC), Aarhus University

[2]Department of Organismic and Evolutionary Biology, Harvard University

[3]Department of Mathematics, Aarhus University

[*]Corresponding author (irg@birc.au.dk)

ABSTRACT

Genome-wide genealogies of multiple species carry detailed information about demographic and selection processes on individual branches of the phylogeny. Here, we introduce TRAILS, a hidden Markov model that accurately infers time-resolved population genetics parameters, such as ancestral effective population sizes and speciation times, for ancestral branches using a multi-species alignment of three species and an outgroup. TRAILS leverages the information contained in incomplete lineage sorting fragments by modelling genealogies along the genome as rooted three-leaved trees, each with a topology and two coalescent events happening in discretized time intervals within the phylogeny. Posterior decoding of the hidden Markov model can be used to infer the ancestral recombination graph for the alignment and details on demographic changes within a branch. Since TRAILS performs posterior decoding at the base-pair level, genome-wide scans based on the posterior probabilities can be devised to detect deviations from neutrality. Using TRAILS on a human-chimp-gorilla-orangutan alignment, we recover speciation parameters and extract information about the topology and coalescent times at high resolution.

AUTHOR SUMMARY

DNA sequences can be compared to reconstruct the evolutionary history of different species. While the ancestral history is usually represented by a single phylogenetic tree, speciation is a more complex process, and, due to the effect of recombination, different parts of the genome might follow different genealogies. For example, even though humans are more closely related to chimps than to gorillas, around 15% of our genome is more similar to the gorilla genome than to the chimp one. Even for those parts of the genome that do follow the same human-chimp topology, we might encounter a last

common ancestor at different time points in the past for different genomic fragments. Here, we present TRAILS, a new framework that utilizes the information contained in all these genealogies to reconstruct the speciation process. TRAILS infers unbiased estimates of the speciation times and the ancestral effective population sizes, improving the accuracy when compared to previous methods. TRAILS also reconstructs the genealogy at the highest resolution, inferring, for example, when common ancestry was found for different parts of the genome. This information can also be used to detect deviations from neutrality, effectively inferring natural selection that happened millions of years ago. We validate the method using extensive simulations, and we apply TRAILS to a human-chimp-gorilla multiple genome alignment, from where we recover speciation parameters that are in good agreement with previous estimates.

Introduction

Orthologous sites in two or more sequences share a unique genealogical history, with coalescent events happening at certain time points in the past. In the absence of recombination, all sites along the sequences follow the same genealogy. In reality, however, ancestral recombination events might have decoupled consecutive sites, generating an array of segments with different yet correlated genealogies, collectively known as the ancestral recombination graph (ARG) [1, 2]. In principle, if inferred accurately, the ARG contains all available information about the demography of the samples, and it can be used to estimate population parameters (such as the recombination rate and the ancestral effective population sizes), historical events (such as introgression and hybridization), and selective processes [3]. The ARG, however, is challenging to infer because the underlying genealogies along the genome alignment cannot be directly observed. Instead, inference of the genealogy along the genome relies on the site patterns of the accumulated mutations.

The ARG can also be formulated as a spatial process along the genomic alignment [4]. This process, however, contains a long-range correlation structure because if two recombination events happen flanking a genomic fragment, the fragment might be surrounded by the exact same genealogy. However, disregarding the fact that the process is non-Markovian in nature, the ARG can be approximated by a hidden Markov model (HMM), where the genealogy of a certain genomic position only depends on the genealogy of the previous position [5, 6]. It has been shown that this approach, commonly referred to as sequentially Markovian coalescent or SMC, is a good approximation of the true coalescent-with-recombination process [7]. Perhaps the simplest of such models is the pairwise sequentially Markovian coalescent (PSMC) [8], in which the ARG between two sequences (typically, the two copies of a diploid individual) is modelled. Here, the hidden states are coalescent events that happen in

discretized time intervals, which correspond to two-leafed gene trees (fig. 1A). The transition probabilities between pairs of hidden states can be calculated using standard coalescent theory, parameterized by the recombination rate and the ancestral effective population sizes (N_e) in each time interval [8]. PSMC, and other SMCs, such as MSMC [9], MSMC2 [10], ASMC [11], and SMC++ [12], allow the use of standard HMM machinery to infer population parameters, and are thus also useful for inferring the most plausible coalescent times from the posterior decoding. However, SMC models are generally restricted to a single coalescent event between a pair of samples, which limits their usefulness. More recently, there have been new developments to model multiple samples explicitly. For example, ARGweaver [13], Relate [14], tsinfer+tsdate [15, 16] or ARG-Needle [17] use techniques such as resampling, threading and mathematical approximations to sequentially build the ARG [18].

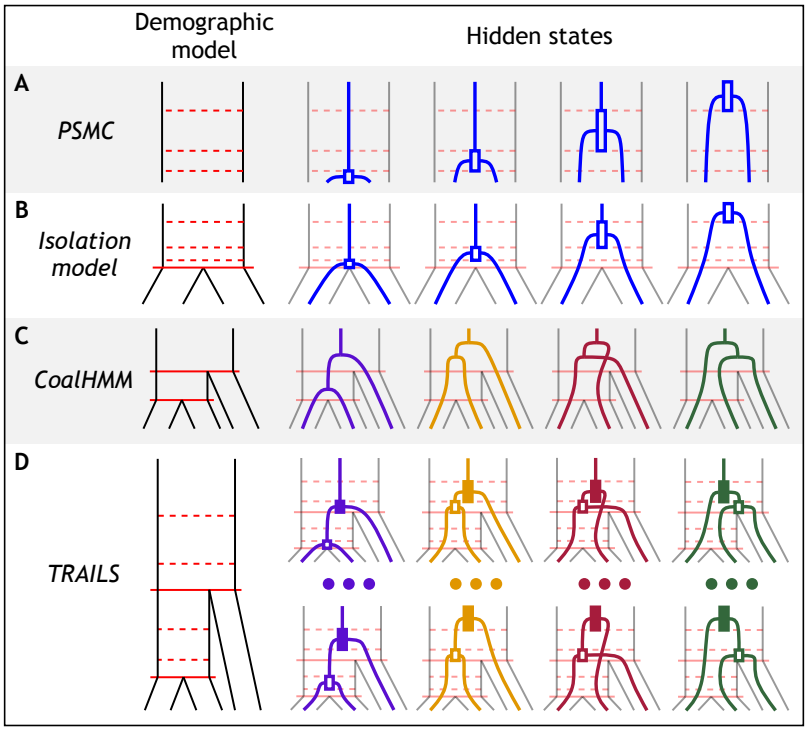


Figure 1: TRAILS is a HMM that reconstructs the time-resolved multi-species ARG for three genomes. TRAILS extends the isolation model (B) [19] to three species, by combining the time discretization of PSMC-like models (A) [8] with the topologically aware hidden states of CoalHMM (C) [20, 21]. The resulting hidden states of TRAILS (D) are three-leaved genealogies with two discretized coalescent events and one of four possible topologies. A full list of the 27 possible hidden states when $n_{\mathrm{AB}} = n_{\mathrm{ABC}} = 3$ can be consulted in fig. S11

These models are typically used to analyze samples from the same species to get within-species information about the ancestral process. Analyzing inter-species coalescent events adds another layer of complexity, since the coalescent events need to be contained within the underlying phylogeny or speciation tree [22, 23, 24]. Moreover, the models described above typically use the presence or absence of a certain mutation to construct haplotypes, but ignore or filter out instances where more than two alleles are observed. This infinite sites model poses a problem for inter-species analysis, because recurrent

mutation is more likely to happen, generating instances of sites that have experienced more than a single mutation [25, 26].

Some other models have tried to extend these concepts for the analyses of multiple species. For example, the coalescent-with-isolation model [19] is conceptually similar to PSMC, but, backwards in time, the two analyzed samples are kept isolated until the speciation event, after which they can coalesce (fig. 1B). This model can be used to estimate the speciation time between the two samples and the N_e of the ancestral species, and an extension of it can be used to model isolation-with-migration [27]. These models, similar to SMCs, can output a posterior decoding of the coalescent times.

Beyond two samples, CoalHMM models the coalescent with recombination of three species [20, 21], where the hidden states are the four possible genealogies that might arise within the underlying species tree (fig. 1C). Two of the four genealogies differ from the species tree, which generate incongruencies that might pose a problem for standard phylogenetic reconstruction. Nevertheless, this phenomenon, commonly referred to as incomplete lineage sorting or ILS, is very informative about the demographic parameters of the underlying species tree, and CoalHMM can thus be used to estimate ancestral N_e and two speciation times. Moreover, CoalHMM uses a substitution model for mutations, so recurrent mutations are allowed. However, unlike SMCs, CoalHMM does not model coalescent events at discretized time intervals and, instead, coalescent times are modelled as single time points within an individual branch. Because of this, some of the parameter estimates of CoalHMM are biased [21], and, although obtaining accurate estimates is still possible [28], the debiasing procedure involves costly coalescent simulations. Moreover, posterior decoding can only be performed on the topology of the gene trees, and not on the coalescent times.

Here we present TRAILS, an HMM that combines modelling the information-rich ILS signal in the style of CoalHMM and the time discretization of SMC-like models to infer unbiased estimates of the demographic parameters (ancestral N_e and speciation times), and to enable the posterior decoding of both topology and coalescent times. In TRAILS, the hidden states are three-leaved gene trees, each with a specified topology and two coalescent events that happen at discretized time intervals on an underlying speciation tree (fig. 1D, fig. S11). The genealogies are rooted by a fourth sample from an outgroup species. The transition probabilities between the hidden states of TRAILS are calculated using coalescent-with-recombination theory for one, two and three lineages that segregate within the branches of the phylogeny. We provide formulas in matrix notation to calculate these transition probabilities for a varying number of discretized time intervals (see Methods for a short explanation, and Supporting Information for an in-depth description of the theory). The emitted states are sites in a four-way multiple

genome alignment, containing the sequences of the three species and the outgroup. The transition and emission probabilities are parameterized by two ancestral N_e, speciation times, and the recombination rate. Keeping the mutation rate at a fixed value, TRAILS allows for the estimation of the other parameters by optimizing the HMM likelihood given the alignment. After fitting the HMM, TRAILS can perform posterior decoding of the hidden states, inferring a posterior probability of coalescent events through time within the speciation tree.

Here we derive the transition and emission probabilities, implement the model and demonstrate its use on simulated and real data. After optimizing the population parameters using TRAILS on a simulated dataset, we show that increasing the number of discrete coalescence intervals reduces the bias in the parameter estimation. We also show how the posterior decoding can accurately reconstruct the true ARG, by inferring the topology of gene trees and the time in which coalescent events occurred. We perform additional simulations to show that the posterior decoding of TRAILS can be used to detect selective sweeps that happened on ancestral branches of the phylogeny. Finally, we analyze a human-chimp-gorilla-orangutan alignment, inferring the demographic parameters of the underlying species tree and performing genome-wide posterior decoding at the base-pair level.

RESULTS

Parameter estimation

The transition and emission probabilities of coalescent hidden Markov models (HMMs) are parameterized by the demographic model, i.e., by the speciation times, ancestral effective population sizes (N_e) and recombination rate (ρ). This means that using standard HMM algorithms, these parameters can be optimized to obtain the model that best explains the observed data. Both in CoalHMM and in TRAILS, numerical optimization is performed on the log-likelihood calculated using the forward algorithm, given a four-way genome alignment of three focal species and an outgroup. The maximum likelihood estimates of the demographic model are then found using a bound-constrained search algorithm that optimizes the likelihood function by evaluating it directly.

Previous work using coalescent HMMs has shown that the estimation of the demographic parameters is challenging. In CoalHMM, for example, the parameter estimates are highly biased [21], especially for the ancestral N_e and the recombination rate. It is possible to obtain close to unbiased estimates, but this requires a costly simulation procedure [28]. The source of the bias seems to be the restrictive state space of CoalHMM [21], which includes the topology of the genealogy but no information on when the coalescents happened within each branch of the tree (fig. 1C).

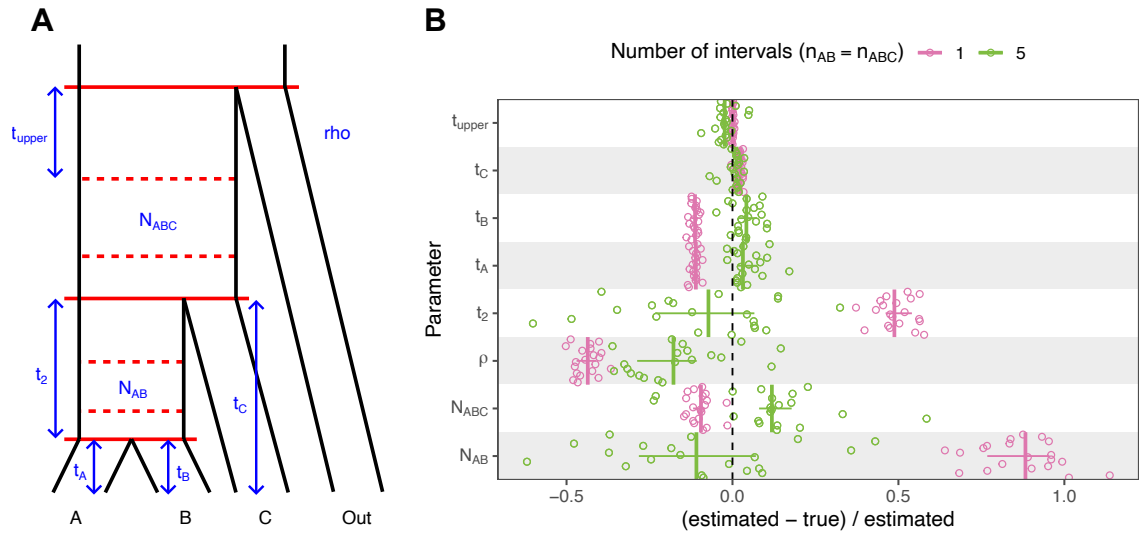


Figure 2: Increasing the complexity of TRAILS reduces the bias of the estimated parameters. (A) Diagram (not to scale) of the demographic model with all the optimized parameters in blue for the non-ultrametric case. In an ultrametric model, t_1 would correspond to the time from present to the shallowest speciation event, where $t_1 = t_A = t_B = t_C - t_2$. (B) Relative error of parameter values estimated from 20 simulated msprime genomes for $n_{AB} = n_{ABC} = 1$ (in pink) and $n_{AB} = n_{ABC} = 5$ (in green). Each independent run corresponds to a dot, vertical lines are median values, and vertical lines correspond to interquartile ranges. Units are normalized as (estimated-true)/estimated to ease comparison across parameters.

TRAILS overcomes this issue by extending the state space to include coalescent events that can happen in discretized time intervals. To demonstrate that TRAILS can perform unbiased parameter estimation, we generated twenty 10-Mb four-way alignments using `msprime` [29] by choosing a demographic model similar to the human-chimp-gorilla-orangutan speciation tree (see Methods, Simulations for details). The simulated sequence alignments were analyzed using TRAILS, estimating the times, ancestral N_e values and recombination rate depicted in fig. 2A. Parameters were estimated for $n_{AB} = n_{ABC} = 1$ and $n_{AB} = n_{ABC} = 5$, where n_{AB} is the number of intervals between speciation events and n_{ABC} is the number of intervals deep in time, in the common ancestor of all three species.

In the model where $n_{AB} = n_{ABC} = 1$, which is equivalent to the original CoalHMM model, parameter estimates deviate from their true values in the simulations, especially for t_2, N_{AB} and ρ (fig. 2B, in pink). Using a larger number of intervals ($n_{AB} = n_{ABC} = 5$) improves the accuracy of the parameter estimation (fig. 2B, in green). An exception to this is the recombination rate, which is still underestimated (albeit less so), possibly due to recombination events that produce small changes in the coalescent tree (e.g., not changing in the topology and only moving the coalescent times by a few generations) and are thus missed from the sequence data. Generally, however, the simulation results demonstrate that the source of the bias in the parameter estimation in CoalHMM can be alleviated with a more flexible model that includes coalescent times at discretized time intervals.

Posterior decoding of simulated data

Posterior decoding using the parameters estimated from the alignment can be performed using the transition and emission probabilities computed by TRAILS for a specific demographic model. In contrast to other coalescent-based HMMs, the resulting posterior probabilities are, however, hard to visualize, since each hidden state will have its own topology, and first (or more recent) and second (or more ancient) coalescent time intervals (fig. 1D, fig. S11). To overcome this, we summarize the posterior probabilities by grouping states that share certain features. For example, the posterior probabilities of all states that share the same topology can be summed. Similarly, the posteriors of all states with the same first or second coalescent times can also be summed.

In order to have a ground truth for comparison, the posterior decoding was performed on 100 kb of an alignment simulated using `msprime`, with a demographic model identical to that used in fig. 2B (see Methods for a full description of the model). The resulting posteriors can capture the true topology and the second coalescent time quite accurately (fig. 3 A and B, respectively), while the first coalescent time (fig. 3C) is harder to estimate. Additionally, V1 segments (fig. 3A) are easily misclassified as V0 segments, since V0 and V1 only differ in branch lengths but not in topology (see fig. 3 and fig. S9), and,thus, the emitted site patterns for V0 and V1 are similar.

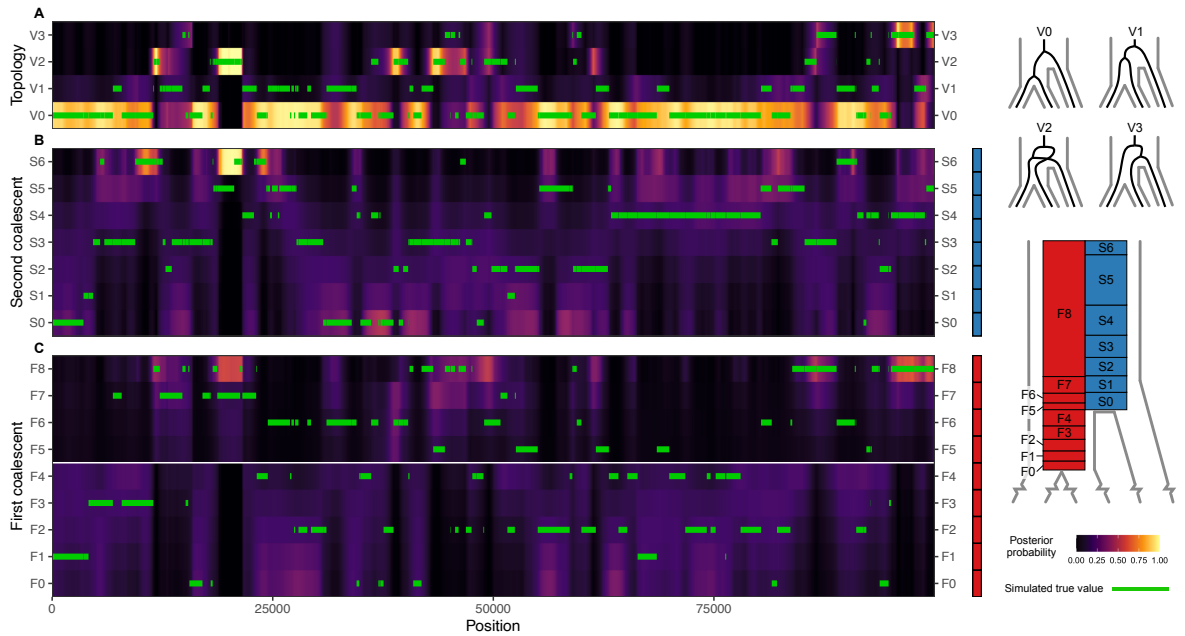


Figure 3: Posterior probability for the topology (A), second coalescent event (B) and first coalescent event (C) of a 100 kb msprime simulation. "First" and "second" refer to the order in which coalescent events happen, backwards in time. The true empirical topology and coalescent times are plotted as green lines.

Posterior decoding from simulated data with selection

To showcase how the posterior decoding could be useful to infer deviations from the standard coalescent, we simulated a 200 kb alignment using SLiM [30] containing a single positively selected variant ($2N_e s = 175$) in position 100 kb of the simulated alignment that arises at the first interval backwards in time, where the first two species merge in the speciation tree (interval S0 in fig. 4A). The site is strongly positively selected, but it lies well within the possible range of values for selection coefficients recorded in humans, with the lactase gene having up to $2N_e s = 1000$ in some human populations [31]. The demographic parameters were the same as those used for fig. 2 and fig. 3.

Both the true empirical values and the posterior decoding show that there is an overrepresentation of second coalescent events happening in interval S0 (fig. 4A), which is qualitatively different from the neutral case (fig. 3B). The positively selected variant confers a big advantage and is fixed rapidly in the population, with an expected fixation time of $\frac{4}{s}\left(\ln(2N_e s) + \gamma - \frac{1}{2N_e s}\right) = 2622$ generations or 65,550 years for an $N_e = 10,000$ and a generation time of $g = 25$ years, where $\gamma \approx 0.577$ is Euler's constant [32]. In contrast, for a neutrally evolving site, the expected fixation time is $4N_e = 40,000$ generations or 1 million years [33]. The effect of such strong selective force is that whatever polymorphism existed at the selected locus is quickly purged from the population, and, with it, linked neutral variants are hitchhiking along, causing a selective sweep. As a result, there is an excess of coalescent events happening in interval S0, which can be discerned from the posterior probabilities.

The signal observed in the posterior decoding can be summarized by computing the mean posterior probability per time interval and comparing it to the theoretical neutral expectation. There is a clear excess of coalescent events estimated to happen at the interval where the beneficial mutation arises (S0 in fig. 4B), although there is also an excess of coalescent events inferred by the posterior in nearby intervals (S1 and S2), where coalescents are misclassified due to close proximity in time. In any case, the pattern observed for the selective sweep in fig. 4B is in stark contrast with the neutral case shown in fig. 4C, where the posterior falls within the expected values. This demonstrates that deviations from neutrality can be inferred using the posterior decoding of TRAILS, and one could devise a windowed genome-wide scan for selection by summarizing the posterior as proposed in fig. 4B and C.

Parameter estimation from a HCGO alignment

ILS happens pervasively on the branches of the tree of life, spanning taxonomically diverse groups such as marsupials [34], birds [35, 36], fishes [37], plants [38], and mammals [39, 40], including primates [28, 41, 42]. For example, there is around 32% of ILS in the human-chimp ancestor, with 16% of the

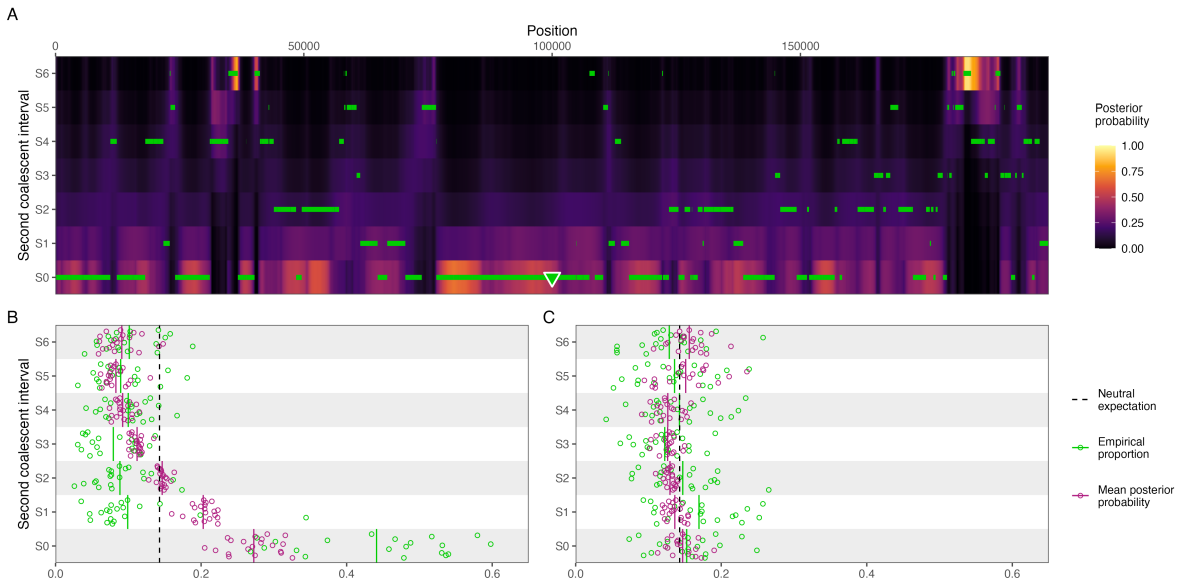


Figure 4: The posterior probability can be used to detect deviations from the neutral expectation. (A) Posterior probability of the second coalescent event for a simulated 200 kb region containing a positively selected variant at position 100 kb that arises in interval S0, represented by a triangle. The true simulated coalescent times are plotted as green horizontal lines. (B) Mean posterior probability for each second coalescent interval (purple), and the empirical true proportion of sites for each interval (green) for 20 simulated replicates with a selective sweep, using the same model as in (A). The theoretical neutral expectation is plotted as a black dashed line, and time intervals are adjusted so that all intervals have equal probability of observing a coalescent event. Continuous vertical lines represent mean values of the simulations. (C) Same as in (B), but for a neutrally evolving region, using the same model as in fig. 3.

genome grouping human and gorilla, and another 16% grouping chimp and gorilla [28]. These estimates were obtained using CoalHMM [21], together with estimates for ancestral N_e and split times, which were debiased using simulations. Here, we apply TRAILS to a 50 Mb human-chimp-gorilla alignment from chromosome 1 with orangutan as an outgroup to infer population genetics parameters and to gain information about the coalescent times and the topology through posterior decoding.

Using MafFilter [43], the alignment was first preprocessed to extract the species of interest (human, chimp, gorilla and orangutan), to merge consecutive alignment blocks, and to filter out small blocks (see Methods for further details). Using `biopython` [44], 50 Mb were extracted from chr1, namely the region from 25 Mb to 75 Mb. This region was used as the input for TRAILS, choosing the parameter values estimated in Rivas-González et al. [28] as starting values for the optimized parameters, setting $n_{\mathrm{AB}} = n_{\mathrm{ABC}} = 3$, and using the L-BFGS-B algorithm for model fitting [45, 46], although other bound-constrained method can also be used. To get more accurate parameter estimates, the optimized parameters were used as starting parameters for a second TRAILS run, where $n_{\mathrm{AB}} = 3$ and $n_{\mathrm{ABC}} = 5$.

The resulting estimates are displayed in fig. 5A. Assuming a mutation rate of $\mu = 1.25 \times 10^{-8}$ per site per generation and a generation time of $g = 25$ years [47, 48], the speciation time estimates are in

good agreement with previously inferred values. Using the human tip branch length, we estimate the time until the HC split at 5.51 million years ago (95% CI: [5.43, 5.54], ∼4-7 MYA from literature [21, 28, 49, 50, 51]), the HCG split at 10.40 MYA (95% CI: [10.27, 10.40], ∼8-12 from literature [21, 28, 50, 51]), and the HCGO split at 18.55 MYA (95% CI: [18.37, 18.73], ∼10-20 from literature [41, 50, 28]). Moreover, ancestral N_e inferred for the HC ancestor (167,400, 95% CI: [165,548, 170,361]) and for the HCG ancestor (101,290, 95% CI: [100,467, 101,492]) are consistent with previous estimates using CoalHMM (177,368 and 106,702, respectively [28]). Using the estimates for t_2 (in generations) and N_{AB}, we get a probability of ILS equal to

$$\mathrm{ILS} = \frac{2}{3}\exp(-t_2/(2N_{\mathrm{AB}})) = \frac{2}{3}\exp(-195,000/(2 \times 167,400)) = 37\%,$$

so our parameter estimates suggest the ((human, gorilla), chimp) topology in 18.5% of the genome and the ((chimp, gorilla), human) topology in 18.5% of the genome. Finally, the recombination rate was estimated to be $\rho = 1.19 \times 10^{-8}$ per site per generation, which matches the rate estimated for present-day humans [52].

TRAILS allows for the independent estimation of each individual branch length, which is useful for non-ultrametric trees. Figure 5A shows that the branch leading to chimps is longer than that leading to humans by around 5.9%, and the gorilla branch is longer than both the human (12.6%) and the chimp (9.1%) branches (calculated from the second speciation event to present). This deviation from the molecular clock is well supported by previous studies [53], and is likely because of different branches accumulating a different number of mutations per year, either due to an acceleration or deceleration of the mutational process, changes in the average time of reproduction, or a combination of these [54, 55, 56].

In summary, we have demonstrated that TRAILS is able to infer demographic parameters that are in agreement with estimates from the literature. More importantly, it does so without the need for any post-processing or corrections, avoiding the use of fossil calibrations [50] or debiasing procedures [21, 28].

Posterior decoding of the HCGO alignment

Posterior decoding was then performed using the optimized parameters and setting $n_{\mathrm{AB}} = n_{\mathrm{ABC}} = 5$. In order to get an understanding of the genome-wide variation of ILS and coalescent times, the resulting posterior probabilities were summarized in 100 kb windows along the chr1 region in three different ways. First, the mean posterior probability was calculated for each of the four possible topologies,

by first summing the posteriors of all hidden states sharing the same topology for each site, and then averaging over all sites on the 100 kb window. The resulting probabilities were then used to calculate a proxy for ILS, by summing the probabilities of observing the ILS topologies (V2 and V3 in fig. 3). Second, using a similar procedure, the mean posterior probability was calculated for each of the six possible intervals for the first coalescent event, by first summing the posteriors of all hidden states sharing the same first coalescent interval for each site, and then averaging over all sites in the 100 kb window. As a proxy for the first coalescent time, integers from 1 to 6 were assigned to each interval in chronological order backwards in time, and a weighted mean of those integers was computed, where the weights were the mean posterior probabilities per window. Third, the same quantity as for the first coalescent was computed for the second coalescent.

After filtering outliers smaller than the 1st percentile and larger than the 99th percentile, these proxies were plotted as heatmaps (fig. 5B). The first coalescent and the ILS proportion show a very strong correlation ($\rho = 0.979$), likely reflecting that V2 and V3 can only happen in the common ancestor of all three species, so when ILS is present, coalescent times are generally deeper (and vice versa). This signal is also captured, although more weakly, by the correlation between the second coalescent and the ILS proportion ($\rho = 0.483$). This can be explained by knowing that, conditional on ILS, the second coalescent follows a convolution of exponentials of rates 3 and 1 [57], while, conditional on V0 (i.e., conditional on the first coalescent happening between speciation events), the second coalescent simply follows an exponential or rate 1. Thus, if more ILS is present in a certain window, then, on average, the second coalescent will tend to happen deeper in time.

Figure 5B also shows how, at the 100 kb level, the genome displays spatial covariation in the amount of ILS and the time to coalescence that exceeds stochastic effects of a neutral coalescent process. This is in line with previous results [28], where ILS proportions are affected by genomic features such as gene density, recombination rate, and the effects of linked selection.

A zoomed-in region of around 41 kb is shown in fig. 5C, which shows a long fragment of the ((chimp, gorilla), human) topology (V3). This fragment is unusually long, spanning 6,800 bp, and it is highly implausible following the demographic model inferred by TRAILS. Thanks to the posterior decoding of the coalescent times performed by TRAILS, we can observe that, for this fragment, the first coalescent event backwards in time happens close to the second speciation time (in interval F5), while the second coalescent event happens in the deepest time interval (S4).

One explanation for such a long V3 fragment is that it might be influenced by selection, which would maintain the alternative topology uninterrupted for a long period of time. Another explanation could

be that this fragment is introgressed, especially given that the first coalescent event is shallow and that the fragment is long. For comparison, another region in chromosome 1, which also shows an excess of V3 topology, has a much more variable distribution of coalescent times, and it is more fragmented (fig. S17). Such detailed information about the timing of coalescent events is only possible thanks to the time discretization of TRAILS, and these details would have been missed, for example, in the posterior decoding of CoalHMM (recall fig. 1C).

The V3 fragment in fig. 5C overlaps with the last exon of the *LDLRAD1* gene, which codes for a lipoprotein receptor (UniProt: Q5T700). LDLRAD1 does not show signals of positive selection in hominids (based on dN/dS values from Rivas-González et al. [28]). Additionally, this gene is not particularly constrained in primates, as measured by PhastCons [58] and PhyloP [59], and it is not enriched for repeat elements, as retrieved from the UCSC Genome Browser [60]. While we were unable to point out a specific cause for the pattern observed for LDLRAD1, fig. 5C showcases how TRAILS can be used to infer the topology and coalescent times of protein-coding genes at the base-pair level across millions of years of evolution. Comparing the posterior decoding with genomic covariates can reveal selective processes affecting the sorting of lineages [28] or solve cases of phenotypic hemiplasy [34].

Discussion

Coalescent-based approaches for analyzing genomic data are essential tools for understanding the ancestral history of species. Here, we have introduced TRAILS, an HMM that models the topology and the two coalescent events for gene genealogies within a phylogeny of three species. TRAILS can accurately infer population genetic parameters (ancestral N_e, speciation times and recombination rate). From the posterior decoding, the three-species ARG can be inferred at the base-pair level, providing insight into the ancestral history of the species at high resolution. Deviations from neutrality can be detected by summarizing the posterior decoding in windows and running genomic scans to find excess of coalescents happening at certain time intervals, such as proposed in fig. 4. Moreover, more coarse-grained summaries of the posterior decoding spanning several kilobases could be used to infer the genome-wide variation of ILS or coalescent times, potentially revealing correlations with other genomic features such as variation in the recombination rate or selection [28, 61].

As demonstrated here, the posterior decoding from TRAILS is a powerful way to infer details of the ARG in the context of speciation, together with departures from the neutral expectation. Recurrent selective sweeps that have happened during the speciation process are hypothesized to be drivers of speciation, and to greatly influence the genealogical landscape of present-day genomes. For example, the

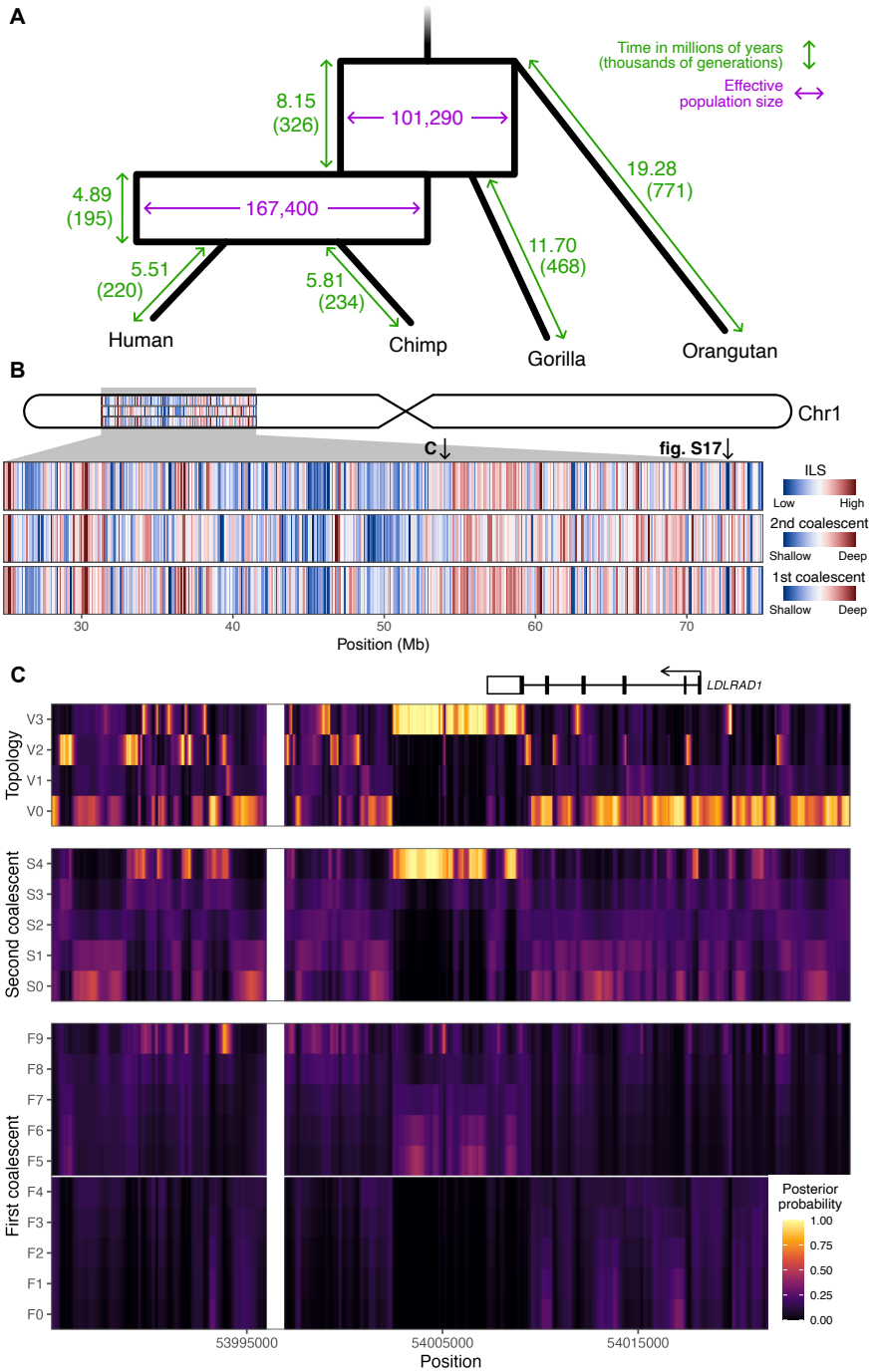


Figure 5: TRAILS output for 50 Mb of chromosome 1 of the HCGO alignment. (A) Estimates for the speciation times (green) and ancestral N_e (purple) of the speciation process, optimized using TRAILS and assuming a mutation rate of $\mu = 1.25 \times 10^{-8}$ per site per generation. To convert time from generations to millions of years, a generation time of $g = 25$ years per generation was used. (B) Genome-wide variation of ILS, and first and second coalescent times. (C) Posterior decoding of the topology, and first and second coalescent events for a zoomed-in region in chromosome 1. As in fig. 3, both V0 and V1 correspond to the species topology (((H,C),G),O);, V2 corresponds to (((H,G),C),O);, and V3 to (((C,G),H),O);. The *LDLRAD1* gene is plotted on top, where exons are represented as boxes, coding regions as filled boxes, and introns as horizontal lines.

human X chromosome contains long haplotypes shared across all non-African populations [62], spanning large genomic regions that are both lacking Neanderthal introgression [63], and showing very low rates of ILS in the human-chimp ancestor [64]. This suggests that the X chromosome has a unique evolutionary history which is greatly affected by gene flow (or lack thereof), and that these low-diversity regions might be related to genetic incompatibilities that arose during the speciation of ancestral hominids. TRAILS can help locate these ancient sweeps and infer when they occurred, potentially illuminating when and how genomes were affected by selection during the speciation process.

Using posterior decoding, regions that show unusually high levels of an alternative topology with very shallow coalescents can also be detected, which could indicate ancient introgression or hybridization events happening between ancestral branches of the species tree. Such ancient introgression events have been reported to be pervasive among some branches in the primate species tree [50], although they can be difficult to distinguish from ILS [65] unless explicitly modelled. TRAILS could be extended to model introgression more directly by including additional hidden states representing introgressed genomic fragments. These would have exceedingly short coalescent times compared to the deep coalescent ILS states [28], and TRAILS provides the mathematical framework to distinguish between these two cases.

TRAILS could also be extended to accommodate variation in N_e along individual ancestral branches in the species tree, conceptually very similar to what is done in PSMC analyses from a single extant genome [8]. Modelling variation in N_e can elucidate how speciation events might have happened. For example, population sizes that are maintained more constant during the speciation event might indicate a cleaner split, while increased ancestral N_e just prior to the estimated time of speciation (here equalled to the total cessation of gene flow) might point to a period of elevated population structure and a prolonged species separation with migration [27]. Modelling changes in the demography around speciation events might also help us detect and characterize instances of complex speciation, as proposed, for example, by Patterson et al. [49].

The current implementation of TRAILS for calculating the transition probability matrix of the HMM is restricted to three species and a relatively few number of hidden states (see section 10 of the Supporting Information for a discussion on the running times). With more efficient algorithms, future extensions of TRAILS could be devised to analyze more than three species, thus allowing for the inference of the speciation tree and the multi-species ARG for more taxa. Based on the parameters estimated by TRAILS for the HCGO alignment (fig. 5A), the proportion of ILS between humans (or chimps), gorillas and orangutans would be around 13%. While this violates one of the assumptions of TRAILS, which is that there should be inappreciable ILS between the outgroup and the rest of the analyzed species, it also

showcases the need for models that are able to accommodate more species. This could help us resolve more complex patterns of ILS, which include phenomena such as anomaly zones [36, 66, 67].

METHODS

The transition probabilities between the hidden states of TRAILS can be calculated from a series of interconnected continuous-time Markov chains (CTMCs) that model the coalescent with recombination of two contiguous nucleotides for one, two or three sequences. The CTMCs are parameterized by the ancestral N_e, speciation times and recombination rate. The transitions for TRAILS are subsequently calculated by conditioning the CTMCs on the topology and coalescent times of the gene trees at those two sites, binning coalescent events into discretized time intervals along the speciation process. Additionally, the emission probabilities for each hidden state are calculated from a CTMC of the mutational process by choosing a certain substitution model. In this section, we provide a summary of the model, and the full explanation can be consulted in the Supporting Information.

Continuous-time Markov chains for the ancestral process

The coalescent with recombination between two sites can be approximated as a continuous-time Markov chain (CTMC). For one sequence, the left and the right sites can be either linked or unlinked, so there are only two possible states for the CTMC. Two linked sites become unlinked when a recombination event happens between them, which happens with a rate of ρ_1. On the other hand, the unlinked left and the right sites become linked when a coalescent event happens between them, with a rate of γ_1. These two transitions can be gathered in a transition rate matrix

$$\boldsymbol{Q_1} = \begin{pmatrix} -\gamma_1 & \gamma_1 \\ \rho_1 & -\rho_1 \end{pmatrix}. \tag{1}$$

From this transition rate matrix, we can calculate the probability matrix P_{A} as $\exp(t\boldsymbol{Q_1})$, which gives the probability of the sites being unlinked or linked at a certain time t given that the chain starts in the unlinked state (first row) or unlinked state (second row).

When two sequences are involved, the state-space of the CTMC becomes more complex. Apart from the coalescent and the recombination events described above, sites can also coalesce irreversibly backwards in time with rate γ_2, which happens when two left (or two right) sites from two different sequences find common ancestry. The resulting rate matrix (fig. S5) for the coalescent with recombination with two sequences then corresponds to a CTMC with 15 states (fig. S4), which was originally described

by Simonsen and Churchill [68]. Due to these irreversible coalescent events, the rate matrix has a block-like structure, and it contains sets of states in which sequences can freely recombine and coalesce until a irreversible coalescent event occurs. Ultimately, both the left and the right sites will have irreversibly coalesced, reaching one of two absorbing states. Note that the matrix is quite sparse, since most of the transitions are not allowed.

Following a similar reasoning, the coalescent with recombination for three sequences can also be modelled as a CTMC. In this case, both the left and the right site will eventually undergo two irreversible coalescent events, which can potentially happen in any order between the three sequences. This creates 203 possible states (fig. S7), the transitions of which can also be gathered in a rate matrix (fig. S8). This matrix also has a block-like structure, and, given enough time, states will transition into one of the two absorbing states.

If all three sequences belonged to the same species, a three-sequence CTMC would be sufficient to model the coalescent with recombination. However, the sequences belong to three different species, so the speciation process has to be overlaid on top. Subsequently, the coalescent with recombination along the speciation tree is modelled as a series of interconnected CTMCs.

Because sequences are sampled in present time, the left and the right sites are fully linked at time 0, meaning that the starting probability vector for the one-sequence CTMCs is $(0, 1)$. Backwards in time, each of the sequences will remain isolated for a certain period of time in which the two sites can recombine and coalesce freely. The sequences for species A and B will remain isolated until the first speciation event at time t_A and t_B, respectively. Then, the final probabilities of the one-sequence CTMCs for A and B are merged to create the initial probabilities for the two-sequence CTMC. After a certain time, where the two sequences are allowed to coalesce and recombine, the final probabilities for the two-sequence CTMC and the final probabilities for the one-sequence CTMC of species C will be merged, thus creating the starting probabilities for the three-sequence CTMC. Finally, given enough time, all sequences for both the right and the left site will eventually coalesce into one of the two absorbing states of the last CTMC.

Transition probabilities of the HMM

The hidden states of TRAILS are genealogies which include a topology and two coalescent events that can happen within discretized time intervals. The breakpoints of the time intervals can thus be used to transform the CTMCs into a discrete-time Markov chain (DTMC). First, the joint probability of observing the genealogies and the left and the right loci can be computed by careful bookkeeping of the

appropriate paths within the CTMC, defined by the corresponding genealogies and the discretized time intervals. The transition probability matrix of the DTMC (and the HMM) can then be obtained upon dividing the joint probability by the discretized marginals. A detailed description of these derivations is given in Supporting Information.

Emission probabilities of the HMM

For each hidden state, the emission probabilities are calculated using the Jukes-Cantor mutational model [69]. Instead of calculating the emitted nucleotides for the three species only, TRAILS also includes the nucleotides emitted by an outgroup, which provides essential information about the ancestral state in each site. This additional species must have a sufficient divergence with the rest of the species such that ILS can be neglected between them.

Parameterization

The transition and emission probabilities are parameterized by the speciation times (t_A, t_B, t_C, t_2, t_{upper}), the effective population sizes (N_{AB}, N_{ABC}), and the recombination rate (fig. 2A). Implicitly, TRAILS is also parameterized by the mutation rate, but this cannot be jointly inferred with the rest of the parameters because the parameter values can be scaled by any factor and still produce the same coalescent model [57, 70]. Instead, the mutation rate in the model is fixed to 1, and all other parameters are rescaled appropriately. The resulting units for the speciation times are number of generations multiplied by the mutation rate, and, similarly, the effective population sizes are number of individuals times the mutation rate. Accordingly, the recombination rate is divided by the mutation rate, so the optimized parameter is the ratio between the recombination and the mutation rate. After estimating the parameters, parameter values with more interpretable units can be obtained by choosing an appropriate mutation rate.

TRAILS allows for two different parameterizations, namely the *ultrametric* model and the *non-ultrametric* model. In the ultrametric model, all sequences are sampled at time 0, so the molecular clock is assumed ($t_A = t_B = t_C - t_2 = t_1$). Instead, in the non-ultrametric model, each sequence (A, B and C) is allowed to be sampled at a different time. This is useful to model deviations from the molecular clock, for example, when the number of generations for each of the species from present time until the speciation event is different, or the mutation rate varies between the species.

Simulations

Parameter estimation

Simulations to validate the model were performed in `msprime` [29]. The underlying demographic model follows a speciation tree with four species, namely A, B, C and D. The time from present to the first speciation event was set to $t_1 = 200,000$ generations. The (haploid) ancestral N_e for the time between speciation events was set to $N_{\text{AB}} = 80,000$. In order to keep an ILS proportion of 32%, the time between the first and the second speciation events was set to $t_2 = -N_{\text{AB}} \log(\frac{3}{2} \times 0.32) = 25,501$ generations. The (haploid) ancestral N_e earlier than the second speciation event was set to $N_{\text{ABC}} = 70,000$, and the time between the second speciation event and the speciation event with the outgroup was set to $t_3 = 1,000,000$ generations. The recombination rate was set to $\rho = 0.5 \times 10^{-8}$ per site per generation. The tree was kept ultrametric in number of generations, meaning that all species were sampled at generation 0, so $t_{\text{A}} = t_{\text{B}} = t_1$, $t_{\text{C}} = t_1 + t_2$, and $t_{\text{D}} = t_1 + t_2 + t_3$. Mutations were then added on top of the simulated genealogies according to the Jukes-Cantor model [69] with a mutation rate of $\mu = 1.5 \times 10^{-8}$ per generation per site.

In order to investigate how the number of intervals in the AB-ancestor (n_{AB}) and the ABC-ancestor (n_{ABC}) affect the parameter estimation, twenty 10-Mb alignments were simulated using `msprime`, and then TRAILS was run to estimate the demographic parameters for $n_{\text{AB}} = n_{\text{ABC}} = 1$ and for $n_{\text{AB}} = n_{\text{ABC}} = 5$, using the bound-constrained Nelder-Mead algorithm. The starting parameters of the optimization were randomly drawn from a normal distribution centered on the true value and with a standard deviation of the true value divided by 5. Convergence was achieved at around 150 iterations, with a runtime of around 10 hours for $n_{\text{AB}} = n_{\text{ABC}} = 5$ per 10-Mb region (see section 10 of the Supporting Information for further details on the runtime of the model).

Posterior probability

The demographic model described above was also used to generate a 100 kb alignment to perform posterior decoding with the true parameters fixed. In TRAILS, the default way of dividing up the coalescent space into discretized time intervals is by taking quantiles of a truncated exponential of rate 1 (measured in units of N_{AB}) for the time between speciation events, while the time previous to the earliest speciation event is divided following the quantiles of an exponential of rate 1 (measured in units of N_{ABC}). This scheme is appropriate for computing the posterior decoding of the first coalescent event when it happens between speciation events, since it is expected to follow exactly a truncated exponential of rate 1 according to the standard coalescent. However, if the first coalescent happens deep in time, then

the coalescent times will follow an exponential with rate 3, since there are 3 lineages present. This means that most coalescent events happen very fast, and there would be an overrepresentation of coalescents in the first interval if the default cutpoint scheme is used.

TRAILS is, however, not restricted to a specific discretization, and it can compute the transition and emission probabilities, and, thus, perform posterior decoding, for user-specified intervals. For the first coalescent in deep time, posterior decoding was performed using cutpoints from the quantiles of an exponential with rate 3, with $n_{\text{ABC}} = 7$ and the true parameters from the `msprime` simulation.

Moreover, the second coalescent event will follow a mixture of an exponential with a rate of 1 (for the V0 states) and a convolution of two exponentials with rates 3 and 1 (for the deep coalescent states) [57], which will happen with probability $1 - \exp(-t_2/N_{\text{AB}})$ and $\exp(-t_2/N_{\text{AB}})$, respectively. Here, t_2 is the time between speciation events in number of generations, and N_{AB} is the effective population population size. The second coalescent time can thus be represented as a phase-type distribution [71], with sub-intensity matrix $\boldsymbol{S}$ and initial probability vector $\boldsymbol{\pi}$, such that

$$\boldsymbol{S} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -3 & 3 \\ 0 & 0 & -1 \end{pmatrix}$$

and

$$\boldsymbol{\pi} = \left(1 - \exp\left(-\frac{t_2}{N_{\text{AB}}}\right), \exp\left(-\frac{t_2}{N_{\text{AB}}}\right), 0\right).$$

Therefore, posterior decoding for the second coalescent was performed using the quantiles of this phase-type distribution using `PhaseTypeR` [72], with $n_{\text{AB}} = 5$, $n_{\text{ABC}} = 7$, and the true parameters used to generate the `msprime` simulation.

Selection

Using the same demographic model, a 200-kb alignment was simulated using SLiM [30], assuming a single positively selected variant in the middle of the region, with population-scaled selection parameter $2N_e s = 175$. Since SLiM is a forward simulator and runs much slower than backward simulators such as `msprime`, all the demographic parameters of the model were rescaled by a factor of 200 in order to increase computational speed. Posterior decoding was performed on the resulting alignment simulated using the same discretization scheme as described above, and the resulting posterior probabilities are plotted in fig. 4A.

To showcase how the posterior of TRAILS can be used as a test to detect deviations from neutrality,

SLiM was used to generate twenty 200-kb with and without a selected variant, and TRAILS was run afterward to calculate the posterior probability for the second coalescent time in 7 discretized intervals. The signal of the posterior decoding was summarized as the mean posterior probability for each discretized time interval, plotted in fig. 4B and C.

Real data

The chromosome 1 multiz alignment of 30 mammalian species (27 primates) was downloaded from the UCSC Genome Browser database in MAF format. Using MafFilter [43], the species of interest were filtered (human, chimp, gorilla and orangutan), syntenic blocks separated by 200 nucleotides or less were merged using human as a reference, and blocks smaller than 2,000 bp were filtered out. The resulting filtered MAF was used as input for TRAILS, using the parameters estimated in Rivas-González et al. [28] as starting values. The optimization was performed using a bound-constrained version of the L-BFGS-B algorithm implemented in `numpy` [45, 46, 73], by setting $n_{\mathrm{AB}} = n_{\mathrm{ABC}} = 3$, and using the L-BFGS-B algorithm for model fitting. To get a more accurate parameter estimation, the optimized estimates were used as starting values for a second TRAILS run where $n_{\mathrm{ABC}} = 5$, optimized using a bound-constrained Nelder-Mead algorithm [74, 75], which showed better convergence for already-optimized TRAILS runs (results not shown).

Confidence intervals for the estimated parameters were computed using parametric bootstrapping. 20 replicates of 50-Mb regions were simulated from the model fitted with the estimated parameters. Afterward, TRAILS was run on the simulated regions to get optimized parameters. For each parameter, a normal distribution was fitted for the 20 replicates, and the 95% confidence intervals were calculated from the fitted normal (fig. S18 and table S2).

Code availability

TRAILS can be downloaded and installed from pip (`https://pypi.org/project/trails-rivasiker/`), and the source code can be browsed at `https://github.com/rivasiker/trails`. The code for reproducing the figures can be found at `https://github.com/rivasiker/trails_paper`. Simulations and analysis of real data can be re-run using a gwf workflow (`https://gwf.app`). Figures were plotted using the `tidyverse` [76], `glue` [77], `GGally` [78], `ggthemes` [79] and `PhaseTypeR` [72] packages in `R` version 4.2.1 [80].

Author contributions

I.R.-G. performed conceptualization, methodology, software development, validation, formal analysis, investigation, writing, and visualization. M.H.S. performed conceptualization, writing (review and editing), and supervision. J.W. performed conceptualization, methodology, writing (review and editing), and supervision. A.H. performed conceptualization, methodology, writing (review and editing), and supervision.

Acknowledgements

We gratefully acknowledge Julien Dutheil and Nick Patterson for useful discussions on the implementation of the model. We also thank GenomeDK for providing the computational resources for performing the analyses. Furthermore, we thank the reviewers that greatly helped improve the manuscript. This work was supported by grants NNF18OC0031004 from the Novo Nordisk Foundation and 6108-00385 from the Independent Research Fund Denmark, Natural Sciences.

References

[1] R. C. Griffiths. The two-locus ancestral graph. *Lecture Notes-Monograph Series*, 18:100–117, 1991. ISSN 07492170. URL `http://www.jstor.org/stable/4355649`.

[2] Robert C Griffiths and Paul Marjoram. An ancestral recombination graph. In Peter Donnelly and Simon Tavaré, editors, *Progress in Population Genetics and Human Evolution (IMA Volumes in Mathematics and its Applications, vol. 87)*, pages 257–270. Springer-Verlag, New York, 1997.

[3] Melissa Hubisz and Adam Siepel. *Inference of Ancestral Recombination Graphs Using ARGweaver*, pages 231–266. Springer US, New York, NY, 2020.

[4] Carsten Wiuf and Jotun Hein. Recombination as a point process along sequences. *Theoretical Population Biology*, 55(3):248–259, 1999.

[5] Gilean AT McVean and Niall J Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393, 2005.

[6] Paul Marjoram and Jeff D Wall. Fast "coalescent" simulation. *BMC Genetics*, 7:1–9, 2006.

[7] Peter R Wilton, Shai Carmi, and Asger Hobolth. The SMC′ is a highly accurate approximation to the ancestral recombination graph. *Genetics*, 200(1):343–355, 2015.

[8] Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011.

[9] Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–925, 2014.

[10] Anna-Sapfo Malaspinas, Michael C Westaway, Craig Muller, Vitor C Sousa, Oscar Lao, Isabel Alves, Anders Bergström, Georgios Athanasiadis, Jade Y Cheng, Jacob E Crawford, et al. A genomic history of aboriginal australia. *Nature*, 538(7624):207–214, 2016.

[11] Pier Francesco Palamara, Jonathan Terhorst, Yun S Song, and Alkes L Price. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics*, 50(9):1311–1317, 2018.

[12] Jonathan Terhorst, John A Kamm, and Yun S Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303–309, 2017.

[13] Matthew D Rasmussen, Melissa J Hubisz, Ilan Gronau, and Adam Siepel. Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10(5):e1004342, 2014.

[14] Leo Speidel, Marie Forest, Sinan Shi, and Simon R Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329, 2019.

[15] Jerome Kelleher, Yan Wong, Anthony W Wohns, Chaimaa Fadil, Patrick K Albers, and Gil McVean. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338, 2019.

[16] Anthony Wilder Wohns, Yan Wong, Ben Jeffery, Ali Akbari, Swapan Mallick, Ron Pinhasi, Nick Patterson, David Reich, Jerome Kelleher, and Gil McVean. A unified genealogy of modern and ancient genomes. *Science*, 375(6583):eabi8264, 2022.

[17] Brian C Zhang, Arjun Biddanda, Árni Freyr Gunnarsson, Fergus Cooper, and Pier Francesco Palamara. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nature Genetics*, pages 1–9, 2023.

[18] Débora YC Brandt, Xinzhu Wei, Yun Deng, Andrew H Vaughn, and Rasmus Nielsen. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*, 221(1): iyac044, 2022.

[19] Thomas Mailund, Julien Y Dutheil, Asger Hobolth, Gerton Lunter, and Mikkel H Schierup. Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. *PLoS Genetics*, 7(3):e1001319, 2011.

[20] Asger Hobolth, Ole F Christensen, Thomas Mailund, and Mikkel H Schierup. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genetics*, 3(2):e7, 2007.

[21] Julien Y Dutheil, Ganesh Ganapathy, Asger Hobolth, Thomas Mailund, Marcy K Uyenoyama, and Mikkel H Schierup. Ancestral population genomics: the coalescent hidden Markov model approach. *Genetics*, 183(1):259–274, 2009.

[22] James H Degnan and Noah A Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24(6):332–340, 2009.

[23] Bruce Rannala, A Leache, S Edwards, and Ziheng Yang. The multispecies coalescent model and species tree inference. Self Published, 2020. URL `https://inria.hal.science/PGE/hal-02535622`.

[24] Siavash Mirarab, Luay Nakhleh, and Tandy Warnow. Multispecies coalescent: theory and applications in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 52:247–268, 2021.

[25] Colm O'hUigin, Yoko Satta, Naoyuki Takahata, and Jan Klein. Contribution of homoplasy and of ancestral polymorphism to the evolution of genes in anthropoid primates. *Molecular Biology and Evolution*, 19(9):1501–1513, 2002.

[26] David B Wake, Marvalee H Wake, and Chelsea D Specht. Homoplasy: from detecting pattern to determining process and mechanism of evolution. *Science*, 331(6020):1032–1035, 2011.

[27] Thomas Mailund, Anders E Halager, Michael Westergaard, Julien Y Dutheil, Kasper Munch, Lars N Andersen, Gerton Lunter, Kay Prüfer, Aylwyn Scally, Asger Hobolth, et al. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLoS Genetics*, 8(12):e1003125, 2012.

[28] Iker Rivas-González, Marjolaine Rousselle, Fang Li, Long Zhou, Julien Y. Dutheil, Kasper Munch, Yong Shao, Dongdong Wu, Mikkel H. Schierup, and Guojie Zhang. Pervasive incomplete lineage sorting illuminates speciation and selection in primates. *Science*, 380(6648):eabn4409, 2023.

[29] Franz Baumdicker, Gertjan Bisschop, Daniel Goldstein, Graham Gower, Aaron P Ragsdale, Georgia Tsambos, Sha Zhu, Bjarki Eldon, E Castedo Ellerman, Jared G Galloway, Ariella L Gladstein, Gregor

Gorjanc, Bing Guo, Ben Jeffery, Warren W Kretzschumar, Konrad Lohse, Michael Matschiner, Dominic Nelson, Nathaniel S Pope, Consuelo D Quinto-Cortés, Murillo F Rodrigues, Kumar Saunack, Thibaut Sellinger, Kevin Thornton, Hugo van Kemenade, Anthony W Wohns, Yan Wong, Simon Gravel, Andrew D Kern, Jere Koskela, Peter L Ralph, and Jerome Kelleher. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3):iyab229, 2022.

[30] Benjamin C Haller and Philipp W Messer. Slim 4: Multispecies eco-evolutionary modeling. *The American Naturalist*, 201(5):E000–E000, 2023.

[31] Laure Ségurel and Céline Bon. On the evolution of lactase persistence in humans. *Annual Review of Genomics and Human Genetics*, 18:297–319, 2017.

[32] Joachim Hermisson and Pleuni S Pennings. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169(4):2335–2352, 2005.

[33] Sarah P Otto and Michael C Whitlock. Fixation probabilities and times. *Encyclopedia of Life Sciences*, pages 1–5, 2013.

[34] Shaohong Feng, Ming Bai, Iker Rivas-González, Cai Li, Shiping Liu, Yijie Tong, Haidong Yang, Guangji Chen, Duo Xie, Karen E Sears, et al. Incomplete lineage sorting and phenotypic evolution in marsupials. *Cell*, 185(10):1646–1660, 2022.

[35] Alexander Suh, Linnéa Smeds, and Hans Ellegren. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLoS Biology*, 13(8):e1002224, 2015.

[36] Alison Cloutier, Timothy B Sackton, Phil Grayson, Michele Clamp, Allan J Baker, and Scott V Edwards. Whole-genome analyses resolve the phylogeny of flightless birds (Palaeognathae) in the presence of an empirical anomaly zone. *Systematic Biology*, 68(6):937–955, 2019.

[37] Fernando Alda, Victor A Tagliacollo, Maxwell J Bernt, Brandon T Waltz, William B Ludt, Brant C Faircloth, Michael E Alfaro, James S Albert, and Prosanta Chakrabarty. Resolving deep nodes in an ancient radiation of neotropical fishes in the presence of conflicting signals from incomplete lineage sorting. *Systematic Biology*, 68(4):573–593, 2019.

[38] Y Zhou, L Duvaux, G Ren, L Zhang, O Savolainen, and J Liu. Importance of incomplete lineage sorting and introgression in the origin of shared genetic variation between two closely related pines with overlapping distributions. *Heredity*, 118(3):211–220, 2017.

[39] Kun Wang, Johannes A Lenstra, Liang Liu, Quanjun Hu, Tao Ma, Qiang Qiu, and Jianquan Liu.

Incomplete lineage sorting rather than hybridization explains the inconsistent phylogeny of the wisent. *Communications Biology*, 1(1):169, 2018.

[40] Celine Scornavacca and Nicolas Galtier. Incomplete lineage sorting in mammalian phylogenomics. *Systematic Biology*, 66(1):112–120, 2017.

[41] Asger Hobolth, Julien Y Dutheil, John Hawks, Mikkel H Schierup, and Thomas Mailund. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Research*, 21(3):349–356, 2011.

[42] Thomas Mailund, Kasper Munch, and Mikkel Heide Schierup. Lineage sorting in apes. *Annual Review of Genetics*, 48:519–535, 2014.

[43] Julien Y Dutheil, Sylvain Gaillard, and Eva H Stukenbrock. MafFilter: a highly flexible and extensible multiple genome alignment files processor. *BMC Genomics*, 15:1–10, 2014.

[44] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25 (11):1422–1423, 2009.

[45] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208, 1995.

[46] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

[47] Kevin E Langergraber, Kay Prüfer, Carolyn Rowney, Christophe Boesch, Catherine Crockford, Katie Fawcett, Eiji Inoue, Miho Inoue-Muruyama, John C Mitani, Martin N Muller, et al. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proceedings of the National Academy of Sciences*, 109(39):15716–15721, 2012.

[48] Richard J Wang, Samer I Al-Saffar, Jeffrey Rogers, and Matthew W Hahn. Human generation times across the past 250,000 years. *Science Advances*, 9(1):eabm7047, 2023.

[49] Nick Patterson, Daniel J Richter, Sante Gnerre, Eric S Lander, and David Reich. Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441(7097):1103–1108, 2006.

[50] Dan Vanderpool, Bui Quang Minh, Robert Lanfear, Daniel Hughes, Shwetha Murali, R Alan Harris, Muthuswamy Raveendran, Donna M Muzny, Mark S Hibbins, Robert J Williamson, et al. Primate

phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. *PLoS Biology*, 18(12):e3000954, 2020.

[51] Aylwyn Scally, Julien Y Dutheil, LaDeana W Hillier, Gregory E Jordan, Ian Goodhead, Javier Herrero, Asger Hobolth, Tuuli Lappalainen, Thomas Mailund, Tomas Marques-Bonet, et al. Insights into hominid evolution from the gorilla genome sequence. *Nature*, 483(7388):169–175, 2012.

[52] Bjarni V Halldorsson, Gunnar Palsson, Olafur A Stefansson, Hakon Jonsson, Marteinn T Hardarson, Hannes P Eggertsson, Bjarni Gunnarsson, Asmundur Oddsson, Gisli H Halldorsson, Florian Zink, et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science*, 363(6425):eaau1043, 2019.

[53] Priya Moorjani, Carlos Eduardo G Amorim, Peter F Arndt, and Molly Przeworski. Variation in the molecular clock of primates. *Proceedings of the National Academy of Sciences*, 113(38):10607–10612, 2016.

[54] Søren Besenbacher, Christina Hvilsom, Tomas Marques-Bonet, Thomas Mailund, and Mikkel Heide Schierup. Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nature Ecology & Evolution*, 3(2):286–292, 2019.

[55] Gregg WC Thomas, Richard J Wang, Arthi Puri, R Alan Harris, Muthuswamy Raveendran, Daniel ST Hughes, Shwetha C Murali, Lawrence E Williams, Harsha Doddapaneni, Donna M Muzny, et al. Reproductive longevity predicts mutation rates in primates. *Current Biology*, 28(19): 3193–3197, 2018.

[56] Lindell Bromham. The genome as a life-history character: why rate of molecular evolution varies between mammal species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366 (1577):2503–2513, 2011.

[57] John Wakeley. *Coalescent Theory: An Introduction*. Roberts & Company Publishers, 1 edition, June 2008.

[58] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005.

[59] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010.

[60] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome Research*, 12(6):996–1006, 2002.

[61] James B Pease and Matthew W Hahn. More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. *Evolution*, 67(8):2376–2384, 2013.

[62] Laurits Skov, Moises Coll Macia, Elise Anne Lucotte, Maria Izabel Alves Cavassim, David Castellano, Mikkel Heide Schierup, and Kasper Munch. Extraordinary selection on the human X chromosome associated with archaic admixture. *Cell Genomics*, 3(3), 2023.

[63] Sriram Sankararaman, Swapan Mallick, Michael Dannemann, Kay Prüfer, Janet Kelso, Svante Pääbo, Nick Patterson, and David Reich. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, 507(7492):354–357, 2014.

[64] Julien Y Dutheil, Kasper Munch, Kiwoong Nam, Thomas Mailund, and Mikkel H Schierup. Strong selective sweeps on the X chromosome in the human-chimpanzee ancestor explain its low divergence. *PLoS Genetics*, 11(8):e1005451, 2015.

[65] Mark S Hibbins and Matthew W Hahn. Phylogenomic approaches to detecting and characterizing introgression. *Genetics*, 220(2):iyab173, 2022.

[66] James H Degnan and Noah A Rosenberg. Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2(5):e68, 2006.

[67] Fábio K Mendes and Matthew W Hahn. Why concatenation fails near the anomaly zone. *Systematic Biology*, 67(1):158–169, 2018.

[68] Katy L Simonsen and Gary A Churchill. A Markov chain model of coalescence with recombination. *Theoretical Population Biology*, 52(1):43–59, 1997.

[69] Thomas H Jukes, Charles R Cantor, et al. Evolution of protein molecules. *Mammalian Protein Metabolism*, 3:21–132, 1969.

[70] Jotun Hein, Mikkel Schierup, and Carsten Wiuf. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, USA, 2004.

[71] Asger Hobolth, Arno Siri-Jegousse, and Mogens Bladt. Phase-type distributions in population genetics. *Theoretical Population Biology*, 127:16–32, 2019.

[72] Iker Rivas-González, Lars Nørvang Andersen, and Asger Hobolth. PhaseTypeR: an R package for phase-type distributions in population genetics. *Journal of Open Source Software*, 8(82):5054, 2023.

[73] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[74] John A Nelder and Roger Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

[75] Fuchang Gao and Lixing Han. Implementing the Nelder-Mead simplex algorithm with adaptive parameters. *Computational Optimization and Applications*, 51(1):259–277, 2012.

[76] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

[77] Jim Hester and Jennifer Bryan. *glue: Interpreted String Literals*, 2022. URL `https://CRAN.R-project.org/package=glue`. R package version 1.6.2.

[78] Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. *GGally: Extension to 'ggplot2'*, 2021. URL `https://CRAN.R-project.org/package=GGally`. R package version 2.1.2.

[79] Jeffrey B. Arnold. *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'*, 2021. URL `https://CRAN.R-project.org/package=ggthemes`. R package version 4.2.4.

[80] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

[81] Charles Van Loan. Computing integrals involving the matrix exponential. *IEEE transactions on automatic control*, 23(3):395–404, 1978.

[82] Wolfram Research, Inc. Mathematica, Version 13.2, 2023. URL https://www.wolfram.com/mathematica.

[83] Joseph Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[84] Michael JD Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2):155–162, 1964.

[85] Stephen G Nash. Newton-type minimization via the Lanczos method. *SIAM Journal on Numerical Analysis*, 21(4):770–788, 1984.

[86] Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

[87] Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[88] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I Jordan, et al. Ray: A distributed framework for emerging AI applications. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, pages 561–577, 2018.

[89] The pandas development team. pandas-dev/pandas: Pandas, April 2023. URL https://doi.org/10.5281/zenodo.7857418.

[90] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[91] Siu Kwan Lam, Antoine Pitrou, and Stanley Seibert. Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pages 1–6, 2015.

SUPPORTING INFORMATION CAPTIONS

List of Figures

S1 Species tree of the speciation process of three species (A, B and C). τ_{AB} and τ_{ABC} correspond to the speciation times, measured, backwards in time, as time since the present. The transition rate matrices for the CTMCs are also plotted in the part of the species tree that are used. These are three 2×2 matrices for the one-sequence CTMCs ($\boldsymbol{Q}_{\mathbf{A}}, \boldsymbol{Q}_{\mathbf{B}}, \boldsymbol{Q}_{\mathbf{C}}$), a 15×15 matrix $\boldsymbol{Q}_{\mathbf{AB}}$ for the two-sequence CTMC, and a 203×203 matrix $\boldsymbol{Q}_{\mathbf{ABC}}$ for the three-sequence CTMC.

S2 States of the CTMC of the coalescent-with-recombination process between two sites for a single sequence.

S3 Transition rate matrix for the CTMC of the coalescent-with-recombination process between two sites for a single sequence. A coalescent event happens with rate γ_A (in pink), while a recombination event happens with rate ρ_A (in green).

S4 States of the CTMC of the coalescent-with-recombination process between two sites for two sequences. The states are grouped by $\boldsymbol{\omega}$ sets. Refer to fig. S6 for an explanation of the colors and shapes.

S5 Transition rate matrix for the CTMC of the coalescent-with-recombination process between two sites for two sequences. A coalescent event happens with rate γ_{AB} (in pink), while a recombination event happens with rate ρ_{AB} (in green). The state indices in the left and bottom axes can be consulted in fig. S7. The right and top axes represent sets of states $\boldsymbol{\omega}$, which define the block-like structure of the transition rate matrix. Transitions with rates equal to zero are shown in white.

S6 Diagram showing the different representations of sites used throughout this report. Colored circles represent uncoalesced sites, while triangles represent sites where at least one coalescent has happened. The numerical representation of each of the sites can be used to define the sets of states $\boldsymbol{\omega}$.

S7 States of the CTMC of the coalescent-with-recombination process between two sites for three sequences. The states are grouped by $\boldsymbol{\Omega}$ sets. Refer to fig. S6 for an explanation of the colors and shapes.

S8 Transition rate matrix for the CTMC of the coalescent-with-recombination process between two sites for three sequences, or $\boldsymbol{Q}_{\mathbf{ABC}}$. A coalescent event happens with rate γ_{ABC} (in pink), while a recombination event happens with rate ρ_{ABC} (in green). The state indices in the left and bottom axes can be consulted in fig. S7. The right and top axes represent sets of states $\boldsymbol{\Omega}$, which define the block-like structure of the transition rate matrix. Again, transitions with rates equal to zero are shown in white.

S9 The four possible topologies that can arise during the speciation process of three species (A, B and C).

S10 The hidden states of TRAILS. This figure shows four of the 27 possible hidden states for when $n_{\mathrm{AB}} = n_{\mathrm{ABC}} = 3$. In this figure, and in the rest of the figures of this report, the first coalescent is always plotted as an empty rectangle within an interval, meaning that the coalescent is allowed to happen at any point within the interval. The same goes for the second coalescent, which is in turn represented as a filled rectangle. Note that $t_0 = 0$.

S11 All 27 possible hidden states when $n_{\mathrm{AB}} = n_{\mathrm{ABC}} = 3$. The states are colored by topology.

S12 Paths and transitions. This figure shows 8 of the 729 possible combinations of L and R when $n_{\mathrm{AB}} = n_{\mathrm{ABC}} = 3$. The lowercase omegas ($\omega$) at the left-hand side of each tree show the set of possible states that are allowed at the beginning and end of each time interval in the two-sequence CTMC, while the uppercase omegas (Ω) at the right-hand side show the same for the three-sequence CTMC. The names for the interval limits marked by red horizontal lines can be consulted in fig. S10.

S13 Diagram for an interval of length t with a single coalescent event at time u. a, b, c and d represent nucleotides at different lineages and times.

S14 Diagram for an interval of length t with a two coalescent events at times u and v. a, b, c, d, e and f represent nucleotides at different lineages and times.

S15 The two possible tree configurations for the emission probabilities, with labelled nucleotides at different time points.

S16 Runtime for calculating the transition probability matrix and the log-likelihood for a 1-Mb region of the HCGO alignment using the current implementation of TRAILS. The data points are averaged over 5 runs per combination of $n_{\mathrm{AB}} \in [1, 2, 3, 4, 5]$ and $n_{\mathrm{ABC}} \in [1, 2, 3, 4, 5]$. All these computations were performed on an 8-core MacBook Pro with Intel Core i5 processor and 8 GB or memory.

S17 Posterior decoding of the topology and coalescent times of a region in chromosome 1 showing an excess of V3 topology.

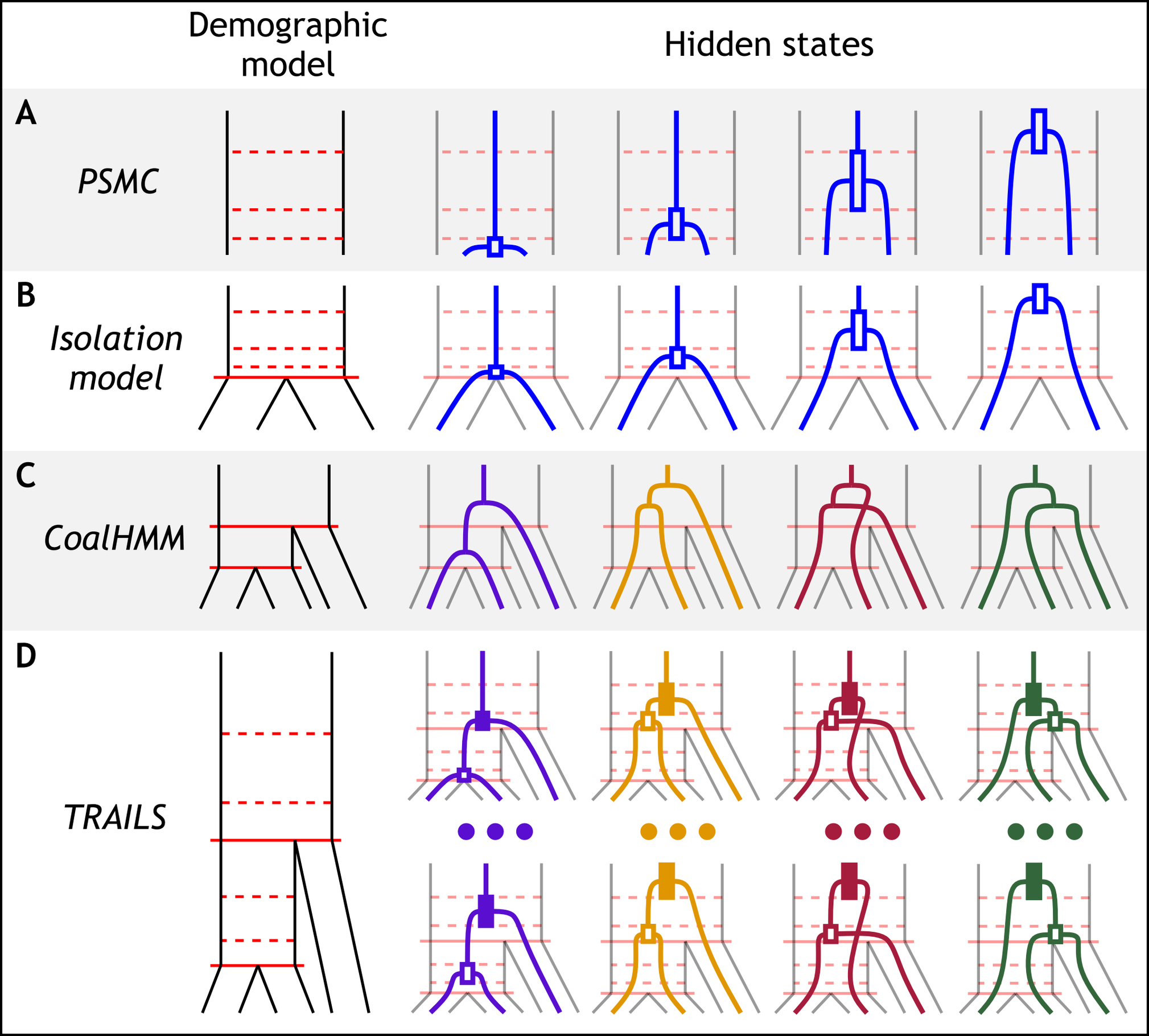
S18 Kernel densities (black) and fitted normal distributions (blue) for the parametric bootstrap replicates. True values are plotted as vertical dashed red lines.

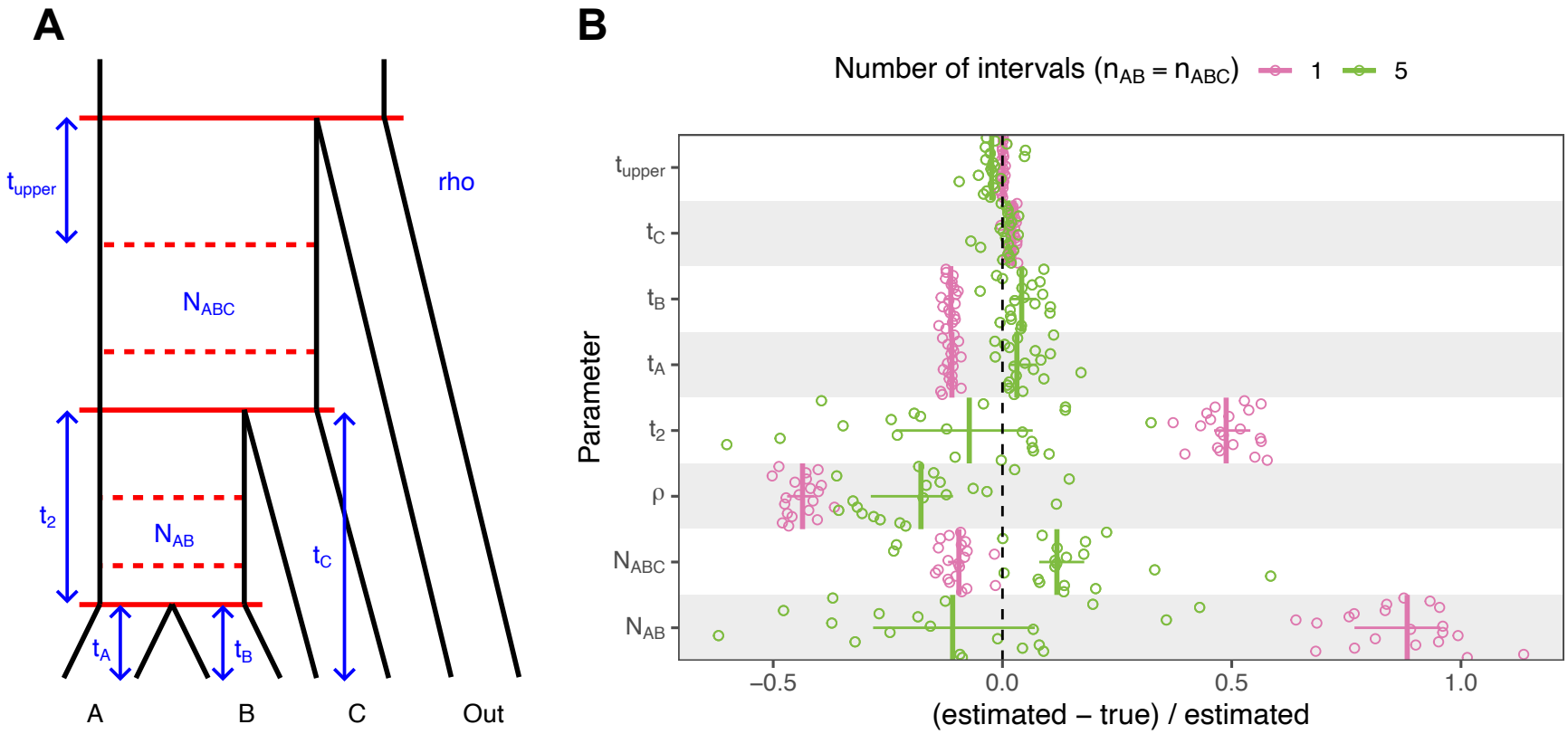
S19 Pairwise correlation analysis for the raw values of the estimates of the parametric bootstrapping. The kernel densities are plotted in the diagonal. The upper diagonal contains correlation coefficients, with significance levels of <0.001 ("***"), <0.01 ("**"), <0.05 ("*"), and <0.1 (".") . The lower diagonal contains pairwise scatter plots, with contour lines corresponding to the density of a fitted bivariate normal distribution.

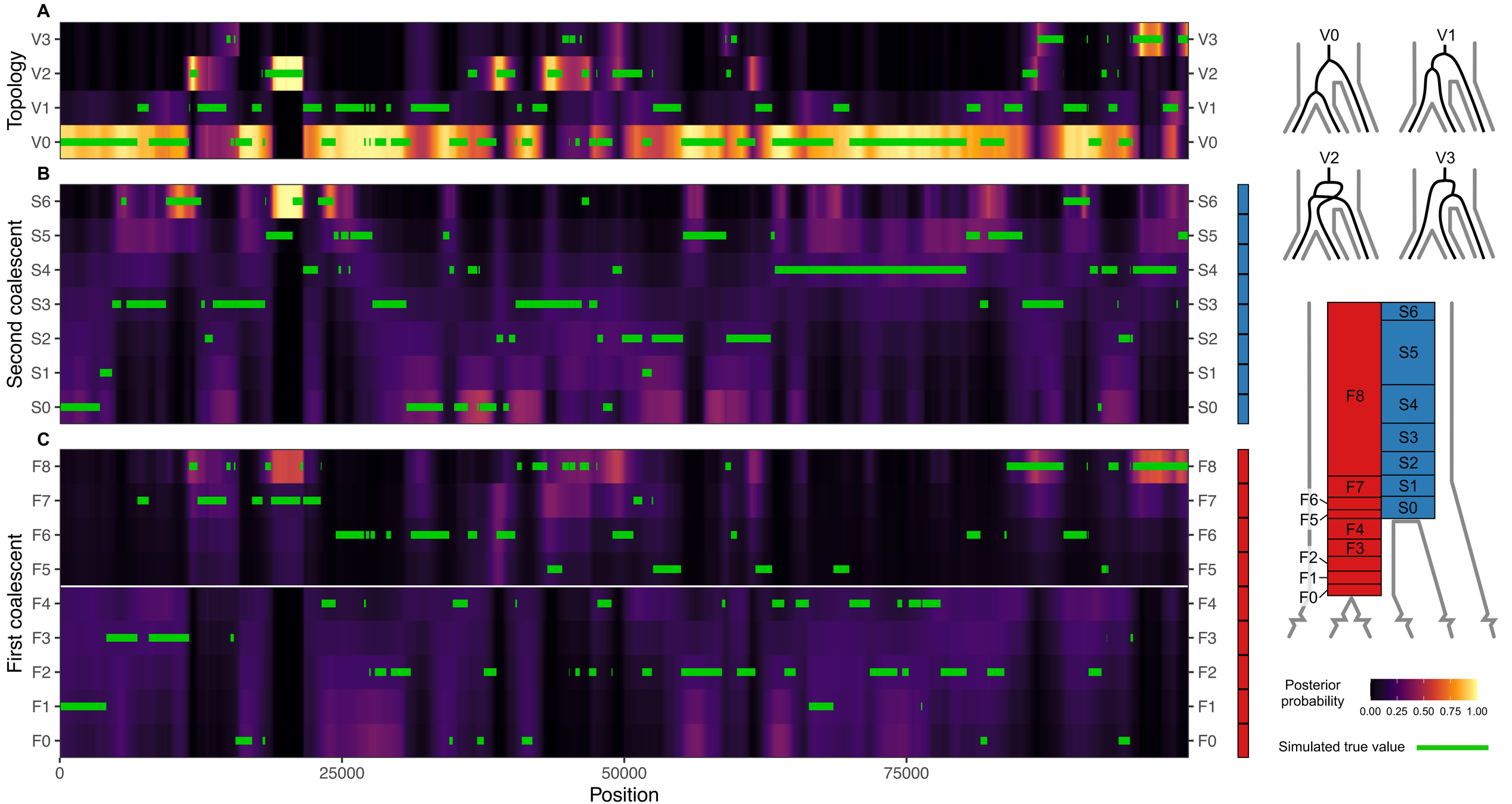
List of Tables

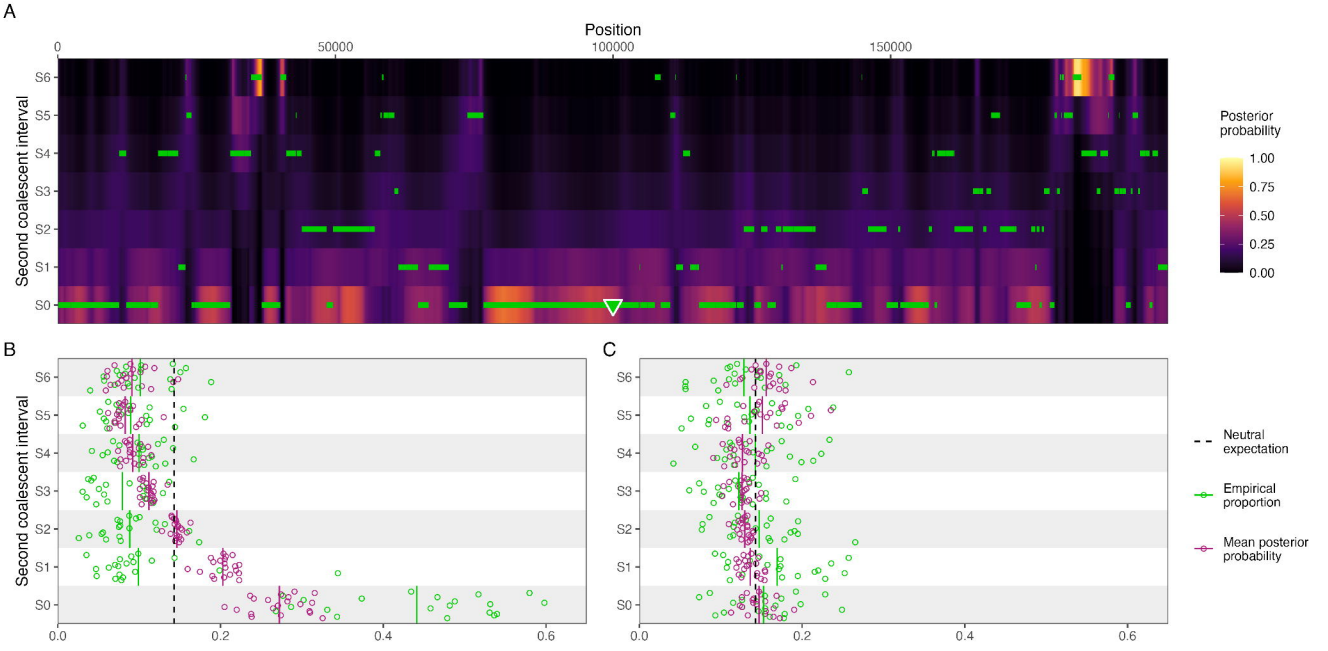
S1 The number of states and proportion of non-zero entries in the rate matrix for a CTMC with n lineages.

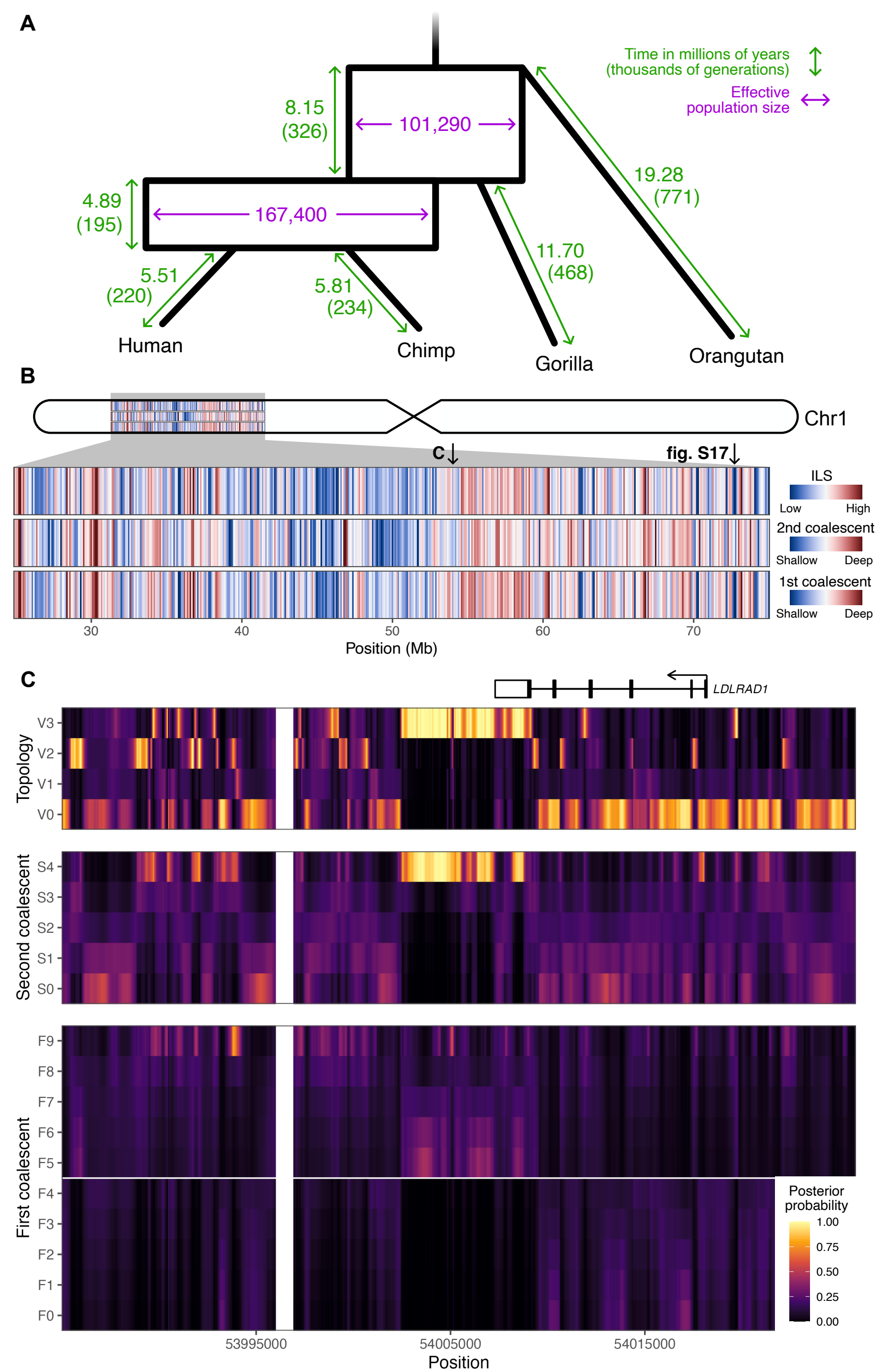
S2 Estimated values, and mean, standard deviation (SD) and 95% confidence intervals (CI) of normal distributions fitted on the parametric bootstrap replicates.


Demographic model
Hidden states
A
PSMC
B
Isolation model
C
CoalHMM
D
TRAILS


A
t_{upper}
rho
N_{ABC}
t_2
N_{AB}
t_C
t_A
t_B
A
B
C
Out
B
Number of intervals ($n_{AB} = n_{ABC}$)
1
5
Parameter
t_{upper}
t_C
t_B
t_A
t_2
ρ
N_{ABC}
N_{AB}
−0.5
0.0
0.5
1.0
(estimated − true) / estimated


A
Topology
V3
V2
V1
V0
B
Second coalescent
S6
S5
S4
S3
S2
S1
S0
C
First coalescent
F8
F7
F6
F5
F4
F3
F2
F1
F0
0
25000
50000
75000
Position
V0
V1
V2
V3
S6
S5
S4
S3
S2
S1
S0
F8
F7
F6
F5
F4
F3
F2
F1
F0
Posterior probability
0.00 0.25 0.50 0.75 1.00
Simulated true value


A
Position
0
50000
100000
150000
Second coalescent interval
S6
S5
S4
S3
S2
S1
S0
Posterior probability
1.00
0.75
0.50
0.25
0.00
B
C
0.0
0.2
0.4
0.6
Neutral expectation
Empirical proportion
Mean posterior probability


A
Time in millions of years
(thousands of generations)
Effective
population size
8.15
(326)
101,290
4.89
(195)
167,400
19.28
(771)
11.70
(468)
5.51
(220)
5.81
(234)
Human
Chimp
Gorilla
Orangutan
B
Chr1
C
fig. S17
ILS
Low
High
2nd coalescent
Shallow
Deep
1st coalescent
Shallow
Deep
30
40
50
60
70
Position (Mb)
C
LDLRAD1
Topology
V3
V2
V1
V0
Second coalescent
S4
S3
S2
S1
S0
First coalescent
F9
F8
F7
F6
F5
F4
F3
F2
F1
F0
Posterior
probability
1.00
0.75
0.50
0.25
0.00
53995000
54005000
54015000
Position